VIRTUAL ACOUSTIC RECORDING: AN INTERACTIVE APPROACH

Marcin Gorzel, Gavin Kearney, Frank Boland and Henry Rice

School of Engineering
Trinity College Dublin
Ireland
gorzelm@tcd.ie, gavin.kearney@tcd.ie

ABSTRACT

In this paper, we present a framework for recording real musical auditory scenes for interactive virtual acoustic reproduction over headphones. The framework considers the parameterization of real-world soundfields and subsequent real-time auralization using a hybrid image source method/measurement-based auralization approach. First Order (FOA) and Higher Order (HOA) Ambisonics are utilized together in a single system to provide an optimized and psychoacoustically justified framework.

1. INTRODUCTION

Virtual acoustic recording refers to the capture of real-world acoustic performances in reverberant spaces and their subsequent plausible reproduction in a virtual version of the original performance space, otherwise known as a Virtual Auditory Environment (VAE). An important aspect in the quest for realism in such auditory scene synthesis is *user interaction*. That is, how the movements of a person listening to the virtual auditory scene directly influences the scene presentation. Such 'walkthrough auralization' presents several challenges for production engineers, the most significant of which is the generation of the correct room acoustic response due to a given source-listener position. In particular, the correct direction of arrival of the direct sound and early reflections must be maintained since these signals contain the most vital cues for localization of acoustic sources.

In recent years, the formation of auditory scenes based on real world spaces has benefited greatly from the use of convolution reverberation techniques, and a significant body of work has been presented illustrating the possibilities and limitations, for example in [1], [2] and [3]. The representation of room responses in this manner assumes that the source-room interaction is one that is linearly time-invariant (LTI). In reality, the room impulse response (RIR) changes significantly with the relative spatial positions of the source and the listener. However, if the geometric properties of the performance space are known (as well as the frequency dependent absorptive properties of the materials in the room), then the acoustic response can instead be computed. Highly accurate results can be achieved using wave-based methods, such as the Finite Element Method (FEM), Boundary Element Method (BEM) [4] or Finite Difference Time Domain (FDTD) method [5]. Due to computational expense, such methods are generally limited to low frequency RIR estimation. Geometric-based solutions to calculating RIRs, such as the image-source method [6] are well-suited to the mid-to high frequency regions, although they do not consider phenomena such as diffraction or scattering. However, calculation of the propagation delays and magnitudes at low reflection orders using image sources is well suited to real-time auralization. Hybrid reverberation algorithms have been proposed which combine computational and measured impulse responses but have largely focused on the synthesis of the diffuse decay as opposed to early reflections [7], [8]. In this paper, we focus on the real-time rendering of the early reflections in conjunction with pre-rendered diffuse field recordings for walkthrough auralization.

The reproduction of the auditory scene is also highly dependent on the spatialization method employed. Many techniques have been proposed in the literature, most notably Vector Based Amplitude Panning (VBAP) [9] and Wavefield Synthesis [10]. However, Ambisonics [11], which is based on the spherical harmonic decomposition of the soundfield, represents a practical and asymptotically holographic approach to realtime soundfield manipulation.

Thus, in this paper, we outline a framework for the capture and reproduction of real-word musical performances which borrows from both measured acoustics and image source computation, as well as the Ambisonic approach. The paper is outlined as follows: First we will address auralization issues for real-time early reflection synthesis using higher order Ambisonics and the image source method. Then, we address the capture and rendering of the diffuse properties of measured soundfields for Ambisonic representation. Finally we discuss an example implentation of the framework.

2. DIRECT SOURCE ENCODING

Direct-field capture of a musical source can be achieved within the critical distance using a spot microphone. The positioning and directional characteristic of the microphone is important not only to the tonal balance, but also to minimize the amount of other instruments (commonly referred to as 'spill') in the recorded signal. Hypercardioid and supercardioid microphones are frequently used in order to maximize rejection, but the cost of increased directional response can often lead to compromised frequency response in lower grade microphones as well as 'proximity effect'. Such frequency response distortions must be corrected, before encoding the signal into Ambisonics format.

