Playing a different imitation game: Interaction with an Empathic Android Robot

Frank Hegel, Torsten Spexard, Britta Wrede
Applied Computer Science
Faculty of Technology
Bielefeld University, Germany
{fhegel, tspexard, bwrede}@techfak.uni-bielefeld.de

Gernot Horstmann
Department of Psychology
Bielefeld University, Germany
gernot.horstmann@uni-bielefeld.de

Thurid Vogt
Multimedia Concepts and Applications
Institute for Informatics
Augsburg University, Germany
thurid.vogt@informatik.uni-augsburg.de

Abstract – **Current research has identified the need to equip robots with perceptual capabilities that not only recognise objective entities such as visual or auditory objects but that are also capable of assessing the affective evaluations of the human communication partner in order to render the communication situation more natural and social. In equivalence to Watzlawick's statement that "one cannot not communicate" [1] it has been found that also in human-robot interactions one cannot be not emotional. It is therefore crucial for a robot to understand these affective signals of its communication partner and react towards them. However, up to now, online emotion recognition in real-time, interactive systems has scarcely been attempted as apparently demands concerning robustness and time constraints are very high.**

In this paper we present an empathic anthropomorphic robot (torso) that mirrors the emotions happiness, fear and neutral as recognised from the speech signal by facial expressions. The recognition component as well as the development of the facial expression generation are described in detail. We report on results from experiments with humans interacting with the empathic robot.

I. Introduction

It has been suggested that anthropomorphic robots serve as an interface between man and technology [2] with the assumption that the more anthropomorphic a robot looks like the more the user will expect the robot to behave like a human counterpart. In accordance with this statement we base our research on the assumption that a human-like behaving robot is the easiest to use interface simply because humans are already highly skilled in having natural interaction with and communication to other humans. Furthermore, because of the communication interfacing function that the robot serves, users do not have to learn a new technical vocabulary in order to reach a goal when interacting with a technical device.

The underlying idea of the work we present in this paper is that motor mimicry is a simple yet powerful means to improve the (perceived) quality of human-robot interation (HRI). On our robots BARTHOC [3] and BARTHOC Jr. we use facial expressions as nonverbal social tools with the potential to improve the interaction. For human-human communication, motor mimicry has been described as a primitive form of empathy. Motor mimicry, which frequently occurs in interactions, has been interpreted to reveal information about relationships between communication partners, in particular about sympathy or empathy. From this point of view, a robot who is capable of mirroring the emotional expressions of a user may be interpreted as showing empathy. If the human counterpart feels emotionally "understood" by the robot it is not unreasonable to expect that the perceived quality of the human-robot interaction will be improved. In order to achieve a real gain in communication, however, the robot will need to develop adequate response mechanisms to the detected user's emotions. Mimicry would be one such response, although mimicry may not be adequate in every situation (e.g., not when the user is angry about the robot). As a first step towards this goal, we investigated the effects of motor mimicry by our robot BARTHOC Jr. on users.

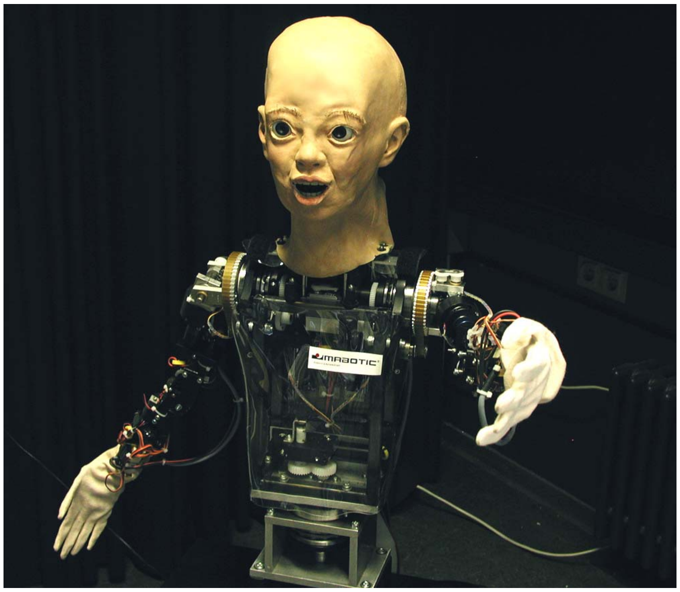


Fig. 1. The anthropomorphic robot BARTHOC Jr.
It has the size of a four year old child.

