A 3D REAL TIME RENDERING ENGINE FOR BINAURAL SOUND REPRODUCTION

Markus Noisternig, Thomas Musil, Alois Sontacchi, Robert Höldrich

Institute of Electronic Music and Acoustics
University of Music and Dramatic Arts Graz
Inffeldgasse 10/3
A-8010 Graz, AUSTRIA
noisternig@iem.at

ABSTRACT

A method of computationally efficient 3D sound reproduction via headphones is presented using a virtual Ambisonic approach. Previous studies have shown that incorporating head tracking as well as room simulation is important to improve sound source localization capabilities. The simulation of virtual acoustic space requires to filter the stimuli with head related transfer functions (HRTFs). In time-varying systems this yields the problem of high quality interpolation between different HRTFs. The proposed model states that encoding signals into Ambisonic domain results in time-invariant HRTF filters.

The proposed system is implemented on a usual notebook using Pure Data (PD), a graphically based open source real time computer music software.

1. INTRODUCTION

The following paper deals with the theory and practice of 3D sound reproduction using headphones.

A review of literature on creating virtual environments using loudspeaker states that methods based on physical reconstruction of the acoustical field, like Ambisonic and the holophonic approach, offer a good localization performance over an extensive listening area. The main advantage of the general Ambisonic approach [1] is the high computational efficiency. Using generalized Ambisonic the reconstruction of the sound field is accurate only over a small listening area. However, the proposed system is related to a binaural synthesis method based on decoding Ambisonic to virtual loudspeakers avoiding the problems caused by a small listening area.

Sound source spatialization in virtual acoustic environments using headphones requires the filtering of the sound streams with head related transfer functions (HRTFs). The HRTFs capture both, the frequency and time domain aspects of the listening cues to a sound position. The measurement of HRTFs has been researched extensively by Wightman and Kistler [2]. In the proposed system generic HRTFs using the KEMAR [3] as well as the CIPIC [4] database have been used. Wenzel *et al.* [5] state that the use of nonindividualized transfer functions leads to a degradation of localization accuracy, increasing the following errors:

- *localization error*, referring to the deviation of the perceived to the synthesized direction of a virtual sound source
- *localization blur*, that describes the “width” of the perceived stimulus
- *externalization error*, also termed as “inside-the-head localization”
- *cone of confusion*, which refers to localization errors caused by contours of constant interaural time difference (ITD) and interaural level difference (ILD) resulting in front-back confusions

Regarding hearing in natural sound fields humans are able to improve sound source localization using small head movements. Begault and Wenzel [6] have shown the importance of incorporating head tracking as well as reverberation in binaural sound reproduction systems to improve localization capabilities. This yields the problem of high-quality time-variant interpolation between different HRTFs. To overcome this problem a virtual Ambisonic approach is used, as is shown in the next section.

2. THEORY

The following section gives a brief introduction into Ambisonic theory. Furthermore a 3D binaural sound reproduction system is developed, incorporating head tracking as well as room simulation.

2.1. The Ambisonic Approach

Ambisonic is a technique for spatial audio reproduction introduced in the early seventies by Gerzon [1]. Further details of Ambisonic are published in [7][8][9][10].

The basic idea of the generalized Ambisonic approach is the expansion of a wave field into spherical harmonics, assuming that the original wave field is a plane wave. However this assumption is not compulsory because any acoustical field can be expressed as a superposition of plane waves. It is claimed in [7],[9] that Ambisonic systems are asymptotically holographic. Holographic theory states that the Kirchhoff-Helmholtz Integral relates the pressure inside a source free volume of space to the pressure and velocity on the boundary at the surface. Deriving the Ambisonic encoding/decoding equations from the Kirchhoff-Helmholtz theory it can be shown that the original wave field may be reconstructed exactly by arranging infinitely many loudspeakers on a closed contour assuming plane wave signals. Using a finite number of N loudspeakers arranged on a sphere a good approximation of the original acoustical field may be synthesized over a finite area (sweet spot). Poletti has shown in [9] that higher order Ambisonic systems are increasingly accurate. Beyond, some indication of the upper frequency limit of the system is given in [9] as well.

