A framework to support multiple levels of interaction

Dario Di Mauro
dario.dimauro@unina.it

Department of Information Technology and Electrical Engineering
PhD in Information Technology and Electrical Engineering
Coordinator: prof. Daniele Riccio
Tutor: prof. Francesco Cutugno
University of Naples "Federico II"

Abstract. The human world is more and more inter-connected and, as consequence, we need to communicate in a faster and smarted way. Many software applications do that, but they usually work in a single context at time. In this paper we present a framework that aims at making the interaction easier and more natural; involved actors are connected in a graph, sending and receiving signals by adjacent nodes. The framework has been developed to work in different contexts, especially in domotic and cultural heritage fruition; the goal is to connect different spaces, such as homes, museums, etc, focusing the attention on interaction perspective, maintaining a natural interface for the user and hiding low-level communication issues. This work will be framed in the context of Natural User Interfaces, so it is centered on the user, exchanging signals with external sensors or devices, such as a smartphone and PC, using gestures, dialogues and augmented reality.

1 Introduction

The human world is more and more inter-connected and, as a consequence, we need to communicate in a faster and smarter way. We want to talk with friends, require information, enjoy a work of art, interact with appliances and live real life. Many softwares and apps do that, but can we develop a general interaction system, composing different communications channels in a single environment? Could we create a mixed reality in which the perceived world is extended and fused with a virtual and interactive one, as more naturally and transparently as possible?

Progresses in HCI have changed the type of interaction between humans and computers. Advances in HCI make the interaction faster and simpler, but a lot of work still needs to be done. With the introduction of smart sensors, low-cost hardware and powerful smartphones, smart-homes were born, interactive museums are going to grow, human-robot interaction is going to be more complex but all of them are treated as different spaces. The present is this, the future could be the connection among these worlds. The cited contexts are clearly

Proc. of CHItaly 2015 Doctoral Consortium, Rome (Italy), September 28th 2015 (published at http://ceur-ws.org).

different, but they could share some interaction ingredients: the exchange of signals and information, for example, or the need to present enough data to handle a communication between humans and machines, without loading too much the interface.

The goal of this PhD thesis/project is to develop a framework that makes easier the communication among cited sensors and actors, managing in the same graph of connections all the parts and activating the essential nodes in the right moment. As you will read in the next sections, the theoretical framework in which this goal will be framed is that of Natural User Interfaces. Due to the recent beginning of this PhD, both the general concept and single modules are in phase of definition. Some parts of this paper, therefore, could require further arrangements in the next future.

This paper is organized as follows: section 2 summarizes the starting point and related works; section 3 explains the proposed framework. Section 4 reports preliminary results about technologies that will be included in the framework; section 6 concludes the work.

2 Related Works

This work takes ideas from many previously published contributions: a similar framework is presented in [18]. That system synchronizes and records different signals in order to recognize human emotions; considered sensors are a microphone, a webcam and a wiimote. The proposed work starts from this concept relying on the synchronization module but, instead of just analyze the signals, it processes them to produce an output, introducing an interaction with other systems. The motivation about this work is in the emerging interests in social signal processing, a new cross-disciplinary research domain that aims at understanding and modeling social interactions and providing computer with similar abilities [17][15].

By interacting with a user, a part of the system will be devoted to analyze implicit feedback from her. This approach is useful helping to do a recommendation [14] or to better understand a user's preference, without asking her an explicit question [12].

Due to the main context of the work, a large part of the work will be devoted to an high-level interaction, based on gestures, augmented reality, dialog systems and 3D audio; this choice aims at making the interaction natural and effective. The cited channels are not new in literature: gesture-based interaction is extensively adopted [13]; regarding 3D audio, a large number of systems has been presented [11][5][8]; usually they limit themself to a 3D audio production but a more interactive and social approach misses. The state of the art about dialog systems could be represented by OpenDial [9]. This system will be used as dialog engine.

Proc. of CHItaly 2015 Doctoral Consortium, Rome (Italy), September 28th 2015 (published at http://ceur-ws.org).

3 Proposed Approach

As announced in section 1, this PhD project aims at developing a framework that simplifies the communication among different agents. The work is based on the concept that the same framework can be used in different scenarios, so the system abstracts from the figure of agents: they can be humans, robots, smart-devices or works of art, for example. They need to interact each other: not-human agents need to communicate by packets; if at least an agent is a human, instead, a different channel is needed. Regarding the presentation of data, a remarkable attention has been oriented on the naturalness of the interaction; in some cases, for example, an interaction based on voices, gestures and expressions can be more effective than a text-based one. In the work a modular architecture has been designed, separating the communication and synchronization modules from the dialog manager and the presentation layers.

