Identifying the Addressee in Human-Human-Robot Interactions based on Head Pose and Speech

Diplomarbeit
Prof. Dr. A. Waibel
Interactive Systems Laboratories (ISL)
Carnegie Mellon University, USA
University of Karlsruhe TH, Germany

by

cand. inform.
Michael David Katzenmaier

Advisors:
Prof. Dr. A. Waibel
Dr. R. Stiefelhagen
Dr. T. Schultz

June 2004

Hiermit versichere ich, die vorliegende Diplomarbeit persönlich und ohne unzulässige Hilfsmittel angefertigt zu haben. Alle verwendeten Quellen sind im Literaturverzeichnis aufgeführt.

Karlsruhe, 30. Juni 2004

Abstract

In this work we investigate the power of acoustic and visual cues, and their combination, to identify the addressee in a human-human-robot interaction. Identifying the addressee in human-human-robot interaction as well as in other human-human-machine interactions is important for building a suitable machine or robot, which is able to interact like a human. Such a robot must be user friendly and should know, when to react and when not. To react in an acceptable human-like way, a voice activated machine has to know if it is the addressee of an utterance or not.

Based on eighteen audio-visual recordings of two human beings and a (simulated) robot we discriminate the interaction of the two humans from the interaction of one human with the robot. This report compares the result of three approaches. The first approach uses poorly acoustic cues to find the addressees. Low level, feature based cues as well as higher-level cues are examined. In the second approach we test whether the human's head pose is a suitable cue. Our results show that visually estimated head pose is a more reliable cue for the identification of the addressee in the human-human-robot interaction. In the third approach we combine the acoustic and visual cues which results in significant improvements.

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Identifizierung des korrekten Adressaten in Mensch-Mensch-Maschine Kommunikation anhand akustischer als auch visueller Merkmale, und schließlich deren Kombination.
Identifizierung des korrekten Adressaten in Mensch-Mensch-Maschine Kommunikation ist eine wichtige Voraussetzung eine intelligente Maschine zu entwickeln, die fähig ist, auf menschliche Weise zu interagieren. Für die Benutzerakzeptanz ist eine einfache Handhabung im Umgang mit der Maschine wichtig. Um auf Elemente zu verzichten, die die Handhabung in unnötiger Weise verkomplizieren – wie z.B. zu drückende Knöpfe, um die Adressierung anzuzeigen –, ist eine Identifizierung des Adressaten von Seiten der Maschine unumgänglich. Die Entwicklung einer Maschine, die fähig ist, auf menschliche Art zu interagieren, wie z.B. ein Roboter, schließt ein, dass dieser nur in der richtigen Situation reagiert. Deshalb muss eine sprach-gesteuerte Maschine wissen, wann sie Adressat einer Äußerung ist und wann nicht, um in einer mensch-ähnlichen Weise zu interagieren und zu reagieren.
Basierend auf achtzehn audio-visuellen Aufnahmen von zwei Personen und einem (simulierten) Roboter, wurde die Interaktion zwischen den Personen von der Interaktion einer Person mit dem Roboter differenziert. Hierzu wurden drei Verfahren separat optimiert. Das erste benutzt lediglich die Akustik. Es finden Merkmale Verwendung, die direkt aus den rohen Aufnahmen als auch aus den Hypothesen verschiedener Spracherkenner extrahiert sind. Im zweiten Verfahren wird getestet, ob die Kopfrichtung ein hilfreiches Merkmal ist. Die Ergebnisse zeigten, dass die visuell erhaltene Kopfrichtung einen größeren Beitrag zur Identifizierung des Adressaten leistet als die akustischen Merkmale. In dem letzten Verfahren werden die Akustik und das Visuelle kombiniert, was zu deutlicher Steigerung der Ergebnisse führt.

Acknowledgments

This work was conducted at the *Interactive System Labs* (ISL) in cooperation with the Language Technology Institute (LTI) of Carnegie Mellon University, USA and the "Institut für Logik, Komplexität und Deduktionssysteme" of the University of Karlsruhe, Germany. I would like to thank Prof. Alex Waibel, Director of the ISL, for giving me the opportunity to perform research at Carnegie Mellon University and thereby improve my understanding of the American culture and language.
I am grateful to my advisor Dr. Rainer Stiefelhagen of the University of Karlsruhe, especially for his advice conducting the vision-based experiments. Along with Dr. Stiefelhagen Dr. Jie Yang of Carnegie Mellon University helped me with the recording equipment for which I am especially thankful. I would also like to thank my co-advisor Dr. Tanja Schultz for her guidance along the way, first of all by the speech based experiments. I would also like to mention that she specially bought furniture to meet my special needs as a paraplegic. For that I am very grateful.
My thanks go also to Susanne Burger and her team for the help in transcribing.
Moreover, I would like to thank all the colleagues at the Interactive Systems Labs, both in Karlsruhe and in the USA, for their collaboration and their support: Sebastian Stüker, Yue Pan, Hua Yu, Ying Zhang, Christian Fügen, Matthias Eck, Hartwig Holzapfel, Petra Gieselmann, Victoria MacLaren, Florian Metze, Kai Nickel, Thomas Schaaf, John McDonough, Kornel Laskowski, Ashish Venugopal and Robert Malkin.
Many thanks to Silke Dannenmaier, Annette Römer, Kelsey Walko, Celine Carraux and Radha Rao for their cheerful assistance in administrative matters. Thanks to Michael Bett, Gopi Flaherty and Norbert Berger for having kept our machines, network and last but not least the recording equipment running.
Special thanks to everyone who participated in the data collection sessions!
I also enjoyed the warm welcome in the labs, especially from my colleagues in the office, who also helped me first of all in infrastructure matters: Fei Huang and wife, Sinaporn Suebvisai and Dominik Raub.
Finally, I would like to thank my wife Verena, for her love and understanding and especially for the help at home and in getting around with my wheelchair in Pittsburgh!

Contents

Chapter 1

Introduction

In this work we investigate the power of acoustic and visual cues, and their combinations, to identify the addressee in a human-human-robot interaction.

Identifying the addressee in human-human-machine interactions is necessary for building a suitable machine/system to interact like a human. For the acceptance of the user, easy use is important. Therefore 'push-to-talk' buttons or similar features are not desirable. Moreover, it is preferable to address the voice-activated machine in a humanlike way by simply talking to it, without any additional operations. To develop such a system, identifying the addressee is necessary, so that the machine knows, when to react and when not to react. To react in an acceptable humanlike way a voice-activated machine therefore has to know, if it is the addressee of an utterance or not.

Identifying the addressee is important for the wide range of human-human-machine interaction: voice-activated video recorders and music stations or any kind of entertainment system as well as the combination of all entertainment systems in one system we could call home-multi-media-terminal. Additionally the following devices could be voice-activated: air conditioning, information desks, automatic doors or gates and navigation systems or other systems in a car or plane. In a room the light switches, electric curtains, shades or blinds and in the future the LCD picture on the wall or even the color of the wall itself and similar features, could be voice-activated. Voice-activated devices would be especially helpful for the elderly and disabled persons. Tetraplegic people[1] would be able to control a voice-activated

[1]tetraplegic people: people disabled on all four extremities; complete tetraplegic: only able to move the head

wheelchair by themselves – as well as opening voice-activated doors etc. – and therefore become mobile. Even for un-/locking their house door, the technology of today has a solution: persons can be identified by their voice [Chil 2004] and therefore the own voice could be used as a 'key'/'code' for the door. But also for paraplegic people, such as myself, voice-activated devices help to make life a lot easier.

As this field is such a wide field, many different human-human-machine interaction scenarios (with a wide variety on voice-activated machines) could serve for recording purposes. Since a humanoid household robot could control all other devices in a human way or even have remote control, every human-machine dialog could be integrated. On this account we could incorporate every command addressed to a speech-controlled system. Thereby we reach a generality and the results of such a scenario promise to be transferred in other scenarios with other conditions.

1.1 Objective

In this work we address the problem of automatically determining when a robot was addressed by a human and when not. This is an important problem, if robots eventually become companions in our daily lives. A household robot for example should know whether a person in the room is talking to him (the robot) or to someone else in the room.

In this work we aim at detecting the addressee of a person's speech in a multi-party human-human-robot interaction scenario, by analyzing the speaker's head pose as well as his or her speech.
On the speech side possible acoustical cues are searched and their discriminative power are investigated. Complex cues as for example parseability by a special designed context free grammar, correlation of the hypothesis of two different speech recognizers and simpler cues as utterance length in ms and number of words to name a few.
To this end, we recorded eighteen multi-party interactions with a simulated robot and analyzed the power of head pose and acoustic cues to discriminate between the addressees of the speakers.

1.2 Outline

The remainder of this paper is organized as follows: in Chapter 2 we review some related work; Chapter 3 describes the data collection setup; Chapter 4 investigates, how well the addressees of a speech act can be determined based on the visually estimated head pose of a person; Chapter 5 describes our experiments with identifying the addressee based on acoustic cues and analyzing the speaker's speech; Chapter 6 presents experimental results for audio-visual determination of the addressee and Chapter 7 summarizes our findings.

- Chapter 2: Review of Related Work
- Chapter 3: Description of the Data Collection Setup
- Chapter 4: Usefulness of visually estimated Head Pose in Identify the Addressee of a Speech Act
- Chapter 5: Identifying the Addressee based on Acoustic Cues and analyzing the Speaker's Speech
- Chapter 6: Experiment Results for audio-visual Determination of the Addressee
- Chapter 7: Conclusion

Chapter 2

Related Work

2.1 Gaze, Head Pose and Body Orientation

Research suggests that gaze, head pose and body orientation play an important role during social interaction and are used and perceived as a signal of attention during human interaction [Argyle 1969, Ruusuvuori 2001, Tankard 1970, Kleinke et al. 1973].

2.1.1 Eye Gaze versus Head Pose

Stiefelhagen and Zhu [Stiefelhagen et al. 2002] have investigated the relation between eye gaze and head orientation in multi-party interaction between four people. They concluded that head pose is a reliable cue to determine at whom someone looked in small meetings.

They could estimate the visual target based only on head orientation with an average accuracy of almost 90%, although the average head orientation contribution was only 68.9%.

They found that head orientation and eye orientation are pointing in the same horizontal direction most of the time (in their case 87% of the time). Additionally they conducted that eye gaze cannot be obtained a fifth of the time due to eye blinks and human varying in using head orientation by changing gaze direction. The test results vary considerably for each person,

because of their different behavior.

Furthermore they point out that in practical scenarios of human-machine interaction the focus of attention is the desired information and the exact gaze is mostly not needed. Therefore calculation of the overall gaze – estimated by adding the eye gaze to the head pose – is often not necessary and head pose is sufficient to determine the visual target in many cases.

2.1.2 Estimate Eye Gaze and Head Pose

Most of the current existing eye gaze and head pose tracking systems cumber users with head mounted equipment or set heavy restrictions on user's behavior:

- eye gaze tracker: [ARR], [ASL], [SRR], [SMI] and [LCT].
- head pose tracker: [Polhemus], [iReality], [Ascension] and [Stiefelhagen 2002].

A non-instructive solution to estimate the head orientation based on a neural network is proposed in [Stiefelhagen 2002]. This system obtains the head orientation of a simple video sequence frame by frame out of the facial image and therefore put no restrictions on user's behavior. It does not cumber users with head mounted equipment [1]. A more detailed description is given in the section 4.3 "Head Pose Estimation" since this system is used in this work.

2.2 Relation between Gaze and Speech

The relation between gaze and speech in multi-party communication between several people recently has been investigated by Vertegaal et al [Vertegaal et al. 2001]. They found that subjects looked about three times more at individuals they spoke to.

Other researchers have investigated how people use speech and gaze when interacting with attentive objects in a smart environment. Maglio et al [Maglio et al. 2000] have for instance shown that people tend to look towards objects with which they interact by speech. In their study they found

[1] such as cameras or special light sources

that subjects nearly always looked at the addressed device before making a request.

This behavior rule 'people look where they talk to' is in the literature called conversational hypothesis. 1977 Argyle & Graham proposed that the first behavior rule may be overruled by the presence of "situational attractors" [2] such as a television set that distracts our attention during a conversation (also proposed in [Bakx et al. 2003]). Most researchers find the conversational hypothesis true for situations, where only one person speaks to a system or machine. But in situations where two or more parties speak to each other and interact with the system, people are often attracted by the system as the situational attractor hypothesis predicts. This fact makes it necessary to use additional cues besides head pose or eye gaze. With visual cues alone, we would therefore make false classification by determination of the acoustical addressee in presence of situational attractors that draw our visual attention to themselves, while speaking to others.

2.3 Combination of Audio and Video

[Zahn et al. 1996] was one of the early works in building a system that models the human ability to separate unknown sound under natural conditions. Besides sound they used acquired knowledge and interaction with other sensory systems as sources. Its design allowed combination of visual, acoustical and other sensory input, they said.

2.3.1 Identifying the visual Target based on Head Pose and Speaking Information

In [Stiefelhagen 2002] the target at which subjects looked or as they say focus of attention is predicted based on head pose and sound. The information, if someone is speaking and who is speaking, is the used cue on the sound side. Additionally they found that the temporal information of who is speaking during the last N frames is a helpful cue.

Identifying or predicting of the visual target based on (acoustic) sound is just

[2] situational attractors: objects or situations in the environment that attract people's eye gaze when they are talking to each other

the other way around as investigated in this work: identifying the acoustical addressee based on head pose (as a visual cue).

They concluded that head pose outperforms the other cues, using temporal sound information leads to better results than using only the actual sound information and last but not least that combining all cues outperforms the single use.

2.3.2 Identifying the acoustical Addressee based on Head Pose and Utterance Length

Bakx et al. [Bakx et al. 2003] have analyzed facial orientation during multi-party interaction with a multi-modal information booth. They found that users were nearly always looking at the screen of the information kiosk when interacting with the system. However, when the user was talking to a friend next to the system, the user was still looking towards the information system in 57% of the time, thereby limiting the discriminative power of facial orientation to find the addressee. Therefore they concluded that facial orientation could be used as a cue to identify the addressee of an utterance only asymmetrically.

Bakx et al. also analyzed using the utterance length of the speaker for discriminating between addressees. They concluded that by combining the acoustic cue with facial orientation, some improvement in detecting the correct addressee can be achieved. Furthermore they noted that face orientation is not only governed by "looking at the addressee" but also by "looking at the speaker". Therefore they included facial orientation of the non-speaking participant as an additional cue and found that some additional improvement can be achieved.

2.4 Conclusion

We have seen that gaze, head pose and body orientation play an important role during social interaction and are used and perceived in particular as a signal of attention during human interaction. Furthermore it was shown that head pose is sufficient to determine the overall gaze and is easier to estimate than eye gaze.

For the relation of gaze and speech it was found that gaze can be predicted based on two hypotheses:

1. The conversational hypothesis: *people look where they talk to*
2. The situational attractor hypothesis: *the first behavior rule may be overruled by the presence of 'situational attractors'*

The combination of all cues outperformed the approaches using fewer cues in recent works as for instance reported in [Stiefelhagen 2002] and [Bakx et al. 2003]. [Stiefelhagen 2002] proposed the results of just the opposite of one experiment conducted in this report: Identifying the visual target based on sound. One part of this work is identifying the acoustical addressee based on vision (head pose). [Bakx et al. 2003] is one of the first proposals combining vision and audio to identify the acoustical addressee.

Identifying the acoustical addressee based only on speech has not been researched yet. In this work is also a system built to identify the acoustical addressee based only on the speech of the speaker (see also Studienarbeit Katzenmaier [Katzenmaier 2003]; our first approach based only on acoustical cues).

Chapter 3

Experimental Setup

To identify if the addressee is a human or a machine, the scenario has to combine human-human and human-machine interaction. More exactly it is desirable that the machine recognizes that it is addressed by a human. – By distinguishing human-human and human-machine dialogs, recognition of machine-human monologs is not interesting, but only human-machine monologs. If it was interesting to detect machine-human monologs, it would be trivial to detect them. Therefore we can focus on human-machine monologs[1].

