Proceedings of the Institute of Acoustics

DESIGN AND IMPLEMENTATION OF 3-DIMENSIONAL SPATIAL AUDIO FOR IMMERSIVE ENVIRONMENTS

DR ANDREW RIMELL (1) & DR MIKE HOLLIER (1)

(1) BT LABORATORIES, MARTLESHAM HEATH, IPSWICH, SUFFOLK, IP5 3RE

1. Introduction

Immersive environments, providing sensory immersion for VR and Telepresence, can be achieved with wrap-around screens and spatial-audio systems. Such systems are currently being developed at BT Laboratories for virtual-meetings, education, medicine, training and entertainment. To maximise naturalness and immersion it is necessary to use high quality spatial audio is provided. In many multimedia systems more effort has been invested in the development of the video component than the audio part. Further development of the audio part is therefore required. Work at BT Laboratories reflects the fact that audio is at least as important as video, particularly in communications - imagine a video-conference with no sound!

This paper discusses the design and implementation of several immersive environments under development at BT Laboratories. Design considerations such as multiple listeners and difficult acoustic environments are considered. Binaural, transaural and adapted ambisonics techniques have been investigated to provide spatialisation of virtual audio sources for a variety of applications. The systems described have been implemented and results of the subjective performance of the immersive audio environments are discussed.

Three applications; the SmartSpace chair, the Network Spatial Server and the BT/ARC VisionDome [1]. The VisionDome [2] is described in detail, with an in-depth look at the audio system and methods of overcoming various sound localisation problems.

Reproduction strategies are introduced before describing the applications together with a brief overview of the strategies employed in each of the applications outlined in Section 3.

2. Reproduction Techniques

A variety of spatial audio reproduction techniques have been exploited and developed at BT Labs to suit a range of applications. The ones relevant to this paper are described in Sections 2.1 to 2.4.

2.1 Binaural

A binaural signal (recorded with a dummy head or synthesised) recreates the soundfield present at

the entrance to the ear canal when listening to an actual sound source [3]. The sound waves impinge the outer ear (pinna) at a given angle, the pinna then filters the sound with an angle dependent transfer function (Head Related Transfer Function - HRTF). The combination of HRTFs and time delay enable the listener to locate sounds: HRTFs have the greatest influence at mid and high frequencies and time-delay at low frequencies. Figure 1 illustrates the encoding process where HRTFs are encoded into each ear's signal. To prevent the sound being filtered by the pinna twice it is necessary to use headphones to listen to the binaural signal

Binaural encoding produces very realistic sounding audio especially when it also encodes the HRTF of each individual reflection. All of the processing is carried out at the encoding stage and the spatialised audio can be reproduced on conventional audio equipment. The principal disadvantage is the necessary use of headphones.

2.2 *Headtracked Binaural*

Because binaural signals are heard with headphones, any movement of the head will result in the same movement of the spatialised audio signal. In telepresence systems it may be desirable to keep the sound location static with respect the physical environment resulting in movement as the listener rotates their head. Applying different HRTFs for different head positions can keep the sound source location static; a typical practical implementation [4] uses 128 different HRTFs.

The solution comprises a pair of headphones with a rotational position sensor mounted on top and DSP based hardware that encodes each of the two audio signals with the appropriate HRTF filters. Figure 2 shows the effect of head movement both with and without head tracking applied.

As with standard binaural systems headphones are required for sound reproduction, however they now require a positional sensor to be attached or built in.

2.3 *Transaural*

With binaural encoding it is necessary to use headphones for playback. Transaural systems use crosstalk cancellation to enable HRTF processing to be used with loudspeakers. Any number of loudspeakers may be used and they may be located at any angle with respect to the listener. Figure 3 describes the principles behind the shuffler crosstalk cancellation network introduced by Cooper & Bauck [5]. The signal S represents the transfer function from the loudspeaker to the nearest ear and the signal A represents the transfer function from the loudspeaker to the opposite ear. The functions S & A are defined as the HRTFs from each speaker to each ear.

Transaural coding gives good auditory localisation and avoids the use of headphones in the playback system. The correct combination of signals from the N speakers only occurs within a relatively small "sweet-spot" - making transaural most suitable for individual workspace implementation rather than in a multi-user environment.