In an experiment we assessed the users' evaluations of the robot after they had interacted with the "empathic" robot that mirrored their perceived emotions via facial expressions. During the interaction the users read an abrigded version of the wellknown fairytale *Little Red Riding Hood* to BARTHOC Jr. They were instructed to read the fairytale in a vital and emotional way so that the robot would be able to correctly classify the emotions from the speech signal. The recognised emotion then triggered the display of the same facial expression. After the interaction the users were asked to fill out a questionnaire to evaluate the interaction with BARTHOC Jr. and to answer questions on whether or not they thought the robot was able to show the correct facial expressions to the different sections of the fairytale.

In the following section the function of motor mimicry is explained in more detail from a psychological point of view. In section three we present the hardware of our anthropomorphic robot we used to realise our experiement. Section four gives an overview of MiCo (Mimic Control) which controls the facial expressions used with BARTHOC Jr. EMO, the software we are using to classify emotions from speech signals, is explained in section five. The integration of MiCo and EMO in an XML-based Communication Framework (XCF) is pictured in section six and our experiment to study user evaluation of the empathic robot is described in section seven.

II. Related Work

A. Motor mimicry as a primitive form of empathy

Motor mimicry is a nonverbal response frequently occurring in interactions, where one person mimics behaviours of another, such as smiling at another's delight or showing pain at his injury [4]. As a matter of fact such a behavior is in some sense curious, because the individual's reaction is not appropriate to his or her actual situation as oberver but to that of the observed person. Sometimes, motor mimicry is essentially a form of mirroring the other persons behavior; for example a mother who is spoon feeding her baby can be observed to open her mouth shortly after the baby had opened its mouth [5]. In other instances, however, the observer responds apparently to the content of a verbal communication, or to other nonverbal cues to an emotional response of the observed person, for example, the tone of voice; these latter forms has been aptly termed cross-modal motor mimicry [6]. Motor mimicry obviously presupposes a process that aims to discern aspects of the private state of another person. Thus, motor mimicry has been described as a primitive form of empathy [5].

From a communicative act theroretical point of view, motor mimicry has been interpreted as revealing relationship information and that such nonverbal and analogical communication serve to define and reinforce the relationship. It is thus an analogically (or iconically) coded illustrator or emblem that is equivalent to the message "I know how you feel", which probably implies similarity: I can feel as you do; I am like you [5].

B. Displaying similarity to foster a close relationship.

Motor mimicry can occur quite frequently. In a study where participants told each other in some detail situations where something bad almost happened or the experience was not as bad as it could have been, motor mimicry was observed at an average rate of about 5 times per minute [6].

III. Hardware

We use the humanoid robot BARTHOC Jr. (see Fig. 1) for the evaluation of human-human communication and human-robot communication. BARTHOC Jr. is able to move its upper body like a sitting human and corresponds to a four year old child with the size of 65 cm from its waist upwards. The torso is mounted on a 65 cm high chair-like socket, which includes the power supply, the actuator controllers called *iModules*, and two serial interfaces to a desktop computer. One interface is used for controlling head and neck actuators, while the second one is connected to all components below the neck. The weight of the robot including its socket is sufficiently small to keep robot and socket easy to transport. The torso of the robot consists of a metal frame with a transparent cover to protect the inner elements. In total 41 actuators consisting of DC- and servo motors are used to control the robot. To achieve humanlike facial expressions ten degrees of freedom are used in its face to control jaw, mouth angles, eyes, eye brows and eye lids. The eyes are vertically aligned and horizontally steerable autonomously for object fixations. Each eye contains one FireWire color video camera with a resolution of 640x480 pixels. Besides facial expressions and eye movements the head can be turned, tilted to its side and slightly shifted forwards and backwards. In addition, two arms are mounted at the side of the robot. Each robot arm can be moved similar to the movement of a human arm. With the help of two five finger hands both deictic gestures and simple grips are realizable. The fingers of each hand have only one bending actuator but are controllable autonomously and made of synthetic material to achieve minimal weight. Besides the neck two shoulder elements are added that can be lifted to simulate shrugging of the shoulders. We used a headset for the audio recording although this is a temporary solution. A pair of microphones will be fixed at the ear positions as soon as an improved noise reduction for the head servos is available. By using different latex masks the appearance of BARTHOC Jr. can be changed for different kinds of interaction experiments from a male youngster to an old woman. For extended experiments we use BARTHOC [3], the second and taller version of the robot with the appearance of an adult.

