A 3D Audio Augmented Reality System for a Cultural Heritage Management and Fruition

Daniela D'Auria[1,2], Dario Di Mauro[1], Davide Maria Calandra[2], Francesco Cutugno[1]
[1]Department of Electrical Engineering and Information Technology
University of Naples Federico II
Via Claudio, 21
[2]Department of Physics
University of Naples Federico II
Via Cinthia SNC
Naples, Italy
{daniela.dauria4, dario.dimauro, davidemaria.calandra, cutugno@unina.it}

Journal of Digital Information Management

ABSTRACT: One of the main features of the physical environment in which humans live is spatial dimensionality. When we think about 3D, we usually refer to 3D video, even if it is not the only existing channel of natural interaction. On the contrary, related work in the field of 3D audio mainly refers to the investigation of acoustic effects by adding reverberation in the context of virtual 3D environments. In this paper, we present an interaction system based on spatialized sounds. We developed an innovative cloud application in the cultural heritage context; a personal guide, in 3D sound, attracting the tourists' attention toward monuments or buildings, offering sound scapes of augmented reality. The designed system interacts with smart headphones that remotely takes the orientation of the listener's head and properly generates an audio output, which also takes into account the listener's position and orientation in the environment. Thus, such innovative headphones using an inertial measurement unit for determining the orientation of user's head have been designed and developed in open-ear mode, in *order to locate the user in the real context.*

Subject Categories and Descriptors
H.5.1 [Multimedia Information Systems]: Artificial, augmented, and virtual realities **I.3.7 [Three-Dimensional Graphics and Realism]:** Animation; **I.2.10 [Vision and Scene Understanding]; 3D/stereo scene analysis**

General Terms:
Augumented Reality System, 3 Dimensional Analysis

Keywords: Augumented reality, 3D Audio, Video Analysis

Received: 13 March 2015, Revised 20 April 2015, Accepted 27 April 2015

1. Introduction

Humans use hearing to detect significant sounds and to interact with the external environment. Moreover, hearing abilities include their capability to estimate the position of a sound source. Like vision, hearing is also three-dimensional. Listeners can hear sounds and easily determine their position both on sagittal and longitudinal planes (with the head position).

Audio-augmented reality is a method to augment the environment and objects in the real world with virtual sounds in a given context. Therefore, position and orientation of the user have to be tracked. The easiest way to personally provide audio to the user would be using headphones; and while one is using headphones, the idea of tracking orientation at the head is not far away. To do so, headphones can be equipped with specific positioning devices. While using headphones with the equipped hardware, the rotation of the head describes the orientation of the user's gaze. For this reason, audio systems have been developed in order to use headphones and to project 3D sound, in which the brain perceives the sounds as coming from a particular direction. Some applications of this technology already exist in both military and general aviation and where most installations were using complex and large hardware (Dell, 2000) (Joffrion, Delsing, Hamilton, Presnar, & Reed, 2004). On the contrary, this approach is very innovative within the cultural heritage context and

there is no use of complex and large hardware.

In order to be effective, 3D audio systems require real time knowledge of head orientation. Thus, this paper describes the development and testing of an integrated inertial measurement unit (IMU)/GPS system that determines real-time head orientation, by also exploiting a 3D audio system. The system incorporates a low cost micro electro mechanical system (MEMS) IMU combined with a single-frequency GPS receiver inside a smartphone. Real-time data from IMU system flow to a microcontroller implemented for determining roll, pitch and yaw. The system communicates with an Android app that receives the attitude information from IMU and implements a 3D sound and video interaction based on a reproduction of spatialized sound and video outputs. The whole scene is focused on the user and is updated to follow the movements.

One of the first field tested with audio-augmented reality as navigational aid was conducted by (Holland, Morse, & Gedenryd, 2002).Their prototype, called AudioGPS, is a spatial audio user interface. They analyzed various audio mappings to represent location and direction. All sounds are non-speech and non continuous because they want to avoid additional load on the human voice channel. They argue that speech sounds will place a large processing and attention burden on the user. To provide direction to the user, they use simple stereo panning to move an audio source around the users head. Their prototype does not use an electronic compass to get the direction of the user, therefore they have a latency of 10 to 15 seconds before the system starts reporting an update.

