A Comparison between Measured and Modelled HRTFs for an Enhancement of Real-time 3D Audio Processing for Virtual Reality Environments

Alejandro Saurí Suárez, Jason-Yves Tissières, Luis S. Vieira,
Reuben Hunter-McHardy, Sam K. Sernavski, Stefania Serafin

Abstract—Sound in Virtual Reality (VR) has been explored in a variety of algorithms which try to enhance the illusion of presence, improving sound localization and spatialization in the virtual environment. As new systems are developed, different models are applied. There is still the need to evaluate and understand the main advantages of each of these approaches. In this study, a comparison of two methods for real-time 3D binaural sound tested the preferences and quality of presence for headphones in a VR experience. Both the mathematical based HRTF and the convolution based on measured HRTF from the MIT KEMAR show a general similarity in the participants sense of localization, depth and presence. Nevertheless, the tests also indicate a preference in elevation perception for the convolution-based measured HRTF. Further experiments with new tools, techniques, contexts, and guidelines are therefore required to highlight the importance and differences between these two methods and other implementations.

Index Terms—3D binaural sound, HRTF, VR

1 Introduction

3D sound rendering is an important element in Virtual Reality (VR) applications. As visual feedback and narrative move towards immersive interaction, investigating how 3D audio rendering tools complement or enhance visual feedback becomes important. 3D sound offers the dimension of elevation to aural perception, which is missing in the stereo or surround formats [10].

In binaural 3D audio sound rendering, several signal processing methods have been developed over the years by modeling and/or estimating Head Related Transfer Functions (HRTFs). Thorough analysis of head-related spectrum as well as their time-domain representation (called Head related impulse response or HRIR), often resulted in complex filter designs, as discussed by Jyri Huopaniemi and Matti Karjalainen [5]. A method that has risen over the past years that seems to be efficient for auralization is structural modeling. In VR, these modeling methods are very attractive as they can be parametrized and have low computational cost.

Moreover, there has recently been a considerable investment and development in this domain. The main actors in VR technology (Oculus Rift, HTC Vive, Sony Playstation VR, Samsung Gear) develop their own audio engines (plug-ins) for rendering 3D sound. This, however, does not have cross-platform compatibility. Additionally, the in-built 3D sound engine in existing game engines such Unity3D is, by empirical observations, less efficient for human auditory cues.

This study implements and evaluates data in order to understand the importance of 3D audio cues in virtual reality and the differences between two methods for 3D audio rendering in real-time described in the following section.

1.1 Measured Head-Related Transfer Function: The MIT Kemar Database

HRTFs or HRIRs contain the static cues of spatial hearing. They describe *"the transmission from a point in the free field to a point in the human subject's or dummy head's ear canal"* [5].

Several HRTF (HRIR) databases are accessible on the Internet, for example, the CIPIC library [6], as well as the MIT KEMAR database [3]. The latter is used here for the implementation of a system capable of navigating through the database in real-time in relation to the Oculus' tracking device.

- *Stefania Serafin is with Aalborg Univeristy Copenhagen. E-mail: sts@create.aau.dk.*
- *Alejandro Saurí Suárez et al. are graduate students in Sound and Music Computing at Aalborg Univeristy Copenhagen*

The problem with real-time implementation of these databases is that they are costly in terms of computation, therefore excessive CPU usage and time-delays may occur. However, for real-time applications, not all of the measurements and samples are needed, thus they can be carefully reduced to a smaller amount. The current project uses the MIT 'compact' database, composed of 368 measurements of 128 samples long HRIRs. The database contains the left and right ear HRIR measurements of the sound emitted from 0° to 180° on the right side of the dummy head with an elevation from −40° to 90° (directly above the head). The density of measurements is higher between elevations −30° and 30° because of the high localization sensitivity of the human ear in this interval.

Elevation	Number of Measurements	Azimuth Increment
-40	29	6.43
-30	31	6
-20	37	5
-10	37	5
0	37	5
10	37	5
20	37	5
30	31	6
40	29	6.43
50	23	8
60	19	10
70	13	15
80	7	30
90	1	N/A

Fig. 1. Number of measurements and azimuth increment at each elevation for the 'compact' database. [3]

1.2 Modelled Head-Related Transfer Function: A Structural Model

The model used is a structural and adaptive HRTF model designed by Brown and Duda [2]. By analyzing the spectral behaviour of HRTFs, they approximate the diffraction effects of sound waves produced by the head, torso, shoulders and pinnae before entering the ear canal. Fortunately, the torso and shoulders have little significance in localization performance, thus they can be omitted and this allows for a decrease in computational cost and a reduction in complexity. The general filter block diagram representation is shown in Fig. 2. The system input is a monaural sound source. The signal is first filtered by a distance model, which is in turn fed to the head model. The stereo output of the head model is then filtered through a pinna echoes model, and finally, a room

model is added to obtain spatialization effects. The overall filter, even though simple, is a robust and efficient way to bypass the costly convolutions of HRTF databases. In the following formulas, elevation ϕ and azimuth θ are measured according to the interaural-polar coordinate system[1].

