TECNIACUSTICA MURCIA · 2014

45º CONGRESO ESPAÑOL DE ACÚSTICA
8º CONGRESO IBÉRICO DE ACÚSTICA
EUROPEAN SYMPOSIUM ON SMART CITIES AND ENVIRONMENTAL ACOUSTICS

ACOUSTICAVE: AURALISATION MODELS AND APPLICATIONS IN VIRTUAL REALITY ENVIRONMENTS

PACS: 43.55-Ka

Guilherme Campos[1,2], Paulo Dias[1,2], José Vieira[1,2], Jorge Santos[3,4,5], Catarina Mendonça[6], João Pedro Lamas[4,5], Nuno Silva[4,5], Sérgio Lopes[7]
[1]Departmento de Electrónica, Telecomunicações e Informática (DETI), Universidade de Aveiro
Campus Universitário de Santiago, 3810-193 AVEIRO – Portugal
Tel: +351 234 370 355 Fax: +351 234 378 157
[2] Instituto de Engenharia Electrónica e Informática de Aveiro (IEETA), Portugal
[3] Departamento de Psicologia Básica – CIPsi, Braga, Universidade do Minho, Portugal
[4] Centro Algoritmi, Guimarães, Universidade do Minho, Portugal
[5] Laboratório de Visualização e Percepção (LVP), Centro de Computação Gráfica (CCG), Guimarães, Universidade do Minho, Portugal
[6] Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
[7] Instituto de Telecomunicações (IT), Aveiro, Portugal
{guilherme.campos, paulo.dias, jnvieira}@ua.pt, *jorge.a.santos@psi.uminho.pt*
{mendonca.catarina, lamas.jp, nunomiguel06, sergioilopes}@gmail.com

ABSTRACT

This communication is an overview of the FCT-funded three-year research project '*AcousticAVE* – Auralisation Models and Applications in Virtual Reality Environments', a collaboration between the Universities of Aveiro (UA - IEETA) and Minho (UM - CIPsi, LVP). The project involved the development of auralisation software based on the image-source method accommodating dynamic scenarios, with real-time tracking of source/listener motion and listener head orientation. This software supported psychophysical research at the CAVE-like facilities of UM's Visualisation and Perception Lab (LVP). This included an investigation on learning effects in spatial audio perception using non-individualised HRTF sets and distance and time-to-passage (TTP) perception experiments.

1 INTRODUCTION

The importance of Virtual Reality (VR) has grown rapidly in recent years, with an ever-increasing range of applications in the most diverse areas. However, most efforts in the design and development of VR systems have been directed at providing visual immersion. The development of increasingly convincing models demands that other senses also be considered, especially hearing – we are still very much in the 'silent era' of VR.

A joint initiative of the Universities of Aveiro (UA) and Minho (UM), the research project '*AcousticAVE* – Auralisation Models and Applications in Virtual Reality Environments' aimed precisely at integrating visual and aural immersion in VR environments. It received funding from the FCT (Portuguese Foundation for Science and Technology) for a three-year period which ended in April 2014 (PTDC/EEA-ELC/112137/2009). Prof. Damian Murphy, from the U. of York (UK) and, at a later stage, Prof. Juan Miguel Navarro, from UCAM (Spain), acted as external project consultants.

The project involved essentially two work packages. At UA, an engineering team from IEETA (Institute of Electronics and Informatics Engineering) tackled the software and hardware implementation of room acoustic simulation and auralisation models.

At the Visualisation and Perception Lab (LVP) of UM's Graphics Computing Centre (CCG) in Guimarães, a team of Psychophysics researchers from CIPsi (Psychology Research Centre)

were primarily concerned with the application side of the project, exploring and testing the developed auralisation systems and software packages at their CAVE-like facilities. These include a 3m-by-9m continuous projection screen comprising 3 panels with a DLP (*Digital Light Processing*) projector per panel for flexible configuration (0°, 90° or 135°), a treadmill synchronised with the 3D visual scene to allow walking on the virtual environment and an infra-red motion capture system for user tracking. It was decided that sound presentation should be binaural, in view of the technical requirements (room acoustic correction, equipment noise control…) and costs of multi-speaker alternatives such as *Ambisonics* or Wave-Field Synthesis (*WFS*).

