Learning to Understand Multimodal Rewards for Human-Robot-Interaction using Hidden Markov Models and Classical Conditioning

Anja Austermann, Seiji Yamada

Abstract—**We are proposing an approach to enable a robot to learn the speech, gesture and touch patterns, that its user employs for giving positive and negative reward. The learning procedure uses a combination of Hidden Markov Models and a mathematical model of classical conditioning. To facilitate learning, the robot and the user go through a training task where the goal is known, so that the robot can anticipate its user's commands and rewards. We outline the experimental framework and the training task and give details on the proposed learning method evaluating the applicability of classical conditioning for the task of learning user rewards given in one or more modalities, such as speech, gesture or physical interaction.**

I. Introduction

THIS work describes a method to adapt a robot to its user through a cooperative training task. During the training phase, the robot learns to understand its user's way of giving positive and negative feedback to it using speech, gestures and the robot's built-in touch-sensors. The paper outlines the two essential elements of our learning method:

- A two-staged learning procedure based on Hidden Markov models and an implementation of classical conditioning. It is employed to learn to recognize the user's multimodal behavior patterns and to associate them with positive and negative rewards.
- A method to gather the necessary training data from the user in a training task without stressing or boring him and without putting the user into a situation, where he has to pre-record behaviors that he wants to use for communication in an artificial "recording" situation.

The proposed way of learning reward patterns and multimodal instructions through a training task has certain advantages over the current practice of using hard-wired commands for controlling a robot and giving feedback to it. It allows the user to give multimodal reward and commands naturally in his or her preferred way without restrictions concerning words, grammar or even the language used. Moreover, it transfers the necessary effort of learning and remembering the correct way of interacting from the user to the robot while at the same time avoiding the pitfalls of fully natural language processing. In contrast to other studies, the proposed approach does not aim at enabling robots to understand everyday conversation but at making a robot learn to deal with strictly task-related communication, like commands and rewards. The proposed method for user adaptation resembles the way, humans teach commands to pet dogs. It is designed to be integrated into a personal service-robot or pet-robot to enable it to adapt to the way its user wants to access its service or entertainment functions by natural multimodal commands or reward.

In our learning method, the robot and the user have to complete a cooperative training task, to adapt to each other. The training phase has to be completed before actually putting the robot into service. While the main goal of the training task is to enable the robot to learn the way, its user interacts with it, it also provides an environment for the user to learn how the robot expresses itself in a simple, easy-to-understand scenario where misunderstandings do not have any negative consequences.

In this paper, we are focusing on learning patterns that the user applies for giving positive and negative reward to the robot's behavior as a first step towards learning more complicated multimodal interaction patterns. Further explanations for this decision are given in section III.B.

In first experiments, we evaluated the second, conditioning based stage of our learning method and investigated on the effects of restrictions in allowed reward behavior.

II. Related Work

Techniques to acquire new words[3][6] or gestures[7] through human-robot interaction have been researched upon in recent years as a part of the research on symbol grounding for natural language acquisition.

Lee et al. described an approach [7] for the online-learning of human gestures in Human-Robot-Interaction. Their system is based on the online-training of Hidden Markov Models using gesture information recorded by a data glove. The system was able to recognize 14 gestures from the sign language alphabet with an error rate of 0.1 after observing 4 training instances of each gesture.

Iwahashi described an approach [6] to the active and unsupervised acquisition of new words for the multimodal interface of a robot. He applies Hidden Markov Models to learn verbal representations of objects, perceived by a stereo camera. The learning component uses pre-trained HMMs as a basis for learning and interacts with its user in order to avoid

Manuscript received March, 15th, 2008.
Anja Austermann is with the Graduate University for Advanced Studies (SOKENDAI) 101-0083, Tokyo, Japan, (e-mail: anja@nii.ac.jp).
Seiji Yamada is with the National Institute of Informatics and the Graduate University for Advanced Studies (SOKENDAI) 101-0083, Tokyo, Japan, (e-mail: seiji@nii.ac.jp)

978-1-4244-1823-7/08/$25.00©2008 IEEE

and resolve misunderstandings.

Kayikci et al. [9] use Hidden Markov Models and a neural associative memory for learning to understand short speech commands in a three-staged recognition procedure. First, the system recognizes a speech signal as a sequence of diphones or triphones. In the next step, the sequences are translated into words using a neural associative memory. The last step employs a neural associative memory to finally obtain a semantic representation of the utterance.

In the same way as the approaches, outlined above, our learning algorithm attempts at assigning a meaning to an observed auditory or visual pattern. However, the system is not trying to learn the meaning of individual words or symbols, but focuses on learning characteristic behavior patterns expressing commands or rewards as a whole. Those expressions can be words or gestures as in the studies above, but also prosodic patterns or utterances consisting of multiple words. Moreover, our proposed approach is not limited to a single modality such as only words or gestures, but tries to integrate observations from different modalities.

