Vision-Enhanced Multi-Modal Interactions in Domotic Environments

Jan Kleindienst, Tomáš Macek, Ladislav Serédi, Jan Šedivý

IBM Česká republika, Voice Technologies and Systems
{jankle, tomas_macek, ladislav_seredi, jan_sedivy}@cz.ibm.com

Abstract. This paper introduces key components of user interaction framework that harness speech recognition, multi-modality and computer vision in the residential environment. The interface technologies presented are being developed as part of "HomeTalk" IST-2001-33507 project. Its main goal is to design a platform to deliver and host user-centric services addressing the needs of a wide range of users spanning from technology enthusiasts to elderly and disabled. The technologies described in this paper will be tested and evaluated as part of HomeTalk field trials involving real users in their homes.

1. Introduction

As computers are becoming smaller and cheaper, we are witnessing there accelerated proliferation to various areas of our everyday life. We already see this in automotive industry. Not only there are dozens of microprocessors engaged in driving process at every moment behind the scene, there are also commercially-available telematic interfaces that allow drivers to control their air-conditions, radios, cell phones, navigation systems, etc. using voice and multimodal interfaces. Such a change on user interface frontier will continue to affect most of the traditional HCI patterns in cars, homes, offices, etc.

The HomeTalk EU project (IST-2001-33507) is oriented towards capturing some of the aspects of this UI paradigm shift in the residential environment, by designing and developing HomeTalk platform that runs various home services. Those user-centric services driven by multi-modal and vision-recognition technologies address the needs of technology enthusiasts as well as elderly and disabled. In this paper we will deal with the user interface technologies used in HomeTalk that make the advanced user interaction possible.

1.1. Background

With the advent of ubiquitous computing [Roth 02, Mattern 01] and smart spaces [Korhonen 01, Miller 01] various descriptions of proposed domotic IT architectures emerged. Most of them are component-based, often employing software agents [Issarny 01]. The user interface of these systems usually involves either roomware [Streitz 01], smart gadgets (everyday household items) scattered throughout the environment [Vanhala 01, Mattern01] or personalized mobile devices [Bielikova 01, Tuulari 01] connected wirelessly to the home network. These mobile devices can be equipped by novel interfaces to create visually appealing user experience [McAlester 02] and to compensate their inferior input capabilities (instead of the tiny keyboard use voice recognition) [Rössler 01]. One of the most important target groups for domotic systems are the disabled and elderly users [Korhonen 01], which too require specially tailored user interfaces.

1.2. Overview

The heart of HomeTalk platform is a residential gateway (RG), a low-resource Linux–powered computer that lacks moving parts such as fans and hard disks to ensure high reliability and low operating costs. Acting as a household communication hub, the RG runs HomeTalk services, while supports connectivity to various outdoor and indoor networks. The indoor network consists of various devices, including sensors, actuators, and white appliances connected in the home environment. The RG integrates data networking, control networking, and voice technology (using IBM Embedded ViaVoice Technology) for speech recognition/synthesis and multi-modal support. Moreover, the RG runs a central control component of the HomeTalk system that serves as a backend that processes and reacts to user interface actions.

The user requests that are captured by the backend are coming from various clients connected to the HomeTalk platfrom, including a VoiceXML browser, a multi-modal PDA (Section MACI - Multimodal Appliance Control Interface), and even from the actuators and white-appliances operated by the user. Apart from these explicit user actions, HomeTalk uses visual recognition software to collect specific cues from the interaction environment (Section VISIONARY). Where such visual context help enhance the dialog in the "interaction zones" in the vicinity of the appliances, the wirelessly connected PDA complementary offers remote monitoring and control capabilities.

1.3. Approach

In the design of HomeTalk platform and service capabilities the project tried to follow the user-centric design methodology based on ISO 13407. At the initial stage, the user requirements from the various target groups including technology enthusiasts, elderly & disabled, as well as their care takers were collected. The result of such user survey was analyzed and served as an input for the selection of HomeTalk supported and their user interface features. The reader can find more details in [Dermo03]. The main conclusion of the user survey pertaining specifically to UI aspects was the user interface should be based on standards and follow the design-for-all principle to ensure maximum accessibility. The most desired residential services across all groups were detection (fire, smoke, water leakage), emergency function, control of actuators, electronic devices remote control, control of actuators, central entrance opening, and white appliances wizard.

