Proceedings of the 2008 International Conference on Cognitive Systems
University of Karlsruhe, Karlsruhe, Germany, April 2-4, 2008

Learning and Verification of Names with Multimodal User ID in Dialog

Hartwig Holzapfel and Alex Waibel

Abstract— **Acquiring new knowledge is a key functionality for humanoid robots. By envisioning a robot that can provide personalized services the system needs to detect, recognize and memorize information about specific persons. Recent work already shows promising results in the area of speech recognition, voice identification and face identification that enable a system to reliably detect and recognize persons, as well as approaches to interactively learn to know new persons in dialog acquiring their names and ID information.**

One problem in this area is verification, namely to detect which person is known versus which person is unknown; a second problem is the learning phase, namely to learn the name of a person and store it in a database with associated face and voice classifier information. This paper presents work to interactively acquire ID information, combining both of the above problems into one learning dialog. In dialog we combine multimodal input including spoken name recognition, name pronunciation (phoneme recognition), name spelling (grapheme representation), face identification and voice identification and seek to build dialogs optimized to verify or learn a person's name and ID. For designing and training of optimized dialogs we use a reinforcement learning approach and propose a multimodal simulation modeling the user's actions and multimodal ID recognition components including stochastic error models.

I. Introduction

In this paper we present work on learning names and person ID information in a multimodal dialog system for a humanoid robot. One part of the dialogs that can be conducted with the robot are dialogs to identify and especially to learn to know new persons. We have conducted experiments with a receptionist scenario, where one task of the robot receptionist was to identify the visiting person or learn the name of the person if unknown. In the following we present efforts on especially this task namely isolated identification dialogs within the receptionist scenario. These dialogs fulfill two purposes: In case the person is known, confirm the name of the person. In case the person is unknown, classify the person as unknown and conduct a learning dialog to obtain the person's name.

The presented experiments make use of standard perceptual components available on a humanoid robot. These components are visual perception with a stereo camera and acoustic perception with distant and close-talk microphones. Visual perception provides face detection and identification. Acoustic perception provides voice identification and speech recognition including name recognition, spelling and phonetic understanding. These components provide recognition hypotheses which are interpreted by the dialog manager.

The challenge of this task is to define a dialog strategy, including when to confirm ID information, when to ask for name pronunciation or spelling. With the goal of optimizing dialogs regarding success, length, and subjective measures, we have implemented a reinforcement learning approach which combines both verification and learning into one dialog integrating the multiple input modalities presented above. For achieving this goal, we implemented a first rule based dialog strategy, and later a reinforcement learning strategy, which was trained in a multimodal user simulation. In the following we present the setup for multimodal integration in dialog, definition of the handcrafted strategy and learning of dialog strategies in the multimodal user simulation. Both dialog strategies are evaluated within the simulation and are compared against each other. First results from a real user experiment are reported.

A. Related Work

The main goal of this work is to build a strategy to learn names and associated multimodal ID information in dialog. Related work can be found in the area of name recognition, multimodal person identification, learning dialogs and optimization of dialog strategies.

Name recognition requires large vocabulary handling which is addressed with unknown word (OOV) detection [1], [2], [3], [4] and dynamic vocabulary approaches. [5] describes a system with dynamic vocabulary that can be updated according to the given context. [6] presents learning of new words in a multimodal setting. [7] uses multiple recognition passes with a phone-based OOV word-model in the first step, and a constrained vocabulary in the second step. Attention is also required for obtaining a phonetic representation of a name which can be used to understand the user's name and to pronounce the name. [8] combines phoneme recognition of spoken input with telephone keypad input to obtain textual representation of names. [9] implements fusion of spoken and spelled names on large vocabularies.

While the main aspect of learning in this work is to obtain a name for a person and associated multimodal features, it has similarities to learning dialogs like names of objects or semantic properties, e.g. [10], [11], [12].

Besides name learning and person identification using names, the presented system also integrates face and voice ID such as used in smart home environment or surveillance tasks [13], [14], [15]. Recent work which uses voice identification in dialog with a dialog strategy trained by reinforcement learning is presented in [16]. More work related to reinforcement learning in dialog systems is described later.