In this work, Hidden Markov Models are employed for the low-level modeling of speech-, prosody- gesture- and touch patterns. As a standard approach for the classification of time series data, Hidden Markov Models are widely used in literature. The use of Mel-Frequency-Cepstrum-Coefficients (MFCC) for HMM-based speech recognition is described in [17] Appropriate feature-sets for emotion/affective intent and gesture recognition are outlined in [4][10][11] and [7] respectively. Those tried and tested feature-sets are used in our work as an input for the HMM-based low-level learning phase.

For the high-level learning of associations between the meaning of commands and rewards and their appropriate Hidden Markov Model representations, classical conditioning is used. Mathematical theories of classical conditioning were extensively researched upon in the field of cognitive psychology. An overview can be found in [2].

The relation of classical conditioning to the phase of learning word meanings in human speech acquisition has been postulated in the book “Verbal Behaviour” by B. F. Skinner [12] and has been adopted and modified by researchers in the field of Behavior Analysis. In [13] Staats et al. describe an early approach to explain human learning of word meanings by classical conditioning. An explanation of more complex phenomena in learning word meanings by conditioning is described by B. Lowenkron in [8]. Our method is based on the psychological background, explained in these works. Our requirements for choosing an appropriate conditioning model to learn multimodal commands are outlined in section IV.C of this paper.

Recently, there have been several studies concerning the way that humans like to teach robots or other artificial creatures, such as virtual characters. However, as far as speech or gesture is used in these studies, it is typically restricted to fixed sentence patterns and a limited vocabulary of pre-trained gestures or words.

Yamada et al. described an approach to mutual adaptation between a human and an AIBO type robot based on classical conditioning using the Klopf neuron model [16]. While the robot learned to interpret the human's commands given by pressing one of the buttons on the robot's back, the human found out in the course of the experiments, how to correctly give commands to the robot.

The use of positive and negative reward from a human instructor to teach a robot was investigated upon in several studies. Lockerd et al. described an experimental setting for assessing human reward behavior and its contingency [14]. The participants of the study could give positive as well as negative reward to teach the virtual character Sophie to bake a cake in the "Sophie's World" scenario. Reward could be given by an interactive reward interface that allowed the user to assign any reward on a scale from -1 to +1 either to a certain object or to the world state. The character learned from a human teacher by this kind of reinforcement. In their experiments they found a strong bias towards positive reward and discovered a phenomenon that they described as anticipatory rewards, positive rewards that were assigned to an object that the character has to use in a later step. This kind of reward can be interpreted as guidance for the character.

III. The Training Phase

The training phase that is used to adapt the robot to its user and vice-versa has to fulfill certain requirements. As the robot does not have any prior knowledge about its user and his way of interacting at the beginning of the training phase, the training task needs to be specifically designed to allow the robot to anticipate what instruction and reward the user is going to give at any given time during task execution.

A. Requirements

Learning the meaning of user behavior by our leaning method is only possible within a task that has the following properties:

- The robot as well as the user know the target state in advance
- The order of steps leading from the current state to the target state is determined unambiguously by giving a target state.

In a scenario like that, the robot can anticipate the user's

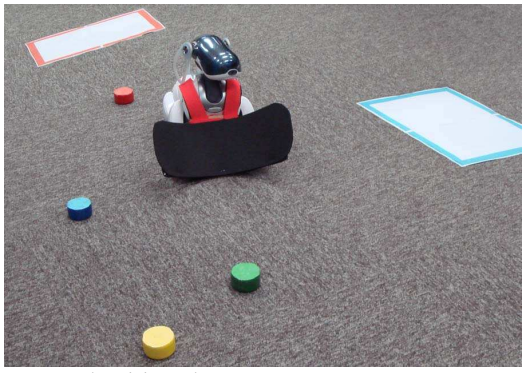

Figure 1: Sample training task

reward based on its knowledge about the target state and the current state of the task execution. It can provoke positive and negative reward by correct/incorrect behavior, that is, behavior that does or does not lead to target state, and it can prompt the user for the next instruction, if no interaction is observed.

The requirement that the order of steps for task execution must be fixed may appear quite strong. However, in a scenario, where the robot does not have any prior knowledge about the user and his way of giving commands, this is necessary, because otherwise, fine-grained instruction from the user might lead to a situation, where an action of the robot does not correspond to the user's instructions although it leads to the desired final state. For example, in a task with the target state "The red and the blue lamp are switched on", the user may give the detailed instruction "First, switch on the red lamp". Now, switching on the blue lamp first, becomes incorrect although it leads to the correct final state. Without understanding the user's instructions, there is no way for the robot to correctly deal with this kind of situation. Therefore the robot has to be able to infer the order of steps from knowing the target state.