IV. Facial Expression Module

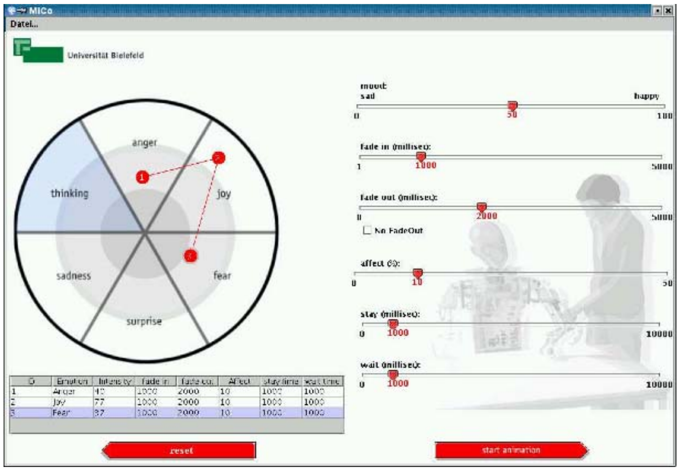
MiCo (Mimic Control) is an interface to control six different facial expressions in applications and to design facial animations with our anthropomorphic robots BARTHOC and BARTHOC Jr. We implemented and evaluated five basic emotional displays (happiness, fear, surprise, anger, sadness) as proposed by Ekman [7] and one further facial expression to represent a symbol for thinking. We assume that this expression will be especially helpful in human-robot interactions and situations where thinking conveys the information that the robot is currently processing the input from a user. Currently, the robot does not issue any reaction until it has finished computing the input and sometimes this causes communicative problems, because the user assume that the robot did not understand her or his messages and repeats or rephrases it. This can bring the interaction out of synchronisation since the robot will answer to the first input. Therfore, we assume that the display of a thinking face will help users in understanding the internal state of the robot much better.

Facial expressions have different proporties and they have to be used differently depending on context, application, and state of emotion. In our model these differences are represented by five parameters describing the facial expressions that have to be surrendered by an application: *FadeIn*, *FadeOut*, *Affect*, *Stay*, and *Wait*. Additionally, based on the idea of [8] we implemented a mood state (happiness versus sadness) and an emotion state representing the basic emotions. Mood and emotion state will be combined, e.g. the facial expression is most happy if the mood is most positive in combination with happiness.

With *FadeIn* the program defines how fast a facial expression will be elicited and with *FadeOut* how slow or fast it will leave. Usually, in a specific context of unexpectedness surprise should be faded in fast in order to be readable and believable. The parameter *Affect* defines the calculation between the mood and the presented emotion. For instance, if the value for *Affect* is low the mood only has little effect on the facial expressions representing an emotion, but if the value is high the mood has a major effect on the facial expression. *Stay* represents the time the specific facial expression will be shown and *Wait* is the time between the different facial expressions. Generally after showing an expression a neutral expression is shown, only if the value for *Stay* is null the expression moves directly from one to another.

In anthropomorphic robots the kind of a movement of facial expressions is an important factor. The movement of an animation is relevant to believable facial expressions and if the robot looks like a human it should nearly move like a human. In preliminary studies we found that it is not appropriate to use only linear or logarithmic movements but to combine both types. We use a linear movement for the first frames of an expression and logarithmical dynamics for fading out.

