Ambisonic sound in virtual environments and applications for blind people

M Cooper and M E Taylor

Institute of Educational Technology/Knowledge Media Institute/Multimedia Enabling Technology Group,
Open University, Berrill Building, Walton Hall, Milton Keynes, MK7 6 AA, UK

m.cooper@open.ac.uk, *m.e.taylor@open.ac.uk*

ABSTRACT

To date there has been much more effort directed to generating credible presentations of virtual worlds in a visual medium than in reproducing corresponding or even self contained worlds with synthesised or recorded sound. There is thus today a disparity in the relative fidelity of these 2 modes in most VR systems. While much work has been done in hi-fidelity and 3 dimensional sound reproduction in psycho-acoustic research and applications for audiophiles this has rarely been taken onboard by the VR community.

This paper describes work ongoing to apply Ambisonic techniques to the generation of audio virtual worlds and environments. Firstly Ambisonics is briefly outlined then principles behind its implementation in this context described. The design of the implementations to date is described and the results of trials discussed. The strengths and limitations of the approach are discussed in the light of this practical experience.

There is a range of applications envisaged for this technique that would be particularly of benefit to disabled people. The 2 principal areas under consideration by the authors is the extended use of the audio mode in HCI for people with disabilities and VR applications for blind and partially sighted people.

Both of these areas are described and specific examples of applications given in each. The limitations given available low cost technology are highlighted and the technology evolution required to make such applications widespread commented on.

The results of the development work and subsequent user trials undertaken to date are given and discussed. Lines of further exploration of this technique and its application are outlined.

1. INTRODUCTION

Ambisonics is a method of recording and reproducing sound in 3 dimensions. Advances in digital audio and computing now hold the possibility to exploit this approach to facilitate computer synthesis of 3 dimensional audio environments. Applications of this approach have been envisaged offering particular benefits for visually impaired people. Detailed feasibility studies, user needs capture and pilot implementations are now being undertaken to prove the validity of the approach and that it will meet identified needs. This is the subject of this paper, which is a description of work in progress together with references to related work of others and notes on the aspirations of potential users that are directing it.

2. AMBISONICS AN OVERVIEW

Ambisonics, originally developed in the 1970s, is a method of capturing, encoding for recording and reproducing sound in 3 dimensions. A very brief, largely non-technical, overview of Ambisonics is given here to set the rest of the paper in context. Full descriptions of the technique with its psycho-acoustic and mathematical justifications can be found in the key papers of the field's founding fathers [Fellgett 72, Gerzon 74, and Gerzon 92]. A paper outlining its advantages in audio virtual reality was given at the first ECDVART conference [Keating 96].

In the capture of sound Ambisonics detects not only the nature of the sound but the direction at from which it arrives at the point the recording is made. The term “soundfield” is coined to describe the sound impinging on an imaginary sphere around the listening position. The goal of the technique is to be able to record a soundfield at one time and place and then reproduce it at another.

In first order Ambisonics, the only degree of complexity implemented to date, the sound is ideally captured by a set of 3 dipole microphones (i.e. with figure of eight response patterns) coincidentally located and lying along the axes of a Cartesian co-ordinate system. This is physically unrealisable, however the commercially available Soundfield Microphone effectively accomplishes this by having 4 microphone capsules arranged in a tetrahedron then combining the outputs of these in such a way as to yield the required response patterns. The fact that the capsules are not truly coincident is compensated for electronically. The microphones output is 4 separate signals conventionally labelled X, Y, Z and W. The X signal represents the sound components from the front minus those from the rear and similarly Y left minus right, and Z up minus down. The W signal is a non-directional reference signal generated from the combined outputs of all of the microphone’s capsules. Within Ambisonics these W, X, Y, and Z signals are known collectively as “B-Format”. In the applications discussed in this paper, a microphone is not used to capture a “live” soundfield but the same B-Format signals are generated from a computer model of the sound sources and their environment.

3. TECHNOLOGY CONTEXT

A key reason for Ambisonics failing to gain popularity within the audio industries in the 1970s, when it emerged, is that they could not at that time foresee the required 4 channels being available for broadcasts or recording media for the domestic market. However in the last year or so this has become readily possible with the arrival of digital broadcasting and the DVD (Digital Versatile Disc) recording format.

The “digital revolution” that has occurred in sound recording and processing, combined with the increase of computing power available in standard desktop PCs, means it is now practical to use Ambisonic techniques for the creation and reproduction of audio virtual environments. Software tools have been developed, which enable a standard PC equipped with a suitable multi-channel sound card to be used both for the creation of the B-Format as digital signals, and the processing needed for their playback over a particular loudspeaker arrangement [Farina 1998]. There is a commercially available and very powerful system, from Lake DSP of Australia, which facilitates the same (potentially alongside the real-time generation of the virtual environment) by using dedicated digital audio signal processors (DSPs), connected to a standard PC. Both these approaches are based on digital algorithms for the fast implementation of a mathematical process called “convolution”. The response of any system to a given input is the convolution of its impulse response with that input. Acoustically the impulse response can be viewed as the result of a gunshot within the acoustic space.