A simple example for a training task, that possesses the desired properties and that was used our first experiment, is moving a colored object to a color-marked place. The experimental setting can be seen in Figure 1. When the object and the target place are known to the user as well as the robot, the steps leading to the goal are fixed and their order cannot be interchanged. While the user instructs the robot feely by means of his naturally used speech and gesture patterns, the robot anticipates the user's commands from its knowledge of the target state and the current state of the task execution and shows either compliant behavior and expects and learns positive reward or shows non-compliant behavior, expecting negative reward.

B. Target of Learning

The goal of our research is, to develop a method to learn commands as well as positive and negative rewards that can be used for controlling the service and entertainment functions of service-robots and pet-robots through the interaction with a user. We chose the learning of positive and negative rewards as the starting point of our work, because they are the smallest useful set of commands, that can be used to teach a robot, for example by reinforcement learning.

There are two main points, that make learning rewards easier than learning typical other commands:

Rewards are typically not parameterized. While commands can take several parameters, such as in "go <forward> <5 m>" or "put <the coffee-cup> <on the desk>" which need to be processed for understanding the command correctly, rewards can be understood as positive or negative without the need for processing parameterization. Therefore, a reward can be modeled as a single Hidden Markov Model while the processing of arbitrary commands, which is the next step in our work, needs to provide a means for combining HMMs modeling commands and their parameters and segmenting commands and parameters during the training phase.

Apart from that, rewards can only have two meanings: praise or punishment and while varying strongly between different people, the number of different positive and negative reward patterns used by an individual is limited. This leads to a reduction of the necessary amount of training data compared to the learning of more complex and numerous instruction patterns.

C. Assumptions

This research relies on the assumption that patterns of interaction between humans and robots range from rather universal ones, like pointing gestures, which are roughly the same between different individuals to highly individual patterns, like giving positive/negative reward. Patterns that are universal can be pre-trained and adapted to a certain user during task execution. Only patterns that vary substantially between users need to be trained in a training phase that precedes the actual use of the robot.

We further assume that each user has a limited inventory of interaction patterns to express a certain command or reward. The interaction patterns, that are typically used, can change slowly over time. Moreover, interaction patterns used by one user for the same instruction do not vary excessively between different tasks. The term "multimodal reward pattern" is used in this paper to refer to a time sequence of observations that possesses the following properties.

- It consists of perceptions in one or more modalities
- It begins by an increase in activity in one of the modalities (e.g. voice onset)
- It ends by a period of inactivity in all modalities
- Actions in different modalities occur in close timely relation, that is, at the same time or in a sequence, quickly following one another.
- The perceptions follow a behavior of the robot that can be clearly attributed a positive or negative value

IV. Learning Method

A. Overview

The learning algorithm is divided into two stages which are executed after each of the user's actions during the training phase. In the "reward recognition learning" stage, Hidden Markov Models are trained to recognize gestures, touch-sequences, utterances and prosodic patterns. In the "reward association learning" stage, the trained Hidden Markov models are associated with either positive or negative rewards, using a mathematical model of classical conditioning.

Extending a HMM-based recognizer by a second, conditioning-based learning stage has certain advantages and addresses problems that cannot be solved by HMMs alone. While it can benefit from the high performance of HMMs which are widely considered state-of-the-art for the classification of time series data, our model has to deal with the problem, that there is no one-to-one relationship between meanings – in this case, rewards – and their expression in

speech, prosody, touch or gesture. Even a single user employs multiple ways of expressing positive and negative rewards and there are expressions that can have a positive or negative meaning depending on the context, such as for instance, calling the robot's name. As an HMM can only represent a direct relationship of a sequence of observations, to its underlying most probable corresponding utterance or gesture, it is necessary to have a second stage of learning, where the connections between the different utterances, gestures and prosodic patterns, represented by Hidden Markov Models, and their meaning, in this case positive or negative feedback, are learned. We chose conditioning as a biologically inspired approach, as it has a number of desirable properties, which are outlined in section IV.C of this paper. Moreover, it is assumed, that a similar form of learning takes place, when dogs learn to understand commands from their caregiver.

An overview of the implemented system is given in Figure 2. The software is separated into the robot control software itself and a trainer program, which is tailored towards a specific training task. It possesses all information, needed to solve the task and evaluates the current situation that the robot is in. Based on its knowledge about the training task and its desired outcome, it sends commands and provides reward signals to the robot.