Ambisonics was originally developed by Gerzon, Barton and Fellgett [11] as unified system for the recording, reproduction and transmission of surround sound. The theory of Ambisonics is based on the decomposition of the the soundfield measured at single point in space into spherical harmonic functions defined as

$$Y_{mn}^{\sigma}(\Phi,\Theta) = A_{mn}P_{mn}(\sin\Theta) \times \begin{cases} \cos m\Phi & \text{if } \sigma = +1 \\ \sin m\Phi & \text{if } \sigma = -1 \end{cases} \quad (1)$$

where m is the order and n is the degree of the spherical harmonic

and P_{mn} is the associated Legendre function. For each order m there are $(2m+1)$ spherical harmonics. In order for plane wave representation over a loudspeaker array we must ensure that

$$s\, Y^{\sigma}_{mn}(\Phi,\Theta) = \sum_{i=1}^{I} g_i\, Y^{\sigma}_{mn}(\phi_i,\theta_i) \tag{2}$$

where s is the pressure of the source signal from direction (Φ,Θ) and g_i is the i^{th} loudspeaker gain from direction (ϕ_i,θ_i). We can then express the left hand side of the equation in vector notation, giving the Ambisonic channels

$$\mathbf{B}_{\mathbf{\Phi\Theta}} = \mathbf{Y}_{\mathbf{\Phi\Theta}} s \tag{3}$$

$$= [Y^1_{0,0}(\Phi,\Theta), Y^1_{1,0}(\Phi,\Theta), ...Y^{\sigma}_{mm}(\Phi,\Theta)]^T s \tag{4}$$

Equation 2 can then be rewritten as

$$\mathbf{B} = \mathbf{C}\cdot\mathbf{g} \tag{5}$$

where $\mathbf{C}$ are the encoding gains associated with the loudspeaker positions and $\mathbf{g}$ is the loudspeaker signal vector. In order to obtain $\mathbf{g}$, we require a decode matrix, $\mathbf{D}$, which is the inverse of $\mathbf{C}$. However, to invert $\mathbf{C}$ we need the matrix to be a square which is only possible when the number of Ambisonic channels is equal to the number of loudspeakers. When the number of loudspeaker channels is greater than the number of Ambisonic channels, which is usually the case, we then obtain the pseudo-inverse of $\mathbf{C}$ where

$$\mathbf{D} = \text{pinv}(\mathbf{C}) = \mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1} \tag{6}$$

Since the soundfield is represented by a spherical coordinate system, soundfield transformation matrices can be used to rotate, tilt and tumble the soundfields. The number of I virtual loudspeakers employed with the Ambisonics approach is dependent on the Ambisonic order m where

$$I \geq (m+1)^2 \tag{7}$$

An important aspect of encoding the direct field recordings is the simulation of source distance. A key element is the introduction of near field compensation filters which cater for the fact that since the reproduction loudspeakers are at a finite radius, there is a low frequency boost due to near-field spherical wave propagation [12]. Consequently near-field compensation filters are required to cater not only for sources outside the array, but also focused sources inside the array.

3. EARLY REFLECTION ENCODING

For a given source-reciever position the magnitude and time-delays of the early reflections can be computed using the image source method. The reflection coefficients should be chosen to give a good approximation of the general absorptive properties of the real room. Lehmann [13] has shown how phase inversion of image sources can lead to more realistic impulse responses, since a given RIR usually displays stochastic noise like properties around a zero mean. Thus, we suggest the use of negative absorption coefficients as a first approximation to the directional portion of RIR over positive only impulse responses. The improvement in RIR approximation over positive only impulse responses, in comparison to real world measurements is shown in Figure 1.

Early reflections can be implemented as time varying tapped delay lines. This approach was first suggested by Schroeder[14]

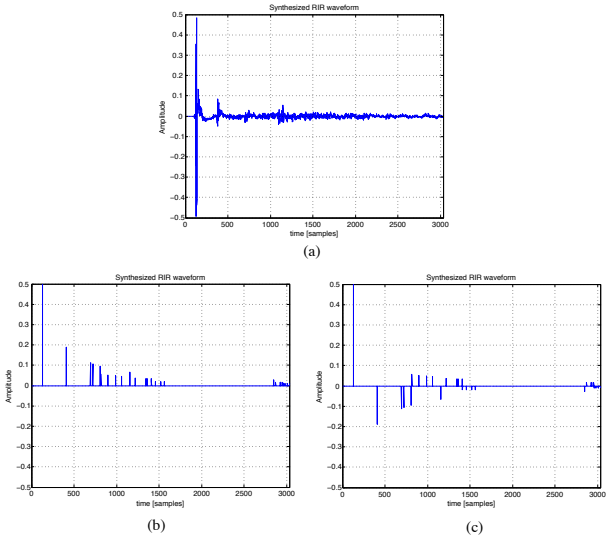


Figure 1: Room impulse response comparison: (a) Measured RIR, (b) Image source RIR (c) Image source RIR with negative reflection coefficients.

and implemented by Moorer [15]. In this way we avoid computationally expensive FIR filter implementation via convolution, but more importantly we are able to represent each reflection in HOA since each delayed version of the signal can be treated as independent. Each respective spherical harmonic component for each delay can be summed up accordingly.