Other systems have emerged for the creation of a 3 dimensional sound effect (e.g. Dolby MP encoding, and Dolby Surround and Pro Logic decoding, DTS, THX, etc.). Theses are all devised to give an impression of sound surrounding the listener but are unable to precisely locate a sound source from the side or the rear. These other techniques are mainly used in cinema where the focus of attention is towards the front. Further, the sound is combined with a wide screen high quality picture, thus the effect of these sound processes on the overall perception of the experience is significant but the brain effectively ignores any spatial imprecision in the sound heard. Thus Ambisonics remains the best available technique for generating of the soundfields for the applications principally directed toward people with a visual impairments described in this proposal.

4. APPLICATIONS FOR BLIND COMPUTER USERS

4.1 Virtual reality for blind people

Virtual reality (VR), in its many guises, has been held up as a potentially powerful tool in education and training. Indeed that potential has now been demonstrated in various commercially available and in house software packages. Much of the work to date in developing virtual reality systems has concentrated on the visual modality. However if the potential of virtual reality is to be extended to blind users then the computer generation of credible audio worlds is required. Further this would be of benefit in a wide range of VR applications for users without a significant visual impairment. There is some evidence to the effect that the perceived veracity of a virtual world is more dependent of the fidelity of the audio rather than visual representation of that world. We appear to be able to much more readily "suspend disbelief" in what we see than what we hear.

4.2 Sound Environments as access to GUIs

The move in personal computing over the last 10 years, almost universally, to Graphical User Interfaces (GUIs) is an example of advances in usability for the majority creating barriers for one user group; blind people. Various approaches have been have been developed to address this problem and have met with mixed results. These include the use of tactile displays, extended speech synthesiser systems and the modal transformation of icons to "audicons", and approaches that seek to combine these. In the audicon approach a sound representative of the icons function is given to the user when the cursor within the GUI is over a particular icon. It is suggested that the usefulness of this approach could be greatly enhanced if the position that this audicon appears to emanate from directly maps to the position of the corresponding icon within the visual display.

Some work has been done to investigate this within the EU TIDE programme sponsored GUIB project [Crispen and. Petrie 1993]. The approach taken in their work was based on the direct calculation of the signals for feeding into the headphones of the user based on a modelling the Head Response Transfer Function (HRTF). If movements of the head are need to be accounted for as would be the case in many practical applications this approach becomes very computationally intensive [See Keating 96]. In terms of using sound to access a GUI the approach taken within the GUIB project was to us a sound cue for the cursor position and a separate one for the icon location. The user was then required to control the cursor through a mouse to bring the first sound to the other. The evaluations undertaken indicated that there were still significant challenges in taking this forward to a practical system. They found that the acuity in the vertical dimension was very much less than the horizontal as would be expected from the theoretical understanding of auditory location. It is not possible to determine from the published results to what degree this was effected beyond the human perceptual limits by their implementation. Certainly inaccuracies in the HRTFs used would have particularly affected the vertical acuity.

The authors of this paper would like to suggest that an alternative model for audio interaction with a GUI than that adopted in the GUIB project would be more powerful and less effected by the limitations of the technology and human perception. The sound synthesised by the computer could be done from the perspective of the user being at the cursor position. Thus the directional information from the available audicons would be in direct relation to their position with respect to the cursor. Thus the relative position of the sound rather than an absolute location becomes the important factor. The scale and mapping of the perceived sound space could be such that accounts for the different spatial perception of individuals and in the different aspects. This is a key direction in the ongoing work and many perceptual and technical factors need further research to validate this approach.

As well as facing particular challenges in the use of GUIs, blind computer users face increased challenges when seeking to learn to use a new operating system or application software. A significant disadvantage for blind computer users compared with their sighted colleagues is the time required for them to learn to such software, particularly when it is GUI based. This is principally due to the difficulty of learning by exploration and experiment. The following indicates a learning strategy typical of sighted computer users learning to use a new piece of software:

"Oh, what does this icon do?"

"Let's select it and see."

The challenge for the blind computer user even given currently available assistive technologies is firstly how to identify the existence of an icon that potentially offers a useful function and then how to evaluate its action. An extension of the spatially located audicon approach could significantly address the first of these and assist with the second.

