Proceedings of the Institute of Acoustics

A SIMPLE ACOUSTIC ROOM MODEL FOR VIRTUAL PRODUCTION AUDIO.

R. Walker BBC R&D Dept., Kingswood Warren, Tadworth, UK

1. INTRODUCTION

In television production, the use of synthetically-generated visual studio features is becoming commonplace. It has obvious potential for saving the costs of constructing, moving and storing studio sets and can also allow studios to appear to be of a different size to their actual size. The technique is generally known as 'Virtual Production' (VP).

2. SYNTHETIC VISUAL IMAGES

The creation of a VP visual image involves two main components. The studio (or, more accurately, the live performance space, since it may not be anything like a conventional studio) must exist in reality. It provides the space in which the live action can take place.

The second main component is a computer model of the virtual studio. In many cases, there will be elements common to the real and virtual spaces. Objects that are located inside the space, rather than on the boundary surfaces, will usually have to exist in both, at least in order to provide cues for the performers.

The link between the real and virtual studios is provided by a computer system, which generates the appropriate synthetic view of the virtual studio, from inputs defining the location and orientation of the cameras and the positions of moving objects.

3. THE HUMAN HEARING SYSTEM

The human hearing system is well adapted to (and therefore highly sensitive at) obtaining spatial cues from the acoustic early reflection pattern. It is also sensitive to the overall acoustic impression of the whole space. These aspects of hearing have been widely studied.

For very close sources, of the order of 2 m distance, a typical early reflection pattern is short, with close-spaced arrival times. The temporal discrimination for those short times, less than 5 ms, is very poor (almost zero). The reflected sounds interfere with the direct component to cause effects that are generally perceived in the frequency domain - as disturbances to the spectral balance. After about 5 ms relative to the arrival of the direct sound, the hearing system becomes sensitive to individual reflections, though they are not perceived as discrete echoes.

After about 50 ms, all of the remaining sound energy is summed into the general reverberation pattern, representing the surrounding space on a global scale. There is virtually no sensitivity to the detail of the later reflection pattern – with the exception of echoes. If an individual reflection has a significantly higher level than the reverberation at that time, it is perceived as a discrete echo.

4. ACOUSTICS OF ROOMS AND ACOUSTICAL COMPLEXITY

Sound emanating from a source travels outwards from the source at the characteristic velocity of sound (≈ 340 m/s). The relative intensities in particular directions are governed by the radiation pattern of the source. Because of the spreading loss, the sound intensity in a particular direction decreases at a rate of 6 dB per doubling of distance.

In any partially or fully enclosed space, that uniform spreading proceeds for only a short time, until part of the sound wave strikes some acoustically significant object. What then happens is always complicated. Sound propagates as a wave function and demonstrates all of the properties usually associated with the interactions of waves and objects – reflection, refraction and absorption. When a sound wave meets a discontinuity in the medium, the results depend on the acoustic properties of the materials and the size of the discontinuity in relation to the wavelength of the sound.

Over the normal audio frequency span, wavelengths range from about 15 mm to 7 m. That nicely encompasses most sizes of objects within rooms, and even the room itself. Thus, the interactions between sound waves and the room and its contents cover the whole gamut of reflection and refraction effects, as well as absorption. It is that complexity which makes real sound fields impractical to treat analytically. Numerical methods, like Finite Element Analysis or Ray Tracing, are also limited to fairly simple approximations.

In a typical room, there is usually at least the floor surface within about 2m of the source. Therefore, from a maximum of about 6 ms onwards, the sound field (even outdoors) contains components which have interacted with some surfaces or objects. After 30 ms in a small room the sound wavefront will have travelled in every direction to the boundaries of the room and will have interacted at least once with every object contained within.

5. ACOUSTICAL SIMPLIFICATIONS

It is obviously impractical to create even a moderately accurate objective model of anything but the simplest acoustic space. It may well be unnecessary anyway. For the purposes of improving the subjective quality of the sound in a VP production, a relatively simple synthesis is sufficient to produce a convincing audio illusion.

The most obvious simplification would be just to add artificially generated reverberation to the 'clean' sound. Done in an artistically sensitive manner, that might provide most of the illusion required. However, it would not be an automatic process and would fail if either the source or the microphone moved close to or behind a large surface (virtual or real).