2.4 Ambisonics

Ambisonic theory presents a solution for encoding directional information into an audio signal [6,7,8]. The signal is intended to be replayed over an array of at least 4 (for a horizontal plane only - pantophonic system) or 8 (for a 3-dimensional space - periphonic system) loudspeakers. As with a binaural system, the sound space can be recorded with a specifically designed microphone [9] or synthesised).

The signal, termed B-Format, consists of 3 components for pantophonic systems (W the ambient component, X the front-back component & Y the left-right component) and 4 components for periphonic systems (W, X, Y & Z the up-down component). The 3-dimensional spatialised sound to be encoded is placed within a notional unit sphere

During the encoding stage no knowledge of the loudspeaker positioning is necessary - the decoder is programmed with the loudspeaker layout. The output at each of N speakers in an equally spaced pantophonic array is given by: $P_n = \frac{1}{N}W + 2X\cos(\varphi_n) + 2Y\sin(\varphi_n)$

Where W X and Y are defined as: $W = S \cdot \frac{1}{\sqrt{2}}$ $X = S \cdot \cos(\varphi)$ $Y = S \cdot \sin(\varphi)$

Ambisonic encoding provides a realistic spatial audio percept, however because the theory is based on wavefront reconstruction the correct summation of the loudspeaker outputs only occurs within a small "sweet-spot", as with transaural systems, making an ambisonic system most suitable for personal workspace environments. An advantage of ambisonic systems over transaural systems is that the loudspeaker layout does not need to be known at the encoding stage, however a greater number of loudspeakers are required.

3. Practical Systems

In this section we shall consider some examples of immersive environments which employ 3-Dimensional spatial audio. All of the examples described are actual working prototype systems that are used for a variety of applications such as teleconferencing, education and entertainment.

3.1 SmartSpace Chair

The SmartSpace chair [10,11] (as shown in Figure 4) combines a video screen, computer terminal, video camera and spatialised audio with a chair. The chair has been developed by BT to suggest an alternative to the traditional office desk. The video screen wraps around the user to provide a feeling of visual immersion for use in telepresence applications. The chair currently has two loudspeakers mounted below the screen at either side. By using transaural processing it is possible

to create a "sound bubble" around the user giving them the feeling of complete audio immersion.

3.2 Network Spatial Audio Server

The Network Spatial Audio Server (NSAS) provides a networked shared meeting space with high quality spatialised audio. Each terminal consists of networked PC connected to two central servers comprising a spatial audio server and virtual world server. Each user's screen shows a view of a computer-generated room with doors to other rooms (as shown in Figure 5). The user is represented by a computer-animation known as an avatar. The audio output from each user is picked up by a microphone which is fed into the Spatial Audio Server which in turn feeds a spatial audio signal to each of the others users. By using the system it is possible to have meetings with other people in virtual rooms and it is possible to move from room to room meeting with different people in each room. The NSAS system includes room acoustic modelling to provide realistic audio spatialisation and localisation [12]. The audio spatialisation and rendering is carried out (obviously in real-time) on a Lake Huron system [13].

3.3 VisionDome

The BT/ARC VisionDome [14] is an interactive immersive environment comprising a large plastic dome (capable of holding ten to twenty users) with a large curved video screen mounted on 180° of the internal surface (the screen is curved in two dimensions) – as shown in Figures 6 & 7. A video projector is mounted in the centre of the dome with special lenses to give the full field of vision. The VisionDome can be used for telepresence, remote control (controlling a satellite in orbit for example) etc. The main difference between the VisionDome and the systems described above is that it is intended to be a multi-user interface which requires both audio and video systems that are able to provide immersion for more than one person at once.

Within the dome there are nine loudspeaker channels: front left, front right, screen left, screen right, screen centre, rear left, rear right, rear centre and a sub-woofer. The screen channel speakers are placed behind the curved projection screen - with suitable equalisation applied. Each channel has a separate feed from the audio spatialisation system (a Lake Huron running a 3-Dimensional panning and a 3-Dimensional ambisonic-decoding program).