H. Holzapfel is with interACT, Faculty of Computer Science, University of Karlsruhe, Germany hartwig@ira.uka.de

A. Waibel is director of the joint center interACT, Faculty of Computer Science, University of Karlsruhe, Germany and Carnegie Mellon University Pittsburgh, PA, USA waibel@ira.uka.de

II. Scenario and System Components

A. Name Learning in Dialog

The term learning dialog refers to the objective of a dialog to acquire some specific information through interaction with the user and update a background knowledge base. The presented dialog task to learn the name of a person is associated with face ID and voice ID information which together is stored in a database. Concerning terminology, we refer to this kind of dialog as a task of knowledge acquisition or as a learning dialog. The outcome of the dialog in our task is to have obtained a label for a person, which can be stored in a database along with collected audio and video data and the person's name. If a known person interacts with the system, the corresponding database entry can be updated along with additional identification data. If the person is unknown, a new entry is created with new ID information for face ID and voice ID, the name is added to the database, speech recognition and understanding grammars, and speech recognition dictionary. For extending the dictionary, the text-to-speech engine's grapheme-to-phoneme conversion is used, which is the same that was used to get the name confirmed in dialog. Note that the presented dialog scheme doesn't allow to fully verify an ID of a person, but only the learned label. That is, if two persons have the same name, they cannot be directly distinguished by having their names confirmed.

The whole dialog system comprises speech recognition, spelling recognition, natural language understanding, voice identification, face detection and face identification, grapheme to phoneme conversion, and text to speech. These components are introduced in this section, most of them only with a brief overview where necessary to understand the remainder of the paper. For more details on these components please refer to the referenced publications.

B. Speech recognition

Speech recognition is performed with the Janus speech recognition software with the Ibis decoder [17] on utterances which have automatically segmented from close-speech or distant speech input.

Standard speech recognition is performed using context free grammars as language models. For phoneme and spelling recognition statistical n-gram models are applied. All recognition models share the same runtime engine and acoustic models, including unsupervised user adaptation during runtime. The speech recognizer is tightly coupled with the dialog manager. This allows to weight grammar rules based on the dialog context to improve speech recognition in context. It is also used to switch between standard recognition and spelling. In previous work, the grammar-based speech recognizer has been extended to also cover unknown word detection [1]. In the experiments presented here, speech recognition and especially name recognition is performed with a vocabulary created from the names of all persons known to the system. We have measured correct detection rates of over 70% for unknown names (OOV-detection) on the given databases and around 95% correct recognition of known names. The databases in these experiments have been 15-30 persons. The Cepstral text-to-speech engine was used for spoken output and also for grapheme-to-phoneme conversion of spelled names, which is stored in combination with phoneme recognition of the spoken name for speech recognition.

C. ID recognition components

Person identification is made possible by the face ID and voice ID components [15], [18] and spoken name recognition. Visual perception is done via a stereo camera mounted on a pan-tilt unit. The robot can track and follow a person with his visual field. We apply single image classification using k-nearest-neighbor classifiers and fusion over image sequences. The approach allows to easily extend the database in an online system without retraining of the whole classifier. Sequence classification only takes into account inherent properties of sequences and doesn't need to be changed when new samples or new persons are added to the database. The two-stage classifier is further used in the simulation, where single images are precomputed and only the sequence hypothesis is computed during runtime. For details about face identifier setup, please refer to [18]. A similar approach is adopted for voice ID. With the difference that in the current setup, sequence classifications over multiple turns is done by concatenating the audio files.

D. Dialog Setup

The scenario for the dialog manager is to control interaction in a receptionist scenario. We conducted a first Wizard-of-Oz experiment with users that were given the task to act as a visitor and to deliver a parcel to a predefined person. The task of the robot was to greet the visitor, ask for the concern (which was to deliver a parcel), who to deliver this parcel to, ask for the visitor's name to be able to announce him to the receiver, and finally to give direction where to go to deliver the parcel.

The Wizard-of-Oz experiment served as a data collection and analysis of the dialog task. From this analysis, the receptionist task was decomposed into the dialog modules greeting, parcel reception, name learning, directions and goodbye. A second, standalone experiment was then conducted without human intervention except starting and stopping recordings. Changes were made only to the dialog implementation, the perceptual components from the Wizard-of-Oz experiment could be used with only slight modifications.

