Virtual Reality Presentation of Loudspeaker Stereo Recordings

by Ben Supper

21 March 2000

ACKNOWLEDGEMENTS

Thanks to:

Francis Rumsey, for obtaining a head tracker specifically for this Technical Project;

Tim Brookes for assuring me that I could cope with it;

Ben Beeson and Richard Wheatley for their continual encouragement, and also for their feedback regarding the quality of the simulation as it took shape, without which it would probably not have sounded quite so convincing;

Steven Singer from the comp.sys.acorn.programmer newsgroup for his moral support when I encountered a particularly obdurate bug;

The Tonmeisters who volunteered for my listening test.

CONTENTS

ACKNOWLEDGEMENTS **i**

CONTENTS **ii**

ABSTRACT **iv**

0 INTRODUCTION **1**

0.1 A CONCISE HISTORY OF STEREO-TO-BINAURAL CONVERSION 2

0.1.1 OETF-BASED SYSTEMS 2

0.1.2 HEAD-TRACKED SYSTEMS 2

0.1.3 TECHNIQUES WHICH DISREGARD INDIVIDUAL AND DYNAMIC CUES 3

0.2 PROJECT AIM 4

1 FACTORS DETERMINING THE PERCEPTION OF SOUND POSITION **5**

1.1 VISUAL STIMULI 5

1.2 LATERAL LOCALISATION 6

1.3 FRONT / BACK DISCRIMINATION OF SOUND SOURCES AND LOCALISATION OF ELEVATED CUES 8

1.3.1 SPECTRAL DIFFERENCES 8

1.3.2 DYNAMIC CUES 8

1.4 APPARENT DISTANCE 11

2 IMPLEMENTING PSYCHOACOUSTIC CUES IN A COMPUTER PROGRAM **12**

2.1 INCLUDING LOCALISATION CUES 12

2.1.1 OBTAINING A USABLE HRTF DATABASE 13

2.1.1.1 EQUALISATION OF INCOMING IMPULSE RESPONSES 15

2.1.1.2 MODIFICATION OF INCOMING RESPONSES TO MINIMUM PHASE 17

2.1.1.3 RE-INTRODUCTION OF TIME DELAY 20

2.1.1.4 INTERPOLATION METHOD 22

2.1.1.5 REDUCTION OF IMPULSE RESPONSE LENGTH 22

2.1.1.6 EXTENT OF THE PROCESSED DATABASE 25

2.2 DISTANCE CUES 27
2.2.1 DISTANCE PERCEPTION — THE CRAVEN HYPOTHESIS 28
2.2.2 IMPLEMENTATION OF EARLY REFLECTIONS 29
3 AURALISE: A LISTENING ROOM SIMULATOR 33
3.1 HANDLING AUDIO FILES 34
3.2 REAL-TIME DSP 35
3.2.1 CONVOLUTION 35
3.2.2 GENERATION OF REFLECTION DATA 38
3.2.3 COMMUTATION OF HEAD-RELATED IMPULSE RESPONSES 40
3.3 HEAD TRACKING 40
3.3.1 COMMENTS ON THE CHOICE OF HEAD TRACKER 40
3.3.2 THE CYBERTRACK-II DRIVER 41
3.3.3 PROCESSING THE HEAD TRACKER DATA 42
4 EVALUATION OF THE SYSTEM 43
4.1 SUBJECTIVE EVALUATION 43
4.2 LOCALISATION 46
4.3 APPARENT DISTANCE PERCEPTION 48
5 CONCLUSION 51
5.1 VIABILITY AS A CONSUMER PRODUCT 52
5.2 VIABILITY AS A PROFESSIONAL PRODUCT 53
5.3 VIABILITY AS A RESEARCH TOOL 54
6 GLOSSARY 55
7 REFERENCES 57
8 BIBLIOGRAPHY 61
A MATHEMATICAL DERIVATION OF POINT ROTATION 62
A.1 ROTATION BY YAW (ROTATION) 63
A.2 ROTATION BY ROLL (PIVOTING) 64
A.3 ROTATION BY PITCH (TILTING) 65
A.4 POINT ROTATION 67
B EXTRACTS USED IN THE LISTENING TESTS 68
C THE LISTENING TEST PAPER 70

ABSTRACT

Virtual reality loudspeaker simulation technology aimed at the recording engineer is a developing field of audio product design. There are many issues behind the implementation of such a system: these are covered in detail, and a software simulation is introduced to illustrate them.

Two separate design stages are discussed. The creation of an HRTF database from an extant set of impulse responses is vital to the successful processing of audio through the system; the nature of this processing is also important. Two of the main problems with existing binaural systems are eliminated: front/back confusion is avoided by tracking the listener's head movements, and in-head localisation is prevented by incorporating early reflections from a simple listening room model into the simulation.