The work is in the context of the Natural User Interfaces (NUI) [19], so it aims at making an interaction as more natural as possible, always centered on the user. The use of the proposed framework is oriented to a non-expert user that need an high-level interaction; the system hides low-level issues to cited user, focusing herself on interaction perspective.

In a connected world, a wide range of sensors should be distributed in the environment. Nowadays, a lot of high-level sensors are low-cost, such as Microsoft Kinect. A single kinect manages a set of users, tracking their skeletons and joints, but in a limited field of view. An array of kinects covers a large part of space, as a room, giving the possibility to track users and moving objects in a wider range and with different orientations. Other kinect-like sensor could be used, but they have not been considered yet. The introduced framework manages different sensors, synchronizing them and combining the flows in a single engine. A recognition module processes this input and extracts gestures information, communicating them to other agents. The same system produces an output, with a bidirectional interaction.

3.1 The framework engine

The core of the system manages a graph in which the nodes are the agents and the arches are direct connections among the agents. Each node has data, requires signals or produces them. An agent communicates with adjacent nodes; it receives inputs from them and produces an output recurring to local abilities or information. As explained above, NUI is the general frame of the work, so a vocal dialog could be a chosen type of interaction; for example, a node receives a vocal input, uses local structures to recognize it and to prepare an utterance or other possible outputs, such as expressions, and recurs to Text-To-Speech (TTS) to talk with others. In a more complex scenario, the framework could support a natural interaction between a human and a virtual assistant and, at the same time, a communication between the assistant and other sensors or actuators.

Because of the nature of interaction, graph connections could change, so flow optimization problem will raise. Given the nature of this framework, different

Proc. of CHItaly 2015 Doctoral Consortium, Rome (Italy), September 28th 2015 (published at http://ceur-ws.org).

signals should be managed and analyzed together. The system includes a multimodal processing block that synchronizes, links and makes readable signals to selected agents.

3.2 Interaction Analysis

As a control system requires to analyze the effects of the output on the world, an interaction framework may require to study the user's feedback in order to obtain a more effective interaction. It is possible, in a fused engine, to track a user's actions in different contexts; these information are very useful for a widespread analysis. The presented framework is born to be applied in different settings trying to connect them, so an extended data analysis system can be provided; type of data are user's position, choices, preferences or personal information. The work aims at providing an integration with external modules, devoted to the analysis of listed signals. An example of behaviour analysis is navigation in indoor environment as offices [10] or museums [2].

In the framework, an integration with the analysis of routines, such as [7], or an user's profiling [16] or an implicit feedback analysis module [14] will be considered. Different solution will be proposed in heterogeneous scenarios, depending on types of signals and interaction: cultural heritage fruition and domotic will be sample contexts. Due to the similarity of coped issues with [3], *domain developers* and *maieuta designer* will be involved in the system creation, especially for the scenario of cultural heritage fruition.

3.3 Natural Interfaces

As explained, the framework will work on different levels of abstraction. Regarding high-level interaction, the system will support gesture recognition, augmented reality or dialog modules; these modules could be used to interact with other agents. Figure 1 reports an instance of the proposed idea: in the example, two distinct environments, home and museum, that are going to be connected. User 1 usually interacts with a Coffee machine, a Smart TV and sometimes goes to the museum. With the use of the presented framework, User 1 could use a smart audio-guide at the museum - details about it are later in this section. She expresses preferences about some pieces of art and a theoretical profiling system shares with the user a sketch about her. Once she is at home, the Smart TV could use the obtained "cultural profile" and possibly recommend cultural channels; the user could change the recommendation interacting by gestures. The same user, every morning, takes a coffee at 8:00. The smart coffee machine could learn this routine and power-on itself at 7:30, optimizing benefits and energy consumption for example; considering the coffee stocks, the machine could estimate when other coffee needs to be bought.

In the proposed example, the user has been interacted with an audio-guide, a smart TV and a coffee machine; many information has been moved, but the framework, working behind all the agents, has hidden this fact. Some of them has been developed in other projects, such as audio augmented reality based on

Proc. of CHItaly 2015 Doctoral Consortium, Rome (Italy), September 28th 2015 (published at http://ceur-ws.org).

