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

Michael Katzenmaier
Interactive Systems Labs
Universität Karlsruhe (TH)
Karlsruhe, Germany

Rainer Stiefelhagen
Interactive Systems Labs
Universität Karlsruhe (TH)
Karlsruhe, Germany
stiefel@ira.uka.de

Tanja Schultz
Interactive Systems Labs
Carnegie Mellon University
Pittsburgh, PA, USA
tanja+@cs.cmu.edu

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. 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. The paper compares the result of three approaches. The first approach uses purely 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.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and presentation**]: User Interfaces

General Terms

Human Factors

Keywords

Multimodal interfaces, attentive interfaces, focus of attention, head pose estimation, speech recognition, human-robot interaction

1. INTRODUCTION

Building human-friendly robots which are able to interact and cooperate with humans has been an active research field in recent years [2, 1]. A major challenge in this field is to develop robots that can interact and cooperate with humans by understanding human communication modalities such as speech, gaze and gestures.

In this work we address the problem of automatically determining when a robot was addressed by human and when not. This is an important problem, when robots should 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 whether this person is talking to someone else in the room.

In this work we investigate to what extent the addressee of an utterance can be determined in a situation where people interact with each other and a robot intermittently. Rather than deciding on the basis of information of the visual and auditory channels separately, we also combine the information from both channels and compare the results with each other. For this purpose, 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.

The remainder of this paper is organized as follows: In Section 2 we review some related work. Section 3 describes the data collection setup. In Section 4 we investigate how well the addresses of an utterance can be determined based on the visually estimated head pose of a person. Section 5 describes our experiments on identifying the addressee based on acoustic cues and analyzing the speaker's speech. In Section 6 we present experiment results for audio-visual determination of the addressee. Section 7 concludes the paper.

2. RELATED WORK

A body of literature suggest that gaze, head pose and body orientation play an import role during social interaction and in particular are used and perceived as a signal of attention during human interaction [3, 9, 5, 7].

The relation between gaze and speech in multi-party communication between several people recently has been investigated by Vertegaal et al. [12]. They found that subjects looked about three times more at individuals they spoke to.

Stiefelhagen and Zhu [11] have investigated the relation of eye gaze and head orientation in multi-party interaction. They concluded that head pose is a reliable cue to determine at whom someone looked in small meetings.

Other researches have investigated how people use speech and gaze when interacting with attentive objects in a smart environment. Maglio et al [8] have for instance shown that

ICMI'04, October 13–15, 2004, State College, Pennsylvania, USA.

people tend to look towards objects with which they interact by speech. In their study they found that subjects nearly always looked at the addressed device before making a request.

Bakx et al. [4] 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. Bakx et al. also analyzed using the utterance length of the speaker for discriminating between addressees. They concluded that by combining the acoustic feature with facial orientation, some improvement in detecting the correct addressee can be achieved.

3. EXPERIMENTAL SETUP

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. Our experiments focus on the recordings of the host, since the goal of this work is to determine if the host addresses the robot or the guest. Therefore, the guest was played by the same person throughout the collection sessions, only the hosts differ for each session and were randomly selected. Subjects included under-graduate students, graduate students and one professor. The main selection criteria was that they are native speakers of English, since all recordings were done in English. Apart from taking part in this data collection, the selected subjects were not involved in this project.

In order to provoke a challenging scenario which includes robot commands directed to the robot and also conversation about the robot, the hosts were given instructions about the task and also an example dialog beforehand. In those instructions the hosts were asked to imagine that they had recently purchased a new household robot which can do many things, such as bring drinks, adjust the lights, vacuum the room, etc. and that they now expect a guest to whom they want to introduce their robot, discuss pros and cons of robots, and give the robot some commands in order to show the capabilities.

Figure 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 robots 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.

We did audio and video recordings of 18 sessions, each of roughly 10 minutes length. The audio data were fully transcribed and tagged on the turn level to indicate whether the host addresses the robot or the guest ([command],[no command]). For training and tuning of speech recognition components, we divided the data into 8 sessions for training, and 5 sessions for development. For the training of the video components, we manually labeled the first 2.5 minutes of the video recordings of four out of these 8 training sessions. For evaluating the speech parts, we used another set of 5 sessions. Table 1 shows the amount of audio and video data and the division into training, development, and evaluation set which was used for the speech-related experiments described in Section 5.