A 'customer-supplier' relationship was established between the two teams, inasmuch as the model development work at UA aimed at meeting the requirements of the VR experiments designed by the UM team. The main research focus was on the role of aural cues in the perception of human motion, commonly referred to as *Biomotion* (see Figure 1). Biomotion perception experiments imply dynamic scenarios; the models must provide control over virtual source movement and respond to changes in listener head orientation and position, tracked in real-time. Also, since perception cues are obviously not only aural but also visual, it must be possible to integrate the two; in fact, their interdependence and sensitivity to synchrony are points of particular research interest. For these reasons, controlling the relative latency of the two channels (visual and aural) is also crucial. These two requirements (real-time tracking and accurate audio-visual synchronisation) posed the most significant engineering and signal processing challenges.

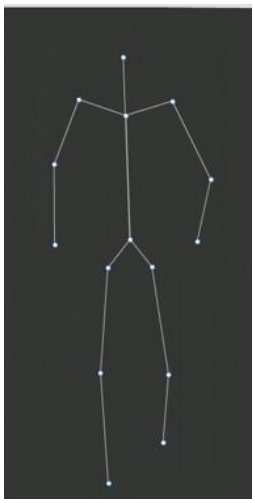

Figure 1 – A 'point-light walker': biomotion perception experiments are often based on video projections using simplified human representations (avatars) of this kind. Integration of aural stimuli (e.g. avatar step sounds) requires not only correct spatialisation (so that they can be perceived to have been emitted at the desired points in space) but also accurate synchronisation with the corresponding visual stimuli.

2 AURALISATION MODEL DEVELOPMENT

2.1 Virtual Microphone Positioning

The first auralisation tool developed in the context of this project [1] was designed to simulate the process of microphone positioning in audio recordings. The application is based on a dense, regular grid of impulse responses pre-recorded on the room region under study for a given sound source position – see Figure 2.

The desired microphone trajectory is specified using the mouse cursor on a diagram representing that region. Each block of output sound is obtained by convolving the anechoic stream representing the sound source with the room impulse response (RIR) corresponding to the position currently selected. The convolution engine uses a very efficient block-based variation on the overlap-add method, especially suited to accommodate the change of RIR filter on successive audio output blocks; a short cross-fade is applied to suppress audible RIR transition glitches.

This program allows real-time operation with room impulse responses of virtually unlimited duration. However, being based on *in-situ* measurement of monaural RIRs, it is does not lend itself directly to model-based binaural auralisation.

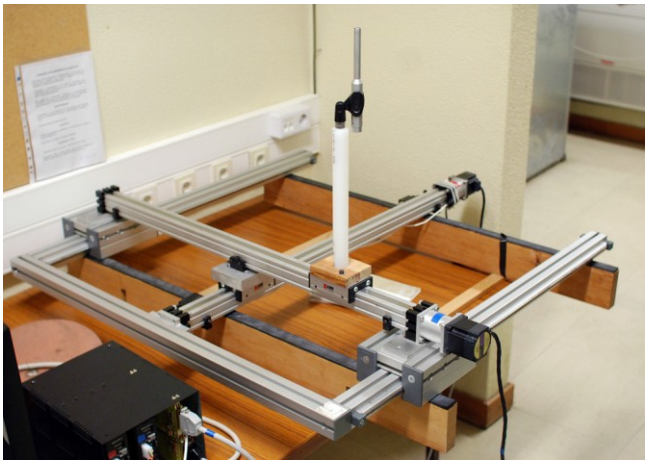

Figure 2 – Mechanical platform used for recording a 2-dimensional RIR grid.

2.2 *LibAAVE* Auralisation Library

The development of the main software package for auralisation in interactive virtual reality environments built on a line of research on audio-visual VR initiated at IEETA with the MEng dissertation project 'Virtual Hall' [2]. A demo from this project using a low-cost head-mounted display (HMD) system was awarded the first prize at the 2007 Audio Technology contest of the Portuguese section of the Audio Engineering Society (AES). An important potential application envisaged for this kind of systems was explored in [3].

In order to ease the gradual refinement of the room modelling and auralisation functions and the integration of new features, the software package was shaped as a library named *LibAAVE* and made freely available under a public license (https://code.ua.pt/projects/acousticave). *LibAAVE* is presented in some detail in [4] and [5]. It is designed to take a geometry model of the virtual room in *.obj* format and allow arbitrary movement of both sources and listener. Its basic operation principles are illustrated in Figure 3. Two functional blocks can be distinguished: acoustic model processing and audio processing.

