Measuring end-to-end delay in real-time auralisation systems

João Lamas
Centro Algoritmi, University of Minho, Portugal

Carlos CL Silva
HASLab, INESC TEC, Porto, Portugal

Department of Informatics, University of Minho, Portugal

Mariana Silva
INESC TEC, Porto, Portugal

School of Psychology, Department of Basic Psychology, University of Minho, Portugal

Sandra Mouta
Centro de Computação Gráfica, Guimarães, Portugal

INESC TEC, Porto, Portugal

José Creissac Campos
HASLab, INESC TEC, Porto, Portugal

Department of Informatics, University of Minho, Portugal

Jorge A Santos
School of Psychology, Department of Basic Psychology, University of Minho, Portugal

Centro Algoritmi, University of Minho, Portugal

Centro de Computação Gráfica, Portugal

Summary
One of the major challenges in the development of an immersive system is handling the delay between the tracking of the user's head position and the updated projection of a 3D image or auralised sound, also called end-to-end delay. Excessive end-to-end delay can result in the general decrement of the "feeling of presence", the occurrence of motion sickness and poor performance in perception-action tasks. These latencies must be known in order to provide insights on the technological (hardware/software optimization) or psychophysical (recalibration sessions) strategies to deal with them. Our goal was to develop a new measurement method of end-to-end delay that is both precise and easily replicated. We used a Head and Torso simulator (HATS) as an auditory signal sensor, a fast response photo-sensor to detect a visual stimulus response from a Motion Capture System, and a voltage input trigger as real-time event. The HATS was mounted in a turntable which allowed us to precisely change the 3D sound relative to the head position. When the virtual sound source was at 90º azimuth, the correspondent HRTF would set all the intensity values to zero, at the same time a trigger would register the real-time event of turning the HATS 90º azimuth. Furthermore, with the HATS turned 90º to the left, the motion capture marker visualization would fell exactly in the photo-sensor receptor. This method allowed us to precisely measure the delay from tracking to displaying. Moreover, our results show that the method of tracking, its tracking frequency, and the rendering of the sound reflections are the main predictors of end-to-end delay.

PACS no. 43.58.Vb, 43.66.Yw

1. Introduction

An interactive immersive environment can be characterized as any kind of environment that is capable of creating a users' illusion of being in a place other than where they actually are, or of having a coherent interaction with objects that do not exist in the real world. In this sense, when we talk about immersive environments we are alluding to all the software and hardware elements, needed to present stimuli to the users' senses, which will have this kind of effect – quoted as *feeling of presence* [1].

In order to successfully conveying feeling of presence an immersive environment should convey an accurately replication of the geometric and temporal characteristic of the real world. In this sense, one of the major challenges in the development of an immersive system is handling the delay between the tracking of the user's head position and the equivalent change in the projection of a 3D image or an auralised sound, also called *end-to-end delay* [2]. This delay is the result of latencies in individual components of an immersive system, including the tracking devices, the signal processing, and displaying [3], and no current interactive immersive system is exempt of end-to-end delay.

As this constitutes one of the main problems in immersive environments implementation, it is highly advisable that developers know the origin and the magnitude of these latencies in order to provide insights on the technological (hardware/software optimization) or psychophysical (perceptual recalibration sessions, for example) strategies to deal with them.

Excessive end-to-end latency has been linked to a set of problems that can be divided in three major types: 1) simulation problems; 2) perceptive problems; 3) user-behavior problems. Di Luca [2] presented a list of these problems for visual virtual reality systems, where we can find simulation problems as the occurrence of motion sickness in the user [4] and the reduction of the subjective sense of presence [5]; perceptive problems as the disruption of multisensory information combination [7] [8]; and user-behavior problems as more errors during reaching, grasping, and object tracking tasks [6]. Unfortunately, these types of problems are not as well documented for virtual acoustic environments (VAEs). Nonetheless, there is some work indicating that end-to-end delays can have a big impact on auditory location [3] [9] [10], thus highlighting the importance of measuring end-to-end delays in interactive VAEs.

