Multimodal Interaction in Distributed and Ubiquitous Computing

Marc Pous

Barcelona Digital Centre Tecnològic
Barcelona, Spain
mpous@bdigital.org

Luigi Ceccaroni

Barcelona Digital Centre Tecnològic
Barcelona, Spain
lceccaroni@bdigital.org

Abstract — **This paper presents a multimodal, service-oriented architecture based on a distributed implementation of W3C's Multimodal Architecture and Interfaces applied to ubiquitous computing. The specific goal is to create a platform able to provide multimodal interactions between people with functional diversity and several semantic services developed by third-party companies and consumed through a display, in order to make content more accessible and interactive. The general purpose of this research is the development and deployment of urban, ubiquitous services for people living in or visiting a city. And the motivating scenario behind the design and development of the multimodal platform and semantic services is one in which people use media poles and confidently interoperate with the surrounding environment in a multimodal way.**

Keywords - Human-computer interaction, multimodal interfaces, multimodality, adaptive interfaces, semantic Web services

I. INTRODUCTION

Information and communication technology systems suffer from an inability to satisfy the heterogeneous needs of many users. For example, media poles present the same static information to people living in or visiting a city, people with widely differing knowledge of the city and widely differing needs. E-stores do not personalize the selection of featured items to customers with different needs, but just take into account (sometimes) user preferences. Virtual museums offer the same guided tour to visitors with very different goals, interests and functional diversity. Health information systems (with very few exceptions) present the same information to patients with different health problems. A remedy for the negative effects of the traditional one-size-fits-all approach is to develop systems with an ability to adapt their behavior to the goals, tasks, interests, needs, functional diversity and other features of individual users and groups of users.

A new generation of context aware systems and ambient intelligence platforms with functionalities such as mobility and ubiquity [14] [2] [5] are appearing in our society. The idea of such environments emerged more than a decade ago in Weiser's [15] seminal article and their evolution has recently been accelerated by improved wireless telecommunications capabilities, open networks, continued increases in computing power, improved battery technology and the emergence of flexible software architectures [10].

Intelligent, multimodal interaction is a relatively young research area. Starting with a few pioneering works on adaptive hypertext in early 1990s [11], it now attracts many researchers from different communities such as knowledge representation, user modeling, machine learning, natural language generation, information retrieval, intelligent tutoring systems, cognitive science, and digital education.

Nowadays, the interaction and the interface between humans and computers are not very natural yet; although they are often lightly multimodal (considering, for instance, the mouse or other pointing devices alongside the keyboard). The appearance of touch screens improved user's navigation through the graphical user interface, but did not introduce a real multimodal interaction between users and computers. Multimodal interactions and multimodality refer to the process in which different devices and people are able to interact aurally, visually, by touch or by gesture. Ubiquitous computing refers to this interaction happening anywhere, anytime, using any device, often in order to increase the accessibility and improve the user experience.

Multimodality is applied to present usable and accessible information to users, including videos, images and texts. One of the main purposes of multimodality is the improvement of user interactions with devices, such as smartphones or public kiosks. Information access and interaction for users with functional diversity, is another main purpose of multimodality [1][1].

User interfaces should allow users to interact with the content or a service through the most appropriate mode or through multiple modes, taking into account user's preferences and context. Interactions can be defined as information exchange among people, technologies (represented by user interfaces) and processes. The user may determine the mode or modes of interaction that he prefers for accessing information more naturally. Multimodality extends and improves user interfaces because it allows the integration of voice, image and other types of data input such as keyboards, mice, webcams, pens, touch screens and Wii remotes[2].

---

[1] See T.V. Raman, "Open, Interoperable Means for Integrated Multimodal Interaction," in W3C Workshop on Multimodal Interaction, 2004, [http://www.w3.org/2004/02/ mmi-workshop/raman-ibm] last visited on February 2010.

[2] See J. Chung Lee, "Low-Cost Multi-point Interactive Whiteboards Using the Wiimote," [http://johnnylee.net/ projects/wii/] last visited on February 2010.

