Fritz - A Humanoid Communication Robot

Maren Bennewitz, Felix Faber, Dominik Joho, and Sven Behnke
University of Freiburg
Computer Science Institute
D-79110 Freiburg, Germany
{*maren, faber, joho, behnke*}*@informatik.uni-freiburg.de*

***Abstract*—In this paper, we present the humanoid communication robot Fritz. Our robot communicates with people in an intuitive, multimodal way. Fritz uses speech, facial expressions, eye-gaze, and gestures to interact with people. Depending on the audio-visual input, our robot shifts its attention between different persons in order to involve them into the conversation. He performs human-like arm gestures during the conversation and also uses pointing gestures generated with eyes, head, and arms to direct the attention of its communication partners towards objects of interest. To express its emotional state, the robot generates facial expressions and adapts the speech synthesis. We discuss experiences made during two public demonstrations of our robot.**

I. Introduction

Humanoid robots have become a popular research tool in recent years. More and more research groups worldwide develop complex machines with a human-like body plan and human-like senses [1], [2], [3]. One of the most important motivations for many humanoid projects is that such robots could be capable of intuitive multimodal communication with people. The general idea is that by mimicking the way humans interact with each other, it will be possible to transfer the efficient and robust communication strategies that humans use in their interactions to the man-machine interface. This includes the use of multiple modalities, like speech, facial expressions, gestures, body language, etc. If successful, this approach yields a user interface that leverages the evolution of human communication and that is intuitive to naive users, as they have practiced it since early childhood.

We work towards intuitive multimodal communication in the domain of a museum guide robot. This application requires interacting with multiple unknown persons. Here, we present the humanoid communication robot Fritz that we developed as successor to the communication robot Alpha [4]. Fritz uses speech, an animated face, eye-gaze, and gestures to interact with people. Depending on the audio-visual input, our robot shifts its attention between different persons in order to involve them into an interaction. He performs human-like arm gestures during the conversation and also uses pointing gestures generated with eyes, its head, and arms to direct the attention of its communication partners towards the explained exhibits. To express its emotional state, the robot generates facial expressions and adapts the speech synthesis.

The remainder of the paper is organized as follows. The next section reviews some of the related work. The mechanical and electrical design of Fritz is covered in Sec. III. Sec. IV details the perception of the human communication partners. Sec. V explains the robot's attentional system. The generation of arm gestures and of facial expressions is presented in Sec. VI and VII, respectively. Finally, we discuss experiences made during public demonstrations of our robot.

Figure 1. Our communication robot Fritz.

II. Related Work

Several systems exist that use different types of perception to sense and track people during an interaction and that use a strategy to decide which person gets the attention of the robot.

Spexard et al. [3] apply an attention system in which the person that is currently speaking/has spoken is the person of interest. While the robot is focusing on this person, it does not look to other persons who are not speaking in order to involve them also into the conversation. Okuno et al. [5] also follow the strategy to focus the attention on the person who is speaking. They apply two different modes. In the first mode, the robot always turns to a new speaker, and in the second mode, the robot keeps its attention exclusively on one conversational partner. The system developed by Matsusaka et al. [6] is able to determine the one who is being addressed in the conversation. Compared to our application scenario (museum guide), in which the robot is assumed to be the main speaker or actively involved in a conversation, in their scenario the robot acts as an observer. It looks at the person who is speaking and decides when to contribute to a conversation between two people.

Scheidig et al. [7] proposed to adapt the behavior of the robot according to the user's age, gender, and mood. They assume the robot to be focused on one person.

978-1-4244-1635-6/07/$25.00 ©2007 IEEE.

Kopp and Wachsmuth [8] developed a virtual conversational agent which uses coordinated speech and gestures to interact with a user in a multimodal way.

