Proceedings of 2004 IEEE/RSJ International Conference on
Intelligent Robots and Systems
September 28 - October 2, 2004, Sendai, Japan

Natural Human-Robot Interaction using Speech, Head Pose and Gestures

R. Stiefelhagen, C. Fügen, P. Gieselmann, H. Holzapfel, K. Nickel and A. Waibel
Interactive Systems Labs
Universität Karlsruhe (TH)
Karlsruhe, Germany
Email: stiefel@ira.uka.de

Abstract— **In this paper we present our ongoing work in building technologies for natural multimodal human-robot interaction. We present our systems for spontaneous speech recognition, multimodal dialogue processing and visual perception of a user, which includes the recognition of pointing gestures as well as the recognition of a person's head orientation. Each of the components are described in the paper and experimental results are presented. In order to demonstrate and measure the usefulness of such technologies for human-robot interaction, all components have been integrated on a mobile robot platform and have been used for real-time human-robot interaction in a kitchen scenario.**

I. Introduction

In the upcoming field of humanoid and human-friendly robots, the ability of the robot for simple, unconstrained and natural interaction with its users is of central importance [1], [2]. The basis for appropriate action of the robot must be a comprehensive model of the current surrounding and in particular of the humans involved in interaction.

To facilitate natural interaction, robots should be able to perceive and understand all the modalities used by humans during face-to-face interaction. Besides speech, as the probably most prominent modality used by humans, these modalities also include pointing gestures, facial expressions, head pose, gaze, eye-contact and body language for example.

In our research labs at the Universität Karlsruhe (TH) and at Carnegie Mellon University, we are developing technologies for the understanding of these human interaction modalities. In particular in the framework of a German research project on humanoid robots (Sonderforschungsbereich Humanoide Roboter, SFB 588) we have been working using and improving such technologies to provide for natural interaction between a humanoid robot and its users.

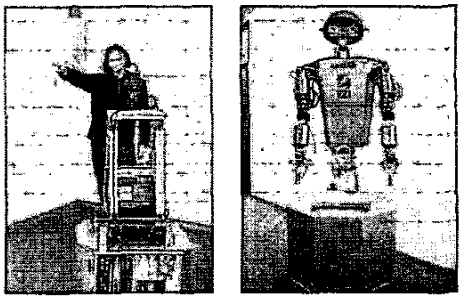
In this paper we present our work in this area. We have developed components for speech recognition, multimodal dialogue processing, visual detection and modeling of users, including head pose estimation and pointing gesture recognition. All components have been integrated on a mobile robot platform and can be used for real-time multimodal interaction with a robot.

The target scenario we addressed is a household situation, in which a human can ask the robot questions related to the kitchen (such as "What's in the fridge ?"), ask the robot to set the table, to switch certain lights on or off, to bring certain objects or to obtain suggested recipes from the robot.

The current components of our system include

- a speech recognizer,
- 3D face- and hand-tracking,
- pointing gesture recognition,
- recognition of head pose,
- a dialogue component,
- speech synthesis,
- a mobile platfom,
- a stereo camera system, including pan-tilt, unit mounted on the platform.

Figure 1.a) shows a picture of our system and a person interacting with it. Part of the visual tracking components have already been integrated in ARMAR [3], a humanoid robot with two arms and 23 degrees of freedom. This robot is depicted in Figure 1b).



a) b)

Fig. 1

FIG.1 A) INTERACTION WITH OUR DEVELOPMENT SYSTEM. SOFTWARE COMPONENTS INCLUDE: SPEECH RECOGNITION, SPEECH SYNTHESIS, PERSON AND GESTURE TRACKING, DIALOGUE MANAGEMENT AND MULTIMODAL FUSION OF SPEECH AND GESTURES. FIG 1B): PART OF THE COMPONENTS HAVE ALREADY BEEN INTEGRATED IN A HUMANOID ROBOT WITH TWO ARMS.

The remainder of this paper is organized as follows: In Section II we describe our JANUS speech recognition system which we use for human-robot interaction and present some experimental results. In Section III, visual perception of the user is discussed. Here we present our approach to visually detect and track a user, his head, hands

0-7803-8463-6/04/$20.00 ©2004 IEEE

and head orientation, as well as our approach for detecting pointing gestures and pointing direction. In Section IV, the dialoge component of the robot is described. In Section V, the integration of all the components on a mobile robot is explained and a typical interaction scenario is described. We conclude the paper in Section VI, where we also give an outlook to future work.

