Chapter 8
Cognitive Dialog Systems for Dynamic Environments: Progress and Challenges

Felix Putze and Tanja Schultz

Abstract In this chapter, we present our existing setup and ongoing research on the development of cognitive dialog systems for dynamic environments like cars, including the main components that we consider necessary to build dialog systems to estimate the user's mental processes (hence, cognitive) and adapt their behavior accordingly. In conducting realistic testing and recording environment to produce real-life data, a realistic driving simulator was used. We also needed to observe the user during these interactions in a multimodal way to estimate the current user state based on this data. This information is integrated with cognitive modeling components that enrich the observational data. We finally needed a dialog management system which is able to use this information for adapting its interaction behavior accordingly. In this chapter, we report our progress in building these components, give an overview over the challenges we identified during this work and the solutions we aim for.

Keywords Cognitive dialog system • Cognitive model • Human machine interaction • User state detection

8.1 Introduction

Spoken dialog systems have matured to a point where they find their way to many real-world applications. However, their application in very dynamic scenarios remains an open and very interesting task. Spoken dialog systems as an interface for in-car services are very desirable and at the same time very challenging. On one hand, they offer eyes-free and hands-free control without visual or manual distraction from the primary driving task. On the other hand, this task uses the user's cognitive capacity, so

F. Putze (✉) • T. Schultz
Cognitive Systems Lab, University of Karlsruhe, Karlsruhe, Germany
e-mail: felix.putze@kit.edu; tanja@ira.uka.de

J.H.L. Hansen et al. (eds.), *Digital Signal Processing for In-Vehicle Systems and Safety*,
DOI 10.1007/978-1-4419-9607-7_8,

we can no longer assume to deal with a fully attentive and perfect interaction partner as in more static environments. Another important aspect is the adaptation to individual preferences. As dialog sessions in driving scenarios may last for several hours, we have to take into account both changing user states, i.e., cognitive workload or emotions, as well as lasting user traits, e.g., gender or personality. Both types of individual differences influence the optimal interaction behavior which the system should use for maximizing user satisfaction, as user studies like [1] show. There is potential for a large range of adaptation measures: One example is reacting to increased cognitive workload by taking the initiative from the user, delaying noncritical information, or reducing its complexity. Another one is adjusting the system to the user's emotional state and personality by selecting appropriate wording, voice, and turn-taking behavior. We propose to use systematic multimodal observation and state classification of the user derived from a variety of different biosignals. This metadata is augmented with a more detailed model-based representation of the user's mental processes and helps to select appropriate adaptation measures. Combining a global model of the user's cognition and affective states for the purpose of building adaptive interaction strategies is new to the field of spoken in-car dialog systems.

After a review of related work, the following sections describe all components which are necessary to develop and evaluate cognitive interaction systems for in-car applications: a driving simulator to create a realistic environment for recordings, an interaction system as a platform for human–machine interaction, a recording setup to collect data for training and testing of systems, a recording software to deal with the challenges of multiple input streams, and a user state detection framework and components to model human cognition.

8.2 Related Work

In the last years, many approaches for user models for application in adaptive in-car dialog systems exist. Like [2], most of them rely on heuristics and indirect user state detection.

The authors of [3] describe a dialog system that bases its handcrafted dialog strategy for a gaming interface on the user's emotional state, derived from prosody, language, and visual features. Together with the history of interaction, the current user command, and other discourse features, the user state can be accessed by the dialog strategy in the form of a decision tree.

Fatma Nasoz and Christin Lisetti [4] describe a user-modeling approach for an intelligent driving assistant. This model is based on a Bayesian network which allows to derive the most useful system action (in terms of driving safety) given the estimated driver state, which consists of emotional state, personality, and other features and is partially derived from physiological measurements like the user's heart rate. The score for each action is calculated using a utility node which measures the probability of safety improvement given the current user state. Similar decision-theoretic, user-model-based action evaluation approaches are used in [5], which

also include an active sensor selection mechanism. Cristina Conati [6] presents an educational dialog system that can decide for different user assistance options, given the user's emotional state (derived from different modalities). This work bases its network on the cognitive OCC (by Ortony, Clore, and Collins) appraisal theory, which relates the users' emotions with their goals and expectations.

