Factors Affecting Sound Synthesis and Presentation Within Virtual Reality Systems

M Griffin (1) & D Keating (2)

Dept of Cybernetics, University of Reading

1. INTRODUCTION

In this paper, the problems associated with sound synthesis and generation within a virtual world environment will be discussed. In this paper the term 'Virtual Reality' has been reduced to 'VR' for the sake of brevity. The field of VR is relatively new in nature and poses some unique problems in audio theory and design. The object of this paper is to identify and discuss these problems and the potential solutions to them.

1.1 General Discussion

The field of virtual reality has recently developed out of research pioneered in the fields of computer graphics and human/computer interaction. The basic concept within this field, is the "immersion" of an operator in an artificial world, such that sensory impressions conveyed to the operator enforce the impression that this artificial world is "real". The model is generated by specialist computing hardware, and suitable displays are used such that the required impressions are impressed onto the operator. Wherever possible extraineous impressions from the normal surroundings are excluded.

A typical VR system composes of a headset system, utilising liquid crystal displays and appropriate optics, such that the wearer can be presented with images from these displays. A set of headphones supplies auditory information to the wearer (whilst also muffling out unwanted extraneous environmental noise). Finally, a tracking system is also included in the helmet, such that the wearer's head position and attitude may be determined (see figure 1).

On some of the more advanced systems, a "dataglove" system is also utilised. This is a special pair of gloves that the operator also wears. They contain similar tracking systems as the helmet, as well as optical fibre flexion sensors. From these elements the hand and finger attitude and position may be determined. Tactile feedback is often included in such gloves, either embodying piezo electrical "buzzers" in the finger tips, or inflatable air pockets. Thus an operator can be given the capability to pick up and touch items within a virtual reality.

A related field to virtual reality, is that of telepresence. In this the visual and auditory information is not generated, but merely captured from a remote location. The helmet and glove system are still utilised, but are used to control a robotic system at the remote site. The robotic system is sufficiently anthropomorphic such that control and information capture are easily facilitated (see figure 2).

1.2 Applications

Currently, most applications of virtual reality lie either in the areas of computer aided design visualisation, military simulation and entertainment. In these, a model of a building, or a battlefield may be presented to the user, such that the salient points within the simulation may be interacted with and tested. A typical example is the Kitchen Design System installed in a Tokyo department store. In this, a representation of the client's kitchen may be created, and various new appliances and arrangements may be introduced. The client, donning the VR system can "see" the arrangements, and even open cupboard and oven doors to check their placement. Once the client is satisfied, the appropriate items may be purchased and installed, just as the client saw in the VR system.

With respect ot the telepresence system, the majority of uses rest in the area of remote manipulation of materials in hazardous environments. With this, it is possible to place an anthropomorphically arranged robot, such that examination and manipulation of materials, is an intuitive and natural operation. A typical application is the NASA anthropomorphic robot, which will permit an operator to assemble complicated structures, without having to resort to donning a space suit.

1.3 Sound in Virtual Reality

So far in this paper, emphasis has been placed on the visual aspects of virtual reality. This is in essence, due to the considerable dominance of the visual system over hearing, and the current body of expertise concentrated in this area. However, it has been noted that the introduction of more subtle stimulations on an "immersed" operator has a profound effect on the perceived level of realism. As hearing is the second most dominant sense, it is vital that an accurate representation of sounds and their position relative to the operator is made.

1.3.1 Sampled Sound Effects

As the mechanism generating the sensory illusions is effectively a computer, it is natural to use sampled sound effects. Most VR systems utilise this in some manner or other. These samples can be either digitised from analog recordings, or synthesised in some manner. The sampled sounds are replayed when appropriate events are triggered within the virtual world. A typical event might be the opening and closing of a door. Upon this event, the system would generate the sound effect of creaking hinges when the door was opened, as well as the appropriate visual imagery.

To convey a correct impression as to the position of sound source relative to the operator, one of two approaches is required. Either the sound is captured using a dummy head system (see later), or appropriate processing is applied to the raw sample such that suitable modification is applied. Once this has been performed, then the sound may be played back to the operator.

In this situation, there is also the problem of modelling the modifications to the sound, by the environment. Typical examples of this are room resonances, absorbtion and reflections. These need to be applied such that they are consistant with the sound source's position and the relation to the observer.

1.3.2 Real Source Sound

In certain instances, it may be necessary to play "live" auditory information to the operator. An example of this is if the two or more people are interacting within the same virtual world. In this case it is desirable that the speach from each party, may be relayed to the other, such that it appears to come from their correct location within the virtual world. Furthermore, any modifications expected, due to environmental effects (as above) should also be created.