In the following, we summarize the approaches to human-like interaction behavior of previous museum tour-guide projects. Bischoff and Graefe [9] presented a robotic system with a humanoid torso that is able to interact with people using its arms. The robot does not distinguish between different persons and does not have an animated face. Several (non-humanoid) museum tour-guide robots that make use of facial expressions to show emotions have already been developed. Schulte et al. [10] used four basic moods for a museum tour-guide robot to show the robot's emotional state during traveling. They defined a simple finite state machine to switch between the different moods, depending on how long people were blocking the robot's way. Their aim was to enhance the robot's believability during navigation in order to achieve the intended goals. Similarly, Nourbakhsh et al. [11] designed a fuzzy state machine with five moods for a robotic tour-guide. Transitions in this state machine occur depending on external events, such as people standing in the robot's way. Their intention was to achieve a better interaction between the users and the robot. Mayor et al. [12] used a face with two eyes, eyelids and eyebrows (but no mouth) to express the robot's mood using seven basic expressions. The robot's internal state is affected by several events during a tour (e.g., a blocked path or no interest in the robot).

Most of the existing approaches do not allow continuous changes of the robot's mood. Our approach, in contrast, uses a bilinear interpolation technique in a two-dimensional state space [13] to smoothly change the mood.

III. The Design of Fritz

Our humanoid robot Fritz has been originally designed for playing soccer in the RoboCup Humanoid League TeenSize class [14]. He is 120cm tall and has a total weight of about 6.5kg. Its body has 16 degrees of freedom (DOF): Each leg is driven by five large digitally controlled Tonegawa PS-050 servos and each arm is driven by three digital Futaba S9152 servos. For the use as communication robot, we equipped Fritz with a 16DOF head, shown in Fig. 1. The head is mounted on a 3DOF neck. The eyes are USB cameras that can be moved together in pitch and independently in yaw direction. Six servo motors animate the mouth and four servos animate the eyebrows. The servo motors are controlled by a total of four ChipS12 microcontroller boards, which are connected via RS-232 to a main computer. We use a standard PC as main computer. It runs computer vision, speech recognition/synthesis, and behavior control.

IV. Perception of Communication Partners

To detect and track people in the environment of our robot, we use the two cameras and a stereo microphone. In order to keep track of persons even when they are temporarily outside the robot's field of view, the robot maintains a probabilistic belief about the people in its surroundings. In the following, we briefly describe the detection and tracking process. Details can be found in [15].

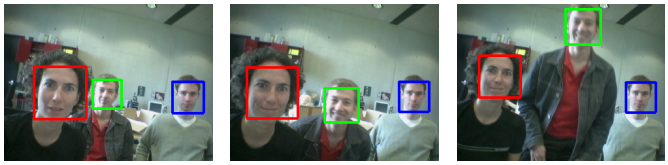

Figure 2. Tracking three faces.

A. Visual Detection and Tracking of People

Our face detection system is based on the AdaBoost algorithm and uses a boosted cascade of Haar-like features [16]. Whenever a new observation is made it must be determined to which person, that has already been detected by the robot, the newly detected face belongs. To solve this data association problem, we apply the Hungarian Method [17] using a distance-based cost function. We use a Kalman filter [18] to track the position of a face over time. Fig. 2 shows three snapshots during face tracking. As indicated by the differently colored boxes, all faces are tracked correctly.

To account for false classifications of face/non-face regions and association failures, we apply a probabilistic technique. We use a recursive Bayesian update scheme [19] to compute the existence probability of a face. In this way, the robot can also guess whether a person outside the current field of view is still there.

B. Speaker Localization

Additionally, we implemented a speaker localization system that uses a stereo microphone. We apply the Cross-Power Spectrum Phase Analysis [20] to calculate the spectral correlation measure between the left and the right microphone channel. Using the corresponding delay, the relative angle between the speaker and the microphones can be calculated [21].

The person in the robot's belief that has the minimum distance to the sound source angle gets assigned the information that it has spoken. If the angular distance between the speaker and all persons is greater than a certain threshold, we assume the speaker to be a new person, who just entered the scene.

V. Attentional System

It is not human-like to fixate a single conversational partner all the time when there are other people around. Therefore, our robot shows interest in different persons in its vicinity and shifts its attention between them so that they feel involved into the conversation. We currently use three different concepts in order to change the robot's gaze direction.

