EuroRegio2016
June 13-15, 2016
Porto ▪ Portugal

Virtual Reality Performance Auralization in a Calibrated Model of Notre-Dame Cathedral

Barteld NJ Postma, David Poirier-Quinot, Julie Meyer, Brian FG Katz

Audio & Acoustic Group, LIMSI, CNRS, Université Paris-Saclay
Email: {first.lastname}@limsi.fr

Abstract
As part of the 850-year anniversary of the Notre-Dame cathedral in Paris, there was a special performance of *'La Vierge'*, by Jules Massenet. A close mic recording of the concert was made by the Conservatoire de Paris. In an attempt to provide a new type of experience for those unable to attend, a virtual recreation of the performance using these roughly 45 channels of audio source material was made via auralization. A computational acoustic model was created and calibrated based on in-situ measurements for reverberation and clarity parameters. A perceptual study with omnidirectional source and binaural receiver validated the calibrated simulation for the tested subjective attributes of reverberation, clarity, source distance, tonal balance, coloration, plausibility, apparent source width, and listener envelopment when compared to measured responses. Instrument directivity was included in the final simulation to account for each track's representative orchestral section based on published data. Higher-Order Ambisonic (3rd order) room impulse responses were generated for all source and receiver combinations using the *CATT-Acoustic TUCT* software. Virtual navigation throughout a visual 3D rendering of the cathedral during the concert was made possible using an immersive rendering architecture with *BlenderVR*, *MaxMSP*, and an *Oculus Rift* Head-Mounted Display. This paper presents the major elements of this project, including the calibration procedure, perceptual study, system architecture, lessons learned, and the technological limits encountered with regards to such an ambitious undertaking.

Keywords: Auralization, Calibration, Virtual Reality.

PACS no. 43.55.Ka

1 Introduction

The use of Virtual Reality (VR) technologies has increased the last decennia due to the improvement of available computing power and the quality of numerical modelling software. This study explored the current potential of VR technologies which combine auralizations and 3D graphics. The global concept of this project was to present a complex VR scene, with numerous acoustical sources, in which the listener could move around having a realistic experience throughout the regarded venue.

Several studies have reconstructed historical sites in terms of audio and visuals. The *ERATO* project [1] constructed acoustical and visual models of archaeological open-air and roofed theatres. Acoustical simulations used the geometrical acoustics (GA) software *ODEON*. Visual reconstructions were created with the *3ds Max* software based on architectural drawings, photos, and videos. The visitor was able to navigate within the visual scene. Auralizations were linked to interactive area triggers, allowing the visitor to perceive and experience the simulated voices from specific positions.

Game engines are a useful platform for combining visuals and audio in VR applications [2]. They offer interactive rendering of visual environments while also enabling the integration of audio and visuals. Lindebrink et al. [3] employed a software platform combining the game engine *TyrEngine* and the room acoustical software *BIM/CAD*. RIRs were calculated and convolved with anechoic recordings offline. When progressing through the visual scene, the audio rendering was performed by playing the sound file of the nearest neighbor.

Another VR application [4,5] created audio-visual scenes employing the game engine *Gamebryo* and rendered the room response in real-time. In order to enable real-time convolution, the RIR was divided into an early and late part. An underlying GA based algorithm computed specular reflections, diffuse reflections, and edge diffraction on a multi-core system. The late reverberation time was simulated by a statistical estimation technique. Physical restrictions were imposed on the motion of source and receiver to generate an artifact-free rendering. In 2010, this application was used to present the visual rendering of the Sibenik cathedral at 20-30 fps in combination with a binaural audio representation of 12 instruments, taking into account the listener's position and orientation.

As with these discussed studies, the current project employed a game engine platform to create an audio-visual reconstruction of an orchestral performance of *'La Vierge'* in the Notre-Dame de Paris cathedral. The cathedral's complicated geometry and considerable dimensions (length: ~130 m, width: ~48 m, height: ~33 m, volume: ~80,200 m^3) as well as the number of musicians results in a complex scenario. In contrast to previous studies, the audio-visual rendering was designed to be suitable for a tracked Head Mounted Display (HMD), requiring a higher frame rate than ordinary desktop screens. The combination of a complex scene with the high technical requirements rendered the audio-visual reconstruction suitable to explore the contemporary potential of VR platforms. As the study proposed an exploration of technology, emphasis was placed on identifying technological limitations and perceptual aberrations.

