# Advances in the development of a cognitive user interface

*Oliver* Jokisch [1, 2, *] and *Markus* Huber [2]

[1] Leipzig University of Telecommunications (HfTL), Institute of Communications Engineering, 04277 Leipzig, Germany
[2] InnoTec21 GmbH, 04105 Leipzig, Germany

**Abstract.** In this contribution, we want to summarize recent development steps of the embedded cognitive user interface UCUI, which enables a user-adaptive scenario in human-machine or even human-robot interactions by considering sophisticated cognitive and semantic modelling. The interface prototype is developed by different German institutes and companies with their steering teams at Fraunhofer IKTS and Brandenburg University of Technology. The interface prototype is able to communicate with users via speech and gesture recognition, speech synthesis and a touch display. The device includes an autarkic semantic processing and beyond a cognitive behavior control, which supports an intuitive interaction to control different kinds of electronic devices, e. g. in a smart home environment or in interactive respectively collaborative robotics. Contrary to available speech assistance systems such as Amazon Echo or Google Home, the introduced cognitive user interface UCUI ensures the user privacy by processing all necessary information without any network access of the interface device.

## 1 Research background

The joint research project Universal Cognitive User Interface (UCUI) 2015–2018 is developing methods, data and a prototype [1] to easily manage connected home appliances (as example scenario) by corresponding intuitive actions of the users. In the framework of our contribution, some preliminary results from the UCUI project shall demonstrate the potential for a novel class of interfaces for the human-machine or human-robot interaction.

With UCUI, the user can control the system via speech, gestures or even a virtual keyboard. The system is designed to operate autonomously, neither using an extensive database nor a network connection. Recent speech dialog systems and cognitive user interfaces allow a verbal, natural human-machine-interaction and achieve an excellent performance. However, the leading commercial solutions heavily rely on transmitting sensitive user information such as personal data or voice recordings through public networks and on processing, storing and analyzing the data on servers of external service providers.

The UCUI demonstrator is realizing a cognitive user interface for intuitive interaction with arbitrary electronic devices and ensures privacy by design. The system collects user-specific data which are processed by a cognitive behavior control to allow an adaptation to the users' communication style and to improve the strategy in problem solving. The underlying paradigm requests a systems' adaptation to the user and not vice versa, assuming the fact, that such systems are mainly used by human beings who are less trained in the use of complex technical devices. In addition, user-specific data are not delivered to other users to avoid possible conclusions from these data. In order to achieve an appropriate system behavior, a variety of possible human-machine interactions needs to be integrated into the UCUI system, since alternative input phrases may have an identical meaning in speech control. Therefore, all input and output modalities are fused on a semantic processing level.

For the data preparation we conducted Wizard-of-Oz (WoZ) experiments to collect typical user inputs [2]. In further steps, the user behavior shall be analyzed and integrated into the system model.

The analysis and classification software in the system is based on the Unified Approach to Signal Synthesis and Recognition [3, 4], hosted by the BTU Cottbus-Senftenberg and Fraunhofer IKTS including a speech recognizer and synthesis engine, both ported to the hardware.

The project partners are mainly focusing on the cognitive processing of meanings and knowledge about the user habits. For the representation of semantic data feature-values-relations [5-8] are used, processed by Petri net transducers (PNT) [9, 10]. Feature-values-relations are treelike non-sequential structures, where a feature has a set of values which themselves can be features again. Petri net transducers are used to translate input signals into such structures and also for translating them into output signals.

The system shall be capable of learning from the behavior of users in order to improve its function. Multiple devices will be able to cooperate (distributed microphone array, task assignment, etc.) over a strongly encrypted wireless connection. The system design is based on studies of user-machine interactions in a real home-automation scenario and takes into account relevant legal and ethical aspects.

For the demonstrator, the project partners reduced the task to the domain of controlling a heating installation.

[*] Corresponding author: jokisch@hft-leipzig.de

The semantic processing transforms all inputs (speech, gestures and touch screen) into a unified representation. By the cognitive behavior control, the representation can be transformed into any output channel (speech, acoustic signals, display). Figure 1 shows the first version of the UCUI demonstrator.