The dialog strategy is implemented in a hierarchy with one top level strategy to control a situation model and selection of different dialog modules. Each dialog module provides its own strategy implementation, which for the experiment was implemented by a separate rule-based strategy. Separating the dialog modules' strategies makes it easy to also develop each strategy separately and later combine the modules into the dialog model for the whole task. The experiments with the system have shown that the modular design was feasible [19] and that the name learning task could be approached in an isolated manner.

The handcrafted rule-based strategy for the name learning module shall serve as a baseline for the experiments presented in this paper with reinforcement learning. The actions which are available for the dialog strategy are the same for both the handcrafted as well as the learned strategies. The following actions are available:

get information	ask_name, ask_name-spelling
confirm	conf_name-asr, conf_name-spelling conf_faceID, conf_voiceID
finish dialog	accept-name, abort

The implementation of the handcrafted strategy follows a simple pattern: Alternatively ask the visitor for his name or for the spelling of his name. If either of both are given, try to confirm the name. When speech recognition reports an unknown name ask for spelling. In the beginning, if face ID produces a hypothesis, try to confirm the associated name. If neither is set but voice ID is given try to confirm the associated name. Do not ask for the same name twice. As soon as the name is confirmed quit the dialog and store the name. If a predefined threshold of turns is reached e.g. 15 (in the simulation), or after 3 unsuccessful confirmation questions (in the online experiment), the dialog is aborted without storing the name.

III. Optimized Learning Dialogs

This section describes a reinforcement learning approach to automatically acquire a dialog strategy which is optimal with regard to a predefined metric, the reward function. The design of single modules separates concerns and allows training of the name learning module, which can be conducted in reasonable training time and in isolation of other dialog concerns.

One promising approach for optimization of dialog strategies in general is with reinforcement learning. The idea of reinforcement learning is that one cannot define correct and incorrect actions for each state as in supervised learning, but rather to expose the system (usually referred to as the agent) to an environment in which it can take a series of actions, where each action is associated with some reward. The agent is supposed to learn from these observations and optimize its expected reward. Reinforcement learning in this definition is a class of learning problems. Various systems exist that apply reinforcement learning, and several algorithms exist to solve the reinforcement learning problem [20].

For dialog systems, reinforcement learning has successfully been employed to learn dialog strategies. A problem with this approach has been that reinforcement learning requires large amounts of data, so that training strategies on real data has usually been conducted with a limited state space and/or action space. Recent work has targeted reinforcement learning in dialog systems using a simulated user and statistical models trained from a data corpus.

A. MDP State Model

One important aspect of defining a model for reinforcement learning is the state model. In theory it has been shown that if the state model fulfills the Markovian Property, Q-Learning converges to the optimal policy. The Markovian Property requires that the state transition probability only depends on the current state:

$$P(s_{t+1}|s_t, a_t) = P(s_{t+1}|s_t, s_{t-1}, ..., s_0, a_t, a_{t-1}, ..., a_0)$$

In practicable applications however, and especially in dialog, this property usually doesn't hold, but still good policies can be found. A tradeoff needs to be found between encoding fine grained information and history versus simple models, to find a model which can be computed with the given data and in a reasonable amount of training runs.

Information Slot	MDP state	MDP state values
ASR name input	Name-ASR	empty, filled, oov
Spelling input	Name-Spelling	empty, filled
Voice ID	VoiceID VoiceIDConf	empty, filled low, medium, high
Face ID	FaceID FaceIDConf	empty, filled low, medium, high

TABLE I

Dialog information slots and mapping to MDP states

MDP state	values	description
nNameFailed	0,1,2+	number of failed attempts to confirm a name
nNameConf	0,1+	number of successful attempts to confirm a name
nASRNameFailed	0,1,2+	number of failed attempts to confirm a name from speech recognition
lastAction	*action*	name of the previous action

TABLE II

Dialog state variables in the MDP state

The MDP state model used in our experiments encodes information about the information state of semantic slots plus information about the progress of the dialog. The dialog manager uses the information slots with associated MDP state values as shown in table I. States representing the progress of the dialog are shown in table II.

Important for the learned strategy is the chosen reward function. It defines which dialogs are 'good'. In our scenario, learning correct names is rewarded (+10), learning wrong names is punished (-10). From experiments we found that some persons accept names which are almost correct. We try to quantify this effect with the Levenshtein distance between learned and correct name (distance = 1 is rewarded +3; distance = 2 is rewarded 0). Each additional turn is punished with -1, so dialog length is kept moderate; repeating the same system action is punished with -0.5. Other functions can be chosen to increase the importance of different factors.

