*INTERSPEECH 2017: Show & Tell Contribution*
August 20–24, 2017, Stockholm, Sweden

# Towards an Autarkic Embedded Cognitive User Interface

*Frank Duckhorn*[1], *Markus Huber*[3], *Werner Meyer*[2],
*Oliver Jokisch*[4], *Constanze Tschöpe*[1], *Matthias Wolff*[2]

[1]Fraunhofer IKTS, Fraunhofer Institute for Ceramic Technologies and Systems, Dresden, Germany
[2]Brandenburg University of Technology Cottbus - Senftenberg, Germany
[3]InnoTec21 GmbH, Leipzig, Germany
[4]Leipzig University of Telecommunications, Germany

frank.duckhorn|constanze.tschoepe@ikts.fraunhofer.de, markus.huber@innotec21.de,
werner.meyer|matthias.wolff@b-tu.de, jokisch@hft-leipzig.de

## Abstract

With this paper we present an overview of an autarkic embedded cognitive user interface. It is realized in form of an integrated device able to communicate with the user over speech & gesture recognition, speech synthesis and a touch display. Semantic processing and cognitive behaviour control support intuitive interaction and help controlling arbitrary electronic devices. To ensure user privacy and to operate autonomously of network access all information processing is done on the device.
**Index Terms**: speech recognition, human-computer interaction, cognitive user interface

## 1. Introduction

Recent speech dialog systems and cognitive user interfaces allow natural verbal human-machine-interaction and achieve an excellent performance. However, leading commercial solutions heavily rely on transmitting sensitive user information (personal data, voice recordings, etc.) through public networks and on processing, storing and analyzing these data on servers of service providers. The demonstrator we present (see fig. 1) is realizing a cognitive user interface for intuitive interaction with arbitrary electronic devices. It ensures privacy by design. That means that all information processing is done on the device and that no user data ever leave the interface. To this end we develop a stand-alone hardware module doing all signal, speech and cognitive information processing. Interaction with the device takes place through speech, acoustic and visual symbols, and gestures. 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 a study of user-machine interactions in a real home-automation scenario and takes into account relevant legal and ethical aspects. For the demonstrator we reduced the task to the specific domain of controlling a heating installation which is also used in the study (see [1]).

The software of the system is developed within the Unified Approach to Signal Synthesis and Recognition (UASR, [2, 3]) maintained by BTU Cottbus-Senftenberg and Fraunhofer IKTS. The speech recognition and -synthesis engine is taken from that tool and ported to the hardware. We focus on the cognitive processing of meanings and the knowledge about the users habits. Semantic processing transforms all inputs (speech, gestures, touch display) into a unified representation. After cognitive behaviour control the unified representation can be transformed into any output channels (speech, acoustic signals, display).


Figure 1: *Current demonstrator*

## 2. Semantic processing and cognitive behaviour control

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.

On the basis of [4] we use *feature-values-relations* (FVR) for representation and processing of semantic information. An example is depicted in fig. 2 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.

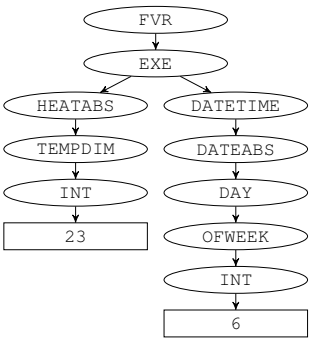


Figure 2: *Feature-values-relation*

In [5, 6] FVRs are equipped with weights – which are omitted in fig. 2 – and related to language modelling, whereas [7] defines several operations on FVRs.

In our multi-modal system any input signals are transformed into FVRs representing the individual semantics. All FVRs get unified resulting in one 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 comparing the new state with a world model – again an FVR – 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. Another memory holding user habits modelled as FVRs is available from where the system can also incorporate missing data.

In [8, 9] so called *Petri net transducers* (PNTs) are proposed, which 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.

## 3. Demonstrator

The demonstrator is realized in form of an integrated device. It has an external power supply and a RS232 interface for communication with arbitrary electronic devices. Four microphones, an loudspeaker and the touch panel are integrated in the case. Fig. 3 shows the demonstrator's main board.

