Cooperative Agents and Recognition Systems (CARS) for Drivers and Passengers

Luc E. JULIA
STAR Laboratory
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025
USA
julia@speech.sri.com

Adam J. CHEYER
Artificial Intelligence Center
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025
USA
cheyer@ai.sri.com

Abstract

In this paper we present SRI's vision of the human-machine interface for a car environment. This interface leverages our work in human-computer interaction, speech, speaker and gesture recognition, natural language understanding, and intelligent agents architecture. We propose a natural interface that allows the driver to interact with the navigation system, control electronic devices, and communicate with the rest of the world much as would be possible in the office environment. Passengers would be able to use the system to watch TV or play games in their private spaces. The final prototype will be fully configurable (languages, voice output, and so forth), and will include speaker recognition technology for resetting preferences and/or for security.

Keywords

Multimodal Interfaces, Speech and Speaker Recognition, Gesture Recognition, Natural Language Understanding, Cooperative Agents.

1. Introduction

New technologies such as Global Positioning System (GPS), wireless phones or wireless internet and electronic controls inside cars are available to improve the way we drive and manage the time spent in our automobiles. To manage this heavy flow of data and to keep the cognitive load as low as possible for the driver, we propose a solution based on our previous developments: a small, speech-enabled, touch display device that provides a combination of the best features of several interfaces we have developed over the past few years. This device can be used according to the specific task that has to be completed by the driver or the passenger.

The interfaces we have developed are the front ends to SRI's powerful framework, the Open Agent Architecture™ (OAA) [18], which allows a community of intelligent agents to work together to achieve user goals. To build multimodal systems, the key agents are those that recognize human signals such as speech or gestures and those that extract the meaning: the natural language understanding agent and the multimodal interpretation agent, for instance.

2. Natural Interfaces

The first prototype we built using Java™ combines different reused interfaces that were chosen according to the task. For each section of the system, we reference the full project for which it was developed. The user can select the tabs using both speech or deictic gestures. Each panel provides its own vocabulary and set of commands in addition to the main commands that allow navigation between the tabs.

2.1. Navigation System

The Multimodal Maps [1] allow the user to navigate maps naturally and query associated databases using speech, handwriting, and 2D gestures in a synergistic fashion on a pen computer. Using the same interface (replacing the pen with the finger) to query the navigation system and to display the GPS information gives the driver or the passengers the ability to plan the route and to get information from local or remote databases displayed on the map ("I want to go to Menlo Park." "Show me the restaurants around here.") (Figure 1). The GPS system guides the car along the chosen route using both the map display and a text-to-speech output. The interactions among all the agents belonging to the system, even if they

do not seem to be in use by the current visual interface, enables a great degree of proactivity from the system. For example, it could ask questions such as: “The tank is almost empty; would you like to find the nearest gas station?”. As well as a multimodal synergistic input interface, the system provides multimedia outputs such as iconic sounds, discriminative talking voices, images, or videos.

Figure 1. Navigation Panel

2.2. Electronic Device Control

Most of the cars will have numerous electronic devices that are accessible through a serial port using a predefined protocol. By connecting a computer and its multimodal interface, it will then be possible for the driver to control critical electronic devices such as cruise control or lights, and for everyone in the car to access comfort devices such as air conditioning, windows, sound, and entertainment (“Play CD 2, track 1.”) (Figure 2).

Priority should be given to the driver, possibly through speaker identification. Moreover, an interesting study [11] has shown that it is also possible to use the touch screen in blind condition (for the driver) to enter simple command gestures (down arrow to turn the volume down for instance).

Figure 2. Sound System Panel

2.3. Communication Center

The communication center is a remote office accessible by voice (Figure 3).

Figure 3. Communication Center Panel

This is an instance of the Automated Office (Unified Messaging), developed to show some capabilities of OAA [18]. The driver or passengers are able to browse the incoming emails by voice (even multipart/multimedia MIME messages), make phone calls, or send spoken notes. As basic features of OAA, filtering and triggering capabilities are included in each connected agent: “If email arrives for me about OzCHI, read it to me.” Plugging in an agent using speaker identification tech-

niques allows commands such as "If voicemail arrives for me from Larry, send an email to Patti" [9]. Intelligent cross-media conversion and adaptability to the current set of available or desired output channels is a key characteristic of the Unified Messaging prototype.