A. Focus of Attention

To determine the focus of attention of the robot, we compute an importance value for each person in the belief. It currently depends on the time when the person has last spoken, on the distance of the person to the robot (estimated using the size of the bounding box of its face), and on its position relative to the

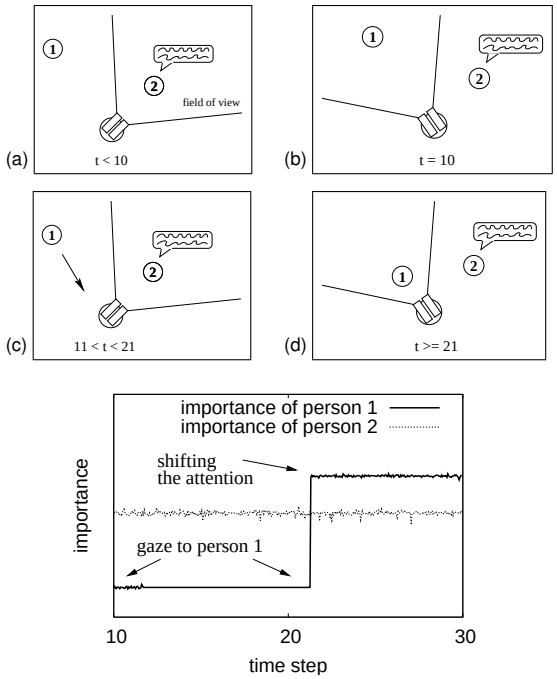

Figure 3. The images (a) to (d) illustrate the setup in this experiment. The lower image shows the evolution of the importance values of two people. During this experiment, person 2 is talking to the robot. Thus, it has initially a higher importance than person 1. The robot focuses its attention on person 2 but also looks to person 1 at time steps 10 and 21 to demonstrate that it is aware of person 1. At time step 21 the robot notices that person 1 has come very close and thus it shifts its attention to person 1, which has a higher importance now.

front of the robot. The resulting importance value is a weighted sum of these three factors. In the future, we plan to consider further aspects to determine the importance of persons, e.g., waving with hands.

The robot focuses its attention always on the person who has the highest importance, which means that it keeps eye-contact with this person. Of course, the focus of attention can change during a conversation with several persons. While focusing on one person, our robot also looks into the direction of other people from time to time to involve them into a conversation (see below).

B. Attentiveness to a Speaker

If a person that is outside the current field of view, which has not been detected so far, starts to speak, the robot reacts to this by turning towards the corresponding direction. In this way, the robot shows attentiveness and also updates its belief about the people in its surrounding.

C. Gazes outside the Focus of Attention

Since the field of view of the robot is constrained, it is important that the robot changes its gaze direction to explore the environment and to update its belief about it. Our robot regularly changes its gaze direction and looks in the direction of other faces, not only to the most important one. This reconfirms that the people outside the field of view are still there and involves them into the conversation.

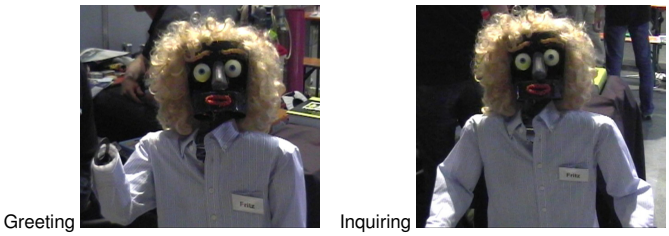

Figure 4. Robot Fritz performing two symbolic gestures with its arms.

D. Example

Fig. 3 illustrates an experiment that was designed to show how the robot shifts its attention from one person to another if it considers the second one to be more important. In the situation depicted here, person 2 was talking to the robot. Since person 2 had the highest importance, the robot initially focused its attention on person 2 but also looked to person 1 at time steps 10 and 21, to signal awareness and to involve him/her into the conversation. When looking to person 1 at time step 21, the robot then noticed that this person had come very close. This yielded a higher importance value for this person and the robot shifted its attention accordingly.

VI. Arm and Head Gestures

Our robot uses arm and head movements to generate gestures and to appear livelier. The gestures are generated online. Arm gestures consist of a preparation phase, where the arm moves slowly to a starting position, the stroke phase that carries the linguistic meaning, and a retraction phase, where the hand moves back to a resting position [22]. The stroke is synchronized to the speech synthesis module.