This paper discusses a multimodal architecture design based on W3C's Multimodal Architecture and Interfaces[3], and an implementation on an Interactive Community Display (ICD) [2][4]. The proposed framework is general and flexible enough to provide interoperability among modality-specific components coming from different service providers, for example: speech recognition from one service provider, sign language recognition from another one and sign language avatar from a third one.

The multimodal ICD experimental platform is a Web interface running on media poles in public spaces such as streets, attractions or libraries. The ICD's screen is a 52-inch vertical touch screen with computer, webcam, speakers and microphone integrated (see Figure 1). The enrichment due to multimodality includes the possibility for the user to interact with the ICD through different sensory channels: using the voice, the Spanish sign-language, and the ICD's touch screen. The system also includes a service able to detect the user's emotion through the analysis of voice and physiognomy.

The structure of this paper is as follows. The next section introduces the INREDIS project, which is the context of the research and development of the ICD, and the service-oriented architecture. Section 3 describes the proposed multimodal architecture, detailing the conceptual model and the architecture components. Finally, Section 4 presents the architecture implementation scenario and a discussion about the possible evolution of the multimodal architecture.

A. *The INREDIS project*

The INREDIS project[5], which is the context of the research described in this paper, constitutes a new approach to applying accessible technology. Until now, technological advances in accessibility were generally product modifications aimed at making the product usable to people with functional diversity. INREDIS aims to take a technological leap based on developing a system able to make devices already available on the market interoperable, using a universal interoperability architecture, which can adapt to new market standards while maintaining compatibility with earlier systems, as well as ensuring ease of use and universal access.

Figure 1: Interactive Community Display (ICD)

II. SERVICE-ORIENTED ARCHITECTURE

These technological developments will have a great social repercussion worldwide as they lead to major advances in accessibility for people with functional diversity and enhance their quality of life. This paper presents the architecture of an experimental platform designed and developed within the human-computer interaction research line of INREDIS.

B. *Service-oriented human-computer interaction*

Multimodal interfaces offer flexibility providing multiple channels for interaction [4] [3] [13] [6][6]. The main idea of the experimental platform developed is to manage distributed services that offer the capability of processing and synthesizing the multiple modalities of interaction of the user interface. These services are enriched with semantics and adapted to user necessities through an orchestrator and a service bus which manage the multimodal, distributed, semantic services depending on user's preferences, necessities and modeling [9].

The aim of the experimental platform is demonstrate that is possible to offer multimodal interfaces in a distributed architecture responsible of the sensory input and output services, orchestrated through a multimodal and interaction manager. The approach presented in this paper is based on W3C's Multimodal Architecture and Interfaces, improves some of its runtime framework sub-components and explores connecting the modality components to the framework as a remote service.

---

[3] See "Multimodal Architecture and Interfaces," W3C Working Draft 1 December 2009. [http://www.w3.org/TR/mmi-arch/] last visited on February 2010.

[4] See also M. Johnston and S. Bangalore, "Multimodal applications from mobile to kiosk," in W3C Workshop on Multimodal Interaction, 2004, [http://www.w3.org/2004/02/ mmi-workshop/mmiwsshopjohnston.pdf], and M. Pous, L. Ceccaroni, M. Palau, and V. Codina, "Ubiquitous, social networks in the street," in W3C Workshop on the Future of Social Networking, 2008, [http://www.w3.org/2008/09/ msnws/papers/Ubiquitous_social_networks_in_the_street.pdf] last visited on February 2010.

[5] See [http://www.inredis.es] last visited on February 2010.

[6] See K. Wang, "From Multimodal to Natural Interactions," in W3C Workshop on Multimodal Interaction, 2004, [http://www.w3.org/2004/02/mmi-workshop/wang-microsoft] last visited on February 2010.

The orchestrator or Interaction Manager (IM) manages the interactions among the ICD user interface, the distributed services and the users. In the ICD interface, content is adapted depending on user's preferences. The service bus or Multimodal Server (MS) manages the distributed services connections, services protocols and languages, using semantic Web services and streaming (or p2p) protocols to send (and receive) user's interactions with the ICD.