The commercial viability of this loudspeaker auralisation system is discussed: it would almost certainly be necessary to improve the simulation by using a faster processor, but a product incorporating this technology would be particularly useful for the film sound and mobile recording sectors of the audio industry.

0 INTRODUCTION

The vast majority of stereo recordings are made with the intention of being replayed on loudspeakers. When they are monitored using headphones, the stereo image will appear to be inside the head, with sound sources tending to cluster around each ear. This can be attributed to the unique experience in headphone listening of hearing each stereo channel at the corresponding ear with very little interaction between the two; a phenomenon which never occurs naturally over the full frequency spectrum.

Early attempts at compensating for this difference between loudspeakers and headphones included the production of binaural recordings, using one of a number of specialist recording techniques. These recordings are intended to be reproduced only via headphones. Binaural recordings reproduce at each ear the pressures incident on the microphones of a head-shaped or spherical stereo microphone placed within the recording environment. While a recording made with a high-quality dummy head simulates extremely realistically all of the directional and spatial cues of an inert listener, the technique disregards any additional cues via head movement. The profound influence of these cues on sound source localisation was proven as early as 1939 [Wallach].

Loudspeaker stereo sound can also be processed electronically to simulate the phenomenon of listening to the left and right stereo signals via loudspeakers in a listening environment, thereby making the original programme material headphone compatible. In most circumstances, however, this will again simulate an idealised inert listener. Exactly the same problems observed when employing binaural techniques will therefore exist for such a system.

The ability to furnish additional cues by changing the nature of the headphone sound as the listener moves their head is a relatively recent development, as the real-time digital signal processing required to simulate these cues with sufficient accuracy has only been feasible for a few years. However, this technique can provide an extremely realistic impression of the stereo signal as it would sound if it were replayed through loudspeakers in a listening room.

Owing to the relatively high price of head tracking hardware, many recent attempts at creating loudspeaker auralisation systems have chosen to disregard dynamic cues,

comprising a static stereo-to-binaural conversion processor.

0.1 A CONCISE HISTORY OF STEREO-TO-BINAURAL CONVERSION

The concept of modifying loudspeaker stereo signals for reproduction through headphones is not new. Bauer [1961] innovatively discussed methods of converting programme material from stereo to binaural format, and vice versa. In 1977, Martin Thomas published a working circuit which combined delayed crosstalk with empirically developed filters in an attempt simulate loudspeaker listening through headphones. Thomas's own evaluation showed that every individual in a sample of listeners preferred listening through this filter structure to hearing the unprocessed audio through headphones [Thomas 1977: 477].

The most recent attempts at producing a realistic impression of a loudspeaker stereo image via headphones have invariably employed digital processing techniques. Each of these attempts takes one of three approaches. These will be analysed individually.

0.1.1 OETF-BASED SYSTEMS

Some systems use a database of 'Own-Ear Transfer Functions', aiming to achieve more accurate localisation by obtaining transfer functions from the ears of the individual who will be using the system [Persterer 1991].

A system which relies on own-ear measurements is cumbersome to implement, particularly if a large number of individuals will be using the same equipment: gathering OETFs is an onerous and time-consuming process, using expensive specialist equipment. For this reason, it is best avoided wherever possible. The biggest advantage of an OETF-based system over one which uses non-individual HRTFs is that confusion between sounds to the front and sounds to the rear of the listener is significantly reduced [Persterer 1991; Richter 1992; Møller et al 1999]

0.1.2 HEAD-TRACKED SYSTEMS

A second category of systems employ a digital head tracker, and process the signal in real-time so that the listener is immersed within a virtual listening room: the positions of the loudspeakers relative to the listener are re-calculated whenever the listener's head is moved.

Head-tracked auralisation, made possible by technological advances in Virtual Reality, is becoming an increasingly popular approach. This technology is also becoming more and more affordable to implement, and many examples of head-tracked loudspeaker auralisation systems are either in development or have been released as products, most notably by Sony, Lake DSP and Stüder [Goodyer 1997; Inanaga et al 1995; McKeag and McGrath 1997b; Horbach et al 1999]. Head-tracked audio also has the advantage that filter databases do not need to be changed either to suit different listeners, or when the listener changes their brand of headphones, and therefore the frequency response of the system. Dynamic cues work irrespective of the individual peculiarities of listeners' pinnae, and localisation performance is reported to be superior to the OETF-type system, especially with regard to front-back discrimination [Jot 1995: 4; Horbach et al 1999: 6].