IV. User Simulation

To build a user simulation a common approach is to model user actions with statistics estimated on collected data. In addition to that, (error-) models are created that describe the behavior of the system's recognition components, i.e. a statistical model of errors. The idea behind this approach is that statistically describing user actions and error models is simpler than directly learning the system's strategy.

A. Multimodal User Models

Existing approaches for training a user model range from simple models, such as the bi-gram model, to more complex models [21], [22], [23]. In previous work [24] we have achieved good results using a simple bi-gram model for statistics on a semantic level. In the addressed restricted task bi-gram statistics provided good estimations already on a small amount of training data, which would not suffice to train more complex models.

The quality of the bi-gram model $p = P(\mathsf{act}_{\mathsf{user}}|\mathsf{act}_{\mathsf{system}})$ highly depends on the defined abstraction granularity of simulated user and system actions and the task restriction. In our work we have adopted the general bi-gram model to a more fine-grained model of bi-grams over semantics of user actions (input speech act + semantic attributes) given the system's speech act. Statistics for user actions have been trained on a single dialog goal, i.e. name learning, in isolation of other dialog goals. To show the feasibility of this approach and collect necessary training data, a Wizard-of-Oz study and a standalone experiment were conducted as described in section II-D.

In addition to speech-only interaction our multimodal system models non-verbal information from voice ID and face ID. Voice ID, like speech input, is computed turn-wise for each spoken utterance. Inspired by recent work [16] which concatenates data from speech snippets to simulate data that is provided for voice ID during runtime, we adopt this approach and simulated recognition input by taking samples of real recorded data.

Face ID at first glance is not turn based. However, since face ID only updates the dialog state during a new turn, this is imitated in the simulation by grouping ten to twenty images for face recognition per turn. To produce a variety of hypothesis values, we use real images from one person taken during data collection at 2 fps, which are cut into sub-sequences of ten to twenty images. From these, the simulation environment randomly picks single sequences.

Problematic with this setting are the high computing requirements. Just considering face ID, given a database of roughly twenty persons, the face ID recognizer can process two to four images per second on a standard 3GHz Pentium processor. A minimum training requirement of 1 million dialogs then poses an impracticable computational burden. The biggest part in time consumption is to detect a face within an image and to produce a per-image ID classification using the nearest-neighbor classifier. Both problems can be pre-computed given a fixed database of known persons, when the state space of the classifier remains constant. The combination of pre-computed single-image hypotheses to a sequence hypothesis is much faster and can adequately be conducted during simulation. A similar approach using audio snippets has been applied to voice ID recognition. With these settings, the system runs a full dialog in simulation (including dialog state update, policy update and action selection) in 1.2 ms at 3.5 turns per dialog on average, which is roughly 0.3 ms per turn. Note that the chosen setup doesn't allow to directly simulate the effects of storing more and more persons in the database, but rather allows to train a strategy for a fixed database setting of known persons.

B. Error Models

Simulation of user actions is not sufficient to model the input for the dialog system. The missing link between user actions and dialog input is described by error models, which statistically simulate typical errors made by the recognition components. For example, the difference between experiments using close speech and distant speech is simply a different error model. Face ID as well as voice ID don't require additional error models since their errors are implicitly modeled by applying real classifiers (partly pre-computed) to simulated data. Speech, in contrast, is modeled statistically and requires additional error models for speech, phoneme, and spelling recognition. Spoken name input is modeled as a speech act with semantic parameters. For spoken name input, the speech act is *informName* with a semantic parameter 'NAME'. The error model first statistically models concept confusion and deletion, i.e. probability for recognizing a wrong concept (confusion) and the probability for not understanding any concept at all (deletion). Secondly, statistics are applied to model confusion and deletion of the semantic parameter(s).