One way to collect human-machine dialogs for our experiment would be to collect data in a Wizard of Oz experiment. A Wizard of Oz experiment is a simulation of such a system, in which its reaction is controlled by a human. The more this person is behind the scene, the better the Simulation or the Wizard of Oz experiment.
As already mentioned in the introduction, many scenarios with different voice-activated machines could be thought of, but a household robot seems to be the most general choice and thereby promises to give the best results by transferring our system to other scenarios.

In our case, an appropriate Wizard of Oz experiment would be to take a robot and control its tasks by a human, depending on the requests given. But unfortunately we do not have a robot or a close simulation. Furthermore a simulation like dressing up a person as a robot always depends on the imagination of the recorded persons somehow. The first plan was that the author plays or controls the robot to accomplish a Wizard of Oz experiment,

[1] later called *commands*

but it turned out that he had to serve in a different role (as the guest in the scenario explained next). Therefore a compromise had to be accomplished: no real robot is used and no dressed up person is in the scenario included, but a standing camera with a face.

Moving or Standing 'Robot'

In case of a moving robot, only the frames and therefore utterances the recorded person is looked at by the *robot* are interesting; respectively the frames, where the recorded person is in the camera view. Therefore all other utterances, where the recorded person is not in the camera view would be useless. On this account it would be necessary to select all data, where audio *and* video data are received. [2] Using only the frames, where the host is looked at, in comparison to let the robot remain standing and let it (the robot) always look to the host therefore only differ in the amount of useful data; in the last case all frames can be used. Thus the camera can be remain standing. Moreover all data could be used and time to select the data, where both video and audio data are received, can be saved.

3.1 Scenario

The data collection setup mimics the interaction between two humans and a robot. One person -acting as the host- introduces another person -acting as his/her guest- to the new household toy, a robot. In order to provoke a challenging scenario which includes *robot commands* directed to the robot and also conversation *about the robot*, the participants were given instructions beforehand about the (simulated) features of the robot (brings drinks, performs household duties) and were asked to discuss these features, the pros and cons of a robot in the household, as well as to give commands to the robot. See the 'Record Guide' in appendix A for further details.

As we wanted the scenario to be as natural as possible, we told the participants in our data collection to behave towards the robot as they wished.

[2] Studies with using only the audio recordings are already done in [Katzenmaier 2003]. Furthermore experiments with only audio or video recordings can be established by simply ignoring the other recordings; the results therefore you find in the chapters 'Identifying the Addressee based on Head Pose respectively based on Speech'

This included that people spoke about different topics in some interjectional conversational parts, mostly at the beginning or end of the session and thereby making the speech recognition challenging. For example they spoke about their jobs, their social interaction (i.e.when to swim, when to meet) and so on.
Only the above mentioned instruction were given to enforce the following:

1. making sure that the planned scenario will be accomplished
2. making the task challenging by perform a demonstrating task

By demonstrating something, you always talk about the subject being demonstrated; in our case the robot. And talking about the robot, i.e. what he can do, makes the differentiation of commands and conversations more difficult as the following example shows:

"Yeah, and when I tell the robot 'bring me something to drink' he does."

"Robot, bring me something to drink!"

The phrase 'robot, bring me something to drink' is in both sentences and thereby making the differentiation to a very difficult challenge. In the first sample the words before and after the mentioned phrase could be recognized as noise and thereby making it to the exact same sequence of words as in the second sample. But the second one is a command and the first one is not.

3.2 Data Collection

Figure 3.1 shows the data collection arrangement. The robot consists of a construction using a Canon VCC-1 camera to simulate the "eyes", and a Sony distance microphone to simulate the robot's "ears". The distance between humans and robot is about 4 meters. Since we expected the far-distance speech recognition performance to deteriorate, we additionally recorded close talking speech using a Sennheiser lapel microphone.

The described experiments focus on the recordings of the host, since the aim is to build a system (for the robot or machine) determining if the owner of the robot addresses the robot or the guest. As it is more natural that the host is the owner of the robot, he is the one giving commands to his robot.

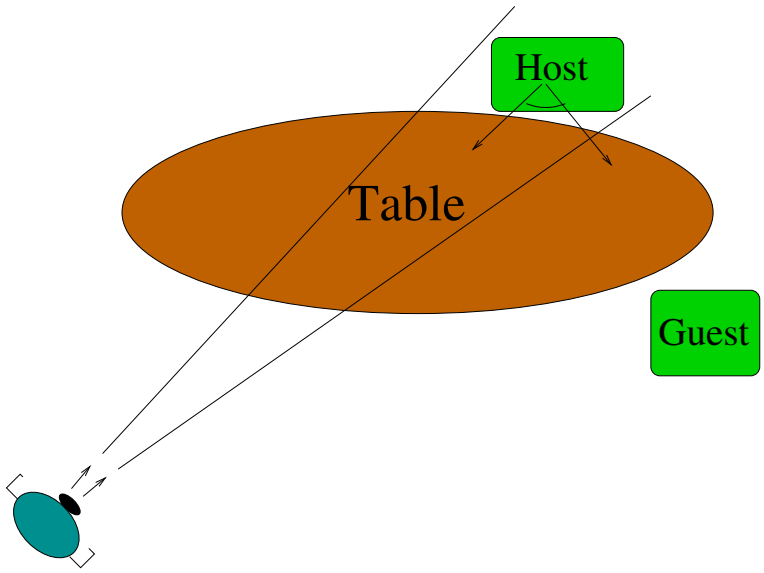

Figure 3.1: Data Collection Setup

He is therefore the one in the point of interest and in the view of the camera as figure 3.1 shows.

All recordings were done in English, the host speakers are native American English speakers. As there were less people available as planned in the beginning, the author, a non-native speaker, had to serve as the guest. For the speech recognition the recorded persons should be native speakers and therefore this is an additional reason, why only the data from the host is used.
The flow of the dialog is normal with no great pauses between each turn of each speaker, so they behave normal. A non-native speaker can be a problem. If for example the native speaker realizes that he is difficult to understand, he tries to speak slower and more clearly. This can among other things be recognized by pauses before answers or many back queries for understanding purposes.

Altogether we have recorded nearly 30 sessions, each of roughly 10 minutes in length. Due to not working microphones and similar problems, 18 recording sets could be used. The audio data was fully transcribed and tagged on the turn level to indicate whether the host addresses the robot or the guest ([command],[no command]). The program TransEdit (version 1.1 beta 11) by Susanne Burger and Uwe Meier was used (see the snapshot in figure 3.2) and the transcribing rules abutting on the ones used for the Nespole project

```
[..]
e002\_1\_0010\_GAH\_00: yeah , yeah he is <uh> pretty great he can
 <uh> <P> basically see and <uh> hear and do everything that <uh>
 that you expect maybe a human to do , so <uh> <P> you know for
example he could go +/and/+ <uh> <B> and <uh> get us some beer for
example , so .

e002\_1\_0011\_GAH\_00: robot , go get us some beer <;r> .

e002\_2\_0012\_SQYZCM\_00: <Smack> <hm> nice <%> <*T>t

e002\_1\_0013\_GAH\_00: two please <;r> .
[..]
```

Table 3.1: Example for Transcribing. The tag <;r> was used to mark a command.

[Burger 2003] was applied. A turn is defined as talking from one speaker with no longer pause than the length of three words. Additionally changed topics define a new turn. Examples are listed in table 3.1.
Every turn starts with an identification key, starting with 'e' for signalizing that it is English data, followed from the session ID (3 numbers long), appended from the speaker number (1 digit) and the turn number (4 digits). Then follows the anonym speaker ID (3 to 6 letters) and finally 2 digits reserved for further decoding of information – these two digits could i.e. serve for decoding a command, but the comment tag <;r> is used instead (all comments within the tag are signalized like <;*command*>; after the turn ending punctuation mark or turn abrupt signal '<*T>t' comment lines with starting ';' can follow). All segments of the turn identification are divided by an underscore. See the reference for further details.

Data	# Session	Length		# labeled visual targets
		[min:sec]	[frames]	[frames]
Training	8	82:37	32491	5024
Development	5	50:35	20435	-
Evaluation	5	51:35	20435	-

Table 3.2: Audio and Video data

For four training sessions we manually labeled the first 2.5 minutes of the video recordings. In each frame the visual target, to which the recorded

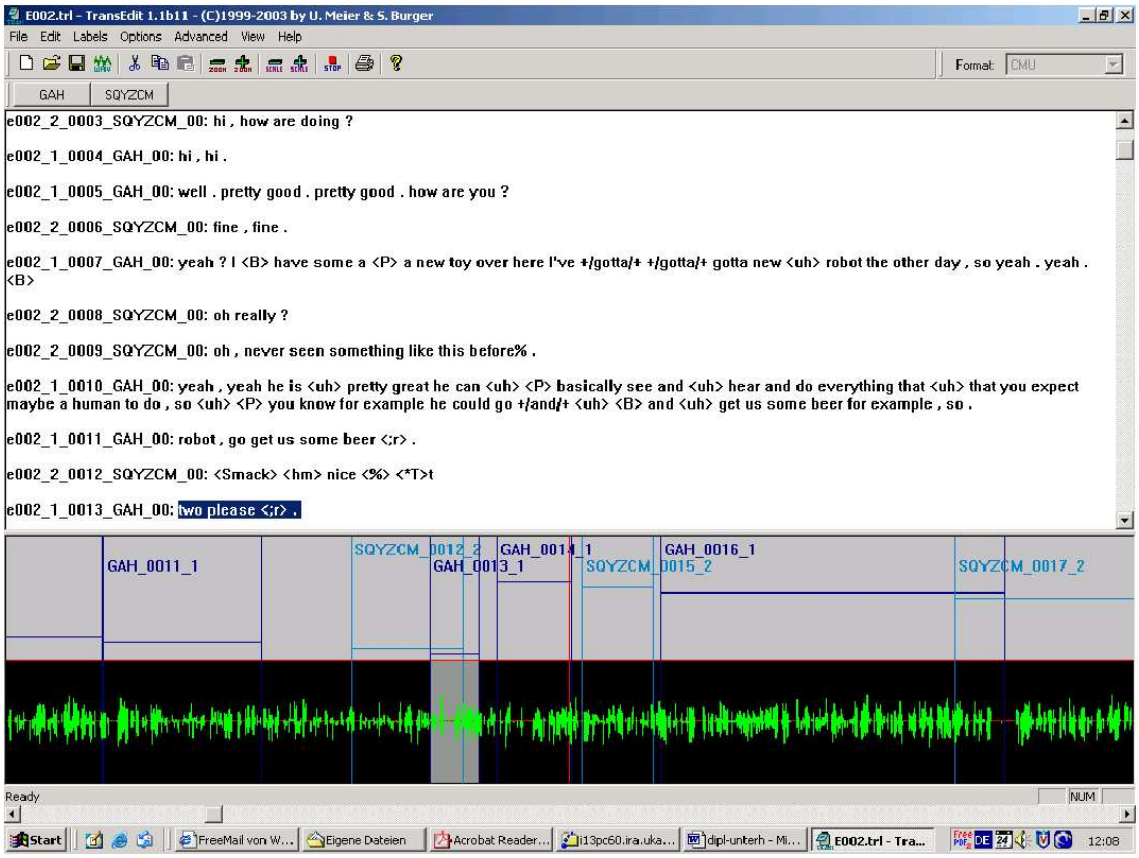

Figure 3.2: Typical transcription part from TransEdit. The gray marked signal down (labeled as GAH_0013_1) corresponds to the highlighted text in the transcription.

person is looking at, is labeled (*Robot*, *Guest* or *Other*). Table 3.2 shows the audio and video data and the division into training, development, and evaluation set for the speech engine developing.

Chapter 4

Identification of the Addressee based on Head Pose

In this section we describe, how we estimate the likely addressee of a speech act, based on visual estimation of the speaker's head orientation.

4.1 Relation of visual target and the addressee of a speech turn

To check, how good gaze or head pose are as an indicator of the (acoustically) addressed target, we first analyze the correlation between the manually labeled acoustic targets - i.e. the addressees of a speech act - and the manually labeled visual target, i.e. the targets that had been looked at.

Humans naturally look at the highest point of interest, to our focus of attention. This is often the person we speak to. Also we can recognize that this is a behavior rule. Parents i.e. teach their children to look to people they speak to or listen to. Therefore our hypothesis was that there would be a high correlation between the visual and acoustic targets or in other words: people look, where they talk to. Here we can formulate more specific:
"Does the host look at the robot, when he talks to it?"
"Does he not look at it, when he speaks with others?"

To see if this behavior rule is also true when humans and machines com-

municate, we manually labeled at which target the host in our experiments was looking for four videos in our data set. Here, the visual targets could be either the "Robot" the other person ("Guest") or anything else ("Other"). The acoustic targets, as in all our experiments could of course be either the "Robot" or the "Guest". Table 4.1 shows the confusion matrix between the acoustic and visual labels. Here, the acoustic targets (the addressee) are indicated with T_A and are given columnwise, the visual targets (at whom did the speaker look?) are given in rows, labeled with T_V.

Audio \ Video	$T_V = Guest$	$T_V = Robot$	$T_V = Other$
session 1 $T_A = Guest$	462	44	202
$T_A = Robot$	3	43	2
session 2 $T_A = Guest$	463	69	136
$T_A = Robot$	0	94	0
session 3 $T_A = Guest$	289	34	221
$T_A = Robot$	0	46	3
session 4 $T_A = Guest$	575	2	5
$T_A = Robot$	6	93	2
Sum $T_A = Guest$	**1969** (73%)	**149** (6%)	**564** (21%)
$T_A = Robot$	**9** (3%)	**276** (95%)	**7** (2%)

Table 4.1: Confusion-matrix between hand-labeled addressees of speech acts (T_A) and the targets at which the speaker looked (T_V).

It can be seen that people mostly looked at the robot when they addressed the robot (95% of time). In 35% of the frames, however, people did not talk to the robot while still looking at him. We also see that when the host looked at the guest, then in almost all cases (1969 occurrences out of 1978 cases, also 99.5% of time) he also addressed the guest.

The hypothesis "The host looks at the robot, when he talks to it" and "He does not look at it, when he speaks with others" can be established; people look at the robot, when they speak to it (95% of time), people look from half of the time (session 1: 51%) to almost never (session 4: 2%) to the robot, while speaking to others (here the guest).

To summarize, in the data that we recorded, looking towards the other human was a direct indication that the other person was addressed. Looking at the robot, however, was not such a clear cue: Here in 65% of the cases the robot was addressed and in the remaining 35% of the cases, the other human was

the addressee of the speech act.

4.2 Using Gaze or Head Pose

Here head pose will be taken to determine the visual target. The overall gaze would be a more accurate way to determine the visual target, but it is not that easy to measure the eye gaze. And besides that, head pose is an accurate way to determine the visual target – see subsection 2.1.1 in the chapter related work.

In most cases the determination of the eye gaze requires special technical equipment, which has to be worn by the recorded person. There are some systems available, which do not need the recorded person to wear the special equipment, but still there is special equipment necessary to perform the recordings (special camera or similar technical equipment). For head pose tracking instead, there are systems not needing any special equipment at all: not worn by the recorded person and not for recording purposes. [1]

As we want the scenario to be as natural as possible, we do not want the recorded person to wear such special equipment. Additionally it would be nice, if we would get accurate results without special recording equipment (no infra red camera and so on), so that standard equipment available almost everywhere would be sufficient. Therefore the aim is to get the results out of recordings taken with a normal video camera and normal microphones.

4.3 Head Pose Estimation

The approach for estimating head-orientation in this work is view-based: In each frame, the head's bounding box - as provided by a skin-color tracker - is scaled to a size of 20x30 pixels. Two neural networks, one for pan and one for tilt angle, process the head's intensity and grey-scale images and output the respective rotation angles. As we directly compute the orientation from each single frame, there is no need for the tracking system to know the user's initial head orientation.