Another audio-augmented reality navigation application is the roaring navigator (Stahl, 2007), a group guide for a zoo with a shared auditory landmark display. An auditory display uses non-speech sounds to present information. This one is related to landmarks too, specifically to the sound of animals around the zoo. A magnetometer was clipped on the back of a baseball cap that users were wearing during the experiment to get the orientation.

(Heller, Knott, Weiss, & Borchers, 2009) built up Corona, an audio augmented reality experience deployed in the historic town hall of Aachen/Germany. In this study the visitor's position is determined by a Ubisense real time location system and a small compass mounted on the headphones communicates the visitor's head orientation to the mobile device. This information is then used by a spatial audio rendering engine to create a plausible audio experience. The audio engine of Corona is very similar to our system, but it misses of interaction, so it requires a constant contact with the phone.

Similar to the roaring navigator, (Vazquez-Alvarez, Oakley, & Brewster, 2012) built up a virtual sound garden placed in a park in Funchal, Madeira. They placed Earcons (non-verbal audio message which uses an abstract mapping to provide information to the user) at specific positions of landmarks of this park. A Nokia N95-8GB connected to a GPS receiver and a JAKE sensor pack was used to run the application and track the position and heading. 3D spatial audio rendering together with Earcons was the most effective technique as their results show.

(Ankolekar, Sandholm, & Yu, 2013) analyzed the performance and emotional engagement of different types of audio-based clues for directing users' attention like the Earcons mentioned above. Users were interrupted by audio clues while walking on a shopping street. The audio clues were played for a minute and the users had then to identify on a map which POI (Point Of Interest) was meant. Their results show that only one of the five different cues played, musicons (fragment of music that could be representative for that place), would be the better choice for serendipitous discovery, pleasure and identification accuracy.

Other very relevant works in the field of multimedia document analysis are (Albanese, d'Acierno, Moscato, Persia, & Picariello, 2013) and (Amato, Mazzeo, Moscato, & Picariello, 2009).

The paper is organized as in the following. Section 2 presents the design process of the used interactive headphones and the app. Section 3 briefly describes the context of use of our system and presents some preliminary experiments. Eventually, Section 4 discusses some conclusions and future work.

2. Design And Modeling of the Interactive Headphones

The overall system consists of different parts. We designed headphones which are composed by an inertial sensor and a bluetooth communication interface module interacting with an Android app developed on a smartphone as we can see in Figure 1

Figure 1. Whole system example

2.1 Hardware Design

We used a system that incorporates three sensors - an ITG-3200 (MEMS triple-axis gyro), ADXL345 (triple axis accelerometer), and HMC5883L (triple-axis magnetometer) - to measure the inertial parameters and to form an Inertial Measurement Unit (IMU) of 9 degrees of freedom. The outputs of all sensors are processed by an on-board microcontroller ATmega328 and sent over a serial interface; thus, this enables the 9DOF IMU to be used as a very powerful control mechanism. Moreover, programmed through the open source Arduino Software, this IMU outputs its roll, pitch, and yaw, data, characterizing the orientation of the head, via serial.

The 9DOF operates at 3.3VDC and for such a reason we used a LiPo batteries as an excellent power supply choice. Furthermore, a bluetooth module was used to establish a bidirectional communication channel enabling the wireless interaction between two endpoints (Figure 2). In this way, we can connect the IMU module and an Android device via serial stream whose communication is asynchronous.