II. Speech Recognition

The probably most prominent interaction modality of humans is their speech. In order to provide for natural human computer interaction, recognition and understanding of spontaneous speech is of utmost importance.

For speech recognition we are using the Ibis decoder [4], which was developed at the University of Karlsruhe as part of our Janus Recognition Toolkit (JRTk) [5]. Using this toolkit we have developed a user-independent speech recognizer for spontaneous human robot interaction.

A. Context Free Grammar Decoding

The Ibis decoder allows us to decode along context free grammars in addition to the classical statistical n-gram language models. Using grammars instead of n-gram language models is especially an advantage in small domains, like in our household scenario. In such domains there is normally less domain dependent data available for the training of robust statistical n-gram language models.

The context free grammar implementation in Ibis has also several other advantages. Rather than compiling one finite state graph out of all the terminals given by the grammars, we use a more dynamic approach, where several rule based finite state graphs consisting of terminals and non-terminals, are linked together by their non-terminal symbols.

Another feature is that the grammars can be expanded on the fly by new rules or terminals without restarting the recognizer. Even new words can be added to the grammar and the search network on the fly. In most cases we work with non-statistical semantic grammars, i.e. each transition to the next word has the same language model score, whereby terminals are grouped by their semantical meaning to non-terminal symbols.

1) Handling of Spontaneous Speech: A major problem when using context free grammars in speech recognition is the modeling of spontaneous speech together with its ungrammaticalities like hesitations or word repetitions. Due to the fact, that these effects can occur at any time in a speech query, it is impossible to model them manually in the grammar. The same applies also to non-human noises.

We are using so-called filler words to cope with such spontaneous speech events. These words consists of special acoustic models trained only on e.g. non-human noises or hesitations and they can potentially occur between any two terminals of the grammar. Instead of asking the grammar for their probability, a predefined filler penalty is applied.

2) Dialogue-Context Dependent Search Space Control: When using speech recognition together with dialogue systems, the dialogue context is always known. This information can be used to improve the speech recognition performance, because less probable answers to a clarification question of the robot can be penalized. Also at the beginning of a dialogue the search space of the recognizer can be restricted by disabling all answers to system questions like "yes" or "no". This is done by activating/deactivating or penalizing specific semantic top-level rules in the grammars given by the dialogue manager during runtime. Penalizing of rules should be preferred, because it still allows user queries in different contexts.

3) Experimental Results: For our experiments we collected a set of nearly 360 user queries of 9 different speakers, which result in around 15min of speech. We measured the word error rate (WER), the sentence error rate (SER), the real-time factor (RTF) and the memory requirements of our recognizer. A low SER is important for a good language understanding. The RTF is measured on a 800MHz PIII.

Our recognition system for the robot interaction consists of about 34,000 gaussian models and was trained on nearly 300 hours of conversational telephone speech (Switchboard). This size of the acoustic model allows us to decode in less than real-time as can be seen in table I, which gives us the ability to run also other components of the human-machine interface on the same computer. Incremental adaptation techniques like vocal-tract length normalisation (VTLN) and constrained MLLR is used to compensate for different speakers, channels and background noises.

Table I also shows, that when adding the filler-words to the dictionary the grammar based system reaches nearly the same WER as the n-gram based system. But the advantage of the grammar based system is besides its higher recognition speed the much lower SER.

TABLE I

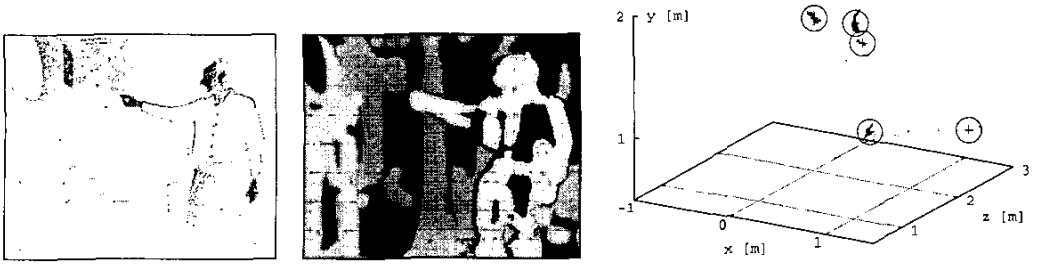
Comparison between grammar based and n-gram based speech recognition.