A disadvantage of most current head-tracked systems, and particularly the one described by Horbach et al, is the amount of real-time processing involved. This makes the system expensive because it requires an enormous database of head-related transfer functions and a number of dedicated digital signal processors to perform the necessary audio filtering.

0.1.3 TECHNIQUES WHICH DISREGARD INDIVIDUAL AND DYNAMIC CUES

In the majority of systems, neither of the techniques above are applied [Begault 1991; Rubak 1991; Robinson and Greenfield 1998; Dolby Laboratories 1999]. With the exception of Dolby, where all of the available literature is intended for marketing, and so does not extend to the shortcomings of their product, designers of this last type of system state problems regarding confusion between sources in front of the listener and those coming from behind. As will be seen in §1, there are a number of psychological reasons for this phenomenon, and a number of effective ways of removing most of them from a system. It has even been suggested [McKeag and McGrath, 1997b] that the addition of a binaural room impulse response to the simulation will reduce front/back confusion. However, this is an isolated statement which is offered no psychoacoustic explanation, and has not yet been proven experimentally.

0.2 PROJECT AIM

This project is based on the observation that it must be both possible and worthwhile to design a professional-quality head-tracked loudspeaker auralisation system which is suitable for two-channel loudspeaker stereo reproduction at least, using a cheap microprocessor with a specification which is not incredibly high, and limited memory resources.

As it is would not be possible to simulate every audible aspect of a real environment under these conditions, it is necessary throughout this project to assess the relative salience of the known ear-brain cues used in sound localisation by drawing upon available literature, and then to use this knowledge as a basis for generating and processing audio data appropriately. A real-time loudspeaker auralisation system is implemented on an Acorn Risc PC personal computer with a 233MHz StrongARM processor.

1 FACTORS DETERMINING THE PERCEPTION OF SOUND POSITION

Before designing any practical system which attempts to convince a listener that they are immersed within a virtual acoustic environment, it is vital to understand the decisions made by the ear-brain mechanism when it attempts to locate a sound source, and the stimuli upon which these decisions rely. This is particularly important in the present case, where the limited availability of computing resources means that decisions have to be made about which of these cues need to be implemented, which may be ignored, and which ones are implicitly built into or left out of the system.

Wightman and Kistler [1997: 2] divide localisation cues into two categories: monaural and binaural. Monaural cues are perceived at each ear individually; binaural cues work by assessing the differences between the signals at each ear. The descriptions below make no such distinction, as monaural phenomena are reinforced when cues from one ear are considered in the light of monaural cues from the other ear. The fact that they may be detected with only one ear is of no relevance when developing a binaural system.

The methods by which a listener may locate a sound source may be divided into five categories. These are covered in order of decreasing significance.

1.1 VISUAL STIMULI

The brain's method of locating sounds by connecting them with visible objects is far more reliable and less ambiguous than its methods of locating sounds solely by hearing them. If visual and aural stimuli conflict, the brain will always favour the visual stimulus. For this reason, visual stimuli are often regarded as the most important localisation cues [Blauert 1989: 193–196].

When there are no visual stimuli, the brain will have to rely purely on aural cues. Whilst this is satisfactory when listening to recorded music through loudspeakers, the inability to see the source of a sound when listening through headphones often causes confusion. The ear/brain mechanism tends to locate an auditory event occurring in front of the listener to the rear of the listener when there is no visual stimulus. In nature, this is where a sound will naturally be placed by the brain when there is nothing within the visible field which can generate it, and the listener's head is not free to move. The opposite reversal in

binaural recordings, where sounds recorded behind the dummy head appear to be in front of the listener, are far less common. [Begault 1991: 2; Wightman and Kistler 1997: 13; Robinson and Greenfield 1998: 7]

Another phenomenon is often reported [Horbach et al 1999: 8] whereby many subjects perceive stimuli as bring elevated artificially. Few subjects, however visualise them as coming from below their heads. This was also discovered by Wallach [1939: 273]

1.2 LATERAL LOCALISATION

A distinction must be made between *lateral localisation* and *lateralisation.* The difference between the two terms were introduced in a paper by Plenge [1974], in which lateralisation was demonstrated as the location of sound inside the head. Localisation is distinct from this, in that it implies that the sound is successfully located outside the head.

The brain is able to gauge accurately, particularly at frequencies from 1.5kHz to about 3kHz [Hartmann 1997: 197], the time difference between a signal reaching one ear and the same signal reaching the other ear. This provides a way of approximating the angle of incidence of the sound to the head.