To illustrate the UI features presented by HomeTalk services, we present a part of a scenario of one of HomeTalk services - White Appliance Wizzard:

…When Mr. N. approaches the oven with the ready-to-cook pizza, an oven-mounted camera detects his face and displays the recommended cooking parameters for the pizza on the oven display. At the same time, it synchronizes the display of a multi-modal PDA assistant held by his mother-in-law. Mr. N. can thus adjust the values using either turning the oven knobs or engaging the multi-modal oven interface on the PDA.

Peeking at the PDA display, his mother-in-law argues that the recommend temperature is too low and the cooking time too short. Mr. N. shrugging in defeat suggests she alter the values. The old lady does so by speaking to the PDA ("Set cooking time to forty-five minutes. Increase the temperature to 200 degrees.") and encouraged, she even tries to fool the PDA assistant with "This should be ready by 1 pm, so set the start time accordingly," astonished that the command actually works as the oven displays the calculated start time.

From her wheelchair she demands Mr.N. to show her how to use the PDA to check the status of the washing machine in the basement. As Mr.N. moves out of the oven direct interaction zone, the oven-mounted camera detects that. Based on this cue, the oven gently notifies him that to initiate the cooking process he must press the START button – either on the oven panel or on the PDA…

The user interface functionality described in the scenario snipped is made possible mainly by two components that we introduce in the rest of the paper: a) Multi-modal Appliance Control Interface (MACI), the multi-modal runtime component that in this incarnation runs on the PDA and b) Visual Context Acquisition Component (Visionary), which drives the appliance-mounted cameras. The central component called Djinn runtime to which the above components are connected, and which thus collects and maintains user and environment context, is not described in this paper.

2. MACI - Multimodal Appliance Control Interface

As we indicated earlier, a successful user interface for a home IT environment must be based on properly formulated user needs, expectations and preferences. As Hometalk is focused not only on early adopters but on elderly users as well as those with various kinds of physical disability we need to examine UI issues from several different perspectives. For example, accessibility of command interfaces has paramount importance for users with mobility problems. Users with cognitive problems (often elderly ones) will need mechanisms to avoid overheating and danger of fire for certain appliances (oven, iron…) as well as regular alarms to remember everyday tasks as taking medicine, making phone calls, etc. Visually impaired users will need speech-driven interfaces while those with hearing loss require visual communication.

Appliances and devices integrated into Hometalk can be controlled not only directly in their interaction zones (assisted with the Visionary discussed in a separate section). They can be also controlled remotely by standard devices – i.e. a PC equipped with www browser, a telephone or mobile phone - or custom-made controllers with specific functionality and user interface.

To experiment with innovative HCI techniques we choose to implement a mobile client running on an off-the-shelf PDA, namely HP iPaq equipped with PocketPC 2002 operating system. The input-output capabilities of these portable devices are rather limited: the size of their screen is relatively small compared to the desktop systems and naturally they lack a full-size keyboard. To overcome these inherent limitations and to acquire better user experience we decided to equip the iPaq with our embedded speech recognition system. In our implementation the speech recognition functionality blends seamlessly with the GUI resulting a so called multi-modal system.

Multi-modal technology allows the interchangeable use of multiple modalities of input and output, such as voice commands, keypads, or stylus -- in the same interaction. With this new blend of UIs it is possible to run speech and GUI interactions either in parallel or exclusively, based on the running application and the actual user's requirements.

MACI has been designed as and general multi-modal runtime built on GUI-rich MacroMedia Flash and IBM VoiceXML technology. The UI design is general enough to accommodate different types of remote control and monitoring.

2.1. MACI: GUI

The PDA screen is divided into five functionally different areas (*see figure 1.*):

1. Speech recognition feedback bar: we discuss this GUI element in the next section.
2. Network connection indicator; if green, MACI is connected an can accept user commands as well as receive notifications from the Hometalk network.
3. Appliance switcher tabs: because of the relatively small screen of PDA we decided to use the well proven tabbed notebook metaphor: only one appliance can be controlled at a time, but switching between control interfaces can be done anytime by tapping the icon on the respective tab.
4. Status line: messages indicating the status of the current appliance as well as system-wide alarms and error conditions are displayed here.
5. Appliance-specific area: control interfaces of particular appliances (including oven, washing machine but also plant irrigation or room light control systems) can be displayed here. To ensure coherent user experience this part of screen closely mimics the existing displays and command dials of appliances. Usually all settings of an appliance can be controlled just by the four arrow keys - the horizontal ones changes the focus (highlighted by a blinking yellow rectangle) the vertical arrows set the value of the item in focus. Tapping the displayed button(s) send commands to the appliance.