In analogy to conditioning in real animals, the reward from the trainer application can be interpreted as some immediately painful or pleasant signal which serves as an unconditioned stimulus (US). The robot software is able to learn the association between the user's behavior, represented by a trained HMM, which becomes the conditioned stimulus (CS) and the reward from the trainer program (US). In Figure 2, positive/negative rewards are denoted by +/-. The conditioned associations between the rewards and the HMMs for reward recognition are shown as dashed lines connecting rewards and HMMs in the right part of the image. An overview of the learning algorithm that is used to train the HMMs and associations is shown in Fig. 3. It is described in detail in sections IV.B and IV.C.

B. *Reward Recognition Learning*

A set of pre-trained HMMs for each of the four modalities speech, speech prosody, touch and gesture is created from pre-recorded sensor, audio and video data in order to minimize the need for training samples from each individual user. The initial HMM-set for speech recognition contains monophone models. The models are based on standard MFCC feature-vectors, extracted from the recorded speech data.

The HMM-set for prosody recognition is based on features extracted from the pitch and energy contours as well as the frequency distribution present in each frame of the speech signal that are typically used for recognizing emotion or affective intent in speech [4][11]. First trials are done with standard left-right HMMs but there is some evidence in literature [10] on emotion recognition from speech, suggesting that ergodic HMMs may be better suited for recognizing prosody.

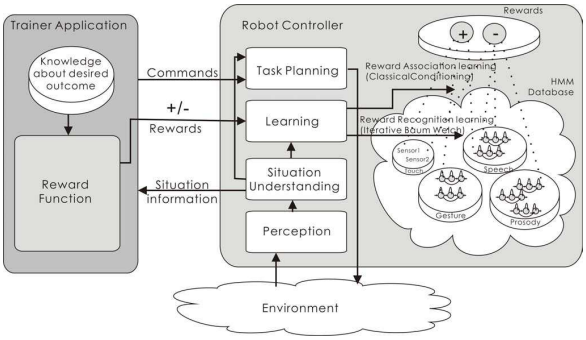

Figure 2: Overview of the system

The HMM-set for gesture recognition bases on features describing the relative positions of the hands and the face of a person. The positions are extracted from the stereo image of two cameras and tracked over multiple frames in order to determine the trajectories of the user's face and hands.

The HMM-set, that is used to learn touching patterns, such as stroking, patting or hitting the robot, uses feature vectors containing the values returned by AIBO's tactile head and back touch sensors, and by the paw touch sensors.

For each of the above described modalities, Hidden Markov Models are pre-trained, using an implementation of the Baum-Welch-algorithm. The pre-trained HMMs are stored separately for each modality in an HMM database.

During the training phase with human instruction, observed reward behavior in each of the modalities, is first processed by the Viterbi algorithm, in order to match it against the pre-trained HMM-Models in the HMM-database. In this stage, the matching is done on isolated "word" level: The full utterance, gesture or touch-sequence is assumed to correspond to one HMM in the database and matched against every single HMM in the HMM database. The output of the Viterbi algorithm is the best-fitting HMM along with a confidence value.

If the confidence value is above a threshold, the HMM is trained with the observed utterance/prosody/gesture and the application proceeds to the reward association learning phase. If the confidence value output by the Viterbi algorithm is below the threshold, the recognizer is executed again. This time, it is used as a continuous recognition based on an EBNF-like grammar describing the possible HMM-sequences for recognition. The sequence of HMMs resulting from this execution of the Viterbi recognition is merged into a new HMM. The new HMM is trained with the utterance/prosody/gesture and inserted into the HMM database for reuse.

The confidence threshold used for deciding whether a HMM fits the currently observed utterance depends on the number of instances already used to train the model and takes into account the associative strength present between the currently expected reward and the candidate HMM. Before comparing the confidence value to the threshold, the value is time-normalized to compensate for different utterance lengths. The HMM that results from this low-level classification and

learning stage, that is, the HMM providing the most accurate available model of the observed utterance/prosody/gesture serves as an input for the reward association learning.

C. Reward Association Learning

The reward association learning phase is based on the theory of classical conditioning, which was first described by I. Pavlov and originates from behavioral research in animals. In classical conditioning, an association between a new, motivationally neutral stimulus, the so-called conditioned stimulus (CS), and a motivationally meaningful stimulus, the so-called unconditioned stimulus (US), is learned. The unconditioned stimulus produces an unconditioned reaction (UR) as a natural behavior. After completing training, which is done by repeatedly presenting the conditioned stimulus just before the occurrence of the unconditioned stimulus, the conditioned stimulus is able to evoke the same reaction, when it is presented alone. This reaction is called the conditioned reaction. Pavlov found this relationship while he was doing experiments investigating the gastric function of dogs and measuring the amount of their salivation in response to food. At first the dog did not show any reaction to the tone of a bell (CS) but when the dog was given food (US), it salivated (UR). After repeatedly ringing the bell just before feeding the dog, the tone of the bell alone was able to make the dog salivate.