MiCo can be used in an XML-based Communication Framework (XCF) [9] or with the graphical user interface (GUI) [see figure 2]. While using XCF, MiCo can be addressed by different applications. Within the GUI we have the ability to design animations that can be saved and modified. This should be used to test the expressions on BARTHOC and BARTHOC Jr. just to know what values have to be specified in order to get the desired facial expression for a specific application or context. The GUI can be used intuitive just by clicking into the circle on the left. The circle represents the five basic emotions plus thinking and by clicking the selected points are connected by an animation path. Each point can be moved afterwards and modified by the sliders on the right to set the different parameters for the properties more precisely.



mood:
sad 0 — 50 — 100 happy

fade in (millisec): 1 — 1000 — 5000

fade out (millisec): 0 — 2000 — 5000

□ No FadeOut

affect (%): 0 — 10 — 50

stay (millisec): 0 — 1000 — 10000

wait (millisec): 0 — 1000 — 10000

ID	Emotion	Intensity	fade in	fade out	Affect	stay time	wait time
1	Anger	40	1000	2000	10	1000	1000
2	Joy	77	1000	2000	10	1000	1000
3	Fear	37	1000	2000	10	1000	1000

Fig. 2. GUI of the MiCo interface. Displayed is an animation path representing different facial expressions.

V. Emotion Recognition

The automatic recognition of emotions is currently a widely discussed topic in human-machine interaction. Speech is an obvious means for conveying emotion and has thus received much attention. However, up to now, online speech emotion recognition in real-time systems has scarcely been attempted. It is a great challenge for current methodology as apparently demands concerning robustness and accuracy are very high. Furthermore, in most related work, some features used to classify emotions rely on manual labeling such as phrase accent annotation or word transcription which is obviously not possible in a fully automatic system. In this section, we present our approach to emotion detection which is suitable for real-time recognition. First, feature calculation and classification are discussed and then we address the topic of segmentation which is particularly important for real-time recognition. Finally, we describe the database that was used for training the classifier.

A. Feature calculation and classification

The task of the feature calculation is to find those features that best describe the properties of the speech signal that convey emotions. As there is no agreement yet on an optimal set of features for speech emotion recognition, most approaches compute a high number of possibly redundant features and then select from this set the most relevant ones for the given task. Here, we computed features based on pitch, energy, MFCCs, the frequency spectrum, duration and pauses which resulted in a vector of 1316 features. Then, in order to reduce dimensionality and to improve and speed up classification, a sequential feature selection was applied ending up with 20 features related to pitch, MFCCs and energy. A more detailed description of the feature calculation exceeds the scope of this paper and can be found in [10]. For classification, a Naive Bayes classifier was used. Though this is a very simple classifier, it has the advantage of being very fast without performing much worse than more sophisticated classifiers such as support vector machines. For these reasons it was chosen for the experiment described in this paper.

B. Segmentation

Feature calculation and classification performed in this work are comparable to any research on automatic emotion recognition. The crucial point when it comes to online emotion recognition is the segmentation of the speech signal. Segmentation must be fast and result in meaningful, consistent segments. An important consideration is how much knowledge should be put into segmentation as it can be performed purely on the signal level, or on a linguistic level. Words or utterances are the most frequent units in offline emotion recognition. However, in online applications, word and utterance information have to be determined automatically, i. e. at least automatic speech recognition, if not even more high-level syntactic and semantic natural language processing is needed. As these systems do not yet perform very well on arbitrary speech and erroneous output could negatively influence the emotion recognition, we opted for voice activity detection as segmentation method, which relies on acoustic information only. In spontaneous speech, this coincides quite well with phrase breaks and a change of emotion is not likely to occur within such a segment. However, in this work, we were dealing with read speech so segments with voice activity tend to be longer and there is a risk to over-segment the speech.

C. Training data

The speech database used for training was recorded at the Technical University of Berlin [11]. It was originally designed for emotional speech synthesis and is thus of very high recording quality. Ten professional actors (five male, five female) were asked to speak ten sentences with emotional neutral semantic content in six different emotions (fear, anger, joy, boredom, sadness and disgust) as well as neutrally. On this database, high recognition accuracy can be achieved, as the recognition of acted emotions is by far easier than the recognition of spontaneous, real-world emotions. Since our setting is more realistic and conditions vary from the database recording conditions, results cannot be expected to be as good. But the design of the users' task in this work implies users speaking very expressively which led us to vote for acted emotions as training materials.