C. *Interface and user's interactions*

The ICD's user interface is browser-based and runs in Gecko, a layout engine currently developed by Mozilla Corporation. The Web-based environment facilitates rapid prototyping and integration of different applications (mash-ups) and languages, such as voice and video players, maps, external content, JavaScript and Ajax, which are very useful to build usable and multimodal interfaces.

In the experimental platform developed, the interactions are represented using a finite-state graph in order to parse the multiple input streams and manage the user's navigation through the ICD interface. The next state of the ICD interface depends on user's state, preferences and the explicit modal interactions. User interfaces available during user's navigation at the ICD are represented as nodes. The transition between nodes depends on the interaction's commands, independently of the modality (by voice, by gesture or by haptic interaction). This system follows event-driven rules, and the event dispatcher is included into the IM, which understands the messages of the multimodal semantic services and merge them with the state diagram to infer the following state, which will be deployed for the user as a new interface adapted to her necessities and abilities [8].

III. MULTIMODAL ARCHITECTURE

The ICD service-oriented architecture enables the development of software that is delivered and consumed on demand depending on the user's preferences. The benefit of this approach lies in the loose coupling of the software components that make up the application. Semantic services discovery mechanisms can be used for finding and selecting the functionality that the ICD, depending on the user that is interacting with it, is looking for. This loose coupling among the services and the ICD platform allows programmers to select the protocol of communication while still being able to access the functionality of services that are using different methods.

The components in the ICD architecture can be divided into three main groups (see next sections and Figure 2), following the W3C philosophy of a system as a model-view-controller (MVC) pattern managing all data and interactions among the users, the system and the external services.

D. *Task Manager*

The main goal of the Task Manager (TM) is to manage user petitions. The TM has all the user profiles serialized, and knows the context and the user's preferences (which are managed by the User Modeler) in order to adapt the system and the content according to user's necessities and context. In order to know the user, the first user's interaction is to get identified to the system.

The TM is able to get the content of third-party distributed services in order for the IM to generate the user interface. The content will be obtained through the Service Broker which would manage Semantic Web Services depending on the information that should show the ICD or the device and the necessities of the user. The Service Broker role would be discovers and manages Semantic Web Services in order of presenting the right information depending on the IM requests.

As an MVC pattern component, the TM is the model, because it interacts with the IM, making transparent all the accesses to services, the message transformation, and the protocol used by the MS. The User Modeler is another component of the model.

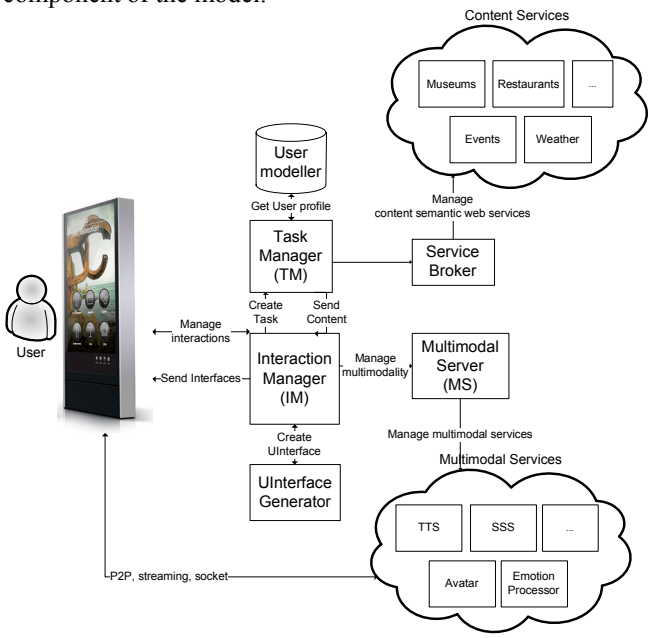

Figure 2: Multimodal architecture

E. *Interaction Manager*

The IM is the controller into a MVC pattern and it is the core of the architecture presented in this paper. The IM's role is the management of the user interactions between the ICD and the user. The TM creates a task, generating an instance of the user's context in order to inform the IM about the user's preferences, and the IM adapts the content and integrates it with multimodality for the user. The user interface is generated with the User Interface Generator (UI Generator) module, which would be the view.