	WER	SER	RTF	Memory
grammar based	25.55%	48.21%	-	37 MB
+ filler words	23.05%	45.18%	0.759	37 MB
n-gram based	22.95%	51.79%	0.801	37 MB

B. Distant Microphones

A well-known problem in the speech recognition community is the difficulty of automatic speech recogntion with remote microphones or even worse, with microphones at variant distances. Therefore, in many cases, head-mounted close-talking microphones are used for speech recognition.

Since we want to develop human-friendly robots that eventually can operate in our daily lifes, we certainly don't want to force people to wear such head-mounted microphones in order to communicate with the robot.



a. Skin-color map b. Disparity map c. 3D skin-color clusters

Fig. 2

FEATURES FOR LOCATING HEAD AND HANDS: SKIN-COLORED 3D-PIXELS ARE CLUSTERED USING A K-MEANS ALGORITHM. THE RESULTING CLUSTERS ARE DEPICTED BY CIRCLES. IN THE SKIN COLOR MAP, DARK PIXELS REPRESENT HIGH SKIN-COLOR PROBABILITY. THE DISPARITY MAP IS MADE UP OF PIXEL-WISE DISPARITY MEASUREMENTS, THE BRIGHTNESS OF A PIXEL CORRESPONDS TO ITS DISTANCE TO THE CAMERA.

Therefore, we need to develop technologies to improve speech recognition under such situations, i.e. with remote microphones at variable distances.

1) Speech Segmentation: Another issue, when using remote microphones is the speech segmentation, because the user is not able to push a button for recording. Therefore we initially developed an energy and zero-crossing based speech segmentation, which transmits the segmented audio signals to the recognizer.

2) Experimental Results: We have performed some initial adaptation experiments to evaluate the sensibility of our real-time recognizer in combination with single microphones at different distances. Due to the lack of enough testing and adaptation material in the houshold domain, we've collected 2hrs of read speech of 25 speakers for adaptation and 15min of read speech of 9 speakers for testing.

For each microphone distance we adapted the codebooks on the adaptation data using MLLR [6], without performing speaker adaptation. As can be seen in table II, we reach significant WER reduction of around 10% - 15% through adaptation for the remote conditions, but the results within a normal action radius of about 4-8ft to the robot are even for the adapted case inacceptable. To see how sensible the

TABLE II

WERS FOR UNADAPTED AND ADAPTED SYSTEMS AT DIFFERENT MICROPHONE DISTANCES.

	close	lapel	4ft	5ft	6ft	8ft
unadapted	26.6%	29.7%	47.7%	51.9%	66.1%	69.3%
adapted	26.5%	28.4%	42.5%	44.7%	59.7%	60.1%

recognizer is to variations in the distance, we ran decoding experiments in which we tested the codebooks adapted on 5ft data on the 4ft and 6ft condition. Therefore, no speaker adaptation (which would also perform channel adaptation) was performed. As can be seen in table III, the stability of the resulting recognizer against moving speakers (changing distance to the microphone) seems to be very good, as there

TABLE III

ANALYSIS OF THE SENSIBILITY AGAINST VARIATIONS IN THE DISTANCE OF THE MICROPHONES FOR AN ALREADY ADAPTED SYSTEM.

	4ft	5ft	6ft
adapted on 5ft	42.4%	44.7%	60.1%

is almost no loss in recognition accuracy for a 20% change in distance when compared to the results in table II.

We are currently working on adapting the model-combination-based acoustic mapping (MAM) [7] developed for car navigation in our lab to the robot szenario.

III. VISUAL PERCEPTION OF THE USER

Knowledge about the users location, posture and focus of attention is an important cue for the understanding of human intention within a dialogue situation. From the images delivered by a fixed-baseline stereo camera head, we extract the following information in real-time: a) the 3D-positions of the users's head and hands, b) the head orientation and c) the direction of the pointing gestures that are performed by the user.

A. 3D Tracking of Head and Hands

Head and hands can be identified by color as human skin color clusters in a small region of the chromatic color space [8]. To model the skin-color distribution, two color histograms (S^+ and S^-) are built by counting pixels belonging to skin-colored respectively *not*-skin-colored regions in sample images. By means of these histograms, the probability of a pixel being skin-color can be calculated. The result is a gray-scale map of skin-color probability (Fig. 2.a). To eliminate isolated pixels and to produce closed regions, a combination of morphological operations is applied to the skin-color map.

Due to the robot's motion, the lighting situation is likely to vary strongly. Thus, it is important to initialize and to update the skin-color model automatically. In order to do this, we incorporate the lighting invariant depth information, and search for a person's head in the disparity map (Fig. 2.b) of each new frame. Following an approach proposed in [9], we first look for a human-sized connected region, and then check its topmost part for head-like dimensions. Pixels inside the head region contribute to S^+, while all other pixels contribute to S^-. By means of this procedure, the skin-color model is permanently kept up to date and no manual initialization is required.