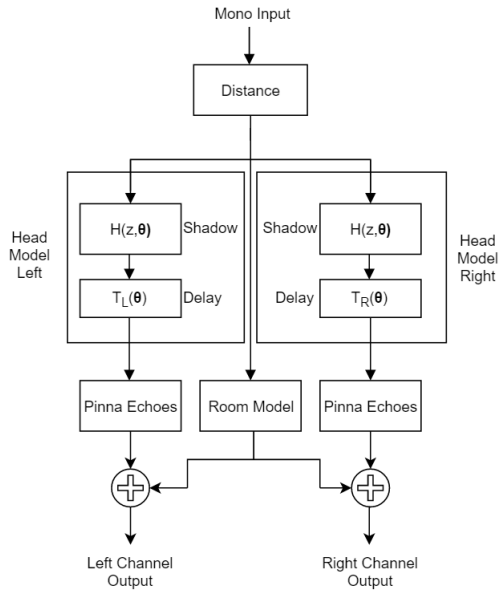


Fig. 2. Filter block diagram representation of the model. [2]

1.2.1 The head model

The head model is based on two physical phenomena that are azimuth dependent: the head shadow effect, which is the diffraction of the incident sound wave around the head, and the Interaural Time Difference (ITD), which is the difference in arrival time of a sound between two ears. The head is represented as a sphere of radius 'a'[2] and an accurate estimate of the time difference between the moment when the sound wave hits the head and when it reaches the ear pinnae is as follows[3]:

Let $T_L(\theta)$ and $T_R(\theta)$ be the difference for the left ear and the right ear, respectively and 'c' the speed of sound (approximately 343 m/s).

$$T_L(\theta) = \frac{a + a\theta}{c} \tag{1}$$

$$T_R(\theta) = \frac{a - a\sin\theta}{c} \tag{2}$$

These equations are for a sound source coming from the right side of the head, i.e. $0° \leq \theta \leq 90°$. If the sound source comes from the left side ($-90° \leq \theta \leq 0°$), equation (1) and (2) are interchanged.

The head shadow effect is implemented as a first order infinite impulse response (IIR) filter which is meant to approximate the Rayleigh spherical head model. It simulates *"the loss of high frequencies when the source is on the far side ear"*, as well as *"frequency group delay at low-frequencies"*[4]. Its analog transfer function is:

$$H(s,\theta) = \frac{\alpha(\theta)s + \beta}{s + \beta}, \tag{3}$$

where $\beta = \frac{2c}{a}$, $\alpha_L(\theta) = 1 - \sin\theta$ and $\alpha_R(\theta) = 1 + \sin\theta$ for the left and right channel, respectively. The coefficient α is a zero location controller in the transfer function depending on the azimuth.

By applying a bilinear transform in order to discretize the filter, the following transfer function is obtained[5]:

$$H(z,\theta) = \frac{(2\alpha(\theta) + \beta) + (\beta T - 2\alpha(\theta))z^{-1}}{(2 + \beta T) + (\beta T - 2)z^{-1}}, \tag{4}$$

where T is the sampling period in seconds.

1.2.2 The pinna echoes model

Avanzini et al. state the following [9]:

"Pinna effects on incident sound waves are of great importance in sound spatialization. Several experiments have shown that, contrarily to azimuth effects which are dominated by diffraction around the listeners head and may be reduced to simple and intuitive binaural quantities, elevation cues are basically monaural and heavily depend on the listeners pinna shape, being the result of a superposition of scattering waves influenced by a number of resonant modes inside pinna cavities."

Therefore, pinna modeling is a complex task to achieve. Brown and Duda's study propose a somewhat simple model that reproduces time-domain 'events' that they found in measured head-related impulse responses (HRIR). These 'events' refer to different significant reflection paths of the sound waves on the anatomical folds and ridges of the pinna before entering the ear canal.

The model is composed of 6 'events' that are azimuth and (strongly) elevation dependent, including the direct sound path into the ear canal. Each path enters the ear canal with different gains and delays, resulting in the model shown in Fig. 3.

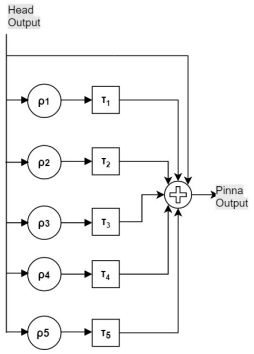


Fig. 3. The pinna model. [2]

ρ_k and τ_k represent the gain and delay of the kth 'event', respectively. The latter is approximated as follow:

$$\tau(\theta,\phi) = A_k \cos(\theta/2)\sin(D_k(90° - \phi)) + B_k, \tag{5}$$

where A_k is an amplitude, B_k is an offset, and D_k is a scaling factor adoptable to each listener. For the purposes of this research, idealized pinna coefficients were taken from [2]. They are shown in table 1.

Table 1. The pinna model coefficients. [2]

k	ρ_k	A_k	B_k	D_k
1	0.5	1	2	1
2	-1	5	4	0.5
3	0.5	5	7	0.5
4	-0.25	5	11	0.5
5	0.25	5	13	0.5

[1] *"The azimuth θ, measured as the angle over from the median plane"* [6], *"is restricted to the interval from $-90°$ to $+90°$, while the elevation ϕ ranges over the full interval from $-180°$ to $+180°$. This means that it is elevation rather than azimuth that distinguishes front from back"* [2].

[2] An average radius for an adult is 8.75 cm [7]

[3] Woodworth model formulas

[4] No deeper analysis is made here, as it goes beyond the scope of this project

[5] see Appendix A for calculation details

1.3 Room Model and Sound Level Distance Damping

One of the major challenges of binaural 3D sound is to produce an externalization effect, i.e. that the sound comes from outside the head, as opposed to lateralization (in-head localization of sound images). Convolving a sound with a HRIR or filtering it with a model is not sufficient to obtain an external sound source at a specific direction and distance.

One of the key ingredients of externalization effect is room reflections and reverberation associated with the distance sound damping effect. One technique of room modeling is to convolve the sound with a room impulse response (RIR). However, RIR are quite sizable and too costly computationally speaking. Filter based reverberation effects are most likely to be used in real-time applications.