1) Relevant features of classical conditioning

For our task of learning multimodal reward patterns, certain properties of classical conditioning are of special importance. The most important features of classical conditioning for our application are blocking, extinction and second-order-conditioning as well as sensory preconditioning:

The term *blocking* denotes the phenomenon that occurs, when a CS1 is paired with a US, and then conditioning is performed for a second CS2 to the same US. In this case, the existing association between the CS1 and the US blocks the learning of the association between the CS2 and the US. The strength of the blocking is proportional to the strength of the existing association between the CS1 and the US.

For the learning of multimodal interaction patterns, blocking is helpful, as it allows the system to emphasize the stimuli that are most relevant. For instance, if a certain user always touches the head of the robot for giving positive reward, and sometimes provides different speech utterances together with touching the robot, then learning an association between positive reward and these speech utterances is blocked if there is already a strong association between touching the head sensor and positive reward.

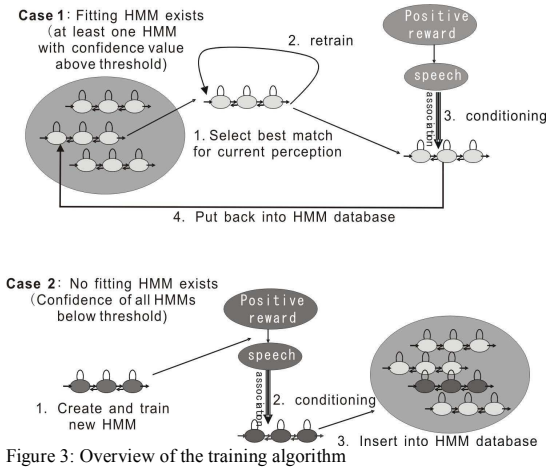

Figure 3: Overview of the training algorithm

Extinction refers to the situation, where a CS that has been associated with a US, is presented without the US. In that case, the association between the CS and the US is weakened. This capability is necessary to deal with changes in user behavior and with mistakes, made during the training phase, such as a misunderstanding of the situation by the human and a resulting incorrect reward.

Secondary preconditioning and *second-order conditioning* describe the learning of an association between a CS1 and a CS2, so that if the CS1 occurs together with the US, the association of the CS2 towards the US is strengthened, too. In *sensory preconditioning*, learning the association between CS1 and CS2 is established before learning the association towards the US, in *second-order conditioning*, the association between the US and CS1 is learned beforehand, and the association between CS1 and CS2 is learned later.

This property is important for our learning method, as it enables us, to learn connections between stimuli in different modalities, as well as to continue learning associations between stimuli given through different modalities even in situations, where no clear positive or negative feedback can be given by the trainer function as long as new stimuli, such as new gestures or commands are presented together with stimuli, that are already known and associated to a reward. E.g. a new positive speech feedback is uttered with a typical, known positive/negative prosody pattern.

2) The Rescorla-Wagner-Model

There are several mathematical theories, trying to model classical conditioning as well as the various effects that can be observed when training real animals using the conditioning principle. The models describe how the association between an unconditioned stimulus and a conditioned stimulus is affected by the occurrence and co-occurrence of the stimuli.

In our work, we employ the Rescorla-Wagner model [2], which was developed in 1972 and has served as a foundation for most of the more sophisticated newer theories. In the Rescorla-Wagner model, the change of associative strength of the conditioned stimulus A to the unconditioned stimulus US(n) in trial n, $\Delta VA(n)$, is calculated as in (2).

$$\Delta VA(n) = \alpha A \, \beta US(n) \, (\lambda US(n) - Vall(n)) \quad (2)$$

αA and $\beta US(n)$ are the learning rates dependent on the conditioned stimulus A and the unconditioned stimulus US(n) respectively, $\lambda US(n)$ is the maximum possible associative strength of the currently processed CS to the US(n). It is a positive value if the CS is present when the US occurs, so that the association between US and CS can be learned. It is zero if the US occurs without the CS. In that case, $\Delta VA(n)$ becomes negative. Thus, the associative strength between the US and the CS decreases. Vall(n) is the

combined associative strength of all conditioned stimuli towards the currently processed unconditioned stimulus. The equation is updated on each occurrence of the unconditioned stimulus for all conditioned stimuli that are associated with it.

One advantage of using conditioning as an algorithm for learning the associations between positive/negative reward and the user's corresponding behaviors is its rather quick convergence, depending on the learning rate.