The task of tracking consists in finding the best hypothesis s_t for the positions of head and hands at each time t. The decision is based on the current observation (the 3D skin-pixel clusters, Fig. 2.c) and the hypotheses of the past frames, $s_{t-1}, s_{t-2}, \ldots$. With each new frame, all combinations of the clusters' centroids are evaluated to find the hypothesis s_t that exhibits the highest results with respect the product of the following 3 scores:

- The *observation score* $P(O_t|s_t)$ is a measure for the extent to which s_t matches the observation O_t. $P(O_t|s_t)$ increases with each pixel that complies with the hypothesis.
- The *posture score* $P(s_t)$ is the prior probability of the posture. It is high if the posture represented by s_t is a frequently occurring posture of a human body. For the calculation of $P(s_t)$, a basic model of the human body was built from training data.
- The *transition score* $P(s_t|s_{t-1}, s_{t-2}, \ldots)$ is a measure for the probability of s_t being the successor of the past frames' hypotheses. It is higher, the better the positions of head and hands in s_t follow the path defined by the preceding positions.

Our experiments indicate that by using the method described, it is possible to track a person robustly, even when the camera is moving and when the background is cluttered. The tracking of the hands is affected by occasional dropouts and misclassifications. We address this problem by applying multi-hypotheses tracking, so that the tracker is free to choose the most likely path through an n-best set of hypotheses for each frame, instead of being tied to a single (and maybe wrong) hypothesis.

B. Head Pose Estimation

Monitoring a person's head orientation is an important step towards building better human-robot interfaces. Since head orientation is related to a person's direction of attention, it can give us useful information about the objects or persons with which a user is interacting. It can furthermore be used to help a robot decide whether he was addressed by a person or not [10]. In our experiments head pose has also proved to be helpful to decide whether a person has performed a pointing gesture, as will be described in section III-C.

Our approach for estimating head-orientation is view-based: In each frame, the head's bounding box - as provided by the tracker - is scaled to a size of 24x32 pixels. Two neural networks, one for pan and one for tilt angle, process the head's intensity and disparity image 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 have a total number of 1597 neurons, organized in 3 layers. They were trained in a person-independent manner on sample images of rotated heads. We collected training data from six users. Users were standing approximately at a distance of two to three meters away from the camera and were free to move around within the camera's field of view (see Fig. 3). We asked people to freely look around and recorded their exact head pose using a magnetic pose tracker. The recorded rotation angles varied from $-90°$ to $90°$. We evaluated the system's performance on a multi-user test set and on new users. For the multi-user evaluation, the system was trained on images from all users and was tested on different images from the same users. The results for new users was obtained by training the system on images from five users and testing on the sixth user. Table IV shows the results for multi-user and new user tests.

Fig. 3

SAMPLE IMAGE FROM THE DATA COLLECTION. A MAGNETIC SENSOR PLACED ON THE SUBJECTS' HEAD PROVIDES GROUND TRUTH FOR HEAD POSE, WHICH WAS USED FOR TRAINING AND EVALUATION.

TABLE IV

MEAN ERROR OBTAINED FOR THE MULTI-USER AND NEW USER TESTS (PAN/TILT ANGLES)

mean error	multi-user	new user
gray values	4.6 / 2.4	15.5 / 6.3
depth info	8.0 / 3.3	11.0 / 5.7
depth + gray	**4.3 / 2.1**	**9.7 / 5.6**

It can be seen that the combined approach of adding depth images to the input feature vector improves the results significantly in both cases.

C. Pointing Gesture Recognition

In the human-robot interaction scenario, we define a pointing gesture as the movement of the hand towards a pointing target. We model this typical motion pattern of the pointing hand in order to detect pointing gestures within other natural hand movements. Therefore, we decompose the gesture into three distinct phases (see Table V) and

model each phase with a dedicated Hidden Markov Model (see [11] for details). The features used as the models' input are derived from the tracked position of the pointing hand. The hand coordinates are transformed into a cylindrical, head-centered coordinate system in order to become invariant against the person's location. We have noticed [11], that people tend to look at the pointing target at an early stage of the gesture. We can exploit this behavior by calculating the absolute difference between the head's azimuth (elevation) angle and the hand's azimuth (elevation) angle, and incorporate these two features to the gesture models.