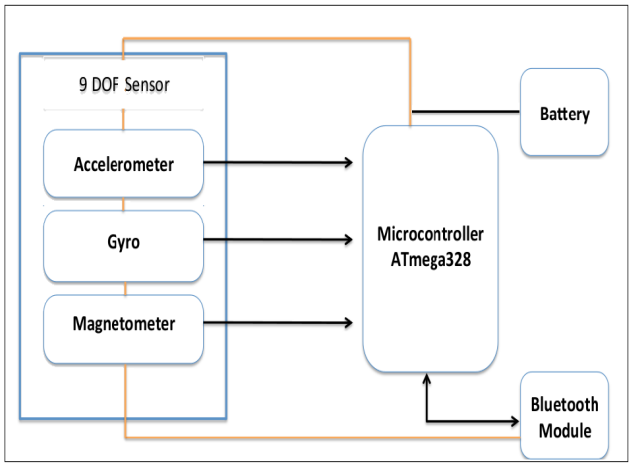

Figure 2. Hardware Schematic Description

2.2 The Headphones CAD Model

Solidworks, a well-known CAD software for 3D model, was used to create a new design of headset. These new headphones have a space that incorporate the electronic components used for calculating the orientation of the head and the bluetooth connection Figure 3.

Figure 3. Headphones CAD model

The CAD prototype of headphones including the hardware used is shown in Figure 4.

Figure 4. Prototype of headphones included the hardware used

The use of headphones is often criticized as it isolates the listener from the world, limiting the interaction with other people. In order to avoid it and to locate the user in the real context, we designed open-ear capsules, allowing external sounds to be reached by the listener and being able to play additional audio as well (Figure 4).

2.3 3D Audio Interaction

We have designed and developed an Android app, called CARUSO, for the augmented reality (AR) experience. The app works like an interactive personal guide, offering the descriptions about monuments and buildings around the user and reproducing virtual soundscape for a different cultural fruition. The app generates 3D audio output, in order to obtain a more realistic effect and more user's participation (Barreto, Faller, & Adjouadi, 2009). This paper presents an evolution of SCA3D (Di Mauro & Cutugno, 2013), (D'Auria, 2014) including a more suitable orientation detection, via smart headphones, and a more voicebased interaction approach.

We retain that developing an app for smartphone is a good choice, taking advantages of different layers: the device supports signal-processing, simplifies internet communication and position detection, offers stable APIs and is popular. The software system is organized in a client-server architecture. Android app, the client, requires a scene giving the phone position in order to get the monuments and sounds around him.The server part manages data about monuments, soundscape and generated and XML on the user request; the exported files follow MPEG-21 standard. A XML collects data about elements (the monuments) and soundscapes for each of them. A sound source is detailed in MPEG-7 standard.

Local scene is updated if current position is very far by the initial coordinates. Figure 5 shows a graphical schema.

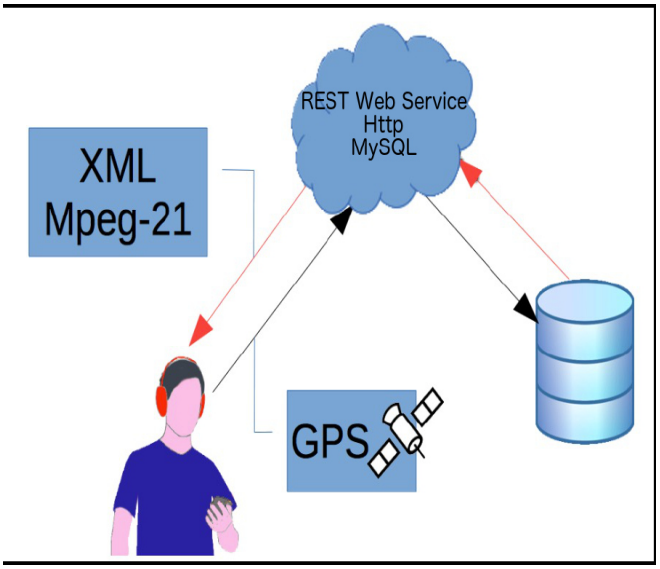

Figure 5. Schema of the client— server architecture

CARUSO takes the position of the smartphone to implement the scene, in order to establish relationship between the listener and the sound sources. All the connections are updated taking into account user's position and orientation. The audio output is managed by an open-source library, called OpenAL-soft (Creative), a software implementation of the OpenAL 3D audio API. It provides capabilities for playing audio in a virtual 3D environment and uses HRTF (head related transfer functions) to simulate 3D effect; the use of HRTF is the best way to get a realistic effect, but it requires a lot of CPU load: for such a reason, we need to limit the maximum number of sources in the scene. For a better result, we also reproduce the reverberation as close as possible to the real one by cited library.

