

## Socially aware conversational agents

**Citation for published version (APA):**

Turnhout, van, K. G. (2007). *Socially aware conversational agents*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Industrial Design]. Technische Universiteit Eindhoven. <https://doi.org/10.6100/IR630335>

**DOI:**

[10.6100/IR630335](https://doi.org/10.6100/IR630335)

**Document status and date:**

Published: 01/01/2007

**Document Version:**

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

**Please check the document version of this publication:**

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

**General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.tue.nl/taverne](http://www.tue.nl/taverne)

**Take down policy**

If you believe that this document breaches copyright please contact us at:

[openaccess@tue.nl](mailto:openaccess@tue.nl)

providing details and we will investigate your claim.

# Socially Aware Conversational Agents

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de  
Technische Universiteit Eindhoven, op gezag van de  
Rector Magnificus, prof.dr.ir. C.J. van Duijn, voor een  
commissie aangewezen door het College voor  
Promoties in het openbaar te verdedigen  
op donderdag 1 november 2007 om 16.00 uur

door

Koen Gerardus van Turnhout

geboren te Tilburg

Dit proefschrift is goedgekeurd door de promotoren

prof.dr.ir. J.H. Eggen

Copromotor:  
dr. J.M.B. Terken

The research of this thesis has been carried out as part of the CRIMI project (Creating Robustness in Multimodal Interaction) and has been funded by the Dutch Ministry of Economic Affairs through the Innovation Oriented Programme Man-Machine Interaction (IOP-MMI).

A catalogue record is available from the Eindhoven University of Technology Library. ISBN: 978-90-386-1137-2

# Table of Contents

---

## Table of Contents (3)

---

---

### Chapter 1: An Introduction to Socially Aware Conversational Agents

---

|     |  |    |
|-----|--|----|
| 1.1 | Introduction                                       | 8  |
| 1.2 | A case for socially aware conversational agents    | 8  |
| 1.3 | The domain of socially aware conversational agents | 10 |
| 1.4 | Methodological considerations                      | 14 |
| 1.5 | A socially aware information kiosk                 | 16 |
| 1.6 | Outline of this thesis.                            | 19 |

---

### Chapter 2: Language as Coordinated Action

---

|     |  |    |
|-----|--|----|
| 2.1 | Introduction                                   | 22 |
| 2.2 | Language as Coordinated Action                 | 23 |
| 2.3 | Sequencing Communicative Acts                  | 25 |
| 2.4 | Coordination within a single communicative act | 26 |
| 2.5 | Coordinating with conversational agents        | 30 |
| 2.6 | Conclusions                                    | 32 |

---

### Chapter 3: Sensing Who is Talking to Whom in Dyad-Kiosk Conversation

---

|     |  |    |
|-----|--|----|
| 3.1 | Introduction                           | 36 |
| 3.2 | Experimental Setup                     | 37 |
| 3.3 | Dialog History                         | 41 |
| 3.4 | Coordination within communicative acts | 45 |
| 3.5 | Speaking Styles                        | 51 |
| 3.6 | General Discussion                     | 53 |

---

### Chapter 4: Automatic Addressee Determination

---

|     |                               |    |
|-----|-------------------------------|----|
| 4.1 | Introduction                  | 56 |
| 4.2 | A description of AAD          | 56 |
| 4.3 | Post-utterance classification | 61 |
| 4.4 | Early estimates               | 63 |
| 4.5 | Conclusions and discussion    | 66 |

---

### Chapter 5: Prototyping a Socially Aware Conversational Agent

---

|     |   |    |
|-----|---|----|
| 5.1 | Introduction                            | 70 |
| 5.2 | Ensuring techno-ecological validity     | 70 |
| 5.3 | General Architecture                    | 73 |
| 5.4 | Head Orientation Tracker (HOT)          | 74 |
| 5.5 | Speech Activity Detection (SAD)         | 76 |
| 5.6 | Automatic Addressee Determination (AAD) | 76 |
| 5.7 | Dialog Management                       | 83 |
| 5.8 | MATIS interface                         | 86 |

---

## Chapter 6: (How) Should Socially Aware Conversational Agents Show Users what they are Doing?

---

|     |                                      |     |
|-----|--------------------------------------|-----|
| 6.1 | Introduction                         | 88  |
| 6.2 | Three Design Questions               | 90  |
| 6.3 | Conceptual Design                    | 95  |
| 6.4 | Design intervention study - approach | 101 |
| 6.5 | Six design interventions             | 107 |
| 6.6 | Conclusions and Discussion           | 135 |

---

## Color Plates (139)

---

## Chapter 7: Towards the Coordinated Development of Socially Aware Conversational Agents

---

|     |  |     |
|-----|--|-----|
| 7.1 | Introduction                                 | 144 |
| 7.2 | Contributions and connections                | 144 |
| 7.3 | (Technological) possibilities and challenges | 149 |
| 7.4 | New challenges for developing SACA's         | 151 |
| 7.5 | A final note                                 | 153 |

---

## Appendix A: Scenarios from the Experiments in Chapters 3 and 6

---

|     |                                   |     |
|-----|-----------------------------------|-----|
| A.1 | Original scenarios (in Dutch)     | 154 |
| A.2 | Translated scenarios (in English) | 163 |

---

## Appendix B: Utterance length in 'logarithmic time'

---

|     |                                 |     |
|-----|---------------------------------|-----|
| B.1 | Utterance Length in 'Log' space | 171 |
|-----|---------------------------------|-----|

---

**Appendix C: The SASSI Questionnaire**

---

- |     |                                      |     |
|-----|--------------------------------------|-----|
| C.1 | Original Items and Factors           | 172 |
| C.2 | Dutch translation of the SASSI items | 174 |

---

**Appendix D: Results of the Quantitative Study of Chapter 6**

---

- |      |                     |     |
|------|---------------------|-----|
| D.1  | Introduction        | 176 |
| D.2. | SASSI               | 176 |
| D.3  | Behavioral Measures | 183 |

---

**Bibliography (185)**

---

---

**(Short) Curriculum Vitae (193)**

---

---

**Summary (195)**

---

---

**Samenvatting (198)**

---

---

**Acknowledgements (201)**

---

# Chapter I: An Introduction to Socially Aware Conversational Agents

*In this chapter we introduce the notion of socially aware conversational agents, our approach for developing them and the outline of this thesis. Socially aware conversational agents support speech as input modality and behave appropriately in a social context, based on an awareness of this context. To contribute to the development of socially aware conversational agents we present a design case: we try to adapt an existing speech-centric multimodal interface for the case of shared use. We approach this design case from a threefold perspective: a social psychological, a systems engineering and an interaction design perspective. With this design case, we aim to explore the (technological) possibilities for developing socially aware conversational agents, to collect the pieces of specialized knowledge needed to develop this type of solution and to uncover interdisciplinary challenges for the domain of socially aware conversational agents.*



---

## 1.1 Introduction

---

This thesis focuses on a solution for shared use of speech-centric multimodal interfaces. The difficulty is this: if there are multiple users, interacting with each other and with a multimodal interface, they will use speech, both to communicate with each other and with the system. Since speech interfaces are designed for single user situations they assume that all incoming speech is intended for them and they respond accordingly. A solution for this problem would be to build systems that have a sense of the ongoing social context. In particular, they would have to keep track of who is talking to whom, and they would need to have a way to use this information in the interaction with multiple users. In this thesis, we call such systems *socially aware conversational agents*. We will describe the design of such a socially aware conversational agent. This serves two goals. First, we aim to explore the (technological) possibilities making conversational agents socially aware. Second, we aim at collecting the pieces of specialized knowledge needed to develop this type of solution and to uncover interdisciplinary challenges for the domain of socially aware conversational agents.

This introductory chapter is organized as follows. First, we expand the notion of socially aware conversational agents. In section 2, we briefly address the question *why* we would want to develop socially aware conversational agents at all. Answering this question leads into constituting requirements: we can say *what* socially aware conversational agents *should* be able to *do*. We do so in section 3, where we will also discuss three (future) scenarios for socially aware conversational agents. The scenarios serve as a thought exercise to give us an idea of the research challenges we face. Having sketched the domain of socially aware conversational agents, we turn to the specific contribution we intend to deliver to the domain. In section 4, we explain our methodological approach and contrast it to other possible approaches. We then turn to the design case we take up in this thesis and its a priori constraints in section 5. We end the chapter, in section 6, with an outline of the rest of this thesis.

---

## 1.2 A case for socially aware conversational agents

---

The need for information is ubiquitous and people often share information needs when they are together. Travelers in public transport need information on departure and arrival times of planes, trains and busses. Museum visitors may want background information on whatever there is on display. Students, collaborating with electronic whiteboards could

benefit from access to materials they have used or produced earlier on. These people could possibly be supported with automated information services, if these systems have the flexibility to accommodate very specific information needs in a potentially large database of information. On a day-to-day basis, language technology proves itself for doing just that. Whether they are aware of it or not, a billion people in the world (NRC 2007, pp 6) use web browsers, employing language technology, for finding information on the internet - the largest information database in the world. It hardly comes as a surprise then that, in designing automated information services for the people in the examples, we opt for language technology. The question becomes whether the way we currently have access to this technology: through screen, keyboard, and mouse, is suitable for them.

Dourish (2001) highlights some serious disadvantages of screen, keyboard and mouse. In ‘where the action is; the foundations of embodied interaction’ he writes (p.27):

‘Interaction with screen and keyboard, for instance, tends to demand our direct attention; we have to look at the screen to see what we’re doing, which involves looking away from whatever other elements are in our environment, including other people. Interaction with the keyboard requires both of our hands. The computer sits by the desk and ties us to the desk, too.’

Thus, in designing for social situations, such as the ones described - where people are together and have shared information needs - we may opt for the alternative way to access language technology: speech. Speech is currently not as established as language technology. Although speech interaction with computers has been popularized by science fiction writers at least as early as Kubrick’s movie ‘A space odyssey 2001’ (Kubrick, 1968), and scientific efforts to enable machines to recognize speech date back to 1949 (Dreyfus - Graf, 1949; Davis et al. 1952) <sup>1</sup>, it has turned out to be hard to reach a performance that enables widespread application of this technology. However, the age of speech interaction may be dawning. At least special purpose speech applications, such as dictation systems, telephone based information services, and speech interfaces for the car environment have reached the market. Note that speech input is currently most employed in situations where typing is inconvenient. Following Dourish’s observations, social situations and shared use - where people interact both with each other and with a system – should actually be listed among those situations.

However, applying speech input technology in social situations is not straightforward. If there are multiple users interacting with one another, they will use speech to talk to

---

<sup>1</sup> Strictly speaking, Dreyfus-Graf did not create a speech *recognizer* but a speech *transcription device*, dubbed the sonograph. The Bell Labs’ device reported in Davis et. al (1952), could be more accurately described as a *recognizer*.

each other. How would a system know how it has to act on what it is hearing? A simple, and currently the only, solution for this problem is to enforce an interaction protocol on the user. We may require users to push a button (push to talk), tap a field on the screen (tap 'n talk) or to use a specific cue word (for example. 'computer') to indicate they are addressing the computer. Then, and only then, the computer will try to recognize the incoming speech and respond. Although this is an acceptable solution, certainly in the absence of alternatives, there is evidence it is not the most natural solution for users. Maglio, Matlock, Campbell, Zhai, & Smith (2000), for example put people in a fake 'intelligent' room and had them command office appliances like a fax and a telephone by speech, without too much instruction on how to do so. They found people to use the name of the appliance they were addressing only sparsely, while they did tend to look at it, just before or after the command they issued. Brummit and Cadiz (2001), present similar results for a study where people were asked to manage lights in an 'intelligent' home environment. Seemingly, inexperienced users, confronted with speech enabled technology do not bother to use the name of the appliance; at least not when it works without such explicit commands. This motivates the search for alternative solutions. How can we equip appliances with the possibility to detect who is talking to whom automatically, *without* enforcing a specific interaction protocol on the user? And, if we are able to do so, how should these appliances employ this information in their interaction with multiple users?

---

### 1.3 The domain of socially aware conversational agents

---

#### 1.3.1 A definition

In section 1.2 we ended with two questions: how can we build appliances with the possibility to detect who is talking to whom automatically and how should these appliances employ this information in the interaction with users? Or, more generally, how can we make conversational agents *socially aware*? Where we take socially aware conversational agents to be able to:

1. Accept speech as a way to access the functionality *of* and information *in* the system.
2. Be aware of the social context; specifically,
  - to assess whether they are used by groups.
  - to assess to what extent incoming speech is intended for them.
3. Behave appropriately for the social context, based on the assessments made in (2).

To illustrate the challenges involved in developing socially aware interfaces we will discuss some scenarios where these interfaces do their work.

### 1.3.2 Scenario I

This first scenario shows a limitation of a current commercial speech application.

A secretary, dictating a letter in the office is interrupted by a co-worker.

The office environment is a social environment. However, dictation systems are designed for single person-single system interaction. Current dictation applications would capture the speech of the secretary during an interruption; thus forcing the secretary to switch off the microphone each time she is interrupted. Research has shown that it is possible to detect interruptions with fairly simple sensors (Horvitz & Apacible, 2003). A socially aware conversational agent should be able to assess whether interruptions occur, and might prevent the application from capturing irrelevant speech.



**Figure 1: (Scenario I) a secretary dictating a letter in the office is interrupted by a co-worker**

Two technological challenges of dictation systems are error-free and timely detection of on- and offsets of interruptions. An agent that fails to detect interruptions as soon as they occur or shuts off the dictation application when there is no interruption, will be more cumbersome to handle than a manual switch for the microphone. Clever interaction design could overcome some of the mistakes of an agent. For example, in the presence of errors it may be more correct to shut off critical tasks (for example ‘delete’) and enable easy undo (for example deletion of recorded side conversations) over disabling full functionality. However, it will remain a challenge to keep the agent’s actions transparent, and to ensure that the secretary feels in control of the system, although it makes errors.

This scenario shows that the technological challenge is to deliver error-free and timely detection of interruptions and the interaction design challenge is to reduce the cost of system errors for the user. These challenges are mutually dependent: the options for interaction design depend on system performance, and the required system performance depends on the desirable interaction design.

### 1.3.3 Scenario 2

This second scenario introduces the design case we will focus on in the rest of this thesis:

Karen and Bob make joint use of a speech centric multimodal travel information service on an information kiosk.

We will return to this scenario in much more detail later this chapter. Here we just treat it like the other scenarios as a free thought exercise to explore the types of challenges involved in developing socially aware conversational agents. The case of shared use presents greater research challenges than the first scenario. All speech of Karen and Bob could principally be intended for both the system and the partner. The interaction with the information kiosk can be interleaved with discussions among Karen and Bob. Therefore, the agent needs to assert to what extent incoming speech is intended for either the system or a partner on an utterance by utterance basis. This is a difficult task that has only got cursory research attention so far (for example. Vertegaal, Slagter, Van der Veer, & Nijholt, 2001; Bakx, Van Turnhout, & Terken, 2003; Katzenmaier, Stiefelhagen, & Schultz, 2004; Van Turnhout, Terken & Eggen, 2005; Jovanović, Op den Akker, & Nijholt, 2006). These studies suggest that we can use a range of linguistic and non-verbal cues to infer the addressee of an utterance, but that no single cue is decisive.



**Figure 2: (Scenario 2) Karen and Bob make Joint use of a travel information service on an information kiosk**

The interaction design challenges can take up different forms. The easiest way to use the agent's assertions about the addressee of an utterance is similar to what a push to talk button would do. The system will only react on speech that is asserted to be intended for the system. But it may well be desirable that travel information services are much more "conversational" than dictation applications (Sturm, 2005 pp 47-71; Den Os, Boves, Rossignol, Ten Bosch, & Vuurpijl, 2005). If so, the challenge is to design suitable system responses in which content and timing are combined in a turn-taking protocol (Thórisson, 1996).

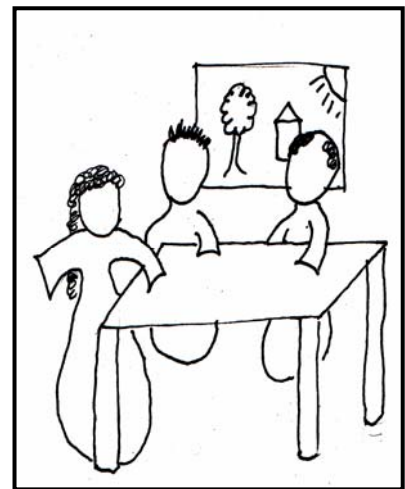
This scenario shows that the challenge of delivering appropriate responses changes shape when the complexity of the interaction design increases. The question whether a system response is appropriate is different for a tap ‘n talk implementation than for a conversational system. Again we see a mutual dependency between the quality of the agent’s assertions, the possibilities for interaction design and the corresponding research questions. Both the technological challenges and the interaction design challenges increase further if we allow for larger groups to use the system. Surely agents in public environments would have to be able to deal with larger groups. We clarify this problem of scalability further in scenario 3.

### 1.3.4 Scenario 3

This scenario presents the largest technological and design challenges in developing socially aware conversational agents.

A group of co-workers is brainstorming about the functionality of a new system using an intelligent brainstorm support room.

In a group meeting a lot of speech is not intended for the system, and the amount of users is not specified upfront. Also, the system may take several roles. For example, a system may serve as facilitator, organizing the structure of the brainstorm, or it may act as inspirator by displaying inspirational pictures about the discussion. It may act as ‘note taker’ supporting the participants in making references to earlier discussed materials (Van Gelder, Van Peer, & Aliakseyeu, 2005), or it could be a social worker supporting equal participation and group interaction (Van Turnhout, Malchanau, Disaro & Markopoulos 2002; Kulyk, Wang, & Terken, 2005; Danninger et al., 2005). What an agent may do with the information about whom is talking to whom strongly depends on the role of the application in the meeting. In other words, at this level of specification the challenges for interaction design are hard to predict.



**Figure 3: A group of coworkers is brainstorming about the functionality of a new system using a brainstorm support room**

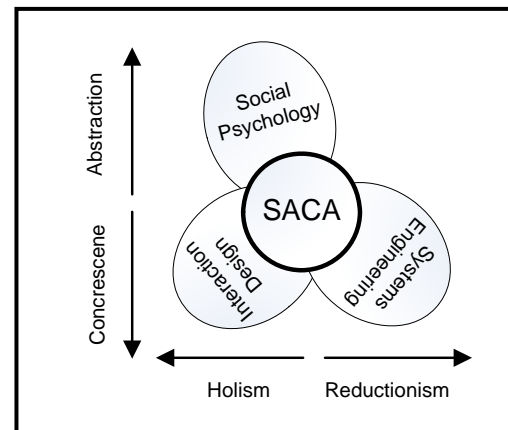
The meeting scenario has received notable scholarly attention. There are many ideas for possible support (see for example: Danninger et al., 2005), there is technology that helps understand who is talking to whom (Stiefelhagen, 2002) and some first attempts to infer who is talking to whom (Jovanović, 2007). However, the scenario shows that bringing those bits together will be far from trivial.

## 1.4 Methodological considerations

In the previous section, we have seen that the challenges for designing socially aware conversational agents can take up different forms, depending on what they are supposed to do. Scenarios 2 and 3 show that, in a way, the *interaction design* of the specific agent frames all other research questions that may be relevant. However, as is obvious in scenarios 1 and 2, the choices interaction designers can make, depend in turn on the technological possibilities. If socially aware conversational agents are to sense who is talking to whom automatically, *system engineers* may provide information on what can be sensed and interpreted, with the state-of-the-art technology. Both interaction designers and system engineers are limited in turn by human behavior in the social context. To describe human behavior and its semantics is traditionally the domain of *social psychology*. Designers may turn to this body of knowledge to learn about which system behaviors are suitable, engineers about what the relevant behaviors to sense may be and what their semantics are. As a collective effort, the development of socially aware conversational agents then takes the coordinated collaboration of three traditionally distinct disciplines (see figure 4, adapted from Bartneck and Rauterberg (2007)<sup>2</sup>).

Within this thesis, we adopt an interdisciplinary, integrative approach with explicit attention to the standards of the contributing disciplines. Although triangulation - combining research strategies from different disciplines - is widely recommended for HCI research (see: Mackay & Fayard, 1997) it is worthwhile to consider the trade-offs we face for this specific case in some detail. For this purpose we introduce two distinctions that relate the contributing disciplines to each other.

First, Rauterberg (2000, 2006), makes a distinction between scientific disciplines (such as social psychology) and engineering disciplines (such as interaction design and systems engineering), based on the extent to which their results (models, theories or artifacts) take part in the domain of description. According to Rauterberg, scientific disciplines deliver



**Figure 4: Three traditionally distinct disciplines contribute to the development of socially aware conversational agents. A (simplified) depiction of the relation of the disciplines to each other is given.**

<sup>2</sup> The original picture differentiates these three disciplines on 3 barriers, being: (1) knowledge representation (implicit or explicit logic), (2) view on reality (understanding or transforming) and (3) main focus (technology or humans). This has its merits, but for the context of this thesis, the two dimensional model presented here suffices.

theories *apart* from the domain they describe eventually intended to *predict* the behavior of the world at the expense of *abstraction*. For example Newtonian mechanics are able to predict the behavior of ice-skates accurately, but Newtonian mechanics do not take part in speed skating or tell us in a straightforward manner how to make better ice-skates. In contrast, engineering disciplines deliver models and artifacts that claim to *change* the world of description, resulting in models and new artifacts (conrescence<sup>3</sup>) at the cost of *predictive power*. For example the invention of clap skates changed the speed skating sports but its design descriptions tell us little about future possibilities for improving ice skating. This distinction between disciplines, chiefly aiming at either conrescence or abstraction is depicted vertically in Figure 4.

Second, Rauterberg (2000, 2006) observes a difference between disciplines with a ‘weak implicit’ logic such as the social sciences and disciplines with a ‘strong explicit’ logic such as the natural sciences and systems engineering. Rather than dividing these disciplines on the basis of the type of knowledge involved (Rauterberg, 2000; Rauterberg 2006; Bartneck & Rauterberg, 2007) or epistemological roots (Dorst, 1997), we consider the strategy for dealing with the complexity of the real world as most important for this thesis. The horizontal dimension in figure 4 contrasts holistic disciplines that try to preserve the complexity of the real world at the expense of using implicit logic, to those following a reductionistic strategy, tackling partial problems with explicit logic, at the expense of reducing the complexity found in the real world. Although this two-dimensional model is a rather coarse simplification of the diversity of the methodological approaches found in HCI research, it does provide us with a conceptual framework to discuss the trade-offs we need to make here and to evaluate our results.

As a whole, the work in this thesis can best be seen as conrescent and reductionistic. We will design and implement a *concrete* example of a socially aware conversational agent, using an artificial laboratory case setting of *limited complexity*. As a result, our results do not straightforwardly generalize to other socially aware conversational agents and we will not deliver an agent that can face the complexity of the real world. Rather, we aim at exploring (technological) possibilities for developing socially aware conversational agents. We compensate these weaknesses by dividing the design into three parts, corresponding with the three contributing disciplines. In chapters 2 and 3, we focus mainly on the social psychological perspective, trying to link our concrete example to more abstract social

---

<sup>3</sup> The word conrescence is used in biology for ‘a growing together, as of tissue or embryonic part’ (The American Heritage® Stedman's Medical Dictionary), we follow Rauterberg in using it as antonym to abstraction.



psychological theory. In chapters 4 and 5, we will take a systems engineering perspective with its reductionism and concrescence. In chapter 6, we take up an interaction design perspective, where we try to get back to the complexity inherent in a working prototype, and preserve this complexity throughout the study. As a result, the three subtasks keep their own methodological face, and deliver different types of results, that we try to relate to each other in the closing chapter. This facilitates our second design goal, to uncover interdisciplinary challenges involved in designing socially aware conversational agents.

---

## 1.5 A socially aware information kiosk

---

In this thesis, we focus on the case of shared use of an information kiosk. We have given this case a cursory introduction by discussing it in scenario 2 of section 1.3. Although some large research projects such as MASK (for example. Life et al., 1996; Bennacef, Bonneau-Maynard, Gauvain, Lamel, & Minker, 1994) and SMART (for example. Reigh, Loughlin, & Waters, 1997) have addressed multimodal interfaces for public information displays, the case of shared use has received little attention in these projects.

In section 4, we introduced our specific approach: adopting a threefold perspective for this design. We divided the design into three subtasks. We adopt a social psychological perspective *to identify the relevant human behaviours* socially aware conversational agents need to attend to, in order sense who is talking to whom. We adopt a systems engineering perspective to determine the *addressee of each utterance*. We adopt an interaction design perspective to *design the interaction (of socially aware conversational agents) with users*. In our treatment of each subtask we try to deliver a contribution both to the underlying discipline and to our own design. Clearly, in order to be able to attend to all three perspectives we have to restrict ourselves within each subtask of the design in breadth and depth. So, we adopt several prior constraints that try to ensure maximum added value for the communities we serve, with minimal effort. Following the disciplines we identified, we use three limiting frames: a social frame, a technological frame, and an interaction design frame.

### 1.5.1 Social Frame

Within this thesis, we will study pairs of users, or more formally: dyads, interacting with an information kiosk in a laboratory setting. The focus on the dyad has two reasons. First, in a small observation study (Bakx, 2002) we have seen that existing interactive information displays in museums are visited by single persons in about 60% of the cases,

by dyads in about 35% of the cases and only in 5% of the cases by larger groups. We exclude the case of single persons because it is out of the primary scope –shared use– of this thesis and we exclude larger groups because they form such a small portion of actual use. Second, in social psychology, especially in the study of non-verbal communication, the dyad has been a popular object of investigation. Presumably because human interactions can become increasingly complex with an increasing number of people involved. In a way, the dyad is the social nucleus. For us, drawing from social psychology, the advantage of studying dyads is that we can make optimal use of existing studies. Using a laboratory setting has obvious practical advantages over a real world setting. But it also enables us to focus on the problem of addressing rather than adopting the wider scope of overhearing. The term overhearing (or ‘off-talk’) refers to all speech that may be captured in the microphone that is not intended for the system including speech from non-users in the vicinity of the system or self-directed speech from users (Lunsford, Oviatt, & Coulston, 2005; Lunsford, Oviatt, & Arthur, 2006; Opperman, Schiel, Steininger, & Beringer, 2001; Siepmann, Batliner, & Opperman, 2001). Since off-talk and self directed speech form a class of speech with decreased relevance for the system, they are technologically related, but the social psychological origins are quite different<sup>4</sup>.

### 1.5.2 Technological Frame

In section 1, we have stated the goal of demonstrating that conversational agents can be made socially aware with current<sup>5</sup> state of the art technology. Now that we have introduced the specific design case of this thesis, we can survey what technology we need. We will look at two aspects: the sensing technology and the interpretation technology.

We have chosen to *sense* only low-level non-linguistic cues. There are three arguments that justify this choice. First, we connect to developments in systems engineering where perceptive components are being developed to sense the presence, proximity, location, orientation, and gestures of people (for example. Danninger et al., 2005). Second, these cues can be sensed *independent of* the speech recognition. Therefore, we can obtain valuable information early (before speech recognition) and this information may be used for mutual disambiguation with both signal-level and semantic fusion (Oviatt & Cohen, 2000). Third, non-verbal cues play an important role for coordination of execution and

---

<sup>4</sup> For example, the amount of self-directed speech plays an important role in developmental psychology. A topic beyond the scope of this thesis.

<sup>5</sup> ‘Current’ is of course a relative term within a project that stretches multiple years. We have chosen to focus on technology that was available in research labs at the start of the project.

attention in human-human communication, as we will explore in more depth in chapters 2 and 3.

In chapter 2, we will see that eye gaze may serve as an important signal for detecting who is talking to whom. However, in our case, people are free to move around, which makes precise eye gaze measurements cumbersome. Luckily, head orientation is easier to measure and can replace eye gaze measurements in several situations (Stiefelhagen & Zhu, 2002). In chapter 3, we will show this is also the case in our situation. A second source of information that we can employ before speech recognition comes into play is prosody. We will not focus on advanced analysis of (prosodic) pitch or other prosodic features, but we will restrict ourselves to on-off patterns of speech. This enables us to detect utterance length, a feature related to the fact that humans use different speaking styles when talking to machines than when talking to other humans (Oviatt, 2000). Finally we consider using dialog events, such as questions from the system as source of information, since those events are both easy to sense and may have a definite effect on the dialog.

These three sources of information: head orientation, prosody and dialog events, need to be *interpreted* in the light of the question ‘who is talking to whom’. There are many parametric and non-parametric techniques available to link sensor input to such outcomes. Popular pattern classification techniques for this type of problem are Bayesian clustering, (multilayer) neural networks and hidden Markov models (Duda, Peter, & Stork 2001). Given the limited amount of sensor information that we collect, we choose a simple clustering technique: Bayesian clustering. Obviously, this is the first technique to try, but in addition, evaluating more advanced techniques would put high requirements on the amount of data we need to collect.

### 1.5.3 Interaction Design Frame

In the interaction design of the kiosk, we restrict ourselves to adapting an existing multimodal interface for providing train table information: MATIS (for example. Sturm, 2005, pp 32-34; Sturm, Bakx, Cranen, & Terken, 2002). The form on the screen contains fields corresponding to the parameters that MATIS needs, to compose a travel advice (figure 5). MATIS takes the initiative by announcing itself and asking the user to specify the information it needs. Users can take over the initiative by tapping fields on the screen. This



Figure 5: A screenshot of the original MATIS system

activates the microphone for speech input (see figure 5). Station names, times and other dates than today or tomorrow require speech input. Users can fill the other fields: today, tomorrow, departure or arrival time by pressing buttons. Whenever a field is filled, the system either, keeps, or takes back, the initiative by prompting for the next empty field. Although MATIS might not have the flexibility or complexity that would justify equipping it with social awareness, it does provide us with a convenient starting point to do research on the interaction design of such systems.

In chapter 6, where we take up the interaction design challenges, we will reconsider in particular the visual design of the interface. Although we need to implement a turn-taking protocol, we do not feel the spoken dialog prompts of the system need adaptation for our context. Neither do we intend to adjust the mixed initiative character of the interface.

---

### 1.6 Outline of this thesis

---

In this chapter we introduced the design case of this thesis and the approach we will adopt.

In chapter 2, we will present a survey of communication theory, to explore to what extent insights about human-human communication can guide the development of socially aware conversational agents. This results in 3 broad strategies for detecting who is talking to whom across a range of social settings. We evaluate these strategies in chapter 3. Here we zoom in on our specific design case and show how we can operationalize these strategies for our social and technological frame. We describe an empirical study to see whether we are able to employ these strategies.

Following this, in chapter 4 we will try to combine the sources of information we identified to be useful in chapter 3 into a stochastic model, a classifier that is able to assert who is talking to whom. This classifier is evaluated with ROC curves. This classifier needs to be implemented and integrated into a prototype that is able to do speech activity detection, head orientation tracking and dialog management. For practical reasons, part of the dialog management is implemented through a wizard of Oz setup. The implementation of this prototype is described in chapter 5.

In chapter 6, we turn to the interaction design of our information kiosk. Here, we will focus on errors, as our classifier turns out to make quite a few of them. We explore different ways to provide feedback on the inferences of the classifier that we built in chapter 4 and evaluate this feedback with users. As a result, we gain insights in the expectations of users when confronted with this technology and in the role feedback

plays in shaping these expectations. This in turn leads to lessons and recommendations for the interaction design of socially aware conversational agents.

We end the thesis in chapter 7 by zooming out again. Here, we place our efforts in the light of the development of socially aware conversational agents in general. We critically evaluate the extent to which we reached our goals and we identify future challenges for the development of socially aware conversational agents.

## Chapter 2: Language as Coordinated Action

*In this chapter we argue that a perspective of language as coordinated action can guide the design of conversational agents. We develop constructs taken from communication theory, most notably Herbert Clark's theory of conversational grounding, and relate them to empirical work on the fine mechanics of conversation. From these insights we try to see how a conversational agent could possibly know who is talking to whom, resulting in three strategies that we explore further in chapters 3 and 4. Once we, in chapter 4, arrive at a quantitative estimate about who is talking to whom, a more difficult question remains. What should we do with this information in the interface? In this chapter, we develop an argument for early feedback, and arguments for exploring a wider range of possible options than the embodied conversational agents that are currently reported on in the HCI literature. These arguments form the basis for the work in chapters 4, 5 and 6.*

---

## 2.1 Introduction

---

In chapter 1 we have introduced the design of socially aware conversational agents as a threefold task: identifying the relevant behaviors and their semantics, developing the sensing and interpretation technology, and designing the appropriate system behavior. We also said that these tasks fall within traditionally distinct scientific disciplines. Now, in order to orchestrate our efforts across these disciplines, we will benefit from a common set of theoretical constructs. Despite the methodologically distinct approaches we may adopt in the different phases of this design case, we can deliver integrative knowledge if we manage to formulate a single theoretical framework to which we can relate the three subtasks of the design. This is what we set out to do in this chapter.

The central question in this chapter is ‘who is talking to whom’, but there are two sides to this question that are relevant for our case. For developing the sensing and interpretation technology we would like to know *how to detect who is talking to whom* and for designing the interface we would like to know *how should we use this information in the user interface*. There is a fairly limited amount of work addressing the question of how to detect ‘who is talking to whom’ automatically. In the context of meetings there is work from Vertegaal et al. (2001) and Jovanović et al (2006). And in the context of mixed human-human-system interaction there are studies by Katzenmaier et. al (2004), and Traum (2004). The second question: how to use this information about users in the interface has received even less attention. This does not mean there is no theoretical guidance to develop socially aware conversational agents. We can turn to more general work on the *fine mechanics* of conversation among humans. There are answers to the general question: ‘how do people organize everyday conversation on a signal to signal basis?’. From those we work our way back to the two specific questions of this chapter.

Two branches of empirical work are of particular relevance for insights into the fine mechanics of conversation. Social psychologists have studied non-verbal signals such as eye gaze, facial expressions, gestures and interpersonal distance, and have tried to relate the use of these signals to traditional social psychological themes such as power, intimacy, sex and interpersonal relationships (see: Knapp & Hall, 2002). Conversation analysts have tried to identify regularities in naturally occurring speech and explain those (inductively and qualitatively) in terms of the participants’ sense making in the conversation (see: Hutchby & Wooffitt, 1998). Parallel to these empirical lines of research a branch of thinking in philosophy emerged examining language *use*. Within

this line of thinking language is studied as a means to achieve goals, a form of action, rather than an abstract symbolic system. The most influential theories are probably speech act theory (Austin, 1962; Searle, 1969) and Grice's theory of conversational implicature (Grice, 1957). More recently these philosophical and empirical lines of research have merged into several integrative theories highlighting aspects of language use (see: Holtgraves, 2002), of which Herbert Clark's theory of conversational grounding (Clark, 1996) is closest to our needs.

Central to Clark's discussion is the notion of language use as coordinated (or joint) action, which we discuss in the next paragraph. After this we examine in more detail how people achieve this coordination across a series of communicative acts in section 2.3 and within single communicative acts in section 2.4. After this discussion of the way humans communicate with each other, in section 2.5 we look at work examining what might happen when we put computers in the loop. Finally, we summarize the most important answers to the two central questions of this chapter in section 2.6.

---

## 2.2 Language as Coordinated Action

---

In his book *Using Language*, Herbert Clark (1996) introduces his theory of language use as *joint action*. Clark's notion of language use encompasses all possible signals people can use to communicate with each other, including non-verbal signals; in this broad view the term language use is roughly equitable with the term communication. The notion of language as *action* has been introduced by speech act theorists. Searle (1979 p.29), for example, claims: "there are a rather limited number of basic things we can do with language". Some examples are: directives (to get somebody to do something), assertives (to state the state of the world), promises (to commit to do something), declaratives (to change the state of the world, for example to declare war) and expressives (to express feelings, for example thanks, apologies). Other taxonomies have been proposed as well (see for example: Bunt, 2000), but within all these theories language is treated as a means to achieve goals rather than an abstract symbolic system. Clark takes this idea one step further, by emphasizing the *joint* nature of such actions.

A joint action is an action that cannot be accomplished by a single individual. It requires two or more individuals to cooperate and coordinate to complete a joint action. The prototypical example is a handshake (Goodwin, 1981). Surely a handshake can be decomposed into individual actions: both individuals offer their own hand, grab the other hand, shake, and release. But if we want to explain the event as a whole, two people shaking hands, we need to consider that each individual has to perform his



actions in such a way that they ‘fit’ into the actions of the other. Both participants have to *anticipate*, *monitor* and *react to* the actions of the other: it involves the coordination of the two participants. Hence, in emphasizing the joint nature of language use Clark poses it as a coordination problem.

In order to see how language use can be described as a coordination problem let us turn to Clark’s layered protocol for communicative acts. Clark argues that communicative acts can be decomposed into at least four ‘levels of action’ composing an ‘action ladder’. Each level requires coordination between the speaker and addressee: execution and attention, presentation and identification, signal and recognition, and proposal and uptake (See Table 1). The different levels in this ladder are not independent: each level can only be accomplished if all the levels below can also be accomplished (*upward completion*) and likewise, to have evidence that a level is completed also means all lower must be completed (*downward evidence*).

**Table 1: Levels of action (action ladder), adapted from (Clark, 1996: pp253)**

| Level | Speaker <sup>1</sup>                       | Addressee                                  |
|-------|--|--|
| 4     | Proposal (propose joint activity)          | Uptake (consider joint activity)           |
| 3     | Signal (intended meaning)                  | Recognition (recognized meaning)           |
| 2     | Presentation (of a signal)                 | Identification (of a signal)               |
| 1     | Execution (of behavior composing a signal) | Attention (to behavior composing a signal) |

We can clarify the use of these action ladders with the following example:

**Example 1: A minimal exchange between Karen and Bob**

Karen (to Bob): Please sit down?  
 Bob: -sits down-

This exchange consists of two communicative acts: Karen’s proposing Bob to (consider to) sit down and Bob’s acceptance of this proposal. The requirements for speaker and addressee were fulfilled in both acts. What happened here in terms of Clark’s action ladders is the following: Karen produced (1) the sounds, heard by Bob (1), that composed a signal (2) recognized by Bob (2), that was intended by Karen (3) for Bob to understand (3), that she proposed (4) he would consider (4) to sit down. Bob moved to a chair, Karen could see this, which indicated by the principle of

---

<sup>1</sup> We will refer to the person producing the communicative act as speaker, even when the act is not of a verbal nature.

*downward evidence*, (1,2,3,4), that Bob had attended to (1), identified (2), understood (3), and accepted (4) her proposal.

Each level required coordination: Karen's responsibility at level 1 for example, went beyond executing a signal, she also had to make sure she was attended to (she had to capture and monitor Bob's attention). Likewise at level 3 she had to convey meaning in a way she anticipated Bob to be capable of understanding and subsequently monitor his understanding. In other words: her utterance was *designed for Bob*; it was composed in a way that took Bob's (perceptive) capabilities, knowledge and willingness to confer into consideration. Bob had to play his part in the exchange as well. He needed to provide evidence for Karen to what extent he was 'with her' on each level, in order to allow Karen to adapt her presentation to his needs. He did so in the second communicative act, but in other exchanges he might have done part of the job during the first act. For example if he had not understood what Karen was trying to say he might have given her a puzzled look. We will refer to these behaviors as *backchannel responses*.

Clark's action ladders presuppose the roles of speaker and addressee. As a consequence, we should wonder how these roles get established in the first place. We examine this question in the next section and argue that these roles are, generally, implicitly allocated as a result of the way participants coordinate the content of their exchanges act by act.

---

### 2.3 Sequencing Communicative Acts

---

Conversation seldom stops at one or two communicative acts: in fact, strings of such actions, discourse, compose most of our day-to-day interactions with other people. Each communicative act imposes *conversation roles* for the participants (Goodwin 1981). Take the following example:

**Example 2: Karen, Marianne and Bob talk about the weather**

Karen: do you think the weather will improve?

Bob: well, they predicted dry weather, sun, and -8 degrees for tomorrow.

Karen: ok that's good!

Marianne: you call that an improvement?

Karen's first utterance was most likely directed at both Bob and Marianne, so 'the addressee' was a group in this case. Bob's second utterance was a response to Karen's question and as such directed at Karen. Marianne was not directly addressed by Bob's utterance, and we call her a *side participant* (see Clark 1996, pp 14, for a more extensive

taxonomy). Likewise, Karen's response was directed at Bob, and Marianne's response was directed at Karen, making Bob a side participant.

This allocation of roles resulted from the coordination of this sequence of communicative acts: at level 4 of Clark's action ladders we could say Karen proposed to Bob or Marianne to give her a value statement about the upcoming weather, Bob showed his uptake of this proposal by giving facts about the weather, subsequently Karen showed her uptake of these facts by expressing the value statement that she was after, a proposal in turn contested by Marianne. So, as a basic observation, we could say that much of our discourse consists of paired communicative acts (Schegloff 1968, called these *adjacency pairs*). The second communicative act of the pair, where a speaker shows his uptake of the previous speaker's proposal, is mostly directed at the previous speaker. Moreover in many cases those responses are the first part of a new pair: they contain a new proposal, directed at the previous speaker, and with it comes the allocation of the conversation roles for the next communicative act (see: Sacks, Schegloff & Jefferson, 1974).

If we want to have a system keep track of who is talking to whom, it seems that a reasonable starting point is to assign the addressee of a communicative act to the previous speaker (Traum, 2004). At the same time, there are of course exceptions to this rule: Marianne contests Karen's expressive speech act but she might as well have contested the facts Bob was presenting, addressing Bob: 'I read in the paper there would be rain'. In this case, participants will recognize the reference Marianne is making to the facts presented by Bob, rather than Karen's expressive speech act, and still know who is addressed. A system, however, would need to be capable of quite complex semantic processing in order to trace such references<sup>2</sup>. Luckily, the way participants coordinate within a single communicative act provide us with additional means of inferring for whom it is intended.

---

## 2.4 Coordination within a single communicative act

---

Within the (hypothetical) examples discussed so far, addressees provided evidence for their attention, identification, understanding, and uptake simultaneously go in their

---

<sup>2</sup> The issue of addressing groups is beyond this thesis, but as Jovanović & op den Akker (2004) argue people could use personal pronouns such as we, you, some of you, to distinguish between a single addressee and a group. While the occurrence of such pronouns may be easy to track, additional contextual information is needed in order to know who is referred to by those pronouns.

response. However, people may want evidence earlier, even if it is less convincing and on lower levels.

Consider the following (real world) example:

**Example 3: Coordination in a single communicative act between Ethyl and Barbara. (taken from Goodwin, 1981: pp 60):**

Ethyl: So they st- their clas ses start around in  
Barbara: .....X\_\_\_\_\_

Ethyl produced an utterance but not fluently: she produces a restart (*they st-* is replaced later by *their classes*). At the same time of Ethyl's restart Barbara turns her head towards Ethyl (marked as dots) and looks at her from the word 'classes' on (marked as X followed by a line).

We may draw two conclusions from the example: Barbara's attention may be *merely drawn* to Ethyl because she produced a restart, or Ethyl produced the restart *in order to* get Barbara's attention. The truth is most likely somewhere in between: in terms of Clark's action ladder Ethyl and Barbara are coordinating their execution and attention. Ethyl is trying to make sure Barbara is attending to her speech, and likewise Barbara is providing attention at a point Ethyl may need it. Barbara only provides evidence for her attention by looking at Ethyl. This is much weaker evidence that she is eventually going to be 'with' her, than what she could provide in the next communicative act, but the advantage is she is able to deliver at an early stage. Rather than *assuming* Barbara was listening, Ethyl could rely on concrete evidence *indicating* Barbara was listening, making it easier to proceed.

Eye gaze is by no means the only way to provide evidence for attention. In the first place eye gaze is part of a hierarchy of displays of attention consisting of: eye gaze, head orientation, upper torso and body orientation (Von Cranach, 1971). Secondly, there are other backchannel responses including nods and spoken attention signals such as 'mhm'. People may coordinate higher levels during a communicative act as well: addressees can use facial expressions (like a puzzled look, to show non-understanding), acknowledgments (such as yes, indeed), laughter, and so on to indicate they are 'with' the speaker. However, most of these signals are accompanied by eye gaze (Kendon, 1967), eye gaze has been most widely investigated and eye gaze is the primary gesture we can detect automatically. As such we will focus on eye gaze during the remainder of this section.

For coordination purposes, eye gaze serves as a two edged sword; we generally look at those things we attend to, and, in doing so, we provide evidence to others about what we are attending to. This also means that when speakers look at their addressees

to monitor their delivery, their addressees should preferably be looking at them (Goodwin, 1981). But in everyday conversation participants do not look at each other all the time. For example, Argyle & Cook (1976) investigated dyadic conversation and report that speakers look at their addressees on average 41% of the time and their addressees to look at the speaker for 75% of the time. There are patterns in the way speakers look at their addressees. Kendon (1967) found speakers to look away from their addressees during the beginning of long (>5s) utterances. He hypothesized speakers do so in order to reduce visual input while organizing their thoughts. Speakers tended to look at their addressees near the end of their utterances. The most likely explanation for this is that they want to monitor their delivery. In line with this interpretation Cassell, Torres and Prevost (1999b) found speakers to look away from their addressees during those elements of an utterance that linked to previous contributions but to look at their addressees during those parts of the utterance that were new or interesting. While speakers withdraw eye gaze at some points their addressees do so as well. Kendon (1967) found addressees to switch between long glances towards the speakers interrupted by short glances away. One mechanism underlying this looking away may be the occurrence of eye contact. Eye contact, or mutual gaze, is an arousing and salient stimulus that is hard to ignore (see: Argyle & Cook 1976; Rutter, 1984) so that one of the participants may want to break eye contact soon after it is established. Indeed periods of eye-contact turn out to be short (Kendon, 1967).

Vertegaal et al. (2001) found these findings in dyadic conversation to translate well to four party conversation. In a laboratory setting they found speakers looked at individuals they were addressing much more than side participants (39.7% versus 11.9%). When speakers addressed a group of three they distributed their attention between their addressees (19.7% each). This percentage is higher than the amount of visual attention for side participants in the case of addressing a single person (19.7% against 11.6%) but also higher than the visual attention for a single addressee divided by 3 (19.7% against 13.2%). The most likely explanation for this effect is that it is harder for speakers to collect feedback on their delivery and as such they need to gaze more. Listeners looked at the speaker 62.4% of the time, but Vertegaal et al. (2001) made no distinction between addressees and side participants. It is hard to predict what sideparticipants may do, as they can alternate their gaze between speaker and addressee (resulting in less gaze) but they are less likely to have eye contact with the speaker (resulting in more gaze). Overall the quantitative results Vertegaal et al. (2001) presented are similar to those obtained in two-party conversation. If there are

differences between the way participants coordinate coordination and execution in two party and multi party conversation they need to be revealed by a much more detailed comparison of the specific timing of looks between the two party and multi party case.