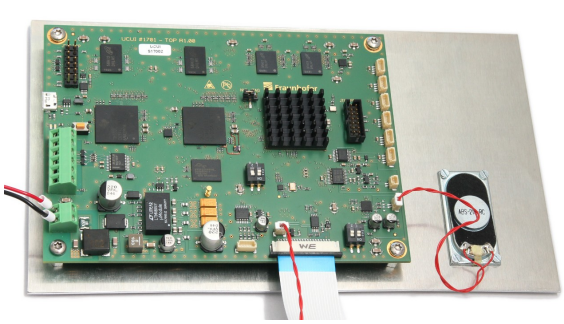

Figure 3: *Hardware board*

The board (100 x 130 mm) features two digital signal processors (DSP), one Field Programmable Gate Array (FPGA), four RAMs, a flash memory, an audio codec and a motion sensor. There are internal connectors for microphones, speakers, the display and debugging interfaces.

The FPGA performs acoustic signal analysis, some other algorithms for speech recognition as well as signal and data routing. One of the DSPs finalizes speech recognition and runs the cognitive processing based on FVRs. Additionally, it executes speech synthesis and controls the display. In further work our partners Javox Solutions GmbH and XGraphic Ingenieurgesellschaft mbH will run beam forming, noise and echo cancellation as well as the gesture recognition on the second DSP.

## 4. Conclusions and outlook

Traditional systems must prematurely *decide* for an input interpretation string (or a set of those strings) and thus cannot benefit from semantic prior knowledge. By using PNTs we can compute all weighted semantic structures corresponding to input signals within a multi-modal hierarchical signal processing system without any premature decision.

The current demonstrator is able to do speech recognition and synthesis, semantic processing and simple cognitive behaviour control all on the embedded hardware platform. Therefore it ensures the privacy of the user by design. In further work an extended behaviour control model will enrich the interaction opportunities. In addition ultrasonic gesture recognition will be implemented by our partners Javox Solutions GmbH and XGraphic Ingenieurgesellschaft mbH.

## 5. Acknowledgements

This work 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). We thank the ministry and our partners Javox Solutions GmbH, XGraphic Ingenieurgesellschaft mbH and Agilion GmbH for their support.

## 6. References

[1] M. Huber and O. Jokisch, "Cognitive Data Retrieval Using a Wizard-of-Oz Framework," in *Proc. of Knowledge Management Conference, Novo Mesto (Slovenia), June 21–24, 2017*. International Institute for Applied Knowledge Management, 2017.

[2] R. Hoffmann, M. Eichner, and M. Wolff, "Analysis of verbal and nonverbal acoustic signals with the Dresden UASR system," in *Verbal and Nonverbal Communication Behaviours*. Springer, 2007, pp. 200–218.

[3] M. Wolff, "UASR: Unified Approach to Signal Synthesis and Recognition (2000–...)," Online: https://www.b-tu.de/en/fg-kommunikationstechnik/research/projects/uasr, last visited: 31.03.2017.

[4] M. Huber, C. Kölbl, R. Lorenz, R. Römer, and G. Wirsching, "Semantische Dialogmodellierung mit gewichteten Merkmal-Werte-Relationen," in *Proc. of "Elektronische Sprachsignalverarbeitung (ESSV)"*, ser. Studientexte zur Sprachkommunikation, R. Hoffmann, Ed., vol. 53. Dresden: TUDpress, 2009, pp. 25–32.

[5] G. Wirsching, "Calculating semantic uncertainty," in *Proc. IEEE 3rd Intern. Conference on Cognitive Infocommunications (CogInfoCom)*, Dec 2012, pp. 71–76.

[6] G. Wirsching and R. Lorenz, "Towards meaning-oriented language modeling," in *Proc. of IEEE 4th Intern. Conference on Cognitive Infocommunications (CogInfoCom), Budapest (Hungary), December 2–5, 2013*. Piscataway, NJ: IEEE, 2013, pp. 369–374.

[7] P. Geßler, "Kognitive Gerätesteuerung," Master's thesis, Brandenburgische Technische Universität Cottbus-Senftenberg, Deutschland, 2017.

[8] R. Lorenz, M. Huber, and G. Wirsching, "On weighted petri net transducers," in *Proc. of 35th Intern. Conference on Application and Theory of Petri Nets and Concurrency, Tunis (Tunisia), June 23–27, 2014*, ser. Lecture Notes in Computer Science, G. Ciardo and E. Kindler, Eds., vol. 8489. Springer, 2014, pp. 233–252.

[9] M. Huber, R. Römer, and M. Wolff, "Little Drop of Mulligatawny Soup, Miss Sophie? Automatic Speech Understanding provided by Petri Nets," in *Proc. of "Elektronische Sprachsignalverarbeitung (ESSV)"*, ser. Studientexte zur Sprachkommunikation, J. Trouvain, I. Steiner, and B. Möbius, Eds., vol. 86. Dresden: TUDpress, Mar. 2017, pp. 122–129.