One can envisage an exploratory mode for interacting with a GUI through audio. In this mode the audicon would be triggered when the cursor was within a given radius of the icon. The volume of the audicon would then increase as the cursor moves toward the icon. The advantage of this is that the user would have a greater awareness of the location of the cursors position within the arrangement of icons and of the existence of the available icons. The potential for this mode resulting in a meaningless noise when the cursor is in the vicinity of multiple icons is largely obviated by the spatial location of the audicons. Human audio perception is very good at focusing its attention at a sound from a given direction even given the presence of a high level of other sounds from other directions (the "cocktail party effect"). The phase "mumbling icons" has been coined to describe this approach. Given the successful implementation of spatial audicons the extension to this exploratory mode is technically relatively trivial but will need extensive user trials to arrive at optimal settings of the various system parameters.

5. WORK IN PROGRESS

5.1 Generation of Simple Audio Virtual Worlds

To date only a basic system for synthesising Ambisonic sound due to a single sound source in a user defined rectilinear world has been developed. This uses simple inverse ray tracing techniques to calculate the resulting B-Format signals at the "listening" position due to a stationary or moving source within the world. Any Wave Format (.WAV) file can be used as the source and the resulting B-Format signals are calculated and stored as 4 individual .WAV files. This is an off line calculation of the audio virtual world. The 4 B-Format signals are then played out through a multi-channel PC sound card (Gadget Labs™ Wave/4) using commercially available sound studio software (Cool Edit Pro). It should be noted that it was not possible to use standard SoundBlaster™ compatible PC sound cards because although it was possible to install 2 cards in a single PC it was not possible to exactly synchronise the outputs from both cards.

The playback facility is a regular array of speakers at the corners of 2 orthogonally bisecting rectangles, one in the horizontal plane, at the level of the listening position, and the second in the vertical plane passing through the listening position from front to back. A key feature of Ambisonics is that the signals recorded, or synthesised, are independent of the configuration of the playback speakers (unlike Dolby 5.1, etc.). Thus the particular array chosen was mainly determined by the to use 2 of the speakers for stereo work at other times and the physical constraints of the sound booth. An 8-speaker array was selected as previous work with Ambisonics had shown this was a practical minimum for a stable reconstruction of a 3 dimensional soundfield. The system was installed in a sound booth that had some acoustic treatment but was only a "semi-dead" acoustically and in no way could be considered anechoic. Currently a speaker decoder, that takes the analogue B-Format signals and derives the individual speaker feeds has been constructed in-house with the parameters as calculated from those given in Gerzon 1980 for the particular speaker array installed. Others, [Farina and Ugolotti 1998] have developed and demonstrated the use of software decoders working on the digital B-Format signals but this for the configuration described then requires 8 soundcard audio outputs. This is perfectly possible with the above listed software by installing a second Wave/4 soundcard but was not selected as the initial route.

The choice of the sound source needs, if fidelity is the objective, to be recorded by close microphone techniques and not subject to any artificial reverberation treatment and the same applies for synthesised sources. That way it is the acoustic of the virtual world that determines the sound as calculated for the listening position. The current version of the software enables any rectilinear space to be modelled, with any number of rectilinear features, of different acoustic properties, on any of its surfaces. The sound source and the listener can be placed anywhere within the room and the sound source moved through any path that can be described as a timed series of Cartesian co-ordinates. Acoustically the modelling is very simple with only surface absorption and inverse square law effects being taken into account.

Trials with multiple users with and without a visual impairment of this system described begin this autumn. The initial trials to date have confirmed that for most people a believable sound image is created but there is some variability between subjects as to the perceived location of a sound within the world. What has been demonstrated to date is the feasibility of the basic approach and useful information gleaned as to the computation levels required. The typical time of calculation of the B-Format signals due to a 1-second source signal on a 133MHz Pentium PC is about 20 seconds. This is of course subject to a lot of variation with programme and virtual world parameters. The most dominant here being the ray step and acceptance angle within the inverse ray tracing algorithm the above is with 0.5 degrees set for each.

There is now an ongoing programme of work seeking both to increase the complexity of the audio worlds modelled and to arrive at optimal calculating of these so that interaction with them can be achieved in real time. This will be based on the use of a Digital Audio Convolution Processor (a multi-channel DSP board and development environment) from Lake DSP (see http://www.lakedsp.com/products/index.html for further information.)

5.2 Justification for Approach

There is an issue here of why apply this high level of, currently expensive, processing power in researching and developing a technology when the vision is for it to become widely used in education and the home? The answer is principally in the lead-time of the proposed work. The research is directed towards applications that may only come into widespread use in 3 to 5 years time. With the current speed of evolution in computer technology it is highly likely that the necessary processing power will be readily available within that time scale on standard PCs. Further developments in computer related technology such as DVDs and PC Sound Card technology will mean the that peripherals that facilitate the implementations that require less processing power will also become available. A significant part of the research and development work outlined will be in the ensuring that the applications will run efficiently on platforms readily available to the target users. Much can be achieved towards making the software more efficient once the effects of the various software parameters on the perception of the users is fully understood but greater levels of processing power are required to fully investigate this in the first place.

