# Real-time, Head-tracked 3D Audio with Unlimited Simultaneous Sounds

*Craig Jin, Teewoon Tan, Alan Kan, Dennis Lin, André van Schaik*

Computing and Audio Research Laboratory,
School of Electrical and Information Engineering,
University of Sydney,
Sydney, Australia.
craig@ee.usyd.edu.au,
teewoon@ee.usyd.edu.au,
akan@ee.usyd.edu.au,
dlin@ee.usyd.edu.au,
andre@ee.usyd.edu.au

*Keir Smith, Matthew McGinity*

iCinema Centre for Interactive Cinema Research
University of New South Wales,
Sydney, Australia

keirs@cse.unsw.edu.au,
mmcginity@cse.unsw.edu.au

**ABSTRACT**

This research presents a novel 3D audio playback method in which real-time head-tracking is maintained with an unlimited number of simultaneous sound sources. The method presented relies on using a 500-900MByte sound buffer which contains binaural data for 385 head orientations and a processing platform with two hard disks in a RAID 0 configuration that can stream data at a rate of 80-100 MBytes/s. We discuss issues related to how the number of head-orientations influences a smooth presentation, how the window length influences smooth transitions between different head-orientations and the file format used for storing the sounds.

The new 3D audio playback method was incorporated into a 3D audio playback engine (3DApe) which can: play a 3D audio soundtrack consisting of an unlimited number of simultaneous sound sources, switch between different 3D audio soundtracks, play back up to 8 simultaneous and instantaneous sound sources on command, use a head-tracker interface via the virtual reality peripheral network (VRPN), supply 3D audio communication using voice over IP, and interface with a Virtools graphical software engine. 3DApe was demonstrated as part of an interactive 3D cinematic artwork, entitled *Conversations*, that was on display at the Powerhouse Museum in Sydney in December 2004 [1].

## 1. INTRODUCTION

For many years, interest in 3D binaural systems has been in the area of scientific research, simulation and entertainment [2]. In more recent years, there has been a growing interest in the application of 3D binaural systems in virtual reality and augmented reality applications [3]. In virtual reality applications, sound sources are often presented over headphones to give precise control over the virtual auditory environment. Head-related transfer functions (HRTFs) are used to create virtual auditory space and have commonly been recorded on humans and manikins to allow externalization and spatialisation of sound sources presented in a binaural audio display.

An HRTF is an acoustic transfer function that describes the sound pressure transformation from a location in the free-field to the listener's eardrum [4]. HRTFs contain the acoustic cues necessary for spatial hearing: the interaural time difference cue (ITD), the interaural level difference cue (ILD) and the spectral transformations applied by the outer ear. More details about the acoustic cues for spatial hearing can be found in [5].

Head-orientation sensors have also been used with binaural audio systems to heighten the experience of the user by allowing head movement. However, head movement increases the complexity of the audio rendering engine as it has to present sound sources in the correct locations relative to the user's head orientation. In current real-time, head-tracked 3D audio systems there is a trade off between keeping the system running in real-time and increasing the number of simultaneous sound sources that are rendered spatially. In one extreme, the number of simultaneous sound sources is limited and the HRTF filtering for spatialisation is performed on-the-fly. At the other extreme, the number of simultaneous sound sources can be unlimited, but the HRTF filtering is performed offline with no head-tracking during playback. Hence the real conflict is between real-time head-tracking and the number of simultaneous sound sources. Dedicated hardware can be used to achieve 3D audio with head-tracking, but this can turn out to be costly, e.g., Lake Technology's Huron processor uses one DSP chip per sound source [6]. This paper presents a method that has been incorporated into a 3D audio playback engine (3DApe) that runs on standard PC hardware and allows real-time head-tracking to be maintained with an unlimited number of simultaneous sound sources

3DApe was developed to meet the audio demands of *Conversations*, an interactive 3D cinematic artwork designed by the iCinema Centre for Interactive Cinema Research at the University of New South Wales [1]. A distributed, multi-user virtual environment, *Conversations* revolves around the escape of Ronald Ryan and Peter Walker from a Melbourne prison on December 19, 1965. During the escape, a prison guard was shot and killed, a crime for which Ryan was subsequently tried. Despite some somewhat incongruous witness accounts (were one or two shots fired?), Ryan was found guilty and sentenced to death.