VI. INTEGRATION

For demonstrating the different classified emotional expressions on our robot the EMO-Module was extended with an interface to the XML-based Communication Framework (XCF) [9]. For any result of EMO a XCF function server of MiCo is invoked, accepting the parameters already described in IV. As EMO provides besides the pure classified emotion a value for the reliability, this parameter is directly used for the intensity of the emotion and after a down scale by a adjustable factor for the affect that it has on the general mood of the robot. The remaining values of MiCo, e.g. the time a mimic is displayed, have been fixed to standards as follows: fadeIn=1ms, fadeOut=500ms, stayTime=1200ms, and waitTime=0.

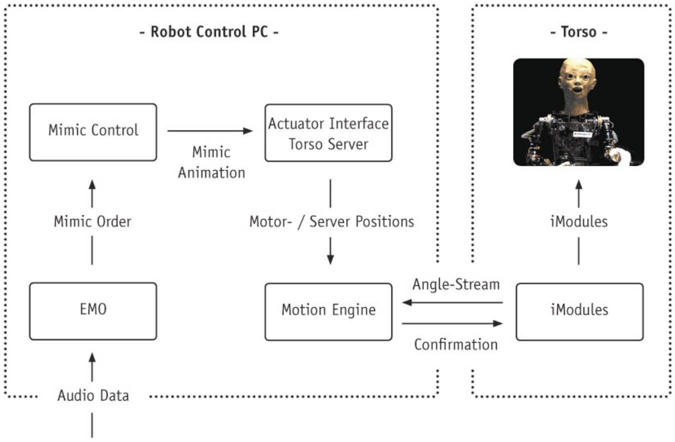


Fig. 3. Communication network for the integrated EMO

The generated mimic animations are send by XCF again to another module called Actuator Interface. It is used as a generic connection to the robot firmware and provides another XCF function server for the different motor commands. The motor commands are transmitted in a robot hardware specific manor to the robot firmware [3].

Using XCF and standard XML datastructures in all software modules, we are able to connect different modules to the robot control and to each other. E.g., an audio and vision based tracking and interaction framework [12] for HRI has already been implemented and will be appended by the described module EMO soon, all running in one integrated framework.

VII. EXPERIMENT

The aim of the presented study was to evaluate the degree of acceptance of the emotion representation with human communication partners. We carried out an experiment with 28 volunteers (13 females and 15 males). The age of the participants ranged from 18 to 35 years, with a mean of 24 years. The main target of this study was to examine whether cross-modal mimicry fosters an impression of a more natural interaction with the humanoid robot. A secondary target was to evaluate the emotion classification in an alternative way to that already carried out in [10].

A. Setting

After a short introduction to BARTHOC Jr., all participants were asked to sit at a table vis-a-vis the robot and to read out the fairytale *Little Red Riding Hood*, imagining they would read the tale to a child (see figure 4). The fairytale was shortened to 13 situations that covered the main plot. Each situation was represented by one or by two sentences. All situations were printed on separate pages. The participants were instructed to read each sentence and then pause to observe BARTHOC Jr.'s reactions. For each page a suggestion was made as to the emotional content of the situation. The suggestions were either neutral, fear, or happiness; these were also the expressions BARTHOC Jr. would show, given that the module EMO categorizes these verbally presented emotion correctly. Because "neutral" was also the robot's default facial display when he did not show any emotional expression (e.g., during the participant's utterances), a short head movement was executed to distinguish the default neutral expression from the classified neutral one. To examine whether a more emotional feedback was preferred by the interaction partner, only 17 of of the 28 participants interacted with BARTHOC Jr. with an active emotion feedback. The remaining 11 participants received only the same short neutral head movement (see above) for any utterance they made. Immediately after the experiment, which lasted on average five minutes including the introduction, the participants were asked to answer a number of questions in a separate room.