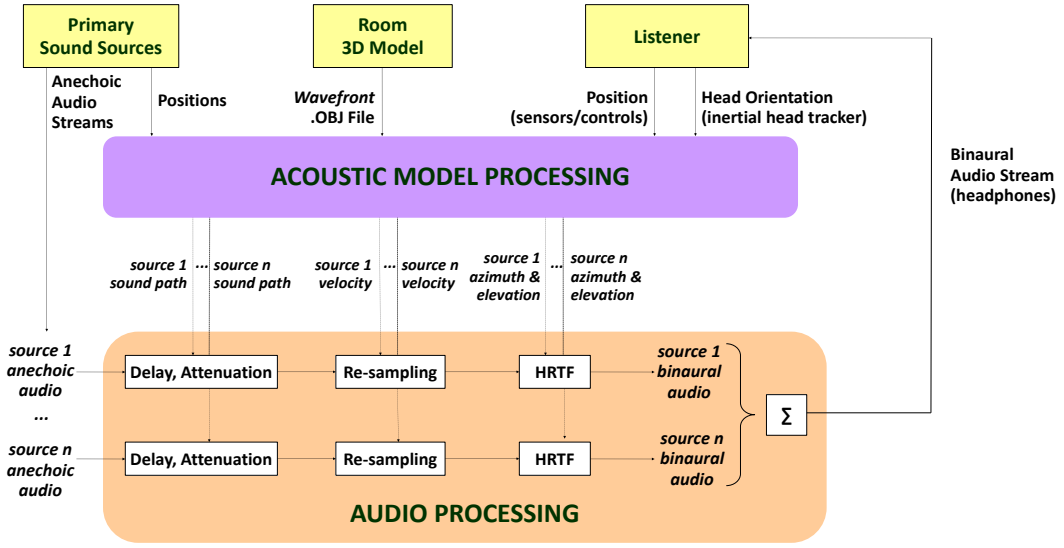

Figure 3 – Overall LibAAVE operation structure.

The acoustic model processing block is based on simple, well-established geometric acoustic models: the direct sound and reflections from each primary source are computed dynamically by the mirror-image source (MIS) method, taking into account the acoustic properties of the room (extracted automatically from its 3D geometry model), the sound source trajectories and listener position, tracked in real time. The result is a set of sources (primary sources plus respective mirror-images up to a certain order specified by the user). Source 'visibility' tests are applied to determine which of these need be considered.

The audio processing block generates binaural output sound, suited for headphone or earphone presentation, by adding the contributions from all the 'visible' sources. The anechoic sound streams from each source are filtered according to their propagation path characteristics. In particular, directional cues are obtained by applying head-related transfer-function (HRTF) filters according to the propagation angles relative to the listener head, worked out in real time with the help of a head-tracking device. An *Intersense InertiaCube BT* was used during the development stage at IEETA. Different HRTF sets can be selected, taken from public-domain databases, namely the KEMAR-based set from MIT Media Lab [6] and CIPIC [7]. As in the virtual microphone positioning system [1], cross-fading is applied to avoid audible HRTF transition glitches.

The MIS method is only suited to real-time simulation of the very early part of the RIR. In other words, for real-time operation the user must specify a sufficiently low maximum reflection order (typically $n<6$ for a single primary source on the platforms tested); this is because the computational cost of checking source visibility increases exponentially with reflection order [5]. The late RIR must therefore be simulated by other means. The reverberation tail solutions adopted in *LibAAVE* are described in [8]. Acceptable sound quality was obtained with the Datorro algorithm, but a feedback delay network (FDN) was preferred, as it allows frequency-dependent RT_{60} configuration.

LibAAVE includes functions to support visualisation of the virtual room based on its *.obj* geometry model. Figure 1 illustrates one of the various interactive applications developed to demonstrate its operation.