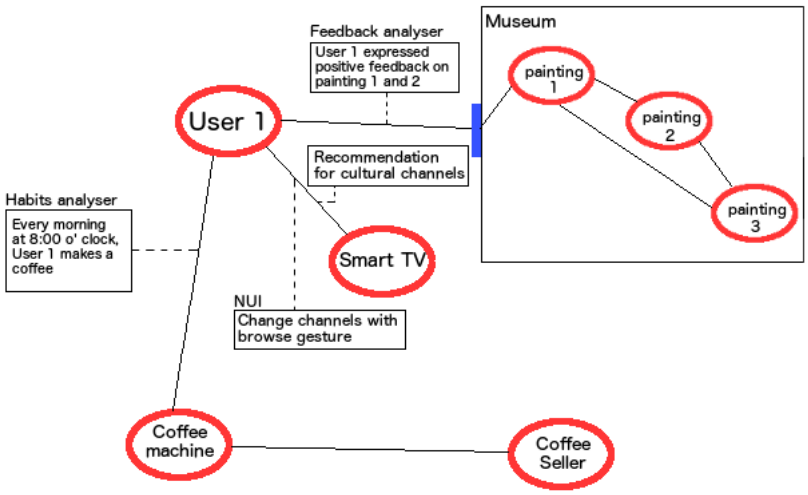

Fig. 1. An example of instance of the framework

3D audio. For privacy reasons, the user can always step in the actions and she has a notification as sensible data are transmitted.

3D Audio Interaction Sound spatiality carries more information than simple stereo audio [1]; by simulating the human hearing, an interaction based on 3D Audio has been tested [4]. Audio-augmented reality is a method to enrich the real world with virtual sounds in a given context and, setting aside the visual interaction, we limit as more as possible obstacles between listener and real life.

The system is an interactive audio-guide developed as an Android app called Caruso. As embryonic phase, Caruso works in a limited area in the historic center of Naples, giving information about historical buildings and churches of that area. As the user is close to a Point Of Interest (POI), a virtual soundscape with a 3D sound composition starts. The stage is dynamic: Caruso follows the user's movements, changing sound's direction basing on listener's position and orientation.

In order to make the interaction as more natural as possible, smart headphones has been designed. These headphones are equipped with dedicated circuits and detects the orientation of the head in the space. A Bluetooth module communicates the orientation to the smartphone; the 3D scene is updated as the angles change. Figure 2 reports a test of Caruso during a test session for the Or.C.He.S.T.R.A. project (`www.orchestrasmartnapoli.it`).

Proc. of CHItaly 2015 Doctoral Consortium, Rome (Italy), September 28th 2015 (published at http://ceur-ws.org).

Fig. 2. A test of Caruso

4 Preliminary Results

This PhD started in November 2014 and, at now, the framework proposed in section 3 is in design phase. No wide experiments has been conducted; preliminary results about 3D sound interaction has been collected. The 3D Audio Augmented Interaction has been tested in a realistic context, with an interactive audio-guide with 3D sounds [6].

An interactive audio-guide: As presented in section 3.3, Caruso has been used as a audio-guide in the historical center of Naples in the experimental process of the Or.C.He.S.T.R.A. project. Figure 3 shows the proposed POIs.

Fig. 3. The POI of Caruso

Proc. of CHItaly 2015 Doctoral Consortium, Rome (Italy), September 28th 2015 (published at http://ceur-ws.org).

The test consists of 8 POIs, composing a path in the historic center of Naples; the proposed path is usually routed by tourists and citizens. Each POI proposes a soundscape in 3D sound, talking about anecdotes and historic personalities. The estimated duration of the visit is about 20 minutes: it includes the duration of each scene and the distance between POIs.

All the scenes have been prepared and recorded by experts people. 32 participants attended the test. They borrowed smartphone and smart headphones and, with poor information about the goal of the test, have gone around. As they pass close to a POI, the soundscape starts automatically, by detecting Bluetooth Low Energy antennas, known as beacons. 16 testers used a version of the app without any global information about the POI distribution. For the other 16, we added the map in figure 3. A blue point showed the user's position.

Just a very small part of the attendees discovered all the POIs, about 10%. Motivations about that are searchable in:

- the participants has no limits of time in the use of Caruso and a large part of them just wanted to try a new technology;
- the same people contributed to other tests, so they did not dedicate all the free time to Caruso;
- the goal of the experimentation was not to find all POIs, but to receive a feedback about the proposed interaction system.