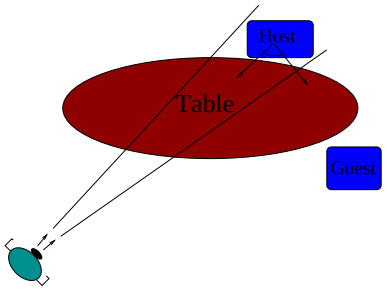

Figure 1: Data Collection Setup

Data	# Session	Length [min:sec]	[frames]	# labeled visual targets [frames]
Train.	8	82:37	32491	5024
Devel.	5	50:35	20435	-
Eval.	5	51:35	20435	-

Table 1: Audio and Video data.

4. IDENTIFICATION OF THE ADDRESSEE BASED ON HEAD POSE

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

4.1 Relation of visual target and the addressee of an utterance

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

We therefore manually labeled at which target the host looked in 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 2 shows the confusion matrix between the acoustic and visual labels. Here, the acoustic targets (the addressee) are indicated with T_A and are given in rows, the visual targets (at whom did the speaker look?) are given in columnwise, labeled with T_V.

It can be seen that people mostly looked at the robot when they addressed the robot (95% of the 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, i.e. 99.5%) he also addressed the guest.

To summarize, in the data that we recorded, looking towards the other human was a direct indication that the other

Audio \ Video	$T_V = Guest$	$= Robot$	$= Other$
set02 $T_A = Guest$	462	44	202
$T_A = Robot$	3	43	2
set03 $T_A = Guest$	463	69	136
$T_A = Robot$	0	94	0
set04 $T_A = Guest$	289	34	221
$T_A = Robot$	0	46	3
set05 $T_A = Guest$	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 2: Confusion-matrix between hand-labeled addressees of speech acts (T_A) and the targets at which the speaker looked (T_V).

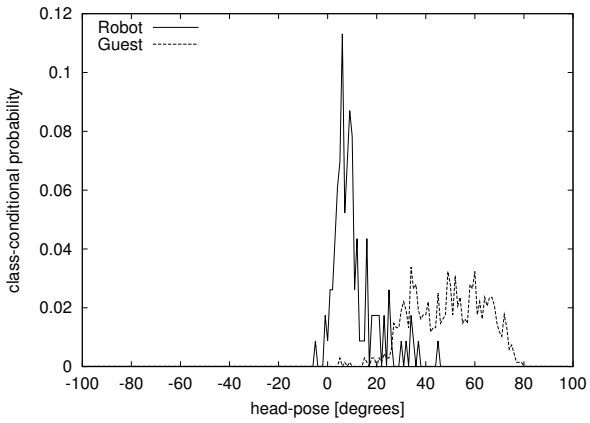

Figure 2: Typical class conditional probability distribution for the classification of the visual target for two targets.

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 utterance. This finding indicates that when we equip a robot with eyes to identify with whom the human counterpart interacts, we need more than just visual cues to solve the remaining 35% of the cases.

4.2 Head Pose Estimation

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

In this work, we only use horizontal head orientation (pan) to distinguish between different addressees of a person. It has to 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 of the system were recorded several years ago in a different lab and under different lighting conditions.

4.3 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 our previous work [10]. We described an approach to find out at whom participants in a meeting had looked, based on their head orientations. Similar to this 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 focus F_V for each target class, first the a-priori probability $P(F_V = target)$, class conditional probability $P(X|F_V = target)$ and the probability $P(X)$ for each head pose X has to be calculated. Once these probabilities are calculated, the a-posteriori probabilities $P(F_V = target|X)$ can be calculated:

$$P(F_V = Target|X) = \frac{P(X|F_V = Target) \cdot P(F_V = Target)}{P(X)} \quad (1)$$

where $Target$ can be either "*Robot*" or "*Guest*" in our case, and X denotes the horizontal head orientation of the host. Figure 2 depicts typical class-conditional probability distributions for a person's head orientation in one of our data sets.