This method creates ambiguities: the *cones of confusion.* These occur because a particular interaural time delay can correspond to one of many locations, which appear geometrically as any point on the surface of a cone extending from the centre of the listener's head, whose axis extends perpendicularly to the ears (*Fig. 1, page 7)*. The most obvious confusion, and the most problematic from the point of view of binaural technology, is that the brain cannot easily discriminate between sounds in front and sounds to the rear of the head as both sounds will have the same interaural time delay. The other problem occurs to a lesser extent in that some people perceive the sound sources to be elevated or dipped. In spite of this plurality of possible source locations, time delay is particularly useful in obtaining directional information because the brain is able to measure and ascertain interaural time differences with considerable accuracy [Blauert 1987: 37].

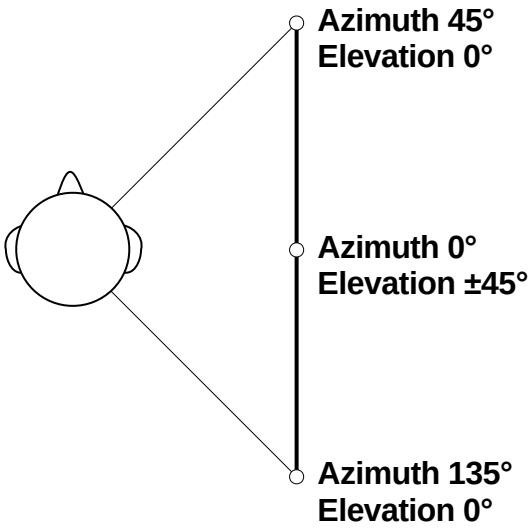


Fig. 1: the 45° cone of confusion, and ,four points on it. The interaural time difference cues arriving from any point on the cone's surface would be identical.

At frequencies greater than approximately 1.5kHz, the brain begins to utilise the head-shadowing effect in which high-frequency interaural level differences play a role in indicating source direction. A sound incident on one side of the head will be perceived as being louder at higher frequencies because incident sound will be reflected from the head, raising the sound pressure immediately around that side of the head. At the other ear, there will be high-frequency attenuation, owing to the presence of the head as an acoustic barrier in the way of the incident sound. Listening tests [Wightman and Kistler 1997: 13] have shown that interaural intensity difference is a weaker cue than interaural time difference: if the two are set in conflict, the brain will always favour time delay.

The exception to this rule is when sound is extremely close to the head: in this case, there will be interaural level differences caused not only by head-shadowing, but also by the greater relative distance of the auditory event from the far ear. Because the sound pressure of an omnidirectional source decays 6dB for every doubling of distance in the free field, this effect can be quite considerable for sounds which occur close to the head, but has little significance for longer distances.

1.3 FRONT / BACK DISCRIMINATION OF SOUND SOURCES AND LOCALISATION OF ELEVATED CUES

1.3.1 SPECTRAL DIFFERENCES

To eliminate the cone of confusion when faced with an unseen auditory event, the brain relies on two methods. The most frequently-implemented cue relies on subtle spectral differences caused by the reflections and shadowing effects of the outer ear, and particularly the conch, which is considered to be of greatest importance for assessing the elevation of sound sources. Front-back discrimination is also possible, and relies on the horizontal asymmetry of the pinna.

There are three immediate problems which are caused by sole reliance on time delay and spectral phenomena. The first is that, without prior knowledge of the nature of the auditory event, the brain cannot discriminate between a sound which is filtered because it is elevated or appears behind the head, or whether the signal's frequency spectrum normally takes the shape of such a cue [Wightman and Kistler 1997: 13]. The second problem is that spectral cues are extremely subtle, and they can be upset by early reflections inside rooms [Hartmann 1997: 200–202]. Lastly, the subtlety of these filtering effects means that they do not transfer well from one listener to another. For example, a dummy-head recording, which relies on a physical model of an idealised listener, will work well only when a listener has very similar pinnae to the ones used for the recording.

1.3.2 DYNAMIC CUES

A far more reliable method which the brain uses to eliminate the ambiguity inherent in lateral localisation involves the extra cues gained during conscious or subconscious head movement. These remained largely uninvestigated in binaural systems until fairly recently, when fast processors became affordable enough to make implementation of these cues practical for binaural synthesis.

When a listener perceives an auditory event, they will almost always move their head, whether or not they are consciously aware of this movement [Thurlow et al 1967: 489]. Changing time and spectral differences between the ears provides a very reliable method of finding the elevation and the location of a sound.

The most stark contrast in dynamic cues occurs when discriminating between front and rear auditory events. This is illustrated in *Fig. 2.* With elevation increasing or decreasing from the frontal axis, interaural time difference becomes less and less pronounced. A subject may use this effect to determine the elevation of a sound as the head is rotated. It is also possible for a listener to decide whether an auditory event is occurring above or below themselves by rolling their head from side to side.