We preferred to set aside the visual interaction, limiting as more as possible its use, because the art should not be lived through a display; CARUSO shows visual information just to trigger the interaction or to exhibit multimedia contents. Humans localize front sound sources also using the sight; the use of the audio channel only, requires a good design of the scene, in order to appreciate the maximum 3D effect. A good scene can influence the listener by implicitly helping him in estimation of the direction.

Position and Orientation

CARUSO exploits the GPS position in order to be able to work also outdoor. In indoor, instead, we tested some positioning system based on Wi-Fi signals (Huang, Millman, Quigley, Stavens, Thrun, & Aggarwal, 2011) or similar approaches (Ni, Liu, Lau, & Patil, 2004). An indoor scenario requires high precision, as the environment of interest is very limited; each indoor positioning system that we have evaluated was not really reliable; thus, we preferred to limit the use of the system to outdoor context. We got some potential results from the use of a preliminary approach based on Bluetooth Low Energy to localize the smartphone in a room or a corridor.

In order to obtain a more realistic effect, CARUSO works with smart headphones calculating the head orientation. Otherwise, in the older approach we used smartphone orientation, binding the user to move the device.

2.4. Dialog System and Data Managing

In order to reach a high level of natural interaction without diverting the user visual attention, we developed a dialog system based on a Finite State Automata and that is able to provide a widespread support for the user. Caruso recognizes voice commands in different areas and with a good level of abstraction: in the application context the user can start or stop a soundscape or can control the volume. Moreover, Caruso operates in a logistic context: the user can search a museum on a map, asking directions to it. She can ask the most important work of a building also. Finally, in the cultural heritage context, the user can query information about a piece of art. Asking, for example, "*Who is the author of this painting?*", the system downloads the required information about the playing element or about the nearest POI in the field of view.

The app stores as less data as possible, taking information as it needs: the user requires a scene or part of it, downloading just the elements in a range around her. In this version the sound files are on the smartphone, but a different approach can be adopted: a simple description, for example, can be stored as text and reproduced in spatialized sound, catching the Text- To-Speech output of the smartphone. By voice commands, Caruso makes HTTP requests to a REST web service devoted to store cultural information - following the Europeana standard - about the elements presented by the app.

3. A Case Study and the Related Preliminary Experimental Results

As mentioned in the introduction, in literature many systems use technology to provide new experiences. Most of them process 3D models in order to simulate the reconstruction of a building or a real life scenario. In this type of system, sight is the main component of interaction and it binds the user to live the real life through a display or other invasive devices.

In this paper we propose a different type of interaction: the user wears the headset and she is embedded in a virtual soundscape augmenting the reality. Our system works on a smartphone normally held into the pocket, so the user is free to move himself around. Moreover, this type of interaction is more natural than one presented in similar works, as it enhances the enjoyment of the art without hindering it. The system can reproduce single voices, calling the listener and attracting her attention; CARUSO can also simulate more complexes audio scenes, recreating a real life scenario or adding virtual audio elements to the classical description.

In order to get an analysis as more complete as possible

we conducted different types of tests. The experiments are composed of a technical part, with a data analysis and a measurement of the errors, and a "social" part, studying the interaction with the users and their reactions. In this section we report experiments and results.

3.1 Hardware Tests

As explained above, we designed headphones which are composed of an inertial measurement unit sensor and a bluetooth module. We must analyze each of these parts to measure the interference, the degree of precision, the load of the device and the user's satisfaction.

