# CHAT – Towards a general-purpose infrastructure for multimodal situation-adaptive user assistance

Carmelo Ardito[1]

Thomas Pederson[1,2]

Maria Francesca Costabile[1]

[1]Dipartimento di Informatica
Università di Bari
70125 Bari, Italy
+39 080 544 3300

[2]Department of Computing Science
Umeå University
SE-90187 Umeå, Sweden
+46 90 786 6548

{ardito, costabile, pederson}@di.uniba.it

top@cs.umu.se

## ABSTRACT

In this position paper, we describe CHAT, an upcoming project aimed at providing multimodal context-sensitive services to mobile users. We specifically describe two conceptual corner stones of the project of relevance to the topic of user assistance in intelligent environments: a) a multimodal interaction framework targeted at providing service access through different modalities for different real-world situations and for improving interaction with mobile devices in general, b) an "egocentric interaction" model for framing interaction with objects in the vicinity of a mobile user, including also other real-world and/or computational entities than the mobile device itself, ranging from computationally "stupid" everyday objects to more advanced interactive devices such as desktop PCs. The final section of the paper is devoted to open issues in the design of the CHAT infrastructure related to the topic of user assistance in intelligent environments.

## Categories and Subject Descriptors

D.3.2 [**Design Tools and Techniques**]: User interfaces; D.2.11 [**Software architectures**]: Domain-specific architectures; H.1.2 [**User/Machine Systems**]: Human factors; I.3.6 [**Methodology and Techniques**]: Interaction techniques; K.8 [**Personal Computing**]: Miscellaneous.

## General Terms

Design, Human Factors.

## Keywords

Multimodal interfaces, mobile human-computer interaction.

## 1. INTRODUCTION & PROBLEM DESCRIPTION

Mobile computing implies computing in more varied physical contexts than desktop computing. Different situations enforce different constraints as to what kind of device interaction and assistance that is needed; technically possible to offer; socially acceptable to perform; etc. To some extent, interaction style is also a matter of preference (e.g. some people prefer headsets when using cellular phones, others prefer using only the phone device itself).

In the CHAT project, "Cultural Heritage fruition & e-learning applications of new Advanced (multimodal) Technologies", we intend to develop a software infrastructure that allows services accessed through thin clients such as cellular phones or PDAs to be a) adaptable to personal preferences of the user, with focus on the choice of interaction modalities, and b) adaptive to the physical-virtual context of the human actor carrying the device. In both cases, the proposed architecture should be open both for channeling interaction between services and user through the mobile device itself, as well as through available input and output facilities in the vicinity. Furthermore, real-world phenomena sensed by the device itself or indirectly through external sensor pools will be made available through the CHAT infrastructure as a resource for service developers to effectively design "intelligent" environments.

The multimodality of interest for us permits users to interact with the system using several input channel simultaneously, classified by W3C as simultaneous co-ordinated multimodality [8]. In the research context there are already similar systems (e.g. [2]) and empirical studies (e.g. [1]) targeting this kind of multimodality, but also commercial tools with these interaction features are currently developed.

The architecture for supporting these kind of multimodal systems is more complex than traditional interactive systems, because we have to consider:

- parallel recognition modules for each input channel: every module produces fragments of the overall input that must be combined to become meaningful;
- a general methodology to interpret the meaning of the input fragments;
- a time-sensitive analysis process to determine which fragments must be combined to become meaningful;
- a module to manage the overall user/system dialogue;
- criteria to adapt the input/output modalities to the users' needs and the environment in which they actually are.

We deliberately try to make the infrastructure as general as possible because we believe that multimodal adaptability and adaptive features are beneficial independent from the area of application. For the purpose of evaluation however, the CHAT project will develop mobile multimodal prototype applications related to two particular activities: e-learning and exploration of cultural heritage.

### 1.1 E-learning

*E-learning* is an area investigating the possibilities of improving the learning process by using information technology. It enables new forms of learning including distance learning and just-in-time

learning. Advantages include new possibilities in distributing learning material through both computer and communications technology. Such devices can include personal computers, CDROMs, Television, PDAs, and Mobile Phones. Communications technology enables the use of the Internet, email, discussion forums, collaborative software, classroom management software. Courses can be tailored to specific needs and asynchronous learning is possible. The "any time, any place" nature of e-learning could be a winning strategy for particular needs, such as decongestion of overcrowded education facilities, support for learners or lecturers who live far from schools and universities, life-long education. Moreover, it could be a valuable opportunity for specific groups of learners, such as disabled learners, if the learning material is actually accessible to them.