During the experience, the user is able to witness a reenactment of the prison break as a 2 minute spherical stereo film. Using a head-tracker and head-mounted display, the user is placed at the scene of the crime and is free to rotate his or her head to choose any point of view (Figure 1). Given the relatively narrow field of view provided by contemporary head-mounted displays (in this case, a Daeyang i-visor DH-4400VPD

3D), it was deemed necessary to provide a soundtrack that would present the user with spatial cues that could draw his/her attention to any action that may be currently out of view. And as a headphone solution was desired to provide a greater sense of isolation and immersion when the piece is exhibited in a noisy environment, binaural spatialisation would be necessary.

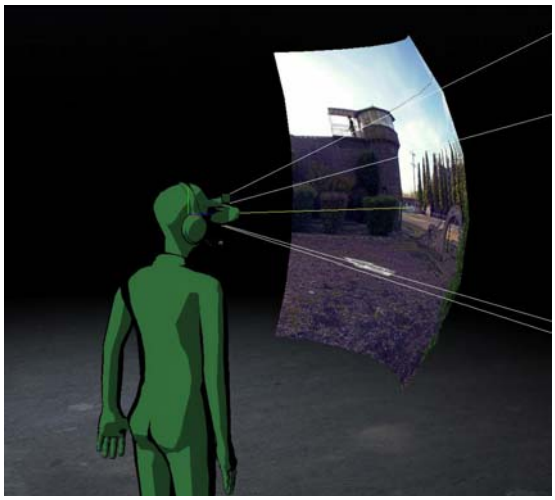

Figure 1. *Wearing a head-mounted display and headphones, a user is able to choose his perspective of a spherical film.*

The soundtrack to the spherical film contains 26 audio sources which follow different spatial paths and need to be acoustically rendered spatially and simultaneously. In other words, the sounds have to be rendered in the correct location with respect to the user's current orientation. An important note is that while the user can change his head-orientation to view different parts of the film, he cannot alter his location. In other words, the film has been rendered for a particular location, but is orientation independent.

In addition to providing the soundtrack to the spherical film, 3DAPE was also required to spatially render up to 8 dynamic sound sources in real-time. In contrast to the soundtrack to the film, which never changes, the position of these dynamic sounds are not known in advance. Finally, 3DApe also provides 3D audio communications via voice over IP, allowing participants positioned at three separate VR stations to communicate with one another. So, in all, *Conversations* required 3DApe to run on standard PC hardware and to perform three functions: provide a soundtrack to a spherical film; spatially render dynamic sound sources in real-time and distribute and spatially render voice data. In addition, 3DApe must communicate with the Virtools graphical software engine on another computer and the virtual reality peripheral network (VRPN) [7].

## 2. METHODS

For *Conversations*, each of the 26 audio sources was supplied as 16-bit mono PCM WAV file at a sampling rate of 44100Hz. All sound files were 1 minute and 54 seconds long with silent intervals when the sources were quiet. The sound files were synchronized with the visual display. The spatial trajectories of the audio sources relative to a fixed viewing angle were provided as a list of Cartesian coordinates. In order to simultaneously render the 26 audio sources in the film, offline audio mixing was employed. However, there is clearly no limit to the number of sound sources that can be rendered offline.

Although the spatial trajectories were provided for only one viewing angle, the soundscape had to be rendered for all viewing angles to allow for head-tracking. With the head-mounted display (HMD), the participant could face any direction, and look up and down to view the panoramic scene and could change their viewing direction at any time. Most head-orientation sensors provide head orientation data in terms of three angles: azimuth, elevation, and roll. Given that it requires approximately 400 azimuth and elevation combinations to cover the sphere with 10 degrees of separation, disregarding directions lower than -45 degrees in elevation, and that there are approximately 19 roll positions in 10 degree steps from -90 degrees to +90 degrees, this implies rendering the soundscape for 400 x 19 or 7600 viewing angles. The sound data corresponding to this number of viewing angles cannot be streamed from a hard disk in real-time with current desktop computers (e.g., it requires a hard disk access rate of 1.25GBytes/second).