Thus, our goal was to develop a new measurement method of end-to-end delay that is both precise and easily replicated and adaptable to different VAEs. This method should allow us to correctly identify the latencies for each component of our VAE and therefore to compute a precise value of end-to-end delay. We should keep in mind that each component of a VAE does not necessary have a constant delay [2]. The latencies on the tracking and the signal processing components can vary with the type of tracking method and simulation complexity [3] (auralisation in free-field vs *n* reflection orders). Therefore, we should compute end-to-end delay values that cover all these possible variations. In this paper, we will begin by describing two VAEs (tracking system, auralisation system, and acoustic display) and our test-bed for measuring the end-to-end delay of both. We will then present results of tracking latencies for different methods of tracking and end-to-end delay for different levels of simulation complexity. Finally we will discuss the implications of these results for VAEs development and the possibility of generalize this method of measuring end-to-end delay to other VAEs.

2. VAE Description and Materials

2.1 – VAE Description

The first VAE is based on free-field auralisation with interpolated Head Related Transfer Functions (HRTF). The database used was the MIT HRTF database [11]. This auralisation system produces a 3D binaural sound-field as output. As input an anechoic sound for each sound source and the listeners position and orientation in real-time is needed.

The second VAE is based on the *libaave* auralisation library [12]. This auralisation library uses several inputs to reach a more immersive virtual audio environment: a room-model, listeners position and orientation, sound sources positions along time, and anechoic sounds for each sound source. This auralisation system works together with an image rendering process using *VTK library (Visualization Toolkit)*.

A virtual audio environment is created in both systems using 3D sound, taking into account the user position and orientation. Furthermore, in the second one an audio-visual virtual environment can be generated and a higher level of immersive sound environment can be reached through sound reflections calculated through the image-source method [13] and sound reverberation.

2.2 - Materials

To perform the auralisation process, two different computers were used, one *MacBook Pro* (*Intel* Core Duo CPU @ 2.4GHz, 8Gb RAM memory) with the first auralisation system, and one *DELL Workstation Precision T3600* (*Intel* Xeon CPU E5 Quad Core @ 3.6GHz, 8Gb RAM) working with the second auralisation system. Both computers were connected through gigabit Ethernet connection to the motion capture workstation to access the orientation segment data.

To analyze the end-to-end delay of our VAE a *Vicon*® motion capture system was used as tracker. The motion capture system is composed by six near infra-red (NIR) 2MPixel cameras and an acquisition module *Vicon*® *MX Ultranet*. This system can reach a high frame rate, up to 500Hz. The motion capture software runs on a dedicated workstation (*Intel* Core 2 Quad Processor @ 2.4GHz, 4Gb RAM memory) with optimized network card settings. The software used allows us to acquire real-time position and orientation data from a set of markers (at least three markers) wherein a single segment was defined to correspond to the listeners' head in the VAE.

To detect a real-time event a voltage trigger was used. At the exact time this trigger voltage is set, a NIR LED with the same wavelength of the motion capture system sensitivity (780nm) was also set. To detect the visual stimulus from the motion capture, a fast response photodiode *BPW21R* with a rise time of 3.1 μs was used.

The sound output from both auralisation computers was presented through a set of flat-response in-ear earphones *Etymotics*® ER-4B. The sound output was captured through a *Brüel & Kjær*® (B&K) HATS 4128-C.

The signal acquisition system used to acquire all the three different kinds of signals (light response, sound and voltage trigger) was a B&K *Pulse Platform*. The recordings were made using the B&K *Labshop* software. The signal acquisition was made using a recording mode that allows collecting data with a sampling frequency of 65.5kHz.

3. Procedure

3.1 – Test-bed for tracking internal latencies

We started by analyze the latency between a real event and the motion capture response. This test-bed was built using the voltage trigger, the NIR LED, and the light-sensing device (photodiode).