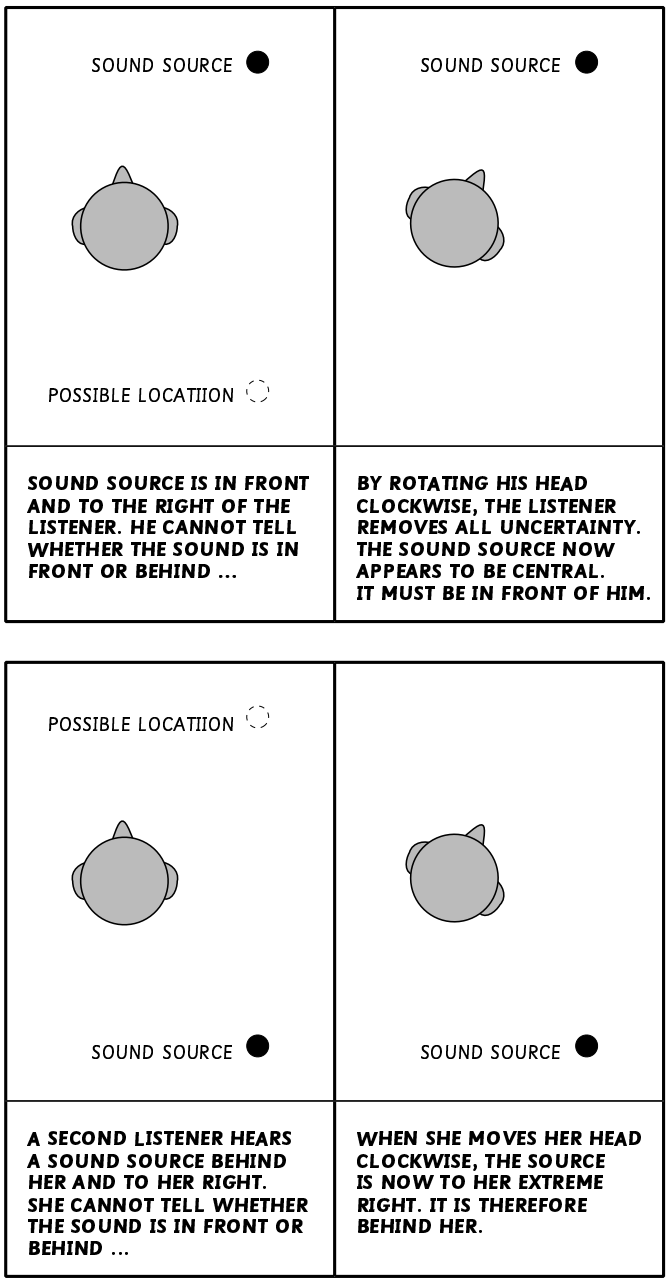


Fig. 2: Successful elimination of front-back confusion through the use of head movement.

The strength of dynamic cues was discovered by Hans Wallach in his experiments of 1939: he could successfully synthesise a stationary source in front, behind or above a listener by switching a signal between an arc of twenty loudspeakers in front of the listener using a rotary switch attached to the listener's head.

If the signal was switched so that the angular displacement of the signal with respect to the listener was twice the angular displacement of the listener's head, the sound appeared to be coming from a point behind them. If the angular displacement of the signal was switched to a loudspeaker at a value equal to or less than the angular displacement of the listener's head, it appeared to be elevated accordingly.

> *Synthetic production experiments in which the direction to be perceived is horizontal were always successful … This experiment was performed with a great number of observers, and never failed.* [Wallach 1939: 273–274]

Wallach notes the fact that his experiment produced overwhelmingly successful results in spite of the incorrect pinna cues: dynamic cues, therefore, play a more important role in the elimination of localisation ambiguity than spectral cues.

1.4 APPARENT DISTANCE

Determination of source distance from the listener relies on a number of approximate factors. A brief list [after Gerzon 1992] must include the following:

- Interaural level differences for small distances, as discussed in §1.2.
- The Craven hypothesis — that the brain is able to assess the distance of a sound source in an enclosed space purely on the relationship between time delay and amplitude of each of the early reflections. This is explained in more detail in §2.2.1.
- Air absorption, which produces a high-frequency roll-off which increases with source distance.
- The angular size of the source: a real sound source will appear to be wider when it is nearer the head than when it is further away.
- The reverberation time of an enclosed space, through which it is possible to achieve an indication of the size and quality of the environment, to place the sound within context.
- Apparent loudness: this is only really useful for familiar sounds including speech and acoustic musical instruments, when the typical level of such a signal is already known by the listener.

These cues are discussed within the context of loudspeaker auralisation in §2.2.