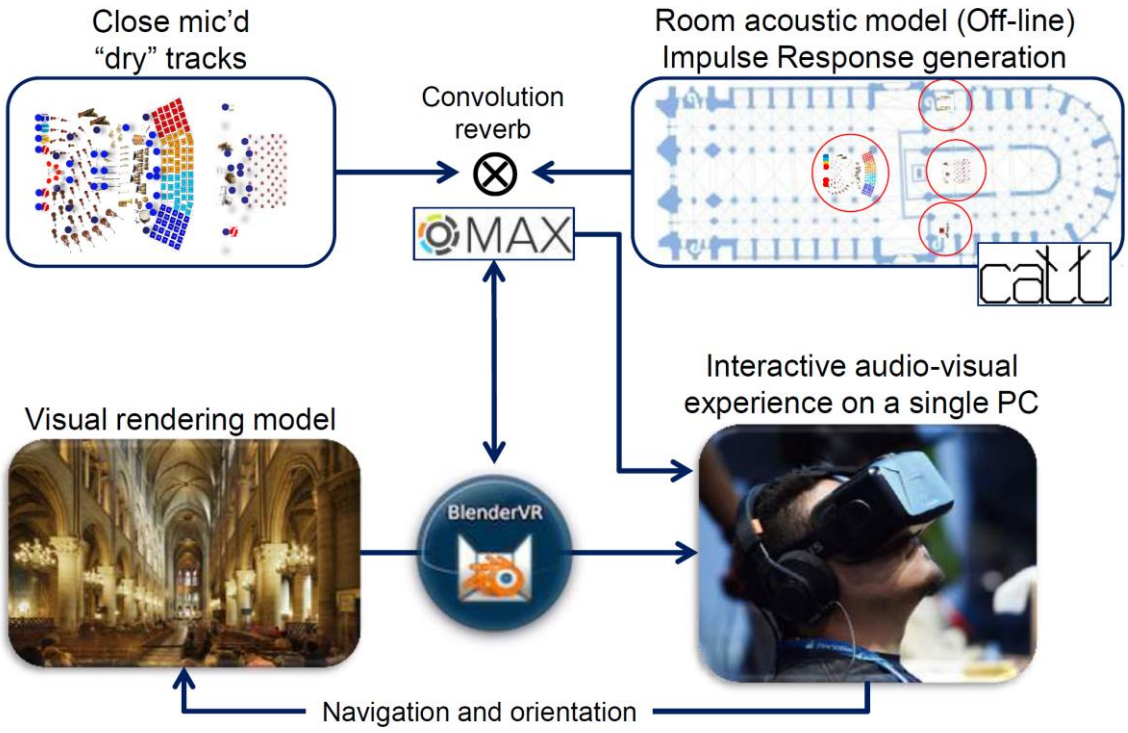


Figure 1 – Schematic diagram of the architecture of the VR experience.