A more realistic, but still much simplified model would be to synthesise the early sound from the direct and first-order boundary surface reflections and a few internal objects and to model the reverberation as a separate process based on the global room parameters.

6. SIMPLE VIRTUAL ROOM MODEL

For the creation of a simple room model, a number of assumptions had to be made. The first one was whether the model should be two-dimensional (2D) or three-dimensional (3D). It is clear that the practical limitations of broadcast production and home reproduction systems will, for some time, constrain the listener's experience to a 2D acoustic space. Budget constraints will mean that, for many years, the 'best' audio reproduction technology that can be anticipated will be the existing 5.0/5.1 multichannel system [1] Even that may take some time to become widespread.

The second main issue was the likely listening environment. Though systems for reproducing virtual sound spaces on headphones (for example, using manipulation of HRTF responses) are available, it is not likely that they will form the principal means of reproduction for the majority of the audience. Such systems also have difficulties with the variability of individual responses. For this work, a more general system, based on a multiplicity of spaced loudspeakers around the listening area and conventional 'amplitude panning', was assumed - even though it is known that simple panning is significantly defective for image presentations to the sides and rear of the listener.

This paper describes the development of a simple room acoustic synthesis system, based on a horizontal, 2-D arrangement of loudspeakers, corresponding to the 5.0 layout of ITU Rec. 775 [1].

The system was also limited to modelling a rectangular room shape, in order to ease the calculations of acoustic response. It will become clear later in this paper that *a priori* concerns about the need for fast recalculation of the acoustic parameters were shown to have been justified, thus making more complex models impractical at the present time anyway.

The development target for the system was limited to a somewhat arbitrary state. Many parameters were not optimised and many additional features could have been incorporated. It was necessary to stop at some point in order to carry out evaluations, which could only be done using a fully functional system, with the accompanying pictures.

7. THE ACOUSTIC ROOM MODEL

The acoustic model was based on current understandings of the behaviour of the human hearing system. The three psycho-acoustic parameters described in the previous section were mapped onto three different aspects of physical room acoustics.

The hearing process can be summarised as being sensitive to three distinct features -

(a) the direct sound (or first arrival),
(b) a number of discrete, delayed arrivals as a result of reflections from nearby surfaces
(c) the overall, diffuse reverberation.

7.1 Direct and 'early' sound

The geometric modelling of the direct sound was essentially trivial and self-evident. The time delay in the direct sound path was modelled (even though it will only rarely it be necessary to model the overall delay in real productions). However, it was necessary in order to implement Doppler shift.

In an enclosure, the boundary surfaces (walls, floor and ceiling), together with any large objects inside the enclosure, will cause a number of discrete reflections of the sound. For the purposes of the experimental system, the space was assumed to be a right rectangular prism, which resulted in six, first-order early reflections and made the calculation of the discrete image locations trivial. No second-order early reflections were included.

The 3-D model was modified to fit the 2-D reproduction paradigm by mapping the reflections from the walls, floor and ceiling onto the horizontal plane at the correct angle and distance relative to the listener. All calculations were carried out in the 3-D space and the mapping only applied to the final result.

All of the seven discrete sound source images were mapped onto the reproduction loudspeaker layout by amplitude panning, using sine-function panning with constant sound power [2]. The same panning law was applied to images to the sides and rear of the listener, where the large angular loudspeaker spacing and the poor acuity of the human hearing mechanism meant that the results

were questionable in principle and worked very poorly in practice (though the listener is not actually constrained to face forwards). For frontal sources, only reflected sound (and reverberation) will come from the sides and rear loudspeakers.

7.2 Reverberant sound

Any complete enclosure will form a reverberant space. The statistical properties of that space can easily be calculated from basic acoustic principles. The reverberation was assumed to be controlled by boundary surface absorption, uniformly distributed on all surfaces, together with air absorption. The required reverberation time was an input control parameter.

The only significant, general feature of reverberation is that it is (theoretically) constant in level throughout a room. Thus, the reverberation amplitude and statistical time-domain properties need only to be calculated once for each new room.