2 IMPLEMENTING PSYCHOACOUSTIC CUES IN A COMPUTER PROGRAM

2.1 INCLUDING LOCALISATION CUES

From the outset, it was decided to include dynamic cues in the project. Many recent experimenters [Horbach et al 1999; Savioja et al 1999; Robinson and Greenfield 1998:9] and writers [Travis 1996: 6; Jot 1995: 4; Blauert 1987; 43, 178–189] advocate the use of dynamic cues, firstly to enhance the sense of reality of the virtual environment, and secondly to help to eliminate localisation ambiguities in headphone simulations of real environments. It was decided that the added expense, processing requirements and development time required by their inclusion would be rewarded by the enhanced realism of the overall result.

It is immediately evident that a computer simulation would also have to provide the listener with interaural time delay and monaural spectral cues in order to sound convincing. This is the method which is employed by all existing binaural processors, whether they use static or dynamic processing.

Interaural time differences are achieved by delaying the signal from each virtual loudspeaker to each ear by a calculated amount. Spectral cues are included by digitally convolving each delayed signal with a position dependent head-related impulse response: in order to find these, it is necessary to model accurately the behaviour of a sound impulse as it travels through the air, and around the listener’s head and ear.

Fortunately, it is not necessary to model the complex diffraction, reflection and delayed paths of sound around a head in order to obtain impulse response data: such modelling would require a hugely detailed computer simulation. The easiest method of obtaining this physical data accurately is to measure the position-dependent impulse response of a real dummy head. Gardner and Martin [1994] have collected and processed a large set of data collected from a KEMAR head in an anechoic chamber, sampled at intervals of ten degrees on the median plane (from –40° to 90° elevation), and at a minimum of five degrees on the horizontal plane (the resolution is reduced away from 0° elevation), taken at a distance of 1.4m from the head. The whole data set comprises 710 impulse responses, sampled at 44.1kHz with 16 bits resolution. Each response is 512 samples (11.6ms) long,

and has been compensated for the frequency response of the loudspeaker used to produce the stimulus.

2.1.1 OBTAINING A USABLE HRTF DATABASE

While the KEMAR data set provides a freely available and convenient starting point to synthesise a set of filters, it is not possible to use it without altering it. Further processing is necessary for three reasons:

- The length of each impulse, at 11.6ms, is far too great to perform a convolution in real time. To do so would necessitate over twenty-two million multiply-accumulate instructions per second for a one-speaker, one-ear system. While a number of binaural processors are available which can handle arithmetic at this speed, they require specialist hardware which is prohibitively expensive and cumbersome.
- The database is too coarse. The resolution of human hearing at its finest is just over 3° [Blauert 1987: 40–41]; this occurs on the horizontal plane at the front of the head. A database with a 5° horizontal resolution will not be sufficient. Ideally, the resolution should be 1° at its finest, so that small angular changes will be unnoticable. Insufficient resolution of the database cannot then present a problem.
- Each impulse response in the database also contains the transfer function of a dummy ear canal. It is not desirable to play sound filtered through one ear canal into another because it will sound overly coloured; a method must be found of removing the canal's transfer function from each impulse response.

Considerable processing needs to be applied to the database of impulse responses before a set is produced that can be used for a head-tracked virtual reality system. A flowchart of the database processing, which is explained in more detail in this section, is shown in *Fig. 3, page 14*. This data manipulation is all performed prior to the simulator being run, and the resulting database is committed to disk, so that the time it has taken to assemble the database will not divert resources from the considerable signal processing which needs to be enacted on the data in real time.

For practical reasons, the database processing is divided between two programs: the first interpolates the original database in the horizontal plane, and the second uses the new data to interpolate in the median plane.

Median plane resolution of the input database is interpolated from 10° to 5°. This is necessary because the minimum localisation blur in the median plane is ±9° [Blauert 1987: 44]: the Gardner and Martin database is therefore slightly too coarse for head-tracked simulation.