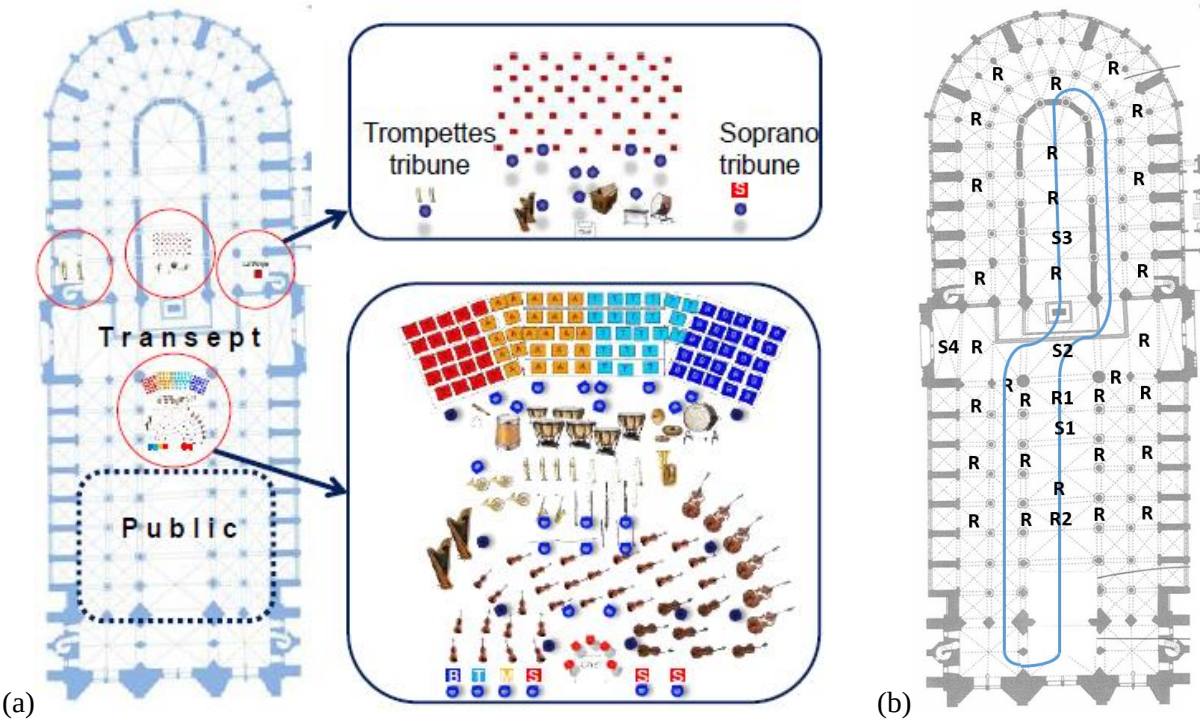
2 Project overview

The first step was to conceive the global project architecture. A recording was made of the *'La Vierge'* concert in the Notre-Dame cathedral. These recordings were convolved with 3rd order Ambisonic RIRs obtained from a calibrated GA model. In parallel, a visual model was created of the Notre-Dame cathedral in *3ds Max*, subsequently ported to the *Blender Game Engine*. The visual and acoustical models were integrated using a platform which combined the interactive VR environments of *BlenderVR* and the audio software *Max/MSP*. Fig. 1 depicts the conceptual architecture of the presented audio-visual VR application.

This paper presents an overview of the different essential elements necessary to achieve the proposed immersive VR experience according to the proposed global architecture.

3 Recordings

On 24-April-2013, a grand concert was organized in the Notre-Dame cathedral, to celebrate its 850th anniversary. A symphonic orchestra, 2 choirs, and 7 soloists performed '*La Vierge*', composed by Jules Massenet in 1880. Fig. 2a depicts the placement of the instruments and section microphones during the concert. The event was recorded by the *Conservatoire de Paris* and made accessible to this study thanks to the *BiLi* project. Each instrument section and soloist were recorded using a total of 44 microphones in close proximity. As the direct-to-reverberant ratio is high for close mic recordings, these were employed as approaching anechoic recordings for the purpose of auralization.



(a) (b)

Figure 2 – (a) Orchestra and microphone (ō) layout for the concert in the Notre-Dame cathedral. (b) Measurement plan in the Notre-Dame cathedral. S# and R represent source and receiver positions (R# positions were employed in the listening test). The blue line depicts the VR trajectory experience.