The input audio signals come from a variety of sources such as hard-disk recordings (ambisonic B-Format), digital and analogue video players (stereo), computer systems (stereo), samplers (multi-channel), telephone (mono) and microphones (mono). Typically for a video presentation a B-Format ambisonic recording is played from a hard-disc recorder using SMPTE frame synchronisation. For a computer generated application (which may be interactive) the computer sends MIDI signals to a multi-channel sampler which plays the desired sample when required and also sends (via TCP I/P) the position of that sample in 3-D space to the Huron which spatialises the sound accordingly. It is thus possible (by continuous update of the sound's location) to make sounds move in any desired trajectory around the listening space. At the heart of the system is a

digital mixing desk (enabling digital recordings to be mixed and spatialised in the digital domain) which facilitates remote operation via MIDI and SMPTE. Stereo and mono sources are usually patched to the corresponding speakers in the dome without any special spatialisation being applied.

Because the VisionDome is a hard plastic structure and because it has a concave screen certain acoustical challenges exist which need to be overcome in order to be able to deliver compelling spatial audio. When the dome is unfurnished and untreated a "whispering gallery" effect is obtained due to the acoustic reflections. This problem is overcome by treating the area behind the screen with acoustic tiles and the other areas with black carpet.

The VisionDome is an immersive environment for multiple users and therefore the audio spatialisation must produce a compelling percept for all of the users wherever they are located. Ambisonic playback, whilst excellent when in the sweet-spot, can fail to correctly spatialise sound when the listener is positioned, say, near a loudspeaker – the sound tends to collapse to that loudspeaker. Consider the case where a user is located near a rear speaker and the sound source is moving from front left to front right loudspeaker. Because the user is listening to mainly the anti-phase signal they hear the sound move from right to left – the opposite direction! When the audio is accompanying a video image the multi-modal mismatch consequences are catastrophic. The authors have proposed a method of warping the ambisonic signal to reduce the level of the anti-phase signal and yet enhancing the spatial percept over the whole listening-area [15].

The proposed method was implemented and tested in the VisionDome with pre-recorded ambisonic B-Format recordings and the spatialisation assessed in a series of informal listening tests. The test subjects agreed that the ambisonic material was preferable when located in the small sweet-spot, however at any other location in the listening area the ambisonic signal failed to spatialise correctly and trajectories reversed depending on listening position – confirming the theoretical expectation. The test subjects also agreed that the warped ambisonic signal gave a realistic spatialisation over all of the possible listening-positions within the VisionDome.

4. Conclusions

Immersive environments for teleconferencing represent an exciting future for interpersonal communication systems. Some systems, such as the SmartSpace chair, are designed for a single user whilst others, such as the VisionDome, are designed for use by a group of users. The combination of spatialised audio and video provide a level of immersion previously unattainable. Immersive environments may be located a few metres or thousands of kilometres apart - using digital networks to transfer visual and auditory information. In single-user applications ambisonic B-Format coding is a convenient method of transmitting N audio channels over M transmission channels - where $N >> M$. In multi-user environments it is necessary to warp the ambisonic signal (presented here) to prevent non sweet-spot listeners from hearing out of phase signals (which could result in a percept of sounds travelling in the opposite direction to that intended).

5. References

[1] Hollier.M., Rimell.A. & Burraston.D. *Spatial Audio Technology For Telepresence*
To be published in British Telecommunications Technology Journal, October 1997

[2] Walker.G., Traill.D.,Hinds.M.,Coe.A. & Polaine.M. *VisionDome: a collaborative virtual environment.*
British Telecommunications Engineering Vol. 16, October 1996
also at: http://www.labs.bt.com/people/walkergr/IBTE_VisionDome/index.htm

[3] Moller.H. Fundamentals of Binaural Technology
Applied Acoustics, Vol. 36, 1992, pp 171-218

[4] Lake DSP
Huron Digital Audio Convolution Workstation User Manual, Section 2-21

[5] Cooper.D. & Bauck.J. *Prospects for Transaural Recording*
Journal of the Audio Engineering Society, Vol. 37, No 1/2, 1989, pp 3-19

[6] Gerzon.M.A *General Metatheory of Auditory Localisation*
Presented at the 92nd convention of the Audio Engineering Society 1992,
Preprint No 3261

[7] Gerzon.M.A. *Surround Sound Psychoacoustics*
Wireless World, Vol. 80, 1974, pp 483-486