4.4 Experimental results

4.4.1 Used metrics

In this work, we are mainly interested in detecting when a robot was addressed. Therefore, we are interested in measuring precision and recall of detected utterances 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 result, the so-called f-measure, which is the geometrical median of the two values:

$$f - measure = \frac{2 * recall * precision}{recall + precision} \quad (2)$$

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

4.4.2 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. We then also learned the priors and class-conditionals automatically with an approach described in [10].

Table 3 shows the average results of the two experiment on the four sequences. With the hand-tuned parameters, we could correctly detect the visual target 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 89%, resulting in an f-measure of 0.82. With learned model parameters, a slightly lower accuracy and f-measure was obtained.

Distributions	Precis.	Recall	F-Measure	Accuracy
True	0.89	0.77	0.82	0.96
Learned	0.74	0.85	0.79	0.93

Table 3: Determination of the visual target - Robot or Guest - based on visually estimated head orientation. Results with true and learned model parameter (distributions and priors) are given.

4.4.3 Estimation of the addressee

Since our previous experiments indicate that visual focus is a good indicator for the addressee of an utterance - especially if the visual target was a human - we can use the estimated visual target as an estimate of the (acoustic) addressee. Table 4 summarizes the results of detection of 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	Precis.	Recall	F-Meas.	Accu.
True	0.61	0.83	0.7	0.93
Learned	0.6	0.8	0.68	0.89

Table 4: Determination of the acoustical addressee, based on (visually) estimated head pose. Results with true and learned model parameter (distributions and priors) are given.

Using the true priors and class-conditional distributions for head pose, we could identify the correct addressee in 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.

5. IDENTIFICATION OF THE ADDRESSEE BASED ON SPEECH

In this section we describe, how we estimate the likely addressee of an utterance, based on features extracted from the speech signal. The goal 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 to recognize 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 done in two phases. In the first phase we conducted a pilot study on the audio recordings of 3 sessions collected in German language [6]. The collection scenario (host-robot-guest) was very similar to the audio setup described above (see section 3), except that we analyzed the recordings of both sides, the host and the guest. In the second phase we recorded a larger set of English speech data as well as video and transferred the findings of the pilot study to the new data.

In the pilot study on German language we investigated the potential of various speech-based features to discriminate between a *command* and a *conversation*. The average of the feature values are given in Figure 3. 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 to more likely contain the term 'robot' or 'robbi' to address the robot. Therefore, we took the *sentence length* $S(X) \in N$ and the *occurrence of 'robot'* $R(X) \in 0, 1$ as discriminating features.

The second set of discriminating features is using the syntactical and semantical differences between commands and conversations. Commands are formulated in imperative form, and are less conversational than the 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 distinguishes the imperative form from others. Furthermore, we used the transcribed material to built two statistical trigram 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* $PP_{cmd}, PP_{cvs} \in R$ 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}.

The third set of features takes the sentence structure and parseability into account. For this purpose we developed a context free grammar (CFG) designed to parse commands, and determined the boolean *parseability* $Z(X) \in 0, 1$ fea-

	Prec.	Recall	F-Meas.	Accu.
Features	Bayes on Hypotheses			
1) PP_{VM}, PP_{Nav}	0.75	0.19	0.30	0.77
2) 1) + S(X), Z(X)	0.75	0.38	0.50	0.80
3) 2) + PP_{com}, PP_{conv}	0.47	0.56	0.51	0.72
	MLP on Hypotheses			
1) PP_{VM}, PP_{Nav}	1.0	0.12	0.21	0.77
2) 1) + S(X), Z(X)	0.56	0.56	0.56	0.77
3) 2) + PP_{com}, PP_{conv}	**0.65**	**0.69**	**0.67**	**0.82**
4) 3) + C(X)	0.43	0.81	0.56	0.67
	MLP on Transcripts			
3) 4 x PP, S(X), Z(X)	0.83	0.63	0.72	0.87
5) 3) + R(X) + I(X)	0.90	0.56	0.69	0.87