In the area of user state detection from biosignals, Liang, Reyes, and Lee [7] developed a real-time workload classifier in the car using facial features, like pupil diameter or gaze direction, extracted from videos of the driver. The ten participants followed a car with varying speed while performing a secondary memory and comparison task. Using support vector machines, the authors achieved a recognition rate of 81.1% on average for the recognition of cognitive workload. Healey and Picard [8] developed a classifier to monitor the stress levels in daily life car-driving tasks. They collected data from 24 real-life drives of at least 50-min duration and used the biosignals electromyography, electrocardiography, and skin conductance for their system. Linear discriminant analysis (LDA) was used for dimensionality reduction, and a classifier using a linear decision function was able to discriminate the three classes with accuracies of 100% (low workload), 94.7% (medium workload), and 97.4% (high workload).

8.3 Driving Simulator

Testing and evaluation of different interaction strategies requires a realistic experimental environment which reproduces all important effects and distractions seen in real-life applications. While recording in a real car in real traffic situations creates the most authentic sessions, the downsides of this approach are safety concerns with early prototypes, the lack of reproducibility, and the missing ability of reliably provoking scenarios which are relevant for the current investigation. Therefore, we decided to build a driving simulator which is designed to create a realistic driving experience. The main focus was not to build a physically correct car test bed but to simulate the most important influences and distractions that occur during real driving tasks, especially in situations where the application of a dialog system plays an important role. We based our driving simulator on a real car and kept the interior fully intact and functional to provide a realistic in-car feeling. The car is surrounded by a projection wall, covering the view of the frontal and lateral windows. The simulator features acoustic feedback via engine sound and environmental surround sound and haptic feedback in the seat (via tactile transducers) and steering wheel (via force feedback).

The simulator software is based on a modified gaming engine[1]. It was extended using a multiscreen display, steering wheel support, and simple ambient traffic control.

[1] MTA:SA: http://www.mtasa.com

Fig. 8.1 The CSL driving simulator in action

Its support for scripting scenarios in LUA allows us to configure individual driving stages: We can position the driver in a wide artificial environment with realistic urban and rural areas, where we define a route represented by navigation directions for the system. It is possible to trigger events at defined points to generate specific traffic situations, position new elements in the environment, or influence the position or driving characteristics of the car (Fig. 8.1).

8.4 Interaction Setup

While the user is driving, they interact (via close-talking microphone to reduce noise) with a dialog system. In our current scenario, this constitutes a virtual co-driver which acts as interactive tour guide and navigation system for the virtual environment. To investigate the phenomena we are interested in, e.g., different levels of workload, we created several scenarios specially designed for studying man–machine interaction. This includes the handling of a variety of secondary tasks, urban and rural routes, and several triggered events.

The virtual co-driver is present on a screen in the cockpit on which it is displayed using the ThinkingHead[2], a morphable 3D avatar, and is equipped with a grammar-based speech recognition system and a speech synthesis component to vocally communicate with the driver. The co-driver is driven by a lightweight interaction manager which was designed especially for the purpose of adaptive dialog systems. The interaction manager uses a rule-based engine which executes one or more rules with preconditions that match the current interaction state, according to the

[2] http://thinkinghead.edu.au

Information State Update paradigm [9]. The interaction state also comprises variables that describe the detected user state to allow adaptive selection of speech acts based on the user's current situation.