2.4. Recreation Area

The recreation center gathers some of the innovative speech-enabled prototypes developed by SRI International. Passenger-oriented, it assumes that each passenger creates a private multimodal/multimedia area (close talking microphone, personal touch screen and headsets). The passenger can play impressive 3D games enhanced with speech commands, search the Internet by voice, talk to an animated avatar, and watch TV in a more interactive way by asking naturally for the available programs with specific features. In addition, speech-based educational systems, such as WebGrader™ [16], provide a fun and effective way to learn foreign languages and their pronunciations (Figure 4).

Figure 4. Recreation Panel

2.5. Technical Information Access

The entire documentation of the car will be available on the Internet, making it easy to keep it up to date and possibly to personalize (via cookies) and control (via certificates) data for the car. This section extends the idea of the dialog with an avatar, or actor, implemented using the Microsoft™ Agent graphics [19].

If a warning message appears from a monitored device, a dialog with an automobile expert, played by the actor, will help to diagnose and fix the problem (Figure 5). The expert may also answer common questions such as "How much air should I put in my tires?" or "How should I talk to you?"

Figure 5. Diagnostic Panel

2.6. Setups

Speaker verification techniques can be used to access the setup panel and private areas (to configure and define the passwords for email and voice mail accounts, for instance). It will also be used to automatically retrieve the preferences for the current driver with respect to seat position, radio selections, temperature, mirror direction, and so forth.

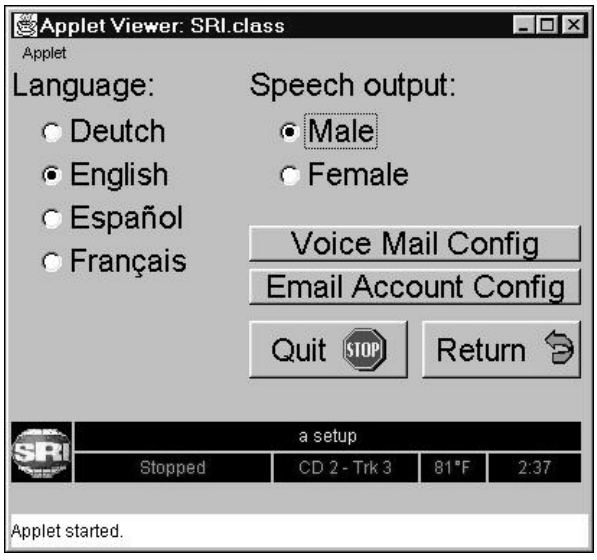

Figure 6. Setup Panel

3. Behind the Scene: the Agents

The functionality described above requires multiple Artificial Intelligence (AI) technologies (e.g., speech and gesture recognition, natural language understanding) to interact with each other and with commercial, off the shelf components such as email systems, map databases, and car electronics. SRI's Open Agent Architecture provides an infrastructure for integrating distributed components in a more flexible way than can be done through other distributed technologies such as CORBA, COM, or Java's RMI. The key difference in OAA's approach is that instead of components writing code to specify (and fix) their interactions and dependencies with other components, each agent (component) expresses its capabilities and needs in terms of a higher-level Interagent Communication Language (ICL). Each request for information or action is handled by one or more "facilitator agents," who break the request into subtasks, allocate subtasks to agents able to perform them, and then coordinate the flow of data and control among the participants. The architecture offers built-in support for creating natural user interfaces to the distributed services, since the logic-based ICL can be translated from and to natural language; users can speak a request in English, and the request can be acted upon by the community of agents, without requiring the user to specify or even know which agents are involved.

The advantages of the OAA approach include true plug-and-play, with new agents able to join the community of services at runtime; managed coordination of cooperative and competitive parallelism among components; heterogeneous computing, with components existing on diverse platforms, written in many programming languages; and enhance code reuse. Since components do not have hard-code dependencies written into them, we will be able to incorporate many existing agents from previous OAA-enabled systems [13, 18].