[1] See the references of the systems listed in subsection 2.1.2 for details

The networks we use are organized in 3 layers and were trained in a person-independent manner on sample images from nineteen users. In previous experiments in our lab we obtained mean angular errors for head orientation estimation of around 10 degrees for pan and tilt on new users [Stiefelhagen et al. 2001a].

In this work, we only use horizontal head orientation (pan) to distinguish between different addressees of a person. It should be noted that the used system for estimating head orientation had been trained on images taken from different persons than those that participated in this study. Furthermore, the images used for training the system were recorded several years ago in a different lab and under different lighting conditions.

To find out, how accurate the head pose tracker works in the new environment, pre-recordings with people asked to look to their left and right as well as straight in the camera were accomplished. Each frame was labeled as left, right or straight. Then the head pose tracker was used to determine the head pose in degrees for each frame as well. After determination of the most likely target (see next section), 80-90% correlation between manual labels and estimated targets was the result. This result was good enough to promise accurate results in our scenario.

Finding the Head's Bounding Box

Since the main differences between skin color of different individuals – including Asian, black and white faces – in the (from most cameras used) RGB representation, are basically brightness, the chromatic color space (r,g) with the absence of brightness by normalization is used:

$$r = \frac{R}{R+G+B},$$
$$g = \frac{G}{R+G+B}.$$

The color blue, defined as $b = \frac{B}{R+G+B}$, is redundant after normalization, because $r + g + b = 1$. Skin color forms a cluster in chromatic color space and is represented by a Gaussian model. After converting the image in chromatic color space and computing the probability of being skin color using

the Gaussian skin color model, the face is located by searching for the largest connected region of skin colored pixels. An additional neural network is used to distinguish faces from other skin color objects such as hands, arms and shoulders.

The facial images are preprocessed in three different ways separately after grey scaling: histogram normalization, horizontal edge detection and vertical edge detection. After this preprocessing each image is down sampled and the pixels of all three images serve as an input vector for two different networks separately, one for pan and one for tilt angle mentioned above.

4.4 Finding the most likely Target

Once a user's head orientation has been estimated, we want to find the most likely person or target at which the user has been looking. To do this, we use an approach that was described in [Stiefelhagen et al. 2002]. This approach was built to find out at whom participants in a meeting have been looking, based on their head orientations. Similar to their approach, we try to identify at which target - the robot or the other human (the "Guest") - the speaker had looked, by finding the target that maximizes the posterior probability $P(Target|Head\ Orientation)$.

To compute the a-posteriori probabilities for the visual target T_V for each class, first the a-priori probability $P(T_V = target)$, class conditional probability $P(X|T_V = target)$ and the probability $P(X)$ for each head pose X has to be calculated. Once these probabilities are calculated, the a-posteriori probabilities $P(T_V = target|X)$ can be calculated:

$$P(T_V = Target|X) = \frac{P(X|T_V = Target) \cdot P(T_V = Target)}{P(X)} \tag{4.1}$$

where $Target$ can be either "*Robot*" or "*Guest*" in our case, and X denotes the horizontal head orientation of the host.

The most likely target for a specific head pose X is found by determination of the maxima of all probabilities for all targets given the head pose X:

$$\max_{Target} P(T_V = Target|X) = \max_{Target} \frac{P(X|T_V = Target) \cdot P(T_V = Target)}{P(X)} \quad (4.2)$$

$$= \max_{Target} (P(X|T_V = Target) \cdot P(T_V = Target)) \quad (4.3)$$

As equation 4.3 shows, the probability for each head pose $P(X)$ is unimportant for determining the maxima posterior probability for all targets. This is true to the fact that the probability $P(X)$ for the same head pose X is the same in the determination of each posterior probability.

Figure 4.1 depicts typical class-conditional probability distributions for the classification of the visual target based on a person's head orientation for two targets on the left and for three targets on the right in one of our data sets.

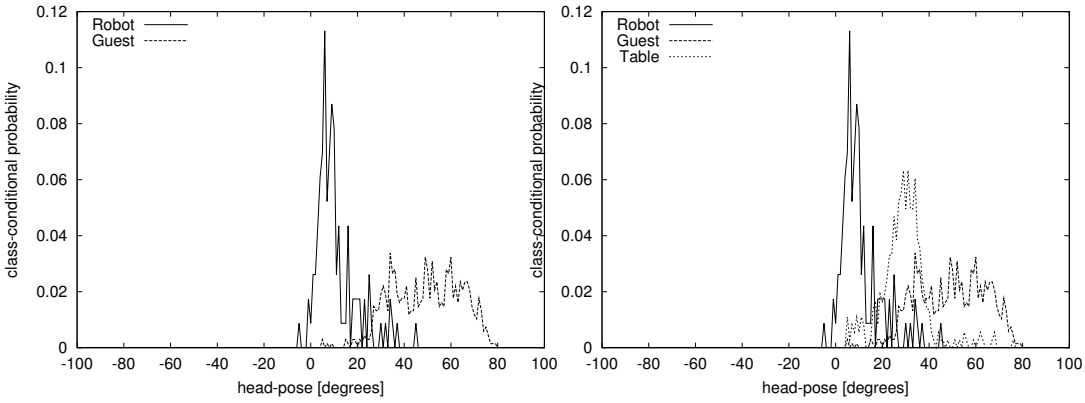

Figure 4.1: Typical class conditional probability distributions for the classification of the visual target for two targets (left) and for three targets (right).

4.5 Unsupervised Learning of the Probabilities

Our approach for unsupervised Adaptation of the model parameters is based on the assumption that the class conditional head pan distribution, such as depicted in figure 4.1, can be modeled as Gaussian distributions. Then the distribution of all head pan observations from a person will result in a

mixture of Gaussians:

$$p(x) \sum_{j=1}^{M} p(x|j)P(j) \tag{4.4}$$

Here the number of Gaussians M is first set to two – modeling the two target classes *Robot* and *Guest*. And then the number of Gaussian is also set to three – modeling the two classes *Robot* and *Other*, where *Other* is the combination of the two former classes *Guest* and *Other* – see section 4.6.2 for more details. Therefore after combining these two classes into one class, it is again a two-class classification (*Robot* and *Other* standing for looking at robot or looking to some other visual target). The combination is implemented by calculating the posterior probability of the class *Other* as the difference from 1 and $P(T_V = Robot|X)$:

$$P(T_V = Other|X) = 1 - P(T_V = Robot|X)$$

where X again is the head pose in degree.

In opposition to previous experiments in our lab [Stiefelhagen 2002], where the number of Gaussians was set to the number of other participants, here the robot is an additional visual target and therefore the number of Gaussians is set to the number of other participants plus one – that is the number of all participants. As before assumed these are the most likely targets that the person, here the host, has looked at during the session and it is desirable to find the individual Gaussian components, that correspond to looking at these targets.

The expectation-maximization algorithm is used to iteratively update the model parameters of the mixture model and thereby adapt them to maximize the likelihood of the pan observations given the mixture model. After adaptation of the mixture model to the data with a maximum of 30 iterations, the individual Gaussian components of the mixture model are used as approximation of the class conditionals $P(X|T_V = Target)$ of our visual target identification model described in equation 4.1. Furthermore the priors $P(j)$ of the mixture model are used as the target priors $P(T_V = Target)$. To assign the individual Gaussian components and the priors to their corresponding target, the relative position of the participants can be used; here the relative position of the target classes *Robot* and *Other* respectively the target *Guest*.

Figure 4.2 shows an example of the adaptation on pan observation from one user. In figure 4.2(A) the true distribution of all head pan observations of the user is depicted along with the Gaussian mixture adapted as described above

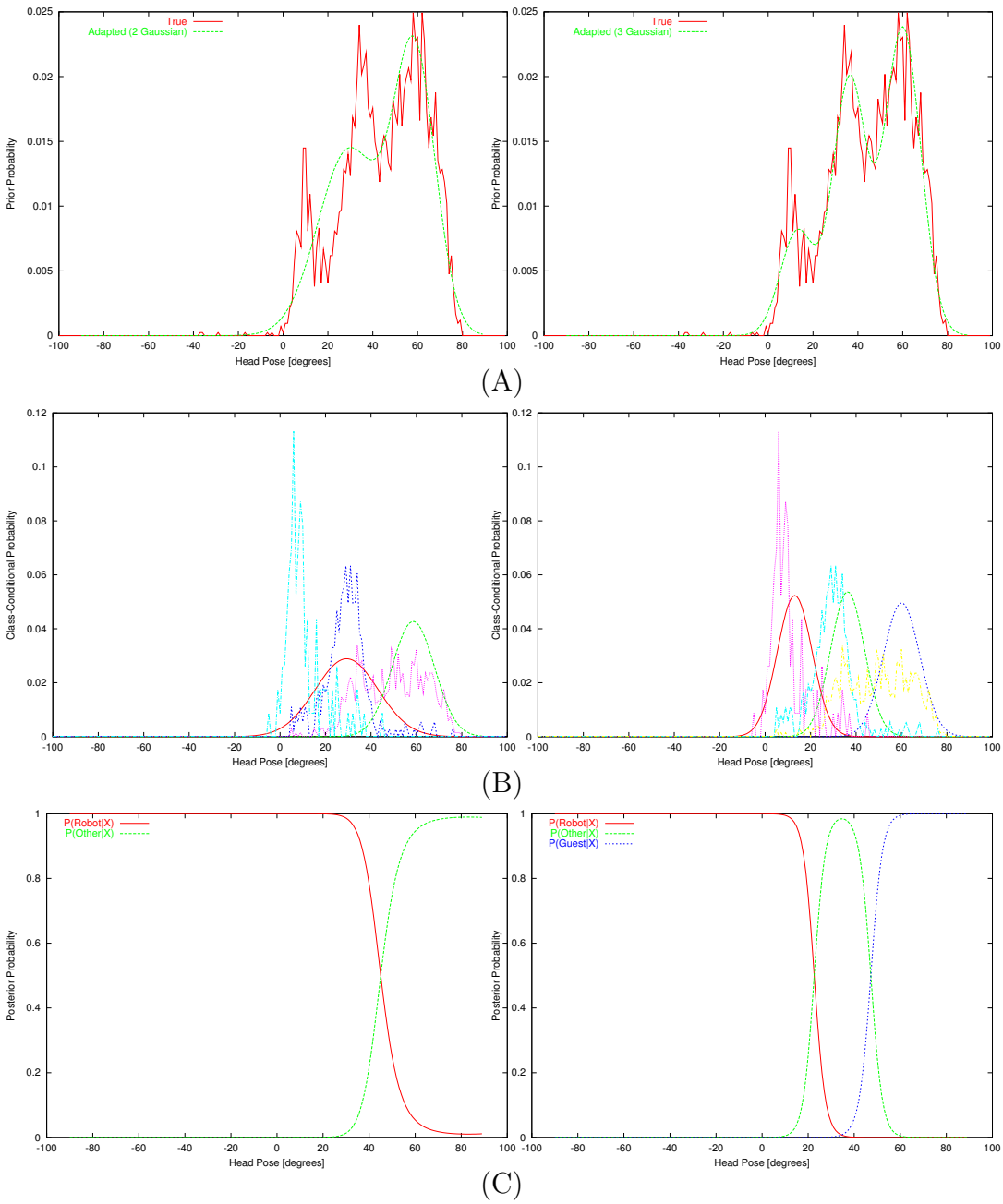

Figure 4.2: A) The distribution $p(x)$ of all head pan observations of one person (True) and the adapted mixture with two Gaussians (left) and with three Gaussians (right). B) True and estimated class-conditional distributions of head pan x for the same subject. The adapted Gaussians are taken from the adapted Gaussian mixture model depicted in A). C) The posterior probability distributions $P(\text{Target}|x)$ resulting from the found mixture models.

for 2 learned (left) and 3 learned (right) Gaussians. Figure 4.2(B) depicts the real class-conditional head pan distributions of that person, together with the Gaussian components taken from the Gaussian mixture model depicted in figure 4.2(A). As can be seen, the adaptation with three Gaussians provide a better approximation of the real class-conditional distributions of the person. (The real class-conditional distributions are not necessary for the adaptation of the Gaussian components – see [Stiefelhagen 2002].) Figure 4.2(C) depicts the posterior probability distributions resulting from the adapted class-conditionals and class priors. In case of three Gaussians the Gaussian adapting the posteriori probability of the target *Robot* is taken to calculate the posterior probability for the target *Other* – $P(Other|X) = 1 - P(Robot|X)$ see subsection 4.6.2.

4.6 Experimental Results

4.6.1 Used Metrics

In this work we are mainly interested in detecting when a robot was addressed.

To detect all such commands we are interested in a high recall value: how many commands of all commands are really detected. Also we want to have as few as possible false alarms. In our case a false alarm – or false accept (fa) – would be, if we say to an utterance it is a command, but it is actually a conversation part. A correct accept (ca) would then be a command, which is detected as a command. False reject (fr) and correct reject (cr) have to be understood respectively. [2] The precision value is higher if the number of false alarms is lower. Or in other words, precision expresses how precise our command detection is: how many of the as command detected utterances are really commands. Figure 4.3 visualize this values in a confusions-matrix.

Therefore, we are interested in measuring precision and recall of detected speech acts that were addressed to the robot. In order to compare the results of different experiments more conveniently, we combine values for recall and precision into one single number, the so-called f-measure, which is the

[2] false reject: not detected command; correct reject: correct rejected conversation or in other words not as command detected conversation

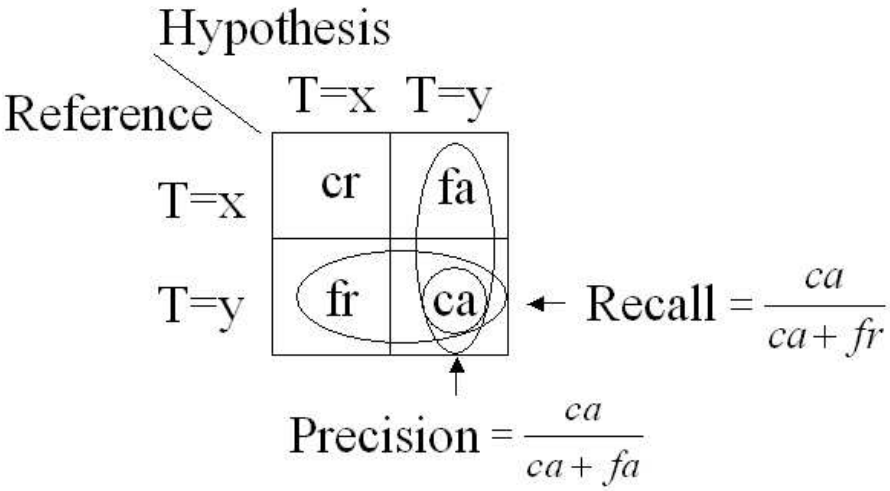

Figure 4.3: Visualization of recall and precision in the confusions-matrix with correct reject (cr), false accept (fa), false reject (fr) and correct accept (ca).

geometrical median of the two values:

$$f - measure = \frac{2 * recall * precision}{recall + precision} \tag{4.5}$$

Since it is also interesting to see how often the correct addressee of a speech act - the robot or another person - was detected, we also indicate the classification *accuracy*, which is the percentage of correctly classified targets.

Discussion: What to optimize?

First we have to decide, whether discarding utterances that are in fact addressed to the system should be avoided (recall) or whether modest levels of discarding relevant utterances are in fact acceptable if it prevents the system from reacting to many irrelevant utterances (precision) (see [Bakx et al. 2003]).

The following examples show that in some situations highest precision is desired: in a voice-activated medical operation system i.e. we would definitely like to have the highest precision possible to prevent unintended actions of the system/machine (Once something is cut, it is cut). In a situation where the system makes drinks, it is always still possible to say I wanted this or that drink instead. Bring this back and bring that instead.

We would suggest that over all in an every day scenario precision and recall values as high as possible are desired. Since f-measure is the geometrical median, this would mean the highest f-measure[3] is the desired aim in such scenarios.