The studies presented so far were stylized in the sense that the only targets of interest were other participants. In contrast, Argyle & Graham (1977) examined the amount of gaze objects may receive in face to face communication. He placed dyads at a table where a map of Europe was present and asked participants either 'to get to know each other' or to 'plan a trip in Europe'. Across different conditions he varied the information density of the map and the surrounding environment, and measured the amount of gaze toward each other, the map and the environment. His main finding was that people looked mostly at the map (>90% of the time) when planning a trip but not when getting to know each other. The information density of the map had some impact on the gaze behaviour when it was relevant for the conversation, but when the map (or the environment) was not relevant to the conversation it had no reliable effect on eye gaze. Argyle and Graham concluded that it was the relevance of the map to the conversation that made people look at the map most of the time and labelled this type of object a *situational attractor*. A recent but small study by Fussell, Setlock & Parker (2003), confirmed the findings of Argyle and Graham. Fussell et al. (2003) classified the gaze of helpers in a collaborative robot construction task. Helpers were found to look at the robot, the pieces, the hands of the participants that constructed the robots but hardly at their face.

People may look at those objects they talk *about* as well. A study about human-human-robot communication (Katzenmaier et al., 2004) tried to distinguish between utterances that were intended for the robot and utterances that were intended for the human partner. In both types of utterances people mostly looked at the robot (played by a camera) but, at the same time, most of the conversation was about what the robot could or could not do. The fact that people look at objects they talk about, or refer to, should not be too surprising. Recall we said that people generally look at what they attend to, and in doing so they provide evidence to others about what it is they are attending to. So if we are talking about objects or using objects as reference points the best evidence we are 'with each other' is provided by looking at the object we are referring to, not by looking at each other. Moreover objects may attract attention at points where people in Kendon's (1967) study tended to look away: they may perform a function in organising one's thought.

In all studies reported there are large individual differences between the amount of gaze participants exhibit to each other. There is a whole line of research identifying

factors that may account for this variance such as: gender (women tend to look more than men), interpersonal relationships (intimates gaze more than strangers), interpersonal distance (people seated far away gaze more than people seated close to each other), attitudes and emotions (positive emotions are correlated with increase of gaze while negative emotions are correlated with decrease of gaze), personality traits (extraverts and dominants look more than introverts and submissives), culture (Americans look more than Asians), topic (strangers look less in discussing controversial topics like abortion) and so on. None of these effects are very strong: most of the variance is left unexplained (see: Rutter 1984). This merely indicates people are flexible in the way they coordinate their communicative acts with each other. Speakers can start by assuming the other is attending, and addressees have multiple ways of providing evidence for their attention: depending on the situation and their own needs for evidence in the process people adopt different styles of communication. When available, eye gaze is a good indicator of attention, but it is by no means exclusive.

To sum up: if we are to infer the conversation roles participants take in conversation from the way people coordinate within a single communicative act, detecting who is gazing at whom is a good starting point. We can also say there is a hierarchy of evidence: the best evidence we are looking at a speaker and an addressee who are coordinating their execution and attention is the occurrence of eye contact between the two. The eye gaze of the speaker follows, since a listener who looks at the speaker may be either a side participant or the addressee. However, when situational attractors are present the evidence we get from detecting eye gaze may be less good. Situational attractors are common in human-human-machine conversation. Most multi-modal interfaces present relevant information; we should expect people to look at that information as it is evidence for both the participants they are attending to the matters at hand. To what extent eye gaze can still be used to infer the addressee in this situation, is a question we take up in Chapter 3.

---

## 2.5 Coordinating with conversational agents

---

So far we have been concerned with the way people communicate with each other. But we may ask if the same principles hold for communicating with computers. In this section we argue that the same basic principles apply, but that we cannot and should not equate human-computer interaction and human-human interaction on a behavioral level. We will illustrate this point of view with two examples: the use of

recipient design and task context as a way to obtain addressee information and the problem of embodiment design.

In Section 2.2 we introduced the notion of *recipient design*: speakers formulate their communicative acts in anticipation of the addressee's attention, identification, understanding and uptake. As we have a lot of experience in communicating with humans we manage to come to a quick understanding of what others know (Fussell & Krauss, 1992), making recipient design relatively easy. Conversational agents However, have much more limited capabilities than humans and if people communicate with them like they do with humans breakdown occurs (Martinovski & Traum, 2003). In anticipation of the limited capabilities of conversational agents, people tend to omit linguistic complexities resulting in shorter utterances (Oviatt, 1999; Katzenmaier et al. 2004). This asymmetry between the way humans communicate among each other and the way they communicate with systems can be used as an indication people are addressing the system (Katzenmaier et al. 2004). However, it is hard to say to what extent recipient design, as it is formed by human expectations of a system, generalizes over different ways the agent is designed, and what the role of factors like previous experience are. A related but slightly different reason people may speak differently to systems has to do with the task context: if systems have a narrow scope such as providing specific types of information, or if they have a certain social role such as 'server', comparing the utterances users use to a language model related to the task or role may help to infer whether the system is addressed. For example Katzenmaier et al. (2004) found users used more imperatives when they spoke to robots than when they spoke to each other. Using task context and social role may help in detecting the addressee but is surely not a full solution. Katzenmaier et al. (2004) found their linguistic cues to be only weak indicators of who was talking to whom each utterance. As people may talk *about* a task to each other, task specific words pop up in utterances that are not intended for the system as well.

A second issue in human-human-computer communication is that of *embodiment*. In section 2.1 we said that the question 'how to use information about who is talking to whom in the interface' has received little attention. However, since we argued in section 2.3 that in multiparty conversation the addressee of an utterance *is* generally known, we may turn to systems that try to coordinate communicative acts with single users, such as 'Gandalf' (Thórisson, 1996) and 'Rea' (see for example: Cassell et al. 1999a) for inspiration. These *embodied conversational agents* are humanoid characters that generate multimodal output such as head nods, facial gestures, and eye gaze closely tied to the behaviors of users: in fact these are the first systems that can be said to



coordinate with, rather than react to users. Impressive as these systems are, we must be cautious to implement human like behaviors straightforwardly into the interface for at least four reasons.

First, as Cassell et al. (1999a) argue, *if* we use the expressive behaviors of humans in the interface they should be more than metaphorical. These behaviors need to be tied to the behaviors of users and the functions they have in the dialog management such as ‘take new turn’ and ‘contribute new information’. Second, as Shneiderman (Maes & Shneiderman, 1997) argues, we must be cautious not to *mislead* the users by raising their expectations of the systems’ capabilities above the actual capabilities of the system. Third, although human modes of expression may be *intuitive*, they are not necessarily the *best* for what the system needs to express: for example if a system displays its real-time speech recognition results in the form of text on the screen it may be said to fulfill the human need for early evidence of their communicative success perfectly well, while humans need to rely on much more limited facial expressions for showing their understanding. Last, not all *tasks* may lend themselves for mediation by a human character: Bolt’s (1980) ‘put that there’ system may be better off by enhancing the replaceable entities on the screen with feedback than to have the communication mediated by a virtual humanoid character, and the (spatial) limitations that go with it. We can conclude that there is nothing wrong with taking human expression as inspiration for designing socially aware conversational agents, but the challenge is to find ways of expressing the right information at the right time given the ongoing dialog context, the capabilities of conversational agents, the limitations of conversational agents and the requirements of the task.

---

## 2.6 Conclusions

---

In this chapter we have worked from the theory of language as coordinated action to preliminary answers to two questions: ‘how can we detect who is talking to whom’, and, once we know, ‘how should we use this information in the user interface’. Let us now review our answers to these questions in turn.

We have argued that in general, the conversational roles imposed by the communicative acts of the participants are implicitly allocated as a result of the way people coordinate the content, their proposal and uptake of these acts. Despite the fact that for human participants it is generally clear who is talking to whom, they use contextual information to extract this information and it is by no means trivial for a

system to keep track of the way people use such information. Still there are three strategies systems can use to infer who is talking to whom.

1. Dialog history: often the addressee of an utterance is the previous speaker. Using this information is no more than a starting point, but combined with other strategies it may bring systems quite far. If the system produces utterances itself, certainly if these are questions, it may expect an answer.
2. Coordination within communicative acts: as the speaker and addressee coordinate their execution and attention, backchannel responses, most importantly eye gaze, provide us with evidence who the speaker and addressee are. Mutual gaze is the strongest evidence, followed by speaker gaze, followed by listener gaze.
3. Speaking styles: because people have lower expectations of systems, and the system has a different role in the dialog, users may speak differently to systems than to their human participants. Most notably they tend to use shorter utterances.

Within chapter 3 we will revisit these strategies, make them operational for the interface we are using and our technological frame, and see to what extent each individual strategy can be used to infer who is talking to whom in our design-case. In chapter 4 we see how a computational model can combine these types of information into a quantitative estimate on who is being addressed.

In section 2.4 we have argued that people, in communicating with each other, seek out and provide early evidence on the lower levels of communicative action. In order to be capable of providing such early feedback we need to have early estimates of who is being addressed. In chapter 4, we will explore to what extent such early estimates are possible. In chapter 6 we will redesign our system *given* the limitations of these estimates. We will take the way people coordinate within communicative acts as inspiration, but explore different ways to embody our agent than merely imitating human expressions.



## Chapter 3: Sensing Who is Talking to Whom in Dyad-Kiosk Conversation

*In this chapter, we cover two tasks that mark the path from the descriptive theory of chapter 2 to our efforts to infer who is talking to whom in each communicative act in chapter 4. First, we will translate the three general strategies for detecting who is talking to whom: making use of dialog history, coordination within communicative acts and differences in speaking styles, to specific tactics that we can employ within our design case. These tactics make use of dialog events, head orientation and utterance length. Second, we will validate these tactics with an empirical study. All three tactics turn out to be useful. Some dialog events: questions from the system and tap to talk actions turn out to have a definite effect on the dialog. After such events participants are likely to address the system, but they do not occur often. We can use head orientation in many other utterances, even though people look at the system most of the time. Utterance length turns out to be the weakest cue<sup>1</sup>*

---

<sup>1</sup> The manuscript of this chapter is partly based on the following publications: Bakx et al. (2003) , Van Turnhout et al (2005) and Van Turnhout (2006).

---

### 3.1 Introduction

---

The way people coordinate their communicative acts in human-human-machine conversation should enable us to infer who is talking to whom in each communicative act. In chapter 2 we claimed that we may employ three general strategies for that. We can keep track of the dialog history, use signals that suggest that people are coordinating within communicative acts or try to capture differences in speaking styles between human-human and human-machine communication. However, so far we have primarily drawn from communication theory and have not considered our specific design case in much detail. Within the design case, we commit - as outlined in chapter 1 - to a social frame and a technological frame. We focus on dyads interacting with an information kiosk and limit ourselves to sensors that measure on-off patterns of speech, head orientation and dialog events of the system. In this chapter, we have to carry out two tasks.

First, we need to operationalize the theory for our specific design case. For each strategy we need to develop a tactic<sup>2</sup> that fits our technological frame. We will discuss how keeping track of dialog events provides us with information about dialog history, how sensing head orientation provides us with evidence about coordination within communicative acts and how sensing utterance length provides us with evidence about differences in speaking styles. Second, we need to validate these tactics. Using data from an empirical study, with a Wizard of Oz simulation of an existing multimodal interface, we try to show to what extent each tactic helps to arrive at estimates about who is talking to whom.

We will start this chapter with describing the setup of this empirical study, in section 3.2. Next, we set out to operationalize and validate each tactic: ‘Dialog history’ in section 3.3, ‘coordination within communicative acts’ in section 3.4, and ‘speaking styles’ in section 3.5. In Section 3.6 we provide a general discussion and outlook.

---

<sup>2</sup> The use of the word ‘tactic’ may seem odd in this context. We use it, analog to the military use, to distinguish unambiguously between the general strategies of chapter 2 and their specific *operationalization* in this chapter. In this chapter we will also speak of ‘a’ tactic and ‘their’ tactics. One may also read: modus operandi, methods, or techniques. The term *approach* is reserved for approaches to other problems than inferring who is talking to whom in our case.

## 3.2 Experimental Setup

---

### 3.2.1 Subjects and tasks

We invited eleven pairs of people that were already acquainted with each other to the usability lab of the university: 3 male only, 4 female only and 4 mixed pairs. All participants were affiliated with our university, either as student or staff<sup>3</sup>. The task was to plan a joint round trip<sup>4</sup> to a tourist attraction in the Netherlands using a multimodal dialog system for obtaining train-table information: MATIS (described further on). First we gave participants a brief explanation of MATIS and the opportunity to act out a single practice trial. Then, all pairs acted out two scenarios of two dialogs each: a round trip to a zoo and a round trip to a museum. They received information about the tourist attractions and a scenario. The scenario included soft constraints such as “you want to be home before dark” (see: appendix A). In order to stimulate discussion between the participants the information about tourist attractions and constraints were different for each participant (although not conflicting). We informed participants about this. Also, we asked them to interleave all negotiations between each other with the interaction with MATIS, rather than discussing everything before use. Participants turned out to be able to empathize with their task. Many pairs planned on trips to tourist attractions that were not in their scenarios, forgot about ‘constraints’, or discussed combinations with other options such as going shopping or staying overnight. For purely technical reasons, a missing log file or video, we excluded the data of 3 pairs, resulting in a total dataset of 8 pairs performing 32 dialogs.

### 3.2.2 MATIS

As we did not have a system with automatic addressee determination we needed to create a Wizard of Oz set-up. The user interface closely mimicked MATIS: an existing multimodal interface (see: Sturm et al. 2002; figure 1a). It contains search parameters: departure station, arrival station, date of travel, time of travel, and a parameter indicating departure or arrival time.

---

<sup>3</sup> Recruited from the department of Architecture, Industrial Design, Mathematics and Technology Management.

<sup>4</sup> A roundtrip takes two dialogs with MATIS: a dialog for the outward journey and a dialog for the return journey.

The screenshot shows a yellow interface with the following elements:

- From:** A dropdown menu with 'amsterdam cs' selected.
- To:** A dropdown menu with 'eindhoven' selected.
- Day Selection:** Three buttons: 'today', 'tomorrow', and 'other day' (with a microphone icon).
- Date:** A field containing 'donderdag' and another field containing '24 -april -2003'.
- Time:** A field containing '19:30'.
- Buttons:** 'departure', 'arrival', and 'search'.

**MATIS:** Welcome at ovis the public transport information system. About what connection would you like information?

**Bob:** from Eindhoven

**Bob:** well, where shall we go?

**Karen:** I would like to go to Burgers Zoo

**Bob:** this sounds cool!

**Karen:** yeah.

**Bob:** where is it?

**Karen:** Arnhem

**Bob:** ok that's not too far, let's go there (presses 'to' field)

**Bob:** to Arnhem

**MATIS:** At what day would you like to travel?

(a)

(b)

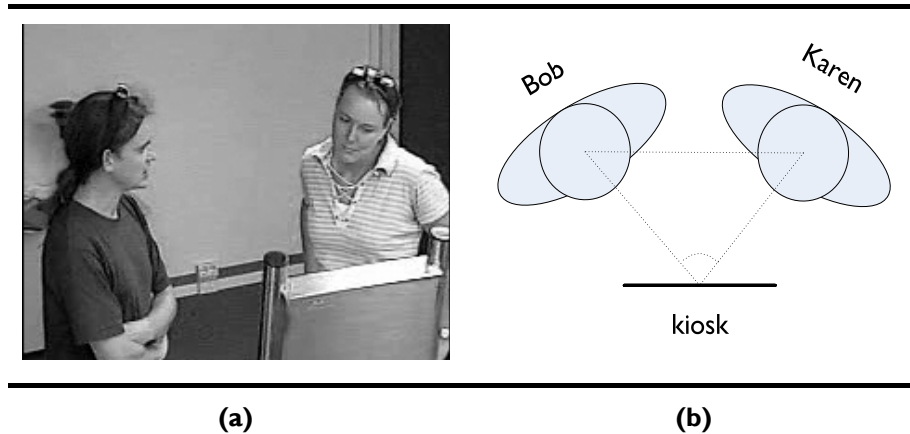
**Figure 1: The form of the MATIS interface (a) and a hypothetical example of a dialog between Karen, Bob and MATIS (b)**

MATIS supports mixed-initiative spoken dialog. It starts with an open question (see: figure 1b) to ask for input, followed by questions prompting for the next empty field, each time a field has been filled. Users are allowed to fill any combination of fields after each prompt, but they hardly ever fill more than one field at a time. Users can also take control of the dialog by pressing buttons of two types. Microphone buttons, such as 'from' and 'to' allow users to fill the specific field by speech (tap 'n talk). Field-fill buttons ('today', 'tomorrow', 'departure', and 'arrival') allow users to fill fields without speech. Once users have filled a field, with either a microphone button or a field-fill button, MATIS proceeds as normal by prompting for the next empty field.

We replaced the speech processing and dialog management modules of the original MATIS system by a wizard using a GUI interface, allowing him to provide the data from the database and to control the dialog prompts. We faked speech recognition errors but did not have a formal protocol to control the number of these errors. Therefore the number of speech recognition errors was low. For one destination city in the museum scenario, Amstelveen, there is no railway station. Users typically interpreted the wrong result in the field as a speech recognition error a few times, before realizing they had to travel to a different station to reach the museum. MATIS prompted for filling the next empty field immediately after the previous field was processed, unless the participants were already involved in further discussion: in that case the prompt was withheld. The decision to accept an utterance as intended for MATIS was up to the wizard.

### 3.2.3 Data collection and transcription

For each pair we recorded a log-file with system events (prompts, button presses, moments text appeared on the screen). In addition we collected head orientation data with a custom made head orientation tracker (described in more detail in chapter 5). And we recorded videos (an example of a video frame is shown in figure 2a).



**Figure 2: Two participants (for example Karen and Bob) interacting with an information kiosk (a), and a schematic top view drawing of this situation (b).**

Because we were not sure about the relation between eye gaze and head orientation we annotated eye gaze for both participants by hand for each video frame. We made a distinction between looking at the partner, looking in the general direction of the kiosk and looking elsewhere (for example at the door behind the pair). However looking elsewhere was so rare (<2% of the data) that in the analysis we treated it as looking at the partner, resulting in a binary classification scheme. Simplifying eye gaze measurements by assuming only a few targets are of interest for participants is in line with the attention approach (Stiefelhagen & Zhu, 2002) that we adopt to replace eye gaze measurements with head orientation measurements. We will discuss this in more detail in section 3.4.1.

The audio recording (taken from the video) was segmented into *utterances*. Although we are mostly interested in communicative acts it is not straightforward to define them on the basis of features that are recognizable by computers. For example, Clark (1996, p130) defines communicative acts as “The joint act of a person signaling another and the second recognizing what the first meant”. Kendon (2004, p7) provides a similar definition of an ‘utterance’: “Any ensemble of action that counts for others as an attempt by the actor to ‘give’ information of some sort.”. Both definitions do not provide a third party with objective criteria that allow to decide when a single communicative act starts and ends. Considering that we would like to settle the addressee of a communicative act independent of the results of any semantic processing, we defined the utterance based on on-off patterns of speech (Definition 1).



**Definition 1: the utterance**

An *utterance* is a stretch of uninterrupted speech from a single speaker followed by a silence of at least 500 ms.

The time-lapse of 500 ms was determined empirically. It represents a tradeoff. We *do not* want to cut up sequences of speech that we would consider a single communicative act based on context and content. But we *do* want to cut up sequences of speech that, based on context and content, we would consider to be composed of a communicative act for a partner, directly followed by a communicative act for the system, or vice versa. There are still some of these combinations in the dataset (about 10%), and we will label those *compositions*. We must note that decreasing the length of the time criterion does not decrease the number of compositions dramatically, but does increase the number of utterances that are cut up unjustly.

We obtained utterances from the audio file by post-processing. The audio file was normalized and microphone noise was removed by applying spectral filtering and gating the file at -30 dB. The audio was then manually divided over three tracks speaker 1, 2 and MATIS. In this process we also removed remaining noise bursts louder than -30 dB. The audio-file was down sampled to 11.1 kHz, sections containing audio were marked and for each of the three tracks we filled gaps shorter than 500 ms. Next, the markers were down-sampled to 10 Hz and combined with the hand marks for focus of attention and log files of the system.

In all, the 32 dialogs contained 925 utterances of the users, 202 directed to the system (21 of those where compositions) and 723 directed to the other member of the pair. The big *class skew*: the large number of utterances for the partner compared with the number of utterances for the system, is only in part a result of our efforts to stimulate discussing among participants. Since there are only five fields to fill in MATIS, the number of utterances for the system is always low, while even a limited discussion between participants takes up a few utterances. This becomes apparent when we compare the class skew in ‘toward dialogs’ with those in ‘return dialogs’. In toward dialogs participants typically discuss the values of all fields, resulting in a huge class skew (114 utterances to the system, 558 to the partner). However, there is still a large class skew in ‘return dialogs’ (88 utterances to the system, 165 to the partner), while in those dialogs participants tend to limit discussion to the values of the date and time field.

In this chapter we will use the overall class skew as reference point. If the class skew of utterances for which we observe certain behavior (for example both participants look at the system) differs significantly from the overall class skew, we consider this evidence that observing this behavior is useful for determining the addressee of this utterance. In other words, say we observe that 90% of the utterances where both participants look at the system are intended for the system, while overall only 20% of the utterances are intended

for the system.. In that case we conclude that observing that both participants look at the system is evidence for the utterance to be intended for the system.

---

### 3.3 Dialog History

---

#### 3.3.1 Dialog History versus Dialog Events

In section 2.6 we said that the addressee of an utterance is often the previous speaker. Thus, tracing dialog history may provide a starting point for detecting who is talking to whom. Here we need to consider how this finding may translate to dyads interacting with MATIS. MATIS is a multimodal system: it will ask questions, print information on the screen and allows users to press tap ‘n talk as well as field-fill buttons (section 3.2.2). This forces us to consider what the effect of each of these events may be in more detail.

When MATIS asks the users a question, it may expect an answer. Thus, utterances from any speaker immediately following a MATIS question may be expected to be intended for the system. At times, however, a MATIS question will serve as a starting point for negotiation with the partner. Say, when MATIS prompts for the arrival station, and participants have not yet decided on that, they will first want to negotiate this field. So, while we do not expect a 100% score, we do expect that more utterances following a question from MATIS are intended for the system than would be expected on the basis of the overall class skew (hypothesis 1).

#### **Hypothesis 1**

Utterances following a question from MATIS are likely to be intended for the system. The frequency of utterances for the system after MATIS questions is higher than may be expected on basis of the overall class skew.

When a participant fills a field by speech, MATIS will respond with displaying the result of the speech recognition engine in the particular slot on the screen. We expect these *text* events to have much less effect on the dialog than MATIS questions for two reasons. First, provided that the speech recognition has an acceptable performance, the text will usually serve as confirmation. MATIS merely communicated it has understood the participants well. In those cases the imperative to react may be much weaker than with questions. Second, the key difference between text on the screen and a spoken confirmation is that text is not evanescent. As a result there is no time imperative for the participants to react on the confirmation. Even if participants feel the need to react on results on the screen there is no principal reason to do so immediately. Following these two arguments we expect that the appearance of text events on the screen bears no relation to the addressee of an utterance following this event (hypothesis 2).

**Hypothesis 2**

Utterances occurring immediately after text events are as likely to be intended for the system as any utterance. The frequency of utterances for the system after text events does not differ from the overall class skew.

There are two ways for participants to communicate explicitly that they intend to address the system. The first way is to press a tap ‘n talk button. By pressing such a button users communicate they intend to fill a particular field with speech. As a result we expect that utterances after pressing such a button are intended for the system (hypothesis 3). The second way is to press a field-fill button. For example, if they intend to travel ‘today’, they can fill this field immediately by pressing the ‘today’ button. With this button users express they intend to fill a particular field, but in one go they fill the field as well. They may also say the value they seek to fill out loud, which we count as an utterance intended for the system, but there is no need to do so. Thus we do not believe pressing field-fill buttons bears a relation to utterances immediately after such an event (hypothesis 4).

**Hypothesis 3**

Utterances immediately following the event of users pressing a tap ‘n talk button are more likely to be directed to the system than utterances not following such an event. The frequency of utterances intended for the system immediately following a tap ‘n talk button is higher than may be expected on the basis of the overall class skew.

**Hypothesis 4**

Utterances immediately following the event of users pressing a field-fill button, are as likely to be intended for the system as utterances not occurring immediately after text-out events. The frequency of utterances intended for the system after pressing a field-fill button does not differ from that expected on the basis of the overall class skew.

**3.3.2 Results for dialog events****Questions from MATIS**

Hypothesis 1 predicts that utterances after questions from MATIS are more often directed to the system than should be expected on the basis of the overall class skew. In total MATIS asked 80 questions. Table 1 shows the addressee of the utterances immediately after these questions, compared to the overall class skew.

**Table 1: The number of utterances intended for either system or partner occurring immediately after a MATIS question, compared with the overall class skew.**

| Addressee | After a question | Overall (chance level) |
|-----------|------------------|------------------------|
| System    | 40 (50%)         | 202 (22%)              |
| Partner   | 40 (50%)         | 723 (78%)              |

$$X^2 = 31.9; df = 1; p < 0.001$$

We see that MATIS questions are just as often followed by an answer intended for the system as by an utterance for the partner. Still this is significantly different from the overall class skew ( $X^2 = 31.9$ ,  $df = 1$ ,  $p < 0.001$ ). Compared to the overall class skew, utterances following a MATIS question tend to be more often intended for the system. Hypothesis 1 is confirmed.

### Text events

Hypothesis 2 predicts that text events have no effect on the dialog. The frequency of utterances immediately after such an event that are intended for the system should not differ from the overall class skew. In total, there are 52 text events in the dataset. Table 2 shows the addressee of utterances following such an event compared to the overall class skew.

**Table 2: The addressee of utterances occurring after a text out event, compared with the overall class skew.**

| Addressee | After a text out event | Overall (chance level) |
|-----------|------------------------|------------------------|
| System    | 12 (23%)               | 202 (22%)              |
| Partner   | 40 (77%)               | 723 (78%)              |

$$X^2 = 0.044; df = 1; p < 1$$

We see that the class skew of utterances after a text out event are indistinguishable from the overall class skew. Hypothesis 2 is thus confirmed.

### Tap ‘n talk buttons

Hypothesis 3 predicts that utterances after pressing a tap ‘n talk button are likely to be intended for the system. We expect there are more utterances for the system after such an event than should be expected on basis of the overall class skew. In total users pressed tap ‘n talk buttons 48 times. Table 3 lists the addressee of utterances following these button presses compared with the overall class skew.

**Table 3: The addressee of utterances occurring after a tap n talk button, compared with the overall class skew.**

| Addressee | After a tap ‘n talk button | Overall (chance level) |
|-----------|----------------------------|------------------------|
| System    | 40 (83%)                   | 202 (22%)              |
| Partner   | 8 (17%)                    | 723 (78%)              |

$$X^2 = 92.4; df = 1; p < 0.001$$

As we expected there is a strong tendency of users to address the system after explicitly indicating they want to do so by pressing a tap ‘n talk button ( $X^2 = 92.4$ ,  $df = 1$ ,  $p < 0.001$ ). The surprising fact in table 3 is that there turn out to be utterances that defy

hypothesis 3. Occasionally it happens that participants press these buttons by accident, or decide to press the button, ask their partner to confirm, for example, the destination station, before filling it by speech.

### Field-fill buttons

Hypothesis 4 states that utterances after pressing a field-fill button do not differ from other utterances: they may be intended for the system or intended for the partner. We expect that the class skew of these utterances is no different from the overall class skew. In total users pressed a field-fill button 20 times. Table 3 lists the addressee of utterances following such an event compared with the overall class skew.

**Table 4: The addressee of utterances occurring after a tap n talk, compared with the overall class skew.**

| Addressee | After a field-fill button | Overall (chance level) |
|-----------|---------------------------|------------------------|
| System    | 8 (40%)                   | 202 (22%)              |
| Partner   | 12 (60%)                  | 723 (78%)              |

$$X^2 = 3.73; df = 1; p < 0,10$$

We see that the class skew of utterances after a field-fill button does not differ significantly from the overall class skew. Hypothesis 4 is confirmed. However there is a trend for these utterances to be intended for the system. We inspected the video for these cases and found out that users often say the value of the field out loud in these cases.

### 3.3.3 Conclusions about dialog history and dialog events

In this section we have outlined a tactic to infer the addressee of an utterance by using dialog events of MATIS. Indeed, some of these events: MATIS questions and users pressing a tap ‘n talk button have a very definite effect of the dialog. As a result we can use them as indicators for the addressee of an utterance. However, there are two points of caution. First, surprising or not, utterances following MATIS questions and those following tap ‘n talk presses are still often followed by an utterance for the partner. Not enough to reject our hypotheses but enough to point out that the existence of an explicit interaction protocol is by no means a guarantee users will follow it. Second, keeping track of dialog events may help to decide who the addressee of an utterance is, but we can only apply it to a limited number of utterances: In our dataset on 128 of the 925 utterances (80 questions from the system plus 48 tap ‘n talk presses). Thus, we need to combine this tactic with other tactics.

## 3.4 Coordination within communicative acts

---

### 3.4.1 Eye gaze versus head orientation a focus of attention approach

In section 1.5.2 we stated that sensing eye gaze can be replaced by -the easier to sense- head orientation. Here we must substantiate this claim. Stiefelhagen & Zhu (2002) report on a study investigating the relation between head orientation and eye gaze in face-to-face meetings between co-workers in a laboratory setting. There were two main findings. First, the head orientation of their 4 participants contributed between 53,0% and 96,7% to their overall line of sight -being the sum of head and eye orientation. Second, they tested a focus of attention approach. Under the assumption that the focus of (visual) attention of subjects could only be other participants they tried to predict this focus of attention with head orientation alone. This approach was successful: for between 82,6% and 93.2% of the video frames the focus of attention could be correctly established on basis of head orientation alone. So substituting eye gaze measurements with the combination of head orientation measurements and a focus of attention approach is profitable where measurements of head orientation are easier to obtain. However, this substitution has its limits. The criterion is that there are only a fairly limited number of targets of potential interest that are spatially sufficiently separated. As Stiefelhagen shows elsewhere (2002) the robustness of the focus of attention approach he uses, decreases already substantially when applied to meetings with five instead of four participants. Likewise, we can imagine if a subject has a laptop in front of him, separating looking at the laptop from looking at the person in front of the subject can be tedious.

Our social situation: two participants interacting with an information kiosk does meet this criterion. Because of the laboratory setting, we were able to make sure the only targets of potential interest for the participants were for the partner and the kiosk. In addition participants turned out to orient towards the kiosk in such a way that they form a triangle with the kiosk. Each participant has to rotate between 40 and 60 degrees to have the other participant in their line of sight (see figure 1b). This is a bigger spatial separation than in the Stiefelhagen and Zhu study (2002). The limitation of adopting a focus of attention approach is that we cannot separate looking at the kiosk from looking in the general direction of the kiosk based on head orientation alone. However making this distinction based on eye gaze would require very precise measurements of eye gaze. These are still very hard to obtain. Thus, given the technological difficulty of tracking eye gaze and the suitability of our situation to apply the focus of attention approach this substitution seems acceptable. The choice for a focus of attention approach is reflected in our choice to use only two targets of interest extracted from the video. We did collect

head orientation data and a reliability test showed a good correspondence with the transcription data<sup>5</sup>. In this chapter, however, we rely on hand transcriptions.

### 3.4.2 Eye gaze versus coordination within communicative acts

In chapter 2 we claimed that keeping track of who is looking at whom may be a good starting point for inferring who is talking to whom. Within our design case we need to distinguish between two mutually exclusive situations. Either the speaker is addressing the system and the partner acts as side-participant. Or the speaker is addressing his partner, who acts as addressee. Since these situations are mutually exclusive, positive evidence for one situation is negative evidence for the other. In other words: if we find evidence that the speaker is addressing the system, this is also evidence that he is not addressing his partner. In this section we will seek out evidence that the speaker is addressing his *partner*. For that we present two arguments.

First, if we set out to seek evidence for one of the two situations we try to distinguish, we take the situation we know best as a starting point. Although we know a lot about how speakers use eye gaze when addressing human conversation partners, it is not so clear how they would use eye gaze in combination with an information kiosk. For example, we know speakers tend to avoid gaze with conversation partners when they need to organize their thoughts. In part because they may want to avoid the arousing stimulus of mutual gaze. Since the speaker and the kiosk cannot engage in mutual gaze, we may ask where speakers look when they organize their thoughts if they address the kiosk. They might turn to their conversation partner, but it seems more likely they would still look in the general direction of the kiosk. Likewise, detecting that the partner is acting as a side-participant is difficult. In chapter 2 we noted the lack of research dealing with the gaze behavior of side participants. In meetings, side participants may alternate their gaze between speakers and addressees. It is unclear how to translate this finding to dyad-kiosk interaction<sup>6</sup>.

The second argument has to do with the notion of *situational attractors*. Situational attractors are objects that are relevant for the conversation and as such attract a lot of eye gaze (see: Argyle & Graham, 1977). The screen of the information kiosk contains relevant information for the dialog, and may act as situational attractor. Recall (from chapter 2) that in the Argyle and Graham (1977) experiment, speakers looked at their situational attractor, a map of central Europe, over 90% of the time. The most important factor for the amount of gaze towards the map was the relevance of the map to the

---

<sup>5</sup> See Van Turnhout (2006), the head-orientation data combined with a focus of attention approach corresponded to the hand transcriptions in 95% of the frames.

<sup>6</sup> Of course it is interesting to find out. However, this is not our primary focus, and it is hard to be conclusive as long as we have so little work on side-participants in human-human communication to compare with.

conversation, not the information content. So, we may expect that in our situation people will look at the screen almost constantly. Against this background, the *event* a participant is looking at his partner is a salient marked act and thus a highly meaningful sign of coordination between those participants. So, because of the situational attractor hypothesis we expect the amount of evidence we find for coordination between a speaker and addressee to decrease but in those cases we find such evidence we can be more sure we are correct about their conversation roles.

In chapter 2, we claimed there is a hierarchy of evidence for coordination between speakers and addressees. The occurrence of *mutual gaze* between a speaker and a conversation partner provides the strongest evidence, followed by *speaker gaze* followed by *listener<sup>7</sup> gaze*. There are utterances where a transition of gaze target occurs. A speaker may look at both the kiosk and his partner during a single utterance. Since, following the situational attractor argument, we put most weight on signs for coordination between partners these utterances are, in first consideration treated as utterances where the participant looks at his partner. This leads to hypotheses 5-8

**Hypothesis 5**

If the speaker and partner engage in mutual gaze during the utterance, it is likely that the utterance is intended for the partner. The frequency of utterances for the partner with mutual gaze is higher than may be expected on the basis of the overall class skew.

**Hypothesis 6**

If the speaker looks at the partner during the utterance, it is likely that the utterance is intended for the partner. The frequency of utterances for the partner where the speaker looks at the partner is higher than may be expected on the basis of the overall class skew.

**Hypothesis 7**

If the listener looks at the speaker during the utterance, it is likely that the utterance is intended for the partner. The frequency of utterances for the partner where the partner looks at the speaker is higher than may be expected on the basis of the overall class skew.

**Hypothesis 8**

Hypothesis 6-8 form a hierarchy of evidence. It is more likely that an utterance is intended for the partner when mutual gaze occurs, than when we only observe speaker gaze. Furthermore, it is more likely that an utterance is intended for the partner when we observe only speaker gaze, than when we observe only listener gaze.

We may spend a few more thoughts on utterances where a transition of gaze target occurs. In hypotheses 5-8 we treated all these utterances in the same way. We consider any glance towards the partner as evidence of coordination between speaker and addressee. But it seems reasonable to involve the type of transition as well. Both Kendon (1967) and Cassell et al. (1999b) noted that speakers in dyads tended to look at their

---

<sup>7</sup> Listener is used as umbrella term for addressee and side-participant. There are cases where it is appropriate to refer to the conversation role of the non-speaking participant, but in general, we do not know which one of the two it is.



partners during the *end* of utterances. Translating this to dyad-kiosk conversation we may expect the gaze target of the speaker during the end of utterances to be more informative about whom the addressee is than his target at the beginning of utterances. This is formalized in hypothesis 9.

#### Hypothesis 9

Utterances where speakers make a transition from the listener to the kiosk, are less likely to be for the partner than utterances where speakers make the opposite transition.

### 3.4.3 Results for coordination within communicative acts

Hypotheses 5-7 propose that mutual gaze, speaker gaze and listener gaze toward the partner are all evidence that the utterance is intended for the partner. We expect that the frequency of utterances for the partner that contain such glances is bigger than may be expected on the basis of the overall class skew. Table 5 lists the number of utterances for the system and partner in each of these categories compared with the overall class skew. There are utterances that fall in multiple categories, for example the occurrence of mutual gaze can coexist with the occurrence of speaker gaze in a single utterance. We counted these utterances only once: utterance where mutual gaze is observed are not counted under speaker or listener gaze; utterances where speaker gaze is observed are not counted under listener gaze. The table has an additional column ‘kiosk gaze’. Here we report those utterances where we found no glances toward the partner from either speaker or listener.

**Table 5: the number of utterances for the system and partner where mutual gaze, speaker gaze, and listener gaze is observed, compared with the overall class skew.**

| Addressee | Mutual Gaze | Speaker Gaze | Listener Gaze | Kiosk Gaze  | Overall   |
|-----------|-------------|--------------|---------------|-------------|-----------|
| System    | 22 (6.5%)   | 18 (11.7%)   | 36 (28.8%)    | 126 (40.9%) | 202 (22%) |
| Partner   | 316 (93.5%) | 136 (88.3%)  | 89 (71.2%)    | 182 (59.1%) | 723 (78%) |

$X^2 = 39.9$ ;  
df = 1;  $p < 0.001$

$X^2 = 17.8$ ; df = 1;  $p < 0.001$

$X^2 = 3.05$ ; df = 1;  $p < 0.1$

$X^2 = 43.0$ ; df = 1;  $p < 0.001$

The table shows that hypothesis 5 (mutual gaze) is confirmed ( $X^2 = 39.9$ ; df = 1;  $p < 0.001$ ), hypothesis 6 (speaker gaze) is confirmed ( $X^2 = 17.8$ ; df = 1;  $p < 0.001$ ), and that hypothesis 7 (partner gaze) is rejected ( $X^2 = 3.05$ ; df = 1;  $p < 0.1$ ). Apparently, the listener can look at the speaker both in the role of side participant or as in the role of addressee. And there is no immediately obvious difference in the way the listener does

this according to his role<sup>8</sup>. Jovanović et al. (2006) mention a similar difficulty with listener gaze for the context of meetings. Note, while listener gaze does not provide evidence about the conversation role of the listener (addressee or side participant), it is still important to measure listener gaze. Looking at the ‘kiosk gaze’ column we see that the absence of any gaze toward the partner from both speaker and listener provides us with evidence the speaker is addressing the system ( $X^2 = 43.0$ ;  $df = 1$ ;  $p < 0.001$ ). Measuring listener gaze is of importance because it enables us to distinguish not gazing at all from the occurrence of listener gaze. Likewise, measuring listener gaze is of importance for the detection of mutual gaze.

A second concern that arises from table 5 is the occurrence of utterances for the system with mutual gaze. It seems weird that people would look at each other when addressing the system. Manual inspection of the video for these utterances revealed that 9 of these utterances were compositions, utterances composed of a part for the system and a part for the partner. So it does happen, but not as often as may seem here.

Hypothesis 8 states that there is a hierarchy of evidence, the strongest evidence is mutual gaze, followed by speaker gaze, followed by listener gaze. Table 5 shows the trends are in the right direction. Mutual gaze has the strongest class skew, followed by speaker gaze, followed by listener gaze. In order to see if this is merely a trend or a robust effect we performed pair wise  $X^2$  comparisons. Comparing the distributions of mutual gaze and speaker gaze shows only a trend ( $X^2 = 3.80$ ;  $df = 1$ ;  $p < 0.06$ ), while comparing the distribution of speaker gaze with listener gaze shows a significant effect ( $X^2 = 12.9$ ;  $df = 1$ ;  $p < 0.001$ ). Hypothesis 8 is thus subject to conflicting evidence. It is reasonable to assume such a hierarchy exist, but decisive evidence is lacking.

Hypothesis 9 deals with the direction of speaker gaze: in those utterances where the speaker looks both at the partner and at the system. The expectation is that if the speaker turns towards his partner during an utterance the likelihood the utterance is intended for the partner is bigger than when the speaker makes the opposite shift. Table 6 shows the number of utterances for the system and partner where the speaker turns towards the kiosk during the utterance, compared with those where the speaker turns towards the partner<sup>9</sup>.

---

<sup>8</sup> Performing significance tests on: (1) only the utterances without transition of gaze target, (2) only the utterances with transition gaze target, (3) only the utterances where listener gaze was observed at the end of the utterance, (4) only the utterances that started with listener gaze; gave no significant results.

<sup>9</sup> The utterances reported in this table are a subset of the utterances reported in the columns ‘mutual gaze’ and ‘speaker gaze’ of table 5. During those utterances it is observed the speaker looks at the listener. However it does not contain all utterances from those columns because in table 6 we only look at utterances with a *transition* of gaze target for the speaker.

**Table 6: the number of utterances for the system and the partner in those cases the speaker turns towards the kiosk and the speaker turns towards his partner during the utterance.**

| Addressee | Speaker turns towards the kiosk | Speaker turns towards the partner |
|-----------|---------------------------------|-----------------------------------|
| System    | 17 (20%)                        | 10 (8%)                           |
| Partner   | 69 (80%)                        | 108 (92%)                         |

$$X^2 = 5.52; df = 1; p < 0.025$$

There is a slight significant difference between these distributions ( $X^2 = 5.52; df = 1; p < 0.025$ ), and the difference is in the right direction. Hypothesis 9 is confirmed: when the speaker turns towards the partner during an utterance the chance the utterance is intended for the partner is larger than when the speaker turns towards the kiosk during an utterance. Or, in general the end of an utterance is more informative for the addressee of an utterance than the beginning. For the sake of completeness we may wonder to what extent these two special cases of speaker behavior provide evidence that the addressee is the partner. Compared to the overall class skew the case where the speaker turns toward the kiosk does not provide such evidence ( $X^2 = 0.19; df = 1; p < 1$ ) while the case where the speaker turns to the partner does provide such evidence ( $X^2 = 11.5; df = 1; p < 0.001$ ).

#### 3.4.4 Conclusions coordination within communicative acts

In this section, we have argued and shown that keeping track of who is looking at whom provides us with evidence about who is talking to whom. We tried to find evidence that the speaker is addressing the partner by seeking out signs of coordination between speaker and partner. Indeed, despite the situational attractor effect, we were able to find such signs. The occurrence of mutual gaze and speaker gaze toward the partner indicate the speaker is addressing the partner. The behavior of the listener is, contrary to our expectations, not such a sign. Apparently both the roles of addressee and side participant allow the listener to look at the speaker. We did find indications that the coordination between speaker and addressee differ in strength, mutual gaze being a stronger sign than speaker gaze, and speaker gaze stronger than listener gaze. However, the evidence for this hierarchy was not decisive. Finally we found that, when the speaker switches his visual attention from the kiosk to his partner during an utterance, this is a stronger indication that the utterance is intended for the partner than when he turns his attention the other way around. The gaze at the end of an utterance is a stronger indication of addressee than gaze at the beginning of an utterance.

There are two points of discussion that mark the path from the evidence found here to the classifier we present in Chapter 4. First, the results of this chapter have placed question marks behind the intention to measure the gaze behavior of the listener. For

example rather than using the distinction between mutual gaze, speaker gaze, listener gaze, and kiosk gaze, we may use the distinction mutual gaze, single gaze (either the speaker or the listener at the other), kiosk gaze. However, from a technological point of view, this has little added value. We still need to measure the gaze of both participants to make the new distinctions, and the difference in the classifier would also be small. The only technological advantage is that we do not have to keep track of who the speaker is. Given the fact that there is a difference between measuring speaker gaze (providing evidence for partner directed speech) and measuring listener gaze (providing no evidence about the addressee of the utterance) excluding the distinction does force us to throw away information. In our view the slight technological advantage is not big enough to justify doing this. A second, more general point is that the findings presented are not all-inclusive. We have tested those aspects of the behavior of the participants for which we felt there was enough theoretical grounding. For example with transition of gaze target we focused on the behavior of the speaker and did not examine the same effect for the listener. Likewise, we could have included a detailed analysis of utterances where mutual gaze occurs during a part of the utterance or we could have analyzed utterances where speaker gaze follows listener gaze and so forth. The possibilities are vast. In a chapter like this, that points back to the behavioral studies presented in chapter 2 on the one hand and forward to the classifier presented in chapter 4 on the other hand, such free explorations have no place. But it would be misguided to use only the distinctions tested here without any further thought. So, rather than taking the specific hypotheses of this chapter as guidance for our classifier we take the main points as guidance. First, making a distinction between mutual gaze, speaker gaze, listener gaze and kiosk gaze is important. Second, in the case of a transition of gaze target, the end of the utterance is most informative for establishing the addressee of the utterance.

---

## 3.5 Speaking Styles

---

### 3.5.1 Speaking styles versus utterance length

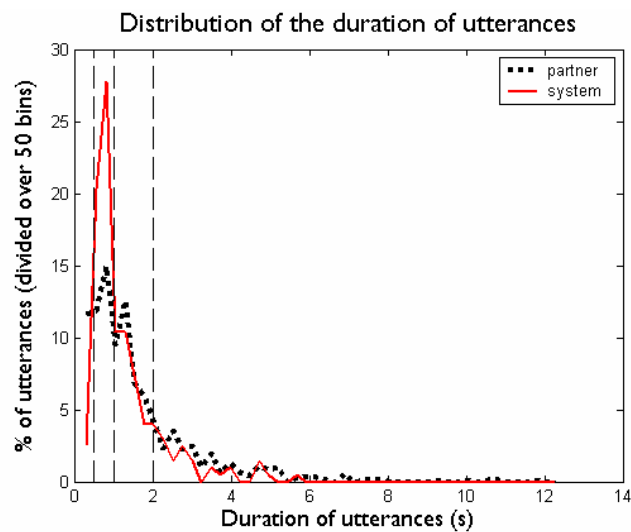
In chapter 2 we cited research claiming that, in anticipation of the limited abilities of conversational agents, people tend to omit linguistic complexities. Thus we expect people to use shorter utterances when speaking to the system (Oviatt, 1999; Katzenmaier et al. 2004). As a result observing a long utterance can be treated as evidence that the utterance is intended for a conversational partner, while observing a short utterance can be considered as evidence the utterance is intended for the system. The difficulty is to decide what short and long utterance durations are. Therefore we will compare utterance duration distributions rather than specific lengths specified in advance (see hypothesis 10).

### Hypothesis 10

There is a significant difference between the distribution of the duration of utterances intended for the system and the duration of utterances intended for the partner.

#### 3.5.2 Results for utterance length

Figure 3 shows the distribution of utterances for the system and partner. The plot shows a histogram of the duration of utterances taken over 50 equal bins. The solid line shows utterances for the system, the dotted line utterances for the partner. Both histograms are scaled in such a way that the area under the graph adds up to 100% of the utterances. In other words this is a relative graph, the fact that there are more utterances intended for the partner is not visible.