Everybody except one person noted the advantages of 3D sounds in the cultural heritage fruition; the use of this channel of interaction and the dynamism of the scene offer a very involving experience. The use of smart headphones improves the transparency of the interface: the listener takes part of the soundscape as an actress, virtually occupying a place in the listened scene and without the need to insert the code of a POI; this is a very different situation compared with traditional audio-guide, in which the sound is static and the user must detract the attention from the cultural context to start a new one. Eventually the map has been very useful for tourists; it offers a global reference in the environment, giving the possibility to plan paths to the Points Of Interest.

5 Acknowledgment

The 3D Audio Augmented Interaction has been funded by the European Community and the Italian Ministry of University and Research and EU as part of the PON Or.C.He.S.T.R.A. project.

6 Conclusions

This paper presented a framework that aims at making the interaction smarter, simplifying it among more agents, in different contexts, mainly focusing the attention on the human perspective in the communication. The framework is designed in multiple layers managing low-level signals, dialogues and high-level

interaction, based on voices, gestures and augmented reality. The presented framework is in a design phase, but separate modules already exists: the audio-augmented reality based on 3D sounds has been applied in different contexts, a suitable gesture recognition module is in developing. Coming work will be devoted to the framework itself, choosing different case studies in cited scenarios: domotic and cultural heritage fruition; other contexts will be considered.

References

1. Armando Barreto, Kenneth John Faller, and Malek Adjouadi. 3d sound for human-computer interaction: regions with different limitations in elevation localization. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, pages 211–212. ACM, 2009.
2. Alessandro Bollo and Luca Dal Pozzolo. Analysis of visitor behaviour inside the museum: An empirical study. In *Proceedings of the 8th International Conference on Arts and Cultural Management, Montreal*, volume 2, 2005.
3. Federico Cabitza, Daniela Fogli, and Antonio Piccinno. Fostering participation and co-evolution in sentient multimedia systems. *Journal of Visual Languages & Computing*, 25(6):684 – 694, 2014. Distributed Multimedia Systems {DMS2014} Part I.
4. Daniela D'Auria, Dario Di Mauro, Davide Maria Calandra, and Francesco Cutugno. Caruso: Interactive headphones for a dynamic 3d audio application in the cultural heritage context. In *Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on*, pages 525–528. IEEE, 2014.
5. William Dell. The use of 3d audio to improve auditory cues in aircraft. *Department of Computing Science, University of Glasgow*, 2000.
6. Dario Di Mauro and Francesco Cutugno. Sca3d: a multimodal system for hci based on 3d audio and augmented reality.
7. Katayoun Farrahi and Daniel Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):3, 2011.
8. Florian Heller, Thomas Knott, Malte Weiss, and Jan Borchers. Multi-user interaction in virtual audio spaces. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 4489–4494. ACM, 2009.
9. http://www.opendial-toolkit.net/. Opendial.
10. Tomoya Ishikawa, Masakatsu Kourogi, Takashi Okuma, and Takeshi Kurata. Economic and synergistic pedestrian tracking system for indoor environments. In *Soft Computing and Pattern Recognition, 2009. SOCPAR'09. International Conference of*, pages 522–527. IEEE, 2009.
11. Jacque M Joffrion. head tracking for 3d audio using a gps-aided mems imu. Technical report, DTIC Document, 2005.
12. Diane Kelly and Jaime Teevan. Implicit feedback for inferring user preference: a bibliography. In *ACM SIGIR Forum*, volume 37, pages 18–28. ACM, 2003.
13. Jani Mäntyjärvi, Juha Kela, Panu Korpipää, and Sanna Kallio. Enabling fast and effortless customisation in accelerometer based gesture interaction. In *Proceedings of the 3rd international conference on Mobile and ubiquitous multimedia*, pages 25–31. ACM, 2004.
14. Douglas W Oard, Jinmook Kim, et al. Implicit feedback for recommender systems. In *Proceedings of the AAAI workshop on recommender systems*, pages 81–83, 1998.

15. Maja Pantic and Alessandro Vinciarelli. Social signal processing. *The Oxford Handbook of Affective Computing*, page 84, 2014.
16. Silvia Rossi, Francesco Barile, and Antonio Caso. User and group profiling in touristic web portals through social networks analysis. In *Proceedings of WEBIST 2015 - 11th International Conference on Web Information Systems and Technologies*, pages 455–465, 2015.
17. Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.
18. Johannes Wagner, Florian Lingenfelser, and Elisabeth André. The social signal interpretation framework (ssi) for real time signal processing and recognition. In *INTERSPEECH*, pages 3245–3248, 2011.
19. Daniel Wigdor and Dennis Wixon. *Brave NUI world: designing natural user interfaces for touch and gesture*. Elsevier, 2011.