4. Recognition and Interpretation

4.1. Speech Recognition

Speech recognition, along with natural language, is a huge component of the multimodal user interface. While it is possible to use any speech recognition product available on the market to make an agent, we prefer the Nuance Communications[1] recognizer. Nuance is a real-time version of the SRI STAR Laboratory's continuous speech recognition system using context-dependent genonic hidden Markov models (HMMs) [4]. This technology recognizes natural speech without requiring the user to train the system in advance (i.e., speaker-independent recognition) and can be distinguished from the few other leading-edge speech recognition technologies by its detailed modeling of variations in pronunciation and its robustness to background noise and channel distortion. We plan to investigate automobile environments in more detail.

[1] SRI spin-off: http://www.nuance.com

4.2. Natural Language Understanding

In most OAA-based systems, prototypes are initially constructed with relatively simple natural language (NL) components, and as the vocabulary and grammar complexities grow, more powerful technologies can be incrementally added. It is easy to integrate different levels of NL understanding, depending upon the requirements of the system, just by plugging in an adequate engine. The available engines are two of our low-end NL systems: Nuance's template-slot tools and DCG-NL, a Prolog-based top-down parser. SRI's GEMINI [5] and FASTUS [7] are more powerful tools, used for complex NL tasks. To design the dialog on the fly, a visual tool is under development (Figure 7). It simulates the behaviors of the NL engine and creates the necessary code and data for the final NL agent.

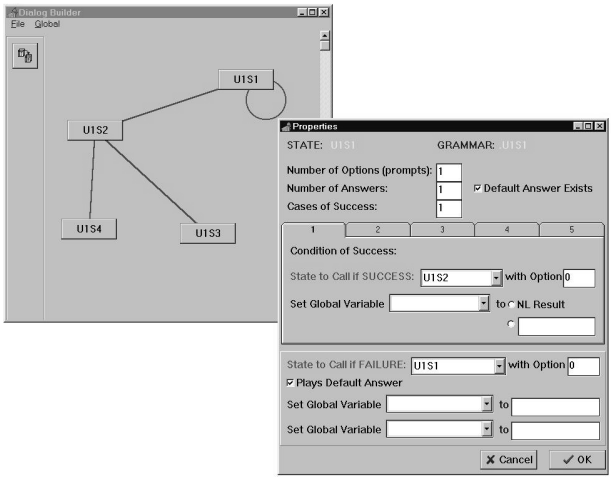

Figure 7. Visual Design Tool

4.3. Speaker Identification

Speaker identification technology has seen significant progress over the past several years. Although good performance can be achieved, several parameters affect accuracy. For example, systems trained on larger amounts of speech from the users will be more accurate. Similarly, variety in the training data (collecting over several days) will improve system robustness, and accuracy is higher for

longer test utterances. Computational limitations of the onboard platform will also be a performance factor. Perhaps the most significant of the factors is the variety in training data. The effects of mismatches between training and testing conditions can be dramatic [17]. A severe example, in the context of cars, of mismatching conditions would occur when a user trains the system with the engine off, in the garage, and then uses it with the top down on the freeway at high speed. We have made significant progress in reducing adverse effects of mismatches between training and testing conditions. In particular, SRI has developed a technology that reduces the effect from a factor of 30 to a factor of less than 3 [6]. The technology enables the user to train in a single session (one acoustic environment).

4.4. Gesture Recognition

The gesture modality is usually used in conjunction with speech to add spatial or deictic data to issue commands to the system. But sometimes a gesture (like a crossout) can carry both a location and semantic content. A set of current gestures (Figure 8) can be recognized using algorithms developed in [8].

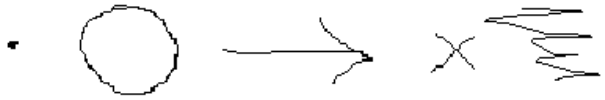

Figure 8. Gesture Set

In our experience [2], most gestures produced by users fall into this set. Since handwriting has rarely been used but we want to provide as many modalities as possible, we incorporated Communications Intelligence Corporation (CIC[2]) recognition routines. The handwriting recognizer is of interest in the navigation task where out-of-vocabulary names may appear, which are normally difficult for speech recognition systems to handle. Both the gesture recognizer and the handwriting recognizer are competing on the same data to find the right meaning.