**Figure 3: The distribution of the duration of utterances for the system (solid line) and utterances for the partner (dotted line), relative to the total number of utterances for the system respectively the partner. Based on a histogram of 50 equal bins.**

The plot shows that the distribution of utterance duration differ in particular for shorter utterances. We have highlighted four areas of interest. Utterances that are shorter than 0.5 second tend to be for the partner. This is probably because of the fact MATIS does not ask for spoken confirmations (yes or no) while participants do use these among each other. Between 0.5s and 1s utterances tend to be intended for the system. Probably short utterances such as those only consisting of a station name fall in to this category. Between 1 and 2 seconds this higher frequency of utterances for the partner is disappearing, to be reversed after 2 seconds. Table 7 summarizes these effects in numbers.

**Table 7: the number of utterances for the system and the partner, with a duration: shorter than 0,5s, between 0,5s and 1,0s, between 1,0s and 2,0s and longer than 2,0s.**

| Duration | System   | Partner   |
|----------|----------|-----------|
| <0.5s    | 19 (9%)  | 126 (17%) |
| 0,5-1.0s | 83 (41%) | 152 (21%) |
| 1.0-2.0s | 73 (36%) | 284 (39%) |
| >2.0s    | 27 (13%) | 161 (22%) |

$$X^2 = 31,1; df = 3; p < 0,001$$

We see that many utterances for the partner are between 0.5s and 1.0s long, while the duration of utterances for partner is more evenly distributed. Since these distributions differ significantly we consider hypothesis 10 confirmed.

### 3.5.3 Conclusions speaking styles versus utterance length

In this section we have argued and shown that measuring utterance length can be used as a tactic to detect differences in speaking styles between speaker and partner. Many utterances for the system are between 0.5 and 1.0 seconds, while the duration of utterances for the partner are more evenly distributed. Therefore utterance length can be used to make inferences about whom the addressee of an utterance is. The strong concentration of utterances for the system in the range between 0,5s and 1,0s is probably because many utterances for the system only contain station names. This is fine for us, with our system, but it raises concerns about the general applicability of this way to use differences in speaking styles. This is an issue we return to in the next section.

---

## 3.6 General Discussion

---

In this chapter we have worked our way from the descriptive theory in chapter 2, to concrete approaches that help to decide who is talking to whom in each communicative act in our specific design case. This has been a successful effort. We have operationalised the 3 strategies for inferring who is talking to whom: dialog history, coordination within communicative acts, and speaking styles into specific tactics for our case. These tactics: dialog events, head orientation and utterance length turned out to be useful. To position the results of this study we need to relate it to the results of the theoretical framework in chapter 2 and the plans for the study in chapter 4. Also, we need to discuss the applicability in other situations than our specific design case.

In this chapter we have used the survey and results of chapter 2 to formulate hypotheses about how to infer the addressee of an utterance in our case. Most of these hypotheses were confirmed. In a way this is support for the theory, but the goal of the

experiments was not to test the theory and the experimental results should not be interpreted this way. The primary value of using the theory of language as coordinated action is that it served as a source of inspiration and as a structuring device. An example of the guidance the theory of language as coordinated action can bring is provided in section 3.4 about coordination between communicative acts. Rather than exploring all possible gaze behaviors speakers and listener might display, we could target towards the most important findings. At the same time this shows the limits of the theory. Because we do not understand the gaze behavior of addressees and side participants well enough, we might have missed behavioral cues of the listener. A blind spot of the theory may become a blind spot in the technology if we are too dogmatic in our uses of this theory. Since we tested inferential hypotheses rather than behavioral hypotheses we have covered much ground for chapter 4. In particular, the results of section 3.3 and 3.5 can be directly implemented in a classifier. As we have mentioned, we will take some more liberty with the implementation of the results about coordination within communicative acts. The main questions remaining for chapter 4 are how well each approach for detecting the addressee of an utterance works and how they can be combined to increase the robustness of a classifier.

We may ask to what extent we can use the strategies that formed the starting point for this chapter and to what extent the tactics that are a result of this chapter hold in other situations. We cannot say much about using the dialog history of human-human interactions because we worked with dyads, but we did show that the events we can measure at the system side were informative. For the application of this tactic in other situations it is important to realize that not all expressions have the same imperative for the users. We have argued that the imperative to react to system expressions depends on the type of dialog act (for example a question or a confirmation) and the modality of expression (spoken words have a time imperative while text on the screen has not). This will probably determine to a large extent how successful a particular dialog event tactic is. Coordination within communicative acts looks like a widely applicable strategy, and there is room for broadening the range of backchannel signals that could potentially be used. The approach to use head orientation rather than eye gaze can only be applied if there is a limited number of potential targets of interest that are sufficiently spatially separated. If a good indicator of gaze between speakers can be found is it likely a useful tactic for many situations, since we were able to employ it despite the situational attractor effect. The differences in speaking styles strategy may be the least generally applicable one. Speaking styles are dependent on user expectations (among other things). This means that if conversational agents become more complex and better the strategy becomes less useful. The success of our utterance length tactic may be because of the form filling interface we used, with its corresponding short utterances for the system. Conversational agents with a more conversational character may not be able to apply this tactic.

## Chapter 4: Automatic Addressee Determination

*In this chapter<sup>1</sup>, we present a naïve Bayes classifier that infers the addressee of each utterance based on information about dialog events, head orientation and utterance length. We evaluate this classifier, nicknamed AAD, by using ROC (receiver operating characteristics) curves. The classifier turns out to have an AUC (area under the curve) of 0,83. Besides inferring the addressee after an utterance has finished we also test if we can infer the addressee of an utterance earlier. To some extent this turns out to be possible during the utterance and even before an utterance has started. Finally we compare our approach to that of others working on similar problems. However, such a comparison is nearly impossible because these related studies build on different corpora, use different classification methods and different evaluation metrics. We conclude there is a need for benchmarking.*

---

<sup>1</sup> The manuscript of this chapter is partly based on the publication Van Turnhout et al (2005) and an internal technical report (Van Turnhout, 2006)



---

## 4.1 Introduction

---

Having settled that we can use dialog events, head orientation and utterance length to arrive at estimates about whom is talking to whom during each communicative act, this chapter focuses on a stochastic model that can deliver such estimates. We present a naïve Bayes classifier, designed to provide estimates about the addressee of each utterance, that we nicknamed AAD (short for Automatic Addressee Determination). We evaluate AAD using ROC (Receiver Operating Characteristics) curves.

First, we treat our classification problem with a traditional corpus based approach. The disadvantage of such an approach is that we deliver a post-utterance classification. From an interaction design perspective this is not ideal, in particular because of classification errors. Say, for example, the system rejects an utterance that is intended for the system. The right time to tell users about this rejection may not be after the utterance. Presumably, users want to know earlier. Besides asking themselves ‘did the system understand what I have said?’, users may also wonder ‘can I speak to the system now?’, and, ‘is the system listening?’. At least with their human counterparts they do seek out such early evidence even if it is of low quality (see Chapter 2.4). To open the possibility to design such early messages, in this chapter we examine the possibility of early estimates of addressee-hood. We will explore to what extent we can adapt AAD to give such early estimates.

This chapter is organized as follows. In section 4.2 we describe the design of AAD. In section 4.3 we discuss the post-utterance classification results and in section 4.4 we discuss the early classification results. In section 4.5 we draw general conclusions and compare our results to related work.

---

## 4.2 A description of AAD

---

### 4.2.1 Feature descriptions

AAD is a naive Bayes classifier. Input for this classifier are features relating to the three approaches for addressee determination we discussed in chapter 3: preceding dialog events, head orientation and utterance length. For each feature we distinguish several mutually exclusive behavioral classes that differ in the class skew of utterances for the system and partner. Table 1 lists these classes.

**Table 1: Feature Definition for Preceding Dialog Event, Head Orientation and Utterance Length**

| Preceding Dialog Event | Head Orientation      | Utterance Length      |
|------------------------|-----------------------|-----------------------|
| No relevant event      | Kiosk gaze            | Shorter than 0,5s     |
| MATIS' question        | Speaker gaze          | Between 0.5s and 1.0s |
| tap 'n talk press      | Listener gaze         | Between 1.0s and 2.0s |
|                        | Mutual gaze           | Longer than 2.0s      |
|                        | Ends in Kiosk gaze    |                       |
|                        | Ends in Speaker gaze  |                       |
|                        | Ends in Listener gaze |                       |
|                        | Ends in Mutual Gaze   |                       |

In chapter 3 we have examined 4 *dialog events*: questions from the MATIS system, tap 'n talk button presses (by users), field fill button presses (by users) and text-out events (from the MATIS system). We found that utterances directly following MATIS questions and tap 'n talk buttons are likely to be intended for the system. It was not possible to say what the intended addressee of utterances following one of the other two other dialog events was. Therefore for the feature *preceding dialog event*, we make a distinction between utterances directly following a MATIS question, utterances directly following a tap 'n talk press, and all other utterances (not directly following a dialog event or following a text-out or field fill button press event).

For *head orientation* we made three distinctions, resulting in 8 behavioral classes. We distinguish between two possible gaze targets for each participant: kiosk or partner. We distinguish between speaker and listener and we make a distinction between utterances where a transition of gaze target occurs and utterances where such a transition does not occur. With *kiosk gaze* we mean that both speaker and listener look at the kiosk. With *speaker gaze* that only the speaker looks at his partner. With *listener gaze* that only the listener looks at his partner, and with *mutual gaze* that both speaker and listener look at their partner. We have already discussed (chapter 3) that some of these classes, for example the class *listener gaze*, may not contain a clear indication about whom the addressee of the utterance is. However, it would be bad for the classifier to combine these classes with classes that do contain such information about the addressee, for example *kiosk gaze* (an indication the utterance is intended for the system) or *speaker gaze* (an indication the utterance is intended for the partner). We elaborate this point later in this section. In line with our finding that, when a transition of gaze target occurs, the gaze

at the end of utterances is most informative, we subdivide those utterances on the basis of the gaze target in the last frame. We have shown this distinction makes sense for speaker gaze, here we assume it makes sense for partner gaze as well.

For *utterance length* we make a distinction between the different regions of interest depicted in figure 3 of chapter 3. An utterance shorter than 0.5s is likely to be for the partner. Utterances between 0.5s and 1.0s are likely to be for the system. For utterances between 1.0s and 2.0s it is undecided and utterances longer than 2.0s are likely to be for the partner. We could use a different basic model for the utterance length feature. The distribution of the duration of utterances for both partner and system is roughly Gaussian when we plot the logarithm of the duration instead of the absolute duration (see Appendix B). By estimating mean and standard deviation for this new distribution we can also arrive at estimates of the probability the utterance is intended for the system based on its duration. However, since the number of parameters we have to estimate from the dataset is the same, and the performance similar to the class model presented here we choose to present only the class model.

In general, the choice of classes is, by default, somewhat arbitrary. For example, there is no fundamental reason why we could not have treated the *preceding dialog events* feature as a feature with only the two classes ‘relevant dialog event’ and ‘no relevant dialog event’. Using fewer classes has the disadvantage of throwing away information, but it reduces the risk of classes with only a few utterances. Since we estimate the probability an utterance is intended for the system from the relative amount of utterances for the system and utterances for the partner on a limited dataset, by chance this ratio may be different from the ‘real’ probability if we have too little data to rely on for our estimates. In other words: classes that contain only a few utterances introduce a generalization error. Therefore, we have chosen to use a fairly small number of classes that incorporate the findings of chapter 3. We have chosen to combine classes that contain little information about the addressee with other classes that do not contain such information but not with classes that do contain such information. So in the *preceding dialog events* feature we combined utterances after the ‘neutral’ text out and field fill button presses with utterances not occurring after a dialog event, but in the head orientation feature we did not combine listener gaze with speaker gaze. In practice, small changes in the choice of classes have little effect on the classification results.

### 4.2.2 Classification

We divide the dataset in a training set and a test set. We use the training set to estimate the (posterior) probability an utterance is intended for the system using Bayes Law (see for example: Duda, Hart & Stork, 2001, pp 22). For each behavioral class ( $x$ ) of each feature ( $f$ ) we calculate the, posterior, probability an utterance that falls in this class is intended for the system, based on the training set (equation 1) .

**Equation 1: posterior probability an utterance is for the system given that we observe it belongs to class  $x$  of feature  $f$ .**

$$P(system | x, f) = \frac{N(system | x, f)}{N(system | x, f) + N(partner | x, f)}$$

$N$  is the number of utterances falling in the observed class ' $x$ '. ' $x$ ' can have the values shown in table 1. Say we observe the utterance follows a MATIS question. The posterior probability the utterance is intended for the system, is the number of utterances for the system after a MATIS question divided by the total number of utterances following a MATIS question.

Next, we combine the features. Under the assumption that each feature gives an equally reliable and independent estimate about the addressee of the utterance we can simply calculate the numerical mean of the three probabilities (Equation 2)<sup>2</sup> .

**Equation 2: probability an utterance is intended for the system given measurements  $f_1, x_1$ ,  $f_2, x_2$  and  $f_3, x_3$**

$$P(s) = \frac{P(s | x_1, f_1) + P(s | x_2, f_2) + P(s | x_3, f_3)}{3}$$

With equation 1 and 2 we can use the data of the training set results to construct a table listing the probability an utterance is intended for the system for all combinations of feature values. For classification we can extract the feature values of the utterances in the test set, and look up the probabilities belonging to those values. Finally, we apply a threshold, classifying utterances with a probability above this threshold as intended for the system and all others as intended for the partner. The value of the threshold can be chosen at will, depending on whether we want a strict or a more lenient classification.

---

<sup>2</sup> We could have used a product instead of a sum. From a mathematical perspective this would be preferable. However, if by accident one of the features delivers the value 0 while others claim there is a possibility the utterance is intended for the system we cannot combine features any more. So there is a practical advantage of using a sum.

### 4.2.3 Evaluation

In a two-category classification problem common measures to evaluate classifiers are precision, recall and the f-measure (Fawcett, 2004). However, we decided to use ROC curves (see Duda et al. 2001, pp 48-51; Fawcett, 2004). The advantage of these curves is that they enable class skew independent evaluation, where precision and recall do not. Since we have many more utterances for the partner than utterances for the system, and because this class skew is, in part, a result of our experimental paradigm (see chapter 3) ROC curves have our preference. ROC curves also allow to evaluate the performance of a classifier across a range of possible thresholds<sup>3</sup>. This is convenient as we do not know if different types of errors have the same effect on the user. Users may find falsely rejected utterances for the system much more disturbing than falsely accepted utterances for the partner.

On the vertical axis of an ROC curve we plot the hit rate (equation 3):

**Equation 3: Hit rate**

$$\text{Hit - rate} = \frac{N(i_s c_s)}{N(i_s c_s) + N(i_s c_p)}$$

**Table 2: Graphical depiction of Hit rate**

|       |             |                   |
|-------|-------------|-------------------|
|       | $c_s$       | $c_p$             |
| $i_s$ | <b>Hit</b>  | Miss              |
| $i_p$ | False Alarm | Correct Rejection |

In equation 3,  $i$  equals the intended addressee and  $c$  equals the prediction of the classifier. The hit- rate is the ratio of the number of utterances intended for the system, that are correctly classified, divided by the total number of utterances intended for the system. In other words: the hit rate is the number of hits divided by the sum of the number of hits and misses (Table 2).

On the horizontal axis we find the false alarm rate (FA-rate, Equation 4).

<sup>3</sup> Clearly this is not a decisive argument against precision and recall, because we could plot precision and recall across different thresholds as well.

**Equation 4: FA-rate**

$$FA - rate = \frac{N(i_p c_s)}{N(i_p c_s) + N(i_p c_p)}$$

 $i_s$  $i_p$ **Table 3: Graphical depiction of FA-rate**

|       | $c_s$              | $c_p$             |
|-------|--------------------|-------------------|
| $i_s$ | Hit                | Miss              |
| $i_p$ | <b>False Alarm</b> | Correct Rejection |

The FA-rate is the ratio of the number of utterances intended for the partner that are wrongly classified divided by the total number of utterances intended for the partner. In other words: the FA-rate is the number of False Alarms divided by the sum of the number of False Alarms and Correct Rejections (Table 2).

We can construct ROC *curves* for our classifier by adjusting the threshold that settles the decision bias. If we want to simplify the evaluation by expressing the quality of the classifier in a single number we can take the *area under the curve* (AUC). The AUC for a random classifier is 0.5 (Fawcett, 2004). Using *N-fold cross* validation, we are also able to estimate the generalization error. We evaluate the classifier in multiple trials each with a different test and training set. The size of training and test set is not critical, as long as we use a bigger training set than test set. If the number of parameters that we need to estimate from the data is small, compared with the number of data points, the reported generalization error is more or less independent of the exact size of test set. (Duda et al. 2001, pp 483-485). We have chosen to obtain the class conditional probabilities from a training set of six out of eight pairs (24 dialogs) and we test classification results on a test set of 2 pairs (8 dialogs). We use all 28 possible combinations of test and training set. The generalization error is visualized in *confidence bands* for our ROC curves using vertical averaging (see: Fawcett, 2004; Macskassy & Provost, 2004).

---

### 4.3 Post-utterance classification

---

Figure 1 contains the ROC graphs for the individual features, dialog events, head orientation and utterance length. Figure 2 shows the combined ROC of these features. The thick line stands for the average performance over all test and training sets, the dotted lines the 95% confidence intervals (vertical averaging) and the dashed line chance performance.

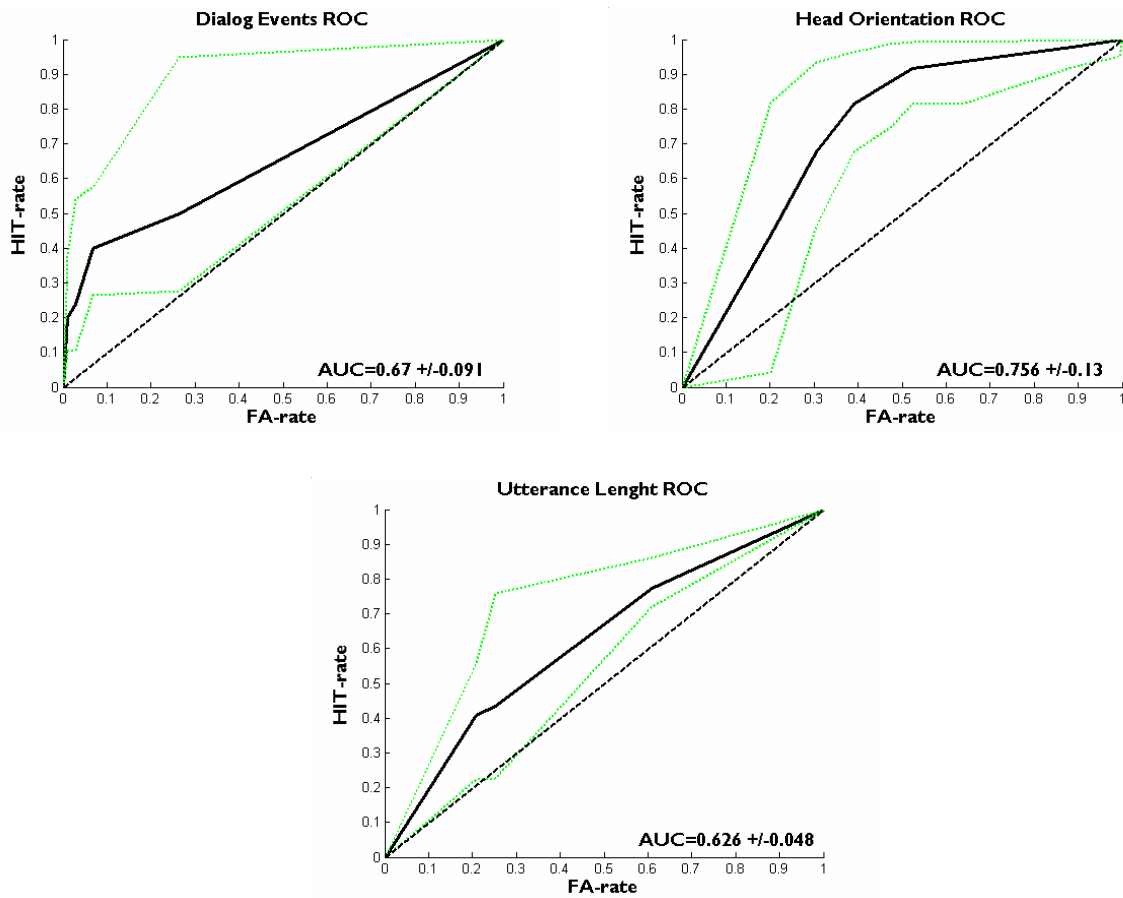


Figure 1: Roc curves for independent features: dialog events (top left), head orientation (top right) and utterance length (bottom).

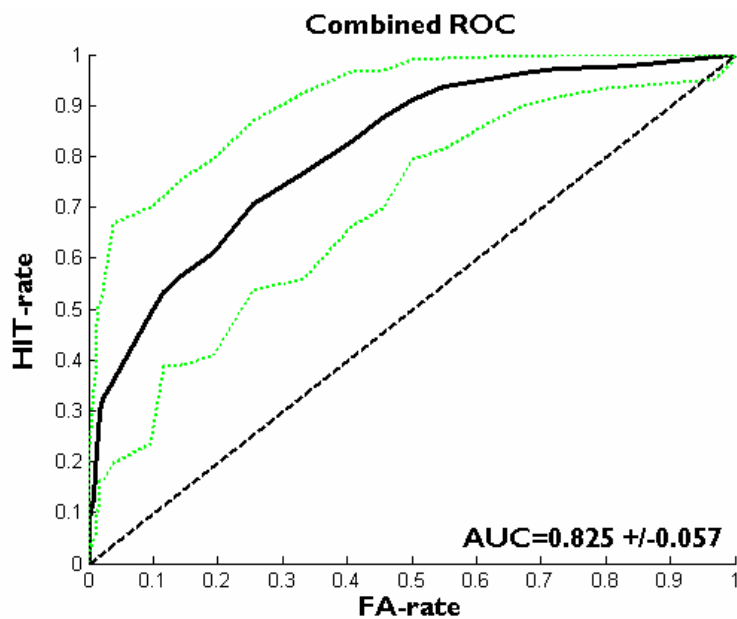


Figure 2: ROC curve for the combined performance of all three features.

The dialog events feature easily produces hits. In other words it is good at correctly identifying utterances that are intended for the system. However, it works only on a few utterances, and therefore avoiding false alarms at higher hit rates is hard. Head orientation is a stronger feature, with a reverse bias. It easily avoids false alarms, utterances for the partner that are identified as utterances for the system, at high hit-rates but it does not produce hits easily at high false alarm rates. These two features complement each other well. The ROC curve for utterance length is symmetric, but utterance length is a weak feature. It has difficulties in producing hits as well as in avoiding false alarms. The combined performance (figure 2) is reasonably symmetric, and can be tailored towards good hit performance as well as good alarm performance. Combining good hit and alarm performance is still difficult.

We may wonder what the added value of each feature for the final classification is. Table 5 lists the independent AUCs of each feature and the quality of the final classifier without this particular feature. For each AUC we list the 95% variance.

**Table 4: independent AUC's and added value for each feature**

| Feature             | Independent         | All but-this feature |
|---------------------|---------------------|----------------------|
| Dialog events       | 0.67 (+0.09)        | 0.78 ( $\pm$ 0.10)   |
| Head Orientation    | 0.76 (+0.13)        | 0.72 ( $\pm$ 0.07)   |
| Utterance Length    | 0.63 (+0.05)        | 0.82 ( $\pm$ 0.07)   |
| <b>All features</b> | <b>0.83 (+0.06)</b> |                      |

In line with the good individual performance of the head orientation feature, excluding it has the largest effect on the combined classification performance. Excluding the utterance length feature has hardly any effect on the performance. In general combining features reduces the variance, meaning that combining information from different sources strengthens the robustness against differences in the behavior of the different pairs. However, this observation should be taken with some care, because we did not estimate the variance of the generalization error. Still we have chosen to include all features in our final classifier (rather than dismissing utterance length) because of this decreased generalization error.

---

#### 4.4 Early estimates

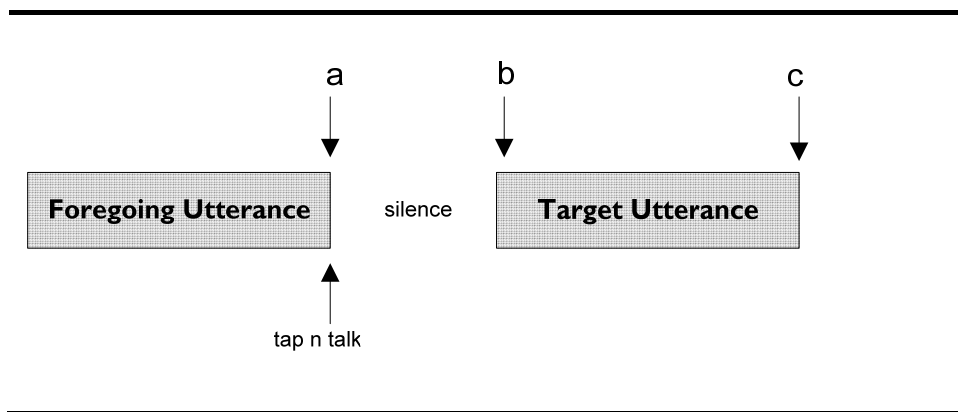
---

In our introduction we mentioned we would like to open the possibility to design early feedback about the inferences of our classifier. Therefore we need to adapt our classifier



to give early estimates about addressee-hood, and we need to assess the quality of this classification at different moments in time. We will look at two types of early estimates. We try to see whether we can make an accurate prediction about the addressee of the utterance *during* the utterance, rather than when it is finished. And we will try to see whether we can predict the addressee of an utterance *before* it has even started. We have chosen to adapt our classifier, rather than designing a new version because we expect that much of the evidence captured in our current features is already available in an early stage.

In the discussion that follows we will label the utterance for which we want to know the addressee the *target* utterance and the last utterance before that, the *foregoing* utterance (see Figure 3). In case users press a tap ‘n talk button, we start our analysis at this button press, in that case there is no foregoing utterance.



**Figure 3: reference frame for discussing utterances.**

We will report the classifier performance at three points: at the end of the foregoing utterance (a), at the beginning of the target utterance (b) and at the end of the target utterance (c). The time stretch between these points varies a lot. Both silence and utterance durations range from zero to a few seconds. Intermediate results, such as the performance a second after the target utterance has started, are therefore too complicated to interpret and report here. Point (c) is, of course, the post-utterance classification reported earlier.

In order to adapt the classifier to predict the addressee of the target utterance (point a) we may train the classifier with the feature values of the foregoing utterance and the addressees of the target utterance. However, we change the feature set in two ways. First, we do not distinguish between speaker and listener gaze. Occasionally the speaker of the foregoing utterance is the MATIS system. So either we should extend the feature set, to include MATIS as speaker or we should omit the distinction. Since the speaker of the target utterance is still unknown when the foregoing utterance is finished we have chosen for the latter, simpler adaptation. Second, we do not use the utterance length feature. It

cannot be used at all in those cases participants press a tap ‘n talk button, but in the other cases the feature turned out to do no better than chance. In other words the duration of the foregoing utterance is not informative for the addressee of the target utterance. If we want to predict the addressee of the target utterance at the end of the silence before the target utterance (point b) we use the feature values of the silence with the addressee of the target utterance. For the same reasons as with ‘point a’ we excluded the utterance length feature<sup>4</sup> and the distinction between speaker and listener gaze. For point c we can of course use the full feature set.

Although we do not report intermediate results between points a,b and c we need to consider how to use the classifier in those intermediate intervals. The problem here is how we deal with the knowledge that we are at an intermediate point. Say, we are 3 seconds after starting (point b) and the target utterance is still ongoing (the speaker has not finished speaking yet). If we act ignorant to the knowledge that the utterance is still ongoing, we can compare the feature values of this utterance at that point with feature values of all utterances. However, it would be more correct to compare feature values of this utterance with the feature values of utterances that are -3 seconds after starting (point b)- still ongoing. But this is a much smaller set of utterances. The practical problem with this ‘correct approach’ is that the amount of training data decreases in time. Therefore we have chosen to use the simple approach and to act always as if we have obtained feature values for a finished utterance<sup>5</sup>.

The shapes of the ROC curves of each feature are similar to those of the matching graphs in figure 2. Therefore table 5 lists only the AUC’s and 95% confidence intervals at the three points of interest for each feature separately and the combined classifier.

---

<sup>4</sup> Applying the utterance length feature on silences, a silence length feature, was explored but performed no better than chance.

<sup>5</sup> We have explored this problem by comparing a range of ROC curves using one of both approaches. In the beginning of utterances there was no difference between the two approaches while the simple approach started to score better after about two seconds. For utterances longer than 3.5 seconds the ‘correct approach’ could not be used at all, because some training sets did not contain utterances this long.

**Table 5: AUC's and 95% confidence intervals at different points in time: a the end of foregoing utterances, b the start of the target utterance, and c the end of target utterances.**

|                  | End of foregoing utterance (point a) | Start of target utterance (point b) | End of target utterance (point c) |
|------------------|--------------------------------------|-------------------------------------|-----------------------------------|
| Dialog Events    | 0.67 ( $\pm$ 0.09)                   | 0.67 ( $\pm$ 0.09)                  | 0.67 ( $\pm$ 0.09)                |
| Head orientation | 0.67 ( $\pm$ 0.12)                   | 0.68 ( $\pm$ 0.12)                  | 0.76 ( $\pm$ 0,13)                |
| Utterance Length | X                                    | X                                   | 0.63 ( $\pm$ 0.05)                |
| Combined         | 0.75 ( $\pm$ 0.04)                   | 0.76 ( $\pm$ 0.04)                  | 0.83 ( $\pm$ 0.06)                |

Clearly all the information about dialog events is available at the end of the foregoing utterance (point (a)). We can use the dialog events feature in all three points with the same results. For head orientation we see much information is available in an early stage. Since baseline (chance) performance would give an AUC of 0.5, we could say that about 70% of the maximum performance of this feature is already be reached in point (a)<sup>6</sup>. During the silence, little is won. The combined performance shows the same trend. 75% of the maximum performance can already be obtained early, and during the silence little is won. Since the utterance length feature gives only a small contribution to the classification results after the target utterance (table 4), the gain during the target utterance is probably because of the better performance of the head orientation feature. In all we can conclude our classifier gives reasonable early estimates of addressee-hood.

---

## 4.5 Conclusions and discussion

---

In this chapter we have described and evaluated a naïve Bayes classifier for the addressee of an utterance in our design case: AAD. We have shown AAD has an AUC of 0.83 when we use it as a post utterance classifier and we have settled that we can reach about 75% of this performance at the end of the foregoing utterance (AUC=0.75). In this section we relate these results to those of chapter 3 and the plans for chapter 5 and 6, and we compare our work to two other studies about detecting the addressee of an utterance.

In chapter 3 we claimed that we can use preceding dialog events, head orientation and utterance length for detecting who is talking to whom. In this chapter we confirmed this

---

<sup>6</sup>  $100 \cdot \frac{AUC(a) - chance}{AUC(c) - chance}$

finding. Now, we also know the relative success of these three tactics. Features relating to head orientation gave the best results, followed by dialog events followed by utterance length. Dialog events and head orientation are complementary features, and as a result combining these features improves the overall classification result. As a feature, utterance length has much less added value. Looking forward to chapter 5 we may say that implementing this classifier should and implementing it in a demonstration platform should be relatively straight forward. We can use a simple lookup method for obtaining the confidences of the classifier in this chapter. For the interaction design study in chapter 6 we have opened the possibility to design early feedback about the inferences of our classifier, by adapting the classifier to give early estimates of addressee-hood. These early estimates are reasonable, compared to the post-utterance classification results. But, the question is of course how good these post-utterance results are. We may ask what an AUC of 0.83 is worth when we employ the classifier in a real conversational agent in interaction with users. Inevitably we will have to postpone this question to chapter 6.

We may try to compare our results to that of related work. We picked two studies to do this. Katzenmaier et al. (2004) tried to identify the addressee of an utterance in human-human-robot communication and Jovanović et al. (2006) did the same for meetings. Katzenmaier et al. report on addressee-hood determination based on head orientation estimates, low-level linguistic cues and weighted combination of those two sources of information. They also found features relating to head orientation to be effective. For their case, head orientation was far more effective than their linguistic features, combining them gave a small improvement on the classification results. Jovanović et al (2006) report addressee classification efforts on a corpus of four participant face to face meetings using a large feature set consisting of features about gaze, language, utterance length, and meeting context. They report gaze to be effective and speaker gaze to be more effective than listener gaze. They also note the negative effect of situational attractors on the use of gaze features, because their results are less good than in stylized meetings (such as Vertegaal, 2001). Finally they note that using utterance length has added value in their dataset, although they do not provide a clear rationale why this should be the case. So, head orientation is generally reported to be successful and combining this with utterance length generally leads to a small improvement in classification. Additional improvements can be reached by including additional features, but the other features tested in these two studies did not have a large impact on the classification results. While comparing these studies provides some insight in the general applicability of the features we may use for addressee identification, the question what the most appropriate stochastic model for this type of problem is, is much harder to answer. We could not answer this question for our

corpus because we only tried a naïve Bayes classifier. For their corpora, Jovanović et al. (2006) report on a comparison of naïve Bayes classifiers and Bayesian networks. There, Bayesian networks outperformed the naïve Bayes classifiers. Katzenmaier et al. (2004) used neural networks for their corpus but did not compare this with other approaches. So to be able to know what is the best stochastic model for addressee identification there is a strong need for benchmark corpora. A final problem in comparing our work to other approaches is the evaluation measures used. Jovanović et al. report accuracy, but they try to distinguish between multiple addressees, which is a more difficult task than our two way classification problem. Katzenmaier et al. have a two way classification problem, but they report precision and recall without mentioning their class skew. Therefore we cannot compare their results on their corpus with our results on our corpus in an honest way. As a community we need to reach consensus on the evaluation measures we use.

## Chapter 5: Prototyping a Socially Aware Conversational Agent

*In this chapter we present a prototype of a socially aware conversational agent. Building this prototype serves two goals. First, by building a prototype we are able to explore the challenges involved in using a classifier such as we described in chapter 4 in a real system. One challenge for real time addressee determination is that a real time classifier needs to be capable of dealing with simultaneous speech and mid-utterance silences. Second, the prototype facilitates the experiments presented in the chapter 6. For practical reasons parts of the system are implemented with different technology than a real conversational agents use and other parts are solved with a wizard of Oz setup. This chapter describes the implementation of this prototype, the wizard interface and protocol.*

---

## 5.1 Introduction

---

So far, we have explored the possibility of creating socially aware conversational agents by analyzing a previously recorded corpus. In this chapter we complement this work by presenting a prototype of a socially aware conversational agent. The prototype serves two goals. First, by building a prototype we are able to explore the challenges involved in using a classifier such as we described in chapter 5 in a real system. Second, in the experiments of the next chapter we intend to confront users with several versions of socially aware conversational agents to find requirements for the interaction design of such agents. In building this prototype we focused on implementing a real-time version of the automatic addressee determination (AAD) module. Other tasks of a socially aware conversational agent, such as speech recognition and dialog management were solved with a wizard of Oz setup.

This chapter is organized as follows. First, in section 5.2, we discuss choices we made about the techno-ecological validity of this prototype. From then on we describe the implementation of this prototype. We discuss the overall architecture in section 5.3, followed by all individual modules. The head orientation tracker is described in section 5.4, the speech activity detection in 5.5 and the AAD module in section 5.6. In section 5.7 we turn to the dialog management module and the wizard protocol. Finally in section 5.8 we discuss how the MATIS interface is integrated in this architecture.

---

## 5.2 Ensuring techno-ecological validity

---

Following the objectives of this prototype: exploring the challenges involved in using a classifier such as we described in chapter 4 and facilitating the experiments in the next chapter, we focused our implementation efforts on a real time version of the AAD. As we have seen AAD needs three types of input: it needs information about the focus of attention of participants (inferred from their head orientation), it needs information about utterance length (inferred from information about on-off patterns of speech) and it needs information about dialog events of the system (received from a dialog manager). In principle, all these types of information can be delivered to the AAD with a wizard of Oz setup. However, in particular to fulfill the second objective, facilitating the experiments of chapter 6, we must consider the limitations of wizard of Oz simulations.

In general, wizard of Oz setups suffer from a lack of techno-ecological validity. In other words: the performance and behavior of the emulation does not resemble that of a

real system. For example, Den Os & Boves (2005) comment on three problems with wizard of Oz simulations. First, the recognition capacities of actual systems are much lower than those of a wizard. Second, the amount of artificial intelligence needed to mimic responses of a wizard is far beyond the current capabilities of intelligent systems. Third, the wizard has generally much stronger turn-taking capabilities, than current generation systems. While Den Os & Boves point to *over performance* of wizard of Oz setups, a *lack of performance* is also an issue. For example, Sturm, Iqbal & Terken (2006) point to the difficulty of replacing perceptual components such as head orientation tracking and speech activity detection with wizards; they found inter-coder reliability of such annotations to be below acceptable standards.

To warrant techno-ecological validity of our prototype, and, at the same time focus our efforts on the implementation of the AAD, rather than a full conversational agent, we face tradeoffs. In this section we present a review the state-of-the art of technological solutions for other tasks than the AAD to see to what extent it is possible to replace this technology with other approaches or with a wizard of Oz setup. We will first discuss head orientation tracking, followed by speech activity detection and speech-recognition dialog management.

### 5.2.1 Head orientation tracking

In recent years, a number of vision based approaches for *head orientation tracking* have been proposed and evaluated (Stiefelhagen, Yang & Waibel, 2000; Seeman, Nickel & Stiefelhagen, 2004; Voit, Nickel & Stiefelhagen, 2006; Horprasert, Yacoob & Davis, 1996). These techniques fall into two basic categories (Stiefelhagen, 2002, pp 26). *Model-based* approaches, such as reported in Horprasert et al. (1996), extract a number of features such as eyes, nostrils and lip corners and reconstruct a 3D model of the head based on those features. This approach has been successful for constrained tasks, but require high resolution cameras (Voit et al. 2006) to extract the landmark points with high reliability (Stiefelhagen 2002, pp 25). The alternative is an *appearance-based* approach, where neural nets estimate the head orientation from facial images, even of low resolution (Stiefelhagen et al. 2000). Variants of this approach include the use of stereo camera images (Seeman et al. 2004) or a combination of multiple camera images (Voit et al. 2006). Disadvantages of this approach are the lower accuracy, and sensitivity to lighting conditions, so that, when used in a new context, a new training session is needed (Stiefelhagen, 2002, pp 89-98).

We felt, using a vision based head orientation tracker such as described in the literature would take up too much of our resources. However, we felt accurate head orientation



tracking was needed to match the capabilities of vision-based head orientations trackers. While focus of attention estimates could be provided by a wizard, these estimates may not be delivered timely because of the reaction time of a wizard. We solved this problem by building our own, low cost head orientation tracker. This tracker requires participants to wear special purpose devices (see section 5.4). The advantage of this approach is that the tracking of such devices is relatively straightforward, accurate and robust against lighting conditions. A disadvantage of this approach, besides asking participants to wear a special purpose device, may be that with this approach we are capable of more accurate head orientation tracking than with the state of the art vision based technology. However, since we combine head orientation tracking with a focus of attention approach (Stiefelhagen & Zhu, 2002) this over-performance is unlikely to harm the techno-ecological validity of our system too much.

### **5.2.2 Speech activity detection**

*Speech -or voice- activity* detection is a well known problem in the field of speech recognition and various other fields. The most straightforward solution is to use the energy of the audio signal to estimate the presence of speech compared to the absence of speech. A disadvantage of this approach is that it can only be used with close- talking microphones in ‘silent’ environments. In noisy environments and with far-field microphones the problem of distinguishing speech from non-speech audio becomes apparent. A number of approaches to this problem have been proposed: computer vision based approaches such as Darrell et al. (2002), and various acoustic approaches (Armani, Matassoni, Omologo & Piergiorgio, 2003; Macho et al. 2005).

Like with head orientation tracking, using the state of the art approaches to deliver information about speech on and off events would take up many resources, while replacing this technology with a wizard would harm techno-ecological validity. We solved this problem by using close-talking microphones, for which we could use the energy of the audio signal to arrive at speech activity data. A disadvantage of this approach is that we cannot distinguish between non-speech audio and audio that is a result of speech. So because of this solution the performance of the AAD module may be impaired compared to state of the art approaches. (Informal) tests with the close-talking microphones we used showed that this problem is small in the lab setting of our experiments.

### **5.2.3 Dialog Management and Speech Recognition**

This project builds on the MATIS project (see: Sturm, 2005) where an operational system for speech recognition and dialog management was used. Unfortunately it

concerned a very old system with a pipeline architecture that is hard to adapt for a new situation. In addition the amount of speech recognition errors and frequent, untraceable, delays, hamper the usability of this system. Because of this, we felt the system did not reflect the state of the art of current speech recognition systems. In fact we felt that using the real system would harm the techno-ecological validity of our prototype. However, we did not have a robust, fast, speaker independent speech recognizer with training data for our domain. So for speech recognition we did opt for a wizard of Oz setup. As a consequence we also needed to solve part of the dialog management within this setup as well<sup>1</sup>.

In order to prevent our wizard setup from outperforming a realistic dialog manager we tried to allocate as many tasks as possible to the wizard application, and use strict protocols for other tasks. As we will describe in (see section 5.7) the wizard had only limited control over the dialog and turn-taking. In parts of the study of chapter 6 we assumed ‘perfect’ speech recognition; in other parts we also faked speech recognition errors and their effect on the dialog.

---

### 5.3 General Architecture

---

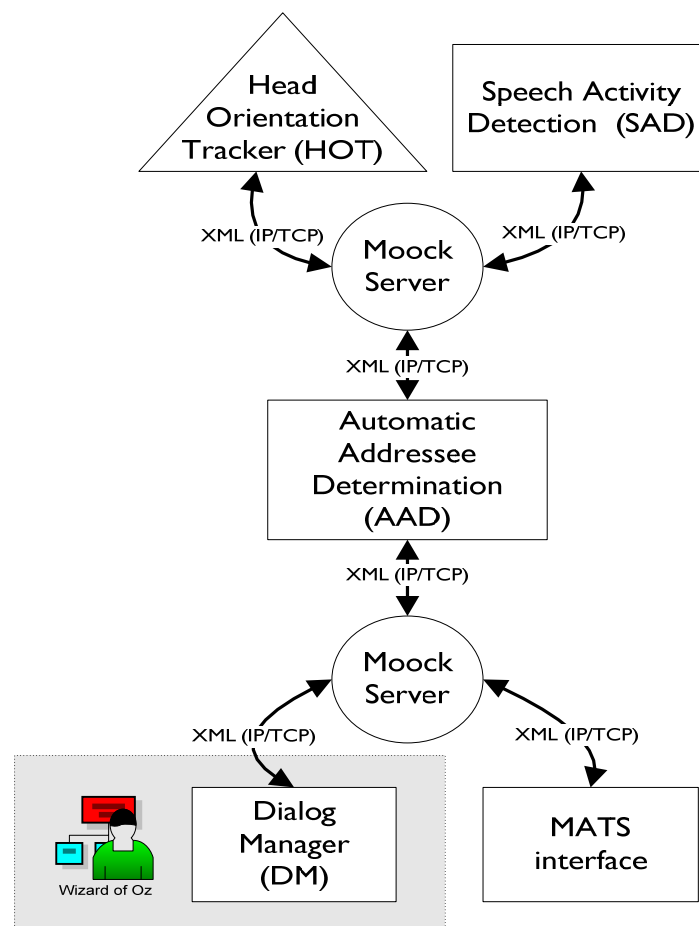
Figure 1 (on the next page) depicts the software architecture of the whole prototype platform. After providing some general information about this architecture we will discuss the individual modules in more detail.

Central in this architecture is AAD, a module that estimates to what extent users are addressing the system. It receives information about the head orientation of the participants from a Head Orientation Tracker (HOT) and about on-off patterns of speech from the Speech Activity Detection (SAD) module. Also, AAD receives information about dialog events from the Dialog Management Module (DM). AAD combines these three sources of information into estimates about the addressee of each utterance, and this is sent to both the DM and the MATIS-interface (including a summary of the raw data). Two JAVA relay servers (Clayton & Mook, 2001) take care of the communication between the modules. These ‘Mook Servers’ send all data they receive to all clients. The modules use XML formatted messages to communicate to one another, containing an ID

---

<sup>1</sup> In principle it is possible to have a wizard type everything that is being said, and feed this into a full fledged dialog manager. However, at the rate of multi-party dialogs a typist cannot match the speed of a speech recognition engine, and a separate module would be needed to deal with typing errors.

from the sender, so that other modules can see whether or not messages are relevant for them. Because the servers are accessible through the internet, all clients can run on a different computer. Time stamp tests showed that communication between the modules does not suffer from delays. Sending and receiving messages took less than 0.1 second, unless many modules ran on a single computer. To prevent delays we ran computationally demanding modules on separate computers.

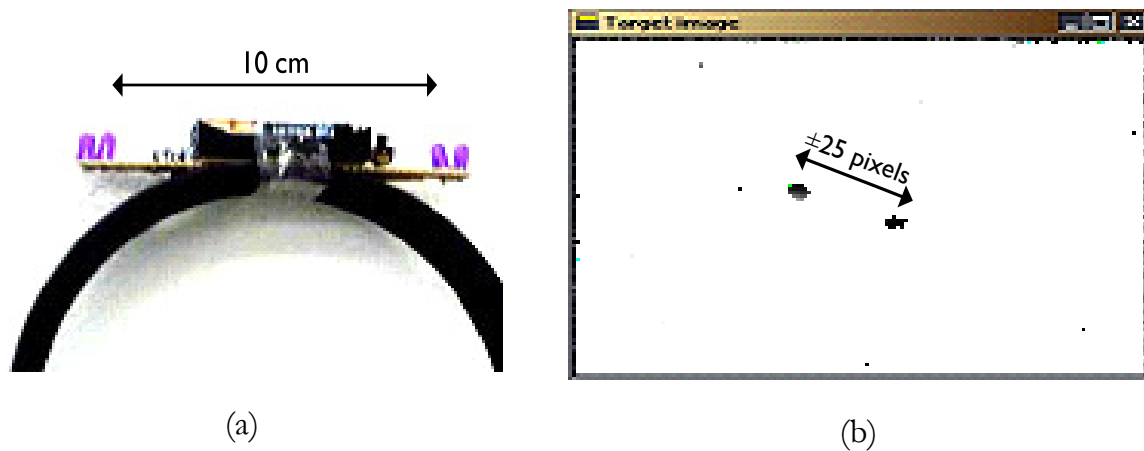


**Figure 1:** a depiction of the general software architecture of the demonstrator platform. We use different shapes to depict the platform we used to implement the different parts. Circles for JAVA, triangles for C++, and squares for Macromedia™ Flash. The dialog management is, in part, solved through a wizard of Oz setup.