The decomposition of the incoming wave field into spherical harmonics can be shown deriving Ambisonic from the homogenous wave equation

$$\Delta p(t,\mathbf{r}) - \frac{1}{c^2}\frac{\partial}{\partial t^2} p(t,\mathbf{r}) = 0 \tag{1}$$

where $p(t,\mathbf{r})$ is the sound pressure at the position **r** and c is the speed of sound. Solving the wave equation for the incoming sound wave and the plane waves of the several loudspeakers yields the so called matching conditions

$$s \cdot Y^{\sigma}_{m,\eta}(\Phi,\Theta) = \sum_{n=1}^{N} p_n \cdot Y^{\sigma}_{m,\eta}(\varphi_n, \vartheta_n) \tag{2}$$

The left side of (2) represents the Ambisonic encoding equation,

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

where $\mathbf{B}_{\Phi,\Theta}$ represents the Ambisonic channels in vector notation,

$$\mathbf{B} = [Y^{1}_{0,0}(\Phi,\Theta), Y^{1}_{1,0}(\Phi,\Theta), \dots Y^{-1}_{M,M}(\Phi,\Theta)]^{T} \cdot s \tag{4}$$

s is the pressure of the original sound wave coming from direction (Φ,Θ) and $Y^{\sigma}_{m,\eta}$ describes the spherical harmonics. On the right hand side of (2) p_n is the signal of the nth loudspeaker at direction (φ_n, ϑ_n). $Y^{\sigma}_{m,\eta}$ can be calculated as follows

$$Y^{\sigma}_{m,\eta}(r) = \begin{cases} A_{m,\eta} P^{\eta}_{m}(\cos\Theta)\cos(m\Phi) & \text{for } \sigma = 1 \\ A_{m,\eta} P^{\eta}_{m}(\cos\Theta)\sin(m\Phi) & \text{for } \sigma = -1 \end{cases} \tag{5}$$

where P^{η}_{m} represents the Legendre polynomials.
Using vector notation (2) can be written as

$$\mathbf{B} = \mathbf{C} \cdot \mathbf{p} \tag{6}$$

where

$$\mathbf{p} = [p_1, p_2, \dots p_N]^{T} \tag{7}$$

is the vector with the several loudspeaker signals.

Now it is possible to calculate the decoder from the encoding equations as follows

$$\mathbf{D} = \mathrm{pinv}(\mathbf{C}) = \mathbf{C}^{T} \cdot (\mathbf{C} \cdot \mathbf{C}^{T})^{-1} \tag{8}$$

The decoding stage will only depend on the actual loudspeaker arrangement. In the 3D case the minimum number N of required loudspeakers is (M+1)², where M is the system order. Considering the reproduction of a 2D field using a finite number of loudspeakers and deriving Ambisonics from the two dimensional spatial Fourier transform, it is shown in [9] that the decoding process may be described by the so called angular sinc functions (Asincs). More precisely the Asinc functions describe the panning of the Ambisonic signals to the several loudspeakers. Furthermore sound source localization may be confused by out of phase signals coming from loudspeakers far away from the intended virtual source position. Like in conventional signal processing techniques using the Fourier transform, windowing may be used to attenuate the sidelobes of the Asinc functions. Therefore, weighting the amplitudes of higher order Ambisonic channel signals, representing higher order spherical harmonics, yields a reduction of the sidelobes as well as a broadening of the main lobe. Therefore the confusing far away speaker signals may be attenuated considering just noticeable difference (JND) thresholds to decrease the localization error. Though as a result of the wider main lobe the localization blur increases. Hence, "windowing" of the Ambisonic channels can be used to improve the capabilities of the decoder to minimize the error of synthesis [10].