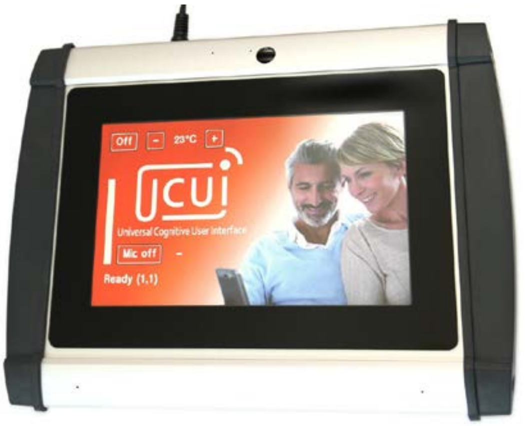

**Fig. 1.** Demonstrator of UCUI, first version 2017 [1].

## 2 User-centered experience

### 2.1 Data retrieval by Wizard-of-Oz experiments

The described cognitive interface is developed user-driven, which poses a challenge, as the overall system is still under construction. The project partners need to evaluate and to optimize some system functions before their implementation. For this purpose, the Wizard-of-Oz (WoZ) method is used in the UCUI project [2]. The main component in Wizard-of-Oz experiments is a human being (the wizard), who simulates the final systems' behavior. During the experiment, the test user interacts with the interface of a simulated technical system. All system reactions to the user are pretended by the wizard.

Wizards have to react accurately and in short time on user inputs. This can be supported with predefined, frequent responses in rapid access, e.g. "please wait, your inquiry is processed" or similar statements, and by a suitable training of wizards to achieve constantly accurate responses. User scenarios and tasks require a known goal of the actions which can be only achieved by the means of the system without restricting the user in his solution strategy, verbal utterances or gestures. The task construction has to consider the interaction variety of the user and should communicate the options to the user.

Within the UCUI tests, the user is receiving written instructions beforehand, and the interface system is demonstrated on the basis of a simple vendor machine application by an investigator (not identical with the wizard). Furthermore, the task assignment is based on hypotheses with regard to the expected user and system behavior.

Finally, a successful, user-driven system construction includes a series of WoZ experiments, whereby the tested system states should increasingly interact in autonomous mode with the user, i.e. less-controlled by the wizard. Consequently, the UCUI project involves three consecutive test runs, followed by the overall evaluation of the optimized system.

A Wizard-of-Oz Framework (WoOF) was built to support the user-driven construction. It allows for the creation of different evolving simulators and serves as an execution environment for these simulators.

The formulated requirements include general ones on frameworks supporting WoZ experiments as well as special ones following from the project specifications. Since the task is to simulate a real system which gives visual and audible feedback to the user, there has to be some mechanism to present visual objects on a monitor and to route audio data to the user. Eventually the system should be controllable via speech and touch-input (among other inputs) which imposes the necessity to interact with the visual objects and to route audio data from the user to the framework. Besides these basic functionalities, an adequate support for the realization of the experiments has to be included. This covers creation of simulators and experiments as well as supporting the wizard during the experiments.

For the UCUI project the first experiments [2] consisted of a series of scenarios which are seen as tasks to the participants. The wizard was able to switch between scenarios. A single scenario is understood as a unit of a user task, the aim of the task, and possible visual and audible feedback.

The outcome of the experiments included a collection of user behaviors. Therefore all interactions with the system were recorded which included the monitor content the participants saw, all touch-events they triggered on it and all spoken input during the experiments. To support the integration of the gesture control in following project phases, the participants were additionally recorded by camera. To ease the evaluation of the recorded data the audio output of the system was recorded as well.

A form of session management to allow for data per participant was also included. To respect the privacy of the participants was explicitly not a requirement on the framework. This had to be assured by the experimenters. The motivation behind was that there cannot be any algorithmic solution appropriate to all applications of the framework. So this burden was left to the users.

A session as the execution of an experiment with a distinct participant included some well-defined events. Also it was possible to add new types of events to the framework.

The preparation of the collected data was semi-automated. All data of a session were cut into pieces corresponding to the scenarios. Audio data were transliterated and phonetically transcribed. All steps were advised by a human being. The construction of the feature-values-relations and Petri net transducers (cf. next section) is an ongoing research.