A. Symbolic Gestures

Symbolic gestures are gestures in which the relation between form and content is based on social convention. They are culture-specific.

- *Greeting Gesture:* The robot performs a single-handed gesture while saying hello to newly detected people. As shown in the left part of Fig. 4, it raises its hand, stops, and lowers it again.
- *Come Closer Gesture:* When the robot has detected persons farther away than the normal conversation distance (1.5-2.5m), it requests the people to come closer. Fig. 5 shows that the robot moves both hands towards the people in the preparation phase and towards its chest during the stroke.
- *Inquiring Gesture:* While asking certain questions, the robot performs an accompanying gesture, shown in the right part of Fig. 4. It moves both elbows outwards to the back.
- *Disappointment Gesture:* When the robot is disappointed (i.e., because it did not get an answer to a question), it carries out a gesture to emphasize its emotional state. During the stroke it moves both hands quickly down.
- *Head Gestures:* To confirm or disagree, the robot nods or shakes its head, respectively.

Figure 5. Fritz asks a person to come closer.

B. Batonic Gestures

Humans continuously gesticulate to emphasize their utterances while talking to each other. Fritz also makes small emphasizing gestures with both arms when he is generating longer sentences.

C. Pointing Gestures

To draw the attention of communication partners towards objects of interest, our robot performs pointing gestures. While designing the pointing gesture for our robot, we followed the observation made by Nickel et al. [23] that people usually move the arm in such a way that, in the poststroke hold, the hand is in one line with the head and the object of interest.

When the robot wants to draw the attention to an object, it simultaneously moves the head and the eyes in the corresponding direction and points in the direction with the respective arm while uttering the object name.

D. Non-Gestural Arm Movements

While standing, people typically move unconsciously with their arms and do not keep completely still. Our robot also performs such minuscule movements with its arms. The arms move slowly, with low amplitude in randomized oscillations.

VII. Emotional Expression

Showing emotions plays an important role in inter-human communication. During an interaction, the perception of the mood of the conversational partner helps to interpret his/her behavior and to infer intention. To communicate the robot's mood, we use a face with animated mouth and eyebrows to display facial expressions and also synthesize speech according to the current mood. The robot's mood is computed in a two-dimensional space, using six basic emotional expressions (joy, surprise, fear, sadness, anger, and disgust). Here, we follow the notion of the Emotion Disc developed by Ruttkay et al. [13].

A. Facial Expressions

Fig. 6 shows the six basic facial expressions of our robot. As parameters for an expression we use the height of the mouth corners, the mouth width, the mouth opening angle, and the angle and height of the eye-brows.

The parameters P' for the facial expression corresponding to a certain point P in the two-dimensional space are calculated by linear interpolation between the parameters E'_i and E'_{i+1} of the adjacent basic expressions:

$$P' = l(\mathbf{p}) \cdot (\alpha(\mathbf{p}) \cdot E'_i + (1 - \alpha(\mathbf{p})) \cdot E'_{i+1}). \quad (1)$$

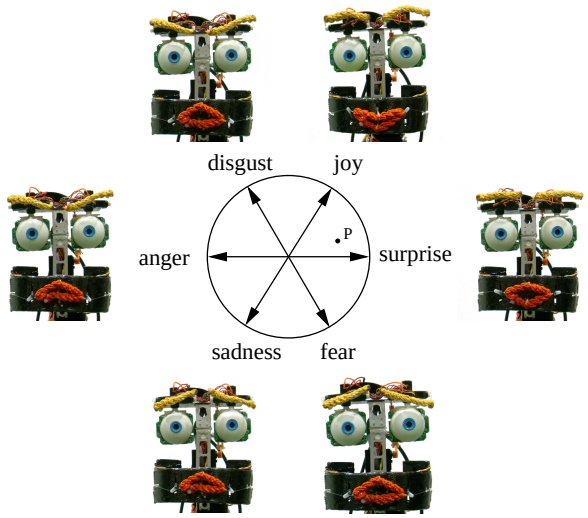

Figure 6. The two-dimensional space in which we compute the robot's mood. The images show the six basic facial expressions of our robot. The mood corresponding to a point P is computed according to Eq. (1).