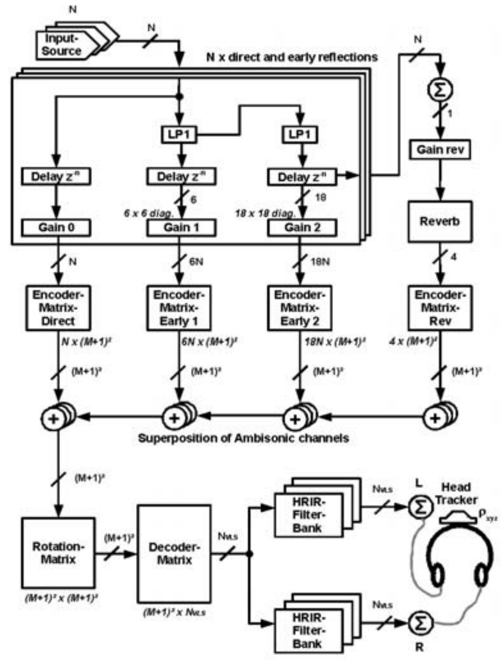

Figure 1: *Block diagram of the 3D binaural sound reproduction system incorporating head tracking and room simulation*

2.2. Binaural Sound Reproduction

In binaural systems, time-varying spatialization of sound sources by direct convolution with HRTFs results in time-variant interpolation between different HRTFs. This interpolation yields artifacts decreasing the localization performance of the system. As mentioned above it is possible to synthesize a sound field by arranging N loudspeakers on a sphere using the Ambisonic approach. The proposed system (figure 1) encodes the virtual sound sources dependent to their position in the virtual acoustic space into the Ambisonic domain. To take the different listener source distances into account, source signals are delayed before encoding using simple tap delay lines. Using 3D Ambisonic of Mth order yields (M+1)² Ambisonic channels. The number of

Ambisonic channels is independent of the number of virtual sound sources to encode. This fact is important for incorporating room simulation as shown later in section 2.3. Now head rotation may be taken into account by using simple time-variant rotation matrices in the Ambisonic domain. Head rotation is identified using a head tracking device mounted on the headphones. Next, the Ambisonic signals are decoded to virtual loudspeaker signals. To create the binaural cues, the virtual loudspeaker signals are filtered with their appropriate HRTFs. Since the system is linear and time invariant (LTI), these signals are superimposed to create the left and right ear signals.

Due to the fact that the decoding matrix is defined by the arrangement of the virtual loudspeakers it is important to distribute them as uniformly as possible over the spheres surface. Otherwise, ill conditioning or even singularities in the decoder matrix may occur.

To further increase the computational efficiency of the overall system psychoacoustical effects are taken into account as well. Humans are able to localize sound sources in horizontal plane more precisely than in vertical directions [11]. Therefore encoding sound sources in vertical directions is done by using Ambisonic of lower order. Using this mixed order Ambisonic approach may reduce the number of required transmit channels.

Filtering with HRTFs is a highly computational task. Listening tests and error analysis of a 2D system have shown that shorten HRTFs up to 128 points yields a satisfactory localization accuracy [12], [13] (figure 2). Furthermore the use of individualized HRTFs will increase localization capabilities. Further investigations on the influence of different HRTFs to the overall system localization performance have not been carried out during this work.

2.3. Room Simulation

To improve localization accuracy and the perceived externality of virtual sound sources, it is important to incorporate room simulation. In this section we focus on the simulation of sound reverberation as a natural phenomenon occurring when sound waves propagate in an enclosed space. Room simulation is divided into two stages of computation. First, the early reflections of first and second order are calculated. Then a model for efficient simulation of the diffuse sound field is introduced.

The early reflections are taken into account using a simple geometrical acoustic approach calculating image sources. We are considering a rectangular room containing omni-directional virtual point sources. To consider the acoustic properties of the reflecting walls, the image source signals are filtered with a low order IIR low pass filter. According to the different distances of the images sources to the listening position, the image source signals are delayed and attenuated as well. Now the several image source signals are encoded to Ambisonic dependent to their position in the virtual acoustic space. Due to the fact that higher order early reflections become more and more diffuse they are encoded with lower order Ambisonic. The loss of localization accuracy may be accepted to increase computational efficiency.

Another approach to decrease computational cost for encoding the image sources is to divide the virtual space into several subspaces or so called regions of influence. Image sources situated in same subspaces are bundled. Henceforth the bundled signals are encoded according to the direction dedicated to their respective subspace.