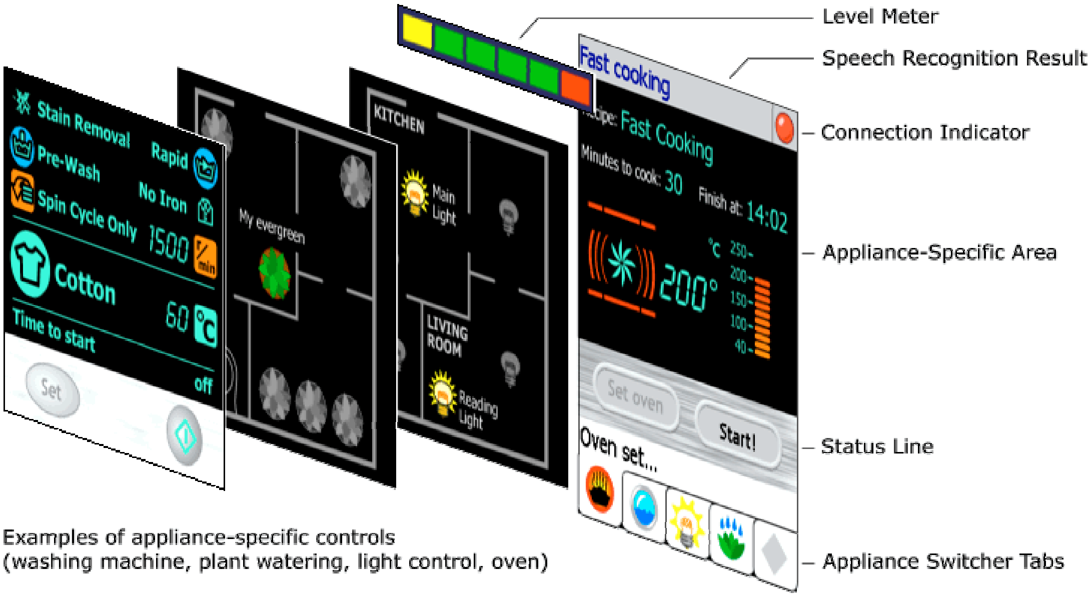
In spite of the relatively small screen we strived to enhance the usability of our interface for elderly/disabled users by using large fonts for displaying messages. The application switcher tabs are as well large enough to be tapped by a thumb eliminating the need for stylus and thus allowing one-handed operation. The pop-up help screen is easily accessible throughout the system by pressing one of the existing five “hardware” buttons of iPaq.

2.2. MACI: Speech recognition

Almost all of functions accessible via GUI (as described above) can be controlled by spoken commands, which in most cases is faster and more efficient than pressing buttons. One of the advantages of the speech input is selecting more than one item in a single utterance. Imagine, for example, a task of setting the oven before cooking: we need to specify the type of cooking, the temperature and the duration. Even though the GUI displays these items as three separate selections, with the help of speech recognition the user can set them all in one utterance, by saying e.g. “Gratin at 250 degrees for twenty minutes”. Through selecting items by a point-and-click interface is typically faster then filling them by voice one-by-one, the above example shows that while designing multi-modal interfaces one must be aware the specific strength (and weaknesses) of each modality.

In our implementation speech recognition works in “push-to-speak” mode: the user presses the record button at the upper-left corner of the PDA utter the command and releases the button. Activating the microphone only during the utterance greatly increase the robustness of speech recognition and in the same time allows comfortable one-handed operation of the PDA.

As we are dealing with multi-modal (and not speech-only) interfaces there are visual feedbacks that enhance the usability of speech recognition gathered on the speech recognition feedback bar. First there is an audio level indicator providing feedback while uttering commands and which helps the user maintain the optimal loudness for recognition. Afterwards the recognized phrase is displayed here (if the phrase is too long, this area can be temporarily enlarged).



Examples of appliance-specific controls
(washing machine, plant watering, light control, oven)

fig. 1. Exploded view of the MACI GUI

3. VISIONARY

Human interaction is based on generating and processing of various visual cues. For example, if a person turns a head toward another individual, it indicates attention which usually prompts the start of the conversation. Using such cues is natural for human to human communication and would enhance human-computer interaction as well.