To acquire the motion capture response, the light sensor was pointing to the screen that was showing a motion capture live camera response. In the moment the voltage trigger was set (by pressing a button), the NIR LED started to emit light and consequently the photodiode was activated through the motion capture response and the differential time between the two signals was recorded (Figure 1).

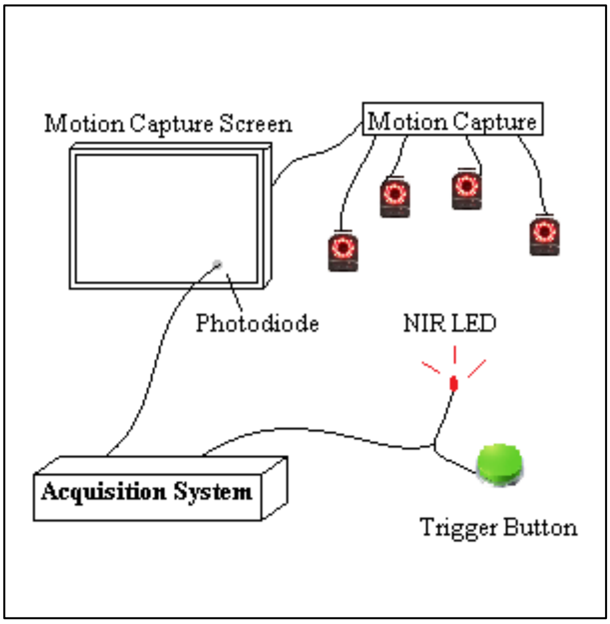

Figure 1 - Test-bed for tracking internal latencies

Several consecutive measures were made with different settings defined in the motion capture system (frame-rate, quality-speed parameter, minimize-latency option, core processor). Our goal with this test-bed was to find the tracking settings that offer a minimal latency.

3.2 – Test-bed for measuring end-to-end delay

The second setup was designed to measure end-to-end delay, between the movement input and the final sound output. In this setup we used a turntable that allowed to change the segment orientation in azimuth, a voltage trigger, the HATS, and the *Pulse Analyzer* to integrate both signals (trigger and audio).

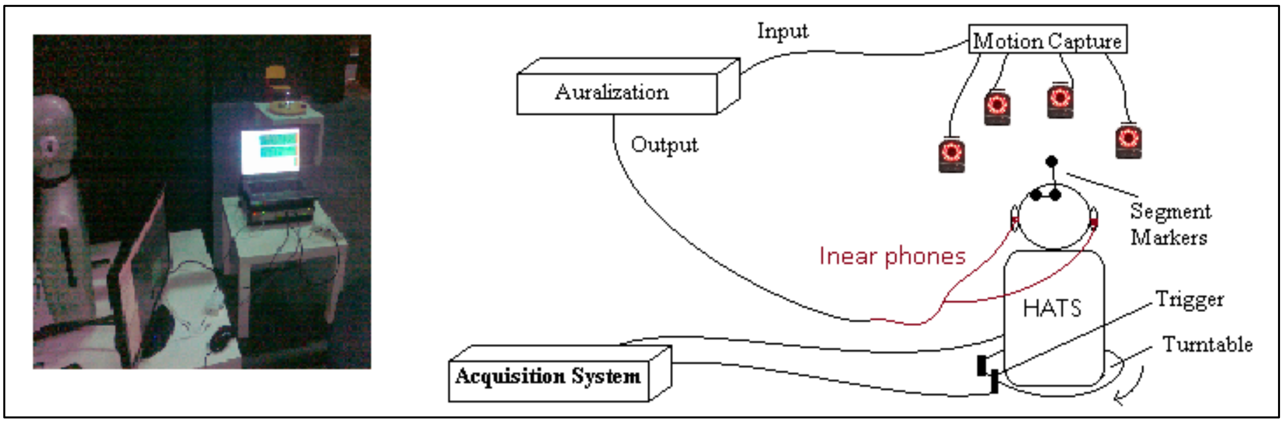

Figure 2 - Test-bed for measuring end-to-end delay

In order to measure this delay, we first placed the turntable center on the origin of the motion capture coordinate system. Then the segment (i.e., markers) was placed on the turntable using concurring reference point as origin, as well as the same motion capture coordinate system orientation. After this procedure, the turntable rotation was able to offer exactly the same azimuth value as the rotation of the segment being tracked.