Figure 2 shows the WoOF graphical interface as seen by the Wizard, mainly on the left side. We replaced the original smart home task in UCUI (the heating control feedback) on the right side by the planned interactive cooperative robot task to collect fruits and to hand them over to a user, suggested by the AI/robotic company 7Bot in [11]. In the future human-robotic experiments, we will use such a low cost platform (in this case ca. 350 USD) with open source software interface to collect our user inputs. Of course, for the WoZ experiments, the robotic arms will be manually controlled by the wizard. This will include actions to surprise users in order to collect more realistic cognitive user data in such a cooperative control task.

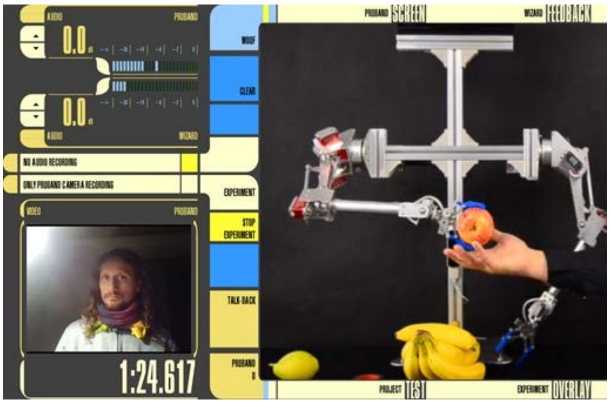

**Fig. 2.** Wizard panel from [2], conceptually extended by an interactive cooperative robot task (right, excerpt from [11]).

## 2.2 Employment of image schemas

The paradigm shift from mainly technology-centered devices – as described in introduction and preliminary WoZ experiments – towards human-centered devices requires the implementation of basic human models for a variety of human experiences.

In [12], we reported amongst others on an experimental investigation of image schemas as basic features of human knowledge that could help to support the development of more intuitive interfaces that should enable effective interactions with a system based on the subconscious application of basic prior knowledge according to Mohs et al. [13] and Turner [14]. One of such a fundamental aspect of human experience, and the focus of this investigation, was the impact of image schemas on human knowledge and language. Image schemas (e.g. up-down, center-periphery) are basic pre-conceptual, universal patterns of human experience that integrate information from multiple generic and perceptual modalities such as visual, acoustic and haptic information, i.a. suggested by Lakoff and Johnson [15]. They serve to structure human knowledge, behavior and experience. As basic building blocks of human knowledge, generated in earliest childhood interactions, image schemas may be available to all potential users. While previous research seems to support this hypothesis (e.g. Hurtienne et al. [16]), surveys on concept formation suggest that some image schemas occur earlier in infancy compared to other image schemas. Mandler and Pagán Cánovas [17] considered path schemas such as up-down, container, location, blockage, into, out of, open as basic image schemas occurring early in infancy at preverbal stages. On the contrary, center-periphery, scale, balance, cycle and other process schemas, near-far, multiplicity/unity as well as attributional image schemas (e.g. big-small, warm-cold) are built upon these basic primitives and, thus, should exert less influence on the development of human thought and knowledge [17].

To investigate whether developmental occurrence of image schemas influences their appliance in human speech interaction with machines/computers, we applied the WoZ paradigm and tested two hypotheses in [12]: First, we expected early, basic image schemas to be employed more often than later image schemas. Second, we expected no impact of individual difference variables on the frequency of applied image schemas, since these basic building blocks of human knowledge should be equally available to all users regardless of age, gender and technical experience. Forty-three German native speakers (20 men, 23 women; mean age 29.2 years, SD 9.9 years) participated in the speech interaction study. To calculate technical experience (TE), participants were asked to indicate the frequency with which they used a variety of technical devices at home (e.g. smartphone, laptop) and in public (e.g. ticketing machine, self-service banking). For seventeen items, participants had to select response options from 0 (I don't know this device) via 5 (almost daily) to 7 (more than once per day). The mean of all responses was then calculated for each participant with a higher score indicating greater technical experience. The TE score in our test persons ranged from 2.35 to 4.24 (mean TE score: 3.34). Both, age and gender were surprisingly not correlated with TE.