8. REFINEMENTS AND FILTER RESPONSES

8.1 Filters

The signals had to be filtered to model a number of acoustic propagation effects. In the simple model, all of the filters were implemented as a combination of a wide-band attenuation and a first-order, IIR low-pass with a single sample of delay, Fig. 1. The filter effects modelled included :-

Short-distance air absorption
Wall reflection.
Source directivity [3].
Reverberation characteristics. (source spectrum and long-term air absorption).

8.2 Internal objects

The basic acoustic model represented only the interior surfaces of the empty room. The addition of a multiplicity of objects within the room would have been relatively simple in principle, but the complexity of the model would very rapidly have become unmanageable. The adverse effects on the calculation speed would also have been substantial. However, it was clearly necessary to model at least one internal object. Such an object could either obstruct the direct sound or add an additional reflection to the six from the boundary surfaces.

The obstruction model also included approximations to the frequency and amplitude domain effects of diffraction around the edges of the object. That was necessary because of the severe high-frequency attenuation which occurs in the shadow zone. When an obstruction was detected, the direct sound was switched off and replaced by two new sources in the directions of the edges of the object and at distances corresponding to the total indirect path lengths.

8.3 Movement

As originally implemented, the simple model produced results for static geometries which were, objectively, of high audio quality. However, when the geometry was changed dynamically, clicks could be heard with speech programme and more severe distortion with music programme (though music is not a very likely signal type as a discrete source in VP).

The reason for the distortion was the quantisation of the time delays in the model by the audio sampling intervals (≈ 21 μs). That created large and clearly audible phase discontinuities. For example, at a modest source velocity of 1 m/s, and an system update speed of, say, 100 Hz, a 4kHz source would suffer a step phase discontinuity of about 40°. It was therefore necessary to quantise the distances/time delays to a much finer resolution by interpolation and to implement a crossfade mechanism to manage the change from one discrete set of parameters to another.

9. IMPLEMENTATION

9.1 Hardware

The experimental hardware consisted of a Lake DSP Huron™ development system. It had four, 40 MHz Motorola® 56002 processors on a single card. It also had an I/O card with input ADCs and output DACs. The dsp calculations were carried out using 24-bit fixed point processing. The audio I/O was 16-bit and system audio sample rate was 48 kHz, which gave (theoretically) 416 programme steps per sample period. The host computer into which the Huron system was installed was an industry-standard IBM PC-type computer. It used a Pentium 160 as the main processor, had 16 Mbyte of RAM and the system ran under Windows 3.11®. The dsp software was written in assembly language and assembled using the Motorola 56000 assembler. The control system was written in C/C++ and was compiled using Microsoft Visual C®.

9.2 System Overview

The signal processing requirements far exceeded the capabilities of a single dsp chip. Therefore, the system had to be partitioned to divide the tasks. Fig. 2 shows the partition. The three modules that resulted were the 'panner', the 'early' processor and the 'reverberation' generator. The audio input signal was applied to both early and reverberation modules. The latter also had a second input for effects or ambience.

The early processor generated the direct sound, six room surface reflection signals and two internal object signals. It provided appropriate time delay and filtering for all of those outputs, but no amplitude control.

The reverberation processor provided four mutually incoherent reverberation outputs, with appropriate filtering and no amplitude control. Though five reverberation signals might have been better, it was thought that four would be adequate and, in any case, the dsp processor did not have the capacity to process five outputs. It meant that one of the loudspeaker signals (the centre front one) did not have any reverberation component. The centre front loudspeaker of the 3/2 layout is also somewhat unsymmetrically placed compared with the other four, which do make a roughly symmetrical, square arrangement around the listener. In practice, the absence of reverberation in the centre loudspeaker was barely detectable, even quite close to it.

The 'panner' module summed all of these 13 inputs, in appropriate proportions, to create the five loudspeaker drive signals. In all, 91 control signals were required from the host to the dsp systems for all of the parameters.

10. CONTROL

The ultimate objective for the acoustic room simulator was for control by a video Virtual Production 'engine'. There had, therefore, to be some system of remote control.

A standard TCP/IP network interface was implemented. It was compatible with the sort of workstations used for the video system and was intended to give the maximum flexibility.