After this alignment, a pre-defined position (azimuth 90º) was defined in the captured area. In the border of the turntable the voltage trigger was set. When the rotation reached this angle, the trigger was set and captured by the *Pulse Analyzer*.

The VAE computers responsible for the auralisation process perform the correspondent real-time auralisation using a special set of HRTFs, which had a value of zero between 90º and 95º azimuth values. The final auralised sound was null at that a specific point, since filtering an anechoic sound with a zero transfer function results in a zero value. Features like interpolation were disabled to prevent acquisition errors. The Figure 2 shows an image and a schematic of this setup.

After setting up all the hardware correctly, consecutive fast turntable rotations were made allowing fast transition triggers to obtain a cleaner definition of the transition point.

4. Results

In Figure 3 we can see the results of an acquisition made with the test-bed for tracking internal latencies: the response of the photodiode

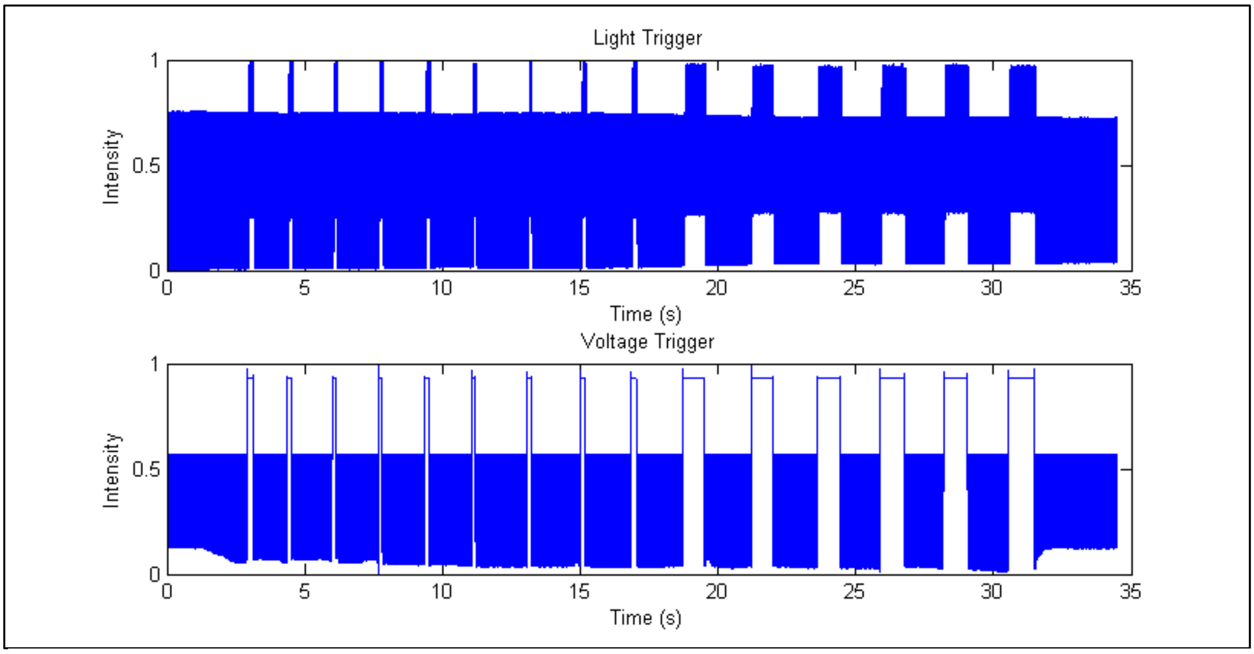

Figure 3 – Acquisition result from test-bed for tracking internal latencies

Table 1 - Tracking latencies

	mean latency (s)	std (s)	min (s)	max (s)
Motion Capture at 100Hz	0.1035	0.0079	0.0941	0.1136
Motion Capture at 250Hz	0.0546	0.0021	0.0516	0.0580