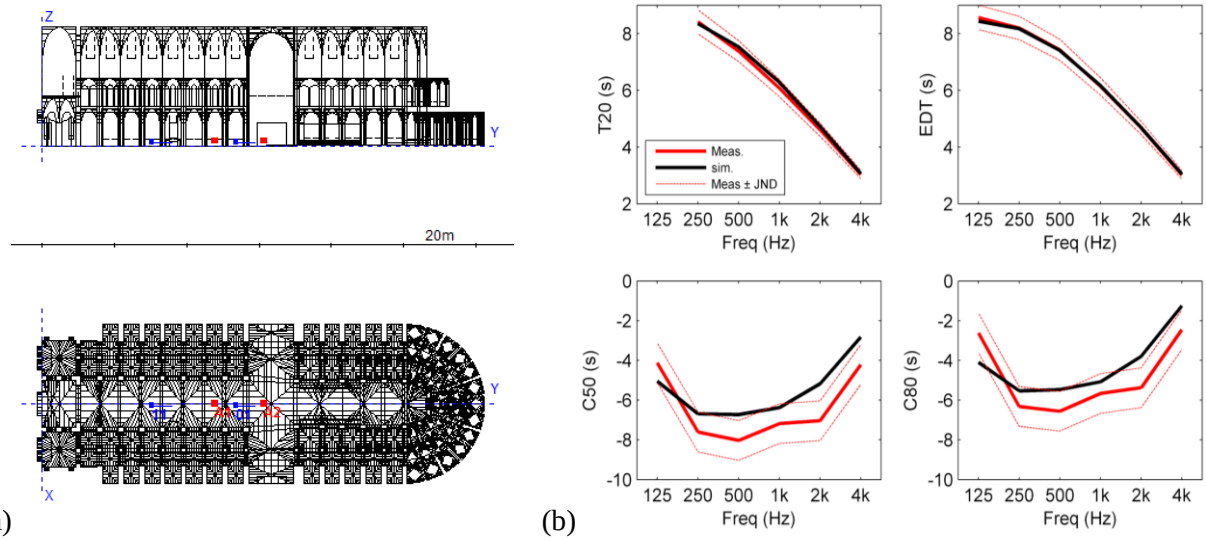
4 Room-acoustic model

4.1 Creation and calibration

The room acoustic model of the Notre-Dame cathedral (see Fig. 3a) was created using the GA software *CATT-Acoustic* (v.9.0.c, TUCT v1.1a) [6]. Calibration was performed according to the 7-step procedure presented in [7]. Room acoustical measurements were carried out to serve as a reference for the calibration. Details of the measurement system are described in [8]. Fig. 2b shows the measurement plan with **S1-4** representing the source positions and **R**'s depicting the omnidirectional and binaural microphone receiver positions. It should be noted that the binaural head was always orientated towards **S2**. T20, EDT, C50, and C80 were calculated for the purpose of this study.

The geometry of the Notre-Dame cathedral was determined from a 3D laser scan point cloud and architectural plans & sections. Surface materials were determined from visual inspection. Initial absorption coefficients were taken from publicly available databases [9,10,11]. Scattering coefficients of surfaces were generally modeled using the CATT option *estimate*, providing frequency dependent estimations based on a given characteristic depth representative of the surface's roughness.

The Notre-Dame cathedral is a venue with a fairly even absorption distribution. As such, simulations were run with CATT Algorithm 1: *Short calculation, basic auralization* with 250,000 rays. Source and receiver positions were defined corresponding to the measurements. The binaural GA simulation incorporated the previously measured Head Related Transfer Functions of the dummy head (Neumann KU80, DPA 4060 microphones) employed during the measurement. Fig. 3b presents a comparison of mean measured T20, EDT, C50, and C80 to those of the calibrated GA model. Simulated *reverberation* parameters EDT and T20 are within one Just Noticeable Difference (JND) of the measured values across all frequency bands. The simulated *clarity* parameters slightly overestimate measurements in the 500 and 2000-4000 Hz octave bands. This was probably caused by the combination of the calibration order (step 5: bringing reverberation parameters within 1 JND, step 6: adjusting the scattering coefficients to calibrate for the *clarity* parameters) and the baseline requirement that material properties should be simulated with physically viable values.



(a) (b)

Figure 3 – (a) 3D GA model of the Notre-Dame de Paris cathedral (~14,800 surfaces).
(b) Comparison between simulated and measured mean (±1 JND) T20, EDT, C50, and C80.

4.2 Source and receiver definitions of the final model

With the GA model sufficiently estimating the acoustical properties, within or near 1 JND, source positions were defined according to the 44 microphones used during the recording. In order to approximate the directivity properties of instruments, directivity patterns were defined in the GA model, based on [12,13,14,15]. Additional details can be found in [16]. All sources were aimed at the conductor position. As no directivity data for a harmonium or glockenspiel was found, these were defined as omnidirectional sources.

To reduce the number of RIRs necessary to compute, and to limit GPU/CPU load at run-time, 1D linear, as opposed to 2D planar, panning was selected for the interactive navigation component of the VR application. A single trajectory path was defined along which the visitor was free to move (see Fig. 2b). Receiver positions were defined along this trajectory at ~3 m intervals, resulting in 88 receiver positions. The use of 3rd order Ambisonic microphone RIRs allowed for real-time binaural conversion of the HOA audio stream, taking into account the HMD's head orientation at run-time.

5 Subjective validation of the calibration

To validate the GA model calibration, a listening test was carried out comparing measured to simulated binaural auralizations. The test protocol was based on a previous study [8]. For this test, the omnidirectional source directivity was used, as it corresponded to the measured configuration.

5.1 Preparation of the auralizations

Prior to the listening test, some additional processing was required concerning the measured RIRs. First, the frequency response characteristics of the measurement system were compensated for by creating an equalization filter. Subsequently, differences in Signal-to-Noise-Ratio (SNR) between frequency bands were compensated for by extending the signal decay beyond the background noise level. These processing steps are described in detail in [7]. Additionally, the previous study indicated that the measured binaural auralizations were judged "brighter" than their simulated counterpart while the monaural auralizations were judged equal for this attribute. Therefore, a secondary equalization filter was generated and applied to the measured binaural RIRs to achieve the same mean spectral response.