Here, $l(\mathbf{p})$ is the length of the vector $\mathbf{p}$ that leads from the origin (corresponding to the neutral expression) to P, and $\alpha(\mathbf{p})$ denotes the normalized angular distance between $\mathbf{p}$ and the vectors corresponding to the two neighboring basic expressions. This technique allows continuous changes of the facial expression.

B. Emotional Speech Synthesis

In combination with facial expressions, we use emotional speech to express the robot's mood. Most speech synthesis systems do not support emotional speech directly; neither does the system we use (Loquendo TTS [24]). However, in this system, we can influence the parameters pitch, speed, and volume and thereby express emotional speech.

Cahn proposed a mapping of emotional states to the relative change of several parameters of a speech synthesis system [25]. She carried out experiments to show that test persons were able to recognize the emotion category of several synthesized sample sentences. In the mapping, she used the same six basic emotions that constitute the axes of the Emotion Disc. We use her mapping for the parameters average pitch, speech rate, and loudness to set the parameters pitch, speed, and volume of our speech synthesizer.

The mapping of emotional states to the relative change of

Table I
MAPPING OF EMOTIONS TO THE RELATIVE CHANGE OF THE SPEECH PARAMETERS. THIS TABLE CORRESPONDS TO MATRIX $\mathbf{M}$ IN EQ. 2. THE PARAMETERS WERE TAKEN FROM [25].

	joy	surprise	fear	sadness	anger	disgust
pitch	-3	0	10	0	-5	0
speed	2	4	10	-10	8	-3
volume	0	5	10	-5	10	0

the speech parameters can be seen in Tab. I. Let $\mathbf{M}^{3\times6}$ be such a mapping matrix, and $\mathbf{e}^{6\times1}$ be an emotion intensity vector of the six basic emotions. We can compute the three speech parameters as a vector $\mathbf{s}^{3\times1}$, as follows:

$$\mathbf{s} = \mathbf{d} + \mathbf{SMe}. \tag{2}$$

The vector $\mathbf{d}^{3\times1}$ contains the default values for the parameters and $\mathbf{S}^{3\times3}$ is a diagonal matrix used to scale the result of the mapping, thereby allowing for an adaption of the mapping to the characteristics of the synthesizer system. The emotion intensity vector contains only two non-zero entries, $l(\mathbf{p})\alpha(\mathbf{p})$ and $l(\mathbf{p})(1-\alpha(\mathbf{p}))$, that correspond to the influence factors of the two adjacent basic expressions of the current mood (see Fig. 6 and Eq. (1)).

Emotions influence many more characteristics of speech, e.g. breathiness, precision of articulation, and hesitation pauses. Hence, the three parameters used in our system can only roughly approximate emotional speech. In spite of these limitations, we experienced that even such simple adjustments can, in conjunction with facial expressions, contribute to the emotional expressiveness of our robot.

VIII. PUBLIC DEMONSTRATIONS

To evaluate our system, we tested our communication robots Alpha and Fritz in two public demonstrations. In this section, we report the experiences we made during these exhibitions.

We chose a scenario in which the communication robot presents four of its robotic friends. We placed the exhibits on a table in front of the robot. Our communication robot interacted multimodally with the people and had simple conversations with them. For speech recognition and speech synthesis, we used the Loquendo software [24]. Our dialog system is realized as a finite state machine (see [15] for details). With each state, a different grammar of phrases is associated, which the recognition system should be able to recognize. The dialog system generates some small talk and allows the user to select which exhibits should be explained and to what level of detail.

A. Two-Day Demonstration at the Science Fair 2005 in Freiburg

The first demonstration was made using the robot Alpha, the predecessor of Fritz. We exhibited Alpha during a two-day science fair of Freiburg University in June 2005. In contrast to Fritz, Alpha did not use emotional speech and performed pointing gestures with his arms but not any other human-like gestures.

Figure 7. Fritz presenting its robot friends to visitors at the Science Days.