## 5.4 Head Orientation Tracker (HOT)

To keep track of the head orientation of both participants we built a low cost head orientation tracker. It requires participants to wear a diadem with two infrared LEDs and a small battery on top (figure 2a). In front of a camera, in the ceiling above the kiosk, we

placed an infrared-pass filter. This way, the camera image contains only the IR-lights of the diadems of both participants (Figure 2b, only one participant is visible in the image).



**Figure 2: Photo of a diadem with battery, and two infrared led's (a) and (b) a (reverse video) screenshot of camera image seeing a single diadem.**

The software (written in C++) is based on software for presence detection using infrared reflectors (see: Raducanu, Subramanian, & Markopoulos, 2004). This version calculates the orientation of the diadems relative to the camera. Using the OpenCv library (see: OpenCV, 2005), it captures and thresholds the camera image, after which it finds connected areas – ‘blobs’, or ‘dots’-. The software rejects dots that are either too large or too small to be part of a diadem. For the remaining dots HOTS calculates a proximity matrix. Based on the heuristic that the distance between two dots belonging to a single diadem is smaller than the distance to any dot belonging to a different diadem, these dots are matched into dot-pairs. For each dot-pair HOTS can calculate a mass centre, a width and an orientation relative to the camera. Each frame the software sends an XML file with data for each diadem to the ‘Mooock Server’<sup>2</sup>. Occasionally a diadem or a single dot is not visible, for example because the participant moves outside the camera image. Therefore, in the pre-processing module of AAD, extra heuristics are used to assess the validity of a measurement and to settle which pair of dots belongs to which participant. Running on a Pentium™ 4 PC, the software enables a tracking frequency of 5 frames a second. The software can track an unlimited number of diadems (we used it for meetings with up to eight participants as well) as long as a single camera can capture all participants.

<sup>2</sup> Thanks to Eugen Schindler for implementing the xml sockets in C++

## 5.5 Speech Activity Detection (SAD)

---

The Speech Activity Detection module converts an audio signal to information about on-off patterns of speech. It was written in Macromedia Flash™. Using AKG™ C420 III close talking microphones and a Terratec™ MIC8 PCI sound card we minimized the risk of capturing cross talk or environmental noise in the audio signal. Participants who picked a position to the right of the kiosk received a microphone that SAD recognizes as right speaker, participants that picked position to the left of the kiosk received a microphone that the software recognizes as left speaker. The software simply thresholds the audio signal for each microphone, to distinguish between speech and non-speech for each speaker. Each time a transition between speech and non-speech occurs on one of the microphones, a speech on (or off) event is generated by the SAD module and an XML file with all speech information is sent to the server. To prevent jitter we applied a time threshold of 100 ms for speech off events. So only after 100ms of silence we send a speech off event to the server.

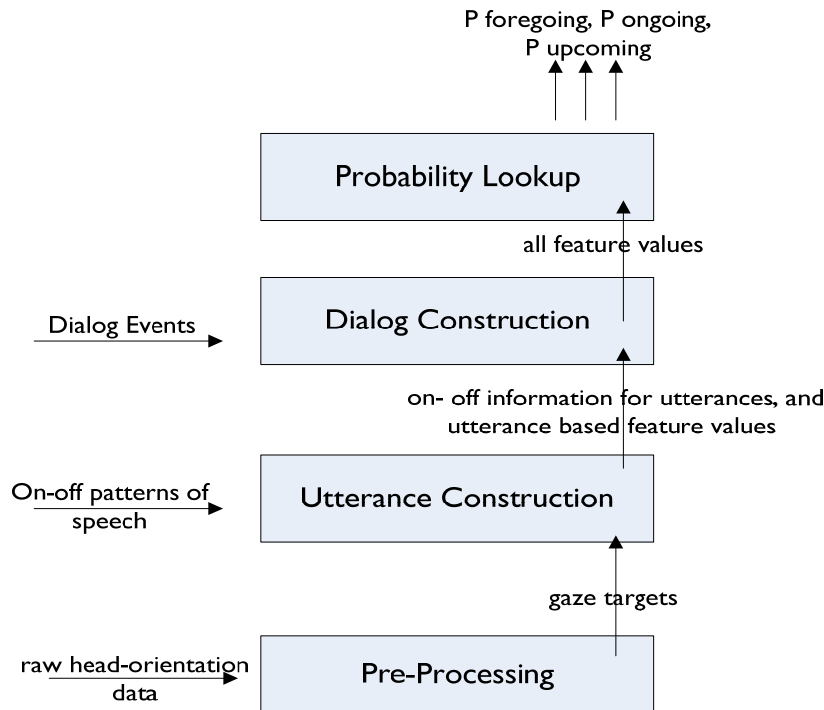
---

## 5.6 Automatic Addressee Determination (AAD)

---

### 5.6.1 An outline of the AAD architecture

The Automatic Addressee Determination (AAD) takes information from the Head Orientation Tracker (HOT), the Speech Activity Detection (SAD), and Dialog Management (DM) modules to deliver three streams of probabilities: the probability the last (or foregoing) utterance was intended for the system, the probability the current (or ongoing) utterance is intended for the system, and the probability the next (or upcoming) utterance will be intended for the system. To be able to do so, AAD needs to construct a knowledge representation from the raw data, containing the feature values for each utterance. It is written in Macromedia Flash™ and it consists of four submodules, depicted in figure 3.



**Figure 3: the four submodules of AAD, and their relation to the external system. AAD receives information about head orientation, on-off patterns of speech and dialog events. This information is parsed through several knowledge representations in order to be able to look-up the probability that the last utterance was intended for the system (P foregoing), the current utterance is intended for the system (P ongoing) and the probability the upcoming utterance will be intended for the system (P upcoming).**

First, AAD needs to transform head orientation data into information about the gaze target of both participants. It also needs to deal with missing or corrupt data from the head orientation tracker. This is arranged in a pre-processing submodule (see 5.6.2).

Second, AAD needs to combine an end-of-utterance criterion with the on-off patterns of speech to be able to use the ‘utterance’ as the unit of analysis. When one of the speakers stops speaking it does not mean the utterance is over. In our definition of an utterance (see chapter 3.2.3), if the same speaker starts speaking again within 0,5 seconds we still consider it to be the same utterance. Two of our features: (utterance length and head orientation) are based on (historical) information about events during an utterance. For example to know whether a transition of gaze target occurred we need to store all gaze targets for both participants during an utterance. The utterance construction submodule deals with these issues (see 5.6.3).

Third, AAD needs to sequence these utterances into a representation of the dialog. For example, if the Dialog Manager asks a question, AAD needs to adjust feature values for the utterance that is yet to come (the upcoming utterance). As soon as the utterance

starts, the feature values need to be maintained with, what now has become, the ongoing utterance. For that AAD needs a historical account of utterances. Also, AAD needs to deal with the possibility that simultaneous speech is occurring. This is arranged in the dialog construction submodule (see 5.6.5).

Last, once all information is organized in a knowledge representation at the dialog level, AAD needs to lookup the probabilities for the foregoing, ongoing, and upcoming utterance. This is arranged in the probability lookup submodule (see 5.6.5). Next, we will discuss these four submodules in more detail.

### 5.6.2 Pre-processing

The pre-processing submodule has the responsibility to check the raw data for validity, to match the head orientation data to the user it belongs to and to transform head orientation data into an estimate of focus of attention. First, it checks the data from the head orientation tracker for validity. The information about a diadem is considered valid if the width parameter is within a realistic range. Second it matches the head orientation data to one of the users. If two diadems are visible AAD orders them from left to right based on information about the mass center in the camera. If only one diadem is visible it calculates the shortest distance to pre-programmed ‘likely’ positions of the left and right diadem. If a diadem has been visible, but disappeared during the interaction, or if it is considered invalid, the last valid values for that diadem are substituted. Third, it converts the continuous head orientation to a binary value indicating the most likely focus of the participants’ (visual) attention: the kiosk or the partner. Therefore it applies a threshold to the head orientation data. We used a double threshold to prevent jitter. We assign the target partner if the tracker reports a head orientation larger  $50^\circ$ . We assign the value kiosk if the tracker reports a value beneath  $45^\circ$ <sup>3</sup>.

### 5.6.3 Utterance Construction

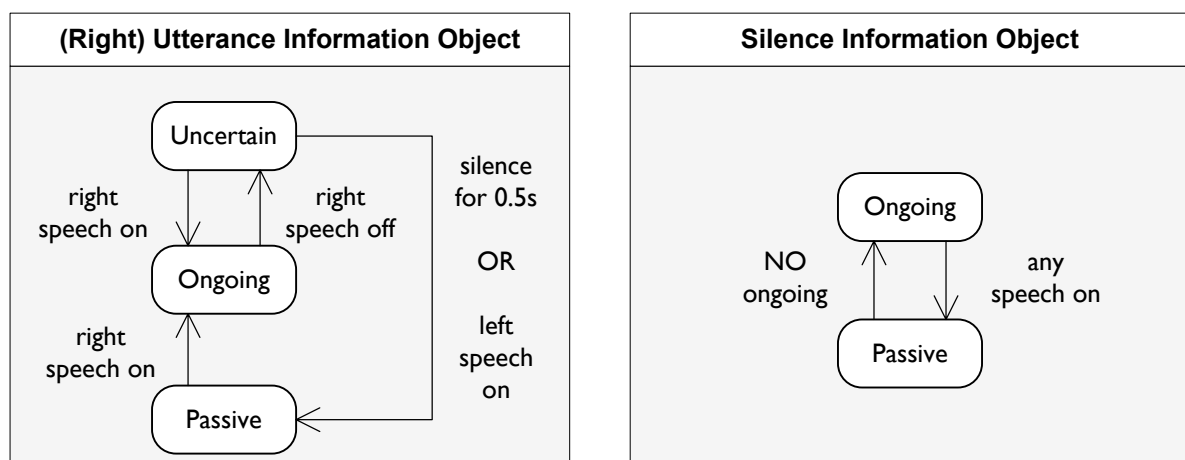
The utterance construction submodule has the responsibility to reorganize the streams of speech-on and speech-off events and head orientation data into a historical account of events during an utterance. This involves deciding when an utterance (of a single speaker) starts and when it ends and keeping a data storage for the stream of focus of attention

---

<sup>3</sup> These values were obtained by comparing the head orientation tracker with the hand transcriptions of the experiment in chapter 3. An absolute boundary anywhere between  $45^\circ$  and  $50^\circ$  performed better on this data than fitting a mixture of two Gaussians on the data per pair. The values can be the same for both participants if the camera is aligned to the kiosk (at  $0^\circ$ ) because the head orientation tracker does not distinguish between left and right orientation.

data occurring in the period between those two events. There are two problems that complicate this task: the possibility of overlapping speech (two speakers may speak at the same time) and mid-utterance silences (a silence shorter than 0.5s during an utterance). The utterance construction and dialog construction submodules share the responsibility to deal with these two problems. The utterance detection module has to make sure no data is lost in these cases (the dialog construction module takes care of sequencing).

The utterance construction module makes sure no data is lost during overlapping speech by keeping a data storage for each speaker and for mid-utterance silences by storing data during silences until the end-of-utterance-criterion is reached. It contains two ‘*utterance information objects*’, one for each speaker and a ‘*silence information object*’. *Utterance information* objects can have 3 states ‘passive’, ‘ongoing’ and ‘uncertain’. An utterance information object, for example, for the right speaker is in passive state when the right speaker is not speaking, in ongoing state when the right speaker is speaking and in uncertain state when the right speaker is not speaking but it is still unclear whether this is a mid-utterance silence or the end of the utterance. Figure 4 shows a state chart for utterance information objects.



**Figure 4: state chart for utterance information objects (the version for the right speaker is shown) and silence information object.**

In *passive* state, utterance information objects wait for a speech-on event of the speaker they listen to. They do not store data. In ongoing and uncertain state, utterance information objects maintain an up to date array of the focus of attention data. The gaze of the speaker they listen to is listed as speaker gaze, the focus of attention of the other participant as listener gaze. A parameter indicating whether the utterance information was in ongoing or uncertain state at the moment of storing the data completes this data storage. Speech-on events for their speaker trigger the utterance information objects to go in ongoing state. At speech-off events utterance information objects can move from



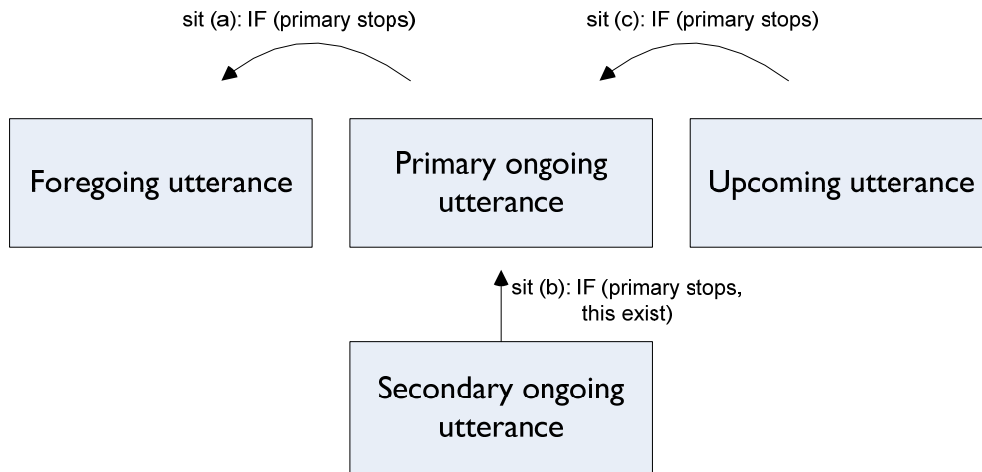
ongoing to uncertain state (see figure 4). In uncertain state, utterance information objects can move back to ongoing state or to passive state. They go to ongoing at a speech on event, and to passive after a time lapse of 0.5s, or when the other person starts speaking during this period. At a move from uncertain to passive state focus of attention data that is marked as uncertain at the end of an utterance is discarded. This way, focus of attention information stored during mid utterance silences is treated as part of the utterance but the focus of attention information of the period needed to reach the end-of-utterance-criterion is not.

*Silence information objects* store focus of attention information during silences (needed for predicting the addressee of the upcoming utterance). They have two states: ongoing (a data storage is maintained) and passive (waiting for a silence to start). They are only in ongoing state when no participant is speaking. They move to passive state when one of the participants is speaking.

Both utterance information objects and the silence information object send an utterance start event to the dialog construction submodule when they move out of the passive state and an utterance end event when they move into passive state.

#### **5.6.4 Dialog Construction**

The dialog construction submodule is responsible for sequencing utterances and for completing the data collection by adding information about dialog events. Like the utterance construction module, it has to be able to deal with the cases of simultaneous speech and mid-utterance silences. To be able to do so the module fills an abstract model for dialog history with concrete utterance information. It has containers for information about utterances that are organized according to their place in the dialog: a foregoing utterance container (for information about the last utterance that has finished), two ongoing utterance containers (for information about utterances that have not finished yet) and an upcoming utterance container (for information about an utterance that has not even started). It needs two containers for information about the ongoing utterance to be able to deal with simultaneous speech. These containers are labeled: primary ongoing utterance container and secondary ongoing utterance container (see figure 5). To be able to deliver probability estimates for utterances yet to come and to be able to deal with mid-utterance silences the ongoing utterance containers can also contain information about silences.



**Figure 5: the dialog construction module has 4 containers for information about utterances. The arrows point out what happens at an ‘end of utterance’ event. The information of the primary ongoing utterance container is copied to the foregoing utterance container (sit (a)), information about the secondary ongoing utterance (if it exist) is copied to the primary ongoing utterance container (sit (b)) and information about the upcoming utterance (if dialog events have occurred) is added to the new primary ongoing utterance (sit (c)). Further information can be found in de text.**

The dialog construction submodule uses the utterance start and end events from the utterance information and silence information objects in the utterance construction module to fill its containers with concrete information about the utterances. To explain how, we will discuss four situations

First, we treat the case of an utterance from a single speaker. Say the right speaker starts an utterance. The dialog construction module will receive an ‘utterance start’ event from the right utterance information object. From then on it will use the ongoing utterance container to keep an up-to-date copy of the utterance information in the right utterance information object until this objects sends an ‘utterance end’ event to the submodule. Then it copies the utterance information to the foregoing utterance container information object (sit (a)) and the primary ongoing utterance container is free to refer to a new utterance.

Second, we discuss the case of overlapping speech. Say, left speaker starts an utterance before the right speaker has finished. Then the dialog construction uses the secondary ongoing utterance object to keep a copy of the utterance information in the left utterance information object. If the left speaker stops before the right speaker, the utterance of the left speaker is simply discarded. If the right speaker stops before the left speaker, besides copying the data of the right speaker to the foregoing utterance container, the dialog construction submodule copies the data of the secondary ongoing utterance container to

the primary ongoing utterance container. In other words: the secondary utterance becomes the primary utterance (sit (b)).

Third, we discuss the treatment of (mid-utterance) silences. The submodule treats silences only slightly different from utterances. To explain, we may go back to the case of a single speaker. When an utterance of a single speaker ends, a silence starts. The primary ongoing utterance container will keep a copy of the silence information object (this is needed to be able to lookup the probability the upcoming utterance will be intended for the system). When one of the speakers starts speaking again the silence ends. Therefore the ongoing utterance container is free for the utterance that has started. Since information about silences is only of importance for the probability the upcoming utterance is intended for the system, the data about the silence is not copied to the foregoing utterance container.

Secondary utterance containers can also contain information about silences. Recall that when a speaker stops speaking, but the utterance information object for that speaker has not finished, the silence information object does start to collect information and it does send a silence start event to the dialog construction submodule. The dialog construction module will keep the information about these silences in the secondary utterance information container. With a mid-utterance silence the silence stops before the primary ongoing utterance has ended and is discarded. With a silence at the end of the utterance (before the end of utterance criterion is reached) the information of the silence is copied to the primary ongoing utterance place holder like in (sit (b)). Since silences only start when no one is speaking there cannot be silences with simultaneous speech.

Finally, we treat the case of dialog events. Dialog events point to the upcoming utterance. When a tap 'n talk button is pressed or when MATIS asks a question, this information is stored in the upcoming utterance container (in fact this is the only information about the upcoming utterance that can be stored). When an utterance ends, this information is added to the primary ongoing utterance container (sit (c)). It stays there until a new primary utterance has ended, then it is copied along with the other utterance information to the foregoing utterance information object. Silence, or secondary utterance, endings do not affect the transfer of dialog events information.

### **5.6.5 Probability Lookup**

The probability lookup submodule reports the probabilities that the foregoing, ongoing and upcoming utterance are intended for the system. A lookup table contains the probabilities a foregoing ongoing or upcoming utterance is intended for the system for each possible feature combination. We used the complete dataset of the experiments in

chapter 3 as training set for these probabilities (see chapter 3 and 4). For upcoming utterances, the probability that the utterance is intended for the system is based on dialog events information and if the ongoing utterance is a silence it is also based on the focus of attention information of this silence. If the secondary ongoing utterance is a silence this information is used rather than that of the primary ongoing utterance. For the ongoing utterance we use the up-to-date feature values, dialog events, utterance length and focus of attention information as if the utterance has finished. Since we cannot distinguish between mid-utterance silences and silences at the end of utterances, we treat these silences as end of utterance silences and neglect the feature-values for this silence. When there is only a silence as primary ongoing utterance, the ongoing probability is set to 'not applicable' (NaN). For foregoing utterances we use all feature values. Foregoing utterances cannot be silences. The probability lookup module sends a continuous stream of xml files containing meta-information about the utterances such as an utterance number, who the speaker was and the probabilities. Clearly the confidence estimates of foregoing utterances do not change until a new ongoing utterance has ended.

---

## 5.7 Dialog Management

---

The responsibility for dialog management is split between a wizard and a Macromedia Flash™ application for Dialog Management (DM). The interface decides whether an utterance should be treated as accepted for the system. The wizard decides whether an utterance contains relevant information for the dialog and controls the prompts from MATIS. We can explain the division of labor between wizard and DM module better if we discuss its interface (figure 6, next page).

The wizard interface consists of two screens. In the first screen (not shown) the wizard is able to control the logger<sup>4</sup>, choose which version of the MATIS interface is presented at the user side and to start the dialog. After starting the dialog, the wizard enters the main screen (Figure 5). On top (1) there is a utterance status display. It is intended to provide the wizard with up-to-date information about the dialog and the inferences of AAD, so the wizard can think ahead on the upcoming decisions he has to make. It consist of two bars. The top bar shows the duration of the ongoing utterance. The length of the bar increases as long as the utterance is ongoing (or in uncertain state) according to AAD. Color coding anticipates the decision the dialog manager is going to make about the

---

<sup>4</sup> Logging is done by the Moock-server that connects AAD, DM and the MATIS interface, it was adapted to do that by Eugen Schindler.

intended addressee of the utterance. It is red as long as the probability that the utterance is intended for the system is below the chosen threshold; it is green if this probability is above this threshold. Once the utterance has finished it moves to the bottom bar.

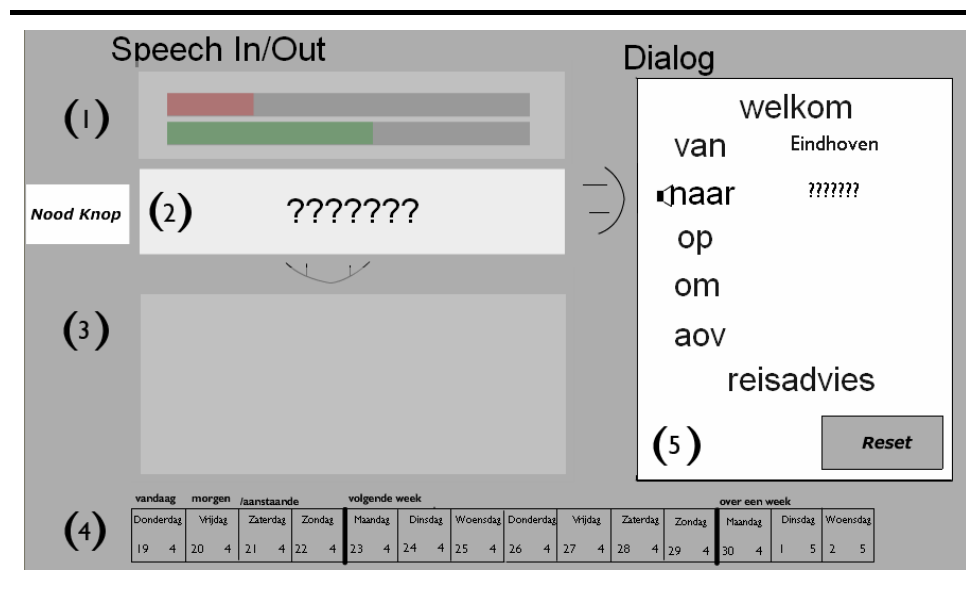


Figure 6: screenshot of the wizard interface. Its main screen holds a display for status information about utterances (1), controls for the dialog (2), for the speech recognition (3), display with dates (4) and dialog information, and a control for ending the dialog (5).

If the utterance is classified as intended for the *system* the dialog application opens the utterance accepted field (2). In the utterance accepted field, a series of question marks appears. At this point, the wizard has to choose: either the utterance did *not* contain words that could potentially be filled in the field that is to be filled (for example a station name or a date) or there *are* such words. If there are *no* such words the wizard waits for 2s. After two seconds the question marks are sent to the suiting fields in the MATIS interface and the wizard has to prompt for the next empty field (with a dedicated key on the keyboard). An exception to the 2s waiting time is when within this time a new utterance has been accepted. At acceptance of a new utterance for the system within the 2s waiting time, the question marks are sent to MATIS immediately. This way the wizard can focus on deciding what to do with this new utterance. If there *are* words that could potentially be filled in the field that has to be filled, the wizard opens the speech recognition field (3, with a dedicated key on the keyboard). This field contains an auto-completion box for typing the relevant station name, date, time, or arrival/departure parameter. The date field requires the wizard to type the date as 8 digit number (ddmmyyyy). Since users can also state the date field in different terms (for example ‘tomorrow’, ‘today’, or ‘next Friday’) a status-bar with dates from today up to two weeks later completes the screen (4). Once the wizard has filled the field, she sends it to the screen of the MATIS interface (using a dedicated key) and prompts for the next empty field unless the participants are already

involved in further discussion. As long as the speech recognition field is open, no new utterances can be accepted and the interface will not send the question marks to MATIS after 2s.

In chapter 6, we introduce a ‘severe error protocol’. In this condition some utterances classified for the system are filled with random values rather than question marks. If the utterance is classified as intended for the system and a random number between 0 and 1 reaches a value above 0.3, the field contains a random value (for example a station name) for the field under discussion, rather than question marks,. The wizard can now choose to wait until the dialog manager sends this random value or to overrule the random value and activate the speech recognition module which in turn prevents the random value to be sent.

If the utterance is classified as being intended for the *partner*, the wizard is not able to activate the speech recognition field or to send anything to the field. This is to ensure the wizard does not overrule false rejections. However, to prevent that an error in the speech activity detection, head orientation tracker or automatic addressee determination frustrates the interaction, the wizard can make use of the emergency button (left of (3) mark). This allows the wizard to act as if there was an utterance classified for the system. We told the wizard to use the emergency button only in those cases there was an obvious problem with the system and users would not be able to finish the dialog unless the wizard would use this button.

To the right of the screen the dialog status is shown (5). The dialog manager considers a field filled when there is a meaningful (thus correct or false but no question marks) value in the field. This can be the case if the wizard has filled the field, if users have filled the field through field fill buttons or -in the case of the severe error protocol- if the DM has substituted the question marks with a random value. With normal proceedings, the dialog is filled field by field from top to bottom but there are exceptions. For examples users may press a tap ‘n talk button or use a field fill button for the date before they have completed the arrival and departure fields. The ‘next’ empty field is the first empty skipped field. Say the departure station is filled and users continue by pressing the today field-fill button, then the DM decides to prompt for the arrival station<sup>5</sup>. When the dialog is finished the wizard has to press the button to prompt for the next empty field once more to show the travel advice. After that, the wizard can reset the dialog manager and

---

<sup>5</sup> The wizard has no choice over which field is filled.

returns to screen 1, while the MATIS application shows a message asking the user for patience before the application is loaded (again).

---

## 5.8 MATIS interface

---

The functionality of the MATIS interface is discussed in some detail in chapter 3.2.2. and the different interface versions we designed are described in Chapter 6. The MATIS interface is implemented in Macromedia™ Flash as a shell that takes care of the communication with the external system, and in which different versions of the interface can be loaded. The DM can tell this shell which version of the MATIS interface is needed. The MATIS interface shell sends messages about button presses (tap ‘n talk or field fill) to the DM and to AAD, so they can act accordingly. Also the MATIS interface accepts messages from AAD to arrange its feedback on the inferences of AAD and from the DM to know when a field has to be filled and with what value and when and what questions to ask the user.

## Chapter 6: (How) Should Socially Aware Conversational Agents Show Users what they are Doing?

*In this chapter, we try to find out how socially aware conversational agents should provide users with feedback about their status. We focus on three intertwined design questions: on what aspects of their behavior should socially aware conversational agents feedback, when should they deliver this feedback, and with what metaphors? To provide answers to these question we present a qualitative, iterative, question oriented design intervention study. Within six design interventions we ask participants to interact with different versions of our prototype. The study shows that naïve users are unaware of the problem of addressing and that feedback in combination with disturbing system errors can make them aware. However, we have not managed to come up with feedback that enables users to make practical use of such an awareness. Therefore, we conclude that, in the presence of disturbing system errors, a new design effort is needed. We propose several research directions that could generate knowledge that makes such an effort easier.*



---

## 6.1 Introduction

---

Socially aware conversational agents do not assume that all users' utterances are addressed to them. If they would, such agents would respond to users even if those users were addressing their conversational partner. Those users may wonder 'why is the system responding, while I did not try to give input?'. Since our conversational agent is not perfect this question may still pop up occasionally. At the same time a system that does not assume all users' utterances are addressed to the system introduces uncertainty for users that are addressing the system. They can no longer be certain the system knows the utterance is addressed to the system and that it will respond. They may wonder 'if I speak to the system now will it understand that I am addressing it', or 'did the system accept my input?'. To answer such questions, socially aware conversational agents may provide feedback about their status to users. In this chapter we seek out to find requirements for such feedback.

Designing feedback for conversational agents is not an easy task because we do not have appropriate models or examples of good designs for this domain to build informed intuitions about the many design decisions we will have to take (Grudin, 1994). One way to go is to draw from *models about human-human communication*. In chapter 2, we have put forward a perspective of language as coordinated action that provides us with a profound understanding of the ways humans coordinate their communicative acts. However, for informing design decisions this theory has two major limitations. First, the theory is very abstract: it captures a wide range of phenomena in only a few comprehensive constructs, and in doing so it abstracts from precisely the kind of situational knowledge we need to inform design decisions. Say, for example, we would try to design a (humanoid) conversational agent using head nods and eye gaze to provide users with feedback about their status. In that case, we are in need of (situational) knowledge about when to use a head nod and when to look where. The (declarative) knowledge that these are both back-channel signals used to coordinate attention and execution within communicative acts is of lesser value. Second, it does hardly account for interaction under technological constraints. In chapter 3, we found differences in gaze and speaking behavior of users in interaction with an information kiosk compared to what we could expect from human-human communication. Similarly, depending on the technological possibilities of a specific conversational agent, we cannot design all system behaviors we might like too, and this might also not be desirable (see for example: Shneiderman & Maes, 1997). So if we work from Clark's theory, we need to build a situational knowledge base including

technological constraints. An alternative way to go about the design of feedback is to draw from *successful examples* of feedback in *other domains*. For example feedback design is extensively researched for desktop computing, resulting in several interaction genres (design conventions that anticipate particular usage contexts) such as command line and GUI interfaces (Bellotti et al., 2002). As Bellotti et al. point out these interaction genres can not straightforwardly be reapplied for the type of problem we address here. Compared to desktop computing we accept a wider range of input, this input is of implicit nature (users are not aware they are giving input) and our system makes autonomous decisions about this input. Existing interaction genres do not anticipate these capacities. So if we choose to work from existing examples of successful feedback we would need to analyze why these solutions are successful and consider how they need to be adapted to be successful in our context. In fact our context could be so different from desktop computing that we may need a different measure for success.

So while there is a lot of material to draw from, we are unsure to what extent it is appropriate for our purpose. Therefore we try to deliver five contributions. First, we try to generate artifacts: concrete examples of feedback suitable for our domain. Second, we try to relate these artifacts to generalizations (typologies of ontological nature) on an intermediate level of abstraction (so in between the abstract theory of language as coordinated action and the concrete artifacts) suitable for our design. Third we try to assess the effect these artifact have on users' perception and appreciation of these artifacts (the interaction space). Fourth we try to assess the suitability of our typologies to understand this interaction space. Fifth, when a generalization is suitable for understanding the interaction space between the artifacts and the users' perceptions we will try to formulate a design advice from this understanding.

We start this chapter by formulating a frame of reference that we may use in the rest of the chapter to describe interfaces. In section 6.2, we try to spell out the issues involved in designing feedback for socially aware conversational agents by discussing three design questions that we can summarize as: *when* to show *what* with *which metaphors*. Next, we try to provide concrete examples of feedback and describe those in terms of our reference frame. We present an, explorative, *conceptual design* study in section 6.3. We end this explorative phase with informal evaluations of those interfaces. This is a first test of our reference frame, we use these evaluations to set priorities of the central study in this chapter. In this *question oriented design intervention* study, we evaluate six new interfaces with users. These evaluations provide insights in the way users perceive interfaces and in the utility of our initial reference frame. We describe the setup of this study in section 6.4, its

results and intermediate discussion in 6.5 and we provide a general conclusion and discussion in section 6.6.

---

## 6.2 Three Design Questions

---

### 6.2.1 Introduction

In this section we discuss 3 interconnected questions that we will use as a frame of reference for reasoning about the design of feedback for socially aware conversational agents: ‘what system behaviors do we need to communicate?’, ‘which metaphors can we adopt to deliver this feedback?’ and ‘when should we deliver this feedback?’. Clearly, in this section we cannot move beyond posing the questions and spelling out the issues involved.

### 6.2.2 Back-channeling versus transparency

The first question we address is which system behaviors we need to communicate to users. Our socially aware conversational agent does three things: it attends to behavioral cues of the user, it interprets these cues in the light of the question who is talking to whom, and it decides to accept or reject an utterance. Within this chapter we will focus on a distinction between *back-channeling* and *transparency*.

When people talk to each other they provide each other with feedback on behavioral cues. This is called *back-channeling*. In chapter 2.4 we have discussed this at some length. We said that people seek out (and provide) early evidence of successful coordination of their communicative acts, even if it is only weak evidence on lower levels of the action ladder (attention and execution). They can do this with signals such as orientation behavior, head nods, vocal acknowledgments (such as ‘mhm’ and ‘yes’) and eye gaze. We may argue that socially aware conversational agents should also provide early evidence about their attention to the users’ behavior. This way these agents provide users with evidence they are attended to. If there is a mismatch between the intention of the speaker and the feedback from a conversational agent speakers may adapt their behavior to attract the agent’s attention or to attract the attention of their human counterparts (see chapter 2.4). A challenge may be to design back-channel signals in such a way that these signals encourage speakers to adjust their behaviors in ways appropriate for the classifier. We must note here, that we define back-channel signals from the system in a general way: all system behaviors that provide evidence the system is attending to and processing the behavior of the users are back-channel signals. They do not necessarily need to resemble human back-channel signals.

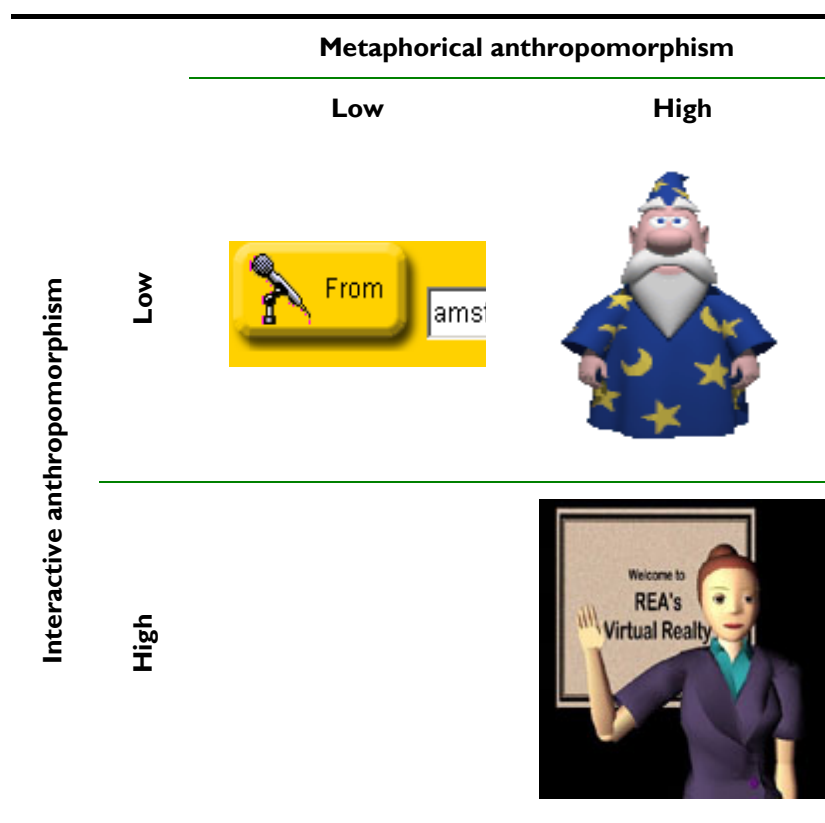
On the other hand, such feedback may be too subtle. We can argue that the task of a socially aware conversational agent is to infer who is talking to whom and that it needs to be *transparent* about its findings. Users may not be aware the agent has the task of separating utterances for the system and for the partner. So if they need to understand what is going on, the feedback of the agent should mark this feature. We can say, rather than mimicking human back-channel signals, we need to find metaphors that communicate to what extent the system is open for input or not. In this line of thinking, the least an agent can do is to tell the user whether it has accepted an utterance or not. Also, the agent may show what it expects to decide later. We may try to design feedback that indicates whether the agents thinks the next utterance will be intended for the system, or whether the agent expects to accept the current, ongoing, utterance. This fulfills the human need for early feedback.

We take the distinction between back-channeling and transparency to be discrete. A single interface can be a back-channeling interface, a transparency interface, or both. But they are not ends to a continuum: we cannot put interfaces on a gradual scale towards more transparency or towards more back-channeling.

### 6.2.3 Interactive versus metaphoric anthropomorphism

The second question to address is: ‘what metaphors can we use to deliver this feedback?’. Back-channeling and transparency suggest different metaphors. For back-channeling signals, anthropomorphic metaphors, such as animated characters, seem attractive. In contrast, transparency suggests borrowing metaphors from other domains that can be open or closed for input, for example a trash bin.

The use of anthropomorphic metaphors is under debate. A disadvantage of animated characters is that they can mislead the users by raising their expectations of the system’s capabilities beyond its actual capabilities (Shneiderman & Maes, 1997). Indeed, users may believe, a character like “Merlin”, that ships with the Microsoft XP™ software package to be more intelligent than it actually is. Therefore, Cassell et al. (1999b) argue that *if* we use the expressive behaviors of humans in the interface they should be *more* than metaphorical. The behaviors of the animated character need to be tied to the behaviors of users and the functions they have in the dialog management such as ‘take new turn’ and ‘contribute new information’. This last argument suggests we can make a distinction between functional or *interactive* anthropomorphism and *metaphorical* anthropomorphism (figure 1). This leads to a two-dimensional typology.



**Figure 1: a two-dimensional typology showing the distinction between interactive and metaphorical anthropomorphism. In the top left box we find a button of the MATIS interface, top right we see Microsoft™’s “Merlin”, and in the bottom right corner: “REA” (Cassell, 1999b).**

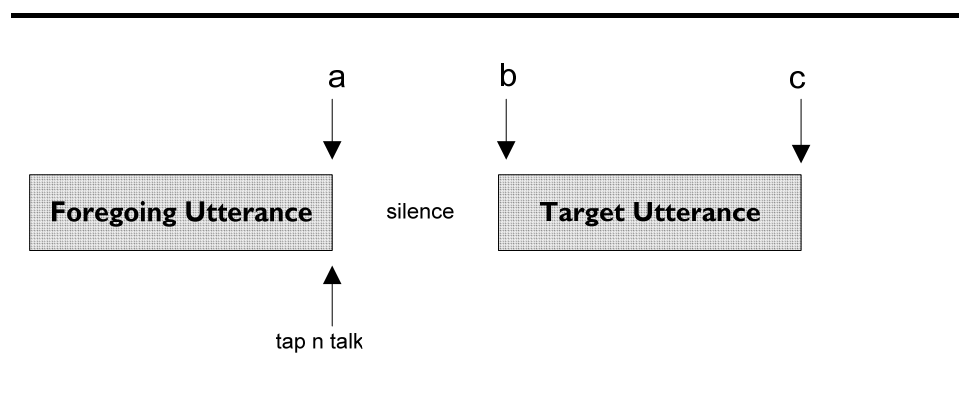
For example, the buttons on the MATIS interface (see figure 1, top left) do not resemble human expressions and are not tied to the back-channel signals humans display in the dialog such as orienting behavior or facial expressions. These buttons are metaphorically and interactively not anthropomorphic. Microsoft’s “Merlin” (see figure 1, top right) displays human like expressions, but these are not tied closely to the back-channel signals of the users. We consider it to be metaphorically anthropomorphic but not interactively anthropomorphic. Cassell’s (1999b) “REA” is an animated agent that has humanoid expressions that are closely tied the behaviors of her human interlocutors. So REA is both metaphorically and interactively anthropomorphic. For now we consider the boxes in the table to be ends on a continuum, thus we allow ourselves to say interface (a) is of higher metaphorical anthropomorphism than interface (b).

This way of looking at the use of anthropomorphic metaphors raises a concern for the transparency option. If we use non-anthropomorphic metaphors that are closely tied to the users back-channel signals, we may introduce a mismatch similar to the mismatch “Merlin” represents. We need to see to what extent this is bad. In the conceptual design study we will explore both anthropomorphic and non-anthropomorphic metaphors for

both the back-channel as the transparency alternative. We will also return to this typology and use it in a generative way in the intervention study reported in 3.4 and 3.5.

#### 6.2.4 Feed-forward, early feedback, and conclusive feedback.

The third question to address is when we need to deliver feedback. In the introduction we mentioned that users may have questions such as ‘if I speak to the system, does it know I am addressing it’ or ‘did I communicate successfully’. But these questions appear at different points in the dialog. In chapter 4 we have introduced a reference frame for discussing dialog history. Here we shortly revisit this reference-frame, to enlist the questions that users may have at different points in the dialog. In the discussion that follows we will look at situations before, during, and after an utterance that we label the *target* utterance (see figure 2). We label the utterance before the target utterance, the *foregoing* utterance.



**Figure 2: reference frame for discussing utterances.**

Say the target utterance is intended for the *system*. At or near the end of the foregoing utterance (a) users may wonder: ‘if I speak to the system, does it know I am addressing it’. The interface should somehow communicate that it is open to receive an utterance for the system soon. We will call this *feed-forward*. At the beginning or during the target utterance (between b and c) the user may wonder: ‘is the system attending to me’. To answer this question the system may provide *early feedback*. At or near the end of the target utterance (c) the user may wonder: ‘did I communicate successfully’. To answer this question the system may provide *conclusive feedback*<sup>1</sup>. Note that the current MATIS interface (see Sturm, 2005) does not provide feed-forward or early feedback but it *does* provide conclusive feedback. If MATIS is processing a request, a rotating hour-glass

<sup>1</sup> We reserve the term *feedback* to refer to the combination of these specific types: feed-forward, early feedback and conclusive feedback.

appears at the screen and as soon as it has recognized a result it is printed on the screen. If this result is correct, it is evidence for successful communication. If it is false or does not appear at all this is evidence of unsuccessful communication. Clearly, if no result appears, for example in case of a miss, this is evidence of bad quality because it fails to be conclusive at some point in time.

If the target utterance is intended for the *partner*, users are not likely to have questions related to the status of the agent at all. In this case, the interface may provide feed-forward and early feedback that points out it is not attending to the user or conclusive feedback about the fact it did not accept input. However, we need to consider the possibility of a false alarm: the utterance is intended for the partner but not recognized as such. In this case users are confronted with the feedback relating to utterances for the system. The system may indicate it is attending to the user, while it should not. This may provoke questions with users (for example ‘What is going on?’). This leads to two considerations for the design. First, in designing the three types of feedback we need to make sure it is clear to users what is going on, both if they intend to address the system and if they intend to address the partner. Second, both the feedback the utterance is assumed to be intended for the system and the feedback the utterance is assumed to be intended for the partner needs to be designed in such a way that, if the system is wrong, the feedback does not disturb the interaction.

Another question is whether these three types of feedback need to be combined or that they deserve different visualizations. One alternative is to combine them into a single representation for openness of the system. Such a representation could, for example, show the likely addressee of the target utterance near the end of the foregoing utterance (a) and the likely addressee of the ongoing utterance (between b and c) when users have started speaking again. This way we reduce the number of elements on the screen users need to understand, but we introduce ambiguity in the meaning of the representation. The alternative, to use two or three representations has the opposite effect. We postpone a decision about this issue until after the conceptual design phase.

---

## 6.3 Conceptual Design<sup>2</sup>

---

### 6.3.1 Introduction

We have chosen to adapt the existing MATIS interface, rather than designing a new interface for our socially aware conversational agent from scratch (see chapter 1). The advantage of this approach is that we can use concrete elements of this interface and our experience with the development of MATIS in the interface design for our conversational agent. However, the disadvantage of this approach is that we may overlook hidden requirements for the case of shared use, that could have drawn our attention if we would have chosen to design a completely new interface (Sturm, 2005, pp 69; Hugunin & Zue, 1997). This conceptual design study aims at overcoming that disadvantage.

The conceptual design study consists of two phases: an idea generation phase and an evaluation phase. In the idea generation phase we focused on generating a wide range of ideas for visual feedback for our agent, for both the transparency and the back-channeling alternative. In the evaluation phase we performed informal evaluations of three working mockups with colleagues. For each interface we invited two pairs of colleagues and asked them to plan (part of) a trip with the system and to comment about their experiences. This helped us to uncover hidden requirements, to provide provisional answers to the questions outlined in section 6.2 and to set priorities for the formal design intervention study described further on.

### 6.3.2 Idea generation

We started out with a short problem statement, asking for visualizations for both transparency and back-channeling. Based on this we brainstormed, structured our solution space, and brainstormed again leading to about 25 ideas. Among these ideas were straightforward solutions such as anthropomorphic characters and less conventional ideas such as depicting the process of recognizing and interpreting speech on the screen or using the orientation of the MATIS form to depict to what extent the agent is open for input (see figure 3).

---

<sup>2</sup> All designs presented here were created by Thomas Noordzij, an internship student, with a background in graphical and interaction design. The designs presented in the next study were created by the first author, making use of Thomas' material.