The resulting measured and simulated binaural RIRs were convolved with an anechoic audio extract of a soprano singing *Abendempfindung* by W.A. Mozart, a stimuli judged appropriate for the acoustic function of the venue. The RMS of the measured and simulated convolutions was used for level normalization.

5.2 Test setup

The measured and simulated binaural auralizations were compared with an AB subjective listening test. Three auralization configurations (**S1R2**, **S2R1**, and **S2R2**) were compared. One configuration was repeated to monitor the repeatability of participant's responses. One pseudo pair (i.e. A≡B) was tested, to determine the reliability of the participants, resulting in 5 tested pairs. Initially, 3 training pairs were presented to the participants under supervision to ensure they understood the task. Results for the training pairs were not tabulated.

Participants were asked to rate the similarity of sample AB pairs according to *Reverberance*, *Clarity*, *Distance*, *Tonal balance*, *Coloration*, *Plausibility*, *Apparent Source Width* (*ASW*), and *Listener Envelopment* (*LEV*). Participants responded using a continuous graphic slider (100 pt) scale, with end points *'A is much more ...'* and *'B is much more ...'*, corresponding respectively to values of -50 and +50, with a center 0 response indicating no perceptual difference. Configuration presentation order and AB correspondence to simulation and measurement auralization files were randomized.

Participants were able to listen to the AB pairs as many times as desired. Auralizations were presented via headphones (Sennheiser, model HD 600) at an RMS level of 75 dBA.

A total of 21 participants (mean age: 35.8 yrs, SD: 11.6) took part. Participants were preselected to have experience with either acoustics or vocal/instrumental performances. 18 participants were tested in an isolation booth located at LIMSI (ambient noise level < 30 dBA) and 3 were tested in a silent office located at the *Institut National d'Histoire de l'Art* (*THALIM*) (ambient noise level ~31 dBA).

5.3 Results

Initial analysis concerns the repeatability of responses, determined from the absolute difference between repeated configuration results. The mean difference between repetitions over participants and attributes was 9.3 pt. This value was used as the tolerance range to judge whether a subjective acoustic attribute differed perceptually between measurement and simulation auralizations. Results (see Fig. 4) for the vast majority of acoustical attributes show mean and median values, as well as principal distributions, near 0, within the response repetition tolerance. However, the simulated auralization of position **S2R1** was judged "clearer" and "closer" than measured auralizations. This could be due to the short source-receiver distance and consequently high direct-to-reverberant ratio whose steep early decay differed between measured and simulated responses, potentially due to local scattering properties.

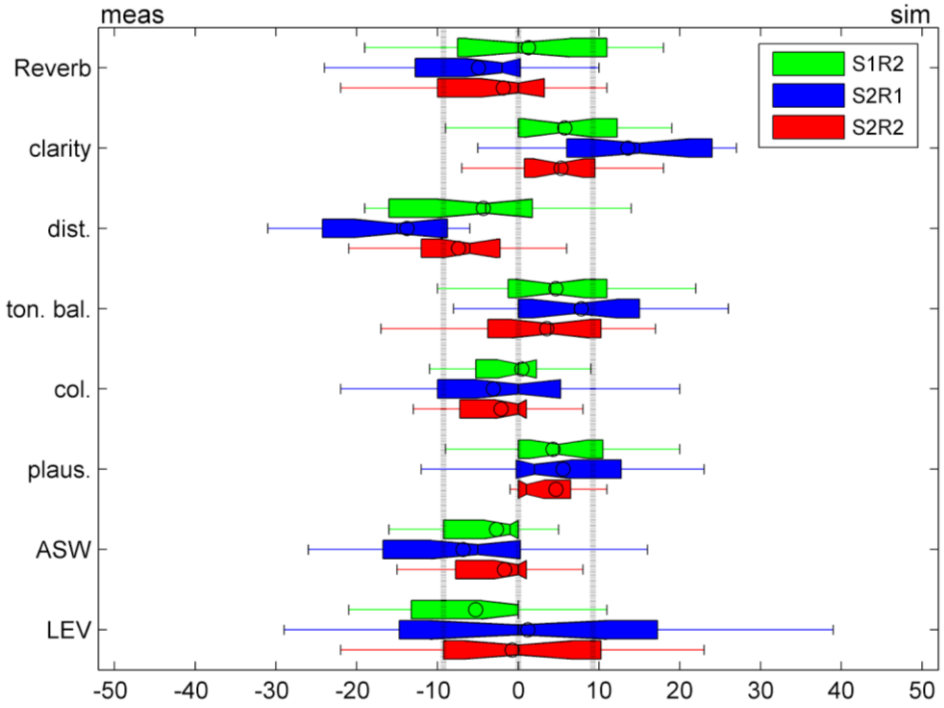


Figure 4 – Perceptual attribute response results of subjective similarity of measured and simulated RIR auralizations, per **S#R#** configuration, over all subjects. Box limits represent 25% and 75% quartiles, (box notch ⏐) median, and (○) mean values. Thick vertical gray reference line at 0 depicts a neutral response, reference lines at ±9.3 indicate repeatability tolerance range.