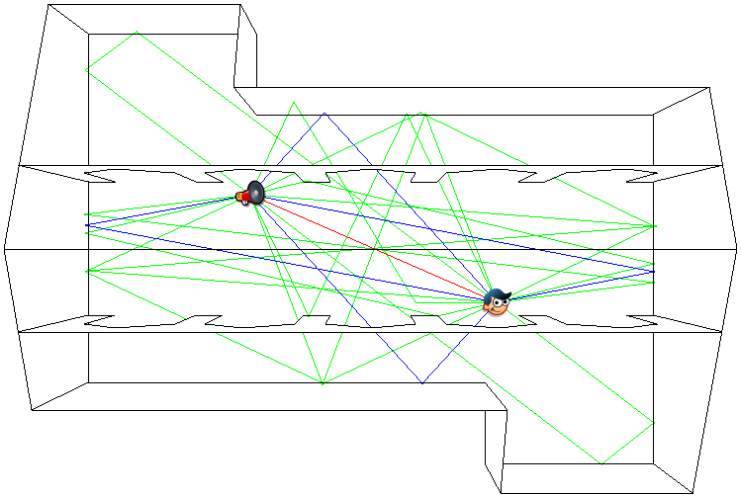

Figure 4 – LibAAVE demonstration GUI: the movements of the virtual source (loudspeaker) and listener (face) within the room are controlled using the mouse; head orientation can be controlled using the mouse or a head-tracking device (Intersense InertiaCube BT). The lines between represent the sound paths from 'visible' sources up to the reflection order specified by the user. The binaural output at the listener position is played in real time on headphones.

Significant effort was put into documentation, to ease future use and refinement. Along with the code, a report including examples of how to build auralisation models is available at UA's software repository (https://code.ua.pt/projects/acousticave). Demonstration code and videos are available at http://sweet.ua.pt/paulo.dias/AcousticAVE/.

2.3 User Tracking and Auralisation with iOS Devices

The integration of *LibAAVE* with an ultrasonic indoor localisation system to track listener position was tested in [9]. The localisation system can be based on mobile devices (e.g. smartphones) and, in contrast with the *Vicon* motion capture system used at LVP (see section 3), requires only a few small, inexpensive ultrasound beacons installed in the room, making it a much more portable alternative, especially useful for audio augmented reality applications. A prototype was built and tested using an iOS smartphone equipped with accelerometer, gyroscope and magnetometer, which allow inertial head-tracking with automatic drift correction. *LibAAVE* is currently being ported to iOS, since the available memory resources also allow the auralisation software to be run from the device itself [10].

2.4 3D Data Acquisition for Room Acoustic Modelling

Practical usage of an auralisation package demands efficient tools of feeding its acoustic model with the relevant room configuration data, namely 3D geometry and acoustic properties of the boundary materials. In the *LibAAVE* case, this means creating an appropriate .OBJ room model.

The problem of acquiring real room data – especially important for validation purposes – was addressed in [11] using the *Microsoft Kinect* sensor and the *Kinect Fusion* application to generate polygonal models of the room boundary surfaces. Tools to help automate the identification of surface materials and assignment of acoustic absorption/reflection coefficients to each polygon, as required in geometric modelling, were developed with the *Visualisation Toolkit* (*VTK*). A voxelisation algorithm was also developed to generate, based on the surface polygonal model, a 3D node grid covering the volume of the room. This is useful for physical room acoustic modelling, whose application is envisaged in future *LibAAVE* developments. With the information on room surface absorption (from the polygonal model) and volume (from the 3D grid), it is possible to estimate reverberation time (RT_{60}) automatically using Sabine's formula and shape the reverberation tail accordingly. A summary of this work can be found in [12]. The data acquisition and modelling algorithms were tested on a meeting room at IEETA, as illustrated in Figure 5.

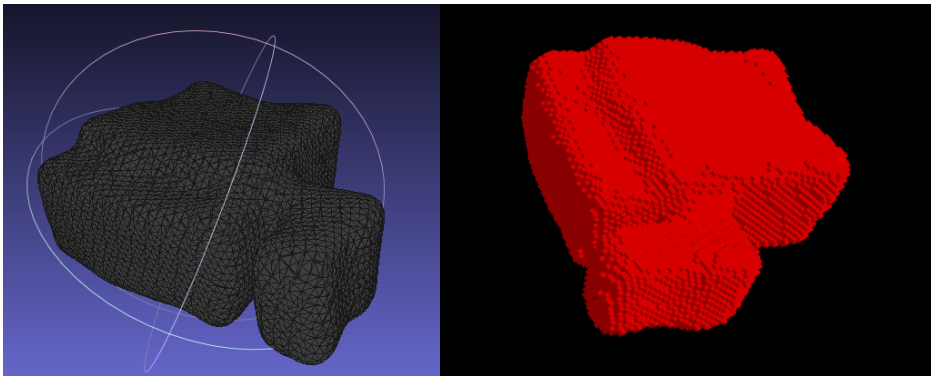

Figure 5 – Polygonal model of room boundary surfaces (left) and corresponding voxelisation (right).