The speech interaction task took place in a quiet and moderately illuminated room. All participants were asked to complete the questionnaire ascertaining their technical experience as well as demographic data. Participants were then asked to stand in front of a multi-touch panel. They were presented with twenty test scenarios, investigating the image schemas-underlying free speech interaction with a heating device. They were not informed about image schemas in any way and were told that there are no right or wrong answers, since they have to test the functionality of a newly-developed heating device. Participants were asked to respond to the various scenarios, as they would also do at home and to simply tell the heating device which changes, if any, they would like the device to carry out. A content analysis was carried out to identify image schemas underlying utterances (see Table 1). Image schemas underlying speech and gesture responses were analyzed by two independent coders. To ensure reliability of coding, one coder coded the entire speech and gesture dataset. A second coder that was blind to the study hypotheses coded 25% of both datasets. In the speech interaction study, the two coders agreed 88% of the time.

Ten different image schemas (up-down, verticality, horizontality, balance, scale, warm-cold, process, near-far, multiplicity/ unit and container) were identified in all scenarios.

Each participant employed on average 6.7 different image schemas (SD 1.0) in speech interaction. The four image schemas up-down, verticality, horizontality and container were regarded as early, basic image schemas, whilst the six image schemas balance, scale, warm-cold, process (including cycle), near-far and multiplicity/ unity were regarded as later image schemas [17]. There were no significant effects for gender, age and technical experience (all p's > .05) on frequencies. In general, basic image schemas were employed significantly more often (67.8%) than image schemas occurring later in development (32.2%). In line with the second hypothesis mentioned, gender, age and technical experience were not correlated with both the frequency of early and late image schemas. This suggests that image schemas as basic building blocks of human knowledge are available to users of all ages and technical experience independently from gender.

**Table 1.** Coding of speech utterances into image schemas [12].

| Image schema | Example speech utterances |
|---|---|
| ***Basic*** | |
| Path: up-down | up/down, to rise/lower, higher/lower (to move/drive) up/down, to rise/to lower, higher/lower |
| Vertical path | to set (German: s*telle ein; stelle auf*) |
| Horizontal path | from…to, along, back/front, backwards/forward |
| Container | to put in, to hold, in the room, in the house, in daytime |
| ***Later*** | |
| Balance | to regulate, to adjust |
| Attribute: warm-cold | make it warmer/colder, to heat (up or down), to temper |
| Process (cycle, iteration) | to turn (up or down), to deactivate |
| Near-far | near xx degrees, next week, next day |
| Unity or multiplicity | whole day, whole week, a few (degrees) |
| Scale | to reduce, to cut, to increase or decrease, how much |

With regard to [13, 14] interaction should be based on (automatically retrievable) prior knowledge available to all potential users. The illustrated survey from [12], however, demonstrated that not all image schemas are equally intuitive in human computer interaction: The developmental occurrence of image schemas impacts upon the frequency of applied image schemas when interacting with technical devices. Early image schemas should, thus, be given preference over late image schemas in the interface design.

# 3 Concepts of semantic processing and behavior control

## 3.1 Feature-value-relation (FVR)

Cognitive user interfaces require a bidirectional translation between input signals and representations of meaning. While low-level signals are sequential, semantics is, in general, non-sequential. In [5], feature-values-relations (FVR) for representation and processing of semantic information were introduced. An example is depicted in Figure 3 showing an FVR for the speech input "Increase the temperature to 23 degrees on Saturday." where the relevant values of the input are related to semantic categories relevant for the system.

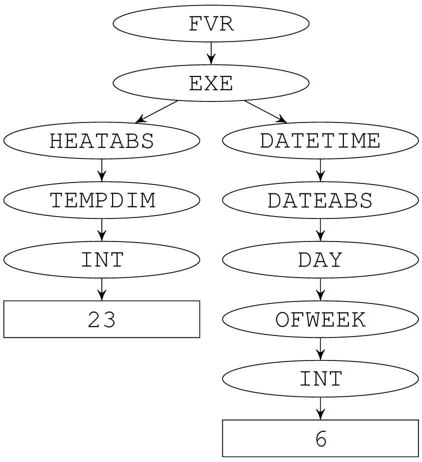

**Fig. 3.** FVR example as part of a heating control from [1].

These categories depend on the available actions of the system and the domains of usage, and are collected in a world model.