4.6.2 Estimation of the visual Target

In our first experiment we tested how accurately we could identify the *visual* target of a person, i.e. the target at which a person looked. To this end we manually labeled the visual targets in four of our recorded sets to obtain ground truth. Estimation of the visual target was then accomplished by using the neural networks for head pose estimation and the Bayesian approach for finding the most likely visual target. For the experiment, we first used the true priors and class-conditional distributions of head pose to determine how well the visual target can be estimated for different amount of targets or classes. We then also learned the priors and class-conditionals automatically with an approach described in [Stiefelhagen et al. 2002].

Visual Target Classification of two Targets – *Robot* or *Guest*

First of all it would be interesting to find out how good the classification works by taking only the two classes of interest: *Robot* and *Guest*. Therefore we here look only at the results of the frames which are hand labeled to *Robot* or *Guest* and do not take the as *Other* labeled frames into account.
The observed a-priori probabilities are listed in table 4.2(A).

Once the a-posteriori probabilities are calculated, the maxima of the posterior probabilities $P(T_V = Robot|X)$ and $P(T_V = Guest|X)$ for the different targets given the head pose X has to be determined as explained in section 4.4. Here this means, a frame is classified as looking to the target with the greater posterior probability.

As we can see in table 4.2(B) the visual target could be correctly detected in 96% of the frames. Occasions when the person was looking towards the robot could be detected 77% of time, with a relatively high precision of 0.89, resulting in an f-measure of 0.82.

[3] see 'Used Metrics' subsection 4.6.1

	$P(T = Rob.)$	$P(T = Gst.)$
sess. 1	0.14	0.86
sess. 2	0.21	0.79
sess. 3	0.18	0.82
sess. 4	0.11	0.89
Avrg.	0.16	0.84

(A)

Precis.	Recall	F-Meas.	Accu.
0.93	0.85	0.89	0.97
0.94	0.94	0.94	0.97
0.87	0.89	0.88	0.95
0.83	0.42	0.56	0.92
0.89	0.77	0.82	0.96

(B)

Table 4.2: A) A-priori probabilities $P(T = Target)$ of the two visual targets – *Robot* and *Guest*. B) determination of the target *Robot* based on visually estimated head pose (head pose estimation: section 4.3).

By comparing the results of the first two sessions, we see that in both cases the classification can be accomplished correctly 97% of the time. But recall and precision are better for session 2.

Reference \ System	$T_S = Guest$	$T_S = Robot$
session 1 $T_R = Guest$	664	7
$T_R = Robot$	17	96
session 2 $T_R = Guest$	1119	17
$T_R = Robot$	20	289
session 3 $T_R = Guest$	470	15
$T_R = Robot$	12	98
session 4 $T_R = Guest$	825	9
$T_R = Robot$	63	45

Table 4.3: Confusion-matrix between hand-labeled references of the video frames (T_R) and the visual targets our Bayesian approach (system) assumes with hand tuned distributions for two targets (T_S).

The confusion-matrix in table 4.3 shows more details. For session one, 96 out of 113 (96+17) commands are detected. This gives us a recall value of 0.85. Here the recall gives the percentage of the frames that the person was looking at the robot detected correctly. In session one, 93% of time the frames our system hypothetical classified the persons visual target as *Robot* the host really was looking at the robot. (96 out of 103 (96+7)).

How to include the Frames, where people look not to the *Robot* or to the *Guest*

The focus of this work is here to recognize looking at the robot (and later talking to the robot respectively detecting commands). So it would be enough to determine whether the target is robot or guest. But because in 20% of the frames people looked at the table or between the robot and the guest, these frames have to be included and therefore get their own class. By doing so, three targets would be obtained: *Robot*, *Guest* and *Other*.

Target =	*Robot*	*Guest*	*Other*
session 1	0.1	0.61	0.29
session 2	0.19	0.69	0.12
session 3	0.12	0.47	0.4
session 4	0.11	0.84	0.05
Average	0.13	0.65	0.21

(A)

Precis.	Recall	F-Meas.	Accu.
0.78	0.73	0.76	0.78
0.77	0.83	0.8	0.85
0.71	0.88	0.79	0.62
0.83	0.42	0.56	0.88
0.77	0.72	0.72	0.78

(B)

Table 4.4: A) A-priori probabilities $P(T = Target)$ of the three visual targets – *Robot*, *Guest* and *Other*. B) determination of the target *Robot* based on visually estimated head pose (head pose estimation: section 4.3).

Now all frames (including the frames manually labeled to *Other*) are taken into account. Since these frames are more difficult to classify correctly, almost all results are slightly worse, except for session 4, that remained as good as before and has now the best accuracy. The recall, precision and f-measure did even not drop at all. The reason is that in that particular session, the recorded person, serving as host, moved very fast from looking to the guest to looking to the robot and vice versa. Therefore only a few frames occur, where the host is looking somewhere else as can also be seen as only 7 frames are manually labeled with *Other* (table 4.5). And in all sessions approximately half or even more than half of the frames, hand labeled to *Other*, are assigned incorrectly. Since in session 4 these are only 7 frames, the results remain as good as before.

In all sessions only a few of the as *Other* labeled frames are assigned to *Robot*. So by adding these frames to the class *Guest*, the errors would be minimized. Additionally you see in table 4.1 that the host almost always has looked somewhere else (labeled as *Other*), when he spoke with the guest. If these frames are added to the target class *Guest*, the errors occurring in the situation $T_R = Other$ and $T_S = Guest$ as well as in the situation $T_R = Guest$

Reference \ System	$T_S = Guest$	$T_S = Other$	$T_S = Robot$
session 1 $T_R = Guest$	570	97	4
$T_R = Other$	98	241	19
$T_R = Robot$	6	24	83
session 2 $T_R = Guest$	1079	43	14
$T_R = Other$	81	118	62
$T_R = Robot$	10	42	257
session 3 $T_R = Guest$	353	117	15
$T_R = Other$	231	205	25
$T_R = Robot$	9	4	97
session 4 $T_R = Guest$	824	1	9
$T_R = Other$	49	5	0
$T_R = Robot$	62	1	45

Table 4.5: Confusion-matrix between hand-labeled references of the video frames (T_R) and the visual targets our Bayesian approach (system) assumes with hand tuned distributions for three targets (T_S).

and $T_S = Other$ would cease to exist. In session 1 for example 196 errors could be remedied.
But in the case $T_R = Robot$ and $T_S = Guest$ 31 errors instead of 9 and in the situation $T_R = Guest$ and $T_S = Robot$ 23 errors instead of 4 occur.

So including the frames manually labeled to *Other* can best be done by assigning them to one class together with the frames tagged as *Guest* (as done in the next section).

Visual Target Classification – *Robot* or *Other* – including all Frames

As explained in the last section – and seen in table 4.1 – almost in all cases, when the utterance is tagged as *Other*, the host speaks with the guest. Therefore we combine cases, when the utterance is tagged as *Other* or as *Guest*, into one class also called *Other* (The nomination is motivated by looking to the robot or to some other target). By doing this, an additional advance would be that the amount of target classes mirror the number of different interesting states (which is looking to the robot or not looking to the robot respectively later addressing the robot or not addressing the robot).

The results in table 4.6(B) show that the assumption[4] in the last section was correct and better results could be reached: all values increased, except the recall value.

Target =	*Robot*	*Other*
session 1	0.10	0.90
session 2	0.18	0.82
session 3	0.10	0.90
session 4	0.11	0.89
Average	0.12	0.88

(A)

Precis.	Recall	F-Meas.	Accu.
0.95	0.73	0.83	0.96
0.96	0.78	0.86	0.94
0.88	0.84	0.86	0.95
0.83	0.42	0.56	0.92
0.9	0.69	0.77	0.94

(B)

Table 4.6: A) A-priori probabilities $P(T = Target)$ of the two visual classes – *Robot* and *Other*. B) determination of the target based on visually estimated head pose (head pose estimation: section 4.3).

The results listed in 4.7 affirm the conclusion in 4.6.2 'How to include the frames looked to another target' that the best result can be reached by first taking three target classes and then combining the classes *Guest* and *Other* to one class: learning three Gaussians and then combining the accordingly two classes to one class the results are significant better than learning only two Gaussians. Note that the results with learned distributions are only slightly lower than the results obtained with true distributions.

sess.	Prec.	Rec.	F-Meas.	Accu.	Precis.	Recall	F-Meas.	Accu.
1	0.3	1.0	0.46	0.66	1.0	0.93	0.96	0.99
2	0.32	1.0	0.48	0.71	0.68	0.83	0.75	0.93
3	0.4	0.91	0.56	0.85	0.43	0.91	0.58	0.87
4	0.28	1.0	0.44	0.65	0.32	0.55	0.4	0.78
Avrg.	0.33	0.98	0.49	0.72	0.61	0.81	0.68	0.9

Table 4.7: Determination of the visual target - Robot or Guest - based on visually estimated head orientation with learned distributions with two Gaussians (left) and with three Gaussians (right).

Since the distributions can be learned automatically without supervision, the manual labels are not needed and the results for all 18 sets can be calculated. The average results for all sets were slightly worse than the average for the first four sessions. For a few sessions using only 2 Gaussians was better than using 3 Gaussians (and combining two).

[4]assumption: 'using three Gaussians and combine two for unsupervised learning is better than using two'

Summarize

Table 4.8 shows the average results of the two experiments in the four sequences. With the hand-tuned parameters, we could correctly detect the visual target in 94% of the frames. Occasions when the person was looking towards the robot could be detected 69% of time, with a relatively high precision of 90%, resulting in an f-measure of 0.77. With learned model parameters, a slightly lower accuracy and f-measure was obtained.

Distribution	Precision	Recall	F-Measure	Accuracy
True	0.9	0.69	0.77	0.94
Learned	0.74	0.85	0.79	0.93

Table 4.8: Determination of the visual target - Robot or Guest - based on visually estimated head orientation with hand-tuned (true) and learned models.

4.6.3 Estimation of the Addressee

Here we investigate if head pose (respectively the looked target) is a good cue to determine the addressee. As conducted in section 4.4, a frame or later turn is classified to the maximum of all posterior probabilities. This means, a frame is classified as speaking to the acoustical target T_A with the greater posterior probability.

$$P(T_A = Robot|X) > P(T_A = Guest|X) \ ? \ command \ : \ no\ command \quad (4.6)$$

Since our previous experiments (section 4.1) indicate that visual focus is a good indicator for the addressee of a speech act, especially if the visual target was a human, we can use the estimated visual target T_V as an estimate of the (acoustic) addressee T_A:

$$P(T_A = Target|X) = P(T_V = Target|X) \quad (4.7)$$

$$= \frac{P(X|T_V = Target) \cdot P(T_V = Target)}{P(X)} \quad (4.8)$$

where X again denotes the hosts horizontal head orientation.

$P(T_A = Robot|X)$ is then assigned to $P(T_V = Robot|X)$ and $P(T_A = Guest|X)$ assigned to $P(T_V = Other|X) = 1 - P(T_V = Robot|X)$. $P(T_V = Robot|X)$ is calculated as explained in section 4.4 – see also equation 4.1.

Table 4.9 summarizes the results of detection the addressee based on estimating the visual target as described in the previous section. Result for both, hand-tuned, true head pose distributions, as well as learned distributions and priors are given.

Distribution	Precision	Recall	F-Measure	Accuracy
True	0.61	0.83	0.7	0.93
Learned	0.6	0.8	0.68	0.89

Table 4.9: Determination of the acoustical addressee, based on head pose. Results with true head pose distributions and learned distributions of the visual target according equation 4.8.

Using the true priors and class-conditional distributions for head pose, we could identify the correct addressee 93% of time. Commands towards the robot could be detected with a recall of 0.8 and a precision of 0.6, resulting in an f-measure of 0.7.
With automatically learned priors and class-conditionals, results only slightly decreased. Here 89% of the addressees were correctly identified. Recall and precision of detecting commands towards the robot almost stayed as good as with hand-tuned model parameters.

4.7 Conclusion

As we have seen the correlation of visual and acoustical target is high, especially if the host looked at the other person (the guest). Therefore the visual target can be used to determine the acoustical addressee.

With a neural network based approach the visual target can be estimated by the visual cue head pose out of the normal video frame (without any special equipment). Finding the most likely target can be formulated in a Bayesian framework, where the goal is to find the target with the highest posterior probability given an observed head pose.

It was shown that f-measure is the value wanted to be optimized. Furthermore we investigated that learning three Gaussians (*Robot*, *Guest* and

Other) instead of two (*Robot* and *Other*) modeled the true probability distributions better and led to significant better results. (People mostly looked to the robot, the guest and the table in front of them.)
This was true for estimating the host's visual target (at whom did the host look) as well as for identification of the (acoustic) addressee (to whom did the host talk).

Chapter 5

Identification of the Addressee based on Speech

In this section we describe how to estimate the likely addressee of a speech act, based on features extracted from the speech signal. The aim is to discriminate between a *command* directed to a robot and a *conversation* between two humans. We see the identification of the addressee as one aspect of understanding the interaction between humans and robots. As a consequence we assume that speech recognition is involved in recognizing the spoken speech. Furthermore, since it is our believe that higher linguistic knowledge is useful to identify the addressee, we extract the speech based features from the speech recognizer output rather than from the raw audio signal. In the next section we first describe the extracted features, then give the main characteristics and performance of the speech recognizers, and finally present the experimental results.

5.1 Feature Extraction

The determination and evaluation of speech based features that are suitable for the identification of the addressee was first done in the Studienarbeit [Katzenmaier 2003]. There the data collection was done in German language for 3 sessions. The collection scenario (host-robot-guest) was very similar to the audio setup described above (see chapter 3), except that the recordings of both sides, the host and the guest were analyzed. In this work (Diplomarbeit)

we recorded a larger set of English speech data as well as video and transferred the findings of the Studienarbeit to the new data.

The following listing gives an overview of all features. The newly found features are cursive.

- *Feature extracted out of raw data*
  1. *Audio Length* $L(X)$ $[ms]$
- For each Hypothesis Separately
  1. Number of words $S(X) \in N$
  2. Occurrence of 'Robot' $R(X) \in N$
  3. Occurrence of Imperative[1] $I(X) \in \{0, 1\}$
  4. Perplexity
     (a) Under Command-LM $\mathrm{PP}_{com}(\mathrm{X})$ (German: add. PP_{Nav})[2]
     (b) Under Conversation-LM $\mathrm{PP}_{conv}(\mathrm{X})$ (German: add. PP_{VM})[2]
  5. *Number of words not in Language Model*
     (a) *Not in Command-LM* $UNK_{com}(X) \in N$
     (b) Not in Conversation-LM $UNK_{conv}(X) \in N$
  6. Parseability Features
     (a) *Number of Parse Trees* $T(X) \in N$
     (b) *Percentage* Parseability $Z(X)$[3]
     (c) *parse Score* $Sc(X) \in N$
- Feature extracted using both Hypotheses
  1. Correlation of Words $C_w(X) \in [0:1]$
  2. Correlation of Letters $C_l(X) \in [0:1]$
  3. *Aligning of words* $A_w(X) \in [0:1]$
  4. Aligning of letters $A_l(X) \in [0:1]$

[1] only for German language
[2] see text for details
[3] $\in \{0, 1\}$ German, $\in [0:1]$ English

As explained in subsection 5.1.4 additional language models from other projects could be used to calculate two more perplexity cues (PP_{Nav}: car navigation task, PP_{VM}: from the Verbmobil task [Verbmobil 2000]) only for the German language. The occurrence of the imperative verb form could be easy extracted only in German language, so that with 7 different kinds of cues 18 cues could be extracted. In English the new ideas brought the amount to 13 different kinds of cues, leading to a total of 23 cues.

5.1.1 Length

The selection of the first set of features was motivated by the observation that commands usually are shorter in length than conversational turns and that we expected commands more likely to contain the term 'robot' or 'robbi' to address the robot. Therefore, we took the *utterance length* $L(X)$ *in ms* and the *number of words* $S(X) \in N$ as discriminating features.