3 MODEL INTEGRATION

Over the course of the project, successive versions of *LibAAVE*, as well as simpler tools for offline generation of binaural sound (e.g. a MATLAB application developed early in the project for synchronised avatar step sound generation) were installed, configured and adapted at LVP to suit the specific needs of each audio or audio-visual perception experiment.

An important task was to feed *LibAAVE* with real-time listener/source position and head orientation data from LVP's *Vicon* motion-capture system, which is based on infra-red cameras tracking reflective stickers fixed to the relevant targets (in this case listeners and sound sources). *Vicon*'s software development kit (SDK), a *C* library allowing communication with Vicon's applications (*Nexus*, *Blade* and *Tracker*), was used for that purpose. Although not ideal

in terms of latency and precision, head-orientation detection with the *Vicon* system is possible using a set of 3 stickers to define the head's reference axes.

In March 2014, at the second of two 'Auralisation Models and Applications' workshops organised to showcase the project, the *LibAAVE-Vicon* motion capture integration was demonstrated with a real-time audio-visual VR simulation of a musical quartet.

As explained in the Introduction, controlling the relative latency of the audio and video chains is a crucial requirement, not least because audio-visual synchrony perception is a research issue in its own right – take, for instance, [17] and [18]. With the help of a *Brüel & Kjær Pulse* data acquisition system and appropriate instrumentation, a device was implemented to analyse the degree of synchrony of different signals (electric triggers, motion-capture events, markers on the auralised sounds and projected images...) and employed in experiments to measure the processing latencies of the audio, video and motion-capture signal chains of the VR environment. The results were documented in an *LVP* internal report and, based on it, a guide was prepared on how to perform this kind of measurements and assess/adjust synchrony.

4 PSYCHOPHYSICS RESEARCH

4.1 Adaptation to Non-Individualised HRTFs

As any auralisation package based on geometric room acoustics, *LibAAVE* relies on HRTF filtering to impart the directional cues that allow 3D source localisation. Since obtaining individualised HRTF filter sets would pose very serious practical difficulties, HRTF sets measured on dummy heads with 'average' characteristics are used instead. As mentioned before, two such HRTF sets were adopted in *LibAAVE* [6][7]. It is therefore extremely important to assess the perceptual effectiveness of HRTF processing and understand to what extent the use of non-individualised HRTFs might hinder spatial sound perception.

The issue was addressed from a learning perspective [13][14]. A set of experiments showed that mere exposure to virtual sounds processed with non-individualised HRTFs did not improve the subjects' performance in sound source localisation, but short training periods involving active learning and feedback led to significantly better results. These findings indicate that using auralisation with non-individualised HRTF should always be preceded by a learning period. This work, on the basis of an earlier presentation at the 129th AES Convention, was selected for publication in the AES Journal [15].

An additional set of experiments were devised to investigate this learning effect in further detail. The experiments involved three groups of subjects and a careful schedule of azimuth/elevation localisation training and test sessions over the course of one month. This made it possible to study the persistence in time (memorisation) of the learning effect, its dependence upon the type of sound source and its decomposition in terms of azimuth, elevation and their cross-dependence (i.e. how azimuth localisation training affects elevation localisation performance and vice-versa). Externalisation effects were also studied. The results evidenced that sound localisation with altered cues is easily trained and subject to generalisation effects across space and sound source type: a brief training session with a restricted set of sounds and source directions is enough to improve localisation performance for trained and untrained sounds in trained and untrained directions. The learning effects are persistent; they can still observed one month after training, especially in azimuth localisation. Externalization levels are also increased by training, although not directly related to localisation accuracy levels [16].

4.2 Sound Presentation

The choice between headphones and earphones is an unresolved debate. In order to obtain some guidance regarding this issue, a few experiments were carried out to compare localisation performance with in-ear phones (Etymotics ER-4B) and headphones (Sennheiser HD 650) already available at LVP. This unpublished work involved training using analogous methodologies to those employed in the HRTF studies. The experiments were repeated under

different noise levels; in both cases, the global average localisation error was lower with headphones.