TABLE V
AVERAGE LENGTH μ AND STANDARD DEVIATION σ OF 210 POINTING GESTURES PERFORMED BY 15 TEST PERSONS.

	μ	σ
Complete gesture	1.75 sec	0.48 sec
Begin	0.52 sec	0.17 sec
Hold	0.72 sec	0.42 sec
End	0.49 sec	0.16 sec

In an evaluation with 12 test persons, this system scored at about 80% recall and 74% precision in recognition of pointing gestures. When head-orientation was added to the feature vector, the results improved significantly in the precision value: the number of false positives could be reduced from about 26% to 13%, while the recall value remained at a similarly high level.

In order to determine the 3D pointing direction, we extract the line from the center of the head to the center of the hand within the hold-phase of the gesture. In our experiments, this turned out to be a reliable estimate for pointing direction. With an average error below 20°, it is possible to disambiguate the possible pointing targets in most cases.

IV. Multimodal Dialogue Processing

The multimodal dialogue management processes the output of the speech recognizer and the one of the gesture recognizer in order to understand what the user wants the robot to do. Currently, the robot can help the user in the kitchen: A user can ask the robot to get cups or dishes and put them somewhere, to switch on or off the lights, to look in the fridge, to tell some recipes, etc. Therefore, results of the speech recognizer and the gesture recognizer are sent to the dialogue manager which evaluates them in the discourse context. The multimodal fusion is based on the semantics of both input modalities [12].

A. Dialogue Management

Our dialogue manager is based on the approaches of the language and domain independent dialogue manager ARIADNE [13]. For the domain-dependent part, we developed different kinds of resources: An ontology, a specification of the dialogue goals, a data base, a context-free grammar and generation templates.

Speech input is parsed by means of a context-free grammar which is enhanced by information from the ontology defining all the objects, tasks and properties about which the user can talk. In our scenario, these objects are the objects in the kitchen and their properties, for example the ability to be switched on or off. The tasks are taking or putting something somewhere, informing the user about the content of the fridge, telling him recipes, etc. The semantic representation created during parsing is then compared against the dialogue goals. If all the necessary information to accomplish a goal is available, the dialogue system calls the corresponding service. But if some information is still missing to accomplish a goal, the dialogue manager uses clarification questions to get this information from the user. This is done by means of generation templates which are responsible for generating the spoken output.

The gesture input is resolved by means of an environment model which is stored in the database. Currently, this environment model consists of different objects in the kitchen, such as cups, dishes, forks, knifes, spoons and lamps. The environment model matches a pointing gesture with possible targets. All objects that meet the matching constraints form an n-best list of pointing hypotheses in semantic representation. These hypotheses are used within the spoken context to disambiguate speech input. Disambiguation is performed by merging speech and gesture in a multimodal parsing process.

B. Multimodal Parsing

We use a constraint based approach to merge speech and gesture. Speech is used as the main modality and gesture events are used to disambiguate input information. This approach has shown to be quite tolerant towards falsely recognized gestures [14]. Parsing rules define constraints on time, context and input information, as well as rules for merging. The multimodal parser is part of the dialogue system and is applied after transforming speech and gesture input to semantic tokens.

For disambiguation of speech input, gesture events have to be assigned to referring speech events. Disambiguation can mean, but is not limited to

- Deixis, for example "switch on this lamp" is ambiguous when looking only at the speech information.
- Speech recognition errors, e.g. "switch on the little lamp" becomes ambiguous after misrecognition of the identifying adjective of the lamp.
- n-best hypotheses from gesture recognition and resolving in the environment model.
- n-best hypotheses from speech recognition when being combined with gesture information.

Therefore we can tolerate gesture recognition errors in the form of false detections, but are interested in missing as few gestures as possible. Incorrect gestures that are not correlated to a speech event can be sorted out by time constraints. In our experiments, we have detected that - for 3d pointing gestures - the the referring spoken word and the gesture are strongly correlated in time [14]. By using a (not very restrictive) 1000 ms boundary, both before the

start and after the stop time of the whole utterance, we can capture the relevant gestures and ignore most falsely detected gestures.

Other constraints test the informational compatibility of the input tokens, and the interpretation of gestures in the spoken context. A speech event such as "please bring me this cup", with the following semantic representation:

$$\begin{bmatrix} act\_bring \\ OBJ[cup] \\ DEICTIC[true] \end{bmatrix}$$

only allows (i) objects that can be carried by the robot and (ii) are instances of the class cup or it's subclasses, which are defined in the ontology.