The number of words is not in relation to the absolute length, since breath and all other kinds of human noise or human fillers are not treated as words.[4] For example 'uh' and 'uhm' are more frequent in conversations than in commands. [5]
Furthermore the utterance length is derived directly from the raw data in ms and the number of words from the hypothesis. The utterance length in ms is the only feature extracted only once and extracted directly out of the audio recordings without applying a speech recognizer's hypothesis. Since the last set of features comparing the two hypotheses applies both hypotheses, it is also extracted only once per turn. All other features are extracted twice: each for context free grammar (CFG) and tri-gram language model (LM) hypothesis.

5.1.2 Occurrence of 'Robot'

In the 'Enterprise' TV show we notice that they always say the word 'computer' before addressing the machine. This would make a command detection easy, if in our case the robot is always addressed with the word 'robot' or

[4]Hesitations and human fillers are modeled separately; see later and for more details: [Verbmobil II 2001].

[5]Maybe the occurrence of these articulations would be another workable cue leading to slightly better results. This would be worth investigating.

'robbi' on the beginning. Usually we address someone by name to specify our addressee and to show our desire to start to talk to this person. Therefore the word 'robot' occurs more frequently after a long period of silence or after talking to a third party.

Unfortunately, in conversation parts, especially if talked about the robot, the word 'robot' occurs as well (– as directly tried to put into the scenario (see chapter 3)). But still the *occurrence of 'robot'* $R(X) \in N$ is a feature giving discriminative information since, most of time at the beginning of some commands, the robot is addressed by name; sometimes even twice.

As some people called the robot 'robot', others 'robbi' and even 'joe', those alias words, which we identified in the training set, were assigned to a placeholder class. An occurrence of the word 'robot' is then defined as an occurrence of one word in this class.

5.1.3 Occurrence of Imperative

This cue is one of the next set of discriminating features using the syntactical and semantical differences between commands and conversations. Commands are formulated in imperative form, and are less conversational than human-human communication. In order to capture this, we used the *number of imperatives* $I(X) \in N$ as a third feature, which could be easily retrieved for German since the German inflexion system differentiates the imperative form from others. On this account this cue is the only one exclusively used in German language.

5.1.4 Perplexity

Perplexity can be looked as the number of hypotheses with equal probability. In other words perplexity expresses the amount of hypotheses from which the final hypothesis has to be chosen. Therefore it is clear that perplexity is as lower as higher the probability of a specific hypothesis is. A low perplexity of an utterance under a specific language model (LM) would mean that the probability of this utterance under this LM is very high.

Here we used the transcribed material to build two statistical tri-gram language models, calculated over the command and the conversational sentences,

respectively. Using the fact that commands should result in a lower perplexity given the 'command LM', while the conversation should result in a lower perplexity given the 'conversation LM', we retrieved two *perplexity* $PP_{cmd}, PP_{cver} \in R$ features. Another two perplexity features were derived from applying language models trained on the German Verbmobil corpus for conversational speaking style PP_{VM} and a car navigation corpus for command style PP_{Nav}.

In figure 5.1 the perplexity calculated given the 'command LM' is plotted on the x-axis and given the 'conversation LM' plotted on the y-axis. To see that the specific calculated perplexities are a powerful feature once the speech recognition is accurate, the perplexity for transcriptions are also plotted (right). We then see that the perplexity scatter plots can be divided by a simple line and make it possible to accurately identify the addressee. The plot on the hypotheses (left) builds one cluster and can hardly be separated. If the speech recognition hypotheses comes close to the transcripts, then it is likely that these two perplexities as acoustical cues are sufficient for an accurate addressee identification.

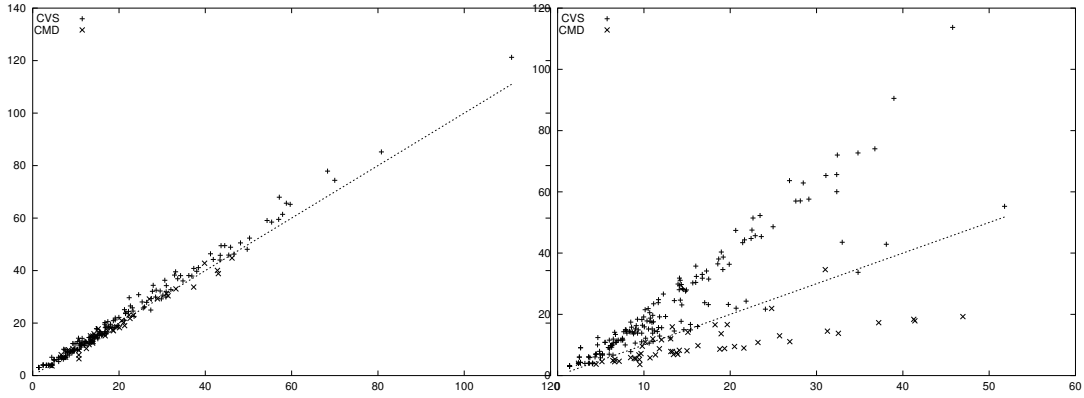

Figure 5.1: Perplexity calculated given the 'command LM' plotted on the x-axis and given the 'conversation LM' on the y-axis. On the left it is plotted for the hypotheses and on the right for the transcriptions.

5.1.5 Negative Occurrence in Language Model

By calculating the perplexity it is necessary to substitute each from the language model unknown word with 'UNK', the representative for all unknown words. But if a word is unknown to a language model, does this not give

the information that it is not common in this language model? So if it is for example not common in the command language model, then it is probably no command – and vice versa for conversations.
On this account, the number of words not in a language model – we will call *negative occurrence (in the language model)* $\in N$ – is added to the feature list.

5.1.6 Parseability Features

Parseability

The third set of features takes the sentence structure and parseability into account. For this purpose we developed in the Studienarbeit [Katzenmaier 2003] a context free grammar (CFG) designed to parse commands, and determined the boolean *parseability* $Z(X) \in 0, 1$ feature that is set to '1', if a sentence could be parsed using the CFG, and to '0' otherwise.

The in the former work [Katzenmaier 2003] used parseability through a CFG is now extended to the percentage of parseability. If only a phrase is parseable, this is getting into account as well. So if the speech recognizer recognizes one part incorrectly or recognizes even just one word incorrectly, no longer no parseability is the result. Instead the part correctly recognized and parseable is counted by taking the percentage of the utterance that is parseable.

Number of Parse Trees

Furthermore the number of parse trees is used as an additional feature. The reason is that a command should be completely parseable in one parse tree. If by any chance a conversation is parseable nearly complete, it is more likely that only phrases were parseable. If it was mostly phrases, then we would have for every phrase one parse tree and therefore a higher number of parse trees.

Parse Score

Last but not least the parse score as a parseability feature is used. It could be that this feature is in relation to the percentage parseability, but this is not investigated so far. If so, no new information is added to the neural network. However, due to the fact that the neural network is able to learn to set weights of unnecessary feature to zero, the performance might remain unharmed.

Example

```
; Parsing utt 1 (line 1)
; " <s> hello robot </s> "
; Interpretation 1.1\\
; !<s> hello robot !</s>
; Coverage 100% (2/2) in 1 tree
[generic:action,VP,\_] ( [robbi:act\_hello,V,\_] ( hello )
[michi:rob\_NT] ( robot ) )
```

Example for parsing the sentence ‘hello robot’ with 1899.88 points parse score.

```
; Parsing utt 1 (line 1)
; " <s> robot please open the window and and bring me something
to drink fork a coke please </s> "
; Interpretation 1.1
; !<s> robot please open the window !and !and bring me something
to drink !fork a coke please !</s>
; Coverage 81.25% (13/16) in 3 trees
[michi:act\_open,V,\_] ( [robbi:start\_NT] ( [michi:rob\_NT\_start]
( robot ) please ) [robbi:act\_open,V,\_] ( open )
[robbi:obj\_openable,N,\_] ( [robbi:obj\_window,N,Sg] ( the
 window ) ) ) [robbi:act\_bring,V,\_] ( bring [michi:whom,NT,\_]
( me ) [michi:sth\_to\_drink,NT,\_] ( something to drink ) )
[generic:act\_command,V,\_] ( a coke please )
```

Example for parsing the sentence ‘robot please open the window and and bring me something to drink fork a coke please’ with 12707 points parse

score. Note that the utterance would not be detected as a command with binary parseability, because of twice recognition of the word 'and' and wrongly recognition of the word 'fork'. With percentage parseability in opposite we have over 80% parseability and therefore the turn is detected as a command.

5.1.7 Correlation and Aligning

The last set of features was derived from the *correlation* $C(X) \in [0 : 1]$ between the hypotheses generated from using the different language models and the CFG for decoding. The specific design of the CFG[6] leads to a high correlation, if it is a command, and a low correlation, if it is not a command. We calculated two different correlation coefficients, one based on the hypothesized words $C_w(X)$, another one on the number of letters $C_l(X)$ of these words. The same process (word based and letter based) is done for the aligning coefficients, a better cue than the correlation – see the following examples for explanation.

Examples

The first example shows why the letter based coefficients are more powerful in some situations: The speech recognizer using the language model for decoding might i.e. recognize Robert instead of robot and the CFG decoding one might recognize robot correctly in a very small sentence like i.e. 'yes robot'. A word comparison leads to a small value (50%), but a letter comparison not (100%). The coinciding letters respectively words are not capitalized and the ones that do not coincide are capitalized in table 5.1.

Letter based	-	Word based
y e s r o b o t	-	yes ROBOT
y e s r o b e r t	-	yes ROBERT
100 % Correlation	-	50 % Correlation

Table 5.1: Example for the advantage of letter based correlation for the sentence 'yes robot', one time incorrectly recognized as 'yes Robert'. (The letters respectively words are capitalized if not coincide and not capitalized if correlated.)

[6]The CFG is designed to parse only commands and therefore we have an accurate recognition only for commands.

In case of letter comparison a high correlation of the hypothesis is quite likely if the sentence is long. This is true because correlation is realized as two directional existence. Does a specific letter (or word) exist in the other hypothesis a "1" is summed otherwise a "0". This sum is divided through the total amount of letters (respectively words). Since the order of the letters in the hypothesis and the reference do not matter for the correlation coefficient, the correlation can be high, even if the words are quite different. Such an example is shown in table 5.2 to demonstrate the disadvantage of the correlation coefficient and the reason why now the aligning coefficient is introduced: The two hypotheses are different, but still the correlation is high. The same flaw is true for the word based correlation, since a similar, accordingly longer example could be shown with the same effect. (Once again coinciding ones are not capitalized and not coinciding letters are capitalized.)

b r I N g m e a C o k e
L o o k W H a T H e b r o U g H T m e
Correlation = 64.5%

Table 5.2: Example for the disadvantage of the correlation coefficient. Even if the two hypotheses are different the correlation is high. (The letters are capitalized if not coincide and not capitalized if correlated.)

Aligning in opposite looks at one hypothesis as the reference and the other one as the hypothesis for aligning. The hypothesis is tried to convert into the reference with the lowest cost. Insertion, deletion and substitution are the only operations beside the correct match that are allowed and has its specific costs. So a letter at the end of the sentence can not be aligned to a letter at the beginning of the sentence. [7] This value of the minimal costs to convert the hypothesis into the reference is called Levinshtein Distance. Now it can be seen (table 5.3) that only e, o and e (not capitalized) are correlated in the sense of aligning – all other words are not aligned (capitalized in the table). This leads to only 25% correct aligning in opposite to the 64.5% correlation and the utterance would be rejected.

[7] The name 'align' is well chosen, since you can look at this process by putting each hypothesis in one line and than try to "align"/ lining up these two lines.

REF: B R I N G M * * * e A C o K * * * * e
HYP: L O O K W H A T H e B R o U G H T M e
Correct = 25.0%
Errors = 133.3%

Table 5.3: Aligning solves this problem of association of a letter or word at the end of the sentence to the beginning of the sentence and has a low correlation in the example, where the correlation coefficient was high. (The letters are capitalized if they do not coincide and not capitalized if they coincide.)

5.2 Feature Evaluation on German Language

We evaluated the features in the Studienarbeit [Katzenmaier 2003] by conducting discrimination experiments using both, the transcribed references and the corresponding first best hypotheses output from the speech recognizer.

We furthermore investigated several classification methods, simple comparison, Bayesian classification, and Multi-Layer-Perceptrons. The results in table 5.4 show that the combination of the above mentioned speech based features outperformed the single features. Here a summary of the results are given, see the Studienarbeit [Katzenmaier 2003] for more details.

In figure 5.2 the best result for each approach is plotted:

- Two simple comparison approaches: correlation of the two hypotheses against a threshold (Corl) with 80% accuracy and command language model (LM) perplexity against conversation LM perplexity (Com-PP) with 74% accuracy
- Bayesian approach (Bayesian) with 80% accuracy
- Multi-Layer-Perceptron (MLP) with 82% accuracy and MLP on transcripts (Trans MLP) with 87% accuracy

Also the specific guessing line for this data is plotted (all points on this line with 50% accuracy can be reached). Since the aim is to reach 100% recall and precision, the f-measure can be looked at as the distance from this point. (See also the section 4.6.1 'Used Metrics')

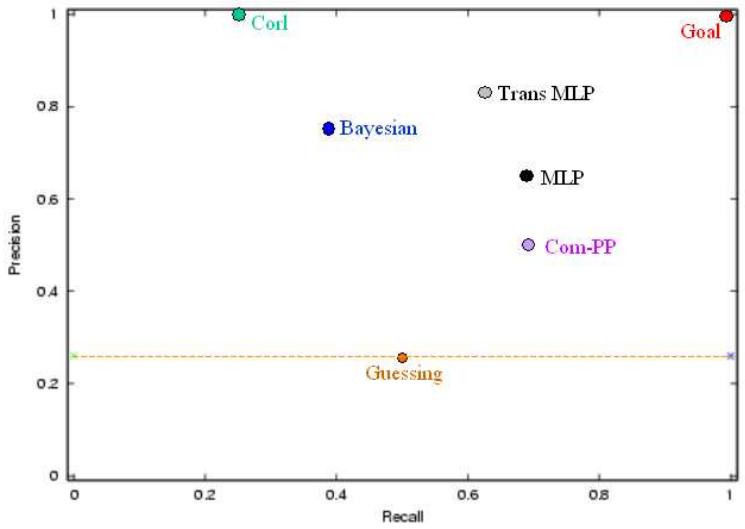

Figure 5.2: Classification method evaluation on the German task. Only the best result for each method is plotted.

The overall best result could be achieved with a simple feed forward MLP with differentiable activation function, one hidden layer, trained using gradient descent on the transcripts giving 87% accuracy, a recall of 63%, a precision of 83%, which leads to an f-measure of 0.72 [Katzenmaier 2003]. These results are encouraging and indicate that speech-based identification of the addressee is possible. However the overall performance in this Studienarbeit suffered from data sparseness and the poor performance of the speech recognition.

To overcome the data sparseness problem we collected a larger data set for the English data collection (see above chapter 'Data Collection'). In addition, a better baseline speech recognizer was applied, which is described in the following section.

5.3 English Speech Recognizer

The baseline English speech recognition system used in this work was trained on the Switchboard corpus. The fully continuous HMM-based system uses 2000 context-dependent acoustic models with a mixture of 16 Gaussians per model. Cepstral Mean Normalization is used to compensate for channel variations. In addition to the mean-subtracted mel-cepstral coefficients, the first

Used Features	Precision	Recall	F-Measure	Accuracy
Bayes on Hypotheses				
$PP_{VM}(X), PP_{Nav}(X)$	0.75	0.19	0.30	0.77
+ S(X), Z(X)	0.75	0.38	0.50	0.80
+ $PP_{com}(X), PP_{conv}(X)$	0.47	0.56	0.51	0.72
MLP on Hypotheses				
$PP_{VM}(X), PP_{Nav}(X)$	1.0	0.12	0.21	0.77
+ S(X), Z(X)	0.56	0.56	0.56	0.77
+ $PP_{com}(X), PP_{conv}(X)$	0.65	0.69	0.67	0.82
+ C(X)	0.43	0.81	0.56	0.67
MLP on Transcripts				
4 x PP(X), S(X), Z(X)	0.83	0.63	0.72	0.87
+ R(X) + I(X)	0.90	0.56	0.69	0.87

Table 5.4: Feature set evaluation on the German Studienarbeit corpus. Only the first best results are listed. Note that they are all using multi features.

and second order derivatives are calculated. Linear Discriminant Analysis is applied to reduce feature dimensionality to 32. The recognizer runs in near real time. In these experiments we customized the vocabulary, dictionary, and language models of the recognizer towards the given task using the transcribed data described in section 3, however we did not re-train or adapt the acoustic models.