TCP/IP messages typically take the form of relatively small blocks of information. The control message structure was therefore based on short packets, actually C/C++ structures. The control messages also included a time parameter to permit actions or changes to be scheduled by clock time rather than as they arrived over the network.

For the purposes of development and demonstration, a Windows™ control interface was developed. Though complete control flexibility would clearly not be necessary in a final product, some aspects of the system will always require manual control. Even if all of the geometrical data were to be derived from the video model, setting the reverberation time would always be a strictly acoustic input

11. PERFORMANCE

11.1 Overall quality

The static noise and distortion (THDN) performance of the complete model was around -78 dB relative to zero level (22 Hz - 22 kHz unweighted). Most of the noise and distortion arose from the fixed-point, 24-bit quantisation limit but was judged to be adequate (at least for an experimental system). The actual noise floor of the base dsp system, measured using a simple mixer example supplied by the manufacturer, was not measurably different (±0.5dB). That showed that the large amount of signal processing (at 24-bit resolution) did not significantly affect the 16-bit output.

Subjectively, the output sounded too reverberant. The relative output level of the reverberation was confirmed by measurement to be objectively correct, so that the effect had to be a psycho-acoustic one. It is well known that the sound quality from a microphone located at a hypothetical viewer's location does not sound 'right' - it picks up to much reverberance and is quite likely to be lacking in spatial information. There is also a significant difference between actually listening in a real space and listening to a simulation of the same thing – even if that simulation is simply the sound picked up by a microphone at the same place.

In the final system, the question would be reduced to the subjectively optimum ratio of 'clean' to 'reverberant' sound. It would be an entirely trivial matter to change the model to adjust that ratio, but it would need trial productions and subjective tests to establish the correct value. For proper subjective assessment, that must be done with matching pictures and sound together.

11.2 Update speed

The final, measured update speed of the complete system was about 1.9 ms for a complete recalculation of the early response without an obstructing obstacle corresponding to an update rate of about 520 Hz. With an obstruction or reflection from the internal object, the rate fell to about 400 Hz. The processing of a new set of reverberation parameters took about 0.8 ms, but only needs to be carried out for new rooms. The amplitude crossfade rate was about 670 Hz. The time-delay crossfading was carried in the dsp processor, at the system clock rate of 48 kHz. That had the beneficial side-effect of also implementing nearly-perfect doppler shift for sources moving at reasonably high velocities.

The crossfade rate was critical because it had a profound effect on the representation of moving objects. The selection of crossfade rate depended on the conflicting requirements of movement speed, source type and desired audio quality. For an equivalent source speed of 5 m/s, the distortion was completely inaudible on all speech programme and many types of music programme.

For a critical test tone, the distortion was just about audible (amongst the other, realistic, effects of source movement).

13. SUMMARY AND CONCLUSIONS

This paper has attempted to summarise the implementation of a simple audio processor for creating virtual sound fields.

The experimental system incorporated models of the three main aspects of the acoustic sound field – the direct sound, the first-order surface reflections from the room boundary and the uniform, diffuse late sound which decays slowly to form the reverberation. The model was based on a simple, rectangular room in order to simplify the geometrical calculations and was designed for a standard, five-channel sound reproduction system.

Many acoustic refinements were included to model aspects of sound propagation such as air absorption and absorption at room boundaries. Also included were variable reverberation characteristics and a model of the directionality of the human voice.

Although the room model was otherwise empty, the facility to have one internal object was included, in order to model obstruction and reflection by large sections of the studio set. That had, perforce, to include a model of the diffraction of sound around an obstacle.

In order to permit the model to be controlled by an external process, the system included a TCP/IP network interface, a simple command 'language' and the ability to receive and interpret messages. For the purposes of experimental control and demonstration, the system also included a comprehensive manual control system, using MS Windows™ dialogue boxes.

The system appeared to perform adequately. Objective tests showed that it was behaving as calculated and that the measured noise and distortion performances were essentially indistinguishable from those of the basic dsp system, at least for a static geometry. The sound quality during geometry changes was lower, but still adequate for human speech sources and most types of music at equivalent source velocities up to about 5 m per second.