1.4 Sound in Telepresence

With a telepresence system, accurate capturing of the the sound relations around the observing unit, must be performed. This must be consistent with the direction of gaze of the observation unit, and accurate rendition of this imagery must be made to the observer's ears. If the direction of gaze changes, the transfer function used to produce the accurate rendition, needs to altered accordingly. Strategies for gathering the sound information, in the first instance, have to be discussed. In essence, there are two recommended:

1.4.1 Dummy Head Recordings

In this, the observation platform is equiped with a recording dummy head. Cameras and microphones are placed, such that they duplicate the normal anthropomorphic arrangement. By fitting artificial pinnae, and ensuring that the arrangement tracks accordingly, then sound is captured, and an accurate binaural representation produced. This approach, ensures accurate capturing of auditory imagery by applying the appropriate transfer function, using mimicry. Although simple to effect, it has the disadvantage that the receiving platform must conform to an anthropomorphic arrangement. In many instances, this may not be possible due to size or other constraints, so an alternative method must be used.

1.4.2 Sound Field Recordings

An alternative approach to using dummy heads for sound capture, would be to use a format like "B-Format" Ambisonics. In this, information is captured using a sound field microphone. This produces four channels of information; front/back, left/right, up/down and omni. All the information required to construct the positions of sound sources to the microphone is present in this format. The soundfield microphone is physically less bulky than the dummy head arrangement, and has the added advantage that much of the repositioning required for direction of gaze alterations, may be performed in relatively simple analog hardware. Further to this, it is possible to "concentrate" the microphone on a potential point of interest, whilst still retaining any background information necessary. Similarly, a distracting noise source may be "nulled" out in a similar manner.

The disadvantage with this technique, is that a conversion from ambisonics to binaural is required for use with a standard VR headset. Work by Gerzon indicates this is possible, and a simple approach has been experimented with. Refinement of this should easily yield a more satisfactory result.

2. Presentation of Auditory Information

So far, sketchy details of the problems to be expected and solved have been laid out. The subject area will now be examined in more detail. Firstly, issues of the presentation sustem will be discussed.

2.1 With Respect to the VR Headset

Typically, a virtual reality headset will be equipped with reference quality, closed back headphones. A typical manufacturer utilised is Senheiser. They are usually moving coil in nature and made an integral element is the headset design. Their fuction is to both present auditory information to each ear, and to block out any outside noise. Due to the fact that closed backed headphones are used, this alters the effective conch resonant frequency. This alteration is unwanted and alters the sound perceived from the headphones. Compensation is therefore required. This compensation takes the form of suppressing the shifted conch resonance, and introducing an "natural" resonance effect in the original position.

The above operation may be performed either as an analog operation, or using digital signal processing technology. It is the last to be performed on the sound information, before presentation to the user. The actual compensation required is headphone dependant, but some manufacturers now already have the details for this. Unfortunately, in most systems, this compensation is not currently implemented.

The essence of VR is not just the ability to visualise information in a "real" manner, but to also be able to alter one's viewpoint within the model being examined and interact with this model in a meaningfull way. For this to be possible, the position and orientation of the observer's body must be determined continuously, and the alterations this causes in the way the sound is modified recomputed. As most VR systems operate at a cycle rate of 50-60 Hz, these modifications are almost a continual problem. If a DSP filter has been used to generate the transfer function to binaural, this will need regular reconfiguring to cope with the constant repositioning. This reconfiguring is likely to cause discontinuities in the sound produced. Suitable filtration of these discontinuities is thus necessary after the binaural conversion.

2.2 Dummy Head Telepresence Operation

With telepresence systems, the binaural conversion is performed by the careful arrangement of microphones on a dummy head. Therefore, a straight presentation of the sound information may be performed, from the microphones, to the headset. As with the virtual reality case, as the operator's head moves, the system has to track this then update the position of the dummy head recording mechanism to match. As with virtual reality, the natural conch resonance of the human ear is altered when utilising the same headphone configuration. To compensate for this, the processing stage listed above must be used. Most telepresence systems do not do this.

2.3 Utilising 3D Ambisonics in Telepresence

From above, it may be seen that although using a dummy head simplifies the sound capturing process, with respect to electronics, it is not terribly suitable for systems where bulk is a constraint. A more elegant solution to the problem must be found. A natural contender for this is to utilise soundfield microphones, and ambisonics. With this a very compact measurement system may be created. Rotation of the observer's head may be matched by rotating the channels appropriately in ambisonics. Thus a mechanical realignment system may be avoided. Unfortunately, the conversion from ambisonics to binaural is still required, as well as the conch resonance compensation noted previously.

