A TOOLKIT FOR 3D-GESTURE AND SPEECH DIALOG IN AUTOMOTIVE ENVIRONMENTS

Timo Sowa[1], Alexander Richter[1], Dietmar Fey[2]
[1] Elektrobit Automotive GmbH, Erlangen, Germany
[2] Friedrich-Alexander Universität Erlangen-Nürnberg, Germany

Keywords: dialog systems, 3D-gesture recognition, multimodality, semantic fusion

Abstract: 3D-gesture recognition has become a matter of growing interest for the automotive industry. Gesture input complements touch displays and the now widely established speech dialog systems in cars quite well, because it distracts less than touch, and facilitates continuous interaction styles not possible with speech. However, tools for implementing such interfaces that take advantage of synergies between the modalities are rare.

We present an extension of an existing toolkit for in-car HMIs [1], adding 3D-gesture (aka mid-air hand gesture) recognition and semantic fusion with speech. It enables non-programmers to easily model multimodal dialogs and develop prototypes. Our approach focuses of three kinds of interaction: Firstly, dialogs containing isolated gestures, *sequences* of speech and gesture, and *alternative* uses of voice commands and gestures can be modeled. This can be used, for instance, to accept or reject an incoming phone call using either modality. Secondly, we support binding a parameter of the hand configuration to a variable in the HMI model. That way, *continuous* gestural interaction, such as modifying the volume by moving the hand in 3D space with simultaneous graphical feedback, can be designed (virtual knob/slider) [2]. Thirdly, we allow modeling and processing of multimodal utterances with semantics distributed over speech and gesture. For instance, when pointing to a warning icon displayed on the dashboard while asking "what does that red icon mean"?

The paper describes how to model multimodal interaction using our tool extension and details the underlying technology of the run-time system. For gesture input we integrated the Leap Motion hand tracking system. The interaction logic is modeled based on state charts with one parallel state machine per modality. Continuous interaction is achieved by temporarily binding a hand model parameter to an HMI parameter within a gesture state. We implemented the fusion of distributed meaning using typed feature structures (TFSs) [3]. Speech and gesture recognizers separately provide interpretations in the form of partial TFSs. Available TFSs are described with a type system (ontology). The partial interpretations are independently sent to a short term memory. An integrator module tries to unify and put them together until a fully-specified TFS is created which can be processed as a user command. We further describe the application of our approach with a prototype model, demonstrating the three abovementioned use cases.

References:

[1] Massonie, D., Hacker, C. & Sowa, T. (2014). Modeling graphical and speech user interfaces with widgets and spidgets. In *Proc. of the 11th ITG Symposium on Speech Communication*. Erlangen, Germany.

[2] Latoschik, M.E. (2001). A general framework for multimodal interaction in virtual reality systems: PrOSA. In W. Broll & L. Schäfer (Eds.), *The future of VR and AR interfaces*. Yokohama, Japan (pp. 21-26).

[3] Johnston. M. et al. (1997). Unification-based multimodal integration. In *Proc. of the 35th Annual Meeting of the Assoc. for Computational Linguistics*, Madrid (pp. 281-288).