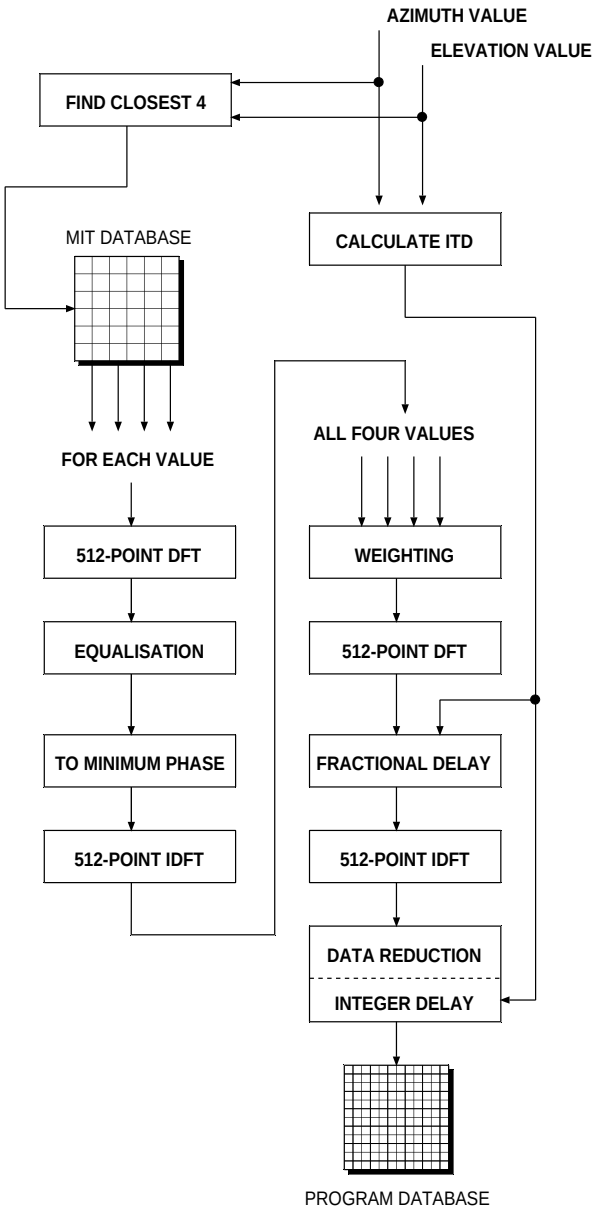


Fig. 3: Flowchart of data processing employed to achieve a usable impulse response database.

2.1.1.1 EQUALISATION OF INCOMING IMPULSE RESPONSES

Each impulse response must be equalised to compensate for the ear canal response of the KEMAR dummy head. This may be done in one of three ways; each one involves performing a discrete Fourier transform on the impulse response to obtain a frequency response, superimposing a particular filter pattern on this response, and performing an inverse discrete Fourier transform to arrive at an equalised impulse response. The three filter patterns which are most often used are:

- The inverted frequency response of the measured 0° elevation and 0° azimuth impulse. [Jot 1995: 7]
- The inverted average of every item of data in the database [Jot 1995: 8; Kistler and Wightman 1992: 2]. This converts a head-related impulse response (HRIR) into a *directional* impulse response (DIR). Rubak [1992] obtains a directional transfer function by equalising the head-related impulse response with the response of an omnidirectional microphone substituted for the dummy head.
- The inverted headphone-to-ear response for a particular brand of headphones on the dummy head.

Of these methods, it was decided that the average response is most suitable for the system. The headphone response is too dependent upon individual manufacturers and types (see a comparison in (*Fig. 4, page 16*) to provide an adequate general response; equalising the impulse responses with the transfer function in front of the head removes any direction-related artefacts from impulse responses taken at this angle, while the ideal procedure should colour the sound in front slightly, and the sound behind slightly: this is what the head-pinna mechanism does. It seems logical that equalising with the inverted average transfer function *(Fig. 5, page 16)* would produce the best overall result. It would also mean that the average response of the database would be flat. Because it is undesirable to tamper too much with the spectral qualities of the sound, this seems to be the best alternative.