Table 2 - End-to-end delays

	mean latency (s)	std (s)	min (s)	max (s)
Auralisation (free-filed)	0.1074	0.0269	0.0658	0.1338
Auralisation (one RO and rendering)	0.3320	0.0533	0.2996	0.4038

(first plot) and the temporal response of a real event as tracked in real-time by the motion capture system (second plot). Using Matlab to calculate the difference between both peaks we get a measure of latency. Subtracting the beginning of the voltage trigger rising transition by the beginning of the light trigger rising transition give us the accurate latency time for tracking.

Different acquisitions with different tracking settings were made and the final average values were used to define the best software features to achieve a faster motion capture response (i.e., quality/speed feature). We run a comparison for frame rate of motion capture, presented in Table 1. Despite measuring a photodiode response from a screen response (which has a specific frame rate and can produce a system latency calculation error) the obtained values are very approximate to the motion capture system latency because by using the mean value of all measures done, we attenuate the possibility for latency calculation errors. According with information provided by motion capture manufacturer, the software used to acquire the camera image is the fastest architecture available.

In the test-bed for measuring end-to-end delay we used the previous fastest tracking settings (see Table 1).

One end-to-end delay measurement example with several turntable movements is presented in Figure 4. We can see the real-time auralisation output in the first plot and in the second plot the voltage trigger, which corresponds to a real event of passing by azimuth 90º.

In this specific case we have free-field real time auralisation, wherein an anechoic sound (an high frequency sinusoid was used) was filtered with the HRTFs. In this figure we can clearly see the null result when the marker segment reached the azimuth 90º and an accurately latency measure is possible to obtain by calculating the temporal differences between the two signals. In Table 2 we present two different measurements: one using the first VAE, obtained with the first auralisation system described, and the second one using the other system which uses reflection orders and an image rendering process.

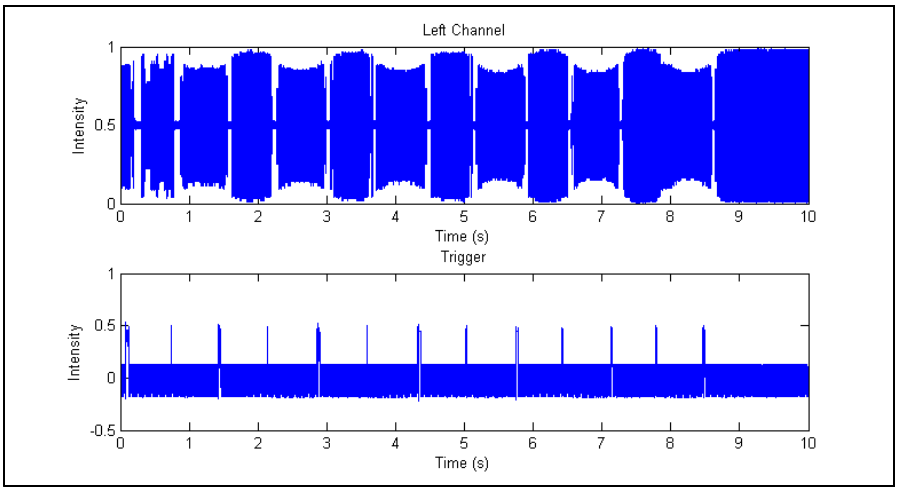

Figure 4 - Result example from end-to-end delay measurement

5. Discussion

The results of the test-bed for tracking internal latencies clearly indicated that, when possible, using a higher frame rate in tracking results in lower latency values. By analyzing Table 1 we can observe that the decrement in latency is not linearly related with the increment of motion capture frame rate. However, these results may indicate that increasing frame rate will decrease the latency values to an asymptotic value (latency resulting from camera photosensor response plus communications).

Despite being possible to track at a higher frame-rate (500Hz) than the ones presented, we did not use this frame-rate in order to reduce marker ghosts that can interfere with the real-time acquisition of the defined segment.