The system is also able to switch its behavior between different styles for the realization of one selected speech act, depending on the user's state. Different behavior styles can change the processing of speech acts in many aspects. For example, the content of a speech act realization can differ in its length and complexity based on the user's workload. It is also possible to adjust the speaking speed, the volume of the voice, and the stress of certain key phrases according to this parameter. Using those parameters, the co-driver realizes a verbose, chatty, and entertaining behavior if it detects a state of low cognitive workload. It presents much information, tells occasional jokes, and shows expressive mimic. For situations with high cognitive workload, the co-driver switches to a different, more concise, and unobtrusive behavior to use the limited available cognitive resources for the transmission of the most critical information. In this style, the system also takes more initiative in the interaction, taking most noncritical decisions from the user.

A user study [10] showed that a behavior which adapts to the changing user's cognitive load is both more efficient and also more satisfying for the user than a nonadaptive one. By changing the information throughput depending on the workload level, the system can optimally use the available cognitive resources of the user without risking overload. This behavior was evaluated as empathic and desirable by the users in a satisfaction questionnaire. It is therefore critical for a cognitive interaction system to provide this kind of adaptation.

8.5 Recording Setup

During the interaction, we employ a variety of signals to observe the user in the car. This is done for multiple reasons. First, an adaptive dialog system needs online data streams from which it can extract meaningful features describing the user's state. Second, to train automatic recognizers that perform this user state classification, we need to provide large amounts of labeled training data. To that end, we installed multiple biosignal sensors in the car to get a reliable, continuous data stream without obstructing or distracting the user too much.

We employ the following equipment to observe the user:

- Small cameras to record videos of the face and the upper body of the driver to catch facial expressions and body pose.
- A close-talking microphone to record the user's utterances
- Brain activity is measured using electroencephalography (EEG) with one of two possible alternatives:
  - A 16-electrode EEG cap with active electrodes for optimal signal quality and coverage of all brain regions
  - A 14-electrode gaming device (Epoc Emotiv) with saline electrodes for increased usability and reduced setup time

- A light sensor glove which measures skin conductance and heart rate
- A respiration belt on top of the clothes to measure respiration frequency
- Two facial electromyography (EMG) electrodes to record facial activity which is not captured by the cameras.

The last three items all use the same recording interface and are either attached to a universal signal recorder[3] or directly connected via Bluetooth, which reduces obstruction to a minimum. In addition, we employ indirect motion monitoring by continuously recording the angle of the steering wheel and the acceleration and brake pedals in the car.

In this recording setup, we already collected more than 100 interaction sessions in the tour guide scenario, interacting with a virtual co-driver controlled by a human wizard. Each interaction session comes with a collection of recorded biosignals, a manual transcription, and the results of several questionnaires on user personality, satisfaction, and task performance. This large collection allows a systematic investigation of interaction behavior under changing workload conditions.

8.6 Recording Software

Metadata extraction for dynamic dialog systems is required to work in real time. To this end, we need to record multiple biosignal streams in a robust, fast, and convenient way, offering interfaces to read data from very different signal sources and output it to very different receivers like recognizers or visualization components. To fulfill all requirements, we developed a new recording software called *BiosignalsStudio* [11]. BiosignalsStudio is designed in a modular fashion and allows to connect arbitrary input modules for data collection from a specific device with arbitrary output modules which write data to files, visualize the data, or send it to an external recognizer software via sockets. All modules share a common generic data format which stores multiple data channels and a meta-information block which contains the sampling frequency, detected errors, etc. Each module can be connected to several receivers, allowing data from one source to be stored to disk and visualized in parallel. There exists a number of intermediate modules which can be installed between input and output modules to augment, filter, or transform the data. Currently, input modules for all connected biosignal recording devices and several others (like gyro and acceleration sensors) are available (Fig. 8.2).