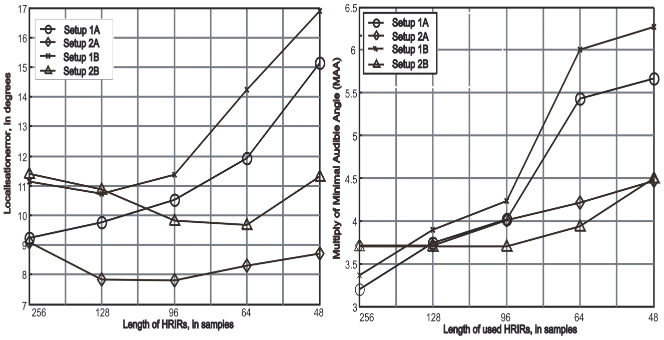

Figure 2: *Localization error (left figure) in degrees and localization blur in multiples of the minimum audible angle (MAA) [11] (right figure) using different HRTFs (1,2) and different interpolation techniques (A,B)*

Late reverberation creates an ambient space in the perception of the listener. Dattorro [14] states that the most efficient implementations of reverberators rely on all pass circuits embedded within very large globally recursive networks (figure 3). To handle the start time of late reverberation the input signal is delayed. Then the signal is low pass filtered to consider the coloration due to the absorption of high frequency signal components at the enclosing walls. Furthermore, the first set of input diffusers quickly decorrelate the incoming sound to prepare it for the next stage. To loop the decorrelated sound indefinitely the second diffuser stages are arranged to feed back globally on themselves. By multiplying a gain factor less than 1.0 in the feedback paths it is possible to control the decay time. The low pass filters in the feedback paths incorporate the texture of materials at the enclosing walls as before.

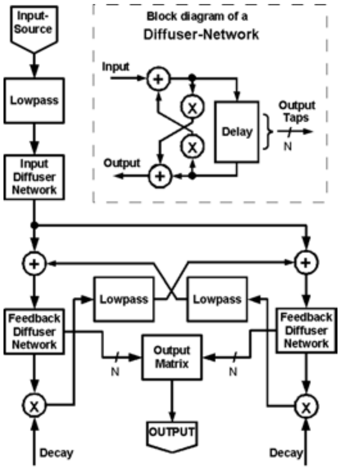

Figure 3: *Recursive reverberation network*

Finally the reverberation signals are encoded into Ambisonic domain. Because of the fact that reverberation signals do not affect localization accuracy, low order Ambisonic is sufficient for encoding.

Incorporating this reverberation network yields a simple way to control parameters of the reverberated sound like the diffusion of the input and decay signals (decorrelation), the reverberation pre delay, the reverberation gain, the decay rate and the cut off frequency of the high-frequency damping.

The effects of the surrounding environment can be used to change the perceived distance of the virtual sound source.

3. IMPLEMENTATION

The proposed system is implemented on a standard notebook using Pure Data (PD). PD is a graphically based open source real time computer music software by Miller Puckette[1].

First a 2D system with 4th order Ambisonic was implemented in PD as well as on a digital signal processor (DSP) running a PC as a host system. The 2D system has been optimized using the results of an objective mathematical model as described in [12]. Furthermore the objective mathematical model as well as the localization accuracy of the optimized system have been evaluated using listening tests [13].

The next step was the implementation of a 3D system using 4th order Ambisonic on a usual notebook running a 1.6 MHz CPU. Because of the fact that for a 2 channel system with room simulation the required CPU performance goes up to 2GHz the computational efficiency of the overall system has been optimized as described above. With the rapid increasing of CPU power it will become possible to run the binaural application as a background task for computer music software.