The above problem may seem to be an insurmountable difficulty. However, observation of participants viewing the display with an HMD indicates that most participants do not roll their heads substantially. Most of the head roll seems to occur when the head is turned to the side by a significant amount. In addition, the HMD, perhaps because of its weight, seems to encourage participants not to roll their head, compared with normal viewing. Thus, in *Conversations* the 3D soundscape was rendered assuming no head roll.

Without head roll, streaming head-tracked sound data from a hard disk becomes realizable for a panoramic video. The critical issue is maintaining smooth playback during head movement. In other words, in a quick action scene such as a prison breakout, the azimuth and elevation angles for the viewer's head orientation can vary by a large amount. In order to maintain smooth playback, given these large variations in instantaneous head-orientation, the sound data for all of the viewing angles were buffered simultaneously for a given time interval, resulting in a very large sound buffer (e.g., 900MBytes). The advantage of this method is that a change in viewing angle only requires shifting a software pointer.

The limiting factor in terms of the number of viewing angles that can be smoothly rendered is the rate at which the sound data can be streamed from disk and the amount of random access memory (RAM) available for buffering the sound. The hardware platform supplied for *Conversations* was a Dual processor 3Ghz Xeon with 3 GBytes of RAM and two Serial ATA hard disks in a RAID 0 configuration. The RAID 0 configuration provided a practical data rate of 80-100MBytes/sec as determined by HD Tune [8]. Figure 2 shows the 385 head orientations that were chosen for *Conversations*. The allowable head orientations were restricted to azimuth and elevation angles with no roll. The azimuth angles spanned a range between -180º and 180º and the elevation angles spanned a range between -45º and 90º, covering the sphere evenly every 10º. It was assumed that the participant would not orient their head below -45º elevation. For 385 head orientations, the minimum disk access rate was 64.77 MBytes/sec[1], which was

[1] For each sample of sound, there are 4 Bytes, 16-bits for the left and 16-bits for the right. Hence, the required data rate can be calculated as (385 x 4 Bytes/sample x 44100 samples/second) / $1024^2$ = 64.77 MBytes/second.

well within the system limits and allowed some headroom for other system processes.

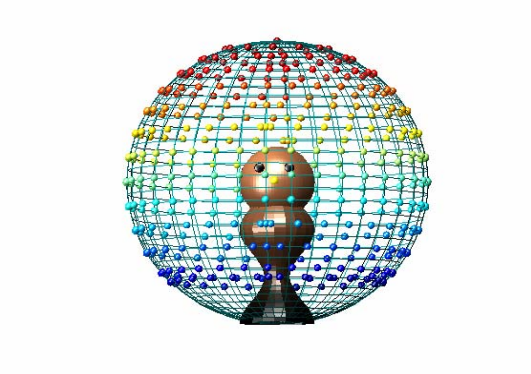

Figure 2. *Pre-rendered head orientations. The dots represent the pre-rendered viewing directions.*

Fig. 3 shows the MATLAB processing steps that were developed to process the sound files based on the spatial trajectories of the sounds. The sound designer composed the sound tracks for a fixed viewing angle and did not apply any Doppler frequency effects. Each mono sound track was divided into blocks of 1024 samples with 50% overlap and a cosine window was applied to each block before and after processing. The overlap and add procedure ensured smooth and continuous movement of sound sources without audible clicks. Doppler shifting and a frequency dependent attenuation to simulate air absorption were applied to each block. Doppler shifting was implemented by resampling the block, according to standard Doppler shift formulas [9]. Attenuation due to air absorption was implemented according to the ISO Standard 9613-1 [10]. The sound tracks were spatialised using HRTFs that had been recorded on a human subject as described in [11]. The recorded HRTFs were interpolated using a spherical thin-plate spline [12] according to the spatial path of the sound. The end result of the MATLAB processing was a separate binaural signal for each sound source.

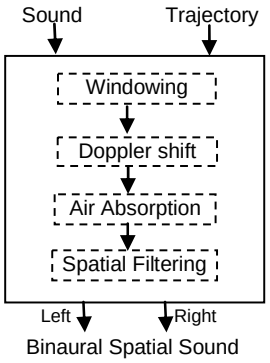


Figure 3. *The steps for processing each sound source in MATLAB are shown. The Doppler shift and air absorption effects were applied using physical models and the spatial rendering was performed using HRTFs.*

The MATLAB processing was applied to each of the 26 sound sources individually and the resulting 26 binaural sound tracks were mixed to a single binaural sound track corresponding to a single viewer orientation. The entire process was repeated for each of the 385 viewer orientations. The 385 binaural tracks were combined into a single spatialised sound source (SSS) file for 3DApe. The structure of an SSS file is shown in Figure 4. The header contains a simple number sequence to indicate that it is an SSS file, a value to indicate the number of sound sources, and a few extra bytes for later expansion. Following the header, pairs of azimuth and elevation co-ordinates are stored for the viewer orientations. Lastly, the audio data are stored with 16-bit precision and at a 44100 Hz sampling rate.