6 Visual model

6.1 Visual model of the Notre-Dame Cathedral

To accompany the virtual acoustic reconstruction, a 3D visual model of Notre-Dame cathedral was created. The visual model was created in *3ds Max* and subsequently ported to the *Blender Game Engine*. The model geometry was based on a 3D laser scan point cloud, plans and sections, as well as visual inspection. The final model consisted of ~500,000 triangles. The textures employed in the model were based on photos taken during on-site visits. Fig. 5 shows a photo of the cathedral and the *Blender* visual model.

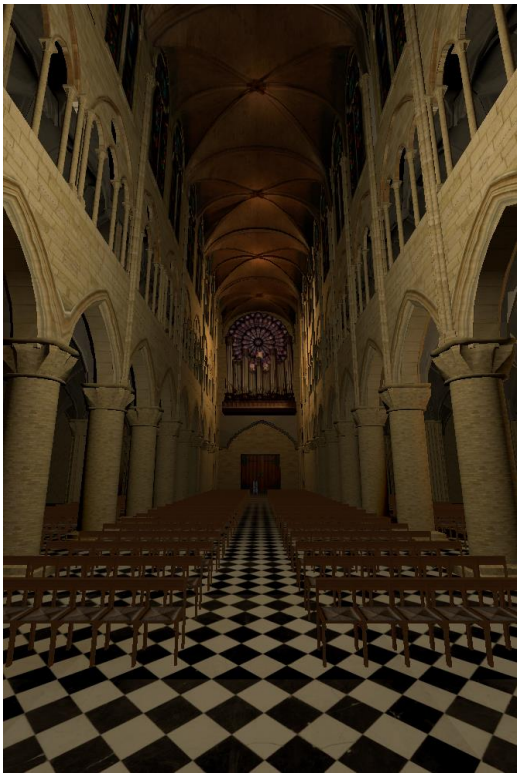
(a)  (b) 

Figure 5 – (a): Picture of the Notre-Dame de Paris cathedral from the altar towards the organ (from [17]). (b): Similar view in the *Blender* VR model.

6.2 Visual animations

In order to create a more engaging immersive VR experience, animations were added to the static cathedral *Blender* model. First, instruments were represented in the 3D environment and positioned in the virtual cathedral to visualize the different components of the orchestra. These instruments included an animated shadow, changing shape as a function of the associated audio track's amplitude in real-time. Second, a 'magic carpet' was added, upon which the visitor sat while exploring the venue. The carpet provided a visual anchor for the 'plausibility' of flying while also avoiding the visual sensation of being suspended several meters in the air with no support, as the HMD allows one to freely look in all directions. On the magic carpet, visitors were free to progress along the predefined trajectory path at a user defined speed and direction using a simple forward-backward joystick controller.

7 Integration of acoustics and visuals to create the VR experience

7.1 Integration architecture

Visual models and room-acoustic simulations were integrated using *BlenderVR* and the audio software *Max/MSP*. The visitor was able to navigate along the trajectory path in real-time within the visual model in the *Blender Game Engine*. Visitor's position (magic carpet) and head orientation (HMD) were communicated to *Max/MSP* which used the position information to perform an amplitude panning between the two nearest HOA receiver positions. Subsequently, head orientation was employed as an HOA rotation prior to decoding the panned Ambisonic stream auralization to the final binaural rendering (see [18,19]).

7.2 Technological limitations

In order to create a smooth audio-visual VR application, it was required that (1) the auralizations were perceived without audible crackles and (2) the visuals were rendered with a sufficiently high frame rate as well as no visual pixelization. In the design process several technological limitations regarding the VR experience were encountered.

The audio was originally conceived to be rendered via real-time convolutions of the 3rd order Ambisonic RIRs with the recordings. In order to prevent audible artifacts due to buffer updating and fast speeds, 5 audio buffers for nearest-neighbor receiver positions needed to be loaded and processed for the 1D linear panning. This resulted in the real-time convolution of 3,520 channels (5 receiver positions × 16 channel HOA RIR × 44 instrument tracks). As this put too great a demand on the CPU for a single PC application, it was decided to perform the convolutions offline. Creating preconvolved HOA audio for each received position lead to a drastic increase in the data storage for the application (88 receiver positions × 24 bit audio wav × 16 channel audio = ~12 GB/min). Consequently, the present VR example demonstration application was limited to a 6-minute audio extract, instead of the entire concert performance.