The IM manages the MS, which will switch on the services involved in user's interactions (through streaming or p2p protocols) and send the messages with the user's request (possibly in natural language) to the IM. This will send the request to the TM as a new user-task in order to analyze (possibly using natural language processing) the user petition and generate the new user-interface.

F. *Multimodal Server*

The MS is included into the controller modules with the IM and it is able to manage all third-party distributed services in order to process and synthesize human-computer interactions. The main goal of this module is to provide communication, in the experimental platform, among the interface and the multimodal services able to understand the user and communicate with her. The MS provides uncoupling between the services and the IM, as well as the reusability of different services from different parties without knowing who is providing the service.

The MS activates all needed interaction processes (recognition and synthesis) when the IM requires a service. The connection between a distributed service and the user uses p2p or streaming protocols directly from the ICD or the user's (mobile) device. For example, if the user is interacting with the system using the voice, the voice recognition process will know how to connect with the user's device because the MS sends it a message with the IP address, port and other security parameters. The voice recognition service gets a direct connection with the user's device (or the ICD), processes the user's interaction and generates a message with the processed interaction, which is sent to the MS. The MS will then send it to the IM that generates a new task to implement the user's command. All messages among all services and system processes are exchanged using XML based language.

G. *User Interface Generator*

The User Interface Generator (UIG) which represents the view module it is able to create an adapted interface with all user's contextual information, needs and preferences [1]. The ICD's user interface (see Figure 3) as been designed as a composition of two types of content:

- The upper part of the interface shows all the multimodal content, represented through an avatar or a map (or both).
- The lower part shows the content related to the sites mapped above, or the content being read or being represented as sign language by the avatar.

The UIG module is also able to adapt the interface of the user's device, depending on the display size, the situation and the context.

H. *Multimodal services*

Multimodal services are managed directly by the MS. Distributed, non-local services currently available in the experimental platform are:

- *Voice service*: a voice recognition service and an emotional voice synthesizer, which receives an XML message[7] (using the SSML format to represent the emotion into the text) and generates an audio file which is sent and played through the user's interface.
- *Sign language service*: a Spanish sign language recognition service, which streams the image from the user's device or ICD's webcam to a remote sign language processor process which translates the signs into a textual message and sends it to the IM.
- *Avatar service*: a service which, from a textual message (possibly with emotions), can generate an avatar which represents the text as Spanish sign language (with subtitles); the signs get then synchronized with the voice service and a video is generated and reproduced through the user's interface using p2p protocols for communication.
- *Emotion service*: a service to analyze and recognize the user's emotion through the ICD's webcam, necessary in order to communicate a correct emotional response to the user; it uses the video feed (with the face image) and the audio signals for the analysis.

Figure 3: Examples of the ICD interface

I. *Content services*

Content services are third-party, semantic, Web services (e.g., city services, metro and bus services, travel agency services, weather services) managed through the Service Broker, which will be in charge of discovering and requesting new services, and adapting the content depending on the context, user preferences and IM requests.

IV. CASE STUDY

In our cities, urban information services are provided in ways which have not changed much in a century: bus-stop shelters, metro-stations panels and screens, maps, and urban furniture. This scenario brings up the opportunity to improve the information services that a city provides to people living

[7] See J. A. Larson, "Standard Languages for Developing Multimodal Applications," [http://www.larson-tech.com/Writings /multimodal.pdf] last visited on February 2010.

in or visiting the city not only making it dynamic but also with the novel possibility of providing ubiquitous, accessible and multimodal access to content. In this scenario the actors are people living in or visiting the city who have some functional diversity or do not know the local language, and need a specific interaction with the local information system, represented by an ICD. The aim is to offer hyper-local media adapted to the user's needs.