5.3 Workshop of Blind Computer Users

A workshop was held with a small group of blind people, on 4 July 1998, to introduce them to the basic concept of 3 dimensional audio and encourage them to brainstorm on potential applications that could be of use to them or other blind and partially sighted people. The majority of the visually impaired participants were current or former OU students and this gave an educational bias to their perspective. Some of the key potential applications to arise from this are listed here but to set them in context the make-up of the workshop is tabulated:

Total number attendees (including facilitators)	14
Number of attendees having a significant visual impairment?	7
Self-identifying as blind	4
Sighted assistants	2
Involved in the development of computer or audio applications for visually impaired people	6
Having no visual memory	1
Number of those with a significant visual impairment describing there current computer usage as at least once per week	5
Of those with a significant visual impairment current usage of assistive technology when using a computer:	
Speech output	4
Braille Displays	-
Enlarged VDU Displays	1
Sighted Assistant	3

Key suggestions for the application of 3D audio virtual environments that emerged from these discussions were:

- Modelling of physical relationships of objects etc. - E.g. Model of the Solar system; Electro Magnetic Fields and their interaction; positions of fielders in a cricket match.
- Use in GUI-controlled Desk Top Publishing and Spreadsheets, etc - Use of 3D sound to give a better indication of where objects were on a page and for moving objects around, flowing text into boxes, etc. Use of sound cues to indicate colour; text attributes such as **Bold** and *Italic*.
- General improvements to the accessibility of a GUI by enhanced sound cues (a need for a common protocol for these cues was identified)
- Simulation of a work or social situations - e.g. in social science study of group situations
- In tele-conferencing position of speaker indicated by position their voice emanates from - turn-taking cues.
- VR - Suggestions of walkthroughs of buildings, London Underground - The idea of "hearing" walls, i.e. introduction or training in the use of echolocation. Navigational training in general.
- Sound maps

5.4 Investigations in Audio Perception and Practical Ambisonics

The limitations of both human audio perception and the practical implementation of Ambisonic theory both need to be further investigated to confirm that the envisaged applications for blind people are in fact viable and practical and to inform further development work.

Key questions to be addressed are summarised in the table below:

In human audio perception	**In practical Ambisonic implementations**
• What is the resolution and variability of human spatial discrimination in sound	• What spatial precision is required in the computer modelling of the soundfield to meet the limits of human perception
• Is the expected sense of "being present" in an audio virtual environment achieved with the implementations made and what are the key success factors for this	• Can virtual environments of sufficient complexity to be believable be created "real-time" given the available technology
• What are the particular requirements for visually impaired users of audio virtual environments where others use the visual modality. (e.g. GUIs)	• What is the most appropriate mapping of a 2D graphical display into 3D audio environment (e.g. GUIs)
• What is the effect of a 3D audio representations on exploration of a GUI by blind computer users	• In an implementation of "mumbling icons" what are the optimal system parameters for different tasks

These questions will be addressed in a series of simple psychophysical experiments undertaken with sighted, blind and partially sighted users. The visually impaired subjects are being recruited from OU students in the regions around Milton Keynes and others involved in local blind and visually impaired groups in the area (e.g. British Computer Association of the Blind). A series of inter-related experiments is being planned over the next year with 20 subjects attending at the laboratory on 3 separate occasions each. It is judged that valid detailed methodologies can be constructed with such sample sizes to meet the research objectives. A challenge for the detailed design of the experimental method is to isolate those factors due to the technology and those due to human perception.

6. REFERENCES

K. Crispen, H. Petrie (1993) Providing access to GUIs for blind people using a multimedia system based on spatial audio presentation. *Pre-prints of the 95th AES (Audio Engineering Society) Convention*, New York

A. Farina, E. Ugolotti (1998) Software Implementation of B-Format Encoding and Decoding *Pre-prints of the 104th AES Convention*, Amsterdam

P.B. Fellgett, (1972 Sept.) Directional Information in Reproduced Sound, *Wireless World*, vol. 78, pp. 413-417

M.A. Gerzon, (1974 Dec.) Surround Sound Psychoacoustics, *Wireless World*, vol. 80, pp. 483-486

M.A. Gerzon, (1980 Feb.) Practical Periphony, The reproduction of full sphere sound. *Pre-print of the 65th Audio Engineering Society Convention*, London

M.A. Gerzon, (1992 Mar.) General Metatheory of Auditory Localisation, *Pre-print 3306 of the 92nd Audio Engineering Society Convention*, Vienna

D.A. Keating, (1996) The generation of virtual acoustic environments for blind people, *Proc. 1st European conference on Disability, Virtual Reality and Associated Technologies,* pp. 201-207, Maidenhead UK