In [6, 7] FVRs are equipped with weights – which are omitted in fig. 2 – and related to language modelling, whereas [8] defines several operations on FVRs. A description of a behavior control – in an instinctive and an adapting version – building upon these operations is given in [12]. In our multi-modal system any input signals are transformed into FVRs representing the individual semantics (cf. exemplary discussion regarding a touchscreen in [12]). All FVRs lead to joint input semantics, which can then be unified with the current state (another FVR). This state serves as memory between dialog turns and contains all data gathered during an ongoing dialog. By comparison of the new state with the world model – and thus identifying the goal of a dialog – the semantics of an appropriate system action can be computed. The world model encodes what data is needed to execute a specific action. Whenever there is not enough data, another dialog turn requesting more data is initiated until execution of an action is possible. Such requests can be routed to different parts of the system, e.g. available sensors, a user model holding the user's habits, initiating a visual or auditory prompt for user input, any module from where the system can get and incorporate missing data. The flexibility of the approach arises from the fact that all processing is done in terms of FVRs and thus independently from the concrete system and any modalities of input and output. So from a behavior point of view it does not matter if the task is controlling a heating or taking part in a collaborative human-robot interaction.

## 3.2 Petri net transducer (PNT)

For the technical realization, the so called Petri net transducers (PNTs) were proposed [9, 10], that process labelled partial orders (LPOs), which in turn can represent FVRs. The application of PNTs to the bidirectional translation between sequences and partial orders allows us to build a seamless signal-to-semantics recognition network. Moreover we are able to prime this network by composition with a semantic structure representing an expectation on the next input. This expectation is a truly semantic one but adjusts the recognizer down to all low-level parts. On the other hand we use the same techniques for the synthesis side, where we can for example inject syntactical restrictions to adapt speech output to the users' wording. As with multi-modality on the input side, we can use different translation units to present the same semantics on different output channels.

### 3.3 Fock space

Based on a further concept described in [18], we are currently entering into a novel theory for mapping FVRs into a so-called Fock space from quantum mechanics. Given $V$ as the set of nodes of an FVR there exists a Hilbert space $V$ with the dimension $|V|$ since every element of $V$ is used as one basis vector. The Fock space is then defined as ⊕_{$n$=0}^{∞} $V$^{⊗$n$} where ⊕ is the direct sum and ⊗ denotes the tensor product. This allows us to use a different branch of mathematics for the processing of semantics. As a side effect we gain new insights as a first approach to discover semantic structures from data or using them for action planning as described in [18].

## 4 Realization as a hardware prototype

As described in [1], the first demonstrator was realized in August 2017 as an integrated circuit device, which still involves external power supply and a RS232 interface for the communication with arbitrary electronic devices. Four microphones, a loudspeaker and the touch panel were integrated. Figure 4 shows the main board.

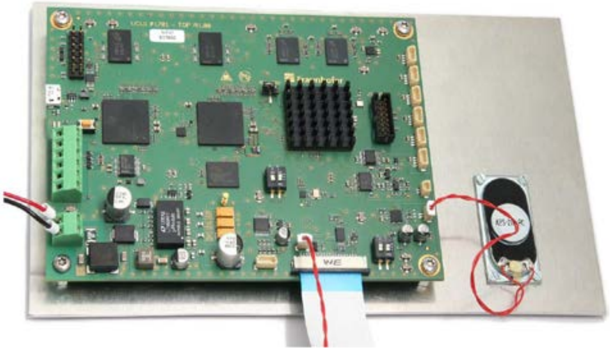

**Fig. 4.** First version of UCUI circuit board from [1].

The board (100 x 130 mm) includes two digital signal processors (DSPs), one Field Programmable Gate Array (FPGA), four RAMs, a flash memory, an audio codec and a motion sensor. The FPGA performs acoustic signal analysis, some other algorithms for speech recognition as well as signal and data routing. One DSP finalizes speech recognition and runs cognitive processes based on FVRs. Beyond it executes speech synthesis and controls the display. In next steps, our project partners Javox Solutions GmbH and XGraphic Ingenieurgesellschaft mbH will run their algorithms for beam forming, noise and echo cancellation as well as the gesture recognition on the second DSP.