In this study, the learning rates for conditioned and unconditioned stimuli are fixed values for each modality but can be optimized freely. They determine how quickly the algorithm converges and how quickly the robot adapts to a change in reward behavior. The maximum associative strength is set to one, in case the corresponding CS is present, when the US occurs, zero otherwise. The combined associative strength of all conditioned stimuli towards the unconditioned stimulus can be calculated easily by summarizing the pre-calculated association values of all the CS towards the US.

The major drawback of the Rescorla-Wagner-Model is that it is not able to model the effects of second-order-conditioning and sensory preconditioning directly. Therefore, we use a second pass of the Rescorla-Wagner-algorithm to learn associations between simultaneously occurring CS. In this second pass, the CS1 serves as the US for the conditioning of the CS2. In a third pass of the algorithm, we update the relation between the US and all CS2, that have an association to the actually occurred CS1, using a learning rate αA2, that is the product of the original learning rate αA and the associative strength between the CS1 and the corresponding CS2

D. Post-Processing

In order to avoid the number of items in the HMM database to grow too large and to improve recognition accuracy, post-processing steps are applied to the HMMs in the database after the training phase. During the post processing HMMs, which are similar in terms of mathematical distance [5] as well as in terms of their associations to the same rewards, are merged. HMMs that do not have a sufficiently strong association with any of the rewards are removed from the HMM database.

V. IMPLEMENTATION

The focus of the actual implementation of the system was to develop a framework for conducting experiments that is easy to extend and to adapt to new tasks. The framework utilizes an AIBO ERS-7 robot. AIBO is a dog-shaped pet robot which has roughly the size of a cat. It possesses 20 degrees of freedom and is able to communicate through sounds, an LED-panel in its face as well as body movements. It perceives its environment through various sensors, including stereo microphones, a 640x480 camera, two proximity sensors, 4 tactile touch sensors - one on its head and three on its back - 4 touch sensors in its paws and an accelerometer. It supports the 802.11b WLAN standard for wireless communication.

For the audio- and video recording we utilize a pair of Logitech Fusion webcams for stereo vision as well as a wireless lavalier microphone. AIBO is controlled using the Sony AIBO Remote Framework [1]. The software uses the HTK [17] as an implementation of Hidden Markov models and the relevant algorithms for recognition and training, as well as functions from the OpenCV[15] for video processing.

VI. EXPERIMENTS

In an experimental study, the participants were asked to instruct the robot to perform the task described in section 3 with different restrictions posed on the reward behaviors that they were allowed to use. The goal of this preliminary study was to get an insight in typical user behavior during a human-robot teaching task and record data for a first evaluation of the conditioning-based second phase of our learning method, which is presented in this paper. We also aimed at finding out, how restrictions in allowed reward modalities affect the frequency of reward given by the users and to understand the variability of rewards given by a single user in response to different robot behaviors. We further investigated in how far reward patterns, that a user selects and pre-records for interacting with a robot, resemble the ones that he employs, if no restrictions are given.

In the experiment, four participants, all of which male computer-literate graduate students aged between 25 and 35 without any prior experience of interacting with a pet robot, were asked to train the robot using the training task described in section 3. During the study, the robot was fully remote-controlled and all in all 109 minutes of audio and video data were collected, containing 141 reward instances – 64 positive and 77 negative ones. Figure 4 shows a sample of the video taken during the experiment.

1) Experimental Setting and evaluation method

The participants were provided with cards, showing which object was to be moved to what place and were told to instruct

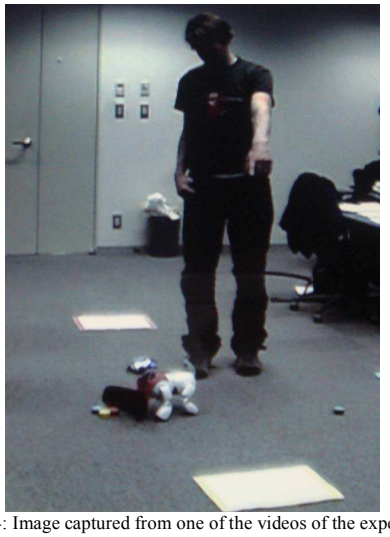

Figure 4: Image captured from one of the videos of the experiment

the robot freely and give rewards according to one of three scenarios at a time. The order of scenarios was changed for each participant to avoid sequence effects:

- *Free reward:* The user could chose freely in which way to give reward to the robot
- *Recorded reward:* The user had to record his preferred way of giving reward in advance and had to stick to it throughout the experiment.
- *Touch reward:* The user had to touch the robot's head sensor for positive and the robot's back sensor for negative reward.