However, for any virtual acoustic recording to be convincing, the directional properties of the source audio and the subsequent effect on the early reflections must also be considered. Several approaches to measuring and synthesizing source directionality have been proposed on the literature, primarily with regard to wave field synthesis reproduction. The capture of source directivity is traditionally achieved using arrays of microphones surrounding a performer and databases of such measurements are available [16]. Directivity filters can then be applied to a single monophonic recording of a performance to simulate the change in frequency response with source/listener movement. A simple approach has been proposed by Giron [17], where the interference of several monopole virtual sources is used to synthesize the directivity of real sources. However, the resulting frequency dependent directivity does not behave like that of real world sources. A further approach is that the array of microphones used to capture the directivity measurements can actually be used to capture the performance (in an anechoic chamber) entirely, and virtual loudspeakers can be synthesized at reproduction using monopoles or virtual cardioids [18]. This is not a practical solution to a real performance situation however.

In this paper we suggest the decomposition of the directional response into spherical harmonics, which has been proposed for computational based auralization in numerous papers most notably by Menzes [19] and Spors [20]. Here, for a given musical instrument, an averaged frequency dependent source directivity measured in an anechoic environment can be encoded into HOA, forming a HOA directional filter. Thus, the recorded direct sound can be pre-rendered before runtime as a directional HOA source. The frequency response of the source audio utilized for each early reflection is therefore dependent on the angle of incidence of that re-

flection. An important aspect of filtering direct-source audio with a HOA directional filter is de-emphasis of the directional responses. Calculation of the directional response functions must not assume that the magnitude spectrum of the recorded musical source audio is flat, and that application of directional filtering will yield the appropriate directional *magnitude* response. A reference source angle should in fact be taken (e.g. the angle of the direct-field microphone to the instrument) and deviations from the recorded source magnitude response used to form the directional filters.

4. DIFFUSE FIELD MEASUREMENT

After the capture of the musical performance, the acoustic response can then be measured in the reverberant space. Unlike previous methods of impulse response capture which focus on preservation of the directional properties of the direct sound and early reflections, here we are only interested in capture (and subsequent extraction) of the diffuse field. For optimal diffuse field capture, the soundfield should be measured at a distance greater than the critical distance of the room using the logarithmic swept sine technique of [21]. An Ambisonic soundfield microphone allows us to capture the pressure (labelled the w channel) and particle velocity (the x, y and z channels) of the soundfield at a single point in space. Processing of the soundfield channels results in impulse responses h_w, h_x, h_y and h_z.

In separating the diffuse field from the measured impulse responses, we adopt the directional analysis method of Pulkki and Merimaa, found in [22]. Here the measured soundfield responses are analyzed in terms of sound intensity and energy in order to derive time-frequency based direction of arrival and diffuseness. The instantaneous intensity vector is given from the pressure p and particle velocity $\mathbf{u}$ as

$$\mathbf{I}(t) = p(t)\,\mathbf{u}(t) \tag{8}$$

Since we are using 1st order Ambisonic impulse response measurements, the pressure can be approximated by

$$p(t) = w(t) \tag{9}$$

and the particle velocity by

$$\mathbf{u}(t) = \frac{1}{\sqrt{2}Z_0}\left(x(t)\mathbf{e}_x + y(t)\mathbf{e}_y + z(t)\mathbf{e}_z\right) \tag{10}$$

where $\mathbf{e}_x$, $\mathbf{e}_y$, and $\mathbf{e}_z$ represent cartesian unit vectors and Z_0 is the characteristic acoustic impedance of air. The instantaneous intensity represents the direction of the energy transfer of the soundfield and the direction of arrival can be determined simply by the opposite direction of $\mathbf{I}$. For 1st order Ambisonics, we can calculate the intensity for each coordinate axis in the frequency domain. Since a portion of the energy will also oscillate locally, a diffuseness estimate can be made which is given by the ratio of the magnitude of the intensity vector to the overall energy density given as