As the VR environment presented in the current work is orchestrated in an open air space, only the ground would reflect the sound. Therefore, the basic reverberation model of Brown and Duda is implemented: a single early reflection added to the output of the two channels, composed of a 15 ms delayed signal of the monaural input with a gain reduction of 15 dB.

The sound level distance damping is modelled by the following equation:

$$L_2 = L_{ref} - \left| 20log\left(\frac{d_{ref}}{d_2}\right) \right|, \tag{6}$$

expressed in dB and where $L_{ref} = 95$ dB and $d_{ref} = 1.4$ m. [6]

Note also that the visual cues strongly help sound localization and distance estimation.

2 Implementation

The implementation was done using *Max 7* for the Audio engine and *Unity 5* for the Visuals. Fig. 4 shows the flowchart connecting all of the main elements of the implementation.

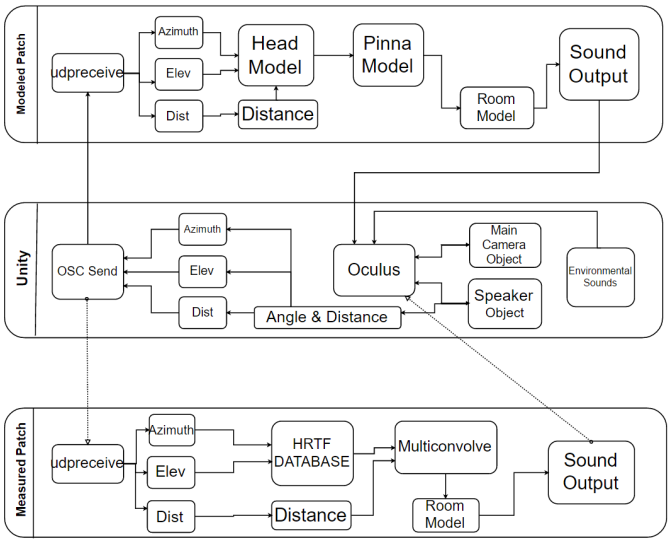


Fig. 4. Implementation overview. The Audio engine must be chosen between the 2 patches.

2.1 Audio Engine

Two main *Max* patches were created: one for the model design and another for the measured HRTF database. Both render the 3D Audio for the environment[7].

The patches both receive, from Unity, the azimuth and elevation angles in interaural-vertical coordinates system, as well as the source-listener distance. More details will follow on how *Unity* and *Max* communicate.

2.1.1 The Model Patch

In the main patch, the input signal is filtered through a series of subpatches which represent the implementation of the Brown and Duda [2] model. Azimuth, elevation and distance were received from Unity using the [udpreceive] object.

In real physics, the sound would propagate through air, travelling a certain distance to finally be perceived in the human auditory system. The same principle was followed in the patch. The input signal was first attenuated using the source-listener distance, and then filtered in the subpatches 'head model' and 'pinna model'. The convolution was done using biquad filters and multiplying them with the delay value for each ear - left and right. (See figure 5)

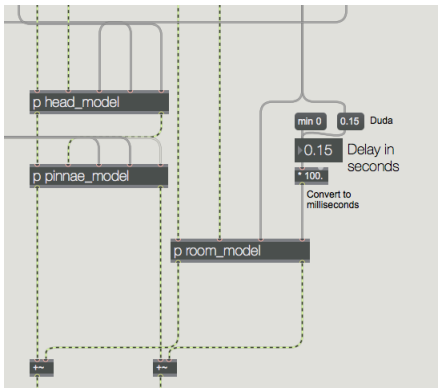


Fig. 5. Modelled HRTF implementation.

In the pinna, the most important computation was the elevation angle convolution. For this process, the input signal was adjusted by manipulating reflection signal gain and time-delay coefficients, this for each ear. The values for this were attributed accordingly to Brown and Duda [2].

2.1.2 The Measured Patch

The use of the MIT KEMAR database implied that the patch needed to access a set of 368 impulse responses to be convolved in real-time. Each convolution represented a specific angle from 0° to 180° in terms of azimuth and −40° to 90° for elevation. All of the stereo HRIR Wave files were stored in a memory buffer (see Fig. 6).

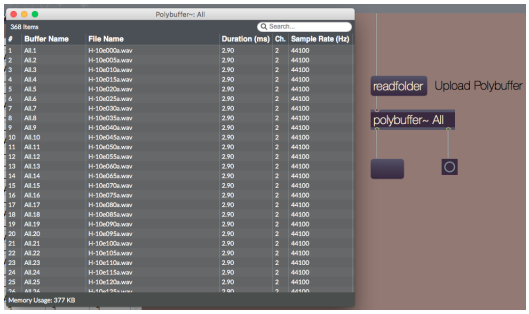


Fig. 6. Polybuffer object stores all HRIRs.

The correct file is called from the translation of the audio file name to its correlated buffer name. This was achieved using two synchronized string matrices. The first one containing all the file names for each possible combination of azimuth and elevation called the corresponding buffer name from the second table. The final message would set the

[6] Based on the MIT KEMAR database measurements.

[7] The background sounds and the footsteps are exceptions

correct file to be chosen for convolution. In order to avoid calling an nonexistent buffer, empty values were replaced with '-1', as in Max this equals to recalling nothing and thus, retaining the previous HRIR for convolution. (See figure 7)

For the convolution, the HISSTools multiconvolve external Max object was used for its low latency in real-time convolution (See figure 8). The use of synchronized tables, that manipulate only string values and the loading of the necessary impulse response helped in the efficiency and speed of the patch.