Visually, it was intended that the visitor could see the entire length of cathedral with the animated orchestra, giving the dynamic option to select the time of the day for the visit, comprising an adjustable lighting scene. As a selectable light scene put too much strain on the GPU/CPU, it was decided to 'bake' the shadows according to a single lighting scheme: night-time conditions. Additionally, the depiction of the entire cathedral from one end to the other lead to pixelization issues for distant elements with the HMD's resolution. Therefore, visuals beyond a distance of ~40 m were 'clipped'. These decisions resulted in a VR experience which can only present night-time conditions, using a dynamic illumination which followed the visitor's progression along the predefined trajectory, lighting only the nearby sections.

7.3 Rendering

The created VR experience demonstration was made available on two platforms with different performance requirements. The first option allowed exploration of the rendering interactively along the path with an HMD (*Oculus Rift DK2*). The visitor was able to control their speed as well as the direction of the 'magic carpet' and the entire field of view was available, providing a highly immersive solution. A second lighter option enabled exploration via a tablet. As the *Blender Game Engine* (the foundation of *BlenderVR*) was unable to run complex environments in real-time on a standard tablet, it was necessary to pre-render the visual part of the scene using *Blender Cycles*. This required predefining the visitor's progression along the path and rendering a high definition 360° video and associated single HOA mixed audio track. This approach allowed for exploration of the cathedral through a tablet equipped with orientation sensors, behaving like an orientable window to the 360° virtual world. The tablet's orientation is used to rotate both the visual 360° rendering and the HOA stream which is then converted to binaural in real-time.

7.4 Observed perceptual artifacts

Several artifacts were observed regarding both the acoustical and visual rendering. Two issues were caused by limitations of the 'dry track' assumption of the recording which actually contained a non-negligible level of cross-talk between audio tracks (i.e. other instrument sections). First, when close to a given instrument's position, the sound from different sections seemed to spatially blur instead of being able to distinguish the positions of separate instrument groups. Second, as the visual avatar animations were based on the audio track levels, during loud passages the visual avatars all appeared active instead of only the active instrumental sections (e.g. kettle drum instances). Finally, when the trajectory passed behind pillars, the expected acoustical variations were absent. For these positions the acoustics varies considerably for relatively small displacements. As such, the chosen RIR calculation/panning step size (3 m) was probably too large for the rate of architectural variations, resulting in an amplitude panning result that did not correctly represent the expected acoustical details.

8 Conclusion

The current potential of VR technologies which combine realistic auralizations and 3D graphics in complex geometries was explored. For this purpose, an ambitious project attempted to reconstruct a large concert in the Notre-Dame cathedral in Paris. Visitors were able to experience this immersive interactive audio-visual VR application on high performance system as well as a portable platform. The presented application enabled realistic audio-visual visits to the complex scene of an extract of the *'La Vierge'* concert.

There remains room for improvements regarding the immersion and accuracy of this and comparable VR scenes. For non-cluster based renderings, today's available GPUs/CPUs limit the presented application in several ways. With increased available computational power, the inclusion of real-time convolution and lighting scene of choice will become possible and consequently comparable VR applications will become more immersive and interactive.

In order to provide a more realistic representation of the reconstructed sound scene, it is recommended that one carefully considers the receiver positions in the room-acoustic model with regards to expected spatial variations of the sound field. A denser receiver grid could be used at locations where acoustical variations are expected to vary considerably with relative small displacements, though the use of a denser grid may impose limitations on movement speed in order to avoid audio buffer switching artefacts. Alternatively, such highly varying areas could be excluded from visitor accessibility.

Additional information regarding this work and *YouTube* videos of the final rendering can be found at groupeaa.limsi.fr/projets:ghostorch.

Acknowledgements

The authors are indebted to a number of persons whom assisted in the realization of this project: Jean-Marc Lyzwa (CNSM) for supervising the concert recordings and his assistance during the acoustical measurements, Cyril Verrechia (LIMSI) for the creation of the visual model of the Notre-Dame cathedral, Marc Emerit (Orange Labs) for the IOS tablet viewer with 360° video and HOA to binaural decoding, as well as Dalai Felinto and Martins Upitis for creating the rendering code for the Oculus and 360° video in *Blender*. Additional thanks to THALIM for their help in hosting the listening test and to all participants of the listening test for their time. Special thanks to the Notre-Dame de Paris cathedral personnel for their assistance and patience during the recordings and measurements. This work was funded in part by the French FUI project BiLi ("Binaural Listening", www.bili-project.org, FUI-AAP14) and the ANR-ECHO project (ANR-13-CULT-0004, echo-projet.limsi.fr).