V. Experiments and Results

A. Training and Evaluation in Simulation

Training of the strategy was conducted with the MDP state model and action model described in section III and in section II-D. Training was conducted with the Watkins-Q-lambda algorithm with exponential cooling of epsilon, and the learning rate alpha. All models were computed with all combinations of a list of 11 equally distributed lambda values: 0.0, 0.1, 0.2.. 1.0 and a list of 11 equally distributed discounting factors: 0.0, 0.1, 0.2 ... 1.0. To test the effect of the number of training runs we experimented with different dialog numbers per training, using 1 million to 100 million dialogs per model. Reasonable training runs are 10 million (10M) dialogs and more, since training with 1 million (1M) runs still contains a couple of state-action pairs that have never been visited, especially in states that occur only seldom. There is still a significant difference between training sizes of 10M and 100M dialogs, so high numbers of training dialogs still means improvement of the dialog strategy. On the other hand, first training runs with 100M dialogs took 32 hours on a Pentium4 3 GHz processor. After a few code optimizations we could lower the training time to roughly 5 hours. Considering training time, all models which have been trained with different configurations, i.e. 121 configurations for all combinations of discount and lambda values per MDP state space, have been trained with 10M dialogs, single configurations have been trained with 100M dialogs for comparison of the best models. All evaluation numbers presented here have been obtained from running 100k dialogs in simulation, which have shown stable results, from which the average reward is computed.

Training and evaluation requires to split data into three parts. The first part is used to train user ID models for voice ID and face ID, and bi-gram statistics. A second data set is used as simulation data, for training of the dialogue strategy, and a third set is used for evaluation of the dialogue strategy. Depending on how the set is split, the dialogue strategy training and evaluation sessions include more or less unknown persons. On a set with a large number of unknown persons, the resulting reward is lower than with a set restricted to known persons, since unknown persons are harder to recognize and to register.

Figure 1 shows evaluation results for the close speech condition of the baseline strategy in comparison to strategies trained with reinforcement learning (RL). The categories shown are 'sim' for the simulation set which was used for reinforcement learning, and 'eval' for the third held out data set. The model abbreviations are 'H' for the handcrafted model, 'F' for RL with face ID only, 'V' for RL with voice ID only and 'M' for RL with multimodal input (face ID + voice ID). Figure 2 shows the name-assignment errors made by the different strategies under close speech condition. An error is an incorrect assignment of a name at the end of a dialog. Almost correct names were counted separately, where the learned name differs by only one letter from the correct name. Both simulation sets included 25% unknown persons. For roughly six percent of all turns no voice ID information was available. On the remaining set, the recognition rates for voice ID are 59% and recognition rates for faceID are 68%.

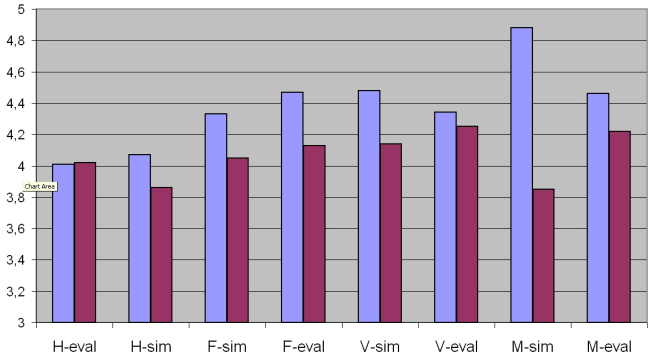

Fig. 1. Evaluation scores for different strategies in the user simulation, showing reward (first column) plus turn numbers (second column).

B. Experiments with Users and Discussion

The results of the simulation show better performance of the reinforcement strategy than the handcrafted strategy. An interpretation is that the reinforcement learning approach learns more complex rules, when to confirm multimodal input, in combination with recognition confidence, dialog length, and failed name recognition. The charts show slight differences between the sets. The 'M' model (multimodal input) performs generally best, which matches our expectations, because it can choose among different modalities. All learned models have a higher number of correct dialogs, at a minimal cost of 0.1 turns more per dialog on average.

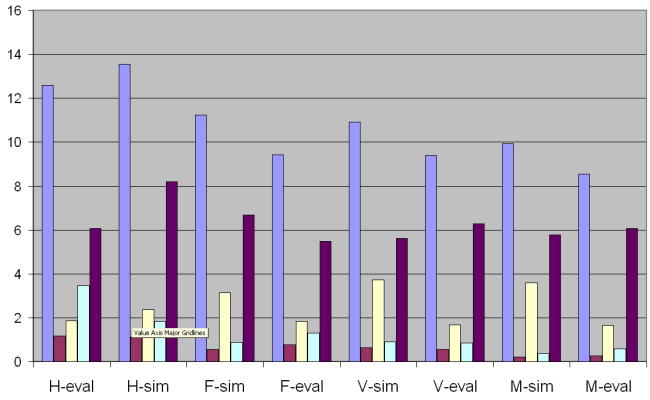

Fig. 2. Incorrect assignments made by the different strategies in the simulation in % (out of 100k dialogs). The bars indicate: total errors, almost correct names for known persons, almost correct names for unknown persons, incorrect names for known persons, incorrect names for unknown persons.