## 1.2 Cultural heritage exploration

*Cultural heritage* is the legacy of physical artefacts and intangible attributes of a group or society that are inherited from our ancestors, maintained in the present and bestowed for the benefit of future generations. The term "cultural heritage" has not always meant the same thing. Having at one time referred exclusively to the monumental remains of cultures, heritage as a concept has gradually come to include new categories such as the intangible, ethnographic or industrial heritage. A noteworthy effort was subsequently made to extend the conceptualization and description of the intangible heritage. This is due to the fact that closer attention is now being paid to humankind, the dramatic arts, languages and traditional music, as well as to the informational, spiritual and philosophical systems upon which creations are based. Physical or "tangible cultural heritage" includes buildings and historic places, monuments, artefacts, etc., that are considered worthy of preservation for the future. These include objects significant to the archaeology, architecture, science or technology of a specific culture. "Natural heritage" is also an important part of a culture, encompassing the countryside and natural environment, including flora and fauna. These kind of heritage sites often serve as an important component in a country's tourist industry, attracting many visitors from abroad as well as locally.

## 2. FRAMEWORKS

The system development is guided by a multimodal design framework for ensuring state-of-the art support for multimodal interaction, and an egocentric interaction design framework for guiding the analysis and design of context-aware services.

## 2.1 Multimodal design framework

We propose a framework inspired by W3C's Multimodal Interaction Framework that identifies major components for any multimodal system (see Figure 1).
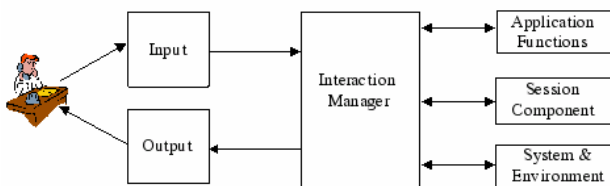


**Figure 1. The W3C Multimodal Interaction Framework [7].**

The W3C multimodal interaction framework is not an architecture, but a modeling framework one step above in abstraction. It describes neither how components are allocated to hardware devices, nor how the communication system enables the hardware devices to communicate. [7]

### 2.1.1 Components of the W3C multimodal interaction framework

*Human user* - A user who enters input into the system and observes and hears information presented by the system.

*Input* - An interactive multimodal implementation will use multiple input modes such as audio, speech, handwriting, and keyboarding, and other input modes.

*Output* - An interactive multimodal implementation will use one or more modes of output, such as speech, text, graphics, audio files, and animation.

*Interaction manager* - The logical component that coordinates data and manages execution flow from various input and output modality component interface objects. The interaction manager maintains the interaction state and context of the application and responds to inputs from component interface objects and changes in the system and environment. The interaction manager then manages these changes and coordinates input and output across component interface objects.

*Application Functions* – The services that should be offered to the users.

*Session component* - An interface to the interaction manager to support state management, and temporary and persistent sessions for multimodal applications.

*System and Environment component* - A component that enables the interaction manager to find out about and respond to changes in device capabilities, user preferences and environmental conditions. For example, which of the available modes, the user wishes to use — has the user muted audio input? The interaction manager may be interested in the width and height of the display, whether it supports colour, and other capability and configuration information.

The framework designed for the CHAT project (see Figure 2) is compatible with the Multimodal Interaction Framework previously discussed.
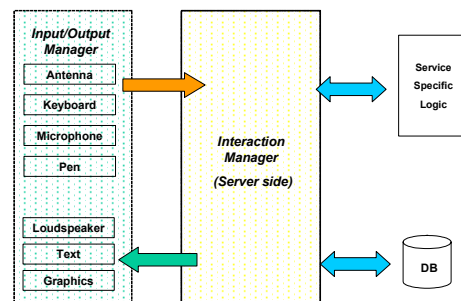


**Figure 2. CHAT Multimodal Interaction Framework.**

The *Input/Output Manager* is a "lightweight" software running on the user device with the responsibility of managing the input and output channels. I/O Manager captures the input fragments coming from several channels and transmits them to the Interaction Manager running on a server using suitable standard protocols.