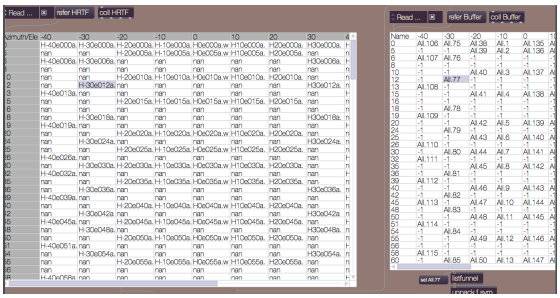


Fig. 7. Jit.cellblock synchronized tables.

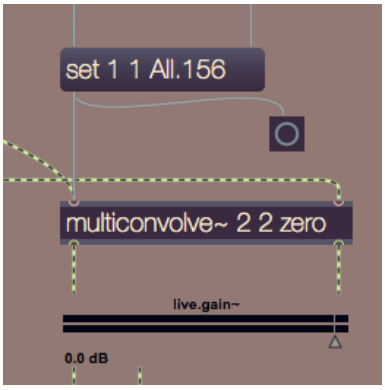


Fig. 8. HISSTools Multiconvolve.

2.2 Virtual Reality Engine

Four different scenes were created in Unity which represented the same outdoor environment with slight differences between them. All the necessary scripts were made in C#. The environment consisted of an almost plain terrain covered by grass and the different sources were represented by speakers mounted on drones. An outdoor environment was chosen to reduce the effect of sound reflections, making it easier to distinguish where the sound was coming from. The first scene presented only one static source and the second scene, sixteen sources spatially distributed in a dome shape. Sample screenshots of these scenes are shown in Fig. 9 and Fig. 10. The third and fourth scenes were a replica of the first two, but with moving sources instead of static.

In order to improve the immersiveness and overall realism of the environment, two different background sounds were included in the scenes; a field recording which consisted mainly of bird noise and light wind and a sound effect of a footstep on grass. These sounds were both pure stereo, i.e. non-3D sound.

2.3 Interconnection between Unity and Max

The communication between *Max* and *Unity* was made using the UDP local subnet. The three pieces of information that the audio engine needs from *Unity*, i.e. azimuth, elevation and distance, are encapsulated in Open Sound Control (OSC) messages and sent to a specific local port number[8].

[8]C# Script: Open Sound Control interface for the Unity3d game engine [8]

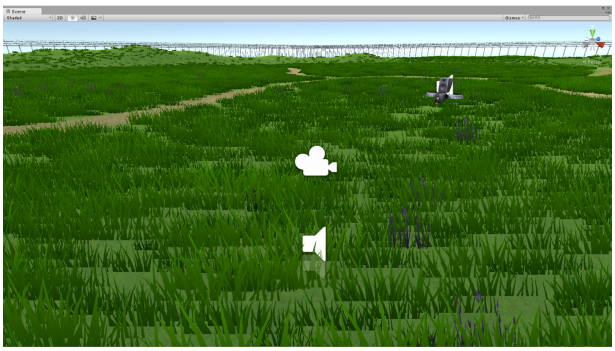


Fig. 9. Outdoor environment with one source.

Fig. 10. Outdoor environment with sixteen sources in dome shape formation.

The distance is the Euclidean distance obtained by the position values from Unity coordinate system. The azimuth and elevation angles are a combination of two angles. The first angle is the angle between the source-listener distance vector and the vector defined by the listener looking at the camera directly in front of them. The second angle is the rotation angle obtained from the head tracking system according to its local coordinate system, which is stored in Unity. The algorithm for obtaining both angles is shown in Appendix B.

3 The Experiment

The experiment was a 20 minutes A/B comparison between the two implementations in 4 different conditions: a static source/static listener; a moving source/ static listener; static source/moving listener and moving source and listener. In the first and third setups, participants were asked to rotate their heads horizontally or vertically in isolation and understand their perception of localization of the sound source in the space. A 5-point Likert scale was used for these questions to determine how accurate the subjects perceived the distance attenuation, and horizontal and vertical rotation to be. The second and fourth setup, which follow respectively setups one and three, were designed as a game where the listener was visually presented with sixteen speaker objects . The task was to locate which color and shape (on top of the speaker object) the sound was coming from. Thirty seconds was given for this task, with a short additional period (a few seconds) where the participants were asked to give their answer or make a guess if they were still not sure. Again, these exercises would enable a comparison of sound sound localization accuracy of both algorithm.

These tasks and questions were repeated four times for the modelled and the measured patch with both static and moving sources after a short training period to help the participants familiarize themselves with the VR system (some participants had never tried before) and the environment. Finally, some more overall and specific questions were asked about the experiment on the whole and desired future developments in the field. Given the extended period of testing, the audio source was selected to have a wider frequency range but still

pleasing or less tiring effect to the participants' ears.

Overall, fifteen participants were tested, with one being used as a pilot test to ensure the experiment could be run efficiently and successfully. This participant was excluded from the final results. Of the participants, six were female (42.9%) and nine were male (57.1%). The participants were also asked whether they suffered from motion sickness, photo-sensitive epilepsy or hearing deficiencies before starting the test to ensure that there were not any health risks to anyone involved. Convenience sampling was used to obtain subjects for testing who would be available at short notice and would not have to travel too far to get to the location of the experiment. Therefore, the data collected cannot be generalised, especially as seven participants were studying Sound and Music Computing and three were studying Medialogy at Aalborg University (Copenhagen). In this case they may have had more experience with 3D audio or VR and have some kind of bias [1].