At the science fair, we asked the people who interacted with the robot to fill out questionnaires about their interaction-experiences with Alpha (see [4] for more details). Almost all people found the eye-gazes, gestures, and the facial expression human-like and felt that Alpha was aware of them. The people were mostly attracted and impressed by the vivid human-like eye movements. To evaluate the expressiveness of the pointing gestures, we carried out an experiment in which the people had to guess the target of the pointing gestures. The result was that 91% of the gestures were correctly interpreted.

However, one limitation that was obvious, is that speech recognition does not work sufficiently well in noisy environments, even when using close-talking microphones. To account for this problem, in our current system, the robot asks for an affirmation when the speech recognition system is not sure about the recognized phrase.

B. Three-Day Demonstration at the Science Days 2006 in the Europapark Rust

In October 2006, we exhibited Fritz for three days at the Science Days in the Europapark Rust (see Fig. 7). Since the people at the previous exhibition were mostly attracted by the human-like behavior, we augmented the number of arm gestures as explained in Section VI. In general, the gestures served their purpose. However, the *come closer* gesture did not always have the desired result. In the beginning of the interaction, some people were still too shy and barely wanted to come closer to the robot. This effect is not uncommon even for human museum guides starting a tour. As soon as the visitors became more familiar with the robot, their shyness vanished and they choose a suitable interaction distance by themselves.

In contrast to the exhibition of Alpha, where toddlers often were afraid of the robot and hid behind their parents, we did not observe such a behavior with Fritz. This is probably due to the different sizes and appearances of the robots.

The kids found Fritz apparently very exciting. Most of them interacted several times with the robot. At the end, some of them knew exactly what the robot was able to do and had fun in communicating with Fritz.

When there were people around Fritz but nobody started to talk to the robot, Fritz proactively explained to the people what he is able to do. While speaking, he performed gestures with his head and arms so that, after the explanation, the

people had a good idea about the capabilities of the robot. This idea resulted from lessons learned of the first exhibition where people often did not know about the robot's actual capabilities.

Due to the severe acoustical conditions, speech recognition did not always work well. The affirmation request helped only if the correct phrase was the most likely one. Hence, for the next exhibition, we plan to employ an auditory front-end that focuses on the fundamental frequency of the speaker, in order to separate it from background noise [26], [27].

A video of the demonstration can be downloaded from http://www.NimbRo.net.

IX. Conclusion

In this paper, we presented our humanoid communication robot Fritz. Fritz communicates in an intuitive, multimodal way with humans. He employs speech, an animated face, eye-gaze, and gestures to interact with people. Depending on the audio-visual input, our robot shifts its attention between different communication partners in order to involve them into an interaction. Fritz performs human-like arm and head gestures, which are synchronized to the speech synthesis. He generates pointing gestures with its head, eyes, and arms to direct the attention of its communication partners towards objects of interest. Fritz changes its mood according to the number of people around him and the dialog state. The mood is communicated by facial expressions and emotional speech synthesis.

We tested the described multimodal dialog system during two public demonstrations outside our lab. The experiences made indicate that the users enjoyed interacting with the robot. They treated the robot as an able communication partner, which was sometimes difficult, as its capabilities are limited.

The experienced problems were mainly due to perception deficits of the robot. While speech synthesis works fairly well, robust speech recognition in noisy environments is difficult. This is problematic, because the users expect the robot to understand speech at least as well as it talks. Similarly, while the robot is able to generate gestures and emotional facial expressions, its visual perception of the persons around it is limited to head position and size. To reduce this asymmetry between action generation and perception, we are currently working on head posture estimation from the camera images and the visual recognition of gestures.

Acknowledgment

This project is supported by the DFG (Deutsche Forschungsgemeinschaft), grant BE 2556/2-1,2.

References

[1] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Mulanda, "Humanoid robots as cooperative partners for people," *Int. Journal of Humanoid Robots*, vol. 1, no. 2, 2004.

[2] D. Matsui, T. Minato, K. F. MacDorman, and H. Ishiguro, "Generating natural motion in an android by mapping human motion," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2005.

[3] T. Spexard, A. Haasch, J. Fritsch, and G. Sagerer, "Human-like person tracking with an anthropomorphic robot," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2006.