References

[1] Magnenat-Thalmann, N.; Foni, A.E.; Cadi-Yazli, N. Real-time animation of ancient roman sites, *Proc. 4th Intl. Conf. Computer graphics and interactive techniques in Australasia and Southeast Asia (GRAPHITE)*, Kuala Lumpur, 2006, pp. 19–30.

[2] Moloney, J.; Harvey, L. Visualization and 'Auralization' of architectural design in a game engine based collaborative virtual environments, *Proc 8th Intl. Conf. Information Visualisation (IV)*, London United Kingdom, 2004, pp 827-832.

[3] Lindebrink, J.; Nätterlund, J. An engine for real-time audiovisual rendering in the building design process, *Proc. Acoustics 2015*, Hunter Valley Australia, 2015, pp. 1-8.

[4] Taylor, M.; Chandak, A.; Antani, L.; Manocha, D. Interactive Geometric Sound Propagation and Rendering, *Intel Academic Spotlight*, 2010.

[5] Taylor, M.; Chandak, A.; Mo, Q.; Lauterbach, C.; Schissler, C.; Manocha, D. *iSound: Interactive GPU-based Sound Auralization in Dynamic Scenes*, Tech. Report TR10-006, Computer Science. University of North Carolina, Chapel Hill, pp 1-10.

[6] Dalenbäck, B.I.; *CATT-A v9: User's Manual CATT-Acoustic v9*. CATT, Gothenburg (Sweden), 2011.

[7] Postma, B.N.J.; Katz, B.F.G. Creation and calibration method of virtual acoustic models for historic auralizations, *Virtual Reality*, vol. 19 (SI: Spatial Sound), 2015, pp. 161-180.

[8] Postma, B.N.J.; Tallon, A.; Katz, B.F.G. Calibrated auralization simulation of the abbey of Saint-Germain-des-Prés for historical study, *Intl. Conf. Auditorium Acoustics*, Paris, 2015, pp. 190-197.

[9] Vorländer, M. *Auralizations Fundamentals of Acoustics, Modeling, Simulation, Algorithms and Acoustic Virtual Reality*. Springer-Verlag, Berlin-Heidelberg (Germany), first edition, 2008.

[10] Beranek, L. *Acoustics.* Acoustical Society of America, New York (USA), 1986.

[11] Beranek, L. *Concert and Opera Halls: How they sound*. Acoustical Society of America, New York (USA), 1996.

[12] Olson, H.F. *Music, Physics and Engineering*. Dover publications, New York (USA), 1967.

[13] Le Carrou, J.-L.; Leclere, Q.; Gautier, F. Some characteristics of the concert harp's acoustic radiation, *J. Acoust. Soc. Am.*, vol. 127(5), 2010, pp. 3203-3211.

[14] Marshall, A.H.; Meyer, J. The directivity and auditory impression of Singers. *Acustica*, vol. 58(3), 1985, pp. 130-140.

[15] Directivity of instruments included in CATT-Acoustic: Dalenbäck, B.I. Instument directivity, http://www.catt.se/udisplay.htm accessed: 2015-06-23, 2015. Based on measurements performed by PTB, Braunschweig, Germany.

[16] Meyer, J. *Auralisation d'une simulation acoustique calibrée de la Cathédrale Notre Dame de Paris*, Master's thesis, Université Pierre-et-Marie-Curie, 2015.

[17] Murray, S.; Tallon, A.; O'Neill, R. Paris, Cathédrale Notre-Dame, http://mappinggothic.org/building/1164 accessed 2016-03-27, 2016.

[18] Noisternig, M.; Musil, T.; Sontacchi, A.; Höldrich, R. A 3D Ambisonic based binaural sound reproduction system. *AES 24th Intl. Conf. Multichannel Audio.* Alberta, 2003, pp. 174-178.

[19] Picinali, L.; Afonso, A.; Denis, M.; Katz, B.F.G. Exploration of architectural spaces by the blind using virtual auditory reality for the construction of spatial knowledge, *Intl. J. Human-Computer Studies*, vol. 72(4), 2014, pp. 393–407.