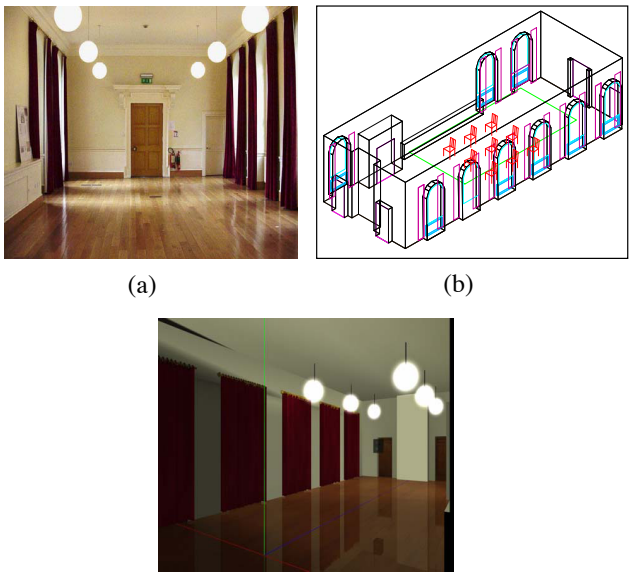
$$\psi = 1 - \frac{||\langle\mathbf{I}\rangle||}{c\langle E\rangle} \tag{11}$$

where $\langle\cdot\rangle$ denotes time averaging and $||\cdot||$ denotes the norm of the vector. The diffuseness estimate will yield a value of zero for incident plane waves from a particular direction, but will give a value of 1 where there is no net transport of acoustic energy, such as in the cases of reverberation or standing waves. Time averaging is used since it is difficult to determine an instantaneous measure of diffuseness.

The output of the analysis is then subject to smoothing based on the Equivalent Rectangular Bandwidth (ERB) scale, such that the resolution of the human auditory system is approximated. The resultant Ambisonic signals are then weighted in each frequency band i according to $\sqrt{\psi_i}$ and a first order decode is then formed. This is justified since it is only vital that the main directional information encoded to higher order. Furthermore, if there exists a general directional distribution to the diffuse field, this will still be preserved in first order form.

5. EXAMPLE IMPLEMENTATION

An example implementation was created by recording a performance of three musicians in an existing small reverberant hall in Trinity College Dublin. The ensemble consisted of vocal, guitar and violin, each recorded in the direct field with unidirectional (supercardioid) microphones. A CAD model of the hall was implemented and imported into the Blender 3D open-source environment [23]. Virtual sound sources were then implemented at matching positions to the real-world situation. Blender was used not only for visual representation of the virtual environment (through adding details to the CAD model such as texturing, lighting, etc.) but primarily for its built in Blender Game Engine (BGE) which accommodates the use of python scripts [24] to add interactivity to the model. The main functionality implemented using the BGE was the control of the position and orientation of the virtual camera within the room. A python script was written which recalculated correct positions of mirror-image sources (up to 2nd order) at run-time by tracking the location of the virtual camera as well as the sound sources at every logic tick interval. In this way, all the reflections were characterized by four parameters: normalized amplitude (0-1), time of arrival in (ms) and horizontal and vertical angles of incidence (ϕ, θ). The arrays of data containing all four



(a) (b)

(c)

Figure 2: (a) Reverberant Hall. (b) Computational Model (c) Interactive Virtual Environment

parameters updated in real-time were being sent over the UDP network protocol to the audio visual programming environment Pure-Data [25] where all the further audio signal processing was done.

The direct sound and early reflections were encoded into 3rd order Ambisonics. A monophonic mixture of the direct field recordings was convolved with the first order diffuse field and triggered at the same time as the direct sound/early reflection playback.

An additional feature implemented in order to provide a higher level of immersion was the 3-degrees-of-freedom (3DOF) head-tracking for headphone reproduction. An Intersense InertiaCube2+ head-tracker was incorporated to update both visuals and soundfield according to the current head orientation of the user. As the user rolled, tilted and/or yawed their head the virtual camera orientation matrix was updated in real-time according to the head-tracker data. This solution is especially effective and desired when using Head Mounted Displays (HMD) since it provides more immersive experience when the full field of view changes with the head movements. Binaural rendering was implemented using the virtual loudspeaker approach outlined in [26], where HRTFs are measured at the 'sweet-spot' (the limited region in the centre of a reproduction array where an adequate spatial impression is generally guaranteed) in a multi-loudspeaker reproduction setup, and the resultant binaural playback is formed from the convolution of the loudspeaker feeds with the virtual loudspeakers. Using the Ambisonics approach here means that instead of recalculating the angle of arrival for every sound source and every reflection it is much easier to rotate the whole soundfield according to the incoming head-tracking data.