3. The Generation of Auditory Information

Having now discussed the presentation of auditory information, the generation requirements must be considered. In this the basic problem lies in the conversion of either ambisonic or monaural information, to binaural format, then adding the necessary compensation to remove headphone artifacts. The two cases where this generation is not a problem, are the dummy head telepresence system, and the use of pre-recorded binaural samples. As well as this, the modification of a basic sound by the artificial environment, must be examined.

3.1 Generation of Binaural Information

In general for humans to localise sound, three phenomina need to be modelled; interaural time differences, interaural intensity differences and spectral shaping by the pinnae. These mechanisms and their precise operation, are still subject to some debate, but a sufficient level of understanding has been gained, such that the effect of these may be synthesised. The overall operation is often refered to as a "Head-related transfer function". This function is different depending on the angle the sound in relation to the observer's head. These functions differ slightly from observer to observer, but there are sufficient similarities such that one generic head related transfer function will give an adequate image for differing observers.

The head related transfer function may be found by exhaustively frequency response testing a dummy head with sources at different locations. From these results a transfer function may be determined. To be of use, this transfer function must be converted into a suitable finite impulse response filter and this applied to the sound using suitable digital signal processing hardware. It may be assumed that for each relative orientation of the sound to the user, the associated transfer function needs to be found. This would imply that a large number of functions would need to be determined to get satisfactory coverage. This is not quite so as the localisation ability of the average person is of the order of 10 degrees. Further to this, as the head is left/right symmetric, transfer functions for the left hand side need only be swapped over for the right to create the correct image. Interpolation between transfer functions, however, is often used to increase angular definition when necessary.

As would be expected from such a system, there are frequently problems with front/back ambiguities, as well as up/down determination. If the system is permitted to allow updating of the transfer function, then by head repositioning, the observer may resolve these ambiguities.

Such an approach suffers from the problem that it can only initially work with individual sources. Multiple sources may be catered for by simple addition of the effects after the binaural calculation stage. If the source is ambisonic the conversion to binaural may be achieved by a similar process. -

Another problem with such an approach, is the amount of digital signal processing necessary to perform the operation in real time. The order of magnitude required to reposition 4 single sources, sampled to 16 bit accuracy, is of the order of 320Mips. Fortunately, as the cost of DSP hardware falls, the processing power required becomes more affordable.

3.2 The Modification of Sound By The Environment

Having discussed how to produce binaural information, it is also important to discuss how to model alterations of the basic sound, by the environment. With telepresence, this is not necessary, as the capturing apparatus, accurately records the natural modification process. With virtual reality, however, the sound received by the observer must be consistent with the environment in which the observer and the source lie. A typical example of this is the modelling of room resonance. In this, the sound received by an observer comes not only from the source, but the resonances set up in the room by the source. The resonance effect is easily calculable and the auditory contribution it represents. As resonances are standing waves, their contribution can be approximated to an omnidirectional signal, applied equally to each ear.

A second environmental effect, important for the perception of depth, is that of reflections. These can be approximated to additional sources, delayed in time and attenuated by a certain factor. This attenuation factor can be frequency dependant. The contribution to the observer would then be that of localised sources. Because of the directional nature, these will require binaural processing. Complicated effects such as dispersion, can be approximated using both of the above operations. Current VR systems only model primary resonances and reflections within a simple cuboid environment.

4. Conclusions

In this paper, we have outlined some of the considerations in producing realistic auditory images in VR and telepresence systems. Due to the interactive nature of the field there is a unique set of problems to be solved. This paper addresses some of these, and the current commerical solutions. The use of ambisonics in telepresence permits a more elegant solution to the problem of sound capture, however this has not yet been persued, despite its obvious benefits.

5. References

[1] B C J Moore An Introduction to the Phsychology of Hearing, 3rd ed, Academic Press, London, 1989.

[2] K B Christensen The Application of Digital Signal Processing to Large-Scale Simulation of Room Acoustics, J. Audio Eng Soc, Vol 40, No 4, April, 1992.

[3] D R Begault Challenges to the Sucessful Inplementation of 3-D Sound, J. Audio Eng Soc, Vol 39, No 11, November 1991.

6. Figures

Figure 1 - Typical VR Headset

Figure 2 - Typical Telepresence System

ISBN 1 873082 40 1