The *Antenna* in the framework symbolizes high-level position information received through for instance GPS, GSM, or RFID technologies.

The *Interaction Manager (IM)* receives the multimodal input fragments from the user's device and processes them to obtain a meaningful input.

The *Service Specific Logic* acts on the input received and potentially produces an output transmitted to the IM.

Finally, the IM generates a multimodal representation of the required service that will be presented to the user by output channels suited to the user's preferences and needs as well as current environmental context and device type.

## 2.2 Egocentric interaction framework

The user interface design in the CHAT project will be guided by the *egocentric interaction* framework [3], inspired by the currently popular view within cognitive science that human individual actions are to a large degree influenced by what the specific individual can perceive of the surrounding environment. Based on a an integrated view on physical and virtual space[1] [4] objects in the proximity of a particular human actor can be categorized as being situated in one out of four spaces, at any given point in time (see Figure 3).
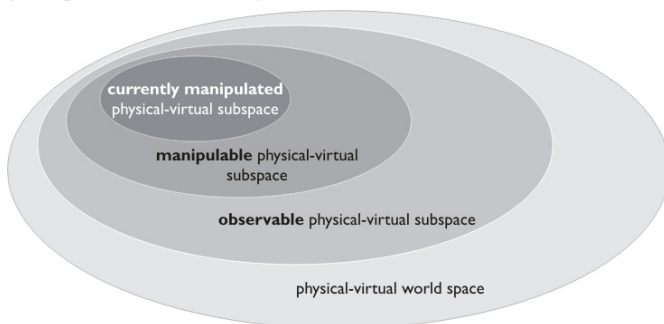


**Figure 3. A situative model of physical-virtual space from the perspective of a specific human actor at a specific point in time. Adapted from [4].**

The borders between the different subspaces are based on assumptions of the specific human actor's perceptive and cognitive experience of the current physical-virtual environment, e.g. real-world objects and/or virtual objects (presented by, and accessed through computing devices) in the immediate vicinity of the specific human actor. The egocentric interaction framework is based on the belief that computer system models should be

[1]  In short, "physical space" denotes the euclidean three-dimensional space of the real world. "Virtual space" is the multi-dimensional hyperspace (locally often euclidean two-dimensional) accessed through interactive computing devices such as cellular phones, PDAs and desktop computers.

closely tied to the cognitive and perceptive models that human actors construct and maintain as part of everyday life.

The term 'egocentric' has been chosen to signal that it is the human body and mind of a specific human individual that (sometimes literally) acts as a centre of reference to which all interaction modeling and activity assistance is anchored.

### 2.2.1 Components of an egocentric interaction system

The conceptual system architecture (illustrated in Figure 4) is based on a wearable computing/sensing hardware configuration consisting of a *private black box* offering computing power and storage space for data generated by an *egocentric interaction sensor pool* monitoring object-centric phenomena within the observable physical-virtual subspace of a specific human actor. Furthermore, the private black box runs a *physical-virtual operating system* hosting both advanced physical-virtual applications developed by software developers as well as simpler programs designed by the user her/himself. Such applications can incorporate the manipulation of both physical objects (e.g. a sculpture at a museum) and virtual objects (e.g. a web page describing the same sculpture). Explicit interaction with the physical-virtual operating system is performed through a *private black box user interface*, either fitted onto the private black box itself, or running on a general-purpose device like a PC. Implicit interaction [5] with the physical-virtual operating system emerges whenever the user interacts with a physical or virtual object inside the manipulable physical-virtual subspace (see Figure 3) monitored by the private black box. The local computing of the private black box can optionally be enhanced by communication with publicly and ubiquitously accessible *shared object knowledge repositories*, distributing anonymous data about objects and their everyday use.
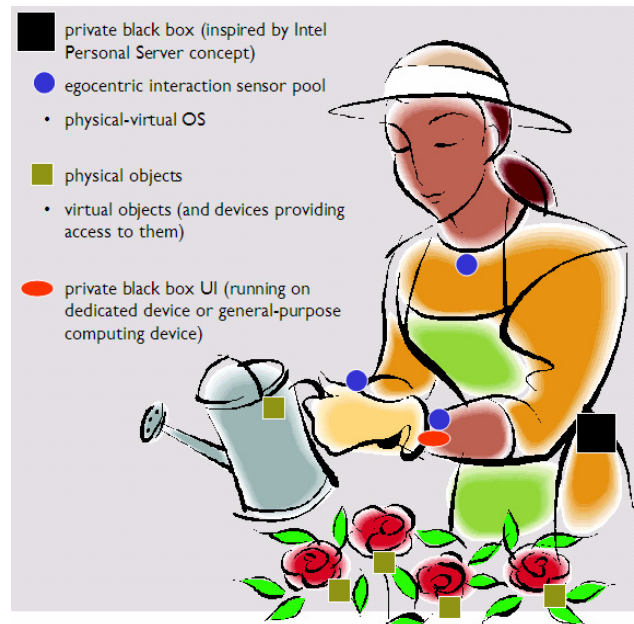