A mixed method was used during the experiment in order to obtain both qualitative and quantitative data. Specifically, the method used could be identified as a more segmented version of an explanatory sequential mixed method (quantitative data collected, followed by qualitative and interpretation) [1]. For example there was a solitary qualitative question (only quantitative questions excluding this) at the end of each of the four tests (i.e. modelled - static/moving and measured - static/moving). More qualitative questions were then asked at the end of the experiment where participants were encouraged to give more in-depth answers, regarding both overall and specific parts of the experiment.

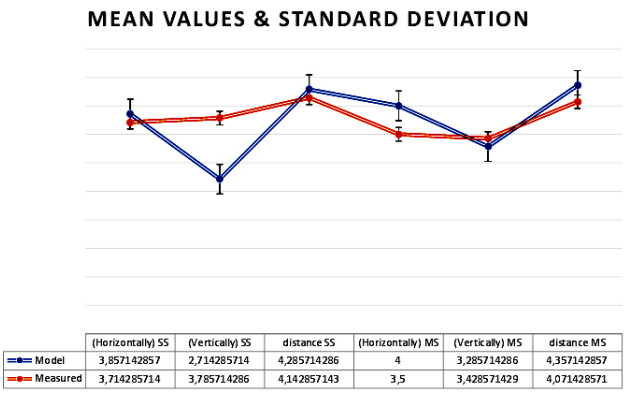
The stimuli were presented in reverse order to check that the participant's familiarity with the system did not influence their responses over time, for example concerning their perception of the accuracy of the rotation of their head in relation to the sound source.

4 Analysis & Interpretation of Results

4.1 Quantitative Data

4.1.1 Perceptive Accuracy of Azimuth/Elevation/Distance in Relation to Sound Sources

Firstly, the data from the questions about accuracy of perception of sound sources (static source, moving source) for both the modelled and the measured patches was arranged in order to compare them: Likert scales (5-point: 1 - strongly disagree, 2 - disagree, 3 - neither agree nor disagree, 4 - agree and 5 - strongly agree) were used to gather quantitative data and the qualitative questions were put in sections - horizontal, vertical and distance. Later the mean values and standard deviations were calculated for all cases, as shown in Fig. 11.



	(Horizontally) SS	(Vertically) SS	distance SS	(Horizontally) MS	(Vertically) MS	distance MS
Model	3,857142857	2,714285714	4,285714286	4	3,285714286	4,357142857
Measured	3,714285714	3,785714286	4,142857143	3,5	3,428571429	4,071428571

Fig. 11. Accuracy of perception of sound sources mean values with standard deviation.

The method used to check the distribution of the results was the two-way ANOVA test. The null hypothesis H_0 refers to a normal distribution of the data, which is explained as if the mean values were equal. Applying the two-way ANOVA test on each combination, enabled the samples to be grouped by one factor, this was done in order to check the relationship between modelled and measured HRTFs (and also the comparison between the different answers concerning horizontal and vertical rotation, and distance).

The results were as follows: By choosing a 95% confidence level, we accepted the null hypothesis between samples of the two methods (modelled/measured). That would infer that there are no statistically significant differences between the results, in another words, there was not a preferred method between the two (modelled/measured). The same regards for the H_0 of the interaction between the two methods, it must be accepted. This implies that there was no independence in the results. As for the experiment source types (static/moving), the null hypothesis was rejected, which infers no relation between them.

Fig. 11 shows that the distance for both methods in both sources is equally rated. Also, in the horizontal sound perception for static source, there was no preferences, but with a moving source, the model had better ratings, though statically insignificant. Vertically, the static sources were rated slightly better in the measured method over the modelled. On the other hand, the moving sources showed no difference for both methods. Nonetheless, there was no statistical significance between these preferences.

4.1.2 Localization

The number subjects was 14 for both methods (modelled/measured) with two different kinds of sources (static/moving), so in total there were 56 combinations. The analysis of the results from the localization test was done by counting the amount of correct answers for the sound source in the environment. Fig. 12 shows the number of correct answers for each method and sources type.

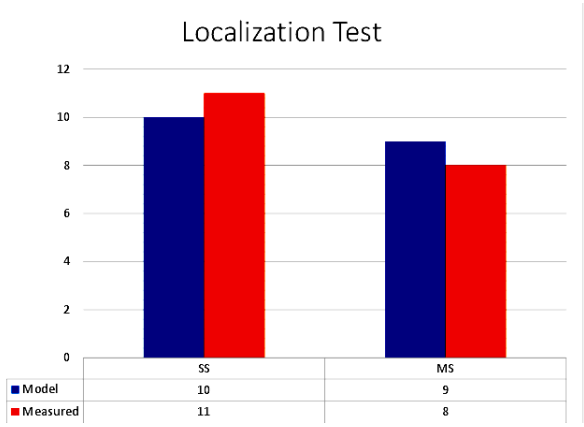


Fig. 12. Localization test results.

A Chi-square test was applied in this situation by assuming that there is normality in the distribution. Thus, with 1 degree of freedom for both sources, the chi-square statistic is 0.106. The P-value is 0.74. This result is not significant at p less than 0.05.

Therefore, the p-value was higher than the error level, this infers that the subjects had no statistically significant different choices.

From Fig. 12, it can be interpreted that subjects were more accurate in their answers for the measured method over the model in the case of static source. On the other hand, in the moving source, the model was easier to distinguish over the measured. Nevertheless, there was no statistically significant preference for neither of them.

4.1.3 Sense of Presence

The mean values were computed from the Likert scale (1-5) data concerning the sense of presence test in the game for finding the correct source in the environment for the two methods (modelled/measured) with both sources types (static/moving). Results are shown in Fig.13.