Figure 4. *The file format for an SSS file is shown.*

| SSS Sequence (4 Bytes) |
|---|
| Number of sound sources = N (4 Bytes) |
| Reserved (4 Bytes) |
| N azimuth, elevation pairs (N x 2 x 4 Bytes) |
| Sound data stripped byte-wise for each orientation<br>Byte 1, Left sample 1, Head orientation 1<br>•<br>•<br>Byte 1, Left sample 1, Head orientation N<br>Byte 2, Left sample 1, Head orientation 1<br>•<br>•<br>Byte 2, Left sample 1, Head orientation N<br>Byte 1, Right sample 1, Head Orientation 1<br>•<br>•<br>Byte 1, Right sample 1, Head orientation N<br>Byte 2, Right sample 1, Head orientation 1<br>•<br>•<br>Byte 2, Right sample 1, Head orientation N<br>•<br>• |

The audio data are stored in a byte-wise fashion across all viewer orientations, and the left and right channels are interleaved. In other words, the first byte of the first sample for the left channel is stored sequentially for all 385 orientations. Following this, the second byte of the first sample for the left channel is stored sequentially for all 385 orientations. Next, the first byte of the first sample for the right channel is stored sequentially for all 385 orientations and similarly with the second byte of the first sample for the right channel. This byte-wise stripping of the data allows easy and continuous streaming of the sound file into the data buffer for audio playback. The data are stripped byte-wise rather than sample-wise because buffer sizes are defined in bytes rather than samples, allowing for simpler data read requests in the code. No data compression was applied to the audio data.

The 3DApe system was developed using Microsoft Visual C++ and relied on the use of a 500-900MByte circular data buffer to read in the sound data from an SSS file. The 500-900MByte circular data buffer is referred to as the SSS data buffer and allows 7.54-13.57 seconds of sound data for each head orientation to be read into memory. The SSS data buffer was then used to fill the DirectX sound buffer for actual playback. In order for 3DApe to play back the correct sound mix for a given orientation, the current azimuth and elevation of the viewer's head orientation was obtained from the head-tracker via VRPN and the nearest matching orientation from the fixed list of 385 orientations was identified using a pre-calculated lookup table. As the user's head orientation changed, 3DApe moved the playback pointer to the appropriate point in the circular data buffer. In order to ensure a smooth transition

between different orientations, a cosine squared window of 50 samples was applied to the sound data for both orientations and an overlap-add operation was performed.

While playing back an SSS sound track, 3DApe could also render on command up to 8 instantaneous and simultaneous mono sounds as spatial audio. At the same time, 3DApe also provided spatial-audio communications using voice over IP to participants in three separate displays. The positions of the participants in the virtual world as well as the locations of the instantaneous sounds were transmitted to 3DApe via the Virtools graphics software engine. The HRTF data used to spatially render the speech and instantaneous sounds were stored in a compressed format and interpolated on-the-fly using a spherical thin-plate spline [12]. These were applied to the sounds as minimum-phase finite impulse response filters.

## 3. RESULTS AND DISCUSSION

3DApe provided the spatial-audio rendering engine for two virtual worlds in the artwork *Conversations*. The first virtual world, *Pentridge World*, portrayed a prison breakout using panoramic 3D graphics and 3D audio. The primary test for 3DApe in this case was synchronization of the head-tracked 3D audio with the visual display as the participant turned to observe the action. In the other virtual world, *Ghost World*, the participant could talk to ghosts of the characters in *Pentridge World* and also avatars of other participants while listening to a 3D audio sound track of whispering ghosts. *Ghost World* tested 3DApe's ability to simultaneously: play a 3D audio sound track with head-tracking, spatially render the speech of the ghosts, and spatially render the voice over IP of other participants.