Table 5: Feature set evaluation on the German pilot study corpus with different classifiers. Feature set 1 includes the two perplexities $PP_{VM}(X), PP_{Nav}(X)$. Feature set 2 includes utterance length $S(X)$ and parseability $Z(X)$ in addition. Feature set 3 additionally includes the perplexities on the the German Verbmobil corpus PP_{VM} and on a car navigation corpus PP_{Nav}. Feature set 4 includes the correlation feature $C(X)$. Finally, feature set 5 includes also the number of occurrences of the word "Robot" $R(X)$ and of imperatives $I(X)$ (see text for details). The result with the best f-measure on hypotheses is highlighted.

ture that is set to '1' if a sentence could be parsed using the CFG, and to '0' otherwise. The last set of feature 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. We calculated two different correlation coefficients, one based on the hypothesized words $C_w(X)$, another one on the correlation of letters $C_l(X)$ of these words.

5.2 Feature Evaluation on German Language

We evaluated the features 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, (1) simple comparison, (2) Bayes classification, and (3) Multilayer Perceptrons. The results in Table 5 show that the combination of the above mentioned speech based features outperformed the single features. The best performance was achieved when using a fully connected Multilayer Perceptron (MLP) with standard backpropagation for the feature combination. We applied this MLP on both, the hypotheses of the speech recognizer and the original transcripts. For the hypotheses output we achieved a command detection accuracy of 0.82 and for the transcripts we achieved a detection accuracy of 0.87 [6]. The comparison shows that the impact of the sub-optimal speech recognition output on the decision accuracy is moderate leading to a 5% absolute degradation.

The overall best result could be achieved by taking all features as an input vector of a simple feed forward MLP with differentiable activation function, one hidden layer, trained using gradient descent. The net gives the probabilities as output, one for being a conversation, one for being a command. On the transcripts this MLP gives an accuracy of 0.87, a recall of 0.63 and a precision of 0.83, which leads to an f-measure of 0.72.

These results are encouraging and indicate that speech-based identification of the addressee is possible, even if the speech recognition accuracy is relatively low. However, the overall performance in this pilot study suffered from data sparseness. To overcome the data sparseness problem we collected a larger data set (see above). 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 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 the training data could be completely parsed. The CFG-based decoder uses filler words, and thus can handle typical spontaneous effects occurring in spoken speech, such as hesitations, false starts, and repetitions. To improve the CFG for commands, we additionally collected 425 commands from 8 people. In total the context free grammar for command parsing consists of 276 rules using 3162 nodes and 4638 arcs based on 434 terminals.

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

Table 6 shows the performances of the English speech recognition on the different data sets and the various customized system and compares the performance of the context free grammar based decoder with the statistical language model based one. The best performance on the evaluation set could be achieved with the n-gram based recognizer and resulted in a Word Accuracy of 0.83.

5.4 Feature Evaluation on English Language

In the following experiments we transferred the speech features to the English. We applied the same features, except for the parseability feature $Z(X) \in [0...100]$ which is no longer a boolean variable for English but express the percentage of parsed output. Figure 3 compares the results for both languages.

Especially the correlation based features show much better results on the English language which can be explained by the improved speech recognition accuracy. Due to the highly conversational style of the English data, the overall utterance length is much higher than that of the German

			Result WER[%] (pp/OOV[%])					
Build on		Vocab.	Train. Set		Dev. Set		Eval. Set	
1) SWB	LM	1045	68.6	(116/0)	67.0	(85/9)	-	
2) Train. (our data)	LM	1045	43.0	(14/0)	67.0	(139/9)	-	
3) SWB + Train.	LM	1165	31.2	(13/0)	78.7	(68/9)	-	
4) **SWB + all our data**	LM	1720	**38.1**	(5/0)	**19.5**	(5/0)	**16.7**	(5/0)
5) SWB + Train. + Eval.	LM	1720	-		57.0	(159/0)	-	
6) SWB + Train. + Devel.	LM	1720	-		-		55.0	(127/0)
7) Train.	CFG	434	34.9	-	65.9	-	-	
8) **all our data**	CFG	632	**39.7**	-	**21.0**	-	**43.0**	-

Table 6: Performance comparison of speech recognizer trained on different training sets. Results with n-gram language models (LM) and with context-free grammars (CFG) for decoding are given. Indicated are the word error rates (WER). In brackets we also indicate perplexities (pp) and the number of out-of-vocabulary words (OOV), where appropriate.

data. In general the results indicate that the selected features can be successfully applied to both languages.