To test the significance of the results, a one-tail T-test with an assumption of equal variances was done. The alpha (level of error) is

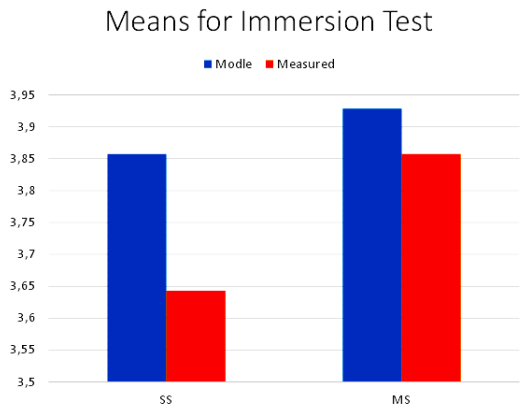


Fig. 13. The mean values of the immersion test.

0.05 and H_0 (the null hypothesis) was that there were no differences between the mean values. As for static source, the T statistic -0.249 is less than the T-critical value 1.771 and the p-value is 0.806. The moving source T-statistic was 0.322 and T-critical value 1.771, p-value 0.752, which was greater than alpha. This implies that H_0 cannot be rejected. Therefore, there was no statistically significant preferences.

Fig. 13 shows that the modelled method received a better rating from the subjects in both cases (for static and moving sources). However, it cannot be considered as a statistically significant preference.

4.1.4 Auditory Sensation of Moving Objects

The same procedure was followed as in the previous data analysis, where the data was also on a Likert scale (1-5). The mean values were computed (Fig. 14) and later a T-test was done to check for any statistical significant difference between the two methods (model/measured). The T-statistic -0.520 was less than the T-critical value 1.771 and the P-value is 0.611, which led to acceptance of the H_0, meaning that there was no preference between methods. Fig. 14 shows that measured method had a better auditory sense of objects moving through the environment, but again it is not a statistically significant preference.

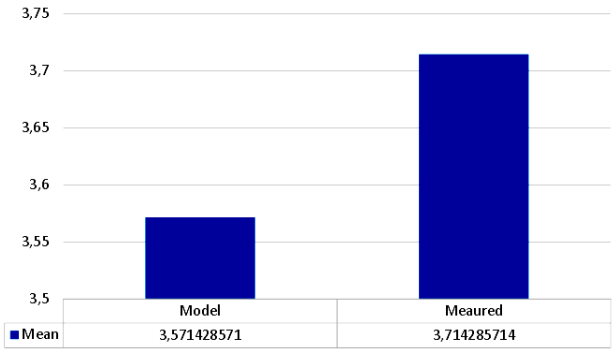


Fig. 14. The mean values of the auditory sensation of moving objects test.

4.1.5 Overall Preferences

The subjects were asked about in which set of tasks they perceived the audio to be more faithful to a realistic behaviour, corresponding each set to a different HRTF method. Participants could also reply that both sets were perceived equally accurate or that they didn't know. With this data, a percentage of preferences for each method was obtained. These results are shown in Fig. 15, where 53.85% of the subjects preferred the modelled method against 46.15% who had chosen the measured.

The order of the methods was alternated for different subjects to reduce the influence of training progression. In order to know how much the training effect influenced the participant's judgement (regardless of which method was presented to them first), the percentages for the presentation order preference show, in Fig. 16, that 25% of the subjects preferred the first set of tasks and 75% of them preferred the second set.

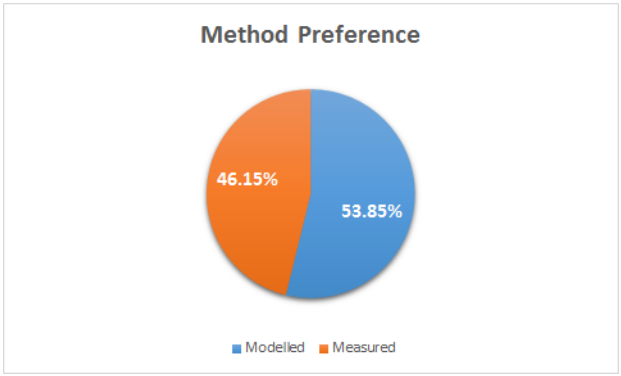


Fig. 15. Method preference results.

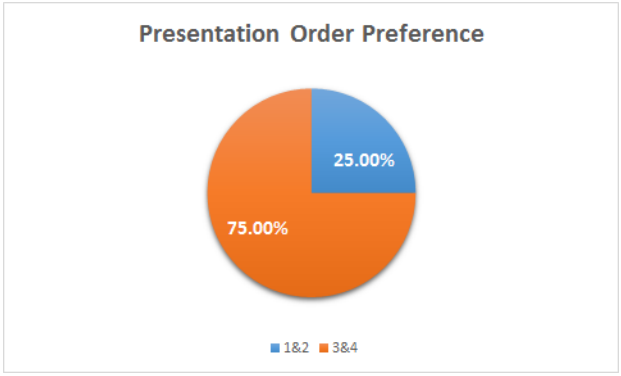


Fig. 16. Experiment presentation order preference results.

4.2 Qualitative Data: Involvement Feedback

The qualitative data gathered during the experiment differed between positive and negative comments, but the more in-depth feedback and justifications for scores given for qualitative questions, were invaluable. For the modelled patch with a static source, subjects mentioned that their sense of involvement within the scene was increased by the presence of the auditory information and that it added another dimension, while some participants mentioned that it sounded like the source was coming from two places at once and that it was easier to navigate based on distance rather than azimuth or elevation.