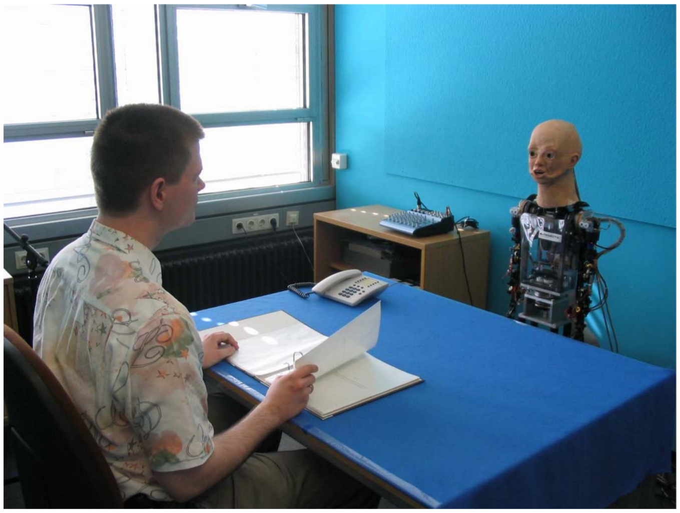


Fig. 4. Setting of the Experiment.
A participant is reading *Little Red Riding Hood* in front of BARTHOC Jr.

B. Results

The questionaire contained basically three blocks of questions. A first block was intended to assess whether BARTHOC Jr.'s responses were adequate to the social situation. Participants rated on separate 5-point scales the degree as to (1) BARTHOC Jr.'s facial gestures overall fit the situation, (2) whether BARTHOC Jr. recognized the emotional aspects of the story, and (3) whether BARTHOC Jr.'s response came close to a human counterpart. (The endpoints of the scales were labeled as not at all fitting / recognized / close, and very good fit / very good recognized, and very close, respectively). Figure 5 shows the mean ratings for each of the questions, separately for the mimicry and the neutral confirmation condition. Averaging over the three ratings, the mimicry condition fared significantly better than the neutral confirmation condition, t (26) = 1.8, p < .05 (one-tailed).

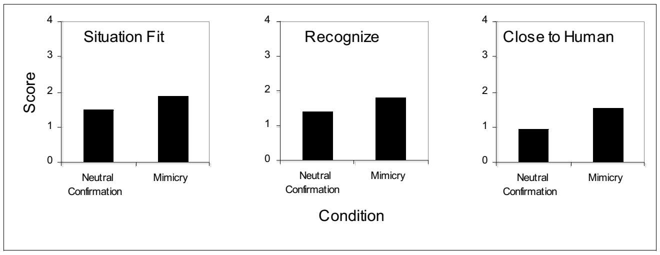


Fig. 5. Ratings of the overall fit to situation (left),the goodness of recognition (middle), and the closeness to human (right).

The second block of questions concerned individual facial expressions. Participants rated on a 5-point rating scale the degree as to which facial expressions happiness, anger, fear, disgust, surprise, and sadness as well a neutral expressions occurred too infrequently (-2), just right (0), or too frequently (+2). Figure 6 shows the results.

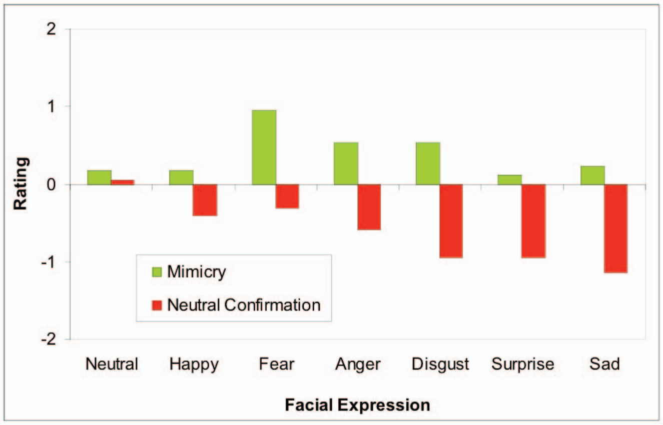


Fig. 6. Evaluation of separate facial expressions with regard to frequency of occurrence (-2=too infrequently, 0=just right, +2= too frequently).

There was a general tendency towards the "too frequently" pole of the rating scale in the mimicry condition and towards the "too infrequently" pole in the neutral confirmation condition, which was, however, only marginally significant, t (26) = 1.4, p = .085 (one-tailed).