## 5 Conclusions

In highly user-adaptive scenarios of human-machine and human-robot interactions, we suggest a decided cognitive modelling including semantic and behavior processing. The introduced UCUI prototype is able to communicate with users in parallel via speech, gesture recognition, speech synthesis and a touch screen. The device operates autarkic and supports a widely intuitive interaction to control different kind of *machines* surveyed with early and late image schemas. It can ensure user privacy by its system design and does not rely on network access.

Conventional interfaces cannot benefit from semantic prior knowledge. By using PNTs, semantic structures – corresponding to input signals within a multimodal hierarchical signal processing system – can be computed without premature decisions.

The current demonstrator is supporting the tasks of speech recognition, synthesis, semantic processing and a simplified cognitive behavior control on the embedded platform. After our next development steps, an extended behavior control model will enrich the interaction opportunities. Furthermore, the ultrasonic gesture control will be implemented.

This contribution has been developed in the project Universal Cognitive User Interface (UCUI) which is partly funded by the German Federal Ministry of Education and Research (BMBF) within the research program IKT2020 (grant #16ES0297). As associated partners we thank the ministry and the main project partners Fraunhofer IKTS, BTU Cottbus-Senftenberg, Javox Solutions GmbH, XGraphic Ingenieurgesellschaft mbH and Agilion GmbH for the decided work and for their previous publications as the foundation of this article.

## References

1. F. Duckhorn et al, Proc. ISCA Interspeech, 3435–3436 (2017)
2. M. Huber, O. Jokisch, Proc. of Knowledge Management Conference, Novo Mesto (Slovenia), June 2017. International Institute for Applied Knowledge Management, 31–40 (2017)
3. R. Hoffmann, M. Eichner, M. Wolff, Verbal and Nonverbal Communication Behaviors, 200–218 (2007)
4. M. Wolff, UASR: Unified Approach to Signal Synthesis and Recognition (2000) Available at: http://www.b-tu.de/en/fg-kommunikationstechnik/research/ projects/uasr
5. M. Huber et al, Proc. of Elektronische Sprachsignalverarbeitung (ESSV), series Studientexte zur Sprachkommunikation, **53**, 25–32 (2009)

6. G. Wirsching, Proc. IEEE 3rd Intern. Conference on Cognitive Infocommunications (CogInfoCom), 71–76 (2012)
7. G.Wirsching. R. Lorenz, Proc. of IEEE 4th Intern. Conf. on Cognitive Infocommunications, 369–374 (2013)
8. P. Geßler, *Kognitive Gerätesteuerung. Master thesis* (BTU Cottbus-Senftenberg, 2017)
9. R. Lorenz, M. Huber, G. Wirsching, Proc. of 35th Intern. Conference on Application and Theory of Petri Nets and Concurrency, **8489**, 233–252 (2014)
10. M. Huber, R. Römer, and M. Wolff, Proc. Elektron. Sprachsignalverarbeitung (ESSV), Saarbrücken, ser. Studientexte zur Sprachkommu-nikation, **86**, 122–129 (2017)
11. 7Bot: a low cost Robotic Arm that can See, Think and Learn! (www.7bot.cc, retrieved on 10/03/2018: https://www.kickstarter.com/projects/1128055363/7bot-a-powerful-desktop-robot-arm-for-future-inven).
12. M. Huber et al, Proc. of Knowledge Management Conference, (to be published)
13. C. Mohs, J. Hurtienne, M.C. Kindsmüller, J.H/ Israel, H.A. Meyer, IUUI Research Group, MMI-Interaktiv, **11**, 75–84 (2006)
14. P. Turner, Behavior & Information Technology, **27**, 475–482 (2008)
15. G. Lakoff and M. Johnson, *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought* (New York: Basic Books, 2011)
16. J. Hurtienne, et. al, Interactive Computing, **22**, 475–484 (2010)
17. J. Mandler, C. Pagán Cánovas, Language and Cognition, 1–23 (2014)
18. P. beim Graben, M. Huber, R. Römer, W. Wolff, Proc. Elektronische Sprach-signalverarbeitung (ESSV), **90**, 167–174 (2018)