One interesting observation during this question was that one subject felt that they were more trained and hence sound localization and perception felt more accurate, but that this may well have increased over time. It was for this reason that the order, of the modelled and measured patches presented to the subjects, was in reverse order for half of the participants. In this participant's case the order of presentation was in reverse, so the observation was very relevant.

For the modelled patch with a moving source, the responses regarding involvement were again quite varied. One interesting response was that there was more consistency during the game task when the subject was standing outside of the circle. Another participants perceive the sound source as if it was a real speaker rotating around them and this reduced the overall feeling of involvement in the environment.

For the measured patch the most constructive piece of feedback mentioned was that the footsteps increased the feeling of involvement in the

scene and made them feel like an actual character walking around. Although, another subject conveyed that the footsteps were not perceived as being on the floor. However, the footstep sound effect only used stereo sound, so no elevation information was received.

Also, with the moving source, one participant mentioned the sense of involvement to be definitely better, while another subject said that the audio from the speaker did not seem as static as in the other examples. There was one subject who also discussed that the elevation and panning was easier to pinpoint with this stimuli. A few subjects did comment however that there were some glitches in terms of where they were perceiving the sound source to be and thus they were relying more on distance than rotation for sound source localization.

5 Conclusion

In this study, two different 3D sound rendering algorithms for a Virtual Reality environment have been presented. The overall goal was to understand the capabilities of alternative designs for audio engines. The comparison was made between a mathematically modelled HRTF and impulse response databased convolution. For a real-time experiment, the study was conducted with Max 7, for the audio, and Unity for the graphical rendering.

None of results evaluated show a statistically significant difference. Therefore, it to be concluded that there is no preference between the two methods regarding accuracy of perception when rotating the head or moving through the environment. Nevertheless, there is a slight preference for the model method in the case of azimuth and moving source, and for the measured method in elevation and static source.There is a preference for the measured method in relation to sound localization, and an overall preference for the modelled method with static and moving sources.

In terms of sense of presence the modelled implementation seems to be the preferred one, and for the auditory sensation of moving objects, the overall preference insights to the measured method.

5.1 Evaluation and Efficiency of the Experiment

The experiment design was confined to a single format. For this reason, further research is highly recommended. The use of an audiovisual setup definitely enhanced the sensation of spatialization and the illusion of immersion in the VR but to achieve a clear and analytical conclusion on the comparison between these two implementations, it is also interesting to evaluate the experiment without the visual cues. Further development to integrate the audio implementation in Unity would also enable to test the participants with less interruptions between each setup and possible interactive data collection (interactive questionnaires inside the VR experience) could bring more insights on the subjects' evaluations of both methods. For example, direct comparison (i.e. presentation of each of the patches with static and moving sources one after another) may have made it easier for participants to compare and differentiate the methods used during the experiment. When asked which set of stimuli they preferred (first or second in terms of presentation order), 75% of participants (see Figure 16) said that they perceived the accuracy to be better in the second set. This supports the need for a longer training period as this did not correlate with the results from the sound localization test or questions regarding the participants' perception of horizontal/vertical rotation and distance (when asked in turn).

The validity of the experiment was affected by the use of convenience sampling, thus generalisations cannot be made as to the the wider population. Also, it cannot be generalised based on the number of participants as there were only fourteen, so this fact also affects the validity of the findings.

References

[1] T. Bjørner. *Qualitative methods for Consumer Research*. Reitzel, Hans.

[2] C. Brown and R. Duda. An efficient HRTF model for 3-D sound. In M. M. House, editor, *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, volume 6, pages 476–488, October 1997.

[3] B. Gardner and K. Martin. The MIT Media Lab: HRTF Measurements of a KEMAR Dummy-Head Microphone.

[4] A. Harker and P. A. Tremblay. The hisstools impulse response toolbox: Convolution for the masses. *ICMC 2012: Non-cochlear Sound, The International Computer Music Association*.

[5] J. Huopaniemi and M. Karjalainen. Comparison of Digital Filter Design Methods for 3D Sound.

[6] T. C. Interface Laboratory. Spatial Sound.

[7] G. F. Kuhn. Model for interaural time differences in the azimuthal plane. Technical report, Institute for Basic Standards, National Bureau of Standards, Washington, D.C.

[8] J. G. Martn. Open Sound Control interface for the Unity3d game engine.

[9] G. M. Spagnol, S. and F. Avanzini. Structural modeling of pinna-related transfer functions. In *Proc. 7th Int. Conf. Sound and Music Computing (SMC 2010)*, pages 422–428, 2010.

[10] Ustwogames. Designing Sound for Virtual Reality, December 2015.

A Derivation of Equation (4)

A bilinear transform is applied to an analog transfer function in order to obtain a discrete-time transfer function of it. It is defined as follow:

$$H(z) = H(s)\big|_{s=\frac{2(z-1)}{T(z+1)}}.$$

Therefore, equation (3) becomes:
Let $\alpha(\theta) = \alpha$,

$$\begin{aligned} H(z,\theta) &= \frac{\alpha\frac{2}{T}\frac{z-1}{z+1}+\beta}{\frac{2}{T}\frac{z-1}{z+1}+\beta} \\ &= \frac{\frac{2\alpha z-2\alpha}{Tz+T}+\beta}{\frac{2z-2}{Tz+T}+\beta} \\ &= \frac{\frac{2\alpha z-2\alpha+\beta(Tz+T)}{Tz+T}}{\frac{2z-2+\beta(Tz+T)}{Tz+T}} \\ &= \frac{2\alpha z-2\alpha+\beta Tz+\beta T}{2z-2+\beta Tz+\beta T} \\ &= \frac{(2\alpha+\beta T)z+()\beta T-2\alpha)}{(2+\beta T)z+(\beta T-2)}\frac{z^{-1}}{z^{-1}} \\ &= \frac{(2\alpha+\beta)+(\beta T-2\alpha)z^{-1}}{(2+\beta T)+(\beta T-2)z^{-1}}. \end{aligned} \tag{7}$$

T is the sampling period in seconds.