4. CONCLUSIONS

In this paper the relations between the virtual Ambisonic approach and time-variant binaural sound reproduction systems have been discussed. The main advantages of using Ambisonic in time varying binaural sound systems are as follows:

- Rendering the sound field using Ambisonic in time-variant binaural sound reproduction systems yields time-invariant HRTFs without the need of interpolation between them.
- The number of required HRTFs is independent of the number of virtual sound sources to encode. This is important because using a geometrical approach for room simulation yields an enormous increase of virtual sound sources to encode.
- Ambisonic provides a decoupling of the encoder and decoder. Hence, the awareness of the playback configuration can be limited to the decoder while only the universal multi channel format is implemented in the encoding stage.
- Head rotation may be taken into account with simple time-variant rotation matrices in Ambisonic domain using a head tracking device for rotation angle detection

The following optimizations have been carried out to reduce the computational cost of the algorithm:

- mixed order Ambisonic
- low order Ambisonic for early reflections of higher order and late reverberation
- bundle image source signals before encoding
- shorten HRTFs up to 128 points

Measuring the CPU performance of the implemented system shows that the optimized system requires about 50% of the computational cost of the non optimized.

As future research, a comprehensive localization error analysis of the proposed system would be interesting using the objective mathematical model of localization as well as listening tests.

5. REFERENCES

[1] Gerzon, M. A., “Ambisonic in multichannel broadcasting and video”, in *J. Audio Eng. Soc.*, vol. 33, pp. 859-871, 1985

[2] Wightman, F. L. and Kistler, D. J., “Headphone stimulation of free field listening I: stimulus synthesis”, in *J. Acoust. Soc. Am.*, vol. 85, pp. 858-867, 1989

[3] Gardner, W. G. and Martin, K. D., „HRTF Measurement of a KEMAR“, in *J. Acoust. Soc. Am.*, vol. 97, pp. 3907-3908, 1995

[4] Algazi, V. R., Duda, R. O., Thompson, D. M. and Avendano, C., “The CIPIC HRTF Database”, in *Proc. IEEE Workshop on Applications of Sig. Proc. to Audio and Electroacoustics*, pp. 99-102, NY, 2001 October

[5] Wenzel, E. M., Arruda, M., Kistler, D. J. and Wightman, F. L., “Localization using nonindividualized head-related transfer-functions”, in *J. Acoust. Soc. Am.*, vol. 94, pp. 111-123, 1993

[6] Begault, D. R. and Wenzel, E. M., “Direct Comparison of the Impact of Head Tracking, Reverberation and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Sound Source”, in *J. Audio Eng. Soc.*, vol. 49, no. 10, 2001 October

[7] Nicol, R. and Emerit M., “3D Sound Reproduction over an Extensive Listening Area: A Hybrid Method Derived from Holophony and Ambisonics”, in *Proc. AES 16th Int. Conf.*, pp. 436-453, 1999

[8] Jot, J. M., Larcher, V. and Pernaux J.-M., “A Comparative Study of 3D Audio Encoding and Rendering Techniques”, in *Proc. AES 16th Int. Conf.*, pp. 281-300, 1999

[9] Poletti, M., “The Design of Encoding Functions for Stereophonic and Polyphonic Sound Systems”, in *J. Audio Eng. Soc.*, vol. 44, no. 11, pp. 1155-1182, 1996 November

[10] Poletti, M., ”A Unified Theory of Horizontal Holographic Sound Systems”, in *J. Audio Eng. Soc.*, vol. 48, no. 12, 2000 December

[11] Blauert, J., “Spatial Hearing”, 2nd ed., MIT Press, Cambridge, MA, 1997

[12] Sontacchi, A., Noisternig, M., Majdak, P. and Höldrich, R., “An Objective Model of Localisation in Binaural Sound Reproduction Systems”, in *Proc. AES 21st Int. Conf.*, St. Petersburg, Russia, 2001 June

[13] Sontacchi, A., Majdak, P., Noisternig, M. and Höldrich, R., “Subjective Validation of Perception Properties in Binaural Sound Reproduction Systems”, in *Proc. AES 21st Int. Conf.*, St. Petersburg, Russia, 2001 June

[14] Dattorro, J., “Effect Design: Part 1: Reverberator and Other Filters”, in *J. Audio Eng. Soc.*, vol. 45, no. 9, pp. 660-684, 1997 September

[1] http://crca.ucsd.edu/~msp/software.html