Each sensor presents some errors due to interferences with the environment noise; in order to calculate it, we compared the sensor data with those of similar devices. We took into account 3 other compasses; our reference value is a mean of them. Several tests in an indoor environment demonstrated that we got a 12° mismatch degree with respect to the real position: this is due to the other electromagnetic fields found into the laboratory. On the contrary, in outdoor environments we got a lower mismatch degree, as we found no conflicting electromagnetic fields; details in Table 1.

SENSOR PRECISION EVALUATION	TEST CONDITION	MISMATCH DEGREE
INDOOR ENVIRONMENT	Presence of EM fields	12±0.8
OUTDOOR ENVIRONMENT	Absence of EM fields	8±0.6

Table 1. Sensor precision evaluation

3.2 Software Tests

In order to know whether *i)* the users prefer 3D sounds to traditional stereo output and whether *ii)* the use of video channel helps users to localize sounds, we set two types of experiments.

3.2.1 Exploiting 3D Audio or Not

In the first test battery, the user hears a set of sounds in the headphones and must indicate the direction. He has been fronted to 3 different scenarios.

In the first, the user can listen many tracks of the same soundtrack; in the second, the user can listen nature environment sounds and in the last we play extraneous sounds for the defined soundscape. In this way, the user focuses herself less on the general content of the sound stage and tends to localize sound sources.

In this kind of test, users are sitting on a wheeled chair in a silent room with a Google Nexus 5 and our headphones.

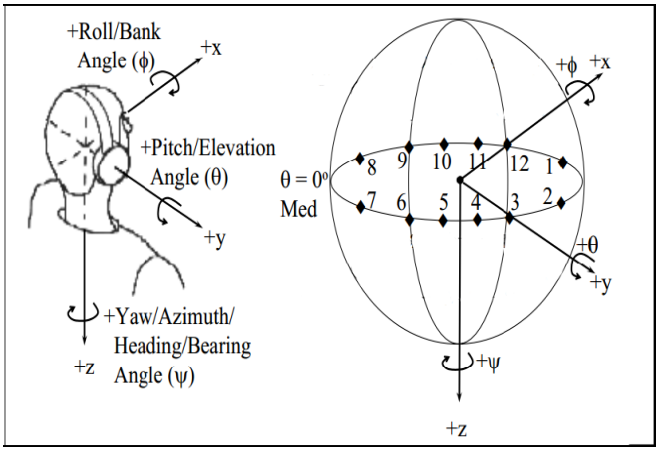

Figure 6. The scene of the test

Around the user there are numbered landmarks. The scene of the test is in Figure 6.

We ask the listener to localize a particular source. As sounds start, the user listens them with closed eyes. Within this tax, the user can help himself turning the chair or his head to rotate sound scene. As he is ready, he taps on the tablet and stops sounds. The user indicates the target corresponding to the direction of the sound. We record his response and the time elapsed from start. The test was attended by 28 users. In music test, two users have not been able to specify the direction of the requested source. In another case, a user was able to distinguish between right and left, but he was not able to add new information. We reproduced the same soundscapes at lower realism, with neither HRTF nor reverb, reaching simple stereo; as regards the type of interaction, all users have preferred 3D sounds to the classic stereo channel. In this section, we report the numerical details of the considered case study: Figure 7 reports the success rate, Figure 8 shows time of response.