4.5. Multimodal Fusion

Even if we consider speech as a privileged modality [10], numerous user studies [e.g., 12] have shown that most subjects prefer combinations of spoken and gestural inputs. In such examples, whereas speech plays a strong role in the acquisition of commands, combining it with a pointing device provides significant (8%) improvement in performance (recognition and understanding) over the use of speech in isolation. Not surprisingly, gestures provide a fast and accurate means of locating specific objects, while voice commands are more appropriate for selecting describable sets of objects or for referring to objects not visible on the screen. Many of these studies also attempt to enumerate and classify the relationships between the modalities arriving for a single command (complementary, redundancy, transfer, equivalence, specialization, contradiction). To model interactions where blended and unsorted modalities may be combined in a synergistic fashion with little need for time stamping, we first proposed a three-slot model known as VO*V* (Figure 9), such that

V or Verb is a word or a set of words expressing the action part of a command.

O* or Object[s] is zero or more objects to which the verb applies (zero if it is a system command).

V* or Variable[s] is zero or more attributes or options necessary to complete the command.

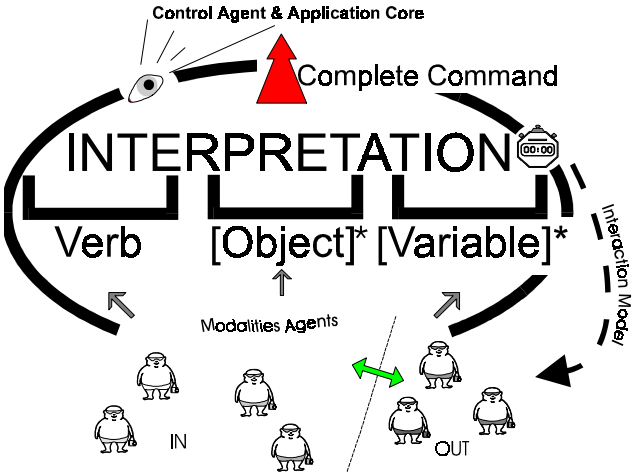

Figure 9. VO*V* Model

Input modalities produced by the user (handwriting, speech, gestures) fill slots in the model, and interpretation occurs as soon as the triplets produce a complete command. A multimedia prompting mechanism is also provided to assist the user in fulfilling an incomplete command. In addition, multiple information sources may compete in parallel for the right to fill a slot, given scored modality interpretations. This model has been shown to be easily generalizable, and has been applied to various application domains

[2] SRI spin-off: http://www.cic.com/

5. Evaluation

When building complex systems, it is important to perform user experiments to validate the design and implementation of the application. As described in [2], we have developed a novel "hybrid Wizard-of-Oz" approach for evaluating how well the implemented system functions for an experienced user, while simultaneously gathering information about future extensions or improvements as dictated by new users. The technique promotes incremental development of a complex system, from initial prototype through tested product, and provides a means for logging user interactions and quantifying system improvements at every stage of development.

6. Conclusions and Future Work

The unique feature of the proposed approach is that we integrate several very distinctive pieces, thanks to the OAA, even though they were not intended for this purpose. Further, we unify those pieces through a common, natural, multimodal interface using as much as possible human-to-human communication to avoid adding cognitive overload to the user. We achieved most of this aim by using good recognition systems and effective fusion and presentation techniques But to improve the reliability and robustness of the speech recognizer in real cars, we still have to address considerable noise and speaker adaptation issues (see, e.g. [3, 14, 15]). Finally, within a short period of time we plan to hook up real GPS and navigation systems and install our system in a moving car so that we can conduct user testing in real-life conditions.

7. Acknowledgments

Many thanks to Patti Price, Director of the Speech Technology and Research lab who spent a lot of time helping to design the CARS system.

8. References

[1] A. CHEYER and L. JULIA, “Multimodal Maps: An Agent-based Approach,” Proc. of Cooperative Multimodal Communication (CMC'95): Eindhoven, The Netherlands, 1995.