In this section we describe the visual component that designed in the Hometalk project. The purpose of the component is to collect visual information from cameras, transform it to more concise higher level knowledge and pass it then to other parts of the system. In this way we achieve to enhance the user interactions and provide better user experience. The Visionary component, based on Intel OpenCV library, currently supports the following features:

Movement detection: Visionary calculates change in subsequent images. If it is larger than a threshold, it rises an event.
Human face detection: Visionary detects presence of human face in front of the camera. It reports if the face appears, disappears or changes the size.
Laser pointer detection: Algorithm detects position of laser pointer spot on a picture and reports its coordinates.

We will illustrate how having these visual cue acquisition mechanism available makes the user interaction more natural. We considered a home environment with various home appliances and tried to enhance their functionality. One of the examples which we implemented is an oven equipped with camera. The camera driven by the Visionary makes the aware of the user presence. For example, if the user approaches the oven, i.e. enters the oven “interaction zone”, the “face-in” event is generated by the Visionary software. If the user stays facing the oven for some time (controlled by a threshold), the event is interpreted by the backend logic as the “user attention” event. The oven reaction to such situation differs based on its operating state. If turned off, the oven display lights up upon user approach. If the oven is in the cooking mode, it gives the user a quick report about the progress of cooking ("there are 30 minutes left"). Upon leaving, it informs the user about how much cooking time is remaining.

The binding of the user state events to appropriate actions is flexible and can be tailored to user needs. For example, the “attention” event can be used to inform a person on a wheelchair of the particular state of the appliance, without requesting the user to touch the appliance control panel. As the Visionary provides the information on the relative distance of the face from the camera we can define the “depth” of the interaction corridor and tune its shape to the specifics of the environment.

Obviously, more visual cues with proper interpretation mean more nationality to the HCI interaction. There are other examples of visual clues which would be useful for human computer interaction, but which not were implemented in the course of HomeTalk include:

Tracking people helps to select microphones, focus cameras or to direct calls to various locations.
Face detection is useful for security application, and also in general to personalize application behaviors.
Lips reading is used to enhance speech recognition.
Emotion detection can be used to modify dialog flow and to make it more users friendly.
Gesture detection. Detection of various gestures allows adding whole new modes of operation to the application. The user can combine speech with gestures. Commands like "move this object from here to there" can be used. Similarly handy is to be able to *recognize laser pointer* on the projected screen.

Visual processing is in general time consuming task and dealing with images requires significant memory and communication resources. The architecture must reflect this fact and keep communication and computation requirements moderate.

We separated visual processing in our architecture to a separate subsystem. This is the decision which has been made based on the type of the applications we are dealing with. The application logic needs to know when certain event happens (image recognized, brightness changed) and parameters of the event (magnitude of the change, ID of the recognized face, distance of the moving object). The video stream itself is not of an interest. In some cases an image or a short video sequence is needed, typically taken shortly before and/or after the moment of the event. For example image of the intruder moving in the house. Instead of sending the images to the mean logic of the application (where it is not needed), it is more convenient to make it available to all other components of the system. Then it can be grabbed directly by output subsystem.

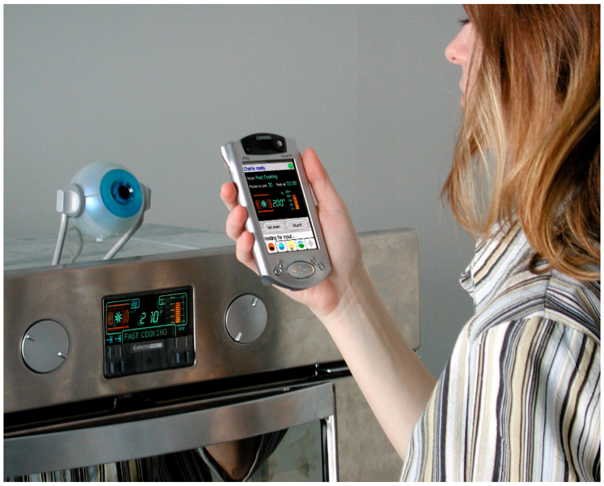


fig. 2. User setting the oven's temperature via MACI

The video stream images are recorded by Visionary to the circular image buffer. The images are kept in the buffer together with their time stamps. Buffer is continuously updated so the oldest images are overwritten by new ones. The application can, therefore, request an image taken at the particular time, presumed that the time is not too far in the history. The specific images (or video sequences) which are of the particular interest of the application can be stored by Visionary on request, independently on image buffer mechanism.

Some of the events are initiated by the Visionary itself. Typically any reporting about something happening in the video stream. For this type of operations we use a subscription mechanism. The logic of the application informs Visionary that it needs certain type of events by sending out a subscription message.