The context free grammar was manually created such that commands from all training data could be completely parsed. The CFG-based recognizer handles filler models to treat hesitations and non-verbal noises during decoding. That is why the CFG allows insertion of filler words of any given position and no further handling is needed to cope with such spontaneous effects in the training data. In addition to the training transcripts we collected 425 commands from 8 people to improve the CFG. In total the context free grammar for command parsing consists of 264 rules using 3162 nodes and 4638 arcs based on 434 terminals. When building it on all transcripted commands including the 425 additional collected ones, 276 rules and 4650 nodes with 632 different terminals respectively vocabulary size.

The statistical n-gram language model was trained on roughly 3 Million word tokens taken from the English Switchboard [Soltau et al. 2003] data, interpolated in a relation of 1:130 with the transcriptions of the collected data.

Table 5.5 shows the performances of the English speech recognition on the dif-

ferent data sets and the various customized systems. It lists the performance of the context free grammar based decoder and compares the statistical language model based speech recognition systems.
The fact that all systems not including the tested data to build the LM have 9% out of vocabulary (OOV) rate shows that people speak at least 9% (likely more) about different topics in the set tested than on that trained on. – Some spoke i.e. about their jobs or when to meet for swimming and so on.
Note that if only the training data is taken to build the language model, only a 67% to 65.9% word error rate (WER) could be reached. Even by taking all data except the one tested on, did not lead to accurate results. This shows the last two listed experiments (only 55% and 57% accuracy).

The context free grammar (CFG) based decoder –remember that it was designed to parse only commands– had only poor recognition on the commands in the development data set, when building it only on the in the training set occurring commands (65.9%). Taking all commands in the whole data set to build the CFG was leading to a recognition giving a good platform (39.7% up to 21% WER) to perform the experiments to identify the addressee. This shows that the English task is more challenging than the German one. Note also that in German language every session is recorded with the same two participants.

The best performance on the evaluation set with the n-gram based recognizer could be achieved by taking all data to build the language model and the vocabulary. It resulted in a Word Accuracy of 83% respectively 17% Word Error Rate.

The latest results of the used IBIS Janus decoder on Switchboard data are reported in [Soltau et al. 2003]. In our experiments we used the same decoder as used in the Switchboard project. The acoustic models were originally trained on the Switchboard task, however, the model size was significally tuned down to provide a smaller and faster system for good turn around time. Without any tuning towards our data, this system gave a 68.6% word error rate (WER) on the training set and 67.0% WER on the development set. On the original data a WER of 43.8% was reached.

	LM		CFG
Build on	Basel. (SWB)	Train Data	Train Cmds
Vocab.	1045	1045	434
	WER(pp/OOV) [%] [%]	WER(pp/OOV) [%] [%]	WER [%]
Train.	68.6 (116/0)	43.0 (14/0)	34.9
Devel.	67.0 (85/9)	67.0 (139/9)	65.9
Build on	SWB + Train	SWB + all own	all own Cmds
Vocab.	1165	1720	632
Train.	31.2 (13/0)	38.1 (5/0)	39.7
Devel.	78.7 (68/9)	19.5 (5/0)	21.0
Eval.	-	16.7 (5/0)	43.0
Build on	total - Devel.	total - Eval.	
Vocab.	1720	1720	
Devel.	57.0 (159/0)	-	
Eval.	-	55.0 (127/0)	

Table 5.5: Performance comparison for English speech recognition in word error rate (WER). Only on Switchboard (SWB) corpus, on training data (Train), all own data (all own) and all available data (total) exept the accordingly tested one (Development or Evaluation) for language model (LM) and context free grammar (CFG).

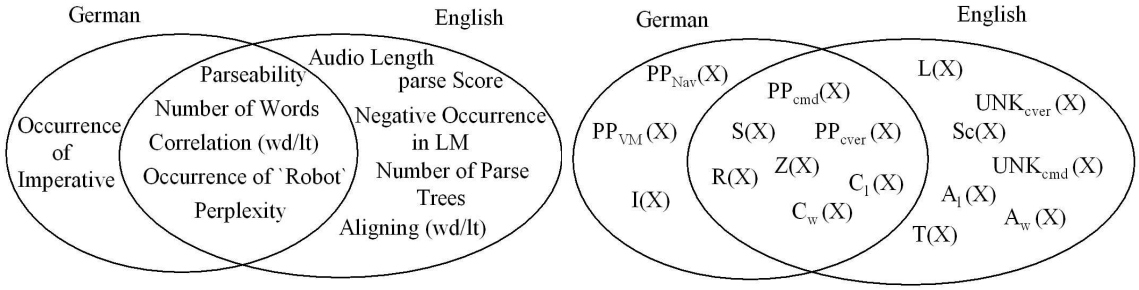

Figure 5.3: Overlapping and differences in feature sets for English and German language. On the left the kind of features are shown and on the right the finally different taken features.

5.4 Feature Comparison between German and English Language

In figure 5.3 you see that we basically took the same features and added some newly found features. On the left the different kind of features are listed and on the right the different finally taken features. As already mentioned the occurrence of imperatives is only applied for German language, since it is easy to recognize on the German reflection verb form. In German language there were also two corpora of other projects useful to build two additional language models for applying the perplexity: one for modeling commands (car navigation corpus) and one for modeling conversational speaking style (German Verbmobil corpus) – see above subsection 5.1.4.
Enhanced for example is the parseability feature $Z(X) \in [0 : 1]$ which is no longer a Boolean variable but expresses the percentage of parsed output. The number of parse trees and the parse score is therefore added to the list of acoustical cues.
The correlation is advanced and further developed in the feature aligning, but the correlation itself is still applied.
The last added cues are negative occurrence in language model and the audio length in ms extracted out of the raw data. – See subsection 5.1.6 for explanations and details of the respectively feature.

In figure 5.4, 5.5 and 5.6 for both languages common features are opposed graphically. Figures 5.5 shows that the absolute difference of the percentage values for conversations and commands are approximately the same for both languages. Only that the percentage for English language is higher in general. This is because people speak more about the features of the robot and its

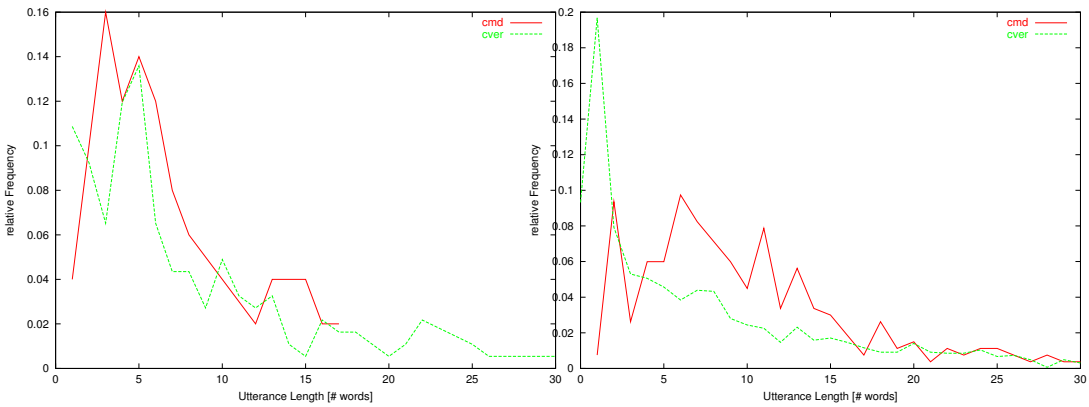

Figure 5.4: Turn length distribution in [# words] for German language (left) and English language (right). As seen most of the conversations (cver) are short and overall they have a wide range. Commands (cmd) in opposite have a small range with a peak around 5 words.

possibilities in the conversations and then try to make it clear, when speaking to the robot. So they use the word 'robot' more in both, in conversations by speaking about it and in commands by addressing (the robot) to signalize the change of addressee. We see that the higher values are due to changed behavior and the features themselves are comparable for both languages. The differences in the features listed in figure 5.6 result from the fact, that in English data the word accuracy of the speech recognition was better than in German speech recognition.

As table 5.6 shows (and graphically in the figures 5.4, 5.5 and 5.6 visualized), the features are comparable in both languages. In figure 5.4 you see the turn length distribution in [# words] for German language (left) and English language (right). As seen most of the conversations (cver) are short and over all they have a wide range. The distribution reminds on the distribution of $\frac{1}{x}$. Commands (cmd) in opposite have a small range with a peak around a number of 5 words and the distribution looks more like a Gaussian. Commands are rarer, so to compare commands and conversations the distribution of the commands should be thought lifted up. Once again it can be seen that most of the conversations are short as the high peak shows. After the high peak at the beginning the distribution for conversations is seceding. Note that in both languages the average length for conversations is higher than the average length for commands –see also figure 5.5(C). On the English collection people in general were using longer sentences for conversation as well as for commands. This can be seen on the higher average values and on the significant wider range.

The occurrence of ‘robot’ has similar range for both languages: Only one occurrence more in the extreme case each for conversation and commands in English data compared to the German data. The average value is higher for commands in both languages. Visualized is this in figure 5.5(A).
For the parseability cue it can be seen that a lot more sentences are parseable in English language. This is true because of the change to percentage parseability to cover more commands. This worked well since 98% of the commands are parseable. For that reason the parseability has a range on English language and on German data not (there it is a binary value: parseable or not parseable). Therefore there is listed the total amount of parseable and not parseable turns. The average parseability relation between commands and conversations is similar: 1:4.3 (0.13:0.56 German) and 1:3.5 (0.28:0.98 English) respectively also binary 1:5 (0.19:0.95). (graphically shown in figure 5.5(B))
The differences in the correlation coefficients between both languages is true to better speech recognition in English language as well as more phrases occur in both conversations and commands on English data set than on the German–see the example at the end of section 3.1. The relation between range and average of both languages behave similar for letter and word based correlation, except that the letter based once are higher.

Since the acoustic cues behave similar in both languages as just discussed, we transferred our results on German language to English language in the following experiments. We adapted the best classification scheme, Multi-Layer-Perceptron and applied basically the same features, except some enhancement of the already researched features as well as additional newly researched features. As explained above, taking more features is never a disadvantage, but only leads to no advantage or even better results.

5.5 Feature Discussion

The features range and average in both languages are listed more detailed in table 5.6. There you see that the perplexity for the two specific build language models for the conversational parts and commands do not differ that much between conversations and commands in English language. This proves once more that by speaking more about the robot, like in a demonstration task, makes the identifying of the addressee more challenging as accomplished for the English recording sessions: The phrase ‘robot, bring me something to

On Transcripts				
	German		English	
	Avrg.	Range	Avrg.	Range
Occurrence 'Robot' R(X)				
cver	0.16	0-3	0.14	0-4
cmd	0.32	0-2	0.77	0-3
Utterance Length S(X) [# words]				
cver	7.84	1-43	10.63	0-152
cmd	6.08	1-17	10.36	1-56
Parseability Z(X)				
	yes no	Avrg.	Range	Avrg.
cver	23 161	0.13	0-1.0	0.28
cmd	28 22	0.56	0.08-1.0	0.98
Perplexity (Conv. LM) $PP_{cver}(X)$				
cver	12.34	1-52	31.94	1-443
cmd	16.64	2-47	31.86	3-167
Perplexity (Com. LM) $PP_{cmd}(X)$				
cver	21.58	3-114	32.21	3-432
cmd	10.46	4-35	33.01	4-183

On Hypotheses			
German		English	
Avrg.	Range	Avrg.	Range
Word based Correlation $C_w(X)$			
0.01	0-0.34	0.24	0-1.0
0.03	0-0.22	0.69	0-1.0
Letter Based Correlation $C_l(X)$			
0.32	0-0.63	0.67	0-1.0
0.38	0.16-0.72	0.89	0-1.0

Table 5.6: Feature Comparison between German and English Language for human-human conversations (cver) and human-robot commands (cmd).

drink' is now as probable in a conversation as in a command. [8] The range for conversations shows that a maximum value of 443 occurs and therefore gives an additional indication that some very improbable parts in the conversational part intent. And thereby conforming the fact that people speak about totally different topics in the human-human conversational part on the English data set and with that making the recognition and identification challenging.

Note that the parseability for some conversations in both languages are 100%, although the CFG is specific designed to parse only commands. This occurs, since turns like i.e. 'thank you' or 'thanks' are sometimes addressed to the robot as well as sometimes to the guest. Therefore although possibly addressed to the guest and so as a conversational part parseable. Sometimes a word or human noise is recognized incorrectly, so that the hypothesis of the

[8] as the example in section 3.1 showed

turn is 'thanks', but truly it is something else. Then the turn is parseable also, even if it is a conversation. That a conversation is 100% parseable occurs in 395 cases out of the 1641 conversational turns in the English data (24%). But not parseable at all are 593 (36%). On the other hand 0% parseable of the commands are 9 from 267 command turns (3%) and 100% parseable are 178 (67%). So over all percentage more commands are parseable (67%) as conversations (24%). On the other hand percentage less commands are not parseable (3%) as conversations are not parseable (36%) – even the absolute amount of not parseable commands are lower than the amount of not parseable conversations (commands 178 and conversations 593) – see table 5.7 below.

conversations	0% parseability	36% (593 of 1641)
	100% parseability	24% (395 of 1641)
commands	0% parseability	3% (9 of 267)
	100% parseability	67% (178 of 267)

Table 5.7: Compare of parseability coefficient between commands and conversations percentage and in absolute terms on English data.

The correlation coefficient behaves similar to the parseability as can be seen in table 5.8.

conversations	0% correlation	43% (709 of 1641)
	100% correlation	11% (180 of 1641)
commands	0% correlation	11% (29 of 267)
	100% correlation	43% (115 of 267)

Table 5.8: Compare of word based correlation coefficient between commands and conversations percentage and in absolute terms on English data.

It is also outstanding that the average utterance length of commands is only slightly higher than for conversations. The reason for the relatively low average turn length for conversations and in the other way relatively high average turn length for commands for especially the English data is the following: On the one hand we have all the single words of the effect to signalize someone's agreement or listening, as: 'uh-hum', 'yeah' and so on (so-called back channel) by the conversations. By the commands on the other hand especially in the English data collection people sometimes give a whole list of commands in one turn for the robot to do: do this, do that and on the way back that and when you are already at this destination, do that also. But these are most of time only exceptions and commands are not that long. It is the other

way around for the conversations, where the feedback channel is a behavior rule. Therefore we can summarize that conversations most likely are even longer or shorter than commands.

As a conclusion we expect that using two thresholds instead of one increases the result. For the neural network approach this could be simulated by also using the – we will call – *contra value* ($= 1 - value$) as an additional input of the input vector. An additional advance would be that in cases when the value was zero, the neural net did not use the information given, since the product with the weight would be zero, too. But now this information is in the contra value and specific new weights can be learned for the "second threshold". This idea can be looked at as cutting out the peak of the command curve in the feature curve as for example in figure 5.4 and therefore promises to be applicable for other features as well.