Subjective testing of 3DApe was carried out using a headphone (Sennheiser ES2200), head-orientation sensor (Intersense InertiaCube2), and head-mounted display (Daeyang I-Visor 3D). It was observed that with 385 pre-rendered orientations and a 500-900MByte data buffer, the auditory scene smoothly matched the visual scene without noticeable lag during the normal head motion associated with participants exploring the prison breakout. Reducing the number of pre-rendered orientations to 283 caused perceptible jumps in the locations of sounds. Also, extremely rapid head movements produce a brief mismatch between the audio and video displays, even with 385 pre-rendered orientations.

The length of the cosine squared window applied during the transitions between pre-rendered orientations influenced the smoothness of the audio playback. Long window lengths resulted in a perceptible lag in the movement of sound sources and short window lengths resulted in some clicking noise. A window length of 50 samples was chosen as optimal, even though clicks were sometimes audible when the head orientation changed quickly. It is believed that the clicks are caused by phase differences in the sounds corresponding to different orientations. The phase differences are likely to arise from the slight variations in the time it takes sound from a fixed source to reach the ear for different orientations.

The size of the SSS data buffer required for smooth 3D audio playback depends on processor load and the number of pre-rendered orientations stored in the SSS file. For the 3GHz Xeon with 3GBytes of RAM, a minimum buffer size of 500Mbytes was required, allowing 7.54 seconds of sound data to be buffered.

## 4. CONCLUSIONS

A novel method for 3D audio playback for an unlimited number of simultaneous sound sources and with real-time head-tracking was presented. By performing the 3D audio processing offline, an unlimited number of simultaneous sound sources can be spatially rendered. By rendering binaural sound tracks for a fixed and closely-spaced set of head orientations, smooth playback with real-time head-tracking was achieved. However, these achievements required a top-of-the-range workstation. In particular, a dual 3Ghz processor with 3 GBytes of RAM and two Serial ATA hard disks in a RAID 0 configuration was used. Also, the software required a 500MByte data buffer in order to store the sound data for all of the head orientations. Currently, ongoing work focuses on examining improvements to the method for smoothly switching between sound tracks for different orientations and also on implementing compression of the audio data for the SSS file such that there is not too heavy a burden on the processor during decompression.

## 5. REFERENCES

[1] D. Del Favero, R. Gibson, I. Howard, and J. Shaw, "Project Conversations," http://www.icinema.unsw.edu.au/projects/prj_conversations.html, 2004.

[2] B. Shinn-Cunningham, "Applications of virtual auditory displays," presented at Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE, 1998.

[3] A. Harma, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, "Augmented reality audio for mobile and wearable appliances," *AES: Journal of the Audio Engineering Society*, vol. 52, pp. 618-639, 2004.

[4] F. L. Wightman and D. J. Kistler, "Headphone Simulation of Free-Field Listening .1. Stimulus Synthesis," *Journal of the Acoustical Society of America*, vol. 85, pp. 858-867, 1989.

[5] S. Carlile, *Virtual Auditory Space: Generation and Application*. New York: Chapman and Hall, 1996.

[6] "Huron Technical Manual," Lake Technology Limited, 1995.

[7] I. Russell M. Taylor, T. C. Hudson, A. Seeger, H. Weber, J. Juliano, and A. T. Helser, "VRPN: a device-independent, network-transparent VR peripheral system," presented at Proceedings of the ACM symposium on Virtual reality software and technology, Baniff, Alberta, Canada, 2001.

[8] "HD Tune 2.10 - Hard Disk Utility," EFD Software, http://www.hdtune.com/, 2003.

[9] D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics*, 5th ed: John Wiley & Sons, 1997.

[10] "ISO 9613-1: Acoustics -- Attenuation of sound during propagation outdoors -- Part 1: Calculation of the absorption of sound by the atmosphere," 1993.

[11] S. Carlile, C. Jin, and V. Harvey, "The generation and validation of high fidelity virtual auditory space," presented at Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE, 1998.

[12] S. Carlile, C. Jin, and V. van Raad, "Continuous Virtual Auditory Space Using HRTF Interpolation: Acoustic & Psychophysical Errors," presented at Proceedings of the First IEEE Pacific-Rim Conference on Multimedia (2000 International Symposium on Multimedia Information Processing), Sydney, 2000.