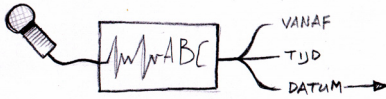
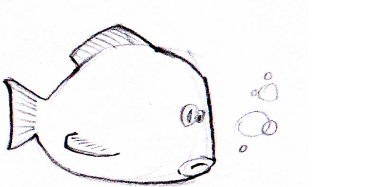
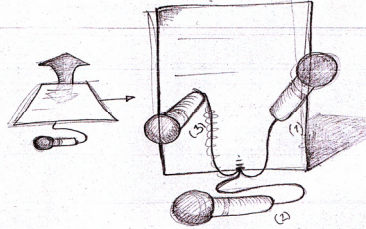
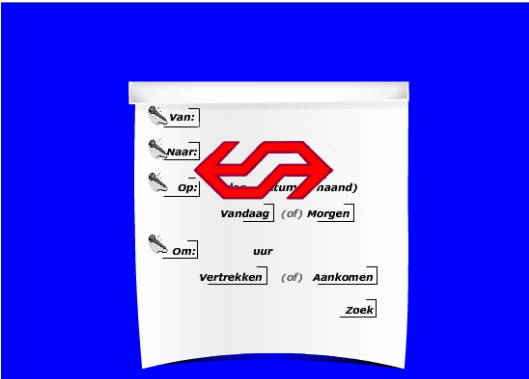
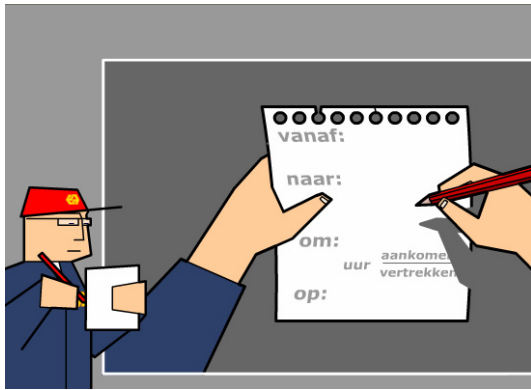
| Drawing of idea   | Description   |
|---|---|
|  | <p>The interface can represent the process of recognizing speech on the screen when it is performed.</p>  |
|  | <p>An anthropomorphic character can visualize to what extent the system expects to be addressed.</p>  |
|  | <p>We can use the orientation of the MATIS form to suggest whether the agent is open for input or not. A microphone attached to the form could provide extra back-channeling information.</p> |

Figure 3: Some early sketches

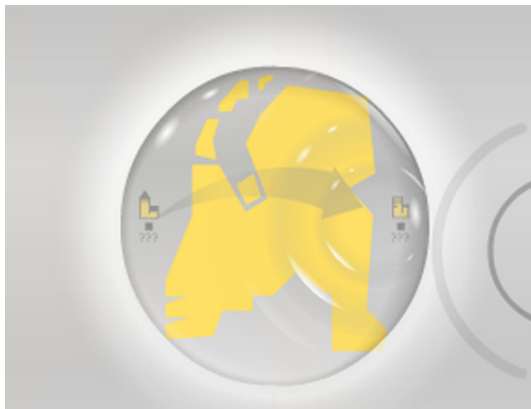
From this set of ideas we selected three ideas that we developed into working mockups. The selection criterion for ideas was that there should be a conceptual link between the elements for giving feedback and the rest of the interface. We show and describe these three ideas in figure 4 (on the next page. On color plate 1 and 2 we provide extra screenshots and explanations). In the next section we provide a discussion of these interfaces in the light of the three design questions of section 6.2 and summarize the reactions of our participants who used these mock-ups.

| Screenshot of Idea  | Description   |
|---|---|
|  | <p>This 'Folding form' folds itself when the system is closed for input. Shown is a screenshot in-between opening and closing.</p> <p>When the system is open for input the interface is the same as that of MATIS (see Chapter 3.1) but it that the lay-out and appearance of the fields and buttons is different. The logo is yellow and on top of the form.</p> <p>When the system is closed for input, the fields and buttons on the form are not visible any more and do not work. The logo is red and lies on the ground.</p> |



This 'Railroad assistant' holds a notebook with the parameters for the system. A cut-out shows the notebook in more detail.

When the system does not expect input, the Railroad assistant holds his arm down, away from the paper. When it does expect input, the arm raises to the paper (shown in this image), ready to write. After accepting input the hand moves over the paper like it is writing until a recognition result is shown.



This 'Water drop' lies on top of a map of the Netherlands. It provides back-channel signals for the speech and gaze of both participants. It can also show whether the agent is open for input or not.

If a participant looks at the screen a backlight is on. When a participant looks at the screen and speaks, waves appear in the air and on the surface of the water drop to point out the agent receives the users' speech. When the agent is closed for input the water drop gets a frozen appearance.

When station names are filled, the map and arrow beneath the water drop scale and rotate to make sure the departure and arrival station are at the ends of the arrow.

Figure 4: the three ideas that we developed into working mockups

### 6.3.3 Evaluation

#### 6.3.3.1 Railroad assistant

The metaphor of a railroad assistant filling a notebook with the essential information provides a natural way to show the status of the system including the social awareness part. It is an example of an anthropomorphic 'transparency' interface. The railroad assistant does not provide back-channel signals, but shows whether it is ready to accept or whether it is processing input. Back-channel feedback could be added to the animated railroad assistant although this would increase the number of elements users need to attend too. Feed-forward and early feedback are combined in a single indicator, the hand comes in 'ready-to-write state' during the silence before the target utterance (provided the agent believes it will be intended for the system). It stays in this state during the utterance (provided the agent still believes it is going to be intended for the system). The railroad assistant metaphor allows for a separate conclusive feedback indicator, because it appears to be writing as long as the system is processing input.

Despite its apparent naturalness, our colleagues in our pilot tests did not understand it. They did not notice the arm disappearing from the screen when the agent was not ready

to accept input. Therefore the interface was not able to show whether it was open or closed for input: the necessary contrast was lacking. The conclusive feedback in the form of a writing hand was also not appreciated unequivocally. Our participants felt it was a funny interface, in particular because the animation of the railroad assistant and the hand in the cut-out were synchronized, but also felt it was annoying because it was moving over the form all the time. Indeed, in practice, this occurred a lot, because of the fairly high number of false alarms of our agent. While the feed-forward and early feedback lacked salience the conclusive feedback was too obtrusive.

### 6.3.3.2 Folding form

The metaphor of a form folding itself open or closed, is an example of a non-anthropomorphic transparency metaphor. It represented just one of several possibilities we thought of, to use the orientation or shape of the MATIS form to indicate whether the system was open or closed. We did not expect users would like that they couldn't see the status of the dialog or use the buttons when the agent was closed for input. But we decided to pilot it anyway because of the salience of this way of pointing out openness of the agent. The interface did not have a separate indicator for conclusive feedback. Feed-forward and early feedback were combined in the same way as in the Railroad assistant interface.

As expected our participants did not like the form being closed at some times, and would at least want to have a way to open it again. It was clear for them they could not speak to the system when the form was closed, but they also mentioned that it was closing and opening at random moments. They were surprised to find out there was a link between their own gaze behavior and the fact the form opened and closed. Indeed the interface does not suggest such a link.

### 6.3.3.3 Water drop

The water drop interface represents the system functionality with a map rather than a form that needs to be filled. The social awareness part, a water drop is a hybrid between the transparency alternative and the back-channeling alternative, both employing non-anthropomorphic metaphors. The transparency information is pictured by making the drop look fluid or frozen. The back-channel signals by back-lighting and surface waves on the water of the drop. This shows the possibility of using non-anthropomorphic metaphors for back-channeling purposes. Also the interface is visually much more appealing than the MATIS interface, because of the quality of the graphics and the scaling, panning and rotation of the map at recognition of a station name. In the test we

only asked participants to fill a departure and arrival station. Integrating buttons and fields for the date and time in a way that fits this interface would have formed a large extra design effort. We wanted to test the social awareness part before doing this.

Like the other two ideas, our participants did not understand the interface elements. Participants noticed and liked the waves and understood that they suggested that one of them was speaking. The backlighting went unnoticed. Participants did not have a clue why the interface would show who is speaking, they interpreted it as a cool animation effect, but nothing more. So in this case participants were able to link the behavior to their own behavior, but did not think about any effects these behaviors might have on the system. Users noticed the bubble changing color (on freezing) but did not recognize it meant that the bubble was frozen or that the system was not open for input then. Participants liked the map as a representation of the system functionality but when confronted with a form as alternative they were not sure they would prefer the map. They did praise the scaling, panning and rotation of the map for its ‘dramatic’ effect.

### 6.3.4 Discussion Conceptual Design

This study aimed at uncovering hidden requirements for the design of transparency and back-channeling feedback for our conversational agent, at providing provisional answers to the three questions of section 6.2 and to further frame the study in the next section. Now, we discuss our findings.

The evaluation results of the three chosen ideas are strikingly similar. These demonstrators seemed clear and reasonable to us. Participants, when planning a trip with the system, However, did not have a clue what was going on. After an explanation of the different states of the system they would usually understand. Thus: users need to learn the meaning of our visualizations. This is hard during a dialog because they focus on the dialog with each other and not so much on the features of the interface. In other words, we have confronted our users with a dual task: plan a joint trip and learn and reason about our interface.

In this study we have shown that it is possible to design anthropomorphic feedback for the transparency alternative and non-anthropomorphic feedback for the back-channeling alternative. In other words there is no necessary link between the type of feedback and the type of metaphor. For both types of feedback we found it is challenging to come up with interfaces that make users aware of the chain of events leading up to the system’s decision. For the back-channeling alternative we have managed to come up with an idea that makes users aware the system knows who is speaking (the water drop), but users were unaware the agent does something with this information. For the transparency

option we came up with a metaphor, that was interpreted as indicating whether the system is open or not (the folding form). But our colleagues called the behavior of this form random, and did not hypothesize a link to their own behavior.

We did not get any hint to answer the question what type of metaphors we should use. We found using an anthropomorphic metaphor (railroad employee) does not automatically mean users attend to the relevant behaviors of the animated character. This may be because the elements weren't clear enough, but it may also be because of system errors, making the interactive anthropomorphism less ideal than we hoped for. We were slightly more successful with non-anthropomorphic metaphors (at least the folding form and the waves in the water-drop were clear) but these interfaces did not suggest a link between the system and the users' behavior. This may be because we used non-anthropomorphic metaphors. But it may also have been the metaphors themselves.

It turned out to be hard to find metaphors where feed-forward, early feedback and conclusive feedback had a natural place. In all three demonstrators we combined feed-forward and early feedback, only in the Railroad assistant idea there was a natural place for conclusive feedback on the acceptance of an utterance. This feedback turned out to be disturbing because of the frequent false alarms of the system.

For the next study we may extract a number of focus points. First, this pilot study was not conclusive around the questions about back-channeling and transparency and around anthropomorphism, therefore in the next study we treat these issues in a more structured way. Second, the study was not conclusive about the type of metaphors we should use. We leave the question open for the next study, but we try to improve on the clarity of the feedback. In particular the 'railway assistant' and the 'waterdrop' may have been too complex. These interfaces had indicators for multiple types of feedback. The railway assistant combined an indicator for feed-forward and early feedback with an indicator for conclusive feedback. The 'waterdrop' combined indicators for gaze and speech and the system interpretation. In the next study we will use only a single indicator per interface. In addition we add the question what is needed to make participants *understand* the feedback to the set of questions we try to answer in the next study. The question *when* we should deliver feedback was answered to some extent. We already combined feed-forward and early feedback in one element, and as we plan to use only a single indicator it seems natural to preserve this combination. Also, we do not give priority to finding new ways of conclusive feedback about acceptance or rejection of an utterance. Such conclusive feedback is provided with the (absence of) appearance of recognition results and it will be hard to make this extra element clear because of the frequent false alarms in our system.

## 6.4 Design intervention study - approach

---

### 6.4.1 Introduction (on intervention methodology)

In this (and the next) section we present an iterative design study. However, the study has a *research* rather than a *design* focus. We do not aim at delivering a ‘best’ design, but we aim at strengthening our understanding of what a good design is in our context of use. For this part of the study we felt intervention methodology was the most effective way to approach this problem. We borrowed the term intervention study from different fields such as educational research (see: Brown, 1992) and behavioral change programs (see: Michie & Abraham, 2004). In both fields the word intervention is used for attempts to effect changes in a complex system (class room settings, or health behaviors of patients). Rather than assessing, for example learning theory in controlled laboratory settings, intervention studies aim at –directly– assessing the applicability of this theory in a real world setting. In turn this fires back to the theory itself, both in terms of content as in terms of relevance and accountability (Brown, 1992). Intervention methodology is also practiced in HCI, be it often less explicitly than we have done (see Cole, Puro, Rossi, & Stein, 2004; Hevner, March, Park, & Ram, 2004; Puro, 2002). We will use the term intervention for a (structured, qualitative) evaluation of a specific example of an interface with users. This is a test of the interface (or rather the interaction space of the interface and the user) as well as a test of our understanding of this interaction space. In this section we will first list the questions of our intervention study (6.4.2), then we will explain the setup of this study in more detail (6.4.3), after which we describe the setup of each individual intervention (6.4.4).

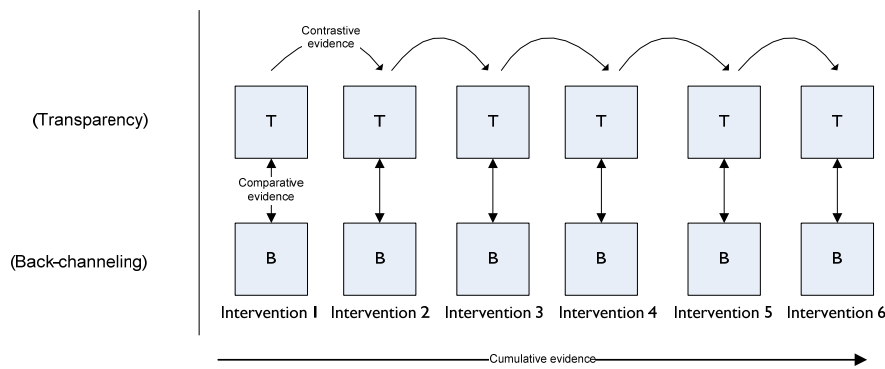
### 6.4.2 Starting questions

The conceptual design study leaves us with 3 questions that we may try to answer in this study. These are listed below.

- *Transparency or Back-channeling*: What are the advantages and disadvantages of the transparency alternative compared with the back-channeling alternative?
- *Anthropomorphism*: What advantages and disadvantages have anthropomorphic metaphors compared non-anthropomorphic metaphors?
- *Understanding*: How can we, given the fact there is a dual task for users, ensure users understand the feedback provided by the socially aware conversational agent?

### 6.4.3 General setup

Figure 5 depicts the setup of our study, and the three types of evidence we can extract from our study.



**Figure 5: a graphical depiction of the design intervention study, across six design interventions we compared interfaces based on the transparency (T) or back-channeling (B) alternative. Between interventions we designed new interfaces, based on questions about understanding and anthropomorphism**

We take the first question, the relative advantages and disadvantages of a transparency interface compared with a back-channeling interface as a central question. *Within* six design interventions we allow users to compare a version of transparency and a back-channeling interface. So from each intervention we may be able extract *comparative evidence*: we may be able to draw conclusions about how the transparency alternative compares to the back-channeling alternative. *Between* interventions we come up with new versions for both the transparency and back-channeling alternative. We do this based on semi-structured interviews with users, that we interpret in the light of the questions about understanding and anthropomorphism above (see: Figure 5). We will refer to a change from one intervention to the next as a *move*. This allows for what we will refer to as *contrastive evidence*. Say that our users in intervention 1 express a preference for a transparency interface and we hypothesize this is because we used non-anthropomorphic metaphors we can try to use anthropomorphic metaphors in intervention 2. Intervention 2 then allows us to draw comparative conclusions about transparency and back-channeling alternatives for the concrete examples we tested in intervention 2 but also to evaluate the move we made. For example we may find that the preference for users is dependent on the type of metaphor we use for both alternatives. In each intervention we invite new subjects so comparative evidence is within subjects while contrastive evidence is between subject. A third type of evidence is *cumulative evidence*. If users express a preference for a transparency interface throughout the study, regardless of the specific interface we can be more and more confident that transparency is preferred over back-

channeling. Clearly the confidence we can have in a conclusion can rely on multiple types of evidence.

This approach allows us to some extent to assess the generality of our conclusions across different implementations of an interface. Also, since we can combine multiple sources of evidence we have several means to ensure rigor. Interpretative conclusions can be strengthened through cumulative evidence or combined with design moves aimed at testing the limits of these conclusions. At the same time this flexibility has its disadvantages. The success of the study depends on finding the right design moves throughout the study. This risk can be reduced by planning the right amount of pairs per intervention, number of interventions and the time between interventions and by specifying heuristics for design moves. Using many participants per intervention increases the quality of that intervention but forces us to plan more time between interventions (all this data needs to be analyzed) before designing a new intervention is possible. Long time intervals between interventions would reduce the iterative power of the experiment. We have chosen to use four pairs per intervention. This way, we try to minimize the risk of having pairs that provide too little information to work with they would not spoil the whole intervention, while the risk of having many pairs commenting on the same aspect remained small. With four pairs we were able to run one intervention per week. The total number of six interventions is a result of this choice.

We specified heuristics for planning the interventions themselves and the design moves from one intervention to the next. First, *the principle of comparability*, relates to comparing transparency and back-channeling interfaces within interventions. When asking users to compare two interfaces we should always try design both interfaces in a way that they are comparable, except for the type of feedback. So we should, for example, not use completely different visuals. Second, *the principle of contrast*, relates to moving from one intervention to the next. We should try to strike a balance between making a small move (changing only a minor thing) which makes our intervention inefficient and making a big move (changing many things) which leaves us in doubt on how to interpret the results. This balance shifts throughout the study. In the beginning we can make larger moves, because we can use later interventions to disambiguate multiple interpretations. Near the end we must make smaller moves. Finally, *the principle of priority*, refers to testing those aspects that are most puzzling. Since the goal of the intervention study is to gain insights in (the utility of) our reference frame we should design interventions in such a way they provide us with new insights. Clearly there is some friction between the principle of priority and the principle of contrast: if the settings differ too much, this entails the risk of making design moves that are too large.

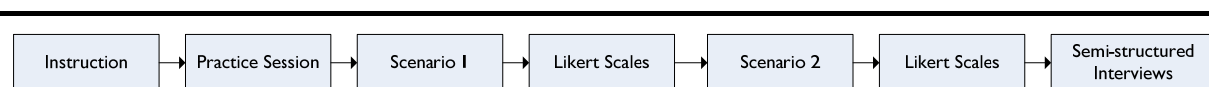


Although the study has a qualitative focus the choice to compare a back-channeling interface and a transparency interface in each intervention does allow us to gather quantitative data that we can compare *across* all 6 design interventions. We collected two types of quantitative data. First we collected data about the subjective assessment of back-channeling interfaces compared to transparency interfaces. For this we used the SASSI (Subjective Assessment of Speech System Interfaces) questionnaire. The SASSI questionnaire aims at providing a valid, reliable, and sensitive measure of users' subjective experience with a wide range of speech recognition systems (Hone & Graham, 2000). It consists of 32 general statements, about 6 underlying factors, with 5-point Likert-scales (strongly disagree; disagree; neutral; agree; strongly agree). Appendix C lists all the questions of the questionnaire. Two examples of SASSI questions are: 'the interaction with the system was slow' and 'it was easy to correct errors'. Second, we collected behavioral data in order to see whether participants behaved differently (for example the amount of kiosk directed gaze) in the back-channeling and transparency alternative. Because both analyses were not informative for the conclusions of this study we report them in Appendix D.

#### 6.4.4 Intervention setup

All interventions were set up in the same way. For each intervention we invited 4 pairs of users to the usability lab of the university (24 pairs in total<sup>3</sup>). These pairs were recruited through a combination of hallway recruitment and making use of the participant database of the research group. This resulted in a diverse group ranging in gender, educational background, and age (12 to 88 years).

Figure 6 enlists the program we carried out with each pair.



**Figure 6: schematic overview of the program for each pair**

In the *instruction* the experimenter and the participants got acquainted and the participants were introduced to the goals and the tasks of the study. First, we introduced them to the research goals. We pointed out examples of existing speech interfaces, such as the possibility to obtain train table or address information by phone, dictation systems and voice dialing and asked them whether they knew of their existence. We told participants that we expect that those speech-enabled systems would be invented for the

<sup>3</sup> For one pair in intervention 2, we have, accidentally, not recorded any video.

public domain, such as: railways, libraries, museums, city halls and so forth, soon. We explained that in those cases shared use is likely and that we intended to find out how to design interfaces for shared use. Then we introduced them to the task. We told them we use a system for obtaining train table information and that they would be asked to plan a trip to a tourist attraction in the Netherlands using two different versions of such a system. After that we would ask them for their opinion. We pointed them to the cameras in the usability labs, and we told them we would record their interaction for research purposes for which we would ask their formal consent later. Next we showed and discussed the scheme of figure 6 with our participants. Finally we asked them to equip themselves with close talking microphones (to get the best speech recognition possible) and with the tracking diadems (to know who is who).

In the *practice* session we asked participants to interact with a version of the system without feedback (just the form). This way users had the possibility to get acquainted with the functionality of the MATIS system, the basic dialog and the different types of buttons. Because of these practice sessions participants were able to get used to the environment, the microphones and the tracking devices. Also we hoped that by explaining the MATIS interface we could focus on the feedback in the interviews. Participants were asked to plan a trip with the system, in two trials. In the first trial we asked them to answer the questions of MATIS by speech, and in the second we showed them the buttons and asked them to try those out. We told participants that in the ‘real’ experiment they would use slightly different versions of the system but that it would ask the same questions and have the same buttons.

For each *scenario* participants were asked to perform a role-play in which they would use the system to plan a joint round trip to a museum or zoo. For this they received a leaflet with a scenario and tourist information about these attractions (see chapter 3, appendix A). Participants were told the scenarios differed for each person and were asked to interleave discussion about their trip with the interaction with the system. Whenever they were ready, they could leave their leaflets behind, walk up to the kiosk, and try the new system. We would be able to see them from the control room and start the system as soon as they were in front of the kiosk.

After planning a round trip they were asked to fill the SASSI *questionnaires*.

After two scenarios, one with a transparency and one with a back-channeling interface (the order of the scenarios and interfaces was counterbalanced) we conducted a semi-structured *interview* with the users. The interviewer started with repeating the fact that they had used two different versions of the system and asked the users whether they had

noticed any differences. If participants noted differences the interview focused on enlisting these differences, before asking them about their experiences and opinion about these two types of feedback. If participants were not able to enlist all differences, the interviewer probed for them in a stepwise process: first telling the users it concerned visual differences, then telling them where the differences were visible on the interface, then explicitly pointing out the differences to users. If users remembered after prompting, the interviews advanced with asking about experiences and opinions, if not the experimenter focused on comments the users gave spontaneously. These spontaneous comments fell into two categories. Some participants had comments about the interface other than the visual feedback, such as asking for extra options or comments about the voice of the system. Others commented about these type of systems in general, such as: whether they thought they would use such a system in a real situation or whether they preferred this system over the current ticket vending machines in the railway stations. These spontaneous comments are summarized in section 6.5.7

---

## 6.5 Six design interventions

---

In this section we will describe the six design interventions. For each intervention we provide five elements. First, we list the specific question(s) for that intervention. This can be one of the questions of section 5.4 or an extra question that rose in the previous intervention. Second, we will provide a rationale for the design of the two interfaces we tested in the intervention. Third, we provide a fact sheet. This is a short summary of the most important aspects of the intervention: the questions, the most important features of both designs (in the form of an icon for the interface; screenshots are provided at color-plate 3 and 4), and focus points for the interviewer. Fourth, we summarize the findings of the interviews. Fifth, we provide a discussion where we interpret these in the light of our design questions, resulting in new questions for the next intervention. Intervention 1 starts with the general question about transparency versus back-channeling.

### 6.5.1 Intervention 1: a rotating form and a color coded logo

#### **Question(s)**

What are the advantages and disadvantages of a transparency interface compared with a back-channeling interface?

#### **Design rationale**

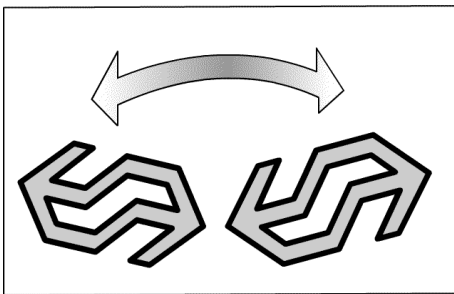
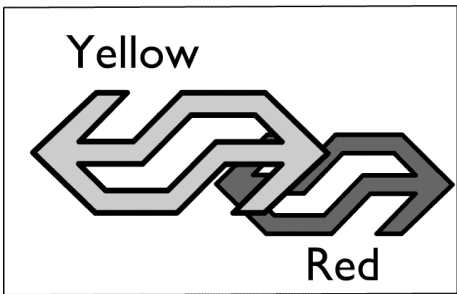
Following the general findings of the conceptual design study we only used a single element to give feedback to users in both the back-channeling and the transparency alternative. We decided to start with non-anthropomorphic metaphors. These seemed most suitable for focusing the users' attention to the single aspect we want to give feedback on.

For the back-channeling alternative we choose to give feedback on the gazing behavior of the participants. Since this has a larger impact on the classifier than speech behavior, it seemed to us that it was a more important aspect of their behavior to provide feedback on than their speech behavior. To do so we animated the form to look like it can orient to either the right or the left participant. This way the whole form could mimic human orientating behaviors, behaviors that provide back-channeling information about intention (see chapter 2.4). There were two different behaviors we could think of: speaker following and gaze following. A *speaker follower* looks at the speaker. A *gaze follower* looks along with the speaker. If the speaker looks at the screen, the form on the screen looks at the speaker while if the speaker looks at his partner the form on the screen also looks at the partner. It is likely that a human operator would use both behaviors at meaningful

moments in the dialog, but with the information we have in the system we have to pick one. We have chosen to use gaze following, because it seemed that by mirroring the speaker's gaze we might encourage the speaker to use his own gaze behavior to signal the addressee of an utterance.

For the transparency interface it seems natural to try a simple metaphor: color coding. With color coding we could indicate open or closeness of the system without the disadvantage that the parameters on the screen disappear like with the folding form (see section 6.3.3, figure 4). We did reuse this form, and kept the folding behavior at the beginning. Thus at start up –during the introduction question of the MATIS interface - the form folds closed and open one time, so that it makes a bow, hopefully this draws the users attention to the form and the logo. From then on, the form remains open and the color of the logo points out whether the agent is open for input or not. If the logo is yellow the system is open for input, if it is red the system is closed for input.

**Fact sheet**

|                     |   |  |
|---------------------|---|--|
| <b>Questions</b>    | <b>(dis)advantages of back-channeling</b>   | <b>(dis)advantages of transparency</b>   |
| <b>Interface</b>    |  |  |
| <b>Behavior</b>     | <b>Gaze following</b>   | <b>Color-coding</b>  |
| <b>Focus points</b> | -   | -  |

### Summary of participants reactions

| Pair                   | Reactions to the feedback  |
|------------------------|--|
| <b>I.A<sup>4</sup></b> | <p>Participants did not notice differences between the two interfaces. Also not when prompted. The only difference they noted was that in the second trial ‘they pushed the buttons more often’. These participants stated explicitly that they were too much involved in planning the trip to pay attention to such things.</p>   |
| <b>I.B</b>             | <p>Participants did not notice differences between the two interfaces. Perhaps the difference was that the first interface did not repeat all kinds of questions and the second did. When prompted about the visual feedback, participants say they did not notice the difference.</p> <p>These participants noted that the system picks up speech that is not intended for the system. They claim the system cannot deal with negotiation and should not be used by pairs.</p>  |
| <b>I.C</b>             | <p>Participants did not notice differences between the two interfaces. When prompted, they say they noticed a difference with the practice session. In the real session they needed to push buttons while in the practice session they did not (this is actually not true). They did not notice the logo change color or the form rotate.</p> <p>These participants had few discussions during the dialog. When asked about this they say they assumed only one person can use the system: ‘you can hardly push together’. When asked how they think the system solves this they think the rule is that the system takes what is said first.</p> <p>These participants ask for conclusive feedback. The system is slow so it should have a light to show the message is transferred. They actually thought they had done it wrongly.</p> |
| <b>I.D</b>             | <p>Participants did not notice differences between the two interfaces. Also not when prompted.</p>   |

---

<sup>4</sup> Pairs are labeled with a digit for the intervention and a letter to indicate which of the four pairs it concerns. Labeling is in chronological order of the session.

## **Discussion**

None of the participants noted a difference between the two interfaces. When prompted to list the differences participants start to talk about the performance of the system (speed, errors) and the options (buttons) rather than the feedback. So we cannot compare the transparency and back-channeling alternatives.

The data confirms an issue that has come up in the conceptual design study. The participants are so involved in the discussion that they are not able to interpret the feedback (1.A state this explicitly). In fact, in this specific design the problem seems to lie deeper than interpretation, participants do not even manage to attend to the feedback. For the back-channeling alternative the rotation of the form may not have been salient enough, but it could also be that a lack of metaphoric anthropomorphism makes the form go unnoticed. There is a potential mismatch between non-anthropomorphic metaphors and providing back-channeling signals (see section 6.2), so that even when people have seen the form rotate they may not have attended to it as something meaningful in the interaction. For the transparency alternative we may try to increase the salience of the feedback. In retrospect, the logo may have been an element of the interface that did not receive much attention in general. In intervention 2 then, we try to increase both salience and metaphoric anthropomorphism.

## 6.5.2 Intervention 2: a pair of eyes and a single eye

### Question(s)

Do people manage to attend to (and interpret) feedback if it is more salient and of higher metaphorical anthropomorphism than the interfaces of intervention 2?

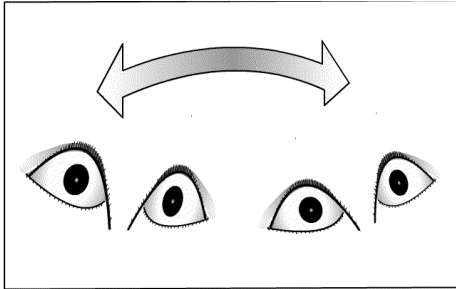
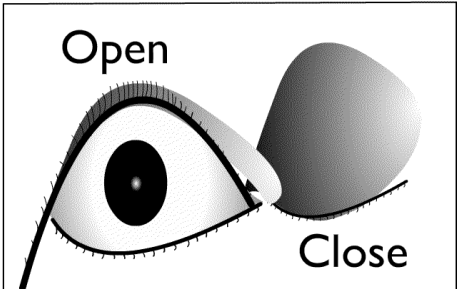
### Design rationale

This intervention aims at increasing the salience of the feedback and its metaphoric anthropomorphism in comparison to the color-coding and rotating form of intervention 1.

For the back-channeling alternative we felt putting eyes on the top of the rotating screen instead of the logo would stress the fact the screen was orienting. In addition this settled a conceptual link between the orienting behavior of the form and human orienting behavior, thus increasing its metaphoric anthropomorphism. We kept the gaze following behavior of the rotating screen.

For the transparency alternative we decided to move from color coding to iconic coding. We put an element on the screen alien to the metaphor of a form (in contrast to the logo). The idea was that if there is something ‘strange’ on the screen users would pay more attention to its behavior. To keep a certain unity with the back-channeling alternative we opted for an eye. We placed a large image of a single eye next to the form. This eye could be either open or closed depending on the status of the agent. The transition from open or closed was animated and during the period it was open its iris and pupil moved around as if looking to see what was happening around them.

### Fact sheet

| Questions    | Does increasing salience and metaphoric anthropomorphism help people to attend to (and understand) the feedback? |  |
|--------------|--|--|
| Interface    |                               |  |
| Behavior     | Gaze following   | Iconic coding  |
| Focus points | -  | -  |



### Summary of participants reactions

| Pair       | Reactions to the feedback  |
|------------|--|
| <b>2.A</b> | (no video was recorded)  |
| <b>2.B</b> | <p>Participants did not notice differences between the two interfaces: 'It might have been the second system was more accurate'. When told that the differences were in the visual design they claimed they had noted them. In one condition the form made a bow and in the other condition there were eyes on the form. They did not notice the big eye near the form.</p> <p>The participants felt the eyes were cute, but did not assume they had a function. They seemed merely decoration.</p>  |
| <b>2.C</b> | Participants did not notice differences between the two interfaces. Perhaps the first system made more errors, or in one of the systems hidden messages appeared on the screen. When prompted they also did not notice the differences listed.   |
| <b>2.D</b> | <p>Participants noticed the first system had two eyes and the second only one. The one eye became bigger and smaller during the dialog (probably they were referring to the eyes being open -small- or closed -big-, clearly the eye did not change size – but it appears to be that way: see color plate 3). These participants had not assigned any meaning to the eyes.</p> <p>Participants felt the eyes were 'scary', they 'looked along'. Felt one eye was better, it was more peaceful.</p> <p>When told that the system did not accept input as long as the big eye was closed, they were surprised. They did not realize the machine might not listen at some points in the dialog.</p> |

## Discussion

In a way our move towards more salient feedback of higher metaphoric anthropomorphism was successful. At least the participants in session 2.B and 2.D have noticed the differences between the two options. However, both these pairs had not assigned any meaning to the eyes in the two conditions. So we might have been successful in making participants *attending* to the feedback, we were unsuccessful in making them *interpret* the feedback.

Users may not be able to *interpret* the feedback because there is a mismatch between users' expectations of the system and the metaphors we use to communicate its status. The participants in session 2.D state explicitly that they did not realize the machine might not listen at some points in the dialog. This remark provides us with two clues to understand why. First, if users do not realize the system has the task of separating utterances for the system and utterances for the partner, it will be harder for them to interpret feedback about this feature. We decided to find out whether other users realize this problem by asking them explicitly in later interviews. Second, these users express the behavior of the system as *listening* (or not listening) rather than *open* (closed) *for input*. Seemingly these participants do think of the system behavior in anthropomorphic terms, but their terminology suggest we have used the wrong anthropomorphic metaphor. In retrospect we have translated the term open for input to *looking* at the participant while we should have communicated the machine is *listening* to the participant. Thus we may stick to anthropomorphic metaphors, but try to use an ear rather than an eye.

For the back-channeling alternative we can interpret the remark that the eyes were 'scary' in a positive way. At least in session 2.D we managed to increase the salience of the interface. But there are two critical remarks to make about this interpretation. First, although a goal was to increase the salience of the form, it was not the goal to introduce 'scary' elements. Second, participants have seen the eyes but do not mention the orienting behavior of the MATIS form (we also forgot to prompt for that). So, while the eyes have caught the attention of users, we may not have been successful in stressing the orientation behavior as we intended. As a first try to tackle this problem we may try to increase the metaphoric anthropomorphism of the interface even further. An extra advantage may be that the 'scary' effect of the eyes may be reduced.

### 6.5.3 Intervention 3: a traffic sign and a rotating bear

#### Question(s)

Are people able to attend to and interpret the feedback in the transparency alternative if we use an icon suggesting whether the agent is listening rather than whether the system is open for input?

Do people notice the orienting behavior of the form if we increase the metaphoric anthropomorphism of the metaphor we use?

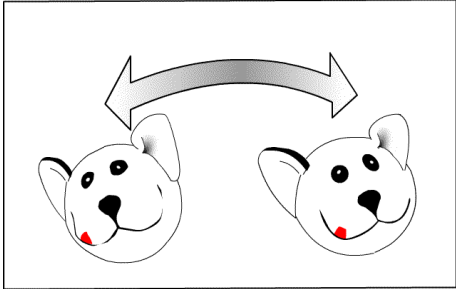
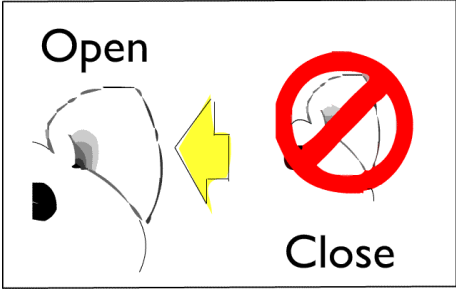
Are people aware the system has the task of separating utterances for the system and for the partner?

#### Design rationale

For the back-channeling alternative we try to increase the metaphoric anthropomorphism. We do this by putting the image of a bear on the side of the form (see color plate 4). We chose a bear rather than a human face for practical reasons: it was easier to animate. This choice may have the extra advantage that it be perceived as more friendly than a human face, so users will not find it ‘scary’. Using a bear at the side of the form allows us to combine gaze following and speaker following. We have chosen to use speaker following for the form. The form now orients to the speaker as if pointing out it expects the speaker to give input. The bear follows the speakers gaze, to show that it is ‘with’ the speaker; thus attending to the same target. To make sure participants attend to the bear, we have made the form and the bear switch sides. The bear is always closest to the speaker, and the form is at the listener side orienting to the speaker. (at color plate 4 the right person is speaking).

We reuse the bear for the transparency alternative (see colorplate 3). Since the bear has ears we can use those to indicate listening or not listening. When the agent is open for input there is a yellow arrow near the right ear of the bear. This suggests the ear is the important element and that it is listening. When the system is closed for input we put a prohibited sign across the ear of the bear. This suggests the bear is not listening.

**Fact sheet intervention 3**

|                     |   |  |
|---------------------|---|--|
| <b>Questions</b>    | <b>Are people aware the system has the task of separating utterances for the system and for the partner?</b>                                    |  |
|                     | <b>Do people notice the orienting behavior of the form if we increase the metaphoric anthropomorphism of the metaphor we use?</b>               | <b>Are people able to attend to and interpret the feedback in the transparency alternative if we use an icon pointing out whether the system is <i>listening</i> rather than whether the system is open for input?</b> |
| <b>Interface</b>    |    |    |
| <b>Behavior</b>     | <b>Gaze following</b>   | <b>Iconic coding</b>   |
| <b>Focus points</b> | <b>Ask whether the users are aware the system has the task of separating utterances intended for the system and not intended for the system</b> |  |

### Summary of participants reactions

| Pair       | General reactions to the feedback  |
|------------|--|
| <b>3.A</b> | <p>Have noticed the side switching of the back-channeling interface but not its rotation behavior and not the traffic sign in the transparency interface.</p> <p>(we forgot to ask whether they thought the system was always listening)</p>   |
| <b>3.B</b> | <p>Have noticed the traffic sign but not the side switching or orientation behavior.</p>   |
| <b>3.C</b> | <p>Participants have noticed the side switching (describe it as ‘flapping’, do not mention the rotation) and –when prompted- the traffic sign.</p> <p>When asked whether they felt they were talking to the animal or to the kiosk, they said they felt they were talking to the kiosk in general rather than to the animal.</p> <p>Participants stress they focused on the content during the interaction (rather than the aspects discussed in the interview).</p> |
| <b>3.D</b> | <p>Participants claim they did not notice any differences. The second system could have been faster, but it may also be that they were more efficient. When prompted they have noticed the bear with the traffic sign.</p> <p>These participants also said they would show they were not talking to the system by speaking less loud when addressing themselves (this is also visible in the video).</p>   |

| Pair       | Reactions to the back-channeling feedback   | Reactions to the transparency feedback  |
|------------|---|---|
| <b>3.A</b> | <p>When asked why they think the interface switches sides, they claim it may be to point out the difference between the from and toward field. Although they have not thought about this during the interaction.</p> <p>They describe the side-switching as annoying, disturbing, restless, needless for this type of information service. They preferred the ‘stable’ version.</p> |   |
| <b>3.B</b> |   | <p>Were able to interpret the traffic sign as listening not listening. They had noted that in one instance they had filled something but had to repeat. This made them realize the system was not always listening.</p> <p>The participants liked the fact the animal showed that it was listening.</p> |
| <b>3.C</b> | <p>When asked why they thought the interface switched sides, participants said they thought it was to attract attention.</p>  | <p>These participants thought the traffic sign was to show listening or not listening. Participants asked themselves why is it not listening?</p> <p>Participants think they want to know whether the system is listening or not but would not miss it if it was absent.</p>                            |
| <b>3.D</b> |   | <p>Think the bear suggested they could speak to the system or not. They claim they have used the sign. If there was one they waited. But also say they didn’t need to do so this session. They could not remember to have thought ‘when does the arrow appear?’.</p>                                    |

## Discussion

For the transparency alternative we have succeeded to make people attend to and interpret the feedback (at least for 3 of the 4 pairs) by switching from the eye to the ear as metaphor for the system status. For the back-channeling alternative, the move to more metaphoric anthropomorphic metaphors was less successful. Some pairs commented on the side switching but not the rotation of the form or the bear. Those people we asked, were indeed aware the system would not always listen. They claimed this awareness was because of either the feedback or the occurrence of errors (false rejections). In fact often it may have been the combination.

These results suggest that we need to improve on the back-channeling feedback. Participants did not notice the rotation of the form or the animal. There may be several reasons for that.

First, we may question the *salience* of this feedback. It could be the bear does not attract enough attention, or directs the attention to the wrong elements. The side switching attracted more attention than the rotation. We may remove the side switching but we must also wonder whether people see the rotation of the form and the bear. So far we have not tested whether the image of the form -facing right or left- was clear to users. It may be that by improving the animation of the form and the bear's rotation we do draw attention to that. Another way to improve salience of the rotation is by having a physical element of the kiosk rotate. The participants of session 3.C noted that they talked to the kiosk, rather than the animal. So it may be better to have elements of the kiosk itself show back-channeling signals, rather than elements on the screen.

Second, we must question the combination of *gaze following* and *speaker following*. Although the combination seemed elegant when we first imagined it, it does form a departure from our early attempts to keep the interfaces simple. We may limit ourselves to gaze following or even move to the simpler speaker following. Surely *if* gaze following is too complex to understand what the rotation is about, it cannot support users in their gaze behavior. Third, in this and in the previous intervention we have tried to enable users to attend to the back-channeling signals by increasing the *metaphoric anthropomorphism*. However, both these attempts were unsuccessful in the sense that the link between the metaphor and the behavior was absent. The metaphors: eyes on top of the form and a bear on the side of the form, were noticed but were not interpreted correctly. We could either continue the line of increasing metaphoric anthropomorphism or focus on other aspects first. At least for the next intervention we decided to focus on the possibilities to

have elements of the kiosk itself communicate back-channeling information and on the salience of this feedback.

For the transparency alternative we have reached a point where most pairs were able to attend to and interpret the feedback. Still the interviews have raised enough concerns to formulate follow up questions. The first is the occurrence of errors as a source of information that makes participants aware the system may not be listening. Participants mention errors in combination with feedback but surely errors have occurred in previous interventions as well. It may be that it is the clarity of the feedback alone that is responsible for the fact users are aware of the fact the system is not always listening. The second concern is the type of errors. Participants tend to mention false rejections: utterances intended for the system but classified as intended for the partner. False alarms in contrast, were mentioned explicitly by only a single pair (session 1B). Are false alarms another potential source of information for users to attend to feedback? Is our feedback as clear during false alarms as it is during false rejections? The third concern is the possible *use* of feedback. The participants in session 3.C claimed they would not miss the feedback if it was absent, and the participants in session 3.D claimed they would wait for the arrow if there was a traffic sign, but had not experienced they needed to do so. We didn't design the feedback in the transparency alternative to provide the user with guidance on how to deal with errors. We may try to gather information on how users expect to be able to correct or prevent errors in combination with the feedback we provide. Three interventions are not enough to spell out these issues in detail. We decided to take up the issue of false alarms first, and keep the other two issues in the back of our heads.

Looking forward to intervention 4 we felt that these two focus points, back-channel signals on the kiosk rather than the screen and false alarms were so different in nature that we could not address them in a single intervention. Therefore we have chosen to split intervention 4 into two half interventions (each with two instead of 4 pairs). This entails the risk of data sparseness but with two interventions left to back up any conclusions we might draw we felt it was worth the risk.



## 6.5.4 Intervention 4: Errors, and a rotating camera

### 6.5.4.1 Intervention 4a: a rotating camera

#### Question(s)

Are users aware of the back-channeling feedback if this feedback is provided by a physical element of the kiosk rather than onscreen feedback?

#### Design rationale

In the search for a physical element of the kiosk that could provide back-channeling responses we considered 3 possibilities: a rotating screen, a rotating camera, or a rotating microphone on the kiosk. We preferred a rotating screen, but because of the practical difficulties of such a choice, such as using a different kiosk frame and making a mechanical construction strong enough to rotate a touch screen we dismissed it soon. There is something to say for both the microphone (a metaphor for the ears of the kiosk) and the camera (a metaphor for its eyes) as a way to show users to whom the kiosk is attending. However, we chose the camera because it would be suitable for both gaze-following and speaker following. In the end we chose speaker following for this intervention because it was the simplest behavior.

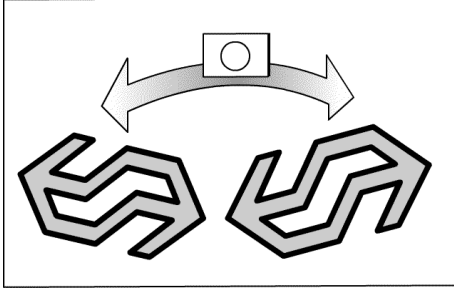
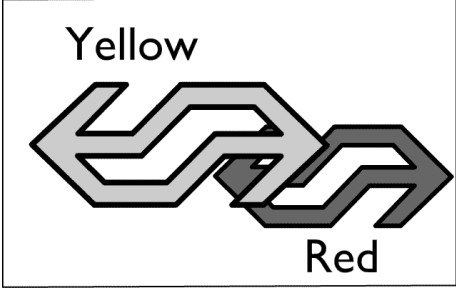
Figure 7 shows two images of the camera on top of the kiosk. We attached it on a Phidget (see: <http://www.phidgets.com>) servo with tape. Although this gave the camera a somewhat unprofessional look, this construction was strong enough to hold the camera in place. When the camera changes position the servo makes a hissing sound.

We reuse the color coding and rotating form of intervention 1 for this intervention. We changed the rotation behavior of the form to speaker following as well.



Figure 7: two images of the camera on top of the information kiosk.

**Fact sheet intervention 4a**

|                     |   |  |
|---------------------|---|--|
| <b>Questions</b>    | <b>Are users aware of the back-channeling feedback if this feedback is provided by a physical element of the kiosk rather than onscreen feedback?</b> |  |
| <b>Interface</b>    |    |  |
| <b>Behavior</b>     | <b>Speaker following</b>  | <b>Color coding</b>  |
| <b>Focus points</b> | -   | -  |

**Summary of participants reactions**

| <b>Pair</b> | <b>Reactions to the feedback</b>   |
|-------------|--|
| <b>4a.A</b> | <p><b>(note – the camera rotated in the wrong direction opposite to the speaker)</b></p> <p>These participants did not notice any differences between the two interfaces. When prompted they say they have noticed the camera rotated in the first (back-channeling) trial and not in the second (transparency). They claimed it was rotating towards the speaker. They did not notice the rotating onscreen form or the logo changing color.</p> <p>They have noticed that the system did not react in one case when they looked at each other. They also experienced a case where the system accepted the wrong input.</p> |
| <b>4a.B</b> | <p>These participants did not notice a difference between the two interfaces. Also not when prompted about the camera, rotating form and color of the logo.</p>  |

## Discussion

In one of the two session participants noted the camera, be it only after prompting for it. They formed correct hypotheses about its intended behavior although it rotated in the opposite direction. The onscreen feedback was not noted by these participants. Given the limited amount of data, and the extensive amount of data where users failed to notice the rotation we take these results as an encouragement of the rotating camera. But this rather positive interpretation of the facts needs to be supported in later interventions. We felt the evidence the camera is an improvement over the virtual feedback to be too weak to move to more complicated behavior of the camera such as gaze-following.

### 6.5.4.2 Intervention 4b: a different error protocol

#### Question(s)

Do disturbing false alarms encourage participants to attend to the transparency feedback?

#### Design rationale

Our conversational agent makes many false alarms. However, users seldom mention them. This may be because of our error protocol. When a false alarm occurs the wizard (see chapter 5.6) has to make a decision. There could be information in the utterance that could potentially be filled in the field, for example a station name, and in that case the wizard fills the field. Users may notice the wrong value has been filled. If there is no information that could potentially be filled in the field the wizard does nothing and the field is filled with question marks. These question marks stay in the interface until an the wizard fills the field. So, if there is a sequence of false alarms without information that could be filled by the wizard only the first false alarm can be noticed by users. We looked back at some of the videos and logs and in practice these unnoticed false alarms were frequent<sup>5</sup>. Therefore we decided to make changes to the wizard interface so that more disturbing false alarms would occur.