Subjective tests, using experimental programmes, will have to be carried out to assess the operating parameters and to optimise the control functions.

14. REFERENCES

1. ITU-R Recommendation BS.775-1, Multichannel stereophonic sound system with and without accompanying picture. (Geneva, 1992-1994).

2. Blumlien, A.D. British Patent Specification 394,325 (Directional effects in sound systems), JAES, 6, pp 91-98, 1958

3. Dunn, H.K. and Farnsworth, D.W., Exploration of pressure field around the human head during speech. *J.Acoust. Soc. Am.*, **10**, 184-199, 1939.

15. ACKNOWLEGEMENTS

This paper is published by permission of the British Broadcasting Corporation.

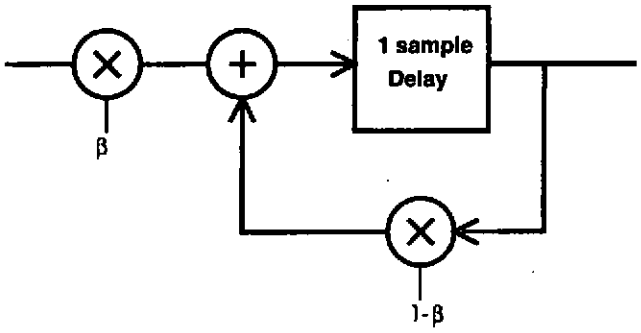

Fig. 1. Basic filter arrangement

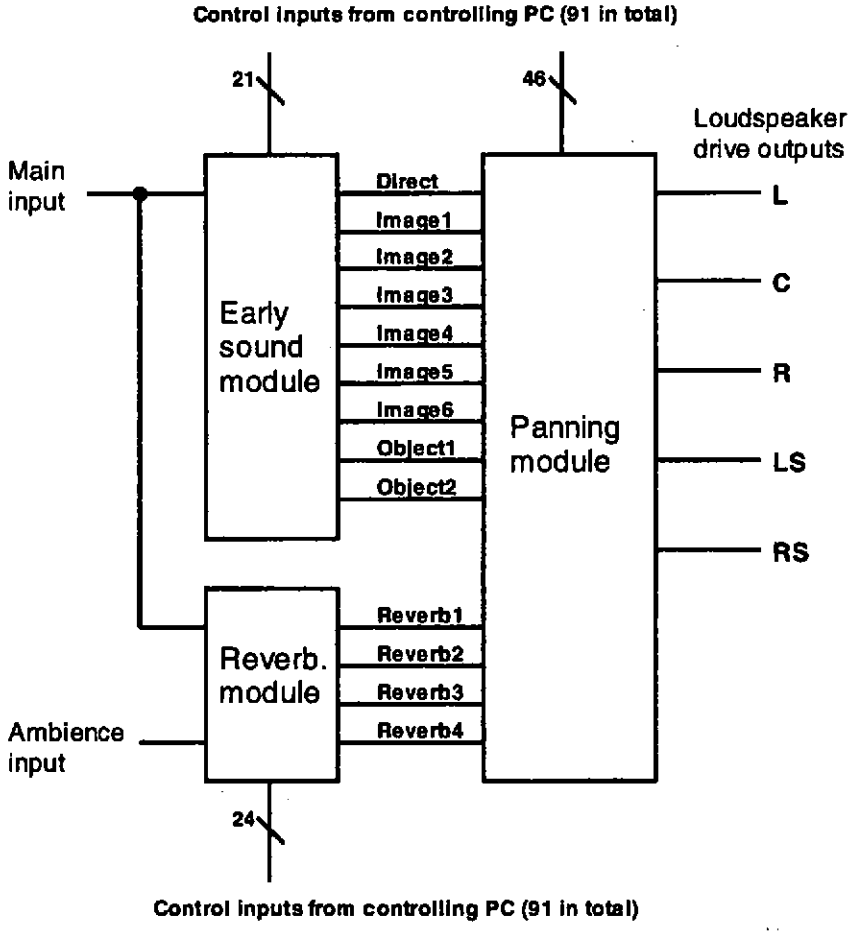

Fig. 2. Final dsp partition.