[8] Malham.D. & Myatt.A. *3-D sound Spatialisation Using Ambisonic Techniques*
Computer Music Journal, Vol. 19, No 4, 1995, pp 58-70

[9] Gerzon.M.A. *The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound*
Presented at the 50th convention of the Audio Engineering Society 1975

[10] Innovation 1997 Web Site http:\\i97.labs.bt.com

[11] Foo.K. & Hawksford.M. *HRTF Sensitivity analysis for three-dimensional spatial audio using the pair-wise loudspeaker association paradigm.*
To be Presented at the 103rd convention of the Audio Engineering Society 1997

[12] Burraston.D., Hollier.M. & Hawksford.M.O. *Limitations of dynamically controlling the listening position in a 3-D ambisonic environment* Presented at 102nd AES Convention March 1997 Audio Engineering Society
Preprint No 4460

[13] Lake DSP Web Site http://www.lakedsp.com

[14] VisionDome Web Site http://www.virtual-reality.com/visiondome.html

[15] Rimell.A. & Hollier.M. *Reproduction of spatialised audio in immersive environments with non-ideal acoustic conditions.*
Presented at the 103rd convention of the Audio Engineering Society 1997
Preprint No 4543

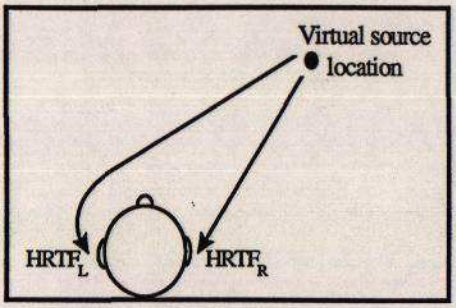

Figure 1: Binaural signal encoding

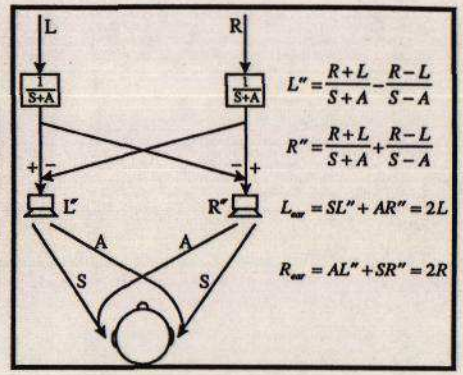

Figure 3: Crostalk cancellation as used in transaural systems

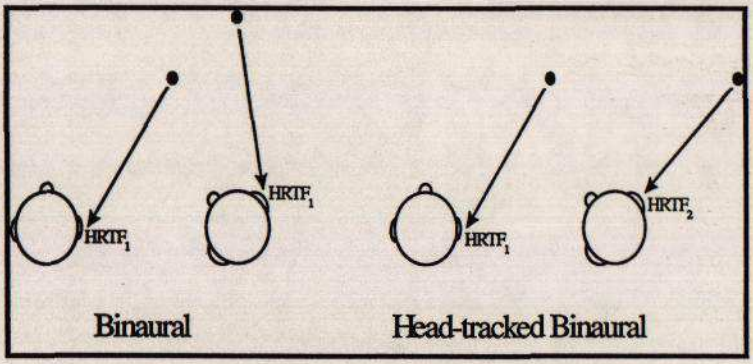

Figure 2: Binaural signal with head-tracking

Figure 4: SmartSpace chair

Figure 5: NSAS terminal being assessed by a prototype multi-modal analysis tool

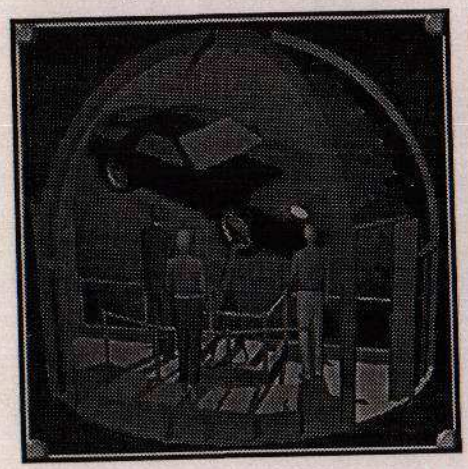

Figure 6: Cutaway drawing of the BT/ARC VisionDome [14]

Figure 7: The VisionDome displaying a pre-recorded historical fly-through