The in this section mentioned cue properties were the same for English and German language except for the correlation. This shows again that the features are comparable in both languages and therefore the results are transferable.
0% word based correlation in *German* language occurred in 173 of 184 and a higher correlation than 10% only 6 times for the conversations. For the commands 0% correlation is calculated 42 times of the total of 50 commands, but all remaining 8 commands had a correlation higher than 10%. With similar letter based correlation we see that by poor speech recognition, the correlation coefficients are only an excluding feature: significant more percent of conversations have very poor correlation (and only some commands have very high correlation, but most of the commands have at least little correlation).
Note also that in German language there were 50 commands and 184 conversations, so that approximately every fifth utterance was a command (21%) and in the English data set only 267 commands against 1641 conversations, one third less percent commands than in German language (14%).
The command detection in the German language is obviously not as difficult as it is in English, since there are significantly less percent commands, additionally people speak more spontaneously, have more false starts and repeatings and talk partly about new topics.

5.6 Results on MLP approach for English Data

As command detection is more difficult on the English data set the result on the Multi-Layer-Perceptron approach is not as good as the MLP approach on the German data set: On the English data set we achieved a recall of 91%, however only 49% accuracy and a precision of 19% could be reached which leads to 0.31 f-measure. We remember recall gives the detection rate, precision the fidelity of detecting commands and f-measure the combining of both values in the geometrical meridian for comparison as mentioned in subsection 4.6.1.

To 'force' the neural net to learn something the rare commands are put in the pattern file more than once. – (Without that, the net always 'learned' to never detect a command, since the accuracy was then relatively high due to the rare occurrence of a command in compare to a conversation.) 4,6,8,10,12,15 and 20 times taken the commands into the pattern, with learn rate 0.1, 0.01, 0.005 and 0.001 and different amount of hidden units 10, 20 and 30 are conducted. (Here also a Multi-Layer-Perceptron as net structure is used.) The best result reported above could be reached with a learn rate of 0.01, 10 times commands and 20 hidden units – see table 5.9. It is also notable that even by apparently bad only acoustic based results, the discriminative power became visible by combining it with the head pose based results –see chapter 6 especially table 6.1 the last two entries. More falsely detected commands could be rejected without any further rejection of correctly detected commands. This was especially interesting in one case as the acoustic results were really poor with 0 recall, precision and f-measure. The acoustic based cues also are increasing the overall performance.

# Hid. Units	x Cmds	Learn Rate	Precis.	Recall	F-Meas.	Accu.
10	8	0.1	0.09	1.0	0.17	0.88
20	10	0.01	0.16	0.89	0.27	0.40
30	8	0.01	0.19	0.91	0.31	0.49

Table 5.9: Acoustic results on English data with Multi-Layer-Perceptron. The number of hidden units, the number of times the commands are taken in the pattern file (x Cmds) and the learning rate is listed for a selection of experiments.

Note that on the German data the two recorded persons are the same for all three sessions and they concentrated on minimizing the effects of spontaneous

speech like: repeatings, false starts, hesitations ('uhm','uh',...), human noise (breath, throat,...) and topic changing. Now no instructions to the recorded persons are given to restrict these effects – see chapter 3. This shows that if people speak less spontaneously, speech recognition and identification of the addressee is easier.

The listed results in table 5.9 show that with more hidden units a better result on the only on acoustic cues based experiments is reachable (higher f-measure).

5.7 Conclusion

It was shown that the features extracted on English and German data coincide in the relation of range and average between conversation and commands relatively and therefore the results from German research study can be transferred to the English data. Furthermore most of the features are more powerful when the speech recognition is more accurate and therefore depend on accurate recognition.

The speech recognition and addressee identification for the English data set was more difficult than for the German data set, since the English task was more challenging.

It was constituted that the introduced *contra value* promises to improve the performance in future work. Furthermore it would be better to use more data for training as for example using 16 sessions for training the neural net and one session each for testing and evaluation.

The [Katzenmaier 2003] was the first system identifying the addressee only speech based and showed that identification based only on speech is possible. It worked accurate speaker dependent, as seen on the German data: trained on the data of two different persons and tested on the data of one of these two participants was possible with 82% accuracy, 65% recall and 69% precision resulting in an f-measure of 0.67. [9] These results show that building a system to identify the addressee only based on speech for daily use from the same user is possible.

On the English data with different recorded persons in every session, we have

[9] with only two same persons recorded data, commands recur more frequently

seen that the neural network approach also works speaker independent.

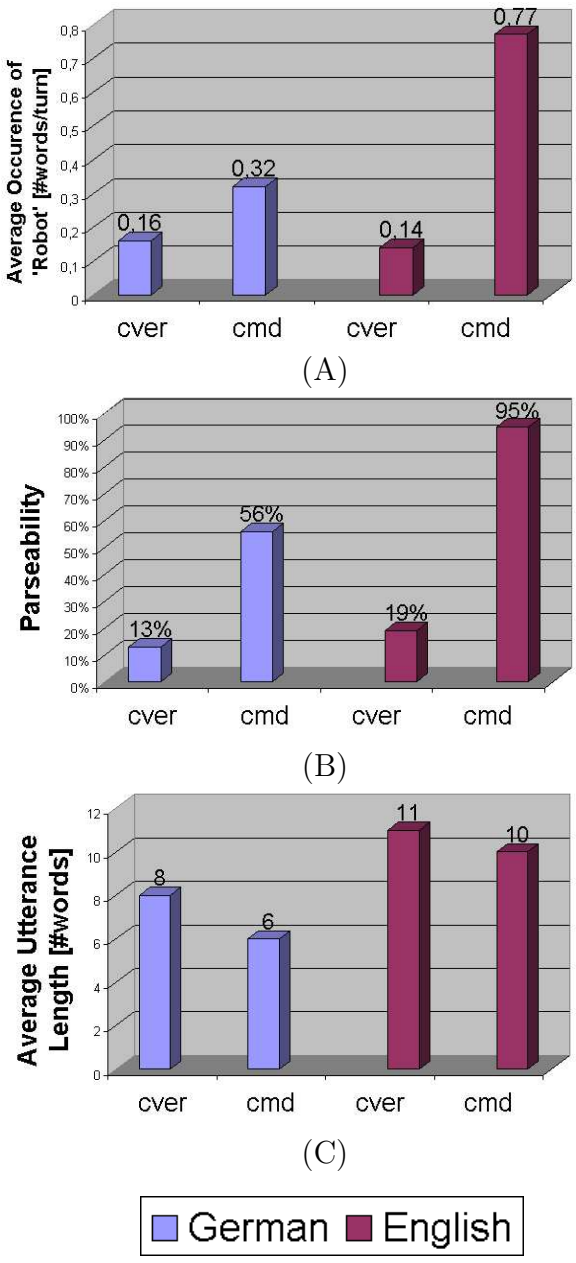

Figure 5.5: Graphical presentation for feature comparison between German (left two bars) and English language (right two bars) for human-human conversations (cver) and human-robot commands (cmd). (A) Average Occurrence 'Robot', (B) binary Parseability and (C) Average Utterance Length.

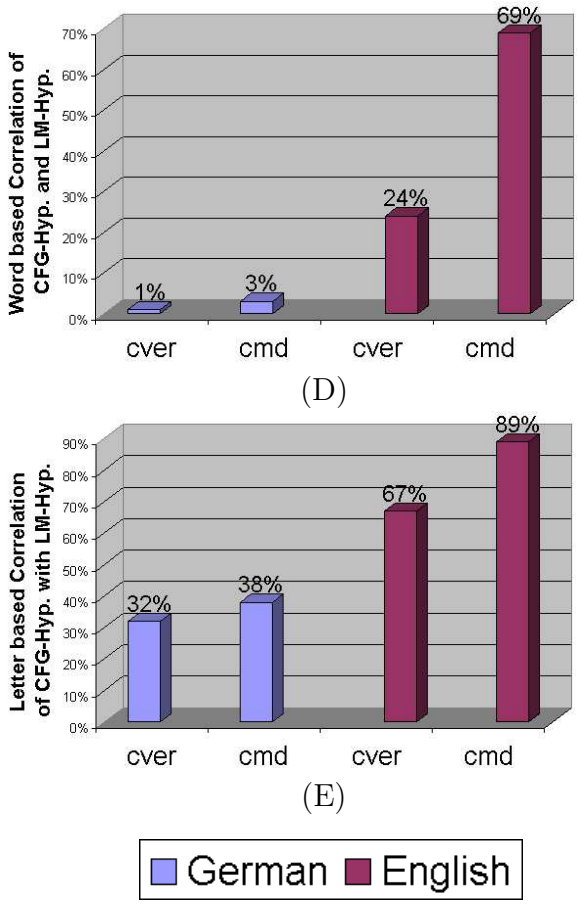

Figure 5.6: Graphical presentation for feature comparison between German (left two bars) and English language (right two bars) for human-human conversations (cver) and human-robot commands (cmd). (D) Word based Correlation and (E) Letter based Correlation. Differences due to better speech recognition on English data than on German data.

Chapter 6

Combined Experiments

In the previous sections we have discussed how the addressee of a speaker can be estimated based on his or her head orientation and based on acoustic cues. Both approaches resulted in posterior probabilities for the possible addressees, either given the head orientation cues - $P(Addressee|HeadPose)$ - or given the acoustic features as input, $P(Addressee|Speech)$.

Improved classification results can be achieved by combining the two classifiers. To this end we computed the weighted sum as well as the weighted multiplications of the two posterior probabilities:

$$P_{Sum} = \alpha \cdot P(Target|speech) + (1 - \alpha) \cdot P(Target|Head\ Pose) \quad (6.1)$$

and

$$P_{Mult} = P(Target|speech)^{\alpha} \cdot P(Target|Head\ Pose)^{(1-\alpha)} \quad (6.2)$$

respectively.

We observed slightly better combined estimation results by using the weighted sum compared to multiplying the weighted probabilities. Figure 6.1 shows a plot of the results when calculating the weighted sum of the probabilities and changing the weight α between zero and one ($\alpha = 0$ corresponds to using only head pose, $\alpha = 1$ corresponds to using only speech). The values for precision, recall, f-measure and accuracy are plotted. The best f-measure was obtained by setting α to 0.7, resulting in an estimation of the correct addressee 92% of the time and a detection of commands towards the robot with precision of 0.65 and recall of 0.81 (f-measure = 0.72).
Note that the better head pose probabilities affect the result more than the

acoustic based probabilities; even until a value of α from 0.9. By an α from 1.0 no head pose probabilities are getting into account anymore and the combines result mirror the worse only acoustic based results.

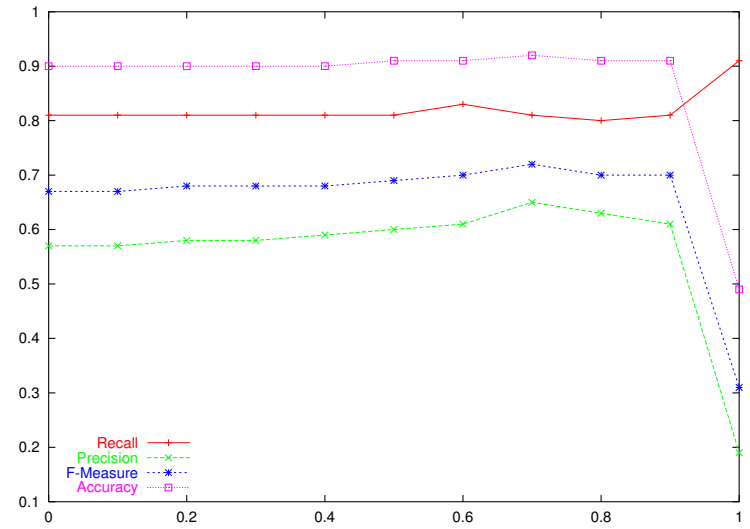

Figure 6.1: Combined estimation results with different weights α (see text). Indicated are accuracy, recall, f-measure and precision (top to bottom).

Figure 6.2: Flowchart for 2-step combination.

Beside weighted sum and power another combination variant was implemented. The flowchart is shown in figure 6.2. First only the head pose is conducted and the classification performed. Then if and only if assigned as command, the acoustic cues are combined and the finally result calculated. This approach takes the fact in account that the guest if looked at was also addressed. Only in case the robot was looked at the discriminative power of head pose was worse and therefore only in this case the combination with the acoustic cues are applied. The result was the same as with weighted sum combination. It can be explained by the fact that the (head pose based) probabilities for command respectively conversation differed a lot in case of the assigning to conversation, but not for commands. On this account the

influence of the acoustic cues were only strong enough in case of questioning the assigning to the class commands.

The last two listed results listed in 6.1 show that even with poorly acoustic results the combined results can be improved. Note that even with apparently bad results of an f-measure of 0.0, the additional information of the acoustic cues help to improve the head pose based experiments. Also outstanding is that the best combined result is reached with the not best acoustic result. The reason for the bad acoustic results was the poor recognition of commands (recall) and not the precision. It was very precise, if it recognized a command. So very accurate probabilities of the acoustic based experiment are combined. If the acoustic based one said it is probable a command, it was and so the combined experiment could be improved in that cases without getting more false alarms.

Combination of acoustic and visual estimation improved the results significantly compared to using only visual or acoustic information: the relative error reduction for estimating the addressee is 20% (10% error with visual estimation, 8% errors combined). In addition, the precision when detecting commands towards the robot could be improved from 0.57 to 0.65 (19% relative improvement) – see table 7.1.

	Precision	Recall	F-Measure	Accuracy
Head Pose	0.57	0.81	0.67	0.9
	# Hidden Units	x Commands	Learning Rate	α
	10	8	0.1	0.8
Acoustic	0.09	1.0	0.17	0.88
Combined	0.68	0.8	0.74	0.93
	# Hidden Units	x Commands	Learning Rate	α
	20	10	0.01	0.6
Acoustic	0.16	0.89	0.27	0.40
Combined	0.66	0.78	0.72	0.92
	# Hidden Units	x Commands	Learning Rate	α
	30	8	0.01	0.7
Acoustic	0.19	0.91	0.31	0.49
Combined	0.65	0.81	0.72	0.92
	# Hidden Units	x Commands	Learning Rate	α
	30	4	0.1	0.5
Acoustic	1.0	0.04	0.08	0.88
Combined	0.7	0.8	0.75	0.93
	# Hidden Units	x Commands	Learning Rate	α
	20	8	0.1	0.4
Acoustic	-	0.0	0.0	0.87
Combined	0.65	08	0.72	0.92

Table 6.1: Acoustic results on English data with Multi-Layer-Perceptron. The number of hidden units, the number of times the commands are taken in the pattern file (x Commands) and the learning rate is listed for a selection of experiments. The last two experiments clearly show, that the acoustic gives helpful information to increase the combined result in compare to using only head pose as listed in the first row, even if the acoustic results itself were real bad. The α value leading to the best weighted sum combination is listed too.

Chapter 7

Conclusion and Future Work

In this work we investigated the power of acoustic and visual cues to identify the addressee in multi-party communication between two humans and a simulated robot.

First, we investigated the correlation between the addressee of a speaker and the user's head orientation: We found that looking towards another person is a very reliable indicator that the other person was addressed. In fact in 99.5% of the cases in our data when a person looked towards the other person while speaking, the person was really addressing the other person.

Looking at the robot, however, could not be used as such a clear indicator: Here, in 65% of the cases when a person looked at the robot while talking, the robot also was addressed. In the remaining 35% of the cases, however, the person was indeed talking to the other human while looking at the robot.

We then investigated how well the addressee can be determined based on visually estimated head pose of the speaker. We employed a neural network based approach to estimate a person's head pose and then use a probabilistic model to find the most likely visual target of the speaker. With this approach and automatically learned priors and class-conditional distributions for a person's head pose, we could correctly identify the *visual* target in 93% of the frames on four recordings. *Looking* towards the robot could be detected with precision of 0.74 and a recall of 0.85.

By using the estimated visual target to determine the *acoustic* addressee of the speaker, the correct *addressee* could be identified 90% of time. Speech

commands towards the robot were (visually) detected with a precision of 0.57 and a recall of 0.89, resulting in an f-measure of 0.67.

We also investigated the power of using cues that are automatically derived from the speaker's speech to distinguish between *conversations* between two humans and *commands* directed to the robot.