Each time an utterance is classified as intended for the system a random generator in the wizard interface produces a random number between 0 and 1. If this number is bigger

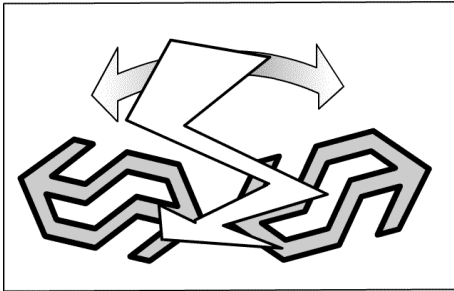
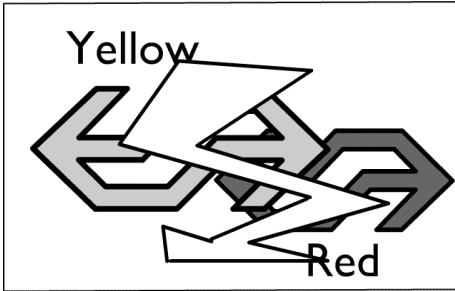
---

<sup>5</sup> There are two reasons for this. First with the ‘perfect’ speech recognition of a wizard the amount of utterances classified as intended for the system that also contain the value of a field such as a station name is very low. Second the wizard has to keep up with a high dialog pace and this makes it hard for the wizard to stick to the protocol. The wizard tends to focus on the utterances intended for the system, resulting in mistakes in judging the content of false alarms. Clearly our wizard has tried to keep the protocol as closely as possible, and we did not perform a post analysis on the amount of mistakes she made.

then 0.3 the dialog manager produces a random (but valid) value for the field that has to be filled. The wizard may overrule this if the utterance is intended for the system but in other cases the field is filled with the randomly selected value. So in about 30% of the false alarms a wrong value is filled and the dialog advances to the next empty field. We asked the wizard to follow the old protocol for the other 70% of false alarms. In practice the threshold of 0.3 resulted in one or two disturbing false alarms each dialog. We will call this protocol the *severe error protocol*.

We have chosen to reuse the color coding form of intervention 1. Since none of the six pairs has noticed the color coding feedback so far, the finding that the pairs subjected to false alarms would notice this feedback would be the strongest evidence we can find that false alarms encourage people to attend to feedback rather than the design of the feedback itself. For the back-channeling alternative we used the matching rotating form.

**Fact-sheet intervention 4b**

|                     |  |   |
|---------------------|--|---|
| <b>Questions</b>    | <b>Do disturbing false alarms encourage participants to attend to the transparency feedback?</b> |   |
| <b>Interface</b>    |                |  |
| <b>Behavior</b>     | <b>Speaker following + severe error protocol</b>   | <b>Color coding + severe error protocol</b>   |
| <b>Focus points</b> | -  | -   |

### Summary of participants reactions

| Pair        | General reactions to the feedback  |
|-------------|--|
| <b>4b.A</b> | <p>These participants thought the difference between the two interfaces was that the first system (color coding) was more accurate, and its questions better timed, than the second (rotating form). When prompted they have noted the logo changed color in the first condition but not the rotating form in the second.</p> <p>When asked the participant thinks the red logo pointed out the system is busy.</p> <p>The participants complain extensively on false alarms, and accuse the experimenter of faking station names. They are content with the error correction possibilities of the buttons but would also want to be able to solve these errors by speech.</p> |
| <b>4b.B</b> | <p>These participants noticed the rotating screen but not the color change of the logo.</p> <p>Users disliked the rotating form because the interface without rotation was more peaceful. They did not understand why the form rotated.</p> <p>These participants have noticed false alarms but did not complain as extensively as the participants in session 4b.A</p>  |

### Discussion

Users complained about false alarms in this intervention, thus we succeeded in making them ‘disturbing’. With some reserve, we can say that participants are encouraged to attend to feedback by disturbing false alarms. At least both pairs noticed the feedback in one of the two conditions. Because participants did not attend to both kinds of feedback we feel that we need to support this conclusion with more evidence.

### 6.5.4.3 General Discussion intervention 4

In intervention 4a we succeeded, for the first time, to make users attend to the rotating form. We managed to do so by using a rotating camera instead of a virtual rotating element and by simplifying its rotation behavior to speaker following, rather the gaze following. It is unclear why participants were able to attend to the camera; it could have been the simpler behavior, it could have been the fact that it is a physical element but it could also have been the hissing sound it makes. In addition, and more important, the evidence that participants are able to attend to a rotating camera is still weak and we do not know how people react to such feedback. In intervention 4b we found that disturbing false alarms make people attend to feedback, both of the back-channeling and the transparency type, be it that the evidence is weak as well. For both questions of intervention 4 we may conclude the answer is ‘yes’ but we need and will use intervention 5 and 6 to support these two conclusions.

### 6.5.5 Intervention 5: a bear with a camera

#### **Question(s)**

Do disturbing false alarms encourage participants to attend to the transparency, and back-channeling feedback?

Does introducing disturbing false alarms lead to a different interpretation and appreciation of the transparency and back-channeling feedback?

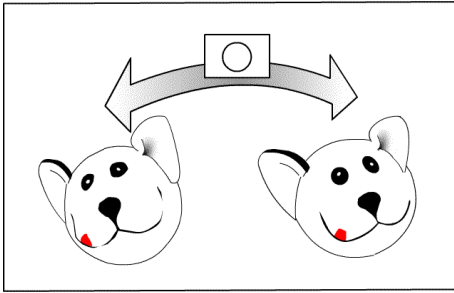
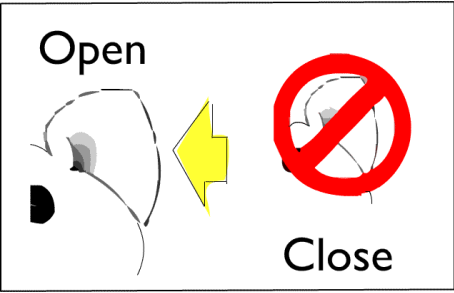
What are the relative advantages and disadvantages of the transparency alternative compared with the back-channeling alternative?

#### **Design rationale**

This intervention aims at backing up the conclusion that a rotating camera encourages participants to attend to back-channeling feedback. In addition we wonder how users feel about this back-channeling feedback. If they notice the rotating camera we may try to find relative advantages and disadvantages of the two feedback alternatives. This last question forces us to use transparency feedback that is clear to users. Therefore we return to the bear interfaces.

We adjusted the bear interface of the back-channeling alternative based on the criticisms of our participants in intervention 3. We omitted the side-switching, the form did not rotate any more, and the bear on the screen showed speaker following behavior rather than gaze-following behavior.

**Fact-sheet intervention 5**

|                     |  |  |
|---------------------|--|--|
| <b>Questions</b>    | <b>Do participants notice back-channeling feedback if physical elements on the kiosk are used rather than virtual elements rotating on the screen?</b> |  |
|                     | <b>How do participants interpret this feedback and do they react to it in any way?</b>   |  |
|                     | <b>(dis)advantages of back-channeling?</b>   | <b>(dis)advantages of transparency?</b>  |
| <b>Interface</b>    |   |  |
| <b>Behavior</b>     | <b>Speaker following</b>   | <b>Iconic coding</b>   |
| <b>Focus points</b> | -  | -  |



### Summary of participants reactions

| Pair       | General reactions to the feedback   |
|------------|---|
| <b>5.A</b> | These participants claim they have not noticed differences between the two interfaces, except that in one condition the camera moved. When prompted, participants also claim they have seen the traffic sign on the animal.   |
| <b>5.B</b> | Participants noticed both the camera rotating in one condition and the traffic sign on the bear in the other condition.   |
| <b>5.C</b> | Notice the camera (in the video they can be seen looking at the camera immediately), but are not able to say how the interfaces are different. Have not noted the traffic sign on the bear. Looked for differences related to the content such as pricing of their trips.                   |
| <b>5.D</b> | Although these participants claim they did not notice any differences between the two interfaces they come up with the camera and traffic sign without prompting.<br><br>They felt it would be bad if the system would pick up speech that belonged to the conversation but now it did not. |

| Pair | Reactions to back-channeling feedback   | Reactions to transparency the feedback   |
|------|---|--|
| 5.A  | <p>These participants thought the camera rotated towards the speaker, they did not know why it was there.</p> <p>These participants felt the moving camera was a nice feature, it made clear the system was attending to them. They did not feel they reacted to the camera in any way.</p> |  |
| 5.B  | <p>Participants thought the camera rotated towards the person speaking. Thought that they could not speak to the system if the camera was not oriented towards them.</p>  | <p>Thought the cross in the bear pointed out they where doing something wrong.</p> <p>Participants did not make use of the feedback, except that they tried to talk in the ear of the bear when addressing the system.</p> |
| 5.C  | <p>Participants state explicitly they looked at the camera because of the hissing noise it makes when it rotates. Thought the camera was orienting to the person speaking.</p>  |  |
| 5.D  | <p>They think the camera rotated towards the person speaking. They did not know why the traffic sign was there.</p>   |  |

## Discussion

The results of this intervention support the conclusion of intervention 4a: that the back-channeling feedback provided by rotating camera can be attended to and interpreted by participants. All participants noticed the camera and were able to interpret its speaker-following behavior. However, participants did not express they reacted to the camera in any way. This is similar to what we found with the transparency feedback in intervention 3. We have managed to enable people to attend to and interpret the feedback provided but not to suggest to them how to react to this feedback.

Like in intervention 3 most participants noted the traffic sign on the bear, be it sometimes only after prompting. Unlike intervention 3, participants did not generally provide us with a correct interpretation of this feedback. This might have been a matter of focus. Since the interviewer focused more on the camera, participants were less inclined to reason about the traffic sign. In that case we also must treat the results of intervention 3 with some caution: the interpretation of the traffic sign might have been a post-hoc rationalization of users. We may not assume it had been clear during the interaction itself.

As an advantage of the camera, participants note that it made clear that the system is attending to them. This may be listed as an advantage of back-channeling feedback, but clearly the message of the traffic sign is much the same. An advantage of the traffic sign is, that it may also communicate the system is *not* attending, while the rotating camera cannot. So far users have not expressed they want the system to indicate it is not attending. However, introducing false alarms may change this.

The question why the camera enabled users to attend to and interpret the feedback remained open, users may have attended to it because of the noise it makes when rotating, but also because it was a physical element of the kiosk.

Although this intervention raises new questions such as: ‘how can we support users in reacting to (back-channel) feedback?’, ‘is the traffic sign feedback clear during the interaction or is it just easy to understand afterwards?’ or ‘is the camera feedback attended to because of the noise or because it is a physical element?’, we decided to use the last intervention to back up the conclusions of intervention 4b. So in intervention 6 we focus on disturbing false alarms. Possibly, we also gain insight about the usefulness of communicating a lack of attention.

### 6.5.6 Intervention 6: a bear, a traffic sign, a camera, and false alarms

#### Question(s)

Do disturbing false alarms encourage participants to attend to (back-channeling and transparency) feedback?

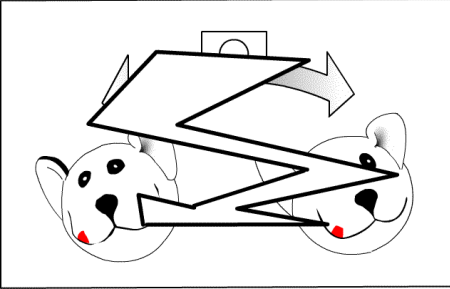
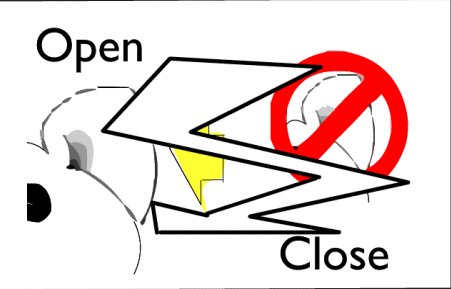
Do disturbing false alarms lead to a different interpretation and appreciation of the (back-channeling and transparency) feedback?

What are the relative advantages and disadvantages of transparency feedback compared with back-channeling feedback?

#### Design rationale

We felt it would be more informative if the results of this intervention would allow a comparison with intervention 5 rather than with intervention 4b. Therefore we used the same feedback as in intervention 5, except for the fact we used a severe error protocol rather than the standard error protocol.

#### Fact-sheet intervention 6

|                     |  |  |
|---------------------|--|--|
| <b>questions</b>    | <p><b>Do disturbing false alarms encourage participants to attend to the back-channeling feedback?</b></p> <p><b>Do introducing disturbing false alarms lead to a different interpretation and appreciation of back-channeling feedback?</b></p> <p><b>(dis) advantages of back-channeling feedback?</b></p> | <p><b>Do disturbing false alarms encourage participants to attend to the transparency feedback?</b></p> <p><b>Do disturbing false alarms lead to a different interpretation and appreciation of the transparency feedback?</b></p> <p><b>(dis)advantages of transparency feedback?</b></p> |
| <b>Interface</b>    |   |    |
| <b>Behavior</b>     | <p><b>Speaker following +severe error protocol</b></p>   | <p><b>Iconic coding + severe error protocol</b></p>  |
| <b>Focus points</b> | -  | -  |

**Summary of participants reactions**

| Pair | General reactions to the feedback   |
|------|---|
| 6.A  | <p>These participant have noticed the camera and the puppet did not move in the second scenario (transparency). They did not notice the traffic sign on the animal.</p>   |
| 6.B  | <p>(When prompted) these participants did notice the camera and the traffic sign.</p> <p>Think the system is always listening except when it asks a question. Seemingly the system cannot listen and talk at the same time. Have not thought about the possibility the system might listen when it was not needed.</p>  |
| 6.C  | <p>Have not noticed the camera movement or the traffic sign. They felt the real system responded less good than the practice system.</p> <p>Have noticed the system gives false station names. They think the system would decide that this station is closer to the zoo they want to go to. It is good they have the possibility to correct errors.</p> <p>Have noted the system fills stations when they are still in negotiation. Think the system tries to capture stations names. If station names appear it will assume it is intended for him. The second scenario was much more difficult because they had much more discussion.</p> <p>The system understood one participant badly. She compensated by speaking louder.</p>  |
| 6.D  | <p>Have noticed the traffic sign and the rotating camera (when prompted).</p> <p>These participants say the system fills places ‘nobody wants to go to’. But did not find this too annoying except for the case when the system filled a field while one participant had only said ‘aha’. This was annoying because they knew the system could not be right. The participants felt error correction went badly. The from and to field where filled, but after correcting the from field, it switched to the date field while the to field was also incorrect. Want to be able to do error correction by voice.</p> <p>Participants want to have feedback after filling a field before text appeared (conclusive feedback about the agents decision). They wondered whether the system was processing or did not hear him or whether he had to repeat or wait.</p> |

| Pair | Reactions to Back-channeling  | Reactions to Transparency  |
|------|---|--|
| 6.A  | <p>The participants thought the camera rotated to the speaker. They were surprised by the rotation of the camera at first but they got used to it, and hardly noticed it was not rotating in the second scenario.</p>   |  |
| 6.B  | <p>The participants thought the camera rotated to the speaker, they felt it was normal the camera pointed to them. They did not respond to the movement of the camera.</p>  | <p>The participants thought the traffic sign was some kind of logo. Did ask themselves why it was there, when the arrow appeared they thought they where not clear enough, but they did not try to solve it.</p> |
| 6.C  | -   | -  |
| 6.D  | <p>These participants thought the camera rotated to the speaker. Felt this was a good idea because then you would know whether you are listened to. One of the participants noted that the camera rotated towards him but did not give him the impression the system looked at him because it was not directed well enough. They also expressed privacy concerns and wanted to know whether the camera really records you or just speech.</p> | <p>The participants thought the arrow meant they could talk then. In the second trial it came too late. They realized only afterwards what the meaning of the stop sign was.</p>                                 |

## Discussion

The participants in this intervention complain about system errors. Users attended to error correction possibilities and some users became aware the system has to separate utterances for the system and user; However, we cannot say these errors make people attend to the feedback.

There is no real difference in the interpretations of the transparency or back channeling feedback provided by participants. We found one ‘new’ interpretation of the transparency feedback: the participants in session 6.B thought the arrow meant they were doing something wrong. But it is questionable if this is because of the different error protocol. Like in previous interventions we found users to attend to and interpret the feedback (correctly) but participants have no clue what to do with it.

We have found no evidence there is an ‘advantage’ of showing the system is *not* listening. This leaves the question about the advantages and disadvantages of transparency compared with back-channeling largely unanswered. One pair (in session 6.D) asked for conclusive feedback on the agent’s decision, this occurred in session 1.C as well. So, although we have put conclusive feedback out of the focus of this study, there may be a need for such feedback.

### 6.5.7 Additional remarks

Although the interviews in the six design interventions focused on the design of feedback for our conversational agent, participants made spontaneous remarks about the MATIS interface we used and about speech interfaces in general. In this section we summarize these comments.

Many comments about the interface addressed the functionality of the underlying system. Participants wanted more options: the possibility to specify an intermediate station, the possibility to see how long the travel lasts, the possibility to get information about busses as well or the possibility to get advice about the nearest station to a zoo. In chapter 1 we argued that the strengths of speech and language technology are to provide such flexibility, here we see users ask for those strengths. A second type of comment addressed the form filling nature of the interface. MATIS requires the user to fill all fields and then produces a travel advise in one go. Some users had difficulties with this. They would like to get information earlier. For example they complained about the arrival - departure parameter. They would like to be able to see immediately when a train arrives given a certain departure time or when a train departs given a certain arrival time. Taken together these comments suggest to explore the possibility to provide information about the trip early. The responses to ‘this type of technology’ fall into extremes. Most

participants claimed the scenario was realistic and that they could imagine such systems would appear in the real world and that they would talk to each other in front of such systems. These participants also praised the advantages for disabled people of not having to type. One pair (4b.B) went as far as claiming they would prefer this system for obtaining information over the ticket counter of the Dutch railways because they could handle this at their own pace. A minority expressed strong opinions against these types of systems. These participants pointed to problems such as hearing in noisy environments.

---

## 6.6 Conclusions and Discussion

---

In this chapter we have tried to answer the question how socially aware conversational agents should show users what they are doing. Users may have questions such as ‘can I speak to the system now’ or ‘did the system understand what I tried to communicate’ and clever feedback might provide users with answers to such questions. To structure our reasoning about this feedback, we worked out three interconnected design questions. *On what* aspects of their functionality should socially aware conversational agents provide feedback? What *metaphors* should they use for this feedback? And *when* should they give feedback? Following experiences in our conceptual design study we added a fourth question: how can we, given that we put users in a dual task situation provide feedback that users *understand*? Now, we revisit these four questions and see to what extent we managed to provide answers to them. In the next section we provide an attempt to specify recommendations for further research.

We have listed 3 aspects of the behavior of socially aware agents for which we may design feedback. First, it may provide feedback about the fact it is attending to the users’ non verbal behavior; we have called this *back-channeling* feedback. Second, it may provide feedback about the extent to which it estimates an utterance is intended for the system; we have called this *transparency* feedback. Finally, it may provide feedback about its decisions about each utterance; we have called this *conclusive* feedback. Much of our efforts went into comparing *back-channeling* feedback with *transparency* feedback with strikingly little result. There was no difference in the appreciation of both types of feedback in the qualitative study or the quantitative study. In part this ‘zero finding’ may be our own fault. In our efforts to create transparency information that users could understand we translated ‘open for input’ to ‘listening or not listening’. In practice, though, by doing this we linked transparency information (who is the intended addressee?) to a metaphor that is interpreted as a back-channeling response (is the system attending to my speech?). In other words, it seems that users only *understand* back-channeling feedback.



The second question was, what metaphors would be best to communicate different aspects of the behavior of the agent to users. To answer this question we made two distinctions: a distinction between *anthropomorphic* and *non-anthropomorphic* metaphors and a distinction between *metaphoric* and *interactive* anthropomorphism. This typology carried the suggestion there should be a match between the metaphoric and interactive anthropomorphism. So we should choose anthropomorphic metaphors for back-channeling purposes while we should choose non-anthropomorphic metaphors for transparency purposes. However, this suggestion might have been misguided. First, in the conceptual design phase we found there is a possibility to deliver transparency feedback with anthropomorphic metaphors (the railroad assistant interface) and back-channeling feedback with non-anthropomorphic metaphors (the water drop). Second, in the first three interventions we made a move to higher metaphorical anthropomorphism without success. In those interventions participants could not attend to and interpret the back-channeling feedback. Third, in intervention 4a and 5 we found a moving camera, arguably of lower metaphoric anthropomorphism but of higher salience, did have this effect. So, at least with our interactive anthropomorphism we found metaphors with low metaphorical anthropomorphism to be acceptable.

We have provided a provisional answer to the question *when* to deliver feedback after the conceptual design study and have let it rest from then on. We started out with a distinction between *feed-forward*, *early feedback* and *conclusive feedback*. It turned out to be hard to find metaphors in which these three types of feedback had their own place and worked with a combination of *feed-forward* and *early-feedback* from then on. The users in our intervention study did not comment on this, although some expressed a need for conclusive feedback. These users wanted to know what was going on between giving input and receiving their answer. These remarks cannot be interpreted as a straightforward call for conclusive feedback because we have not subjected these users to the disadvantages of such feedback when there are many false alarms.

We found the question of providing feedback that users can *understand* in a dual task situation to be of three-fold nature. First, users need to be able to *attend* to the feedback. We can achieve this making it salient enough: so a seamless integration of the feedback in the overall metaphor of the interface may be a bad idea. Second users need to be able to *interpret* the feedback. There has to be a match between the users expectations of a system and its behavior. Third, users need to be able to *use* the feedback. We have not succeeded to deliver feedback, that participants report to be useful.

So, how *should* socially aware conversational agents show users what they are doing? The hidden assumption of this chapter has been that it is good to make users aware the

system is not always listening. But this assumption may have been wrong. We started the chapter with the assumption users had questions such as ‘if I speak to the system now will it understand that I am addressing it’, or ‘did the system accept my input?’ and ‘why is the system doing this while I did not give input’ but these questions only come up if the interaction is hampered because of system errors. We found evidence for this in the interventions employing a severe error protocol (4b and 6). In these interventions, users spontaneously commented on the way an agent may separate utterances for the system and utterances for the partner. In other interventions they commented on this aspect only if we asked them for it (with exception of the participants in session 1.B). In other words: users expect the system to solve the problem of addressing, and errors and feedback (possibly the combination) can make the users aware of the fact the system fails to do so. Although we have succeeded in raising this awareness, the open question remains how users should *use* this feedback. As long as we do not answer this question the design of feedback for conversational agents remains a challenge. This leads to different recommendations depending on the system performance. If the agent does not make many errors, or the agent is designed in such a way that system errors do not hamper the interaction, we may omit feedback altogether. In that case there is no need to make users aware of the problem of addressing and we can provide a good match to their belief the system will solve it. If the number of errors is high and the system cannot be designed in a way these errors are not disturbing we need a new design effort to come up with feedback that users can attend to, interpret and *use*. In the remainder of this section we will try specify recommendations for such an effort.

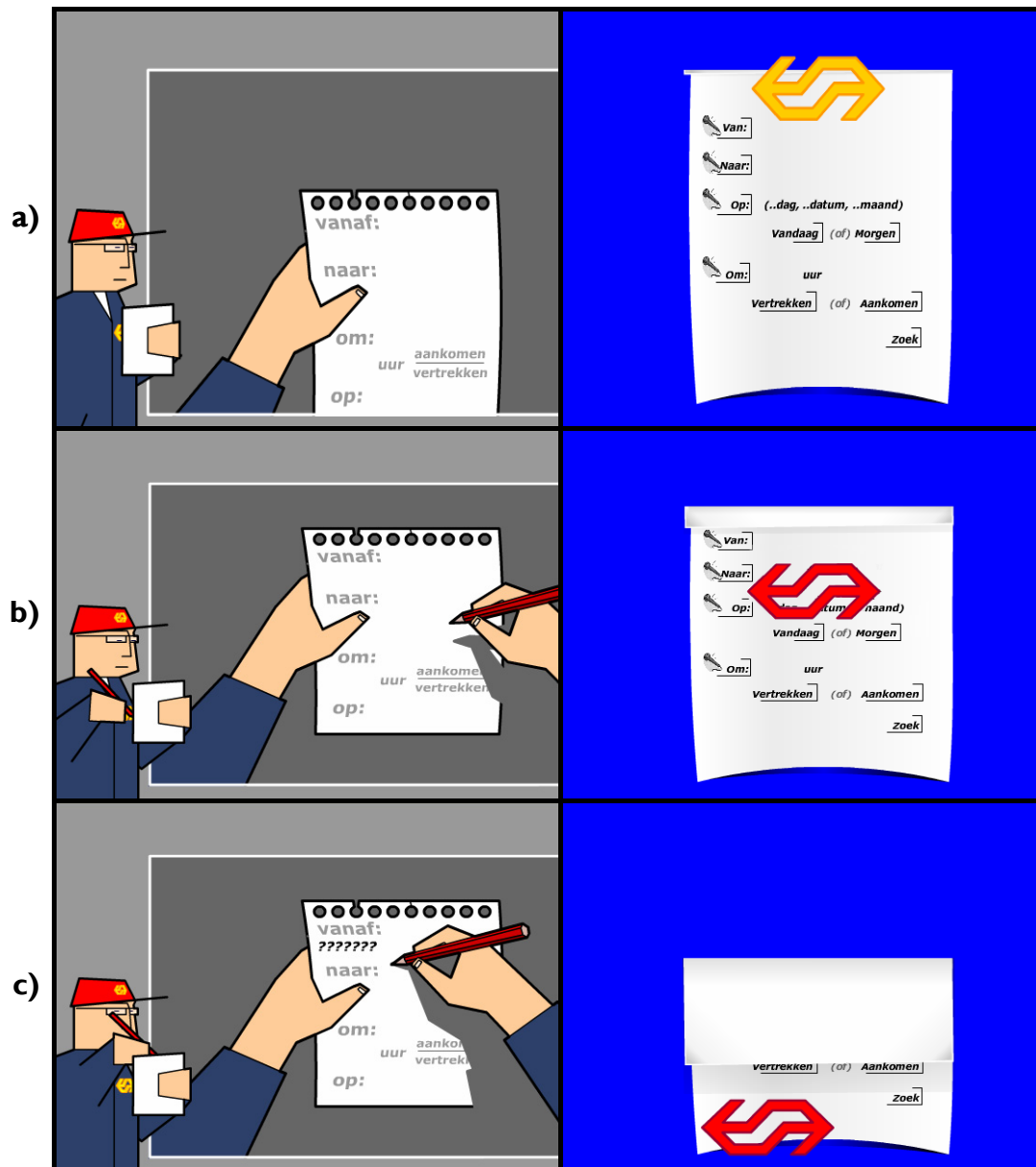
In a nutshell, such a new design (and research) effort should focus on delivering better back-channeling responses, with higher interactive anthropomorphism and with explicit attention to, but not entirely tied to, limited system performance. The choice for *back-channeling* responses follows more or less directly from our comparison of back-channeling with transparency feedback. We said back-channeling feedback is what users understand (in contrast to transparency feedback). However in our framing of the problem we defined back-channeling responses as feedback about the fact a conversational agent attends to users’ behavior. If we go back to the way humans deliver back-channeling responses we might say there is much more intelligence in the way humans use such responses. For example humans may differentiate between the type of responses they use according to their conversational role (side participant or addressee). We have not tried to tie the type of back-channeling responses to the inferred addressee (the transparency information), but this may be a key to arrive at an higher amount of interactive anthropomorphism of our interface. This may in turn lead to a better users’

understanding of the feedback we deliver. At the same time, we learned from our study that we cannot (and should not) straightforwardly copy human behaviors within an interface that is less good at inferring the intended addressee as humans are. So we need to pay attention to the occurrence of system errors as well. In summary then, a design effort to come up with better back-channeling responses suggests that multiple lines of research should be combined.

First, we do not know in sufficient detail how humans use back-channel responses according to their conversational role. So descriptive studies with a social psychological focus on the mapping between the use of back-channeling signals among humans and their conversational role (see: Terken, Joris, & De Valk, 2007) might provide us with insights on how to deliver more realistic back-channel responses for the multi-user case. However, such studies ‘merely’ describe human-human communication, and do not account for interaction with limited system performance.. Therefore such studies need to be combined with concrescent studies like the one described in this chapter, that take such constraints into account. In this second line of research, we might expand on the idea of a rotating camera, explore a rotating screen and modify its interactive anthropomorphism, for example using different rotating behaviors. Alternatively we might explore imitations of other human back-channeling signals such as nods or vocal acknowledgements. Such a line of research should shed light on what the minimal needs are for humans to get the feeling they can coordinate with the agent. Rather than showing how humans coordinate, such a line of research might show what it takes to get humans to react to system responses, ultimately allowing designers to put such reactions to good use in a ‘real’ conversational agent. Finally, we argued system feedback is needed when the amount of system errors is high. However, currently we do not understand this relation very well. To gain understanding about the triangle system errors, feedback, system flexibility we might need more systematic studies about system errors. In our case we have made assumptions about the way system errors affect the interaction with users, with a single interface with limited complexity. We tried two error protocols for the speech recognition (and dialog management) but we did not affect the performance of the AAD. However, with a wizard of Oz setup we have both the possibility to upgrade and to degrade the performance of such components. Upgrading can be done by combining the interpretation of a human operator with an automatic decision, degrading by combining the automatic decision with a random decision. So there is room to explore what flexibility can be delivered, with what system performance, and what feedback, although the system to deliver these things is not yet available.

## Butler

## Rolling Form



A butler is shown holding a notebook with the parameters for the system. A cutout shows this notebook in more detail.

a) When the system starts up, or does not expect input the butler has his hands down

b) The system expects input. The butlers' hand moves to the notebook, indicating it is ready to write

c) The system has accepted an utterance. The butlers' hand moves over the notebook as if it is writing. In this case the utterance could not be recognized, so question marks appear.

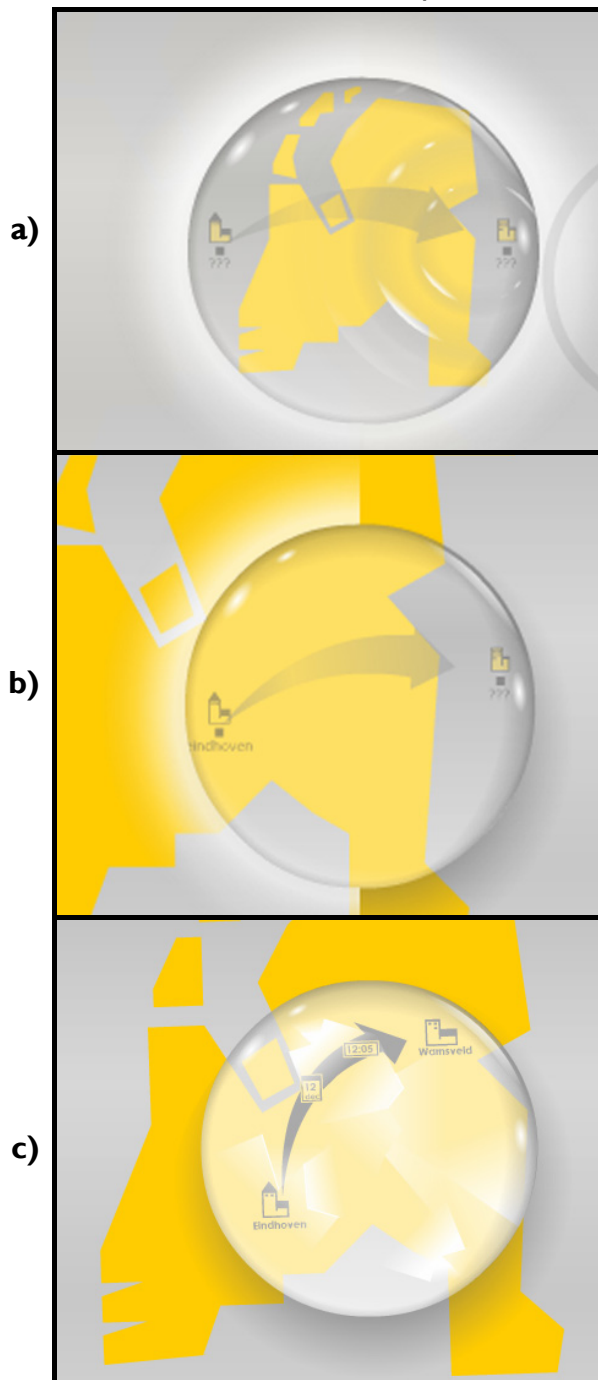
This form can fold itself open or closed depending on whether it is open or closed for input.

a) The system is open for input. Users can press the buttons and see values that are already filled.

b) The system is closing, the animation looks a bit like the form is making a bow. The logo falls off and changes color.

c) The system is closed for input, the values in the fields and the buttons are not visible.

### Bubble with map



A water drop lies like a magnifying glass on a map of the Netherlands.

a) the participant on the right of the kiosk is speaking. The system provides back-channeling signals by showing waves in the air and on the surface of the water drop.

b) The departure station is filled, the map moves so that the location (Eindhoven) is at the beginning of the arrow beneath the water drop.

Only the left backlight is on (in contrast to screenshot a)). This means only the participant on the left of the kiosk looks at the system. The participant on the right could be speaking, but the surface waves only appear when that speaker is also looking at the kiosk.





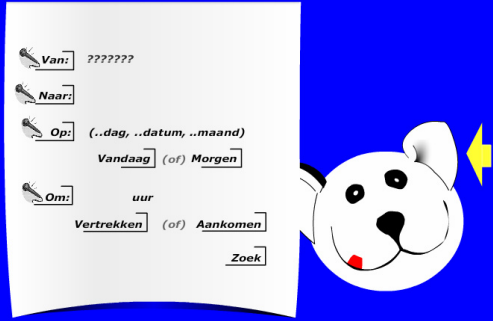
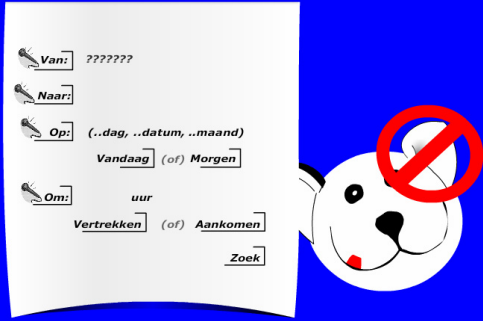
c) the arrival station is also filled, the map scales and the arrow (and the map) rotate so that the arrival station is at the end of the map.

The backlights are off, indicating both speakers look at each other. The system is also closed for input. This is visible because the water drop gets a frozen appearance.

Open for Input

Closed for Input

Color Coding

|  |   |
|--|---|
|  <p>Van: [ ]<br/>Naar: [ ]<br/>Op: [ ..dag, ..datum, ..maand ]<br/>Vandaag (of) Morgen<br/>Om: [ uur ]<br/>Vertrekken (of) Aankomen<br/>Zoek</p>          |  <p>Van: [ ]<br/>Naar: [ ]<br/>Op: [ ..dag, ..datum, ..maand ]<br/>Vandaag (of) Morgen<br/>Om: [ uur ]<br/>Vertrekken (of) Aankomen<br/>Zoek</p>          |
|  <p>Van: [ ]<br/>Naar: [ ]<br/>Op: [ ..dag, ..datum, ..maand ]<br/>Vandaag (of) Morgen<br/>Om: [ uur ]<br/>Vertrekken (of) Aankomen<br/>Zoek</p>         |  <p>Van: [ ?????? ]<br/>Naar: [ ]<br/>Op: [ ..dag, ..datum, ..maand ]<br/>Vandaag (of) Morgen<br/>Om: [ uur ]<br/>Vertrekken (of) Aankomen<br/>Zoek</p>  |
|  <p>Van: [ ?????? ]<br/>Naar: [ ]<br/>Op: [ ..dag, ..datum, ..maand ]<br/>Vandaag (of) Morgen<br/>Om: [ uur ]<br/>Vertrekken (of) Aankomen<br/>Zoek</p> |  <p>Van: [ ?????? ]<br/>Naar: [ ]<br/>Op: [ ..dag, ..datum, ..maand ]<br/>Vandaag (of) Morgen<br/>Om: [ uur ]<br/>Vertrekken (of) Aankomen<br/>Zoek</p> |

One-eye

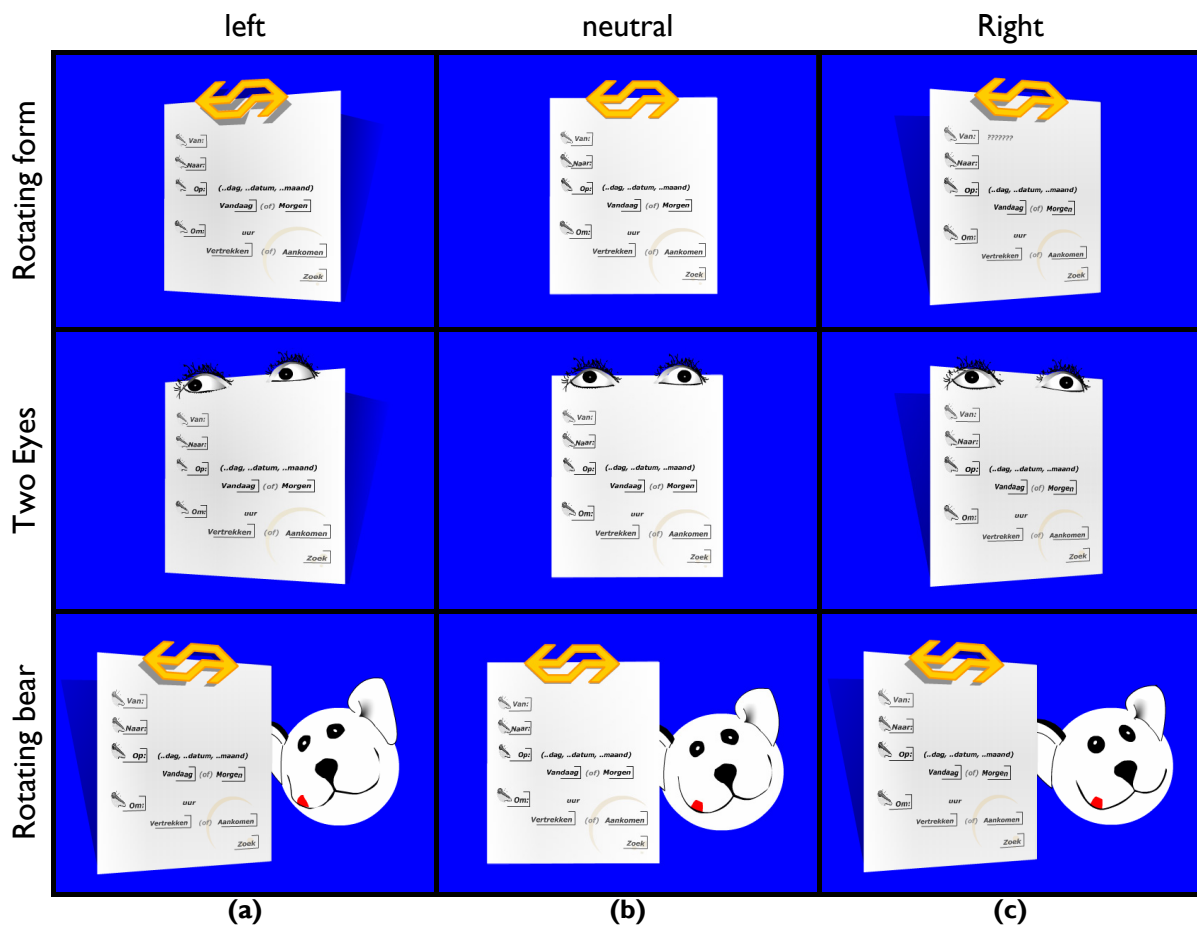
Bear with traffic sign

This table illustrates three ways to indicate the system is open for input or not.

In the 'color coding' option the NS logo on the top of the form changes color.

In the 'one-eye' option the opening and closing of a big animation of a human eye indicates open or closed

In the 'bear with traffic sign' option, an arrow appears near the ear of a bear on the screen to indicate the system is open for input, while a traffic sign across the ear of the bear shows it is closed for input.



This color plate shows three way's to indicate the agent is aware of the (visual) focus of attention of participants.

In the 'rotating form' interface the form is animated to point to the user at the left or right user, or to hold a neutral position.

If *gaze following* behavior is used the form orients to the speaker, unless the speaker orients to its partner, then the form also orients to the partner. If *speaker following* is used the form orients to the speaker, irrespective of the visual target of the speaker.

In the 'two-eyes' interface the form looks the same but two eyes on top of the form mark its orienting behavior.

In the 'rotating bear' interface the form and bear can rotate independently. This allows us to combine *speaker* and *gaze following*. Shown are:

(a) the right speaker speaks (the form orients right) but the speaker looks to the left participant (the bear orients left),

(b) there is a silence (both form and bear orient to the center), and

(c) the right speaker speaks and looks at the system (form and bear orient right).

Color plate 4: three examples of back-channeling feedback

## Chapter 7: Towards the Coordinated Development of Socially Aware Conversational Agents

*In this final chapter, we put the design case of this thesis in the perspective of the future development of socially aware conversational agents. First, we evaluate our design objectives for this case. We summarize the findings of all chapters and identify difficulties we encountered in moving from a social psychological to a systems engineering to a interaction design perspective. Also we revisit the original arguments for developing socially aware conversational agents and identify remaining challenges for developing agents that justify these original arguments. Next, we place the results of our work in the perspective of the development of other socially aware conversational agents than our case. We anticipate on the possibilities of a coordinated development effort of socially aware conversational agents across the underlying disciplines we visited in this thesis. For that we identify four new interdisciplinary challenges: specifying boundary conditions for communicative acts in technological terms, moving towards projective capabilities, finding situational determinants for human back-channeling responses and developing multidisciplinary roadmaps. We end the chapter with a final note on the added value of our interdisciplinary approach.*



---

## 7.1 Introduction

---

In chapter 1, we presented the design case of this thesis as an example of a larger class of solutions for shared use of speech-centric multimodal interfaces. We labeled those solutions socially aware conversational agents (SACA's). From then on, we concentrated on the design of a single example of such an agent from a threefold perspective. We have worked on our case from a social psychological, a systems engineering and a interaction design perspective. In this chapter, we zoom out and try to identify challenges for the development of socially aware conversational agents in general.

To do so, we start by evaluating the objectives for our design case. In chapter 1, we have outlined two main objectives. First, we aimed to explore the (technological) possibilities for making conversational agents socially aware. Second, we aimed at collecting the pieces of specialized knowledge needed to develop this type of solution and at uncovering interdisciplinary challenges for the domain of socially aware conversational agents. For the purpose of this chapter, it is convenient to review these objectives in reverse order. In section 7.2 we discuss the contributions and limitations of these contributions from the perspective of the three disciplinary fields that were involved in this study. Furthermore, we will look at the transitions from one chapter to the next. In section 7.3 we reflect on the (technological) possibilities of developing socially aware conversational agents. These two sections prepare the ground for section 7.4, where we pose new research challenges for the development of future socially aware conversational agents. A final note, in section 7.5, addresses the added value of the interdisciplinary nature of our work.

---

## 7.2 Contributions and connections

---

We divided the design case of this thesis into three subtasks. First, we tried to identify the relevant behaviors that socially aware conversational agents may attend to, in order to infer who is talking to whom and the semantics of these behaviors. For this subtask, we have borrowed methods and existing knowledge from the field of social psychology. Second, we tried to build a system that could automatically detect who is talking to whom in each utterance. We approached this problem from a systems engineering perspective. Third, we tried to uncover how socially aware conversational agents should provide feedback to users. We have addressed this last question from an interaction design

perspective. In this section, we critically review the contributions we delivered to these fields and we discuss the moves from each subtask to the next.

### 7.2.1 Identifying relevant human behaviors

In chapters 2 and 3, we have taken up a social psychological perspective to identify relevant (human) behaviors for inferring who is talking to whom in each utterance. In chapter 2, following Clark (1996), we positioned language use as a form of coordinated action. Following this view, we claimed that we can infer the addressee of an utterance indirectly from the way people coordinate their communicative acts. We identified three strategies to do so: keeping track of dialog history, seeking evidence for coordination within communicative acts and identifying differences in speaking styles between humans and machines. In chapter 3, we supported this claim by showing that we could operationalize these strategies for our technological frame and that they could indeed be used to identify the addressee of each utterance. Clearly, the value of the link between the theory of coordinated language and the findings from our situation is that we are, to some extent, capable to say how our findings generalize to other situations. One area of friction between the communication theory in Chapter 2 and the experiments in Chapter 3 is the difficulty of specifying its constructs and distinctions in technological terms. Clark's constructs (communicative acts, common ground) and distinctions (speakers, addressees, side participants) are just a few in a large ontological (see Jovanović, 2007, pp. 11-54, for a summary) and epistemological (see for example Keysar, 1997) debate. In this debate, the support for face-to-face communication with technology plays only a minor role. For example, the question how a *machine* may decide something is a single 'communicative act' has received little attention. A good, intermediate, step would have been to compare hand transcriptions using (analytic) constructs and their automatically derived counterparts. This way, we can strengthen the link between the results of chapter 2 and 3 and we might learn to what extent finding ways to achieve a closer match between the automatic and hand transcriptions should be a priority for the development of socially aware conversational agents. In summary, we may conclude that the social psychological body of knowledge turned out to be sufficiently apt for this subtask of the design case, although a challenge remains in concretizing its constructs. This can also be an serious hindrance in moving from a social psychological to a systems engineering perspective.

### 7.2.2 Determining the addressee of an utterance

In chapters 4 and 5 we have taken up a systems engineering perspective. The problem was to arrive at a system that could detect who is talking to whom automatically. In

chapter 4, we concentrated on the pattern classification and in chapter 5 on the real-time implementation of this classifier and the integration in a working demonstrator.

In chapter 4, we have come up with feature definitions grounded in the results of the experiments in chapter 3. We evaluated a single stochastic model to infer the addressee of an utterance: a naïve Bayes classifier. We saw that our features are complementary and we could assess the relative strength of these features for our case. Also, we demonstrated that our classifier was able to provide us with early estimates of addressee-hood. Unfortunately, it is difficult to compare our approach to other possible approaches. Colleagues working on related problems (Katzenmaier et al., 2004; Jovanović, 2007; Traum, 2004) applied different methods, on different corpora, evaluated with different evaluation measures. We concluded that there is a need for benchmarking. The step from chapter 3 to chapter 4 was fairly straightforward. However, there may be a weakness in the way we did this. We have opted for an –almost- one-to-one mapping between the metrics we used to settle that a strategy could be used to identify the addressee of an utterance and the feature definitions that we used in chapter 4. However the most suitable way to encode these data in the classifier may not necessarily be the most suitable way to find out whether a strategy works. Alternatively, we could have explored different ways to encode these data. Also, we may ask to what extent separating these two steps has been an efficient way to tackle the problem. If we have shown that a feature definition related to a strategy works, this is also evidence for the conclusion that the strategy works.