4.3 Depth and Time-To-Passage (TTP) Perception Cues

Significant work was dedicated to the investigation of distance (depth) perception [17][18]. As mentioned before, the interplay between aural and visual cues highlights the importance of controlling audio-visual synchrony. The selective control of reflection orders implemented in the auralisation tools was an important feature in the experimental work leading to [19].

The work on the perception of “time to passage” (TTP) and “time to collision” (TTC) of looming sounds involved experiments with sources of various types travelling different distances at different velocities and with different occlusion rates [20][21]. This required the auralisation tools to be configured with an HRTF database including near-field measurements; the choice fell on the database from the TU of Berlin which comprises measurements at 0.5m, 1m, 2m and 3m [22].

5 FUTURE WORK

Possible refinements to the acoustic modelling algorithms used in *LibAAVE* include, for example:

- Adoption of frequency-dependent acoustic absorption coefficients;
- Dynamic adjustment of the maximum reflection order;
- Quantisation of source path delays so that only primary source FFTs need be calculated;
- Optimisation of source visibility checking algorithms;
- Partial pre-calculation and/or less frequent updating of source visibility;
- Combination of MIS with ray-tracing or beam-tracing techniques;

Since the goal is to maximise model accuracy while ensuring real-time operation, any modification must assessed in terms of both perceptual and computational impact.

LibAAVE can benefit enormously from parallel processing through functional decomposition into two threads (room model and audio) and/or data decomposition within each thread. Both are highly parallelisable, since sources can be processed independently from one another.

We also plan to test potentially more accurate techniques (namely physical modelling) and develop novel hybrid models through combination of techniques. One possibility along those lines is adapting the Virtual Microphone Positioning algorithm to use a grid of *Ambisonics* RIRs (obtainable in a single run of a DWM model) instead of a grid of measured monaural RIRs. HRTF processing can then be applied to perform *Ambisonics*-binaural conversion and allow headphone presentation. The MEng dissertation [23] represented an initial step in that direction.

The PhD project already underway to build on the work developed on audiovisual perception [24] is an example of the numerous threads that can be pursued on the Psychophysics research front. The results on HRTF learning effects (arguably the most significant novel contribution from *AcousticAVE*) call for larger-scale experiments to allow further investigation on the underlying mechanisms and influential factors. Pursuing the (yet unpublished) work on headphone vs. earphone sound presentation (section 4.2) could be very valuable in this regard.

On the application front, we intend to create a permanent demo with the prototype mentioned in 2.3, which shows huge practical application potential. Porting *LibAAVE* to iOS is the most immediate task; an MEng dissertation was proposed to tackle it.

Developing room model configuration tools along the lines discussed in 2.4 is essential to promote practical applications. Acoustic archaeology and walkthrough auralisation in cultural heritage sites are among the most promising.

REFERENCES

[1] Barker T, Campos G, Dias P, Vieira J, Mendonça C, Santos J (2012) 'Real-Time Auralisation System for Virtual Microphone Positioning'. 15th International Conference on Digital Audio Effects (DAFx-12), York, UK, September 17-21, pp. 137-143.

[2] Casaleiro R (2008) 'Sala de Espectáculos Virtual'. MEng dissertation. Dept. of Electronics, Telecommunications and Informatics, University of Aveiro.

[3] Dias P, Campos G, Casaleiro R, Seco R, Santos V, Santos B S (2008) '3D Reconstruction and Auralization of the "Painted Dolmen" of Antelas'. Electronic Imaging Conference 2008 (EI 2008), SPIE Vol. 6805, 6805OY, Three-Dimensional Image Capture and Applications 2008, San Jose, California, USA, January 28-29.

[4] Oliveira A, Campos G, Dias P, Vieira J, Santos J, Mendonça C (2013) 'Aplicação de Auralização em Tempo Real'. 11th Congress of AES Brasil, S. Paulo, Brasil, May 7-9, pp. 98-101.

[5] Oliveira A, Campos G, Dias P, Murphy D, Vieira J, Mendonça C, Santos J (2013) 'Real-Time Dynamic Image-Source Implementation for Auralisation'. 16th International Conference on Digital Audio Effects (DAFx-13), Maynooth, Ireland, September 2-6, pp. 368-372.