The final block consisted of only one question that concerned the timing of BARTHOC Jr.'s response. The average rating on a 5-point scale (-2 = too early, 0 = just right, +2 = too late) was 0.4 and 0.1 for the mimicry and the neutral confirmation condition, respectively. Neither rating deviated significantly from zero (just right), although the mean ranting for the mimicry group approached significance, t (16) = 1.8, p = .08 (two-tailed). This result indicates that the timing of BARTHOC Jr.'s responses was quite good, but might appear more natural if BARTHOC Jr. responded a bit quicker.

VIII. Conclusion & Outlook

In this paper we presented an anthropomorphic robot able of cross-modal mimicry, that is, to recognize emotional content (happiness, fear and neutral) from speech and to mirror it by facial expressions. The user ratings obtained from user studies with 28 subjects interacting with the robot indicate that the emotional mimicry is perceived as the robot being able to react more adequately to emotional aspects of a situation ("situation fit") and to recognise emotion ("recognise") better as compared to a robot reacting without emotion recognition.

Based on this first experiment with emotional mimicry we are now able to (1) study in more detail psychological questions pertaining to the effects of facial expressions in communicative situations and to (2) build a more complex model of emotional communication in human-robot interaction. For these goals we will combine the emotion recognition and production modules with our grounding based dialog module, that is already running on the robot, in order to combine emotional with pragmatic information. With such a system at hand it will be possible to improve, and to gain deeper insights in the interactions between contextual factors as mirrored in the pragmatic dialog information and emotional interaction.

Acknowledgments

The research of this paper has been partially funded by the German Research Foundation (DFG) with the specific research area SFB 673 (C2 - Communication and Emotion).

References

[1] P. Watzlawick, J. Beavin Bavelas, D. D. Jackson, "Pragmatics of human communication: a study of interactional patterns, pathologies, and paradoxes", London: Faber, 1968.

[2] B. Duffy, "Anthropomorphism and the social robot, robot as partner: An exploration of social robots", in Proc. IEEE/RSJInt. Conf. on Intelligent Robots and Systems. EPFL, Switzerland: IEEE, 2002.

[3] M. Hackel, S. Schwope, J. Fritsch, B. Wrede, and G. Sagerer, "A humanoid robot platform suitable for studying embodied interaction", in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pages 56-61, Edmonton, Alberta, Canada, August 2005

[4] J. Beavin Bavelas, A. Black, C. R. Lemery, and J. Mullett, "I show how you feel?", in Journal of Personality and Social Psychology, vol. 50, pp. 322-329, 1986.

[5] J. Beavin Bavelas, A. Black, C.R. Lemery & J. Mullett, "Motor mimicry as primitive empathy", in N. Eisenberg & J. Strrayer (Eds.) Empathy and its development, pp. 317-338. Cambridge, UK: Cambridge University, 1987.

[6] N. Chovil, "Social determinants of facial displays", in Journal of Nonverbal Behavior, vol. 15, pp. 141-154, 1991.

[7] P. F. Ekman, W. V., and E. P., "Emotion in the human face: guidelines for research and an integration of findings". NewYork: Pergamon Press, 1972.

[8] J. J. Bryson, E. A. R. Tanguy, and P. J. Willis, "A layered dynamic emotion representation for the creation of complex facial animation", in Intelligent Virtual Agents, pp. 101-105, 2003.

[9] S. Wrede, J. Fritsch, C. Bauckhage, and G. Sagerer, "An XML Based Framework for Cognitive Vision Architectures", in Proc. Int. Conf. on Pattern Recognition, no. 1, 2004, pp. 757-760.

[10] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition", in Proc. IEEE Int. Conf. on Multimedia & Expo (ICME). 2005

[11] F. Burkhardt, A. Paeschke, M. Rolfes, and W. S. and Benjamin Weiss, "A database of german emotional speech", in Proc. of Interspeech, Lisbon, Portugal, 2005.

[12] T. Spexard, A. Haasch, J. Fritsch, and G. Sagerer, "Human-like person tracking with an anthropomorphic robot", in Proc. IEEE Int. Conf. on Robotics and Automation (ICRA). Orlando, Florida: IEEE, pp. 1286-1292, 2006.