During the experiments, the robot was remote-controlled to make different kinds of mistakes: simulating a technical failure to pick up an object, simulating a misunderstanding from the speech recognition and some unexpected behaviors such as deliberately sitting down, to provoke negative feedback. The different types of mistakes were balanced within every pass of the experiment. Positive feedback was expected for picking up the correct object, delivering it to the right place and recovering from an error. After the experiment, a questionnaire was provided to each participant to evaluate his experience throughout the interaction.

The data from the experiments in the "Free Reward" setting was transcribed manually and used for training the associations between actions and their meanings in the "reward association learning phase" of our learning method. Transcribing the data by hand replaces the training of the HMMs, which requires a larger amount of training data than we obtained during our first experiments. The data for training the actual HMMs is currently being gathered during a second series of experiments. Each of the user's actions was transcribed by a starting and an ending timestamp corresponding to the onset and the end of the voice/gesture/touch-based stimulus, the name of the modality (speech/gesture/touch) and the contents of the stimulus, that is, which word/sentence was uttered, which type of gesture was performed and which sensor was touched. We did not include information on prosody in our transcriptions.

The reward from the trainer function was described by a start timestamp, end timestamp and the keyword "positive", "negative" or "neutral". Reward from the trainer function was inserted after each observed reward pattern, independent from the observed user behavior, depending only on the state of the task execution at that time. This corresponds to the information that the robot can access in a real training task.

The keyword "positive" was assigned to the rewards given within the first 10 seconds after a subtask has been successfully finished. The keyword "negative" was assigned to every reward given while the robot was in an error state and had not yet started to correct it. In all other cases, the stimulus returned by the trainer function was "neutral"

For example, a situation, where the user says "good" and touches the head sensor of the robot within the first 10 seconds after successfully finishing a task or sub-task, such as the successful delivery of an object to its target place, is transcribed as follows:

```
0000001  0000210  speech  good
0000003  0000007  touch  head
0000015  0000025  feedback  positive
```

2) Results

The conditioning algorithm has been trained and evaluated separately for every user, using "leave-one-out"-cross evaluation. The total number of stimuli, which included any kinds of utterances, gestures or touch-actions, was 183 ranging from 26 to 56 between different users. Out of these 183 stimuli, 74 were neutral, 63 were negative and 46 were positive. These numbers differ from the number of positive and negative rewards, given above, as one reward can consist of multiple stimuli. The average accuracy for classifying between "positive", "negative" and "neutral" utterances was 81.38%. The confusion matrix can be seen in table 1. Most misclassifications occurred with neutral stimuli which were misclassified as either positive or negative while confusions between positive and negative stimuli were least frequent.

TABLE 1: CONFUSION MATRIX FOR THE CONDITIONING STAGE.

	Positive	Negative	Neutral
Positive	21,72%	1,32%	5,57%
Negative	0,00%	31,17%	4,31%
Neutral	2,88%	4,89%	28,51%

Left to right: expected classifications, Top to bottom: actual classifications

As for the user's reward behavior in the described training tasks, we found from the experiments, that the rewards which were recorded by the users for being used in the "recorded" scenario did not correspond as much as expected to the rewards given in the "free" scenario. Only 8 of 22 positive and 5 of 27 negative rewards were given in the same way as the one that the user considered as his "favorite" way when recording rewards for the "recorded" scenario.

But even within the touch and recorded scenarios, the participants did not stick to the designated reward behavior. Although they were clearly instructed not to use different rewards in the "recorded" and "touch" scenario, all participants gave additional or fully incorrect positive and negative reward. In case of the "touch" reinforcement principle, 44 rewards were given by the participants, 19 of them correct rewards, 17 rewards that contained an additional speech utterance or gesture and 8 rewards that did not at all include touching the head or back sensor, most of them providing a speech utterance only. Out of the 48 rewards given with the "recorded" reinforcement principle, 23 were correct, that is, using the same words and roughly the same gestures as recorded, 18 rewards contained additional speech utterances or gestures and 7 rewards were completely different from the recorded ones.

The amount of reward given varied with the allowed reward modalities, as can be seen in Figure 5. Most reward was given in the free reward scenario, where 47.9 percent of the positive situations and 96.4 percent of the negative situations were rewarded, while least reward was given in the touch reward scenario where a reward was given for 34.7

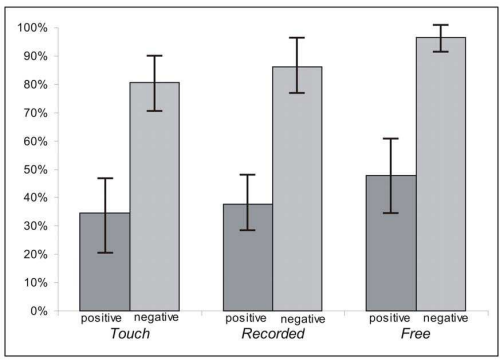

Figure 5: Percentage of positive and negative behaviors of the robot that result in positive/negative reward

percent of the positive and 80.5 percent of the negative situations. The reason for the apparent negative bias in rewards is that most of the time positive reward was only given for reaching the final goal but not for reaching sub-goals like approaching the correct object or correcting a mistake. The robot did not stop and wait or prompt the user for reward in these situations. On the other hand, incorrect performance of the robot was quite obvious and therefore typically resulted in a negative reward.