4. CONCLUSION

We have introduced a work-in-progress HCI technology that uses speech, multi-modality and visual cue acquisition software to interact with users in residential environment. We have described the two main interface components and shown how multi-modal and vision support the design-for-all principle.

HomeTalk technology will be evaluated in real user homes in Athens and Madrid run by TID and TEMAGON, respectively. Since this is a first-of-a-kind set of trials we expect to receive valuable feedback in acceptance of this novel user interface in general, and voice as well as multimodal technology in particular.

5. AKNOWLEDGEMENTS

The authors would like to thanks all HomeTalk partners IBM, MERLONNI, WRAP, ANCO, TID, TEMAGON, and INACCESS NETWORKS for their contributions during the various activities presented in this paper. Further information about the project can be found at http://www.hometalk.org.

6. REFERENCES

[Weiss 02] S. Weiss, *Handheld Usability*, John Wiley & Sons, 2002.

[VoiceXML 01] *VoiceXML 2.0,* W3C, http://www.w3.org/TR/2001/WD-voicexml20-20011023, W3C Working Draft, Oct 2001.

[Multimodal 00] *Multimodal Requirements for Voice Markup Languages*, W3C, http://www.w3.org/TR/multimodal-reqs, W3C Working Draft, June 2000.

[Oviatt 00] S. Oviatt, *Ten myths of multimodal interaction*, Available at: http://www.cse.ogi.edu/CHCC/Papers/sharonPaper/Myths/myths.html, 2000.

[Ramaswamy 99] G. Ramaswamy, J. Kleindienst, D. Cofman, P. Gopalakrishnan, C. Neti, *A pervasive conversational interface for information interaction*, Eurospeech 99, Budapest, Hungary, 1999.

[Kleindienst 02] J. Kleindienst, L. Seredi, P. Kapanen, J. Bergman, CATCH-2004 *Multi-Modal Browser: Overview Description with Usability Analysis*, IEEE Fourth, International Conference on Multimodal Interfaces, Pittsburg, USA, October 14-16, 2002.

[Kleindienst 03] J. Kleindienst, L. Seredi, P. Kapanen, J. Bergman, *Loosely-coupled approach towards multi-modal browsing, Special Issue "Multimodality: a step towards universal access"*, Springer International Journal, Universal Access in the Information Society, 2003.

[McAlester 02] D. Mc Alester, M. Capraro, *Skip Intro: Flash Usability and Interface Design*, New Riders Publishing, 2002.

[Carroll 02] J. Carroll, *Human-Computer Interaction in the New Millennium*, Addison-Wesley, 2002.

[Rössler 01] H. Rössler, J. Sienel, W. Wajda, J. Hoffmann, M. Kostrzewa: Multimodal Interaction for Mobile Environments, International Workshop on Information Presentation and Natural Multimodal Dialogue, Verona, Italy, December 2001.

[Roth 02] J. Roth, *Patterns of Mobile Interaction*, Personal and Ubiquitous Computing, Vol. 6, Issue 4, Springer, 2002.

[Streitz 01] N. A. Streitz, *Roomware: Towards the Next Generation of Human-Computer Interaction*, ECRIM News, No 46, European Research Consortium for Informatics and Mathematics, July 2001.

[Bielikova 01] M. Bielikova, T. Krajcovic, *Ambient Intelligence within a Home Environment*, ERCIM News No.47, October 2001.

[Vanhala 01] J. Vanhala, *A Flood of Intelligence - the Living Room Project*, ERCIM News No.47, October 2001.

[Korhonen 01] I. Korhonen, *Ambient Intelligence and Home Networking for Wellness Management and Home Automation*, ERCIM News No.47, October 2001.

[Miller 01] F. Miller, *Wired and Smart: from the Fridge to the Bathtub*, ERCIM News No.47, October 2001.

[Tuulari 01] E. Tuulari, *Enabling Ambient Intelligence Research with SoapBox Platform*, ERCIM News No.47, October 2001.

[Issarny 01] V. Issarny, *Offering a Consumer-Oriented Ambient Intelligence Environment*, ERCIM News No.47, October 2001.

[Mattern 01] F. Mattern, *Ubiquitous Computing Infrastructures*, ERCIM News No.47, October 2001.

[Dermo 03] V. Dermosoniadis, G. Georgopulos *Smart Homes: a user perspective*, 19th International Symposium on Human Factors in Telecommunication, Berlin, 2003.