**Figure 4. General components of an egocentric interaction system. (Virtual objects and shared object knowledge repositories not pictured.)**

### 2.2.2 Potential application areas for egocentric interaction systems

- (physical and virtual) object logistics support — make sure you have everything you need, make sure unnecessary distracting objects are out of the way

- activity sequence support — make sure you do it in the order necessary

- physical-virtual gap bridging support — allow for smooth transitions between performing actions in the physical world and in the virtual world

- provide access to virtual environments and objects, including to place virtual objects in physical space, to place physical objects in virtual space

- provide manual search and recall functionalities within the log database of previously performed activities

- provide simple support for end-user development / programming of physical-virtual applications

- extension of the user's senses: providing info on when the bus will arrive, how far away the person you should meet is located

## 3. FUTURE WORK

The first step of the project will be to identify a set of concrete use cases in the two application areas of interest. We hope to acquire knowledge and hints from this workshop that can help us make important design decisions, e.g. when in the studied activities user assistance would be useful and in what form (e.g. which kind of assistance, through what modality, and when).

## 4. OPEN ISSUES AND WHY WE WOULD LIKE TO ATTEND THE WORKSHOP

We would come to this workshop with more questions than solutions although we do believe that the multimodal framework and the egocentric interaction model could contribute and help framing the discussion on user assistance in intelligent environments.

Of the questions listed in the workshop call, we find the following very relevant for our aim in CHAT:

- **Which user activities and tasks require assistance?** What method or heuristic should we use to identify the most useful and practically achievable assistance for e-learning and exploration of cultural heritage?

- **How should the designer choose the best sensing and interaction technologies for a scenario?** Having a small mobile device as computation and communication hub between user, ("intelligent") environment, and server (see Figure 4): how should the service logic (see Figure 2) be developed in order to seamlessly cope with ad-hoc appearance and disappearance of sensors and actuators external to the device itself?

- **How should multiple users with concurrent activities be supported?** E-learning and exploration of cultural heritage are activities which often dynamically change from being performed in isolation to being performed in groups. When is synchronous communication in a community better then asynchronous? When experiencing a common phenomena or facing a common problem, in what way can a group of users be encouraged to assist each other in order to increase the knowledge level and experience quality as a group?

- **How should the current state of the user assistance system be represented, especially when dealing with multiple tasks?** We believe that the egocentric interaction perspective, although still very abstract, constitutes an interesting step towards modeling complex and concurrent tasks because a) it attempts to capture what the user views as important at any given time, b) it does so by making no distinction between objects of interest residing in the physical world and objects in the virtual world (i.e. inside computing devices).

Of the key topics mentioned in the workshop call we have the following starting points for discussion:

- **How to unify the complementary concepts of public and personal devices in IEs**. Being inspired by the Intel Personal Server concept [6], We propose to regard interaction *data* as private, and interaction *devices* (including mobile and wearable ones) as potentially public. Distribution of interaction devices among individual users is transparent and can follow established social rules, while distribution of digital data is by nature invisible and thus has to be restricted. From another point of view, the sought-for seamless interplay between (personal) mobile devices and more stationary computational "intelligence" is an important requirement, because without smooth such mechanisms, knowledge of individual preferences will have to be inferred on the fly rather than retrieved (at least in an "occasional user" scenario), making the environment look significantly less "intelligent".