In chapter 5, we tackled the problem of implementing a real-time version of our demonstration platform. Here we followed a very pragmatic path, implementing a version that worked for our specific design case only. Still, in the real-time version of AAD we faced a problem of more general importance. Using the utterance as the unit of analysis has the advantage of hindsight. However, to be able to deliver estimates of addressee-hood before we can be certain the utterance is over, we need to transform the sequence of on-off events of speech into a model of the utterances and a model of the places of those utterances in the dialog. We need to do this in a way that we can account for simultaneous speech and mid-utterance silences. There is room for improvement of our solution, but the problem to be able to decide (or even predict) to what relevant unit the incoming speech or silence belongs, seems as urgent as the problem to decide what the intended addressee of the unit is. Identifying this problem may have been the primary contribution of the work in chapter 5. Possibly, there is relevant work in this area but for us it was a problem we initially overlooked.

In summary, the body of knowledge about pattern classification is large and well-established, but the application of this knowledge to our specific problem is still in an

exploratory stage. This makes it hard to compare the different approaches researchers working on this problem employ. There is a need for benchmarking. Also in our reasoning and working towards the interaction design of our system we identified two new challenges for this field. First, simultaneous speech and mid-utterance silences are a complicating factor for using a classifier in a real system. A central problem is to decide whether a particular sequence of speech belongs to the current or a new unit of analysis. Second, in providing feedback to users, timing plays an essential role. We should not only assess the quality of our estimates but also when we can arrive at these estimates.

### 7.2.3 Designing the interaction with users

In chapter 6, we have taken up the perspective of interaction design. Here we faced the problem of what Grudin (1994) called ‘the breakdown of intuitive decision making’. We found there is a lack of models and examples to guide the design of feedback for socially aware conversational agents. At the level of models we faced two limitations of Clark’s theory of language as coordinated action: the theory abstracts away from the situational knowledge needed to design feedback for conversational agents and the theory does not account for interaction under technological constraints. The other source of knowledge we could base our designs on: concrete examples of successful feedback in other domains, fell short as well. These examples do not scale to this domain because they have not been designed for systems that accept a range of ‘implicit’ input and have a form of agency. We contributed to the debate by devising a frame of reference for discussing design in our context, with distinctions like transparency versus back-channeling, interactive versus metaphoric anthropomorphism, and feed-forward, early feedback, and conclusive feedback. Subsequently, we delivered concrete examples of feedback that could be described in this frame of reference and we evaluated those examples with users. It turned out to be hard to come up with feedback that users could understand, in part because questions we posed reflected the users’ perception of our interfaces poorly. To reopen the debate, we proposed new studies in this field could focus on the way humans and systems can deliver effective back-channel responses.

The move from the work in chapters 4 and 5 to the work in chapter 6 has been far from trivial. We found that both the possibilities and the need for feedback on the inferences of the AAD depend on the performance of the classifier. The possibilities for feedback are constrained by the performance of the AAD. Because of system errors, the interpretation of the feedback may be hindered: the system behaviour may appear random. The need for feedback is dependent on the performance of the AAD, because human expectations of these types of systems rise far above the actual capabilities of

current systems. We proposed that feedback on the inferences of the AAD could solve this problem to some extent, by lowering human expectations on the one hand, and gently supporting users in behaving appropriately for the system on the other hand. However, we failed to show that this is possible. One resolution to the dependency of the possibilities for interaction design on the number of errors is to take some distance from the actual errors of system. A partial wizard, where the interpretation of a classifier is combined with the interpretation of a wizard, might allow to test the interaction design for better (future) classifiers, without compromising techno-ecological validity<sup>1</sup> too much. This may enable us to get a better grip on the minimum performance of a classifier.

In summary, in trying to answer the question of how we should design the interaction of conversational agents, we faced an immature scientific debate. We found that the possibilities for the interaction design and the needs that the interaction design has to fulfil, depend strongly on the technological possibilities. To progress, the field both has to take into account the technological possibilities and limitations, and it has to abstract away from those technological limitations.

### **7.2.4 An integrative theory for socially aware conversational agents?**

In chapter 2, we argued that the theory of language as coordinated action could enable us to ‘orchestrate our efforts across different disciplines’ by providing a ‘single theoretical framework’ to which we could ‘relate the three subtasks of the design’. We may conclude that the theory did not live up to this expectation. It had a strong impact on the work in chapter 3, where we used it to arrive at tactics for finding out who is talking to whom, but the impact on the other chapters was much smaller. In chapter 4, the theory served as inspiration for developing the possibility of early estimates, and our feature definitions were also based on this theory. The problem of finding an appropriate knowledge representation could not be covered by this theory. There have been efforts to create computational dialog models based on the theories proposed by conversation analysts and Clark’s grounding hypothesis (Thórisson, 1996), but so far these have not covered the problem of addressee determination. In chapter 6, we also found the theory of language as coordinated action to be of limited value. In devising the three design questions we took inspiration from the theory, but also from the technological possibilities and the constraints of our classifier. There is a need for models on an intermediate level of abstraction that can account for interaction under technological

---

<sup>1</sup> With techno-ecological validity we refer to the amount a prototype resembles ‘the real’ system

constraints. In retrospect, the ambition to ground all three different subtasks in a single theoretical framework may have been too high.

---

### 7.3 (Technological) possibilities and challenges

---

In chapter 1, we presented arguments for making conversational agents socially aware and claimed that it is possible to do so with current state-of-the-art technology. In the rest of the thesis we focused on a case study, partially intended to explore this possibility. The case study can be seen as a, modest, first step towards developing socially aware conversational agents that have the qualities needed to live up to the original arguments. In this section we summarize the original arguments, and then see what improvements are needed.

To justify the development of socially aware conversational agents, we presented four arguments. First, in many social situations people may be supported with automated information services. These allow them to find very specific information in a potentially large database. Second, providing this possibility is a strength of language technology. Third, in social situations, typing may be inconvenient, so speech provides a convenient way to interact with this technology. Fourth, socially aware conversational agents provide a better match to human expectations than explicit interaction protocols such as push to talk do. The design case of this thesis was not intended to test these arguments, but in chapter 6, we did find some support for, in particular, the first and the fourth argument. These two arguments form the basis for the discussion about the challenges for developing socially aware conversational agents.

In chapter 1, we said that the current MATIS system was a convenient starting point for developing conversational agents, but also somewhat rigid and limited. Indeed, in chapter 6, participants using our version of the MATIS system asked for more functionality, such as obtaining information about buses, and the location of specific tourist attractions. This supports the assumption that participants would like to have a system providing access to a larger database. However, if we would want to increase the functionality of the system, we need improvements of the performance of the AAD. In our demonstration, false alarms have been relatively harmless, because many utterances that were intended for the partner did not contain information that could potentially be filled in the field. For a system with more functionality, and possibly a more flexible dialog, we need a larger language model. This means that more utterances, intended for the partner, will contain speech that could be relevant for the system at that point, and false alarms become more disturbing.

In the fourth argument, we positioned socially aware conversational agents as an alternative to explicit interaction protocols such as push to talk. We claimed that socially aware conversational agents may provide a better match to human expectations than systems employing such a protocol do. Indeed, like in the studies of Magglio et al. (2000) and Brummit & Caditz (2001) we found that users did not expect that they would need to use an explicit interaction protocol. The expectation of our users was that the system would solve the problem of addressing automatically. But, we also found users expect the system to do it *right*. So, ideal socially aware conversational agents provide a better match to users' expectations but real socially aware conversational agents do not necessarily do so. This does not automatically mean that these agents need to be perfect before they can be a worthy alternative to explicit interaction protocols. If the interaction design can be arranged in such a way that the agent's errors do little harm to the interaction, such errors might be quite acceptable. For example, in chapter 6 we concluded that feedback on the interpretation of the AAD was not necessary in the light error protocol, while it was needed in the severe protocol (see: Chapter 6.4). However, the possibilities for interaction design do depend on the performance of the AAD. So in order to develop conversational agents that are a good alternative to explicit interaction protocols, improving the performance of the AAD and improving the interaction design should go hand in hand.

---

### 7.4 New challenges for developing SACA's

---

Having discussed the design case of this thesis in some detail, we address the more general domain of socially aware conversational agents. We have delineated this domain with constituting requirements and illustrated it with three scenarios in chapter 1. Besides illustrating the domain, we used these scenarios to make a general point: the challenges involved in developing socially aware conversational agents cross traditional disciplinary boundaries, and these challenges are interrelated. Indeed, the design case presented in this thesis shows there are such interrelated challenges. The scenarios also ranged in complexity and the level of detail of our specification. Not surprisingly, it is harder to imagine both the challenges and their interdependencies if the scenario is more complex and less specified. In this section, we try to extrapolate some of the –unsolved– interdisciplinary challenges we faced to projects that concern the field in general. We will only enlist those challenges that we feel to be of importance for the general class of socially aware conversational agents.

*Specifying boundary conditions for communicative acts in technological terms.* Within this thesis, we have used the *utterance*, specified in terms of on-off patterns of speech, as unit of analysis.

This was a concession to our technological frame in the sense that, within this frame, we did not have the possibility to use the communicative act as unit of analysis. We did not have a way to recognize something was a ‘single’ communicative act. The problem, however, remains if we use a more sophisticated technological frame. At least within different branches of the social psychological discipline, different units of analysis are used (compare: Goodwin, 1981 pp. 2 with Kendon, 2004, pp. 7). The units are usually specified in high-level terms. In contrast, machines will need low-level characteristics such as on-off patterns of speech, prosodic pitch, speech rate, gestures and so on. Clearly a high-level, semantic matching is also possible but only if its results can be obtained fast enough. If the challenge to specify boundary conditions in low-level terms turns out to be too hard, it is of importance to see how using a different unit of analysis affects the classification results of the addressees of these units.

*Moving towards ‘projective’ abilities.* In chapter 4 we have examined to what extent our stochastic model is capable of delivering early estimates of addressee-hood. This was inspired by the theory of language as coordinated action and our thoughts about the interaction design. In human-human communication, people seek out early evidence of successful communication. Apparently, such early feedback is a human need. Therefore, we felt the interaction design of our system should support such needs. It must be noted that we did not test directly whether such a need exists and to what extent our design solutions (should) support this need. Still, the general point remains that in communication with a system timing is of importance, in particular for smooth turn taking. Human turn taking is hypothesized to occur with almost no gap and only short periods of simultaneous speech because humans are capable of deciding early and precisely when an utterance is likely to be over (the projective view of communication, see Sacks et al. 1974; De Ruiter, Mittere, & Enfield, 2006). So to have systems that smoothly take part in human-human dialogs, it is not only important *what* we can infer from our feature set, but also *when* these inferences become available. Predicting what is yet to come is a fundamental ability we need to work on.

*Situational determinants of human back-channeling responses and ‘minimal needs’.* In the study of chapter 6, one line of investigation concerned mimicking human back-channeling responses. This line of investigation was loosely inspired by the work of Cassell et al. (1999b), that showed that embodied conversational agents have their merits for human computer interaction. We think it is valuable to explore the merits and limits of imitating human back-channeling responses further. From the perspective of socially aware conversational agents, we like to draw the attention to two central questions. First, we do not know when to imitate what behavior, when it comes to an operator that assumes it is



addressed or not (Terken, Joris, & De Valk, 2007). Second, we must test to what extent backchannel responses of a system need to resemble human responses, in order to be functional. In other words: what are the minimal needs for back-channeling responses to support humans in coordinating their communicative acts with systems.

*Multidisciplinary roadmaps.* In this thesis we have encountered the mutual dependency between interaction design and systems engineering challenges involved in developing socially aware conversational agents. We have seen that for developing addressee determination for interactive systems, it is valuable to have a corpus in which human interactions with (an emulation of) such a system are already present. People respond differently to interactive systems than they do to their human counterparts and knowledge of the system behavior, certainly if it is proactive, helps the classification. But we have also seen that the challenges for interaction design, in turn, depend strongly on the technological possibilities. When creating an emulation of an interactive system, it is of importance to account for the (expected) limitations of the technology that can be developed on basis of the emulation. Therefore, in planning the multidisciplinary development of socially aware conversational agents, both the interaction design and the systems engineering perspective need to be taken into account. A multidisciplinary roadmap might facilitate such plans. Road-mapping is a common technique in the field of engineering. Roadmaps provide an educated guess on the future by spelling out what intermediate technologies enable what final technology. We propose that, for developing socially aware conversational agents it is also of importance to map what intermediate technology enables what intermediate interaction designs for socially aware conversational agents. This allows for regular cross-checks on the inevitable assumptions each field must make about the other.

---

### 7.5 A final note

---

In chapter 1, we have described the work in this thesis as an interdisciplinary, integrative approach with explicit attention to the standards of the contributing disciplines. Inevitably, adopting such a broad approach to the problem, also meant we needed to make compromises about the depth in which we treated each discipline. Some of the limitations of this thesis originate from this lack of depth. Still, we felt there was much added value of going full cycle in this early stage of developing socially aware conversational agents. By doing this, we have been able to identify connections between the different contributing disciplines. Connections that, we feel, should be the main course rather than the dessert of the meal. We feel that the future development of

conversational agents requires both unidisciplinary and interdisciplinary projects. Unidisciplinary projects are needed to reach depth, interdisciplinary projects are needed to map out the many things one tends to forget when reasoning from the standards, problems and methods of a certain discipline.

## Appendix A: Scenarios from the Experiments in Chapters 3 and 6

---

### A.1 Original scenarios (in Dutch)

---

#### **Instructie**

Met behulp van de informatiezuil die in deze ruimte staat kunt u treinreis informatie opvragen. Het is een computer waar u tegen kunt praten en waarvan u het scherm kunt gebruiken als touchscreen om op te klikken. Wij zijn geïnteresseerd in de vraag hoe je zo'n systeem het beste kan ontwerpen als er meerdere personen gebruik van maken. Daarom willen we jullie vragen een rollenspel uit te voeren, en je mening te geven over verschillende versies van het systeem.

Het 'rooster' voor het experiment staat op de volgende bladzijde. Nadat jullie de gelegenheid hebben gehad om met het systeem te oefenen vragen we je twee keer om samen een dagje uit plannen met behulp van een versie van dit systeem. Na elke reis volgt een kort interview en een vragenlijst.

Voor elk scenario krijgen jullie beiden informatie over de reis die jullie gaan maken, het is de bedoeling dat jullie gaan overleggen over die reis. De informatie die jullie krijgen is verschillend. Het lijkt misschien efficiënter om voordat jullie beginnen de hele reis al met elkaar te overleggen, maar het is voor ons het meest informatief als jullie tijdens het plannen van de reis met het systeem het overleg voeren. Succes!

## Rooster

---

**Instructie**

**Oefenen**

**Scenario 1**

**Vragenlijst**

**Scenario 2**

**Vragenlijst en Interview**

**Administratie**

---

## Dierentuin: Proefpersoon I

Jullie willen morgen samen een dagje uit, jullie willen graag naar een dierentuin gaan in Nederland. Hieronder staan een aantal dierentuinen, u kunt hier een keuze uit maken, maar u kunt ook een andere dierentuin bedenken. Jullie gaan met de trein en willen redelijk op tijd daar zijn. Probeer samen een zo gunstig mogelijke reis te plannen waar jullie het beide mee eens zijn. Probeer ook alvast de terugreis te plannen.



**Rotterdam**

Diergaarde Blijdorp is met ruim 1,75 miljoen bezoekers per jaar (2001) één van de meest populaire attracties in Nederland.

### **Nieuwe Bizonprairie**

Eind april 2002 is de nieuwe bizonprairie geopend. Dit is het eerste gebied van "Noord-Amerika" dat in het uitbreidingsgebied van Blijdorp ligt (tegenover het nieuwe waterwerelddeel Oceanium). Het nieuwe verblijf is van u gescheiden door roosters. Zo staat u oog in oog met de imposante runderen.

---



## Emmen

In het *Noorder Dierenpark* maakt u in één dag een reis om de wereld. Dieren zijn er gehuisvest in het werelddeel waar ze van oorsprong vandaan komen. Uw wandeling begint bij het Biochron, een spannend museum dat het verhaal vertelt van het ontstaan van het eerste leven. Vervolgens ontdekt u Azië, Afrika, Amerika, Australië en Europa.

### Nieuwe Zoo-show over pinguïns

Voorzichtig komen de kinderen dichterbij om even aan de snavel te voelen hoe scherp de punt is. Spannend is het om te kijken of je groter bent dan een keizerspinguïn. In de splinternieuwe theatershow van het Noorder Dierenpark in Emmen staat dit jaar de pinguïn centraal. Het podium is omgebouwd tot een strand aan de rotsachtige Peruaanse kust. Op dat strand staat een Peruaanse visser, die de kinderen uitnodigt op het strand naar allerlei voorwerpen te komen zoeken.



## Harderwijk

Sinds de opening in 1965 is *Dolfinarium Harderwijk* uitgegroeid tot een uniek themapark. Het grootste zeedierenpark van Europa biedt een indrukwekkende collectie zee(zoog)dieren, waaronder dolfijnen, walrussen, zeeleeuwen, zehonden, roggen en haaien. Het Dolfinarium is ook toonaangevend op het gebied van onderzoek naar zeezoogdieren.

### Feestelijke activiteiten rond Lagune

De Lagune, een uniek natuurlijk leefgebied voor dolfijnen, zeeleeuwen en vele andere zeedieren in Dolfinarium Harderwijk, bestaat dit jaar vijf jaar. Ter ere van dit jubileum staat de maand juli in het teken van een aantal feestelijke activiteiten voor jong en oud. Men kan onder meer genieten van schitterende zandsculpturen van zeezoogdieren op het strand en van een bijzondere jubileumpresentatie in De Lagune.

## Dierentuin: Proefpersoon 2

Jullie willen morgen samen een dagje uit, jullie willen graag naar een dierentuin gaan in Nederland. Hieronder staan een aantal voorbeelden van dierentuinen in Nederland, u kunt hier een keuze uit kunt maken of zelf een andere dierentuin voorstellen die u graag zou willen bezoeken. Jullie gaan met de trein en willen redelijk op tijd daar zijn. Probeer samen een zo gunstig mogelijke reis te plannen waar jullie het beide mee eens zijn. Probeer ook alvast de terugreis te plannen want u wilt graag voor het donker thuis zijn.



**Amsterdam**

In *Artis*, de tuin van het leven, vind je van alles over het ontstaan van onze aarde, ons melkwegstelsel, het eerste leven op aarde en de geschiedenis hiervan, in het heden en het verre verleden, op onze aarde of ver daarbuiten.

### **Ontdek Artis in de palm van je hand**

Wie lekker door Artis wil dolen, kan dat voortaan doen met de Elektronische Artigids, ontwikkeld door Yellow Brick Road Design - gelieerd aan de TU Delft. De handcomputer vertelt je waar je bent, stippelt de kortste route uit naar iedere bezienswaardigheid, attendeert je tussentijds op het feit dat over 10 minuten het programma Varen op de Sterren in het Planetarium begint, of dat zo meteen de speciale rondleiding bij de Apenrots start.

---



### Apeldoorn

Zoals u de dieren op *Apenheul* ziet, ziet u ze in geen enkele dierentuin ter wereld. Tal van soorten; van de reuzengrote gorilla tot het allerkleinste aapje ter wereld, bewonen hier hun eigen ruime bosgebieden. Een groot aantal apen klimt en klautert zelfs helemaal vrij tussen de bezoekers. En...wist u dat u tussen de apen ook nog een heleboel andere dieren op Apenheul tegenkomt?

#### Geboortegolf doodshoofdaapjes

Onlangs is op Apenheul de geboortegolf bij de doodshoofdaapjes op gang gekomen. De dierverzorgers gissen voorlopig nog naar het aantal baby's dat dit jaar ter wereld komt, maar het aantal kan oplopen tot 20 à 30 baby's binnen enkele weken. De baby's worden bijna altijd 's nachts geboren en zijn de volgende ochtend al op moeders rug te zien, in de bomen en tussen de bezoekers.



### Arnhem

Burgers zoo probeert het natuurlijke gedrag van dieren te tonen, waar mogelijk in relatie tot een zo natuurlijk mogelijk leefgebied. Er zijn kunstmatige eco-systemen waarin de bezoekers op ontdekkingsreis kunnen gaan. In 1968 opent Burgers' het eerste Safaripark, gevolgd door het chimpanseeterritorium, het wolvenbos, de Bush (een overdekt oerwoud), de Desert, en de Ocean.

#### 's Avonds geopend

In de zomer blijft Burgers' Zoo tot 22.00u open voor bezoekers. Vanaf 19.00u worden er diverse rondleidingen georganiseerd, is er een kijkje in de centrale keuken van het park mogelijk, zijn er informatiestands te bezoeken en kan er genoten worden van een poppentheater of een verhalenverteller.

## Museum: Proefpersoon I

Jullie willen in het weekend samen een dagje uit, en willen graag naar een museum in Nederland gaan. Hieronder staan enkele musea, probeer daar een keus uit te maken of verzin zelf een museum waar u graag naar toe zou willen gaan. Jullie gaan met de trein, een halve dag lijkt wel voldoende om een museum te bezoeken. Probeer samen een zo gunstig mogelijke reis te plannen waar jullie het beide mee eens zijn. Probeer ook alvast de terugreis te plannen.

---

### **DE PONT**

#### **Tilburg**

De Pont is vernoemd naar de jurist en zakenman mr. J.H. de Pont (1915-1987) uit wiens nalatenschap in 1988 een stichting 'ter stimulering van de hedendaagse kunst' kon worden opgericht. De Pont is gevestigd in een voormalige wolspinnerij in Tilburg die door bureau Benthem Crouwel Architecten is verbouwd tot een ruimte waar hedendaagse kunst optimaal tot haar recht kan komen. De monumentale oude fabriek met de grote, lichte zaal en de intieme 'wolhokken' vormt een prachtige omgeving voor de vele kunstwerken. De Pont is sinds september 1992 voor het publiek geopend.

#### **Expositie Anton Henning**

Werk van de Duitse kunstenaar Anton Henning is in Nederland nog maar weinig te zien geweest. Toch neemt de internationale belangstelling voor zijn schilderijen en installaties de laatste tijd sterk toe. Henning wordt beschouwd als een van de jongere kunstenaars die, na gearriveerde grootheden als Richter, Polke en Baselitz, de schilderkunst weer met nieuw elan tegemoet treden.

---





huis Marseille

### **Amsterdam**

Met de opening van huis Marseille in september 1999 is er voor het eerst in Amsterdam een plaats voor de fotografie gerealiseerd waar permanent foto-exposities zijn te zien. De rijke geschiedenis en zeer diverse toepassingen van het medium bieden de mogelijkheid om aan het in fotografie geïnteresseerde publiek een breed scala van onderwerpen te tonen. De criteria die hierbij worden gehanteerd zijn de beeldkwaliteit en zeggings

#### **Eddy Posthuma de Boer en Juul Hondius**

Deze zomer besteedt het huis Marseille aandacht aan twee sociaal bewogen fotograferen die zich concentreren op het visualiseren van de problemen in de maatschappij door fotografie. Zij representeren niet alleen twee generaties, maar ook twee verschillende visies op geëngageerde fotografie. Zij bieden beiden op hun eigen manier stof tot discussie over de problemen in de derde wereld en de westerse kijk daarop.



GRONINGER MUSEUM

### **Groningen**

Als een schip voor anker ligt het Groninger Museum in het water. Het is een jaren tachtig labyrint waarin de grenzen tussen design, architectuur, kunst en de populaire media wordt doorbroken. Vier paviljoens zijn gegroepeerd rond een goudgele toren, het depot, de schatkamer van het museum. Ze hebben elk een eigen karakter dat de sfeer van de verschillende collecties weerspiegelt.

#### **VOC tentoonstelling**

In 2002 wordt herdacht dat 400 jaar geleden de VOC werd opgericht. De tentoonstelling laat objecten zien uit de periode van de VOC zoals: keramiek uit het scheepswrak, textiel, prenten en schilderijen.

## Museum: Proefpersoon 2

Jullie willen in het weekend samen een dagje uit, jullie willen graag naar een museum in Nederland. Hieronder staan een aantal voorbeelden van musea in Nederland, probeer hier een keuze uit te maken of verzin zelf een museum. Jullie gaan met de trein en willen redelijk op tijd daar zijn. Probeer samen een zo gunstig mogelijke reis te plannen waar jullie het beide mee eens zijn. Probeer ook alvast de terugreis te plannen.



**Amstelveen**

Het Cobra Museum voor Moderne Kunst Amstelveen werd in november 1995 geopend. Het ambitieuze museum beschikt over de omvangrijke en wereldberoemde Cobra collectie Van Stuijvenberg en streeft daarnaast naar een eigen verzameling kunstwerken van de internationale Cobra beweging.

### Expositie Herman Brood

De eerste oeuvre expositie van schilderijen van Herman Brood. De expositie startte op de dag van het overlijden van de kunstenaar precies een jaar geleden. De tentoonstelling is een primeur; niet eerder heeft een museum voor moderne kunst aandacht besteed aan het schilderkunstig oeuvre. Hiermee gaat een grote wens van Herman Brood in vervulling.

---

**RIJKS MUSEUM**  
a m s t e r d a m

**Amsterdam**

Het Rijksmuseum is het grootste museum van Nederland, zowel wat omvang van de collectie, oppervlakte, budget als het aantal werknemers betreft. Ieder jaar bezoeken meer dan een miljoen mensen het museum. Er werken ruim 400 mensen, waaronder 45 conservatoren met allerhande specialismen.

#### **Tentoonstelling: de Haas en de Maan**

Van het Japanse porselein is vooral het exportporselein bekend, dat met de VOC-schepen naar Nederland werd gebracht. Veel minder bekend is het verfijnde porselein met typisch Japanse motieven en vormen dat voor de Japanse markt werd gemaakt. In Oost-Azië wordt de haas stevast geassocieerd met de maan, en dit is dan ook een van de terugkerende motieven in de versieringen.

---

**BONNEFANTEN**  
**MUSEUM**  
MAASTRICHT

**Maastricht**

Het jonge (1995) en indrukwekkende gebouw van de Italiaanse architect Aldo Rossi is gezichtsbepalend voor de nieuwe en internationaal bekende Maastrichtse wijk Céramique. Aan de rechteroever van de Maas gelegen straalt het Bonnefantenmuseum met de opvallende koepeltoren zijn bezoekers tegemoet. Het museum heeft een bijzondere sfeer mede door het vele daglicht en het gebruik van natuurlijke materialen.

#### **Expositie: Philip Guston**

Het museum toont een dertiental schilderijen van de Amerikaanse schilder. Het gaat om werk uit de laatste periode van zijn leven. Deze presentatie van Guston is de tweede gewijd aan "klassieke moderne meesters" die zich tegen het mainstream gedachtegoed binnen de kunst gekeerd hebben. De werken van Philip Guston vormen daarin een exceptioneel gebaar, waarin het karikaturale tot het uiterste gedreven wordt. Ze trekken diepe sporen gevoed door zeer geprononceerde emoties.

---

---

## A.2 Translated scenarios (in English)

---

### **Instruction**

You can use the information kiosk in this room to obtain train table information. You can use the kiosk by speaking to it or by touching the screen. We are interested to find out how to design such a kiosk if multiple people make use of it. Therefore we would like to ask you to perform a role play and to give your opinion about different versions of the system.

The schedule for the experiment is on the next page. After you have had the opportunity to practice with the system we will ask you to plan a day out for the two of you. We will do this two times with different versions of the system. After each trip, a short interview and a questionnaire will follow.

For each scenario you both get tourist information about the trip you are going to make. We want you to discuss the trip. The information we give you differs per person. It may seem more efficient to discuss all details about the trip before starting, but for us it is most usefull if you discuss the trip during the interaction with the system.. Good luck!

### **Schedule**

---

**Instruction**

**Practice**

**Scenario 1**

**Questionnaire**

**Scenario 2**

**Questionnaire and Interview**

**Administration**

---

## Zoo: Participant I

Tomorrow you and your partner would like to go to a zoo in the Netherlands. Below you find information about a number of zoo's. You may pick one of these or think of another zoo to go to. You will go by train and you want to be back at a reasonable time. Try to plan an optimal trip that you both agree on. Try to plan the return trip as well.



Diergaarde Blijdorp is with more than 75 million visitors a year (2001) one of the most popular attractions in the Netherlands.

### Rotterdam

#### New Bizonprairie

At the end of April 2002 the new bison prairie was opened. This is the first area of "North-America", situated in the new area of Blijdorp (opposite to the new water world Oceanium). The new stay is separated from you with gratings. This way you are standing eye-to eye with the impressive cows.



### Emmen

In the *Noorder Dierenpark* you are able to travel around the world in a single day. The animals are situated in the continent they originate from. Your walk starts at the Biochron, an exiting museum telling the story about the origin of life. Then you discover Asia, Africa, America, Australia, and Europe.

#### New: Zoo-show about penguins

The children approach carefully to feel the sharpness of the top of the bill of the penguin. It is exiting to see whether you are bigger than emperor penguin. The new theater show of the Noorder Dierenpark in Emmen centers around the penguin. The stage looks like the rocky beach of Peru. At the beach there is a fisherman inviting the children to search for all kinds of objects on the beach.

---



## Harderwijk

Since its opening in 1965, *Dolfinarium Harderwijk* has grown into a unique theme park. The biggest sea animal park of Europa offers an impressive collection sea animals. Among others: dolphins, walruses, sea lions, seals, thornbacks, and sharks. The Dolfinarium is also renowned for its research about sea animals.

### Celebrations around the Laguna

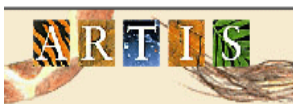
The Laguna, a unique natural living area for dolphins sea lions and many other sea animals in Dolfinarium Harderwijk, exists for five years. In honor of this jubilee the month of July is filled with celebrations for young and old. There are beautiful sand sculptures of sea mammals on the beach and there is a special jubilee presentation in the Laguna.

---

## Zoo: Participant 2

Tomorrow you and your partner would like to go to a zoo in the Netherlands. Below, you find information about a number of zoos in the Netherlands. You may pick one of those or chose a zoo you would like to go to yourself. You will go by train and want to arrive a reasonable time. Try to plan a trip to getter, to which you both agree, also plan the return trip because you want to be home before dark.

---



## Amsterdam

In *Artis*, the garden of life, you find al kinds of things about the origin of earth, the milky way, life on earth and the history of it all.

### Discover Artis in de palm of your hand

If you want to stroll around in Artis, from now on, you can use the electronic Artis guide. The Artis guide is developed by Yellow Brick Road Design – affiliated with Delft university of technology. The palm computer tells you where you are, shows the shortest route to every attraction, and it notifies you of the start of ‘sailing the stars’ in the planetarium or the special guided tour at the monkey rock.

---



**Apeldoorn**

There is no zoo in the world that allows you to see the animals like you see them in *Apenheul*. A lot of ape-species, from the huge gorilla to the smallest monkey in the world inhabit their own spacious forest areas. A large amount of monkeys climbs and clammers completely free between the visitors. And... did you know that between the monkeys you will find a lot of other animals in the Apenheul?

**Birth wave death's-head monkeys**

Recently the birth wave of the death's-head monkeys started. The animal care takers are still guessing about the number of babies that will be born this year but it could be as much as 20 to 30 babies within a few weeks. The babies are usually born at night and you can see them on the backs of their mothers the next morning, in the trees and between the visitors.



**Arnhem**

Burgers zoo tries to show the natural behavior of animals, if possible in relation to their natural environment. There are artificial eco-systems where the visitors can go on a discovery journey. In 1968 Burgers' opened the first safari park, followed by the chimpanzee territory, the wolves' forest, the Bush (an inside rainforest.), the Desert, and the Ocean.

**Opened at night**

During the summer, Burgers' Zoo stays open for visitors until 10 pm. From 7 pm on there are several guided tours, one can take a look in the central kitchen of the park, there are information stands and one can enjoy a puppet theatre or a story teller.

## **Museum: Participant I**

This weekend you would like to go out and visit a museum in the Netherlands. Below you find information about several museums. You can pick one of them or go to a museum you can come up with yourself. You travel by train, and you want to be there at a reasonable time. Try to plan a joint trip on which you both agree. Try also to plan the return trip.

---



**Tilburg**

De Pont is named after the lawyer and businessman mr. J.H. de Pont (1915-1987). In 1988 an organisation for stimulating contemporary art was started from his inheritance. De Pont is located in a former textile mill in Tilburg. Architecture agency Benthem Crouwel Architects has renovated this to a space where contemporary art can be shown to full advantage. The monumental old factory with the large light hall and the internal wool storages forms a beautiful environment for the many pieces of art. Since September 1992, De Pont is opened for public.

### **Exposition Anton Henning**

So far little of the work of the German artist Anton Henning has been on display in the Netherlands. Still, the international interest in his paintings and installations, is growing strongly. Henning is regarded as one of the young artists that approach the art of painting with new élan, after arrived masters like Richter, Polke en Baselitz.

---





**Amsterdam**

Since the opening of “huis Marseille” in September 1999, there is for the first time, a permanent place in Amsterdam for photography. The rich history and divers possibilities of the medium offer the possibility to show the public, interested in photography, a broad range of topics. The criteria that are used are the image quality and expressiveness.

**Eddy Posthuma de Boer and Juul Hondius**

This summer, “huis Marseille” has attention for two engaged photographers that concentrate on visualizing the problems of society through photography. They do not only represent two generations but also two different perspectives on engaged photography. Both offer material in their own way to discuss the problems in the third world and the western way of looking at that.



**Groningen**

Like a ship at anchor the Groninger Museum lies in the water. It is an eighties labyrinth where the borders between design, architecture, art and popular media disappear. Four pavilions are grouped around a gold yellow tower, the treasure room of the museum. Each has its own style reflecting the atmosphere of the different collections.

**VOC exposition**

In 2002 there is a remembrance about the start of the VOC, 400 years ago. The exhibition shows objects from the period of the VOC like ceramics from ship wrecks, textile, drawings and paintings.

---

## Museum: Participant 2

This weekend, you would like to go to a museum in the Netherlands together. Below you find examples of musea in the Netherlands. You can pick one of these or think of a different museum you would like to go to. You travel by train and want to be there at a reasonable time.. Try to plan a trip that you both agree on, also plan a return trip.



**Amstelveen**

The Cobra Museum for modern Art in Amstelveen was opened in November 1995. This ambitious museum has the large and world famous Cobra collection of Van Stuijvenberg and tried to collect its own collection of artworks of the international Cobra movement.

### **Exposition Herman Brood**

This is the first oeuvre exposition of paintings of Herman Brood. De exposition started on the day of the death of the artist exactly a year ago. Never before, a museum for modern art has exhibited paintings of Herman Brood. This way a big dream of Herman Brood has come true.



**Amsterdam**

The Rijksmuseum is the largest museum of the Netherlands, in terms of size of the collection, floor area, budget and number of employees. Each year the museum gets more than a million visitors. It has more than 400 employees, among which there are 45 conservators with all kinds of specialisms.

### **Exhibition: the Hare and the Moon**

Of the Japanese porcelain the export porcelain that came to the Netherlands in VOC-ships is most well known. Much less known is the delicate porcelain with typical Japanese motifs and forms that was made for the Japanese market. In east Asia the hare is often associated with the moon, and this is one of the recurring motifs in the decorations.

---

**BONNEFANTEN  
MUSEUM**  

---

**MAASTRICHT**  
**Maastricht**

The young (1995) and impressive building from the Italian architect Aldo Rossi marks the new and international well know Maastricht district Céramique. The Bonnefantenmuseum lies on the right bank of the Meuse and radiates towards the visitors with its striking dome tower. The museum has a special atmosphere, in part because of the large amount of daylight and the use of natural materials.

**Expositiion: Philip Guston**

The museum shows thirteen paintings from the American painter. It concerns work from the last period of his life. This presentation from Guston is the second exposition dedicated to ‘classical modern masters’ that have turned against the mainstream thought in the art world. The works of Philip Guston form an exceptional gesture where the caricature is drawn in extreme. They pull deep tracks fed by pronounced emotions.

---

## Appendix B: Utterance length in 'logarithmic time'

### B.1 Utterance Length in 'Log' space

Figure 1 shows a histogram (20 bins) of the log duration of utterances for the system and the partner.

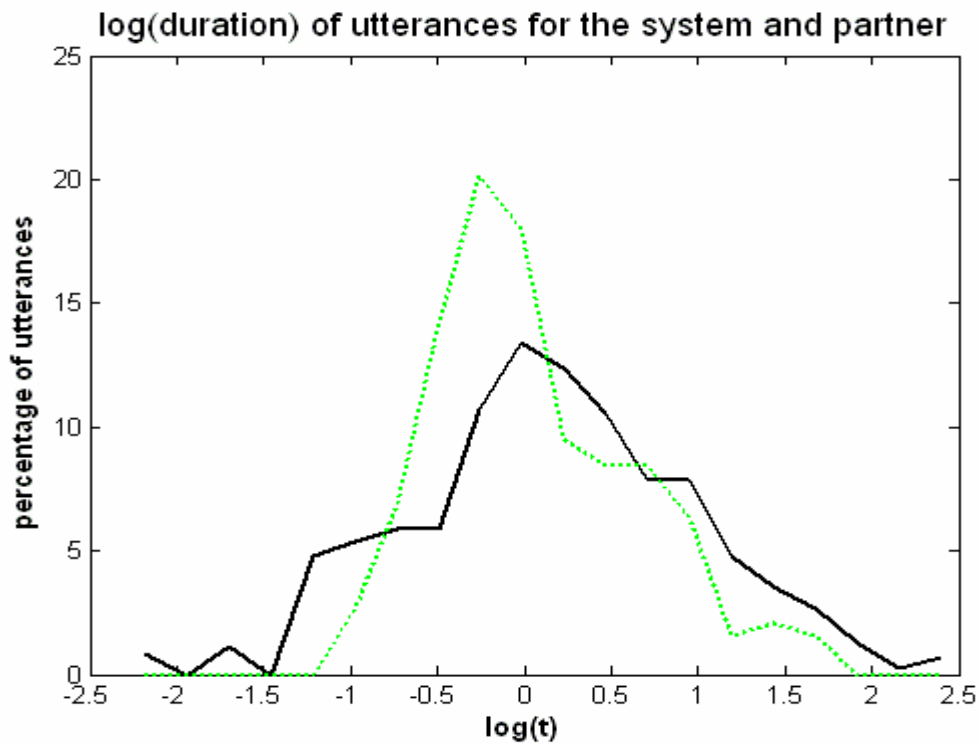


Figure 1: A histogram of utterances for the system (dotted line) and partner (line) when log duration is subtitled for the duration.

## Appendix C: The SASSI Questionnaire

### C.1 Original Items and Factors

In this section we list the statements of the SASSI questionnaire (Hone & Graham, 2000), Table 1). Followed by a short explanation of the six factors (Table 2). In the next section we present our Dutch translation (Table 3).

**Table 1: Items of the SASSI questionnaire. For each of the items, participants had to choose between: 'strongly agree', 'agree', 'neutral', 'disagree' and 'strongly disagree'. (continues on the next page)<sup>1</sup>**

|                          |   |  |
|--------------------------|---|--|
| System Response Accuracy | 1 | The system is accurate                           |
|                          | 2 | The system is unreliable                         |
|                          | 3 | The interaction with the system is unpredictable |
|                          | 4 | The system didn't always do what I wanted        |
|                          | 5 | The system didn't always do what I expected      |
|                          | 6 | The system is dependable                         |
|                          | 7 | The system makes few errors                      |
|                          | 8 | The interaction with the system is consistent    |
|                          | 9 | The interaction with the system is efficient     |

<sup>1</sup> The items were presented to participants in the following order: 24, 31, 8, 21, 25, 15, 27, 30, 26, 2, 29, 22, 16, 19, 10, 33, 7, 4, 28, 3, 11, 14, 32, 6, 12, 23, 9, 5, 18, 13, 34, 17, 1.

(continuation of) Table I

|              |    |   |
|--------------|----|---|
| Likeability  | 10 | The system is useful  |
|              | 11 | The system is pleasant  |
|              | 12 | The system is friendly  |
|              | 13 | I was able to recover easily from errors                                    |
|              | 14 | I enjoyed using the system  |
|              | 15 | It is clear how to speak to the system                                      |
|              | 16 | It is easy to learn to use this system                                      |
|              | 17 | I would use this system   |
|              | 18 | I felt in control of the interaction with the system                        |
| Cognitive    | 19 | I felt confident using the system   |
|              | 20 | I felt tense using the system   |
|              | 21 | I felt calm using the system  |
|              | 22 | A high level of concentration is required when using the system             |
|              | 23 | The system is easy to use   |
| Annoyance    | 24 | The interaction with the system is repetitive                               |
|              | 25 | The interaction with the system is boring                                   |
|              | 26 | The interaction with the system is irritating                               |
|              | 27 | The interaction with the system is frustrating                              |
|              | 28 | The system is too inflexible  |
| Habitability | 29 | I sometimes wondered if I was using the right word                          |
|              | 30 | I always knew what to say to the system                                     |
|              | 31 | I was not always sure what the system was doing                             |
|              | 32 | It is easy to lose track of where you are in an interaction with the system |
| Speed        | 33 | The interaction with the system is fast                                     |
|              | 34 | The system responds too slowly  |

Hones & Graham (2000) summarise the meaning of each of the six factors as follows:

**Table 2: A short description the six factors in the SASSI questionnaire**

| Factor                   | Description   |
|--------------------------|---|
| System Response Accuracy | The extent to which the users feel the system recognizes input correctly and does what is intended and expected.              |
| Likeability              | The extend to which participants felt working with the system was pleasant.   |
| Cognitive Demand         | The perceived level of effort and the feelings arising from this effort   |
| Annoyance                | The amount users find a system annoying, or disturbing the interaction  |
| Habitability             | Speech equivalent of visibility; there is a good match between the users conceptual model of the system and the actual system |
| Speed                    | The perceived speed of the system   |

## C.2 Dutch translation of the SASSI items

**Tabel 3: Nederlandse versie van de SASSI vragenlijst. Voor elk item werd gevraagd te kiezen tussen ‘helemaal eens’, ‘eens’, ‘neutraal’, ‘oneens’, en ‘helemaal oneens’. (vervolg op de volgende pagina)<sup>2</sup>**

|                          |   |   |
|--------------------------|---|---|
| System Response Accuracy | 1 | Het systeem is accuraat/nauwkeurig.                   |
|                          | 2 | Het systeem is onbetrouwbaar                          |
|                          | 3 | De interactie met het systeem verloopt onvoorspelbaar |
|                          | 4 | Het systeem deed niet altijd wat ik wilde             |
|                          | 5 | Het systeem deed niet altijd wat ik verwachtte        |
|                          | 6 | Het systeem is betrouwbaar                            |
|                          | 7 | Het systeem maakt weinig fouten                       |
|                          | 8 | De interactie met het systeem verloopt consequent     |
|                          | 9 | De interactie met het systeem verloopt efficiënt      |

<sup>2</sup> In de vragenlijsten voor de deelnemers stonden de items in de volgende (willekeurige) volgorde: : 24, 31, 8, 21, 25, 15, 27, 30, 26, 2, 29, 22, 16, 19, 10, 33, 7, 4, 28, 3, 11, 14, 32, 6, 12, 23, 9, 5,18, 13, 34, 17, 1.

(Vervolg van) Tabel 3

|                  |    |   |
|------------------|----|---|
| Likeability      | 10 | Het systeem is nuttig   |
|                  | 11 | Het systeem is plezierig  |
|                  | 12 | Het systeem is vriendelijk  |
|                  | 13 | Als er iets fout ging kon ik het gemakkelijk oplossen                     |
|                  | 14 | Ik vond het leuk om het systeem te gebruiken                              |
|                  | 15 | Het is duidelijk hoe ik moet spreken tegen het systeem                    |
|                  | 16 | Het is gemakkelijk om het systeem te leren gebruiken                      |
|                  | 17 | Ik zou het systeem gebruiken  |
|                  | 18 | Ik had het gevoel controle te hebben over de interactie met het systeem   |
| Cognitive Demand | 19 | Ik voelde me zeker van mezelf tijdens het gebruik van het systeem         |
|                  | 20 | Ik voelde me gespannen tijdens het gebruik van het systeem                |
|                  | 21 | Ik was kalm tijdens het gebruik van het systeem                           |
|                  | 22 | Ik moest me goed concentreren tijdens het gebruik van het systeem         |
|                  | 23 | Het systeem is gemakkelijk te gebruiken                                   |
| Annoyance        | 24 | Het systeem valt vaak in herhaling  |
|                  | 25 | De interactie met het systeem is saai                                     |
|                  | 26 | De interactie met het systeem is irritant                                 |
|                  | 27 | De interactie met het systeem is frustrerend                              |
|                  | 28 | De interactie met het systeem is niet flexibel genoeg                     |
| Habitability     | 29 | Ik vroeg me soms af of ik het juiste woord gebruikte                      |
|                  | 30 | Ik wist altijd wat ik tegen het systeem moest zeggen                      |
|                  | 31 | Ik wist niet altijd zeker waar het systeem mee bezig was                  |
|                  | 32 | Je raakt gemakkelijk de draad kwijt tijdens de interactie met het systeem |
| Speed            | 33 | De interactie met het systeem is snel                                     |
|                  | 34 | Het systeem reageert te langzaam  |



# Appendix D: Results of the Quantitative Study of Chapter 6

---

## D.1 Introduction

---

In this appendix we describe the result of an analysis of two types of quantitative data we collected in the experiments described in chapter 6. First, we collected subjective SASSI (Subjective Assessment of Speech System Interfaces, see Hone & Graham, 2000) scores for all participants for all interfaces and compared how participants assessed the transparency and back-channeling interfaces. Second, we extracted behavioral measures from the logs and compared the behavior of participants when interacting with back-channeling or transparency interfaces. Both analyses failed to show any differences. We report them here for the sake of completeness.

---

## D.1 SASSI

---

### D.1.1 Testing SASSI for reliability and validity

For all interfaces we have asked both members of all pairs to fill the SASSI questionnaire. The SASSI questionnaire aims at providing a valid, reliable, and sensitive measure of users' subjective experience with a wide range of speech recognition systems (Hone & Graham, 2000). The questionnaire contains 32 statements with 5-point likert-scales (strongly disagree; disagree; neutral; agree; strongly agree). It intends to provide quality measures for a speech interface for 6 factors: system response accuracy, likeability, cognitive demand, annoyance, habitability and speed (Appendix C list the items and a short description of the factors). In this appendix we report our efforts to translate the SASSI, to test its validity and reliability, and the results of our planned comparisons, most notably a comparison between the transparency and back-channeling interfaces.

We translated the SASSI questionnaire to Dutch with a 2 step procedure: first all items were translated to Dutch and subsequently translated back to English by independent usability specialists. For five items the back-translated version showed too little

correspondence and these items were discussed and decided on by two other usability specialists. The items were presented to users in randomized order.