B Azimuth and Elevation Angles

In this appendix, the algorithm for the calculation of the azimuth and elevation angles in Unity is presented. The azimuth and elevation angles between the source and the listener are not straightforwardly given from Unity. They have to be derived, by simple trigonometric manipulations, from the 'Transform' (transform.position and transform.eulerAngles) of the Unity sound source object and main camera object. The azimuth is obtained by calculating the angle between the source and the listener in the horizontal plane, and the same for elevation in the vertical plane.

First, let us consider the XY plane. Let $O(0,0)$ be the origin and $S_0(x,y)$ the sound source located in front of $L(x,y)$, the listener, so that the direction vector listener-source has the direction of y-axis. Let α be the angle if the source moves to the right (positive angle) or to the left (negative angle), as shown in Fig. 17. α always refers to the angle between vertical line and the actual position of the source.

The simple inequality $S_{L,R,x} - L_x < 0$ gave information as to whether the source has moved to the right or to the left: if this holds true, the source is on the left of the listener, therefore α is negative; otherwise, it is on the right and α is positive. α is simply obtained as follow:

$$\alpha = \arccos\left(\frac{|L_y - S_{1,y}|}{d_{S_1,L}}\right), \tag{8}$$

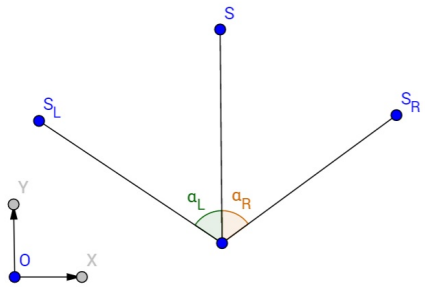


Fig. 17. Source moves to the right or to the left in the horizontal plane. α_L and α_R are the angles if the source moves to the left or to the right, respectively.

if the source moves to the right and $-\alpha$ if it moves to the left. $d_{S_1,L}$ being the Euclidean distance between the listener and the new position of the source.

In Unity, the horizontal plane is the XZ plane and the third dimension is given by Y. For the azimuth, we consider the rotation of the head around the Y axis, and for the elevation the rotation of the head around the X axis. The script receives these angle rotations, say ϕ_y and ϕ_x, in the interval $[0;360[$. The azimuth θ is found by adding α to ϕ_y and idem for the elevation. Note that the Audio engine expects to receive the angles as positive ($[0;180]$) for the right side and negative ($[-180;0]$) for the left side. The following **Algorithm 1** for the azimuth was implemented in the C# script attached to the main camera object in Unity:

Algorithm 1 Find Azimuth θ

Inputs:
$L(x,y,z)$, $S(x,y,z)$, ϕ_y
Initialize:
$\phi_y \leftarrow -\phi_y$, $\alpha \leftarrow \arccos\left(\frac{|L_z - S_z|}{d_{xz}(S,L)}\right)$
if $S_{1,x} - L_x < 0$ **then**
    $\alpha \leftarrow -\alpha$
end if
$\theta \leftarrow \alpha + \phi_y$
if $|\theta| > 180°$ **then**
    $\theta \leftarrow 360° - |\theta|$
end if
Output: θ

The same algorithm is used to calculate the elevation by changing the parameters.

C Max 7 External Object

For the purpose of real-time convolution, there was no in-built object in Max that could manage the appropriate efficiency and minimal latency required for the Measured HRTF patch. The HIRT External library developed by Alexander Harker and Pierre Alexandre Tremblay [4] offered a set of tools that can address this problem in a variety of context. For the purpose of this research, the multiconvolve object was the best choice.

The multiconvolve is a max object designed for zero latency convolution. This object is a combination of time domain convolution for the early portion of an impulse response (IR) with the FFT-based partitioned convolution - technique for efficiently performing time domain convolution with low inherent latency - for the latter part of the IR. The output is addressed for both multichannel loudspeaker system and binaural setup.

The current implementation uses the convolution matrix mode where the object invokes a matrix of convolvers. Each input-output path can have an individual IR, set independently from the polybuffer. This way, it was possible to invert the stereo channels of each KEMAR's IRs for angles between 180 and 360 degrees, see figure 18.

There were three special cases. The angles of 0, 180 and 360 degrees imply the convolution of both channels equally, otherwise, the auditory space would be uncorrelated with the motion perceived in Unity.

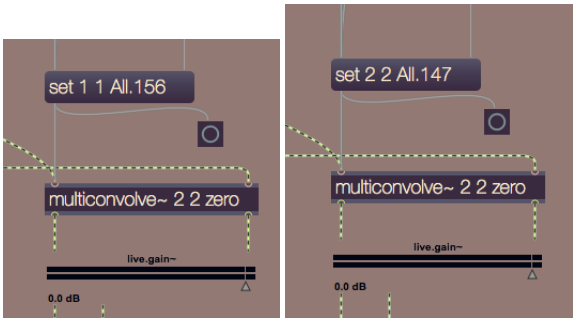


Fig. 18. Multiconvolve message set for convolution below $-180°$, on the left and above $-180°$ on the right.