To evaluate the system with real users we have deployed the strategy to a small user study experiment. Since there are high variations in results depending on the user, the experiment can only be interpreted as a tendency but not be treated as a reliable comparison of strategies. Further and larger evaluations are necessary for this.

The experimental setup was the same as for the Wizard-of-Oz experiment with the same microphones and cameras. This time the users were told that the only purpose of the interaction was the name learning dialog. We conducted dialogs with 11 persons, 7 persons using distant-speech, some of them haven't used speech recognition before, and 4 persons using close-speech. The experiments were conducted with two conditions: known and unknown. A person at first was unknown, and after completing the learning dialog, name and ID information was stored in the database to conduct further dialogs with the known condition. To obtain more dialogs with the unknown condition the ID and name information was removed from the database, to iteratively conduct dialogs in the unknown and known conditions.

The following shows two sample dialogs. The first dialog shows a sample dialog trained with reinforcement learning:

User: Hello ***[act_hello] ; FaceID: Peter***
Sys : Am I right that your name is Peter?
User: No, my name is John. ***[inform_name [NAME "OOV"]]***
Sys : Please spell your name!
User: J O H N. ***[inform_spelling [SPELLING "JOHN"]]***
Sys : Have I understood correctly that your name is John?
User: Yes. ***[yes]***

The next dialog shows a simple confirmation dialog with only one user turn to confirm the correct recognition.

User: Hello ***[act_hello] ; FaceID: John***
Sys : Am I right that your name is John?
User: Yes that's right. ***[yes]***

The close-speech experiment produced similar results as evaluation in simulation. Average dialog length was 4.7 in the unknown condition and 2.6 in the known condition, excluding greeting and goodbye. The distant speech experiment produced worse results than predicted by the simulation.

While first interactions with speech experts went very well, naive users had more problems to complete the dialogs. For example, users spelled their names too slowly, which was not handled correctly by automatic speech segmentation. After an introduction users could complete the task more easily. Additional errors were caused by spelling recognition performance which mostly was not 100% correct. All numbers from the experiment are shown in table III. The problems are rather to be assigned to system conditions than to the dialog strategy. The numbers also show that the task to register an unknown person is much harder than identifying a known person. Unknown persons can neither be recognized by face ID or voice ID, or, if they could be recognized, but during previous interactions no name was stored, this cannot be communicated by the system. So currently the only way to get known by the system was by spelling one's name, which was easier to complete when used to the system.

	1	2	3	3	4	5	6	7
unknown	5,3,4	3,7,5	15,15,15	14,4,5	15,15,12	5,4	15,11	15,15,6
known	3,8,1	1,4,2	1,3		1,6		1,3	3

TABLE III

NUMBER OF TURNS DURING DISTANT SPEECH DIALOGS. THE COLUMNS MARK SUBJECTS 1 TO 7. 15 TURNS MARKS UNSUCCESSFUL DIALOGS.

VI. CONCLUSIONS AND OUTLOOK

Reinforcement learning in multimodal user simulation produces results comparable to a handcrafted strategy. It furthermore has the advantage that it can be obtained automatically and be retrained for new environments. ID hypotheses from different recognition components are integrated in dialog, and depending on the trained conditions (error models, distant speech vs. close speech) selects which hypothesis to trust and thus implicitly implements a confirmation strategy over multiple modalities. Confidence measures evaluated for face ID provide additional improvements. The system combines identification and learning tasks within one dialog. Directions for future work could consider different reward functions for identifying known persons and learning unknown persons, or training of the dialog strategy on top of multimodal fusion of ID hypotheses.

In a distant speech experiment, a small user study has shown that some kind of learning barrier exists for some users. Bad acoustic conditions during the studies with a speaker distance of one to two meters harm interaction quality measurably and increase the requirements to spell in a certain way that can be understood by the system.