[2] A. CHEYER, L. JULIA and J. C. MARTIN, “A Unified Framework for Constructing Multimodal Experiments and Applications,” Proc. of Cooperative Multimodal Communication (CMC'98): Tilburg, The Netherlands, 1998.

[3] V. DIGALAKIS and L. NEUMEYER, “Speaker Adaptation Using Combined Transformation and Bayesian Methods,” Proc. of Intl. Conference on Acoustics, Speech and Signal Processing (ICASSP’95): Detroit, USA, 1995

[4] V. DIGALAKIS, P. MONACO and H. MURVEIT, “Genones: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognizers,” IEEE Transactions of Speech and Audio Processing, Vol.4, Num. 4, 1996

[5] J. DOWDING, J.M. GAWRON, D. APPELT, J. BEAR, L. CHERNY, R. MOORE and D. MORAN, “GEMINI: A natural language system for spoken-language understanding,” Proc. of 31st Annual Meeting of the Association for Computational Linguistics (ACL’96): Columbus, USA, 1996.

[6] L. HECK and M. WEINTRAUB, “Handset dependent background models for robust text-independent speaker recognition,” Proc. of Intl. Conference on Acoustics, Speech and Signal Processing (ICASSP’97): Munich, Germany, 1997.

[7] J. HOBBS, D. APPELT, J. BEAR, D. ISRAEL, M. KAMEYAMA, M. STICKEL, and M. TYSON, “FASTUS: a cascaded finite-state transducer for extracting information from natural-language text,” in Finite State Devices for Natural Language Processing (E. Roche and Y. Schabes, eds.) MIT Press, Cambridge, USA, 1996.

[8] L. JULIA and C. FAURE, “Pattern Recognition and Beautification for a Pen Based Interface,” Proc. of Intl. Conference on Document Analysis and Recognition (ICDAR'95): Montréal, Canada, 1995.

[9] L. JULIA, L. HECK and A. CHEYER, “A Speaker Identification Agent,” Proc. Audio and Video-based Biometric Person Authentication (AVBPA'97): Crans-Montana, Switzerland, 1997.

[10] L. JULIA and A. CHEYER, “Speech: A Privileged Modality,” Proc. of EuroSpeech'97: Rhodes, Greece, 1997.

[11] J. F. KAMP, F. POIRIER and P. DOIGNON, “A New Idea to Efficiently Interact with In-Vehicle Systems. Study of the use of the Touchpad in “Blind Condition,” Poster Proc. of Human Computer Interaction (HCI'97): San Francisco, USA, 1997.

[12] B.A. MELLOR, C. BABER and C. TUNLEY, “In goal-oriented multimodal dialogue systems,” Proc. of Intl. Conference on Spoken Language Processing (ICSLP’96): Philadelphia, USA, 1996.

[13] D. MORAN, A. CHEYER, L. JULIA, D. MARTIN and S. PARK, “The Open Agent Architecture and Its Multimodal User Interface,” Proc. of Intelligent User Interfaces (IUI'97): Orlando, USA, 1997.

[14] L. NEUMEYER and M. WEINTRAUB, “Probabilistic Optimum Filtering for Robust Speech Recognition,” Proc. of Intl. Conference on Acoustics, Speech and Signal Processing (ICASSP’94): Adelaide, Australia, 1994.

[15] L. NEUMEYER and M. WEINTRAUB, “Robust Speech Recognition in Noise Using Adaptation and Mapping Techniques,” Proc. of Intl. Conference on Acoustics, Speech and Signal Processing (ICASSP’95): Detroit, USA, 1995.

[16] L. NEUMEYER, H. FRANCO, V. ABRASH, L. JULIA, O. RONEN, H. BRATT, J. BING and V. DIGALAKIS, “WebGrader: A multilingual pronunciation practice tool,” Proc. of Speech Technology in Language Learning (STiLL'98): Stockholm, Sweden, 1998.

[17] NIST Speaker Recognition Workshop: Linthicum Heights, USA, 1996.

[18] SRI International web site on the Open Agent Architecture: http://www.ai.sri.com/~oaa/applications.html

[19] Microsoft web site about their agents and their animations: http://www.microsoft.com/workshop/imedia/agent/