Speech Based, Multimodal Information Systems

L.J.M. Rothkrantz, B.Sletterink and M. Andre
Delft University of Technology, Knowledge Based Systems
Zuidplantsoen 4, 2628 BZ Delft, The Netherlands
Phone +312787504 Fax: +312787141
L.J.M.Rothkrantz@cs.tudelft.nl

August 11, 1999

Abstract

At Delft University of technology there is a project running on multimodal interfaces.

The objective of the research project is to establish, quantify, and evaluate techniques for designing synergistic combinations of human-machine communication modalities in the dimensions of sight, sound and touch in a multi-user environments. As a first step a simple prototype has been developed that could serve as a basis for complex systems.

When designing the system, one goal was to maintain a degree of flexibility. This means that the application can be changed easily and that new input devices are easy to add to the system. Each of the three modalities has a processing agent who communicates with the fusion agent. Spoken language is chosen as the common means of communication. Each modality module sends a text string, possibly combined with an XY coordinate pair, to the fusion agent. The fusion agent fuses the different inputs and produces either a command or feedback. The command needs to be executed by the application, while the feedback is processed through a text-to-speech synthesizer to produce spoken feedback graphical output.

Students recruited from the population of the DUT tested the system. The results of testing will be discussed in the paper.

Our approach to multimodal fusion is a variation of the slot-filler method, which is well known in artificial intelligence, see Winston (1992) or Allen (1994). Information necessary for command or language understanding is often encoded in structures called frames. In its most abstract formulation, a frame is simply a cluster of facts and objects that describe some typical object or situation. In our application, frames contain information about commands given by multiple modalities, such as information from touch screen and speech inputs.

The principal objects in a frame are assigned names, called slots. The slots can be viewed as functions that take values to instantiated the actual frame. Thus a particular instance of a frame represents a structured set of knowledge. Our approach to multimodal fusion will be illustrated in the paper.

At the Dutch International Airport arrive daily many thousand of travelers who want to have information on hotels, cities, restaurants. Human operators to provide this information are not always available. So an automatic information system was required. Recently Philips Company introduced Access Point, a voice controlled, interactive information system. This system can be used to develop a voice controlled multimedia information terminal. We developed and tested a prototype of such a system.