For example, if an American visitor is in front of an ICD located in Barcelona's city center, he can identify himself with his American passport (using RFID). Because of the profile John made public on some integrated system or social network, the ICD system will know that John have some aural functional diversity and can only understand Spanish sign language (SSL), so it will adapt all content using an avatar which "speaks" SSL and will activate the ICD's webcam to record John's gestures and let him navigate through the user interface.

Another visitor, a Japanese old woman who speaks neither Catalan nor Spanish, uses her tourist card (issued at the airport or at her hotel, for example) to get identified by the ICD, which then adapts all content using a third-company Japanese translation service, and also increases the size of the text characters in real-time.

Another person, living in the city, is blind and wants to interact with the ICD using the voice. All the content would then be synthesized using a Catalan voice from an external service. He also can interact within the ICD interface touching the corners of the touch screen to navigate through the content.

In most scenarios, the user gets identified by the ICD through some identification card (with RFID or NFC). In these cases the IM creates a new task at the TM for the identification of the user and adapts the content and the interaction to the user (after TM sent the user's profile to the IM). Then, the IM creates an instance of the user model and adapts the modal channel and the content synthesis accordingly. After showing the content adapted to user necessities (read, using an avatar with SSL or just shown in the touch screen) the IM will capture user interaction and activate a suitable service to recognize its meaning; finally it will update the user's state. In any of the scenarios the user could change the interaction mode and update her profile through the ICD.

V. FUTURE WORK AND CONCLUSIONS

This paper presented a multimodal architecture based on a distributed architecture paradigm using third party services and non-local processes. The main goal was to create a platform able to provide multimodal interactions for everybody but especially for people with functional diversity to better access several services through a display in an urban context.

The ICD platform presented in this paper which is based on real time applications to interact with the users over a service-oriented architecture has to improve in some particular features. The main inconvenience with this is in that the consumed services are time-wise very expensive compared with a real time interaction between people. The waiting time between interactions would be shorter if all processes (recognizers and synthesizers modules) were local rather than remotely distributed. Nevertheless, improving wireless telecommunications and the processors capabilities will normalize the response time in interactions between user and machine [12].

One of the planned future developments of the platform presented in this paper is the deployment of the services and the adaptive multimodal interfaces methods on user mobile devices. However, the problem with current mobile devices and all software developed for mobile platforms is that applications are not interoperable among the platforms that exist nowadays (iPhone, Android, RIM, Windows Mobile, among others). Web technology would solve this problem, and progress is being made in this sense, but there is no framework yet that allows access from a Web browser to the device hardware of all the mobile operating systems.

Another task identified into the multimodal research is the determination of the user's interaction duration in order to capture the complete interaction or begin the processing without losing the context at the end of the interaction which is a human-computer interaction problem.

A fundamental measure of progress in computing involves rendering it as a way to improve our everyday experience while simultaneously making it disappear. Radical improvements in microprocessor cost-performance ratios have pushed this process forward, enabling us to embed computers in many parts of our environments. In 10 years this change would transform the early bus-stop shelters, metro-stations panels and screens, maps, and urban furniture into devices that can potentially enable, mediate, support and organize our daily activities.

ACKNOWLEDGMENTS

Marc Pous and Luigi Ceccaroni carried out the most part of the research related to this paper at TMT Factory. The research described in this paper arises from a Spanish research project called INREDIS (Interfaces for the relation between the environment and people with functional diversity), which is funded by CDTI, under the CENIT program, in the framework of the Spanish government's INGENIO 2010 initiative. The opinions expressed in this paper are those of the authors and are not necessarily those of INREDIS project's partners or CDTI.

REFERENCES

[1] J. Abascal, B. Boanil, L. Gardeazabal, A. Lafuente, and Z. Salvador, "Managing Intelligent Services for People with Disabilities and Elderly People," Lecture Notes in Computer Science, vol. 5615/2009, pp. 623-630, July 2009.