On the German data set the usefulness of various features that were derived from the hypothesis of a speech recognizer were investigated in the Studienarbeit [Katzenmaier 2003]. These features included sentence length, the number of imperatives, the perplexities on different language models as well as the parseability of a sentence by a grammar for commands. Best classification results where obtained using a Multi-Layer-Perceptron as classifier.

A similar set of speech-related features was then also used to discriminate the addressees on our English data set that was collected for the Diplomarbeit. On this data, 49% of the utterances could be correctly classified solely based on speech related features. Commands towards the robot were detected with a recall rate of 0.91 and a precision of 0.19.

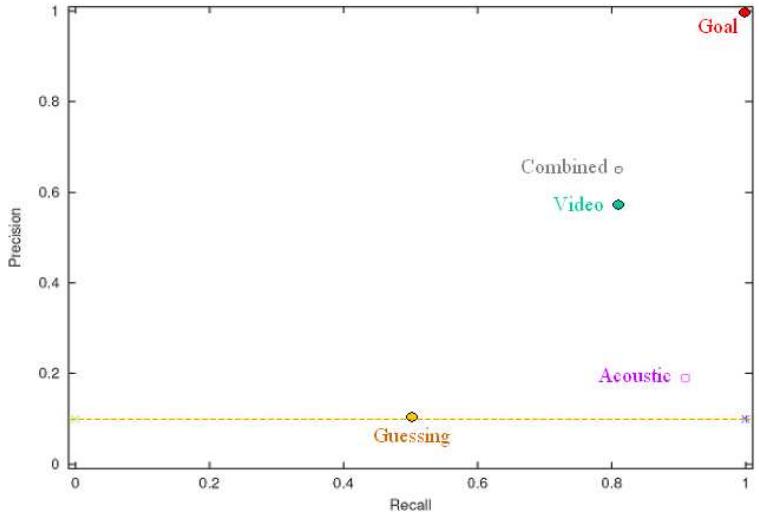

Figure 7.1: Acoustic, visual and combined estimation of the addressee in compare to guessing and the goal of perfect estimation. The recall value is plotted on the x-axis and the precision value is plotted on the y-axis. The f-measure can be looked at as the distance to the lower left corner.

Finally we combined the poorly speech-based and head pose based approaches for discriminating the addressees. This resulted in significant improvements,

despite the comparably worse results of the speech-based discrimination approach: By using the weighted sum of the acoustic and visual posterior probabilities for the addressees, correct classification of the addressee increased from 90% (visually estimated) to 92% accuracy. Furthermore the precision of the detection of commands towards the robot was improved from 0.57 (visually) to 0.65 (combined), while keeping a recall rate of 0.81 – see table 7.1.

Estimation	Precision	Recall	F-Measure	Accuracy
Acoustic	0.19	0.91	0.31	0.49
Head Pose	0.57	0.81	0.67	0.90
Combined	0.65	0.81	0.72	0.92

Table 7.1: Best acoustic and best visual results and their combined estimation of the addressee.

In figure 7.1 the best result for each experiment is plotted. The recall value is plotted on the x-axis and the precision value is plotted on the y-axis. As the f-measure is the geometrical meridian, it is related to the distance to the goal. So as closer the plotted result to the goal as higher is the f-measure. On this account it can be seen that the acoustic experiments alone are significant better than guessing. Guessing the addressee of all turns would lead to a recall of 0.5 (half of the commands detected) and since the data contains only 10% commands, the randomly as commands assigned turns comprehend only 10% commands as well. Consequently the precision is 0.1, which leads to an f-measure of 0.17 by an accuracy of 0.5.

Every point on the line through the guessing point can be reached by choosing the amount of data randomly assigned. – The not randomly assigned data would be strictly assigned to one class i.e. a command.
In an extreme case all turns would be assigned to one class. Assigning all to commands would on this data give an accuracy and precision of 0.1 and all commands would be detected (1.0 recall), which concludes in an f-measure of 0.18. And assigning all to the conversation class would lead to a high accuracy of 0.9 by a poor precision of 0.1 as a limit in this extreme case. None commands would be detected (recall 0.0) and an f-measure of zero would be the result.

91% of all commands could be detected correct with acoustic alone by a precision of 0.19 and this results in an f-measure of 0.31 by an only slightly worse accuracy of 0.49 and shows that command detection based only on acoustic cues is possible. Furthermore the plot shows that combining the

acoustic result with the better video results improved the over all result in evidence.

Appendix A

Record Guide

Remember we tried to provoke a challenging task by giving instructions to speak about the robot and to discuss its features. The record guide given to the participants is shown here:

My name is Michael and I am a visiting researcher from Germany working with the Language and Technology Institute (LTI). I am working on my final project for my Masters' degree in computer science here on CMU. To this end, I require your assistance in making some recordings; my hope is that this will be as entertaining as it is educational. Here is what we will be doing:

The Scenario: Imagine the LTI lab is your living room, and you have recently purchased a robot. The robot looks like a human being (we will be simulating this with cameras, which should resemble the eyes of a human, and a microphone, which will substitute for human ears). The robot can do many things, such as fetch drinks, vacuum your living room, adjust the lights, manage the stereo and TV, etc. You're excited to hear the doorbell ringing, as a visitor has just arrived, a good friend of yours. What an opportunity! Now you can brag about your new robot and show your friend what the robot can do!

Keywords & Guiding:

- introduce/ welcome each other
- themes to talk about with your guest:

- what the robot can do
  - get beer, soda, coffee, tea, candies, cake, newspaper, shoes, ...
  - clean up the table, clean the floor (using a wet rag and a vacuum cleaner), clean the windows, ...
  - turn on/off the lights, dim/brighten the lights,
  - increase/decrease temperature, change air humidity
  - turn on/off music, change settings of equalizer, change loudness ('turn down the volume', ...)
- the robot is good at:
  - monitoring your house
  - helping disabled people
  - helping the elderly
  - helping with cooking
- the cons of a robot:
  - getting in the way of foot-traffic
  - it still has some bugs

- what to say to the robot
  - see what the robot can do

examples (not absolute necessary to read!):

I

- You: hey, hello Michael, how are you (doing)?

o Your guest: Uh, hello <your name>. I am fine, thanks. How are you?

- You: Fine. Thanks.

o Your guest: I thought I come and visit you. Some news?

- You: Yes, come in. I have got a robot a couple of weeks ago.

o Your guest: Uh, interesting. A real robot? What can it do?

- You: It/He can do all sorts of things. He can bring me beer and whatever. All what you can imagine or so. Clean windows, cut the grass, clean up, vacuum...

o Your guest: And when you ... when you call him, does he come?

- You: Yes, sure. I can show you.

- You: Robbi, robbi come here!

o Your guest: Uh, he comes what's interesting.

- You: Robbi, bring me a beer, no bring me two beers.

o Your guest: And this does he also. Can he also suck the living room?

- You: Yes, sure, but let him first get the beer, otherwise he starts hoover the living room and don't bring us the beer.

o Your guest: There, there he comes.

- You: Uh, thanks robot. So and now hoover the living room.

o Your guest: Does he actually know there the vacuum cleaner is?

- Your: Yes, of course, he has placed it there.

o Your guest: Uh, then he has a real unit location plan of the house and the stuff in it?

- You: Yeah, some similar.

o Your guest: Uh-huh, that's interesting.

- You: He shouldn't run over all my stuff then he is moving around, but this he can also handle with his camera for visual detection - for see where the barriers are.

o Your guest: Uh-huh.

o Your guest: And now, during the robot vacuums the living room, you can turn up the sound of the stereo system? Can the robot that also?

- You: Yes, the robot is connected to the stereo system and the room and therefore you can say him turn on/off the light, uh more bass and so on.

o Your guest: Great.

- You: Robot turn up the sound.

o Your guest: Uh, that's nice. That's really nice. And surely you can let down the shutter, or not?

- You: Sure. (Robbi) Please, let down the shutter!
[..]

II

o Your guest: Hello <your name>, how are you?

- You: Hello Michael, I am fine, how about you?

o Your guest: Fine also, thanks.

- You: I guess you can smell than something goes on. Guess what!

o Your guest: What? Hey, you make me curious.

- You: You would never guess it, I just got a human robot.

o Your guest: What? A human robot? You make a fool out of me.

- You: No, I am not. Look, I will show you.

- You: Robot, come here. We have a guest, who wanna see you.

o Your guest: Really, unbelievable, a human robot. What can he do?

- You: Oh, he can do almost everything. I don't really know what he all can do, I guess everything what you expect a human robot to do. Such as vacuum the living room, clean things up, all the stuff that you don't like to do. Even cook.

o Your guest: That's great. Even cook. Hey show me something.

- You: Ok.

- You: Robot, switch on the lights.

o Your guest: That's easy.

- You: Yeah, what do you wanna drink.

o Uh, an orange juice. Cold, please.

- You: Robot, bring us an orange juice and an apple soda. Cold, please.

o And I guess he never gets angry of command him around. Isn't it?

- You: Absolutely not.

o That's great. No more a bad answer when you ask for something.

- You: That's really comfortable. And you don't have to stand up and get your newspaper, your shoes, your laptop and whatever. Just relax and sit on the couch and let him get the things or let him do the dishes, when you relax.

o Your guest: Sounds that I now, what I buy next. Where can you get this?

- You: Actually, this is a beta version. I have it just for testing and make it better. The problem is still the speech recognition and the problems included with it. For example how should the robot know that I said to you, that he brings me beer, when I say "Robot, bring me beer" or that I commanded it to him directly. In one case he should get it and in the other one not.

o Your guest: Sounds complicated.

- You: Yeah, and this isn't all. Then he understands that he is addressed, he also have to understand, what he has to do. And because of the great functionality he has, he can do a tons of things.

o Your guest: hm.

- You: But many people work on things like that, so the dream should become truth soon.

o Your guest: Hey, that sounds better. Let me know. I will get one.

- You: Ok, sure.

o Your guest: Hey, really great news. So see you soon. Have a great time.

- You: All right. See you than, take care.
[..]

List of Figures

Bibliography

[Argyle 1969] Michael Argyle, *Social Interaction*, Methuen, London, 1969

[Ruusuvuori 2001] Johanna Ruusuvuori, *Looking means listening: coordinating displays of engagement in doctor-patient interaction*, Social Science & Medicine, volume 52, pages 1093-1108, 2001

[Tankard 1970] J.W.Tankard, *Effects of eye position on person perception*, Perc. Mot. Skills, volume 31, pages 883-93, 1970

[Vertegaal et al. 2001] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, Anton Nijholt, *Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes*, SIGCHI'01, ACM, Seattle, March 2001

[Kleinke et al. 1973] C. L. Kleinke, A. A. Bustos, F. B. Meeker, R. A. Staneski, *Effects of self-attributed and other-attributed gaze in interpersonal evaluations between males and females*, Journal of experimental social Psychology, number 9, pages 154-63, 1973

[Humanoids 2003] *Proceedings of the Third IEEE International Conference on Humanoid Robots - Humanoids 2003*, IEEE, Karlsruhe, Germany, 2003

[Maglio et al. 2000] Paul P. Maglio, Teenie Matlock, Christopher S. Campbell, Shumin Zhai, Barton A. Smith, *Gaze and speech in attentive user interfaces*, Proceedings of the International Conference on Multimodal Interfaces, volume 1948, series LNCS, Springer, 2000

[JRSJ 1998] *Special Issue on Human-Friendly Robots*, Journal of the Robotics Society of Japan, volume 16, number 3, 1998

[Stiefelhagen 2002] Rainer Stiefelhagen, *Tracking and Modeling Focus of Attention in Meetings*, Dissertation, Universität Karlsruhe, Fakultät für Informatik, 2002

[Stiefelhagen et al. 2001a] R. Stiefelhagen, J. Yang, A. Waibel, *Tracking Focus of Attention for Human-Robot Communication*, IEEE-RAS International Conference on Humanoid Robots - Humanoids 2001, Tokyo, Japan, November 22-24, 2001

[Stiefelhagen et al. 2002] R. Stiefelhagen, J. Zhu, *Head Orientation and Gaze Direction in Meetings*, Proceedings of ACM CHI 2002, Minneapolis: ACM

[Stiefelhagen et al. 2001b] R. Stiefelhagen, J. Yang, A. Waibel, *Estimating Focus of Attention Based on Gaze and Sound*, Proceedings of PUI 2001 (Orlando, FL, Nov. 2001)

[Katzenmaier 2003] M. D. Katzenmaier *Verfolgen der Sprecheraufmerksamkeit mit Hilfe der Ausgabe des Spracherkennners*, Studienarbeit, Universität Karlsruhe, Fakultät für Informatik, 2003

[Bakx et al. 2003] I. Bakx, K. v. Turnhout, J. Terken, *Facial Orientation During Multi-party Interaction with Information Kiosks*, Proceedings of the Interact 2003, Zurich, Switzerland, 2003

[Zahn et al. 1996] T. Zahn, R. Izak, K. Trott, *Mixed analog-digital Neurochip for acoustical Attention*, Graduiertenkolleg "Informatik und Technik", TU Ilmenau, 1996

[Yang et al. 1997] J. Yang, W. Lu and A. Waibel, *Skin-Color Modeling and Adaption*, Technical Report, School of Computer Science CMU-CS-97-146, CMU, USA, 1997

[Stergiou & Siganos 1996] C. Stergiou and D. Siganos, *Neural Networks*, SURPRISE 96 Journal, Department of Computing, Imperial College of Science Technology and Medicine, London, 1996

[Bishop 2000] Christopher M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 2000

[Perzanowski et al. 2001] D. Perzanowski et al., *Building a multimodal human-robot interface*, IEEE Intelligent Systems, pages 16-21, 2001

[Agah 2001] A. Agah, *Human interactions with intelligent systems: research taxonomy*, Computers and Electrical Engineering, pages 71-107, 2001

[Koku et al. 2000] , A.B. Koku et al., *Towards socially acceptable robots*, Proceedings of 2000 IEEE International Conference on Systems, Man and Cybernetics, pages 894-899, 2000

[Adams et al. 2000] , B. Adams et al., *Humanoid robots: a new kind of tool*, IEEE Intelligent Systems, pages 25-31, 2000

[Matsusaka et al. 1999] , Y. Matsusaka et al., *Multi-person conversation via multi-modal interface - A robot who communicates with multi-user*, Proc. Eurospeech 99, pages 1723-1726, 1999

[Soltau et al. 2003] , Hagen Soltau, Hua Yu, Florian Metze, christian Fügen, Quin Jin and Szu-Chen Jou, *The 2003 ISL rich transcription system for conversational telephony speech*, IEEE International Conference on Speech, Acoustic and Signal Processing - ICASSP 2004, Montreal, 2004

[Verbmobil 2000] Alex Waibel, Hagen Soltau, Tanja Schultz, Thomas Schaaf and Florian Metze, *Multilingual Speech Recognition*, chapter in: Verbmobil - Foundations of Speech-to-Speech Translation, Wolfgang Wahlster (Hrsg.), Springer Verlag, 2000

[Verbmobil II 2001] Hagen Soltau, Thomas Schaaf, Florian Metze and Alex Waibel, *The ISL Evaluation System For Verbmobil - II* IEEE International Conference on Speech, Acoustic and Signal Processing - ICASSP 2001, Salt Lake City, 2001

[SNNS] *Stuttgart Neural Network Simulator User Manual*, Version 4.2, http://www-ra.informatik.uni-tuebingen.de/SNNS

[ARR] Arrington Research. http://www.arringtonresearch.com

[ASL] Applied Science Laboratories. http://www.a-s-l.com

[SRR] SR Research. http://www.eyelink.com

[SMI] SensoMotoric Instruments. http://www.smi.de

[LCT] LC Technologies. http://www.eyegaze.com

[Polhemus] Polhemus, http://www.polhemus.com

[iReality] iReality, Inc. http://www.genreality.com

[Ascension] Ascension Technology Corporation, http://www.ascension.com

[Burger 2003] Conventions for the Transcription of Spontaneous Speech, 2003 http://www.is.cs.cmu.edu/trl_conventions/projects/nespole.html

[Chil 2004] Computers in the Human Interaction Loop (CHIL), 2004 http://chil.server.de