As we operate with very different and asynchronous data streams, it is important to store timestamps with each data block to ensure that only data which belongs together is merged in the multimodal fusion of the recognition engine. These timestamps are

[3] Varioport, Becker MediTec

Fig. 8.2 (Part of) the recording setup with EEG cap, audio headset, and sensor glove in the driving simulator

generated at the earliest point possible which is usually when receiving the data block from the hardware interface (some devices are able to generate hardware timestamps, which is preferable). Timestamps within blocks of data are linearly interpolated. They are stored together with one data file for each modality and detailed log files in one directory per session, allowing easy and standardized access for all components, regardless of the specific recording setup. For distributed recording on multiple machines, timestamps are automatically synchronized via the NTP protocol. In this situation, the software is also able to remotely control the recording from one machine which starts and monitors the recording on the others.

8.7 User State Detection

The collected biosignal streams are passed on to a generic biosignal classification framework that performs the following steps. First, the data is filtered and cleaned to remove technical and physiological artifacts from the signals. For this purpose, we employ several source separation techniques, e.g., independent component analysis (ICA) to remove eyeblinks from the EEG signal or canonical correlation regression (CCR) to deal with EMG artifacts. From the cleaned signals, we then calculate features to describe them. Features are extracted on overlapping windows of varying length, depending on the signal type and on characteristics of the user state in question. For the biosignals, we extract features both from the time and the frequency domain. Typical time domain features are mean, variance, or zero-crossing rate, calculated on the raw feature or on the first or second derivative. Frequency domain features are especially relevant for EEG signals. Classical features here describe the band power in the α-,β-,γ-,δ-, and θ-bands

(see [10, 13] for details), but other features, e.g., derived using the Wavelet transform, are also available.

For the speech signal recorded during the interaction, we use the software Praat[4] to extract prosodic features like pitch, jitter, or shimmer from the user's voice. To capture linguistic features, we use the Linguistic Inquiry and Word Count[5] that categorizes each word in its vocabulary in one or more groups, e.g., "negative emotional word" or "self reference." Active Appearance Models [12] are used to capture information on the facial expression and activity of the user as recorded by the camera in the car.

To arrive at a person-independent system, features are normalized using range normalization or z-normalization. The normalization statistics are calculated on additional holdout data which is not used for other steps of training and evaluation. This kind of data can also be collected in an unsupervised fashion as enrolment data to bootstrap the system for a new user.

As we generate a very large initial feature set, we employ Forward Feature Selection during the training step to reduce the dimensionality of the feature space, preceded by a correlation-based filtering to reduce the runtime of the selection.

For classification, the final feature vectors are then passed into a statistical classifier of which multiple variants are available, e.g., a support vector machine (SVM) using Radial Basis Function kernels or a classifier based on linear discriminant analysis (LDA). More exactly, there is one classifier for each modality as this allows dynamic weighting of input channels, e.g., to account for noise or defective sensors. To arrive at a final classification result, the output of all classifiers is combined using majority voting.

One important application for user state detection is the recognition of cognitive workload from multiple biosignals. In a large evaluation, we developed a user-independent classification system that discriminates between low and high workload. For each participant, we record a number of different sessions in a driving scenario. Relaxing phases or simple driving tasks are labeled as low workload while driving sections with different secondary tasks (visual and auditive cognitive tests) are labeled as high workload sessions. From a prestudy [13], we know, from evaluation of subjective workload using the NASA TLX questionnaire, that this assignment corresponds to experienced load levels. Figure 8.3 summarizes the recognition rates achieved in the evaluation using a cross-validation scheme to classify data from relaxing phases and high workload phases induced by driving with secondary task. We see that a person-independent discrimination of the two conditions is possible, and that a decision fusion approach yields the best results.