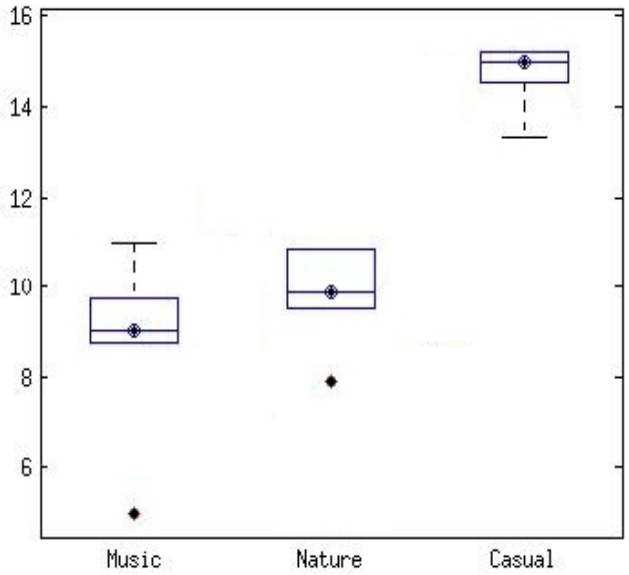

Figure 7. Success rate of the test

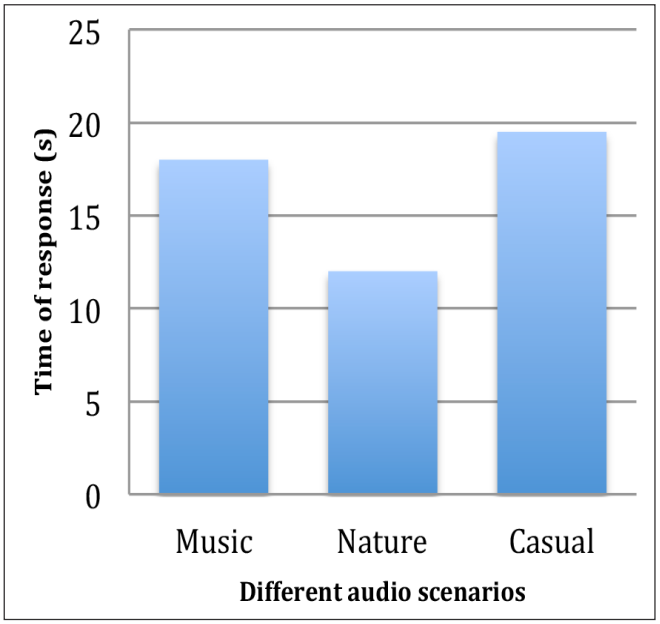

Figure 8. Mean time of response of the user to the request

3.2.2 Video Integration

In a second category of experiments, the user helps himself with the video channel. He points the device toward a target and sees information about it on the screen. The 3D sound comes from the target and video channel helps the user to locate the sound source; in this way 91% of the users reached the goal very soon. Actually, this is a realistic situation: we use multiple senses simultaneously to perform a task (Blattner & Dannenberg, 1992); human hearing, in fact, is more capable to catch sounds coming from behind, using also sight to locate front sound sources; as explained above, we want to avoid visual obstacles between humans and real life, so we design audio scenes preferring rear sounds, hidden sound sources or recommending to close eyes.

4. Conclusion

3D sound research is a newly developing field compared to the highly developed 3D visual research field. Newly conducted research in 3D sound might even bring great help to visually impaired people.

In this paper we showed how 3D-audio can be considered a very interesting interaction channel. Moreover, the choice of estimating the head orientation makes the interaction much more natural, as the user is not forced anymore to carry a smartphone, but he is consequently completely embedded into the new reality much more efficiently than in other existing systems.

Furthermore, the developed head orientation tracking system demonstrated that a low cost MEMS IMU can provide an audio-augmented reality extending the real world with virtual sounds.

Moreover, the audio coming from the virtual sound sources is altered depending on the users' position and orientation. On the other hand, in our work the immersion is quite strong, allowing the user to get the impression that the sounds were emitted from the real world.

This is not the first experiment combining 3D sound and AR (Sodnik, Tomazic, Grasset, Duenser, & Billinghurst, 2006), but we try to use this interaction system in a cultural scenario, that needs to be reinvented: furthermore, CARUSO works in a highly interactive context, recurring to voice commands to understand the user's needs and to dialog with him. In such a way, CARUSO recognizes voice commands; thus, it is possible to put the phone into the pocket. Future work will be devoted to further improve the vocal interaction and a personalization of the contents by a profiling based on social networks (Caso & Rossi, 2014). We will focus our attention on the social component of the app. The users could communicate in the same environment forming a graph: two nodes are adjacents taking into account the euclidean distance or a similarity degree between them. Each user could store a part of information to share with others in a high socially interaction experience.