V. System Integration

We have integrated the components described in this paper to demonstrate multimodal human-computer interation using speech, gestures and dialogue processing. Currently all components run on two laptops. One laptop is mounted and connected to the mobile platform; this laptop is used for real-time image processing tasks. The other laptop is used for speech and dialogue processing and is currently not on board of the platform. The computers are connected via a wireless LAN. All components of the human-machine interface communicate through a blackboard architecture with a socket communication over a central communication server. Each module has to register with its ID, whereby it is in addition also possible to subscribe to specific message groups. Figure 4 gives an overview over all the needed components. The scenario we addressed in our current

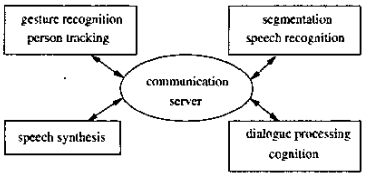


Fig. 4
Components of the system.

demonstration is a household situation, in which a user can ask the robot questions related to a kitchen, such as "'What is in the fridge ?", "What recipes would you recommend with the available items ?". A user could also ask the robot to set a table, to switch some lights on or off or to bring certain objects, such as cups. In our scenario the robot can locate and follow a user using the vision-based tracking system described in Section III, as soon as a person appears in the field of view of the robot's cameras.

VI. Conclusions and Future Work

In this paper we have presented our ongoing work in building technologies to improve natural human-machine interaction with human-friendly robots. We presented components for spontaneous speech recognition, multimodal dialogue processing and visual perception of a user. This included the recognition of pointing gestures of a user as well as the recognition of the user's head orientation, which is an important cue to deterime a person's direction of attention. We described how the components were integrated on a mobile robot platform and have been used for real-time human-robot interaction in a kitchen scenario.

Some of the presented components for human-computer interaction have already been integrated in a more sophisticated humanoid robotic plattform with two arms [3]. Within the German Humanoid robotics project, we are now working on improving the robustness of the presented components as well as we plan to integrate all these systems on a new humanoid torso with two arms. Other ongoing work involves the integration audio-visual person recognition and the development and integration of an attentional mechanism for the robot.

Acknowledgement

This research is partially funded by the German Research Foundaition (DFG) under Sonderforschungsbereich 588 - Humanoid Robots.

References

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

[2] *Special Issue on Human-Friendly Robots.* Journal of the Robotics Society of Japan, 1998, vol. 16, no. 3.

[3] T. Asfour, A. Ude, K.Berns, and R. Dillmann, "Control of armar for the realization of anthropomorphic motion patterns," in *The second IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS 2001)*, Waseda Universtiy, Tokyo, Japan, November 22-24 2001.

[4] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment," in *Proceedings of the ASRU*, Madonna di Campiglio Trento, Italy, December 2001.

[5] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal, "The Karlsruhe-VERBMOBIL Speech Recognition Engine," in *Proceedings of the ICASSP*, Munich, Germany, 1997.

[6] M. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," Camebridge University, Tech. Rep. CUED/FINFENG/TR291, 1997.

[7] M. Westphal and A. Waibel, "Model-Combination-Based acoustic mapping," in *Proceedings of the Internation Conference on Acoustics, Speech and Signal Processing ICASSP*, Salt Lake City, May 2001.

[8] J. Yang, W. Lu, and A. Waibel, "Skin-color modeling and adaption," Carnegie Mellon University, School of Computer Science, Tech. Rep. CMU-CS-97-146, 1997.

[9] T. Darrell, G. Gordon, M. Harville, and J. Woodfill, "Integrated person tracking using stereo, color, and pattern detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, 1998.

[10] R. Stiefelhagen, J. Yang, and A. Waibel, "Tracking focus of attention for human-robot communication," in *IEEE-RAS International Conference on Humanoid Robots - Humanoids 2001*, 2001.

[11] K. Nickel and R. Stiefelhagen, "Pointing gesture recognition based on 3d-tracking of face, hands and head orientation," in *International Conference on Multimodal Interfaces*, Vancouver, Canada, 2003.

[12] P. Gieselmann and M. Denecke, "Towards multimodal interaction with an intelligent room," in *Proceedings of Eurospeech*, Geneva, 2003.

[13] M. Denecke, "Rapid prototyping for spoken dialogue systems," *Proceedings of the 19th International Conference on Computational Linguistics*, 2002.

[14] H. Holzapfel and R. Stiefelhagen, "Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3d pointing gestures," in *Sixth International Conference on Multimodal Interfaces (ICMI)*, 2004.