[4] http://www.praat.org

[5] http://www.liwc.net

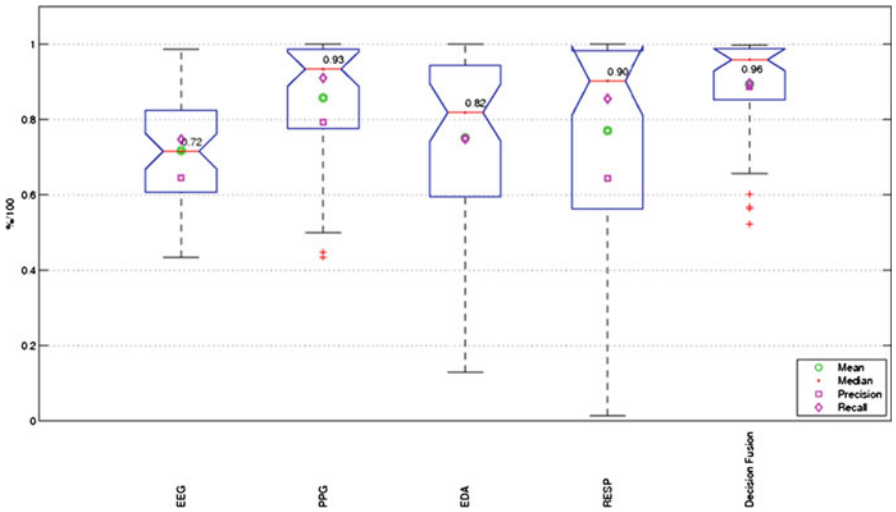

Fig. 8.3 Recognition rates of a multimodal biosignal classifier for discriminating two conditions of low and high workload in a driving scenario. We show recognition rates for EEG, photoplethysmography (PPG), skin conductance (EDA), respiration (RESP), and decision fusion

8.8 Cognitive Modeling

Cognitive architectures like ACT-R [14] aim to provide a general model of human cognition for simulation or prediction. For the use in adaptive dialog systems, they help to represent and estimate nonobservable user states and are also able to predict future user behavior from a given state. This is very useful for two purposes. On one hand, a cognitive model can support the empirical, biosignal-based classification of user states by proving information derived from the evaluation of more formal models of cognition which are backed with a priori knowledge from psychology and cognitive science. On the other hand, a cognitive model is able to simulate human behavior in situations where no real user is available; this is a typical use case in evaluation and training situations in early phases of the development of a new system.

As a first cognitive modeling component, we implement a memory and interest model to represent the user's activation of, and interest in, the actual and potential discourse items. Our focus here is to reflect the fact that the user cannot remember all discourse items correctly and with the same intensity. This is of special importance in situations where the dialog system interrupts an ongoing dialog for more important information or during time-critical situations.

The memory model represents for each time slice an activation value for each possible discourse item in the domain ontology, including relations between those items. The activation determines how present each item currently is in the user's memory and how it can be used to derive the chance of successful retrieval of this

item and the time necessary to perform such a retrieval process. We based our system on the connectionist approach presented in the LTM^c model [15], which was proposed to solve some issues with the memory model of ACT-R. Here, each item is represented as a node, connected with edges to other items that are semantically, linguistically, or hierarchically related. These edges are used to spread activation between nodes when one becomes activated, e.g., through a system speech act. We also extended the LTM^c model to better reflect the dynamics of a memory system which is important to model topic switches in an interaction.

The interest model reflects the user's current interest in each item. This is a dynamic variable that depends not only on the situational context (spatial proximity, expressed interest) but also on more general, static factors. To represent this variety of influences, we employ a Bayesian network for the interest model.

Both models are currently used to determine a general value of importance of giving additional information to the user. This value allows us to weigh the speech act of information presentation against other goals like navigation pointers or entertainment. We do this by summing up the negative activation for all items, weighed by the interest values of each item. This score is called *competence urge*, based on the more general concept of urges that describe the needs of an individual and that influence its emotions and actions [16]. This score is also used to determine the items the system will present next to the user as they maximize the reduction of the competence urge.

In a user study in the tour guide scenario [17], we showed that it is possible to simulate plausible interactions using cognitive models. Utterances of the user where generated using the memory model which was stimulated from the perception of external stimuli and queried for the most highly activated items. The system in those simulations generated its utterances using its own model of the user's memory with a similar structure than the generative model, but separate activation scores to track what is going on in the user's mind. The behavior of both the system and the simulated user was learned in a Reinforcement Learning-based manner, using the urge mechanism to weight the goals of the agents. The generated interactions were played back to human judges and perceived as similar to a handcrafted gold standard and as significantly better than the baseline behavior.