In a questionnaire, the participants could rate their agreement with different statements concerning their interaction with the robot on a scale from 1 (completely agree) to 5 (completely disagree). For the statement "I was able to instruct the robot in a natural way", the free reward received a rating of 1.25 (SD=0.3), recorded reward was rated 3.25 (SD=0.3) and touch reward was rated 4.35 (SD=0.6). For the statement "I would like to interact with a real service robot in the same way", the ratings were 1.5 (SD=0.3) for free reward, 2.0 (SD=0) for recorded reward and 4.5 (SD=0.6) for touch reward.

VII. Discussion

Although the small number of participants does not allow for a sound statistical analysis, the results suggest that users are quite sensitive to restrictions in applicable reward behaviors. Therefore techniques that allow reward and instructions to be given to a robot as freely as possible would be desirable.

Moreover, our data shows, that pre-recording fixed patterns for giving reward does not suffice to enable a user to provide reward to a robot in his or her preferred way.

As results from the HMM based first stage of our learning method, are still missing, the results from the conditioning phase can only be seen as an upper boundary for the performance of the whole learning method. The main cause of misclassifications of the conditioning phase were utterances, that were used only once within our data. This problem was caused by the small amount of training data, we had gathered within our first experiment. It is being addressed by currently ongoing experiments with a larger number of participants and a changed experimental setting. The new setting is based on a game task that allows the user to give more positive and negative rewards in a short time by minimizing the interaction-free task execution time, which is mainly caused by the slow walking movements of the AIBO pet robot.

REFERENCES

[1] AIBO Remote Framework http://openr.aibo.com
[2] C. Balkenius, J. Morn. "Computational models of classical conditioning: a comparative study." Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior, 1998
[3] D. H. Ballard, C. Yu, "A multimodal learning interface for word acquisition" , 2003 Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003.
[4] C. Breazeal, Recognition of Affective Communicative Intent in Robot-Directed Speech Autonomous Robots Volume 12 , Issue 1, January 2002, 83 - 104
[5] M. Falkhausen, H. Reininger, D. Wolf, "Calculation of Distance Measures between Hidden Markov Models", In Proc. Eurospeech, pages 1487--1490, 1995
[6] N. Iwahashi, "Active and Unsupervised Learning for Spoken Word Acquisition Through a Multimodal Interface", RO-MAN 2004 13th IEEE international workshop on robot and human interactive communication
[7] C. Lee, Y. Xu, "Online Interactive Learning of Gestures for Human/Robot interfaces", IEEE Int. Conf. on Robotics and Automation, pp 29822987, 1996.
[8] B. Lowenkron, "Word meaning: A verbal behavior account", Annual convention of the Association for Behavior Analysis, Washington DC, May, 2000
[9] Z. K. Kayikci, H. Markert, G. Palm, "Neural Associative Memories and Hidden Markov Models for Speech Recognition" , Proceedings of the IJCNN 2007
[10] A. Nogueiras, A. Moreno, A. Bonafonte, José B. Marino, "Speech Emotion Recognition Using Hidden Markov Models", Proceedings of Eurospeech 2001
[11] T. L. Nwe, S. Foo, S. Wei, L. De Silva, "Speech emotion recognition. using hidden Markov models", Speech communication 41,4, 2003
[12] B. F. Skinner "Verbal Behavior" Copley Publishing Group, 1957
[13] C. K Staats, A.W. Staats, "Meaning established by classical conditioning". Journal of Experimental Psychology, 1957, 54, 74-80
[14] A. L. Thomaz, C. Breazeal. "Reinforcement Learning with Human Teachers: Evidence of feedback and guidance with implications for learning performance." In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI), 2006.
[15] The Open Computer Vision Library (OpenCV) - http://opencvlibrary.sourceforge.net
[16] S. Yamada, T. Yamaguchi, "Training AIBO like a dog - preliminary results", Proceedings of the IEEE Workshop on Robot and Human Interactive Communication, 2004. ROMAN 2004, 431- 436
[17] S. Young et al., "The HTK Book" HTK Version 3, 2006 http://htk.eng.cam.ac.uk/