[2] L. Ceccaroni, V. Codina, M. Palau, and M. Pous, "PaTac: Urban, Ubiquitous, Personalized Services for Citizens and Tourists," in The Third International Conference on Digital Society (ICDS 2009), Y. Takahashi, L. Berntzen, and A. Smedberg, Eds. USA: IEEE Computer Society, 2009, pp. 7-12.

[3] L. Deng, K. Wang, A. Acero, H. Hon, J. Droppo, C. Boulis, Y. Wang, D. Jacoby, M. Mahajan, C. Chelba, and X. D. Huang, "Distributed Speech Processing in MiPad's Multimodal User Interface," IEEE

Transactions on Speech and Audio Processing, vol. 10(8), pp. 605-619, November 2002. [http://research.microsoft.com/pubs/75215/2002-deng-trans2.pdf]

[4] L. Deng., Y. Wang, K. Wang, A. Acero, H. Hon, J. Droppo, C. Boulis, M. Mahajan, and X. D. Huang, "Speech and language processing for multimodal human-computer interaction," Journal of VLSI Signal Processing Systems, vol. 36(2-3), pp. 161-187, 2004. [http://research.microsoft.com/pubs/75372/2004-deng-vlsi.pdf]

[5] A. Greenfield, Everyware: The dawning age of ubiquitous computing. New Riders Publishing, 2006.

[6] X. Huang, A. Acero, C. Chelba, L. Deng, J. Droppo, D. Duchene, J. Goodman, H. Hon, D. Jacoby, L. Jiang, R. Loynd, M. Mahajan, P. Mau, S. Meredith, S. Mughal, S. Neto, M. Plumpe, K. Stery, G. Venolia, K. Wang, and Y. Wang, "MIPAD: A Multimodal Interactive Prototype," in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, Utah, USA: Institute of Electrical and Electronics Engineers, Inc., 2001. [http://research.microsoft.com/pubs/75373/2001-huang-icassp.pdf]

[7] Y. Jung, and A. Anttila, "How to look beyond what users say that they want," in CHI '07 extended abstracts on Human factors in computing systems, M. B. Rosson, Ed. New York, NY, USA: ACM, 2007, pp. 1759-1764.

[8] A. Kobsa, J. Koenemann, and W. Pohl, "Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships," The Knowledge Engineering Review, vol. 16(2), pp. 111-155, 2001.

[9] D. Küpper, and A. Kobsa, "Tailoring the Presentation of Plans to Users' Knowledge and Capabilities," Lecture Notes in Computer Science, vol. 2821/2003, pp. 606-617, September 2003.

[10] K. Lyytinen, and Y. Yoo, "Issues and challenges in ubiquitous computing," Communications of the ACM, vol. 45(12), pp. 63-65, December 2002.

[11] N. Negroponte, "Talking to Computers: Time for a New Perspective," Wired, vol. 2(2), February 1994. [http://web.media.mit.edu/~nicholas/Wired/WIRED2-02.html]

[12] L., Nigay, and J. Coutaz, "A design space for multimodal systems: concurrent processing and data fusion," in Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems, B. Arnold, G. van der Veer, and T. White, Eds. New York, NY, USA: ACM, 1993, pp. 172-178.

[13] S. Oviatt, A. DeAngeli, and K. Kuhn, "Integration and synchronization of input modes during multimodal human-computer interaction," in Proceedings of the SIGCHI conference on Human factors in computing systems, S. Pemberton, Ed. New York, NY, USA: ACM, 1997, pp. 415-422.

[14] J. Vázquez-Salceda, L. Ceccaroni, F. Dignum, W. Vasconcelos, J. Padget, S. Clarke, P. Sergeant, and K. Nieuwenhuis, "Combining Organisational and Coordination Theory with Model Driven Approaches to develop Dynamic, Flexible, Distributed Business Systems," Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 21, pp. 175-184, 2010.

[15] M. Weiser "The computer for the 21st century," in Scientific American, Sep. 1991, pp. 94-104.