Future applications for the memory model comprise its influence on the user understanding model, by making the chance for misunderstandings dependent on the activation level of the relevant items and the application for coherent user simulation in evaluation and training of interaction strategies.

8.9 Conclusions

The development of flexible, generic, and natural adaptation mechanisms for cognitive interaction systems has seen great progress as reported in this chapter. We implemented and tested a realistic driving simulator which will allow a large number of experiments under controlled but nevertheless authentic conditions.

We presented an adaptive dialog system that can change its behavior depending on the state of its user. We have implemented a framework of biosignal recording components and statistical classifiers that are able to determine the user's current state, for example his cognitive workload. We investigate cognitive modeling architectures to structure the user's adversarial desires and to model the user's memory. The next step will bring all components together to create a system which uses both biosignal-based user state detection and predictive models to a dialog strategy which can adapt flexibly to changes in the user's state.

References

1. Nass C et al (2005) Improving automotive safety by pairing driver emotion and car voice emotion. In: Proceedings of CHI, Oregon
2. Hassel L, Hagen E (2006) Adaptation of an automotive dialogue system to users expertise and evaluation of the system. Lang Res Eval 40(1):67–85
3. Gnjatović M, Rösner D (2008) Emotion adaptive dialogue management in human-machine interaction: adaptive dialogue management in the NIMITEK prototype system. In: 19th European meetings on cybernetics and systems research, University of Vienna, Vienna
4. Nasoz F, Lisetti C (2007) Affective user modeling for adaptive intelligent user interfaces. In: Proceedings of 12th HCI international conference, Beijing
5. Li X, Ji Q (2005) Active affective state detection and user assistance with dynamic Bayesian networks. IEEE Trans Syst Man Cybern 35:93
6. Conati C (2002) Probabilistic assessment of user's emotions during the interaction with educational games. Appl Artif Intell 16:555–575
7. Liang Y et al (2007) Real-time detection of driver cognitive distraction using support vector machines. IEEE Trans Intell Transp Syst 8(2):340–350
8. Healey J, Picard R (2005) Detecting stress during real-world driving tasks using physiological sensors. IEEE Trans Intell Transp Syst 6(2):156–166
9. Larsson S, Traum DR (2002) Information state and dialogue management in the TRINDI dialogue move engine toolkit. Nat Lang Eng 6:3–4
10. Heger D, Putze F, Schultz T (2010) An adaptive information system for an empathic robot using EEG data. In: 2nd international conference on social robotics, Singapore
11. Heger D, Putze F, Amma C, Wielatt T, Plotkin I, Wand M, Schultz T (2010) Biosignals studio: a flexible framework for biosignal capturing and processing. In: 33 rd annual German conference on artificial intelligence 2010, Karlsruhe
12. Gao H (2009) Robust face alignment for face retrieval. University of Karlsruhe (TH), Karlsruhe
13. Putze F, Jarvis J-P, Schultz T (2010) Multimodal recognition of cognitive workload for multitasking in the car. In: 20th international conference on pattern recognition, Istanbul
14. Anderson J et al (2004) An integrated theory of the mind. Psychol Rev 111:1036–1060
15. Schultheis H et al (2006) LTM-C – an improved long-term emory for cognitive architectures In: Proceedings of the 7th international conference on cognitive modeling, Trieste
16. Bach J (2003) The micropsi agent architecture. In: Proceedings of international conference on cognitive modeling, Bamberg
17. Putze F, Schultz T (2009) Cognitive memory modeling for interactive systems in dynamic environments. In: Proceedings of 1st international workshop on spoken dialog systems, Kloster Irsee, 2009
18. Waibel A et al (2001) A one pass-decoder based on polymorphic linguistic context assignment. In: Proceedings of ASRU, Trento