The results of the test-bed for measuring end-to-end delay allowed to measure end-to-end delay in two different VAEs. These measures indicated that the higher VAE complexity, the higher end-to-end delay. This relation is also true for the standard deviation values of the end-to-end delay (see Table 2). We should point out that the second VAE tested included an image rendering process that might have contributed for a higher end-to-end delay.

6. Conclusion

Our goal was to develop a new measurement method of end-to-end delay that is both precise and easily replicated.

This paper describes a method to measure end-to-end delays in VAEs. We successfully applied this method to different VAEs, which allowed us to directly compare general auralisation processes and also different tracking settings. Our results show that the method of tracking, its tracking frequency, and the rendering of the sound reflections are the main predictors of end-to-end delay.

In order to get a better idea of the implications of simulation complexity for end-to-end delay, in future studies we pretend to apply this same method to measure a single VAE system capable of different degrees of simulation complexity.

Acknowledgement

This work has been supported by FCT – Fundação para a Ciência e Tecnologia: EXPL/MHC-PCN/0162/2013; FCOMP-01-0124-FEDER-022674 and within the Project Scope: PEst-OE/EEI/UI0319/2014; FCOMP-01-0124-FEDER-037281 and PEst-C/EEI/LA0014/2013. This work was financed by FEDER grants through the Operational Competitiveness Program – COMPETE and by NORTE-07-0162-FEDER-000076-cave@CCG, ON.2 – Programa Operacional Regional do Norte].

References

[1] T. Schubert, F. Friedmann, H. Regenbrecht: The experience of presence: Factor analytic insights. Presence 10(3) (2001) 266-281.

[2] M. Di Luca: New method to measure end-to-end delay of virtual reality. Presence 19(6) (2011) 569-584.

[3] E. M. Wenzel: The impact of system latency on dynamic performance in virtual acoustic environments. Target 135 (1998) 180.

[4] J.A. Ehrlich: Simulator sickness and HMD configurations. Proc. Intelligent Systems & Advanced Manufacturing 1997, 170-178.

[5] R. Held, D. Nathaniel: Telepresence, time delay and adaptation. Pictorial communication in virtual and real environments (1991) 232-246.

[6] R. Teather, A. Pavlovych: Effects of tracking technology, latency, and spatial jitter on object movement. 3D User Interfaces. 3DUI 2009. IEEE Symposium on. IEEE, 2009.

[7] M. Di Luca, B. Knörlein, M- O. Ernst, M. Harders. Effects of visual–haptic asynchronies and loading–unloading movements on compliance perception. Brain research bulletin 85.5 (2011) 245-259.

[8] B. Knorlein, M. Di Luca, M. Harders: Influence of visual and haptic delays on stiffness perception in augmented reality. Mixed and Augmented Reality. ISMAR 2009. 8th IEEE International Symposium on. IEEE, 2009.

[9] D. S. Brungart, B. D. Simpson, R. L. McKinley, A. J. Kordik, R. C. Dallman, D. A. Ovenshire: The interaction between head-tracker latency, source duration, and response time in the localization of virtual sound sources. Proceedings of ICAD 04 - Tenth Meeting of the International Conference on Auditory Display, Sydney, Australia, July 6-9, 2004.

[10] E. M. Wenzel: Effect of increasing system latency on localization of virtual sounds. Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction. Audio Engineering Society, 1999.

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

[12] G. Campos, P. Dias, J. Vieira, J. Santos, C. Mendonça, J. Lamas, N. Silva, S. Lopes: Acousticave: Auralisation Models and Applications in Virtual Reality Environments . European Symposium on Smart Cities and Environmental Acoustics. (2014) URL: https://code.ua.pt/projects/acousticave

[13] E. A. Lehmann, A. M. Joahnson: Diffuse reverberation model for efficient image-source simulation of room impulse responses. Proc. of the IEEE Trans. on Audio, Speech, and Language, 2010, vol.18 no.6 1429-1439.