- **How to model user activity (terminology, structure, notation) for the design of IEs.** We believe to offer some ideas towards framing activity in intelligent environments by proposing the egocentric interaction framework which centres the attention of the modeler to the users body and mind, and which regard physical and virtual objects as residing in the same physical-virtual space. This stance should be interesting because intelligent environments are almost per definition relying on a combination of real-world and computational phenomena. There is of course more to human activity than mere object manipulation and we look forward to getting feedback and ideas for complementing the model.

## 5. SHORT AUTHOR BIOGRAPHIES

*Carmelo Ardito* is, since November 2004, Ph.D. Student of the School of Informatics at the Dipartimento di Informatica, University of Bari. He got the laurea degree in Computer Science at the University of Bari in June 2002 and the laurea specialistica degree in Computer Science at the University of Bari in October 2004, discussing a thesis on "an experimental evaluation of the use of audio in virtual environment navigation". Since 2001 he has been research collaborator of the Department of Computer Science in various projects, sponsored by European Union and Italian organizations. His current research interests are in the Human-Computer Interaction field, particularly Information Visualization Usability Techniques, Mobile Systems, Web-based systems.

*Thomas Pederson* received his PhD from the Dept. of Computing Science at Umeå university, Sweden, in December 2003. Faculty opponent at the dissertation event was professor William Buxton, University of Toronto, Canada. His PhD thesis proposes a uniform and integrated view on physical and virtual objects, environments, and activities. Thomas did his MSc project at Ericsson Media Lab in Stockholm, and has spent six months as visiting research assistant at Fraunhofer IPSI in Germany prior to his PhD studies as well as two months in the ACES research group in Rennes, France, after his PhD dissertation. He is currently project manager of easyADL, a two-year project intended to investigate the use of Ubiquitous Computing technology to support activities of daily living for individuals suffering dementia disease. In addition to the managing of the easyADL project in Sweden, Thomas holds a post doc fellowship at Bari university in Italy for 2006 and 2007.

*Maria Francesca Costabile* is full professor at the Computer Science Department of the University of Bari, Italy, where she teaches HCI for the computer science curriculum. Her current research interests are in both theoretical and application oriented aspects of visual formalisms for information representation and querying, visual system design, visual data mining, adaptive interfaces, user models, multimodal and multimedia interaction, web interfaces, system usability. She is regularly in the program committees of international conferences and workshops and has been Program Co-Chair of INTERACT 2005. She has been General Chair of the International Working Conference on Advanced Visual Interfaces (AVI) in 2004 and Program co-Chair of AVI '96 and AVI'98; she is in the AVI steering committee (AVI is organized in cooperation of ACM SIGCHI). She is senior member of IEEE and member of ACM and ACM SIGCHI. She was founding member of the Italian Chapter of ACM SIGCHI (SIGCHI Italy) and served as chair from 1996 to 2000.

## 6. REFERENCES

[1] Oviatt, S., De Angeli, A. and Kuhn, K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. In Proceedings of CHI'97, Atlanta, Georgia, USA, 18-23 Apr, 1997, pp. 415-422.

[2] Paternò, F., & Giammarino, F. (2006). Authoring interfaces with combined use of graphics and voice for both stationary and mobile devices. Proceedings of the International Conference on Advanced Visual Interface 2006 (AVI 2006), Venice, Italy, May 23-26, 2006, pp. 329-335.

[3] Pederson, T. (2006). Egocentric Interaction. Workshop on What is the Next Generation of Human-Computer Interaction?, CHI2006, April 22-23, Montréal, Canada.

[4] Pederson, T. (2003). *From Conceptual Links to Causal Relations — Physical-Virtual Artefacts in Mixed-Reality Space*. PhD thesis, Dept. of Computing Science, Umeå university, report UMINF-03.14, ISSN 0348-0542, ISBN 91-7305-556-5. Permanent URL: http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-137

[5] Schmidt, A. (2002). Ubiquitous Computing – Computing in Context. PhD thesis, Computing Department, Lancaster university, U.K.

[6] Want, R., Pering, T., Danneels, G., Kumar, M., Sundar, M., Light, J. (2002). The Personal Server: Changing the Way We Think about Ubiquitous Computing. Proceedings of UBICOMP 2002, Göteborg, Sweden, Springer Verlag, September 2002, pp. 194-209.

[7] W3C Multimodal Interaction Framework. http://www.w3.org/TR/mmi-framework/

[8] W3C Multimodal Interaction Activity. http://www.w3.org/2002/mmi/