For English, we adopted the classification scheme, which worked best in the German case, the multilayer perceptron. We achieved a recall of 0.91, but only 0.49 accuracy and a precision of 0.19 on English, which leads to an f-measure of 0.31.

6. COMBINED RESULTS

In the previous sections we have discussed how the addressee of a speaker can be estimated based on their head orientation and acoustic features. 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:

$$P_{Sum} = \alpha \cdot P(Target|speech) \cdot (1-\alpha) \cdot P(Target|HeadPose) \quad (3)$$

and

$$P_{Mult} = P(Target|speech)^{\alpha} \cdot P(Target|HeadPose)^{(1-\alpha)} \quad (4)$$

respectively.

We observed slightly better combined estimation results by using the weighted sum compared to multiplying the weighted probabilities. Figure 4 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 in 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).

The results show that the combination of acoustic and visual estimation improved 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% error combined). In addition, the precision when detecting commands towards the robot could be improved from 0.57 to 0.65 (19% relative improvement).

Estimation	Precis.	Recall	F-Meas.	Accu.
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: Acoustic, visual and combined estimation of the addressee.

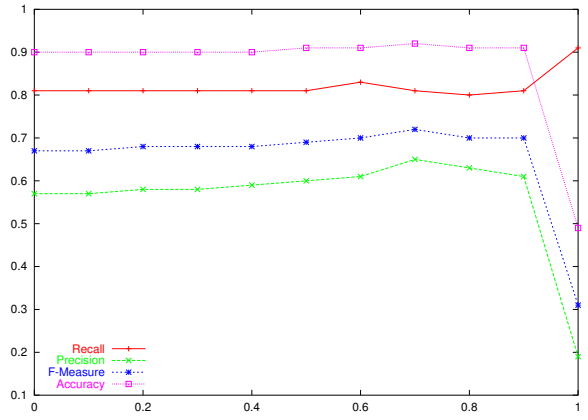

Figure 4: Combined estimation results with different weights α (see text). Indicated are accuracy, recall, f-measure and precision (top to bottom).

7. CONCLUSION

In this work we investigated the power of acoustic and visual cues to identify the addressee in multiparty 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