[6] Gardner B, Martin K (2000) 'HRTF Measurements of a KEMAR Dummy-Head Microphone'. URL: http://sound.media.mit.edu/resources/KEMAR.html.

[7] Algazi, V. R., Duda, R. O., Thompson, D. M. (2001) 'The CIPIC HRTF database'. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York, October 21-24, pp. 99-102. URL: http://interface.cipic.ucdavis.edu/sound/hrtf.html.

[8] Silva N, Oliveira A, Dias P, Campos G, Vieira J, Santos J (2014) 'Auralização em Tempo Real para Ambientes Virtuais Dinâmicos'. 12th Congress of AES Brasil, S. Paulo, Brasil, May 13-15.

[9] Lopes S, Oliveira A, Vieira J, Campos G, Dias P, Costa R (2013) 'Real-Time Audio Augmented Reality System for Pervasive Applications'. 11th Congress of AES Brasil, S. Paulo, Brasil, May 7-9.

[10] Lopes S, Vieira J, Campos G, Dias P (2014) 'Sistema de Realidade Aumentada Áudio 3D para Dispositivos iOS'. 12th Congress of AES Brasil, S. Paulo, Brasil, May 13-15.

[11] Pereira J (2013) 'Aquisição e tratamento de dados 3D para modelação acústica de salas'. MEng dissertation. Dept. of Electronics, Telecommunications and Informatics, University of Aveiro.

[12] Pereira J, Silva N, Dias P, Campos G, Vieira J (2014) 'Aquisição e tratamento de dados 3D para modelação acústica de salas'. 12th Congress of AES Brasil, S. Paulo, Brasil, May 13-15.

[13] Mendonça C, Santos J, Campos G, Dias P, Vieira, J (2012) 'On the Adaptation to Non-Individualised HRTF Auralisations – a Longitudinal Study'. AES 45th International Conference, Helsinki, Finland, March 1-4.

[14] Mendonça C (2012) 'Audiovisual Perception of Biological Motion'. PhD thesis. School of Psychology, University of Minho.

[15] Mendonça C, Campos G, Dias P, Vieira J, Ferreira J, Santos J (2012) 'On the Improvement of Auditory Accuracy with Non-Individualized HRTF-based Sounds'. J. Audio Eng. Soc. 60(10), pp. 821-830, October.

[16] Mendonça C, Campos G, Dias P, Santos, J (2013) 'Learning Auditory Space: Generalization and Long-Term Effects'. PLoS ONE 8(10). *doi: 10.1371/journal.pone.0077900*

[17] Silva C (2011) 'Perceiving Audiovisual Synchrony as a Function of Stimulus Distance'. MSc dissertation. School of Psychology, University of Minho.

[18] Silva C, Mendonça C, Mouta S, Silva R, Campos JC, Santos J (2013) 'Depth Cues and Perceived Audiovisual Synchrony of Biological Motion'. **PLoS ONE** 8(11). *doi:10.1371/journal.pone.0080096*

[19] Mendonça C, Lamas J, Barker T, Campos G, Dias P, Pulkki V, Silva C, Santos J (2013) 'Reflection orders and auditory distance'. 21st International Congress on Acoustics (**ICA 2013**), Montreal, Canadá, June 2-7 (POMA Vol. 19, 050041).

[20] Silva R (2013) 'Judging Time-to-Passage of looming sounds'. MSc dissertation. School of Psychology, University of Minho.

[21] Silva R, Mouta S, Mendonça C, Lamas J, Silva C, Santos J (2013). 'The role of acoustic cues in time-to-passage judgments: Judging time-to-passage of looming sounds'. 5th Iberian Conference on Perception (**CIP**), A Coruña, Spain.

[22] Wierstorf H, Geier M, Raake A. Spors S (2011). 'A Free Database of Head-Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances'. URL: https://dev.qu.tu-berlin.de/projects/measurements/wiki/2010-11-kemar-anechoic

[23] Santos F (2012) 'Auralização Binaural com HRTF e Descodificação de Ambisonics'. MEng dissertation. Dept. of Electronics, Telecommunications and Informatics, University of Aveiro.

[24] Silva, C. (2012). 'Audiovisual Perception in a Virtual World: An Application of Human-Computer Interaction Evaluation to the Development of Immersive Environments'. PhD Thesis Proposal. School of Psychology, University of Minho.