5. Acknowledgment

This work has been funded by the European Community and the Italian Ministry of University and Research and EU under the PON OR.C.HE.S.T.R.A. project.

References

[1] Vazquez-Alvarez, Y., Oakley, I., Brewster, S. A. (2012). Auditory display design for exploration in mobile audio-augmented reality. *Personal and Ubiquitous computing* , 987-999.

[2] Albanese, M., d'Acierno, A., Moscato, V., Persia, F., Picariello, A. (2013). A multimedia recommender system. *Transactions on Internet Technology (TOIT)* , 3.

[3] Ankolekar, A., Sandholm, T., Yu, L. (2013). Play it by ear: a case for serendipitous discovery of places with musicons. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* , 2959-2968.

[4] Amato, F., Mazzeo, A., Moscato, V., Picariello, A. (2009). A system for semantic retrieval and long-term preservation of multimedia documents in the egovernment domain. *International Journal of Web and Grid Services* , 323-338.

[5] Barreto, A., Faller, K. J., Adjouadi, M. (2009). 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* , 211-212.

[6] Blattner, M. M., Dannenberg, R. B. (1992). Multimedia interface design.

[7] Caso, A., Rossi, S. (2014). Users ranking in online social networks to support pois selection in small groups. *Extended Proceedings of the 22nd Conference on User*

Modelling, Adaptation and Peronalization- UMAP , 5-8.

[8] Creative, L. (s.d.). Tratto il giorno January 10, 2015 dahttp://kcat.strangesoft.net/openal.html

[9] Dell, W. (2000). The use of 3D audio to improve auditory cues in aircraft. *Department of Computing Science, University of Glasgow* .

[9] Di Mauro, D., Cutugno, F. (2013). Sca3D: a Multimodal System for HCI Based on 3D Audio and Augmented Reality. *it AIS 2013 Proceedings X Conference of the Italian Chapter of AIS Empowering society through digital innovations* .

[10] D'Auria, D., et al., Di Mauro, D., Calandra, D., Cutugno F. Caruso: *Interactive headphones for a dynamic 3D audio application in the cultural heritage context. IRI* 2014: 525-528

[11] Huang, J., Millman, D., Quigley, M., Stavens, D., Thrun, S., Aggarwal, A. (2011). Efficient, generalized indoor wifi graphslam. *Robotics and Automation (ICRA)* , 1038-1043.

[12] Heller, F., Knott, T., Weiss, M., Borchers, J. (2009). Multi-user interaction in virtual audio spaces. *CHI'09Extended Abstracts on Human Factors in Computing Systems* , 4489-4494.

[13] Holland, S., Morse, D. R., Gedenryd, H. (2002). AudioGPS: Spatial audio navigation with a minimal attention interface. *Personal and Ubiquitous Computing*, 253-259.

[14] Joffrion, J. M., Delsing, F. A., Hamilton, C. E., Presnar, M. D., Reed, S. M. (2004). Inertial Head Tracking for 3D Audio (Project "Sound Advice"). *AIR FORCE FLIGHT TEST CENTER EDWARDS AFB CA.*

[15] Ni, L. M., Liu, Y., Lau, Y. C., Patil, A. P. (2004). LANDMARC: indoor location sensing using active RFID.*Wireless networks* , 701-710.

[16] Sodnik, J., Tomazic, S., Grasset, R., Duenser, A., Billinghurst, M. (2006). Spatial sound localization in an augmented reality environment. *Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments* , 111- 118.

[17] Stahl, C. (2007). The roaring navigator: a group guide for the zoo with shared auditory landmark display. *Proceedings of the 9th international conference on Human computer interaction with mobile devices and services* , 383-386.