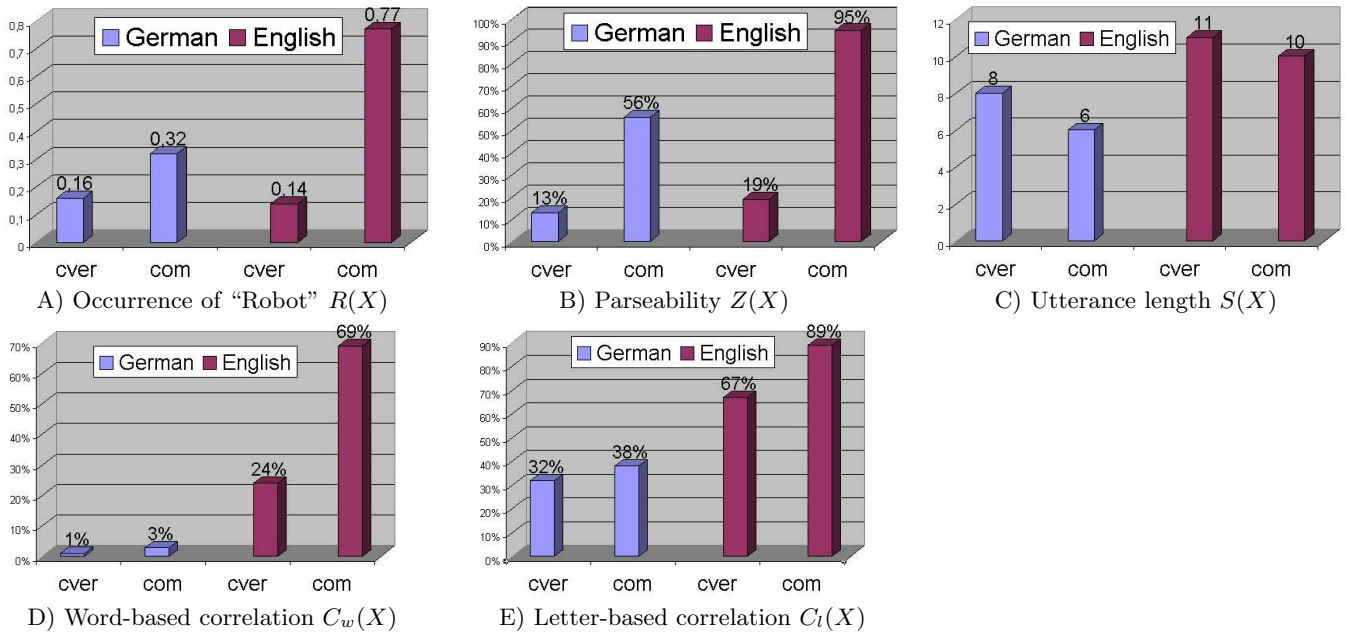

Figure 3: Comparision of different features on German and English language for human-human conversations (cver) and human-robot commands (com).

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, indicating that visual cues by itself might not be sufficient.

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 89% of time. Speech commands towards the robot were (visually) detected with a precision of 0.6 and a recall of 0.8.

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 a German pilot study we investigated the usefulness of various features that were derived from the hypothesis of a speech recognizer. 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 this study. 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.

Finally, we combined the purely speech-based and head pose based approaches for discriminating the addressees. This resulted in significant improvements, despite the comparably poor results of the speech-based discrimination approach: By using the weighted sum of the acoustic and visual posterior probabilities for the addressees, correct classification rate of the addressees increased from 0.9 (visually estimated) to 0.92. 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.

Acknowledgments

The first author would like to thank all colleagues in the Interactive Systems Labs for their support and also would like to thank everybody who participated in data collections.

This research was supported by the German Research Foundation (DFG) within SFB 588 "Humanoid Robots" and the European Commission under contract no. 506909 within the project CHIL (Computers in the Human Interaction Loop; http://chil.server.de).

8. ADDITIONAL AUTHORS

Additional authors: Ivica Rogina and Alex Waibel, Interactive Systems Labs, Universität Karlsruhe.

9. REFERENCES

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

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

[3] M. Argyle. *Social Interaction.* Methuen, London, 1969.

[4] I. Bakx, K. van Turnhout, and J. Terken. Facial orientation during multi-party interaction with information kiosks. In *Proceedings of the Interact 2003*, Zurich, Switzerland, 2003.

[5] J.W.Tankard. Effects of eye position on person perception. *Perc. Mot. Skills*, (31):883–93, 1970.

[6] M. Katzenmaier. Determining the adressee in spoken human robot interaction, studienarbeit. Technical report, Fakultät für Informatik, Universität Karlsruhe (TH), 2003.

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

[8] P. P. Maglio, T. Matlock, C. S. Campbell, S. Zhai, and B. A. Smith. Gaze and speech in attentive user interfaces. In *Proceedings of the International Conference on Multimodal Interfaces*, volume 1948 of *LNCS*. Springer, 2000.

[9] J. Ruusuvuori. Looking means listening: coordinating displays of engagement in doctor-patient interaction. *Social Science & Medicine*, 52:1093–1108, 2001.

[10] R. Stiefelhagen. Tracking focus of attention in meetings. In *International Conference on Multimodal Interfaces*, pages 273–280, Pittsburgh, PA, October 2002. IEEE.

[11] R. Stiefelhagen and J. Zhu. Head orientation and gaze direction in meetings. In *Conference on Human Factors in Computing Systems (CHI2002)*, Minneapolis, April 2002.

[12] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt. Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *SIGCHI'01*, Seattle, March 2001. ACM.