Comparison of the different strategies has been conducted in simulation, where a large number of dialogs can be evaluated with relatively low efforts. To reduce over-fitting on the training set, held out data was used for training of a simulation for evaluation. The presented experiments with real users show that the strategy produces realistic results. For comparison of different strategy types additional experiments are planned, which can also be used to analyze if user simulations can be realistic replacements for real user experiments. Recognizing an arbitrary name is a challenge for speech recognition, since a list of all possible names is too large too keep all names in the vocabulary at the same time. In the current setup the name vocabulary is restricted to known persons, unknown names are detected as OOV and can be spelled for learning. In the future we plan to integrate larger name vocabularies and weighting of common names to enable understanding names of unknown persons as well.

VII. ACKNOWLEDGMENTS

This work was supported in part by the German Research Foundation (DFG) as part of the Collaborative Research Center 588 "Humanoid Robots - Learning and Cooperating Multimodal Robots".

REFERENCES

[1] T. Schaaf, "Detection of oov words using generalized word models and a semantic class language model," in *Proc. Eurospeech*, 2001.
[2] L. Hetherington, "A characterization of the problem of new, out-of-vocabulary words in continuous speech recognition and understanding," Ph.D. dissertation, MIT, 1995.
[3] T. Slobada and A. Waibel, "Dictionary learning for spontaneous speech recognition," in *Proc. ICSLP*, 1996.
[4] A. Park and J. Glass, "Unsupervised word acquisition from speech using pattern discovery," in *Proc. ICASSP*, 2006.
[5] G. Chung, S. Seneff, C. Wang, and I. Hetherington, "A dynamic vocabulary spoken dialogue interface," in *Proc. ICSLP*, 2004.
[6] E. Kaiser, "Using redundant speech and handwriting for learning new vocabulary and understanding abbreviations," in *Proc. ICMI*, 2006.
[7] O. Scharenborg and S. Seneff, "Two-pass strategy for handling oovs in a large vocabulary recognition task," in *Proc. Interspeech*, 2005.
[8] G. Chung and S. Seneff, "Integrating speech with keypad input for automatic entry of spelling and pronunciation of new words," in *Proc. ICSLP*, 2002.
[9] U. Meier and H. Hild, "Recognition of spoken and spelled proper names," in *Proc. Eurospeech*, 1997.
[10] S. Dusan and J. Flanagan, "Adaptive dialog based upon multimodal language acquisition," in *Proc. ICMI*, 2002.
[11] J. Carbonell, "Towards a self-extending parser," *Annual Meeting of the ACL*, 1979.
[12] S. Young, "Learning new words from spontaneous speech: A project summary," in *CMU Tech Report, CMU-CS-93-223*, July, 1993.
[13] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Science*, pp. 71–86, 1991.
[14] G. Aggarwal, A. Chowdhury, and R. Chellappa, "A system identification approach for video-based face recognition," in *Proc. ICPR*, 2004.
[15] H. K. Ekenel and Q. Jin, "Isl person identification systems in the clear evaluations," in *Proc. CLEAR Evaluation Workshop*, 2006.
[16] F. Krsmanovic, C. Spencer, D. Jurafsky, and A. Ng, "Have we met? MDP based speaker id for robot dialogue," in *Proc. Interspeech*, 2006.
[17] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A one pass-decoder based on polymorphic linguistic context assignment," in *Proc. ASRU*, 2001.
[18] S. Könn, H. Holzapfel, H. Ekenel, and A. Waibel, "Integrating face-ID into an interactive person-ID learning system," in *Proc. ICVS*, 2007.
[19] H. Holzapfel and A. Waibel, "Behavior models for learning and receptionist dialogs," in *Proc. Interspeech*, 2007.
[20] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.
[21] W. Eckert, E. Levin, and R. Pieraccini, "User modeling for spoken dialogue system evaluation," in *Proc. IEEE ASR Workshop*, 1997.
[22] E. Levin, R. Pieraccini, and W. Eckert, "A stochastic model of human-machine interaction for learning dialog strategies," *IEEE Transaction On Speech and Audio Processing*, vol. 8, no. 1, January 2000.
[23] O. Pietquin and S. Renals, "ASR system modeling for automatic evaluation and optimization of dialogue systems," in *Proc. ICASSP*, 2002.
[24] T. Prommer, H. Holzapfel, and A. Waibel, "Rapid simulation-driven reinforcement learning of multimodal dialog strategies in human-robot interaction," in *Proc. Interspeech*, 2006.