After the experiment we tested the SASSI questionnaire for reliability and validity. We tested reliability by calculating Cronbach's alpha for each factor with our data. (Table 1)

**Table 1: Cronbach's alpha in our dataset**

| Factor                   | $\alpha$    |
|--------------------------|-------------|
| System Response Accuracy | 0.85        |
| Likeability              | 0.70        |
| Cognitive Demand         | 0.70        |
| Annoyance                | 0.65 (<0.7) |
| Habitability             | 0.54 (<0.7) |
| Speed                    | 0.75        |

A conventional threshold for accepting a factor as reliable is 0.7. Four factors match this criterion: System Response Accuracy, Likeability, Cognitive Demand and Speed. Two do not: Annoyance and Habitability. In interpreting results for these two factors we need to take this lack of reliability into account.

We tested validity by performing a confirmatory factor analysis. We used the same values as Hone & Graham (SPSS varimax rotation) and specified the number of factors to 6. Unfortunately the items loaded on completely different factors than the original ones (table 2, next page) and we concluded the results for all original SASSI factors should be interpreted with care, as they failed to show validity for our data. We decided to continue the analysis with a new exploratory factor analysis. This allowed us to see if there were different factors in our dataset that we could use for our comparisons. Any results of this analysis needed to be taken with care as well, because of limited item to subject ratio (1:3) in our study (see Costello & Osborne 2005). We used SPSS principle components with varimax rotation (table 2).

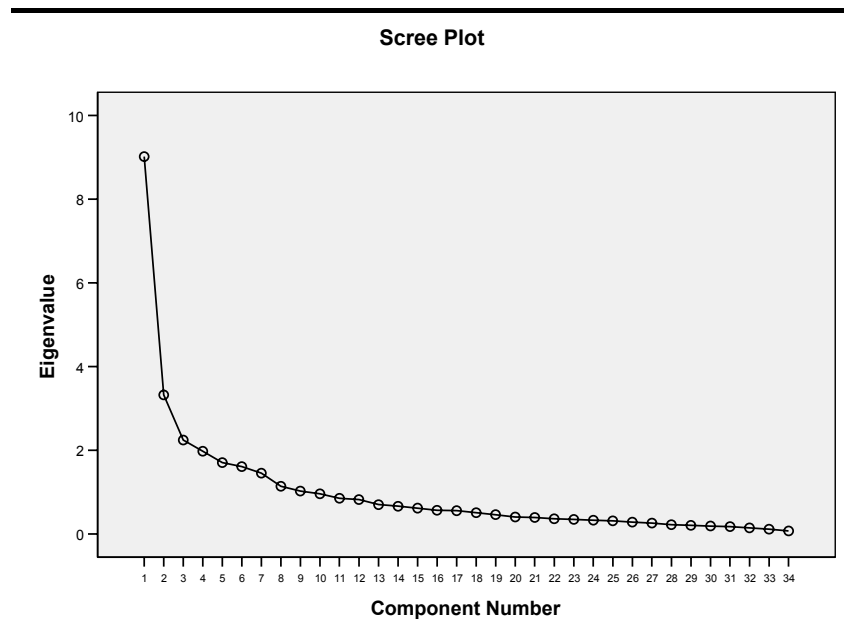
**Table 2: Items, and results for confirmatory and exploratory factor analysis (continues on the next page).**

| Factor                   | Item | Statement  | Confirmatory Factor Analysis |                    | Exploratory Factor Analysis |             |
|--------------------------|------|--|------------------------------|--------------------|-----------------------------|-------------|
|                          |      |  | Loads on                     | loading            | Loads on                    | Loading     |
| System Response Accuracy | 1    | The system is accurate                               | 1-3                          | 0.51- 0.45         | 1                           | <b>0.66</b> |
|                          | 2    | The system is unreliable                             | 1-3                          | 0.59 - 0.53        | 1                           | <b>0.59</b> |
|                          | 3    | The interaction with the system is unpredictable     | 1-4                          | 0.52 -0.42         | 1                           | <b>0.70</b> |
|                          | 4    | The system didn't always do what I wanted            | 1                            | 0.64               | 1                           | <b>0.49</b> |
|                          | 5    | The system didn't always do what I expected          | 1                            | 0.66               | 1                           | <b>0.63</b> |
|                          | 6    | The system is dependable                             | 1                            | 0.64               | 1                           | <b>0.67</b> |
|                          | 7    | The system makes few errors                          | 1                            | 0.77               | 1                           | <b>0.67</b> |
|                          | 8    | The interaction with the system is consistent        | 4                            | 0.45               | 1                           | <b>0.48</b> |
|                          | 9    | The interaction with the system is efficient         | 1-3-5                        | 0.53 - 0.46 - 0.45 | 1                           | <b>0.64</b> |
| Likeability              | 10   | The system in useful                                 | 3-5-6                        | 0.46 - 0.47 - 0.45 | 2                           | 0.64        |
|                          | 11   | The system is pleasant                               | 3                            | 0.67               | 2                           | 0.52        |
|                          | 12   | The system is friendly                               | 3                            | 0.65               | 1-2                         | 0.51-0.56   |
|                          | 13   | I was able to recover easily for errors              | 1-2                          | 0.49-0.50          | 1                           | <b>0.62</b> |
|                          | 14   | I enjoyed using the system                           | 3                            | 0.50               | 4                           | 0.40        |
|                          | 15   | It is clear how to speak to the system               | 4                            | 0.48               | 8                           | 0.53        |
|                          | 16   | It is easy to learn to use this system               | 2                            | 0.64               | 1                           | <b>0.62</b> |
|                          | 17   | I would use this system                              | 3                            | 0.64               | 1-2                         | 0.45        |
|                          | 18   | I felt in control of the interaction with the system | 2                            | 0.62               | 1                           | <b>0.48</b> |

(Continuation of ) table 2

| Factor           | Item | Statement   | Confirmatory Factor Analysis |         | Exploratory Factor Analysis |                |
|------------------|------|---|------------------------------|---------|-----------------------------|----------------|
|                  |      |   | Loadings                     | Loading | Loadings                    | Loading        |
| Cognitive Demand | 19   | I felt confident using the system   | 4                            | 0.74    | 1-2                         | 0.45<br>- 0.48 |
|                  | 20   | I felt tense using the system   | 2                            | 0.65    | 1                           | <b>0.48</b>    |
|                  | 21   | I felt calm using the system  | 4                            | 0.66    | 1-3                         | 0.73           |
|                  | 22   | A high level of concentration is required when using the system             | 6                            | 0.63    | -                           | 0.65           |
|                  | 23   | The system is easy to use   | 2                            | 0.75    | 1                           | <b>0.63</b>    |
| Annoyance        | 24   | The interaction with the system is repetitive                               | 1                            | 0.60    | 1-9                         | 0.48<br>- 0.42 |
|                  | 25   | The interaction with the system is boring                                   | 3                            | 0.64    | 2                           | 0.46           |
|                  | 26   | The interaction with the system is irritating                               | 1                            | 0.44    | 1                           | <b>0.46</b>    |
|                  | 27   | The interaction with the system is frustrating                              | 6                            | 0.41    | 1                           | <b>0.71</b>    |
|                  | 28   | The system is too inflexible  | 5                            | 0.51    | 1-5                         | 0.43<br>- 43   |
| Habitability     | 29   | I sometimes wondered if I was using the right word                          | 6                            | 0.65    | -                           |                |
|                  | 30   | I always knew what to say to the system                                     | 2                            | 0.58    | 1                           | <b>0.58</b>    |
|                  | 31   | I was not always sure what the system was doing                             | -                            |         | 1                           | <b>0.43</b>    |
|                  | 32   | It is easy to lose track of where you are in an interaction with the system | 2                            | 0.6     | 1                           | <b>0.70</b>    |
| Speed            | 33   | The interaction with the system is fast                                     | 5                            | 0.73    | 2                           | 0.44           |
|                  | 34   | The system responds to slowly   | 5                            | 0.77    | 2-4                         | 0.42           |

The new exploratory factor analysis revealed 9 factors with an eigenvalue higher than 1. Figure 1 shows the scree plot for the exploratory factor analysis. From factor 2 on, the inclination of the graph drops strongly indicating only factor 1 should be used.



**Figure 1: scree plot for the exploratory factor analysis, on the horizontal axis the components are listed vertically the Eigen values of these components**

As an additional factor we introduce ‘overall system quality’. We excluded double loading items resulting in the bold items in the table for this factor: 1, 2, 3, 4, 5, 6, 7, 8, 9, 13, 16, 18, 20, 23, 30, 31, 32. This factor consisted of 16 of the 32 items and had a cronbach’s alpha of 0.9.

In summary, the reliability and validity tests for the SASSI questionnaires suggested any result for comparisons with this questionnaire should be treated with care. Two original factors failed the reliability test, the confirmatory factor analysis raised doubts on the validity of the original factors, and while a new factor could be identified the contents of this factor can not be straightforwardly trusted because of the problem of data sparseness. In a way, these results suggest the SASSI questionnaire should be rejected as an instrument for measuring the type of differences in our dataset. But to be absolutely sure we decided to carry on with the tests as planned with the means of the original factors, the new 16 item factor and the grand mean of all questions.

### D.1.2 Preplanned and post experiment comparisons

Before the experiment we planned to do a within subjects ANOVA (SPSS repeated measures) for the difference between first and second scenario for which we expected no

result and a within subjects comparison for transparency versus back-channeling for which we expected there might be a difference. After the experiment we decided to complement this analysis with a comparison for the subset of interfaces where we used a camera for the back-channeling alternative, because in those sessions participants noted a difference between the two types of interfaces (intervention 4a, 5, 6). This last comparison may suffer from data-sparseness, but after the experiments it seemed unlikely that the SASSI-scores would differ if participants were unable to express differences between the systems. This may have obscured an effect in our original comparison.

Table 3 lists the results of the within subjects ANOVA for scenario, for the original factors, the new ‘overall system quality’ factor, and the grand mean of all SASSI questions. We expected to find no differences in this comparison.

**Table 3: within subjects ANOVA between scenarios (48 cases, df=1)**

| Factor                   | Scenario 1        |          | Scenario 2 |          | F    | p    |
|--------------------------|-------------------|----------|------------|----------|------|------|
|                          | Mean <sup>1</sup> | $\sigma$ | Mean       | $\sigma$ |      |      |
| System Response Accuracy | 2.7               | 0.6      | 2.3        | 0.6      | 0.03 | 0.87 |
| Likeability              | 2.3               | 0.4      | 2.2        | 0.4      | 0.13 | 0.72 |
| Cognitive Demand         | 2.3               | 0.5      | 2.3        | 0.6      | 0.26 | 0.61 |
| Annoyance                | 2.6               | 0.5      | 2.6        | 0.5      | 0.01 | 0.92 |
| Habitability             | 2.5               | 0.6      | 2.5        | 0.6      | 0.63 | 0.43 |
| Speed                    | 3.0               | 0.8      | 3.1        | 0.9      | 0.56 | 0.45 |
| Overall System Quality   | 2.5               | 0.5      | 2.5        | 0.5      | 0.11 | 0.75 |
| General Mean             | 2.5               | 0.4      | 2.5        | 0.4      | 0.01 | 0.90 |

As is visible in the table, in line with our expectation the comparison failed to show a difference for any of the tested factors.

---

<sup>1</sup> The SASSI scores can range from 1 and 5, the items were converted so high scores indicate high appreciation (in contrast to the questionnaires)

Table 4 lists the within subjects analysis comparing back-channelling with transparency across all interventions, for all chosen factors. We expected a difference between these two types of feedback but had no hypothesis about the direction of the difference.

**Table 4: Comparison of SASSI scores for Back-channelling and Transparency feedback (48 cases, df=1).**

| Factor                   | Back-channelling |          | Transparency |          | F    | P    |
|--------------------------|------------------|----------|--------------|----------|------|------|
|                          | Mean             | $\sigma$ | Mean         | $\sigma$ |      |      |
| System Response Accuracy | 2.6              | 0.6      | 2.7          | 0.6      | 0.51 | 0.48 |
| Likeability              | 2.3              | 0.4      | 2.2          | 0.4      | 0.07 | 0.79 |
| Cognitive Demand         | 2.3              | 0.5      | 2.3          | 0.6      | 0.38 | 0.54 |
| Annoyance                | 2.6              | 0.5      | 2.5          | 0.5      | 2.84 | 0.10 |
| Habitability             | 2.5              | 0.5      | 2.5          | 0.6      | 0.64 | 0.43 |
| Speed                    | 3.1              | 0.8      | 3.0          | 0.9      | 0.04 | 0.85 |
| Overall System Quality   | 2.4              | 0.5      | 2.5          | 0.5      | 0.24 | 0.62 |
| General Mean             | 2.5              | 0.4      | 2.5          | 0.4      | 0.04 | 0.85 |

As can be seen from the table, this analysis failed to show any significant differences. Since transparency scores higher on the factor annoyance than back channeling, a friendly interpretation could be that there is a trend for back-channeling interfaces to be more annoying than transparency interfaces. But the small difference in the means on this factor, in combination with the low reliability of the Annoyance factor ( $\alpha = 0.65$  see table 1), suggests a chance effect is more likely.

Table 5 lists the comparisons for those interventions where we used a rotating camera for back-channeling feedback.

**Table 5: Comparison of SASSI scores for Back-channelling and Transparency feedback in those cases there was a camera present (20 cases, df=1).**

| Factor                   | Back-channelling |          | Transparency |          | F     | P    |
|--------------------------|------------------|----------|--------------|----------|-------|------|
|                          | Mean             | $\sigma$ | Mean         | $\sigma$ |       |      |
| System Response Accuracy | 2.6              | 0.6      | 2.6          | 0.6      | 0.254 | 0.62 |
| Likeability              | 2.2              | 0.5      | 2.2          | 0.4      | 0.028 | 0.87 |
| Cognitive Demand         | 2.3              | 0.6      | 2.4          | 0.5      | 2.021 | 0.17 |
| Annoyance                | 2.7              | 0.5      | 2.5          | 0.5      | 7.630 | 0.12 |
| Habitability             | 2.5              | 0.5      | 2.7          | 0.6      | 5.429 | 0.31 |
| Speed                    | 3.0              | 0.9      | 3.1          | 0.9      | 0.388 | 0.54 |
| Overall System Quality   | 2.4              | 0.5      | 2.5          | 0.5      | 0.119 | 0.73 |
| General Mean             | 2.5              | 0.4      | 2.5          | 0.4      | 0.333 | 0.57 |

Again the analysis failed to show any differences.

### D.1.3 Conclusions SASSI

We conclude the SASSI questionnaire is not suitable for measuring the type of differences we hoped to find in this study. Both the reliability and validity analysis raised serious doubts about the questionnaire, and all comparisons with the questionnaire failed to show any differences. The problem may be that there were no differences in the subjective assessment of the concepts we tried to compare. This is in line with findings of the qualitative study reported in chapter 5. However in itself this does not explain the limited reliability and validity of the SASSI factors. So in addition to concluding there are no differences in the subjective assessment of these alternatives, we may conclude the SASSI questionnaire lacks the sensitivity to measure differences between variants of a single interface with a limited set of subjects.

---

## D.2 Behavioral Measures

---

At several places in chapter 6 we have touched on the possibility that participants can use the back-channeling feedback to adapt their behavior when things are going wrong (see section 5.2.2). The speaker and gaze following behavior of the rotating form, bear and camera, aimed, in part, at reminding speakers to look at their partners when they were addressing their partners. We have not found any evidence in the qualitative study that this worked. Still, users may have behaved differently in the transparency and back-channeling conditions without being aware of that. Therefore we extracted behavioral measures from the log to see if there were differences between these conditions.

We extracted a measure for speakers' gaze behavior. The logs do not contain information about the *intended* addressee of each utterance, so without a full transcription of the data we cannot produce the type of tables we have shown in chapter 3. Therefore we simply have extracted the percentage of speaking time speakers looked at their partner. Since many more utterances are intended for the partner speakers should look more at the partner if the back-channeling feedback encourages them to look at their addressee. We have used this measure in a within subjects ANOVA (SPSS repeated measures) for all interventions and only the interventions with a rotating camera. Table 6 lists the results.



**Table 6: A within subjects ANOVA on the % of speaking time, speaker's head orientation is interpreted as being oriented towards the partner by the AAD module in the back-channeling condition compared to the transparency condition.**

| % speaker gaze towards the system            | Back-channeling |          | Transparency |          | F     | P    |
|--|-----------------|----------|--------------|----------|-------|------|
|  | Mean            | $\Sigma$ | Mean         | $\Sigma$ |       |      |
| All 24 cases                                 | 38,7            | 24,7     | 31.9         | 22.0     | 3.653 | 0.07 |
| Only the 10 cases where a camera was present | 27,6            | 20,9     | 24,9         | 17,2     | 0.254 | 0.62 |

Both comparisons failed to show any significant differences. A friendly interpretation of the results on all 24 cases could be that there is a trend for speakers to look more at their partners in the back-channeling condition. But as this difference becomes smaller rather than larger in those interventions where a camera was present, and participants noticed the differences, we do not believe this is a robust effect.

---

## Bibliography

---

- Argyle, M., & Cook, M. (1976). *Gaze and Mutual Gaze*. Cambridge University Press.
- Argyle, M., & Graham, J. (1977). The central europe experiment – looking at persons and looking at things. *Journal of Environmental Psychology and Non-verbal behaviour*, 1, 6-16.
- Armani, L., Matassoni, M., Omologo, M., & Piergiorgio, S. (2003). Use of a CSP-based voice activity detector for distant-talking ASR. *Proceedings of Eurospeech 2003*, (pp. 501-504).
- Austin, J. L. (1962). *How to Do Things With Words?* Oxford: Oxford University Press.
- Bakx, I. H. M. (2002). *Use of information kiosks and facial orientation*. Unpublished manuscript, Eindhoven University of Technology, The Netherlands.
- Bakx, I. H. M., Van Turnhout, K. G., & Terken, J. M. B. (2003). Facial orientation during multi-party interaction with information kiosks. *Proceedings of Interact 2003*, (pp. 701–704). IOS Press.
- Bartneck, C., & Rauterberg, M. (in press). HCI reality –‘an unreal tournament’?. *International Journal of Human-Computer Studies*.
- Bellotti, V., Maribeth, B., Edwards, K. W., Grinther, R. E., Henderson, A., & Lopes, C. (2002). Making sense of sensing systems: Five questions for designers and researchers. *Proceedings of CHI 2002*, (pp. 416-422). ACM.
- Bennacef, S. K., Bonneau-Maynard, H., Gauvain, J. L., Lamel, L., & Minker, W. (1994). A spoken language system for information retrieval. *Proceedings of ICSLP*, 3, 1271-1274.
- Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the learning sciences*, 2, 141-178.
- Brumitt, B., & Cadiz, J. J. (2001). Let there be light! Comparing interfaces for homes of the future. *Proceedings of the 13<sup>th</sup> International Conference on Human-Computer Interaction, INTERACT 2001* (pp. 375-382). IOS Press

- Bunt, H. (2000). Dialogue pragmatics and context specification. In: H. Bunt & W. Black, (Eds.), *Abdunction, Believe and Context in dialogue; studies in computational pragmatics* (pp. 81-150). John Benjamins, Amsterdam.
- Cassell, J., Bickmore, T., Billingham, M., Campbell, L., Chang, K., Vilhjálmsón, H., & Yan, H. (1999a). Embodiment in conversational interfaces: REA. *Proceedings of the CHI 1999 Conference*, (pp. 520-527). ACM.
- Cassell, J., Torres, O., & Prevost, S. (1999b). Turn taking versus discourse structure: How best to model multimodal conversation. In Y. Wilks (Ed.), *Machine Conversations* (pp. 143-154). Boston: Kluwer.
- Clark, H. H. (1996): *Using Language*. Cambridge University Press.
- Clayton, D., & Moock, C. (2001). *JAVA relay server and example flash applications*. <http://moock.org/unity/>
- Purao, S. Cole, R. Rossi, M. Sein, M. 2005. Being Proactive: Where Action Research meets Design Research. International Conference on Information Systems. (ICIS) Las Vegas, NV, December 11-14, Download: [here](#) .
- Cole, R., Purao, S., Rossi, M., & Sein, M. K. (2005). Being proactive: where action research meets design r esearch. *Proceedings of the International Conference on Information Systems, ICIS 2005*, (pp. 11-14), Association for Information Systems.
- Costello, A. B., & Osborne J. W. (2005). Best practices in exploratory factor analysis: Four. recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10.
- Danninger, M., Flaherty, G., Bernardin, K., Ekenel, H. K., Kóhler, T., & Malkin, R. et al. (2005). The connector-facilitating context-aware communication. In G. Lazzari, F. Pianesi, J. L. Crowley, K. Mase & S. L. Oviatt (Eds.), *Proceedings of the 7<sup>th</sup> International Conference on Multimodal Interfaces*, (pp. 69-75). ACM.
- Darrell, T., Fisher, J. W., & Viola, P. (2000). Audio-visual Segmentation and “The Cocktail Party Effect”. *Proceedings of the 3<sup>rd</sup> International Conference on Multimodal Interfaces*, (pp. 32–40). Springer.
- Davis, K. H., Biddulph, R., & Balashek, S. (1952). Automatic recognition of spoken digits. *The Journal of the Acoustic Society of America*, 24, 637-642.
- De Ruiter, J. P., Mittere, N., & Enfield, J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82, 515-535.

- Den Os, E., & Boves, L. (2005). User behavior in multimodal interaction. *Proceedings of HCI 2005*. Lawrence Erlbaum Associates
- Den Os, E., Boves, L., Rossignol, S., Ten Bosch, L., & Vuurpijl, L. (2005). Conversational agent or direct manipulation in human–system interaction. *Speech Communication, 47*, 194–207.
- Dorst, K. (1997). *Describing design – A comparison of paradigms*. Unpublished doctoral dissertation, Delft university of technology, The Netherlands.
- Dourish, P. (2001). *Where the action is – the foundations of embodied interaction*. Boston: MIT press.
- Dreyfus-Graf, J. (1949). Sonograph and sound mechanics. *The Journal of the Acoustic Society of America, 22*, 731-739.
- Duda, R. O., Hart, P. E. & Stork, D.G. (2001). *Pattern Classification* (2<sup>nd</sup> edition.). John Wiley and Sons Inc.
- Fawcett, T. (2004). *ROC graphs: Notes and practical considerations for researchers*. <http://www.cs.iastate.edu/~honavar/ROC101.pdf>.
- Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effect of speakers' assumptions about what others know. *Journal of Personality and Social Psychology, 62*, 378-391.
- Fussell, S. R., Setlock, L. D., & Parker, E. M. (2003). Where do helpers look? Gaze targets during collaborative physical tasks. *CHI 2003: Extended abstracts (pp. 768-769)*. NY: ACM Press.
- Goodwin, C. (1981). *Conversational organisation, interaction between speakers and hearers*. South Carolina: Academic Press.
- Grice, H. P. (1957). Meaning. *Philosophical Review, 66*, 377-88.
- Grudin, J. (1994). Eight challenges for developers. *Communications of the ACM, 37*, 93-104.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS quarterly, 28*, 75-105.
- Holtgraves, T. M. (2002). *Language as social action: Social psychology and language use*. Lawrence Erlbaum Associates Inc.
- Hone, K. S., & Graham, R. (2000). *Towards a tool for the subjective assessment of speech system interfaces in natural language engineering*. NY: Cambridge University Press.

- Horvitz, E., & Apacible, J. (2003). Learning and reasoning about interruptions. *Proceedings of the 5<sup>th</sup> International Conference on Multimodal Interfaces* (pp. 20-27). ACM.
- Horprasert, T., Yacoob, Y. & Davis L. S. (1996.). Computing 3-D head orientation from a monocular image sequence. *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*. IEEE Computer Society.
- Hugunin, J., & Zue, V. (1997). On the design of effective speech-based interfaces for desktop applications. *Proceedings of EuroSpeech 1997*, (pp. 1335-1338).
- Hutchby, I., & Wooffitt, R. (1998). *Conversation Analysis*. Polity Press.
- Jovanović, N. (2007). *To whom it may concern: Addressee identification in face-to-face meetings*. Unpublished doctoral dissertation, University of Twente, The Netherlands.
- Jovanović, N., & Op den Akker, R. (2004). Towards automatic addressee identification in multy-party dialogues. *Proceedings of the 5th SIG Dial Workshop on Discourse and Dialogue* (pp. 89-92). The Association for Computational Linguistics
- Jovanović, N., Op den Akker, R., & Nijholt, A. (2006). Addressee identification in face-to-face meetings. *Proceedings of 11th conference of the european chapter of the ACL* (pp. 169-176). The Association for Computer Linguistics
- Katzenmaier, M., Stiefelhagen, R., & Schultz, T. (2004). Identifying the addressee in human-human-robot interactions based on head pose and speech. In R. Sharma, T. Darrell, M. P. Harper, G. Lazzari & M. Turk (Eds.), *Proceedings of the International Conference on Multimodal Interfaces ICMI 2004, State College, PA, USA*, (pp. 144-151).
- Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, 26, 22-63.
- Kendon, A. (2004). *Gesture, visible action as utterance*. Cambridge University Press.
- Keysar, B. (1997). Unconfounding common ground.. *Discourse Processes*. 24, 253-270.
- Knapp, M. L., & Hall, J. A. (2002). *Non-verbal communication in human interaction*. Thomson Learning Inc.
- Kubric, S. (Director). (1968). *2001: A space odyssey*. [Motion Picture]. United States: Metro-Goldwyn-Mayer Pictures Inc.
- Kulyk, O., Wang, C., & Terken, J. M. B. (2006). Real-time feedback on nonverbal behaviour to enhance social dynamics in small group meetings. In S. Renals & S. Bengio (Eds.), *Proceedings of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms 2005*, (pp. 150-161). Springer-Verlag.

- Life, A., Salter, I., Temem, J. N., Bernard, F., Rosset, S., Bennacef, S., et al. (1996). Data collection for the MASK kiosk: WOz versus prototype system. *Proceedings of The Fourth International Conference on Spoken Language Processing, 1996*, (pp. 1672-1675). IEEE.
- Lunsford, R., Oviatt, S. L., & Arthur, A. M. (2006). Toward open-microphone engagement for multiparty interactions. In F. K. Quek, J. Yang, D. W. Massaro, A. A. Alwan & T. J. Hazen (Eds.), *Proceedings of the 8th International Conference on Multimodal Interfaces*, (pp. 273-280). ACM press.
- Lunsford, R., Oviatt, S. L., & Coulston, R. (2005). Audio-visual cues distinguishing self-from system-directed speech in younger and older adults. In G. Lazzari, F. Pianesi, J. L. Crowley, K. Mase, & S.L. Oviatt (Eds.), *Proceedings of the 7th International Conference on Multimodal Interfaces* (pp. 265-272). ACM press.
- Macho, D., Padrell, J., Abad, A., Nadeu, C., Hernando, J., McDonough, J., Wolfel, M., Klee, U., Omologo, M., Brutti, A., Svaizer, P., Potamianos, G., & Chu, S. M. (2005). Automatic speech activity detection: Source localization, and speech recognition on the CHIL seminar corpus. *Proceedings of ICME 2005*, (pp 876- 879). IEEE.
- Macskassy, S. A., & Provost, F. (2004). Confidence bands for ROC curves: Methods and an empirical study. *Proceedings of the 1st Workshop ROC Analysis in AI* (pp. 61-70). Berkeley: University of California.
- Maes, P., & Shneiderman, B. (1997). Direct manipulation vs. interface agents: A debate. *Interactions*, 4, 42-61.
- Maglio, P. P., Matlock, T., Campbell, C. S., Zhai, S., & Smith, B. A. (2000). Gaze and speech in attentive user interfaces. In T. Tan, Y. Shi, & W. Gao (Eds.), *Proceedings of the 3rd International Conference on Multimodal Interfaces*, (pp. 1-7). Springer
- Martinovski, B., & Traum, D. (2003). Breakdown in human-machine interaction: The error is the clue. *Proceedings of the ISCA tutorial and research workshop on Error handling in dialogue systems* (pp. 11–16).
- Michie, S., & Abraham, C. (2004). Interventions to change health behaviours: Evidence based or evidence inspired? *Psychology and Health*, 19, 29-49.
- NRC (2007), W. Oosterbaan & L Starkink (Eds), *M – het wonder van het web (internet special)*, Januari 2007. PCM uitgevers, Rotterdam
- OpenCV (2005): <http://sourceforge.net/projects/opencvlibrary/> .
- Opperman, D., Schiel, F., Steininger, S., & Beringer, N. (2001). Off-talk: A problem for human-machine-interaction? *Proceedings of Eurospeech* (pp. 2197-2200).

- Oviatt, S. L. (1999). Ten myths of multimodal interaction. *Communications of the ACM*, 42, 74-81.
- Oviatt, S. L. & Cohen, P. R. (2000). Multimodal systems that process what comes naturally. *Communications of the ACM*, 43, 45-53.
- Oviatt, S. L. (2000). Multimodal system processing in mobile environments. *Proceedings of the Thirteenth Annual ACM Symposium on User Interface Software Technology (UIST'2000)*, (pp 21-30.). ACM Press.
- Purao, S. (2002). *Design research in technology of information systems: Truth or dare*. School of Information Services and Technology, The Pennsylvania State University.
- Raducanu, B., Subramanian, S., Markopoulos, P. (2004). Human presence detection by smart devices. *Proceedings of the 4th International ICSC Symposium on Engineering of Intelligent Systems*. ICSC Academic Press.
- Rauterberg, G. W. M. (2000). How to characterize a research line for user-system interaction. *IPO annual progress report*, 35, 69-83.
- Rauterberg, G. W. M. (2006). HCI as engineering discipline: To be or not to be? *African Journal of Information and Communication Technology*, 2, 163-184.
- Reigh, J. M., Loughlin, M., & Waters, K. (1997). Vision for a Smart kiosk. *Computer vision and pattern recognition*, 2 , 690-696.
- Rice, D. (2003). Cannonball. On O [CD]. Los Angeles: 14<sup>th</sup> Floor.
- Rutter, D. R. (1984). *Looking and Seeing: The Role of Visual Communication in Social Interaction*. John Wiley & Sons Ltd.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Journal of the linguistic society of America*, 50, 696-735.
- Schegloff, E. A. (1968). Sequencing in conversational openings. *American Anthropologist*, 70, 1075-1095.
- Searle, J. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University press.
- Searle, J. (1979). *Expression and Meaning*. Cambridge: Cambridge University Press.
- Seeman, E., Nickel, K., Stiefelhagen, R. (2004). Head pose estimation using stereo vision for human-robot interaction. *Proceedings of 6th International Conference on Face and Gesture Recognition* (pp. 626- 631). IEEE.

- Shneiderman, B., & Maes, P. (1997). Direct manipulation versus interface agents. *Interactions*, 4, 42-61.
- Siepmann, R., Batliner, A., & Opperman, D. (2001). Using prosodic features to characterize Off-talk in human-computer interaction. *Proceedings of the Workshop on Prosody and Speech Recognition* (pp. 147-150).
- Stiefelhagen, R. (2002). *Tracking and modelling focus of attention in meetings*. Unpublished doctoral dissertation. Universität Karlsruhe, Germany.
- Stiefelhagen, R., Yang, J., & Waibel. (2000). A simultaneous tracking of head poses in a panoramic view. *Proceedings of International Conference on Pattern Recognition* (pp 3726 ). IEEE
- Stiefelhagen, R., & Zhu, J. (2002). Head orientation and gaze direction in meetings. In L. Terveen (Ed.), *Proceedings of the CHI 2002 Conference on Human Factors in Computing Systems*, (pp. 858-859). ACM
- Sturm, J. (2005). *On the usability of multimodal interaction for mobile access to information services*. Unpublished doctoral dissertation. Radboud Universiteit Nijmegen, The Netherlands.
- Sturm, J., Bakx, I., Cranen, B., & Terken, J. (2002). The effect of prolonged use on multimodal interaction. *Proceedings of the ISCA Workshop on Multi-modal Dialogue in Mobile Environments*,
- Sturm, J., Iqbal, R. & Terken, J (2006): A Wizard of Oz Approach towards Designing Aware Environments. *Unpublished Manuscript*
- Terken, J. M. B, Joris, I., De Valk, L. (in press). Multimodal cues for addressee-hood in triadic communication with a human information retrieval agent. *Proceedings of the 9th International Conference on Multimodal Interfaces*. ACM
- Thorisson, K. (1996). Communicative humanoids: A computational model of psychological dialogue skills. Unpublished doctoral dissertation. MIT, USA.
- Traum, D. (2004). Issues in multiparty dialogs. In F. Dignum (Ed.), *Advances in agent communication*. (pp. 201-211). Springer-Verlag.
- Vaishnavi, V., & Kuechler, W. (2006). Design research in information systems. <http://www.isworld.org/Researchdesign/drisISworld.htm>. January 20, 2004.
- Van Gelder, J., Van Peer, I., & Aliakseyeu, D. (2005). Transcription table: Text support during meetings. In M.F. Costabile & Fabio Paternò (Eds.), *Proceedings of INTERACT 2005*, (pp. 1002-1005). Springer..



- 
- Van Turnhout, K. G., Malchanau, A., Disaro, R. M., & Markopoulos, P. (2002). The idea-collector: A device for supporting creative face-to-face meetings. *Proceedings of Human Computer Interaction, vol. 2.* (pp. 74-78)
- Van Turnhout, K.G., Terken J.M.B., Bakx, I, Eggen J.H., (2005). Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. *Proceedings of the 7th international conference on Multimodal interfaces.* (pp 175 – 182). ACM
- Vertegaal, R., Slagter, R., Van der Veer, G., & Nijholt, A. (2001). Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. *Proceedings of the SIGCHI conference on Human factors in computing systems,* (pp. 301-308). ACM
- Voit, M., Nickel, K., & Stiefelhagen, R. (2005). Multi-view head pose estimation using neural networks. *Proceedings of the 2nd Workshop on Face Processing in Video,* (pp. 347-352)
- Von Cranach, M. (1971). The role of orienting behavior in human interaction. In H. Aristide & AH Esser (Eds.), *Behavior and Environment, the use of space by animals and men,* (pp. 217-237). New York: Plenum Press.

## (Short) Curriculum Vitae

---

---

|            |   |
|------------|---|
| 1988-1993  | HAVO<br>Paulus Lyceum, Tilburg  |
| 1993-1997  | Teachers training Physics (2nd Degree)<br>Hogeschool Katholieke Leergangen Tilburg, Tilburg         |
| 1997-2000* | Teachers training Physics (1st Degree, equivalent to MA)<br>Fonty's Educatie, Tilburg               |
| 2000-2002& | User System Interaction<br>Stan Ackermans Instituut<br>Technische Universiteit Eindhoven, Eindhoven |
| 2002-2007& | PhD Candidate<br>Technische Universiteit Eindhoven (Industrial Design)                              |

\* Within this period Koen van Turnhout worked part-time as physics teacher at several high schools

& Within this period Koen van Turnhout was also active, in several roles, within the voluntary organisation Anderwijs



# Socially Aware Conversational Agents

---

## A summary in English

---

This thesis focuses on shared use of speech-centric multimodal interfaces. The difficulty is this: if there are multiple users, interacting with each other and with a multimodal interface, they will use speech both to communicate with each other and with the system. A solution to this problem would be to build systems that have a sense of the ongoing social context. In particular these systems would have to know who is talking to whom, and they would have to have a way to use this information in the interaction with multiple users. In this thesis, we suggest that it is already possible to build machines that are able to do this and we call them socially aware conversational agents.

We intend to contribute to the development of socially aware conversational agents by designing an example of such an agent: a multi-modal information kiosk specifically intended to be used by pairs, rather than individual people. This has two goals. First, we aim to explore the (technological) possibilities for making conversational agents socially aware. Second, we aim at collecting the pieces of specialized knowledge needed to develop this type of solution and to uncover interdisciplinary challenges for the development of socially aware conversational agents. For the second purpose we take an integrative, interdisciplinary approach: we work at this design case from a social psychological, a systems engineering and an interaction design perspective, respectively. This enables us to deliver a balanced design case and it allows us to evaluate to how the research questions about socially aware conversations are connected and intertwined.

In chapters 2 and 3 we adopt a social psychological perspective. We argue that looking at language as a form of coordinated action (Clark, 1996) can guide the design of conversational agents. This theory suggests there are three strategies that may help sensing who is talking to whom. First, we can make use of the way people sequence communicative acts – of what we know about the dialog history. Second, we can sense behaviors that humans use to coordinate within communicative acts - in particular their eye gaze. Third, we can make use of differences in speaking styles between human – system and human – human conversation. We operationalize these three strategies into tactics that apply to our specific design case: dialog events of the system, head orientation,

and utterance length, could be indicators for addressee-hood in our case. Using a dataset collected with a Wizard of Oz setup, we show that this is indeed the case: of these three tactics using head orientation delivers both the most information and is probably most easily applied in other design cases.

In chapters 4 and 5 we adopt a systems engineering perspective. Here, we specify features for the 3 forenamed tactics and combine them into a single statistical model. We nickname this naïve Bayes classifier AAD – short for Automatic Addressee Determination. AAD is able to make a reasonable estimate about to what extent it is likely that users are addressing the system. In addition we show that AAD can be used to provide early estimates of addressee-hood. Not only can we estimate whether an utterance was intended for the system or not, *after* the utterance, AAD is also, be it to a lesser extent, suited to predict whether the upcoming utterance may be directed to the system and whether the ongoing, unfinished, utterance is intended for the system. These early estimates open up possibilities for interaction designers, as humans can be alerted earlier when things may be going wrong. In chapter 5 we provide a description of a prototype of a socially aware conversational agent. We created a real time version of AAD that is able to deal with simultaneous speech and mid-utterance silences, and we integrated this version into a partial wizard of Oz setup. We used this prototype in the last study of this chapter.

In chapter 6 we adopt the perspective of interaction design. We try to uncover how socially aware conversational agents should provide users with feedback about their status. We focus on three intertwined design questions: on what aspects of their behavior should socially aware conversational agents provide feedback, when should they deliver this feedback and with what metaphors. To provide answers to these questions we present a qualitative, iterative, question oriented, design intervention study. Within six design interventions we ask participants to interact with different versions of our prototype. Users are subsequently probed about their experience with this system and their expectations of such systems. The study shows that naïve users confronted with speech technology are not aware of the problem of addressing, and assume the system solves this automatically. Feedback in combination with excessive system errors makes users aware of the ‘problem,’ but they do not succeed to make practical use of the feedback in the interaction. These results indicate that when the amount of system errors is low, feedback on the status of a conversational agent should be used with restraint. However, when the amount of system errors is high, there are also large challenges for interaction design. We suggest that new research lines should focus on the way humans and systems can deliver effective backchannel responses.

In chapter 7 we provide a general discussion, and identify new interdisciplinary challenges for the development of socially aware conversational agents. In moving from a social psychological, to a systems engineering, to an interaction design perspective we

encountered limitations of these bodies of knowledge related to their applicability for the other domains. Clarks theory of language as coordinated action is suitable to identify behaviors that indicate who is talking to whom, but from a systems engineering's perspective there is a need to be able to specify its constructs in technological terms. - in particular we need to come up with boundary conditions for communicative acts. Knowledge about pattern classification can be successfully applied to the problem of addressing, but from an interaction design perspective there is a need for early results. Following this, we need to develop classifiers that can deliver fast result or even predict future events. For interaction design there is a need for concrete examples of successful interfaces, and for models that provide situational knowledge about human communication behaviors and needs, that can take technological constraints into account. The successful development of socially aware conversational agents will depend largely on the extent that these interdisciplinary challenges can be tackled.

# Socially Aware Conversational Agents

---

## Een samenvatting in het Nederlands

---

In dit proefschrift richten we ons op gedeeld gebruik van spraakgestuurde systemen. De moeilijkheid bij gedeeld gebruik van deze systemen is dat gebruikers met zowel het system als met elkaar kunnen spreken. Eén mogelijke oplossing is om systemen te maken die een idee hebben van de sociale context en hun dialoog met gebruikers daaraan aan kunnen passen. We denken dat de technologie beschikbaar is om deze systemen te ontwikkelen en we noemen deze systemen *socially aware conversational agents* (letterlijk vertaald: sociaal bewuste spraaksystemen), of SACA's.

Als bijdrage aan het ontwikkelen van SACA's beschrijven we het ontwerp van een voorbeeld van zo'n system: een informatiekiosk special ontworpen voor het gebruik door paren in plaats van individuen. Dit ontwerp dient twee doelen. Ten eerste willen we de (technologische) mogelijkheden om spraaksystemen 'sociaal bewust' te maken verkennen. Verder zijn er om SACA's te kunnen ontwikkelen verschillende disciplines nodig: sociaal psychologen, systeem ingenieurs, en interactie ontwerpers zullen daarvoor moeten samenwerken. Het tweede doel is dan ook om de kennis uit de verschillende disciplines die nodig is bij elkaar te brengen en juist de uitdagingen die tussen de grenzen van de traditionele disciplines vallen bloot te leggen. Daarom bekijken we in dit interdisciplinaire proefschrift de ontwerp casus vanuit de drie genoemde disciplines.

In hoofdstuk 2 en 3 bekijken we de casus vanuit een sociaal psychologisch perspectief. We beargumenteren dat een theorie die taalgebruik benadert als een vorm van gecoördineerde actie (Clark, 1996) richting kan geven aan het ontwerp van SACA's. De theorie van Clark suggereert dat er drie strategieën zijn die SACA's zouden kunnen gebruiken om vast te stellen of een uiting voor hen bedoeld is of niet. Ten eerste kunnen systemen gebruik maken van de manier waarop mensen uitingen op elkaar laten volgen (dialooghistorie). Ten tweede kunnen ze letten op (non-verbale) signalen die mensen tijdens uitingen gebruiken om te zorgen dat de communicatie vlot verloopt. In het bijzonder kunnen ze bijhouden waar mensen naar kijken tijdens uitingen. Ten derde kunnen ze gebruik maken van het feit dat mensen anders spreken tegen systemen dan tegen elkaar. In hoofdstuk 3 proberen we deze strategieën operationeel te maken voor het

systeem dat we willen ontwerpen. Hierdoor komen we tot drie tactieken die ons systeem kan gebruiken. Met een ‘wizard of Oz’<sup>1</sup> studie laten we zien dat die tactieken inderdaad werken.

In hoofdstuk 4 en 5 kijken we naar de casus door de ogen van een systeem ingenieur. In hoofdstuk 4 stellen we, gebaseerd op de resultaten van hoofdstuk 3, een statistisch model op dat kan inschatten of een gebruiker tegen het systeem spreekt of niet. We noemen deze patroonherkenner ‘AAD’ –een afkorting voor Automatic Addressee Determination (letterlijk vertaald: automatische geadresseerde bepaling). AAD kan een inschatting maken of een gebruiker het tegen het systeem heeft of niet. Bovendien kan AAD gebruikt worden om al op een vroeg moment te aan te geven of een uiting voor het systeem bedoeld is. Het model kan niet alleen *na* een uiting zeggen of deze voor het systeem bedoeld was, maar ook, zij het minder betrouwbaar, *tijdens* of zelfs *voor* dat een gebruiker daadwerkelijk begint te spreken. Deze vroege inschattingen zijn belangrijk omdat interactie ontwerpers dan de mogelijkheid hebben gebruikers al vroeg te kunnen informeren of alles nog goed gaat. In hoofdstuk 5 beschrijven we een prototype van een SACA. De patroonherkenner, AAD, is hier geïntegreerd in een systeem waar gebruikers treininformatie mee kunnen opvragen. Dit systeem kan ook omgaan met stiltes tijdens een uiting en gelijktijdige spraak. Voor sommige delen van het systeem (spraakherkenning en dialoogmanagement) gebruiken we nog altijd een ‘wizard of Oz’ oplossing.

In hoofdstuk 6 bekijken we de casus door de ogen van een interactieontwerper. We proberen er achter te komen hoe SACA’s gebruikers feedback zouden moeten geven over de inschattingen die het systeem maakt. Drie samenhangende ontwerp vragen spelen daarbij een rol: over welke aspecten van hun gedrag moeten SACA’s feedback geven, wanneer moeten zij deze feedback geven, en met welke metaforen moeten zij dat doen. Om antwoorden te kunnen geven op deze vragen presenteren we een kwalitatieve, iteratieve, vraag-georiënteerde, ontwerpinterventie studie. Gedurende zes ontwerpinterventies vragen we gebruikers met twee verschillende versies van ons prototype te werken. Elke versie geeft andere feedback. Daarna ondervragen we deelnemers over hun ervaring met en hun verwachtingen over dit soort systemen. De studie toont aan dat gebruikers zich niet bewust zijn van het probleem, - dat het systeem er achter moet komen wie tegen wie spreekt - en dat ze aannemen dat het systeem dit op weet te lossen. Feedback kan mensen bewust maken van het probleem, maar gebruikers weten vervolgens niet hoe ze die feedback moeten gebruiken. Daarom lijkt het beter om, wanneer het aantal systeemfouten laag is, helemaal geen feedback te geven. Wanneer het aantal systeemfouten hoog is zouden er een nieuwe ontwerp studies moeten komen.

---

<sup>1</sup> In een ‘wizard of Oz’ studie, worden gebruikers gevraagd met een system te werken dat niet echt bestaat. In werkelijkheid zit er achter de schermen een ‘tovenaars van Oz’ die alle systeemacties aanstuurt.



In hoofdstuk 7 richten we ons op een algemene discussie en het identificeren van nieuwe interdisciplinaire uitdagingen voor het ontwikkelen van SACA's. Bij de verandering van een sociaal psychologisch naar een systeem ingenieurs' perspectief en vervolgens naar een interactie ontwerp perspectief zijn we beperkingen van die drie typen kennis tegengekomen voor het toepassen in een ander gebied. Clarks theorie van taalgebruik als gecoördineerde actie is geschikt om gedrag te identificeren dat samenhangt met de vraag wie tegen wie spreekt, maar door de ogen van een systeem ingenieur zouden de begrippen van de theorie een technologische grondslag moeten krijgen. In het bijzonder is het van belang om randvoorwaarden op te stellen voor het begin en einde van een uiting. Kennis over patroonherkenning kan goed worden toegepast op het probleem van het vaststellen wie tegen wie spreekt, maar vanuit het perspectief van een interactie ontwerper is er een behoefte aan vroege resultaten. Daarom moeten we patroonherkenners bouwen die snelle resultaten opleveren of zelfs toekomstige gebeurtenissen kunnen voorspellen. Interactieontwerp voor SACA's is het minst ver ontwikkeld. Hier is behoefte aan voorbeelden van goede feedback en aan modellen over communicatie die minder abstract zijn dan die van Clark (1996) en die technische beperkingen kunnen meenemen. De succesvolle ontwikkeling van SACA's zal in hoge mate afhangen van de mate waarin we deze interdisciplinaire uitdagingen aan blijken te kunnen.