Informal listening tests with several subjects demonstrated that a very good match between the real world soundfield recordings of the performance, and the interactive virtual walkthrough model. Example soundfield renderings of this performance, in comparison to real world walkthroughs can be found at `http://www.mee.tcd.ie/~gkearney/DAFX/DAFX10.html`.

6. CONCLUSIONS

In this paper we have presented a methodology for the capture of real world musical presentations for playback in interactive VAEs. The methodology utilizes HOA encoding of the direct field recordings and the incorporation of source directivity. Diffuse field synthesis is achieved by convolving a mono mix of all direct-field recordings with the extracted diffuse field from the measured spatial impulse response. Further work will focus on subjective assessment of the model as well as higher order reflection synthesis.

7. ACKNOWLEDGMENTS

This work was supported by Science Foundation Ireland.

8. REFERENCES

[1] V. Pulkki and J Merimaa, "Spatial Impulse Response Rendering I: Analysis and synthesis," *Journal of the Audio Engineering Society*, vol. 53, 2005.

[2] Ralph Kessler, "An optimised method for capturing multidimensional acoustic fingerprints," in *118th Convention of the Audio Engineering Society*, May 2005.

[3] G. Kearney and J. Levison, "Actual vs. virtual multichannel acoustic recording," in *134th Convention of the Audio Engineering Society*, May 2008.

[4] Andrzej Pietrzyk, "Computer modeling of the sound field in small rooms," in *15th International Conference of the Audio Engineering Society*, 1998.

[5] L. Savioja, T. Rinne, and T. Takala, "Simulation of room acoustics with a 3-D finite difference mesh," in *International Computer Music Conference*, 1994, pp. 463–466.

[6] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *JASA*, vol. 65, pp. 943–950, 1979.

[7] S. Browne, "Hybrid reverberation algorithm using truncated impulse response convolution and recursive filtering," 2001, `http://www.music.miami.edu/programs/mue/mue2003/research/sbrowne/`, University of Miami, FL, USA.

[8] R. Stewart and D. Murphy, "A hybrid artificial reverberation algorithm," in *Proc. of the 122nd AES Convention, Vienna, Austria*, 2007.

[9] V. Pulkki, "Virtual sound source positioning using Vector Base Amplitude Panning," *Journal of the Audio Engineering Society*, vol. 45, pp. 456–466, 1997.

[10] A. J. Berkhout, "A Holographic Approach to Acoustic Control," *Journal of the Audio Engineering Society*, vol. 36, pp. 977–995, 1988.

[11] M. A. Gerzon, "Periphony: With-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, pp. 2–10, 1973.

[12] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new Ambisonic format," in *23rd International Conference of the Audio Engineering Society*, 2003.

[13] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, July 2008.

[14] M. R. Schroeder, "Digital simulation of sound transmission in reverberant spaces," *The Journal of the Acoustical Society of America*, vol. 47, no. 2A, pp. 424–431, 1970.

[15] James A. Moorer, "About this Reverberation Business," *Computer Music Journal*, vol. 3, no. 2, pp. 13–28, 1979.

[16] Physikalisch-Technische-Bundesanstalt, "Directivities of musical instruments," 2009, `http://www.ptb.de/en/org/1/17/173/richtchar.htm`, accessed Jun. 1, 2009.

[17] F. Giron, "Investigations about the directivity of sound sources," 1996, PhD Thesis, Ruhr-Universität, Bochum, Shaker Verlag.

[18] R. Jacques, B. Albrecht, F. Melchior, and D. de Vries, "An approach for multichannel recording and reproduction of sound source directivity," in *119th convention of the Audio Engineering Society*, 2005.

[19] D. Menzies, "Nearfield synthesis of complex sources with High Order Ambisonics, and binaural rendering," in *Proceedings of the 13th International Conference on Auditory Display*, June 2007.

[20] J. Ahrens and S. Spors, "Implementation of directional sources in Wave Field Synthesis," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, June 2007.

[21] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *108th Convention of the Audio Engineering Society*, 2000.

[22] J. Merimaa and V. Pulkki, "Spatial impulse response rendering i: Analysis and synthesis," *Journal of the Audio Engineering Society*, vol. 53, 2005.

[23] Blender Foundation, "Blender," `http://www.blender.org/`, acessed 5th April, 2010.

[24] Python Software Foundation, "Python programming language," `http://www.python.org/`, acessed 5th April, 2010.

[25] Miller Puckette, "Pure data," 2009, `http://puredata.info/`.

[26] A. McKeag and D. McGrath, "Sound field format to binaural decoder with head-tracking," in *6th Australian Regional Convention of the AES*, 1996.