[4] M. Bennewitz, F. Faber, D. Joho, S. Schreiber, and S. Behnke, "Towards a humanoid museum guide robot that interacts with multiple persons," in *Proc. of the IEEE/RSJ International Conference on Humanoid Robots (Humanoids)*, 2005.

[5] H. Okuno, K. Nakadai, and H. Kitano, "Social interaction of humanoid robot based on audio-visual tracking," in *Proc. of the Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE)*, 2002.

[6] Y. Matsusaka, S. Fujie, and T. Kobayashi, "Modeling of conversational strategy for the robot participating in the group conversation," in *Proc. of the European Conf. on Speech Communication and Technology*, 2001.

[7] A. Scheidig, S. Müller, and H.-M. Gross, "User-adaptive interaction with social service robots," *KI Themenheft Learning and Self-Organization of Behavior*, no. 3, 2006.

[8] S. Kopp and I. Wachsmuth, "Model-based animation of coverbal gesture," in *Proc. of Computer Animation*, 2002.

[9] R. Bischoff and V. Graefe, "Hermes – a versatile personal robot assistant," *Proc. IEEE – Special Issue on Human Interactive Robots for Psychological Enrichment*, vol. 92, no. 11, 2004.

[10] J. Schulte, C. Rosenberg, and S. Thrun, "Spontaneous short-term interaction with mobile robots in public places," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 1999.

[11] I. Nourbakhsh, J. Bobenage, S. Grange, R. Lutz, R. Meyer, and A. Soto, "An affective mobile robot educator with a full-time job," *Artificial Intelligence*, vol. 114, no. 1-2, 1999.

[12] L. Mayor, B. Jensen, A. Lorotte, and R. Siegwart, "Improving the expressiveness of mobile robots," in *Proc. of IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN)*, 2002.

[13] Z. Ruttkay, H. Noot, and P. ten Hagen, "Emotion Disc and Emotion Squares: Tools to explore the facial expression space," *Computer Graphics Forum*, vol. 22, no. 1, 2003.

[14] S. Behnke, M. Schreiber, J. Stückler, H. Strasdat, and M. Bennewitz, "NimbRo TeenSize 2006 team description," in *RoboCup 2006 Humanoid League Team Descriptions, Bremen*, 2006.

[15] M. Bennewitz, F. Faber, D. Joho, M. Schreiber, and S. Behnke, "Multimodal conversation between a humanoid robot and multiple persons," in *Proc. of the Workshop on Modular Construction of Humanlike Intelligence at the Twentieth National Conferences on Artificial Intelligence (AAAI)*, 2005.

[16] R. Lienhard and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2002.

[17] H. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1, 1955.

[18] R. Kalman, "A new approach to linear filtering and prediction problems," *ASME-Journal of Basic Engineering*, vol. 82, no. March, 1960.

[19] H. Moravec and A. Elfes, "High resolution maps from wide angle sonar," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 1985.

[20] D. Giuliani, M. Omologo, and P. Svaizer, "Talker localization and speech recognition using a microphone array and a cross-powerspectrum phase analysis," in *Int. Conf. on Spoken Language Processing (ICSLP)*, 1994.

[21] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. Fink, and G. Sagerer, "Providing the basis for human-robot-interaction: A multimodal attention system for a mobile robot," in *Proc. of the Int. Confonference on Multimodal Interfaces*, 2003.

[22] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought: What Gestures Reveal About Thought*. University of Chicago Press, 1992.

[23] K. Nickel, E. Seemann, and R. Stiefelhagen, "3D-Tracking of heads and hands for pointing gesture recognition in a human-robot interaction scenario," in *International Conference on Face and Gesture Recognition (FG)*, 2004.

[24] Loquendo S.p.A., "Vocal technology and services," http://www.loquendo.com, 2007.

[25] J. Cahn, "Generating expression in synthesized speech," Master's thesis, Massachusetts Institute of Technology, 1989.

[26] D. Joho, M. Bennewitz, and S. Behnke, "Pitch estimation using models of voiced speech on three levels," in *32nd Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[27] S. Roa, M. Bennewitz, and S. Behnke, "Fundamental frequency estimation based on pitch-scaled harmonic filtering," in *32nd Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.