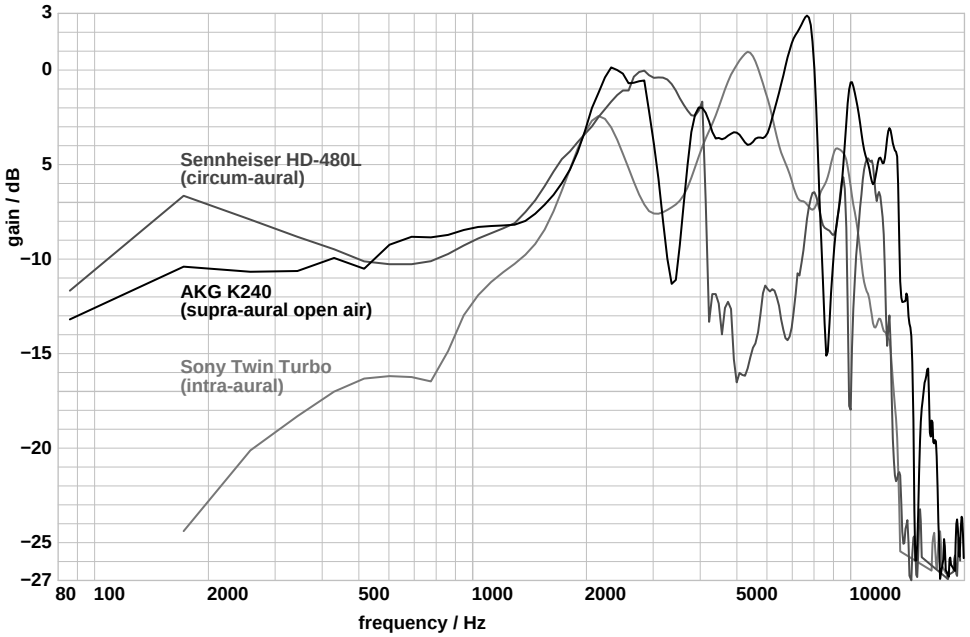


Fig. 4: A comparison of the headphone transfer functions supplied with the Gardner and Martin database.

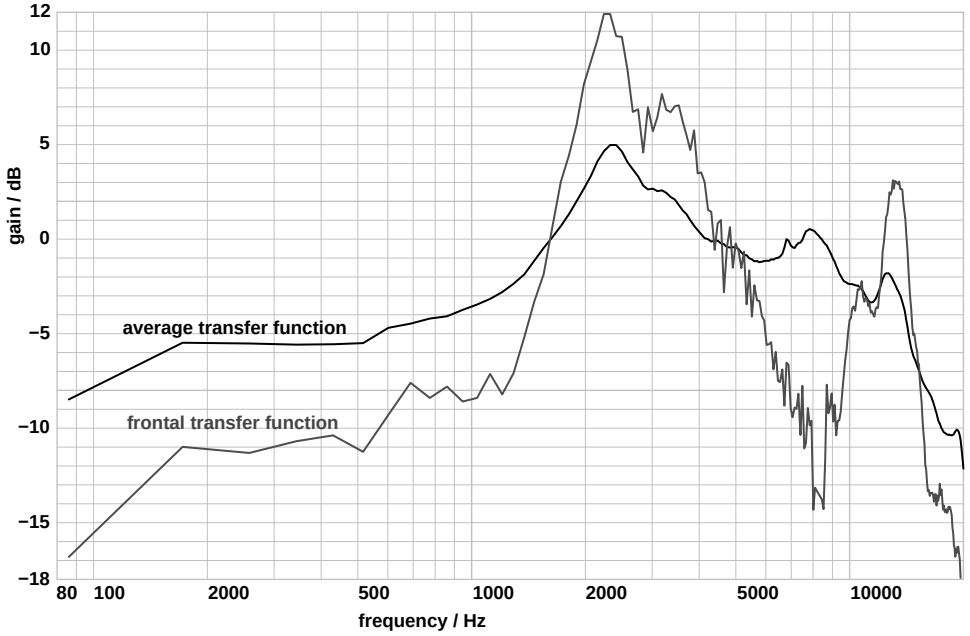


Fig. 5: Frontal (0° azimuth, 0° elevation) transfer function compared with the average transfer function of the data set.

2.1.1.2 MODIFICATION OF INCOMING RESPONSES TO MINIMUM PHASE

In order to combine a number of head-related impulse responses using a standard weighting algorithm, they must be coincident. If they do not all start at exactly the same time, the result achieved by mixing them in various proportions will not be one averaged impulse response, but a single attenuated response followed by three early echoes. This will disrupt the magnitude and phase relationships of the resulting signal. Before combination, each of the impulse responses is therefore reduced to minimum phase with no additional delay; a suitable delay may then be inserted after the responses are combined. Using minimum phase transfer functions does not affect the perceived quality of filtered audio [Kistler and Wightman 1992]. A convenient way of reducing an impulse response to minimum phase is by passing it through a discrete Fourier transform, and then setting the imaginary part of the frequency response to equal zero, and the real part of the frequency response to equal the old magnitude response. This represents a function with the same frequency response as the transformed impulse, but with no phase shift at any frequency. Passing this through the inverse discrete Fourier transform produces a phase-linear impulse response around the impulse response graph's origin, and is wrapped around by the transform so that, for a 512-point inverse transform, the –1st sample appears at the 511th position. The first half of the graph is a purely causal, minimum-phase impulse response. This processing is all demonstrated in *Fig 6, pages 18–19.* If every impulse response is treated in this way, the interpolation algorithm may successfully combine them simply by using weighted averaging.

It can also be seen in *Fig. 6* that converting an impulse response to minimum phase will create a new impulse response which contains levels significantly higher than those in the original sample. It would be disastrous if a number of interpolated impulse responses were clipped as they were stored in the database. To compensate for this, every impulse is attenuated by 12dB in the database pre-processor. This is taken into account by amplifying the audio within the simulator by 12dB after it has been convolved.

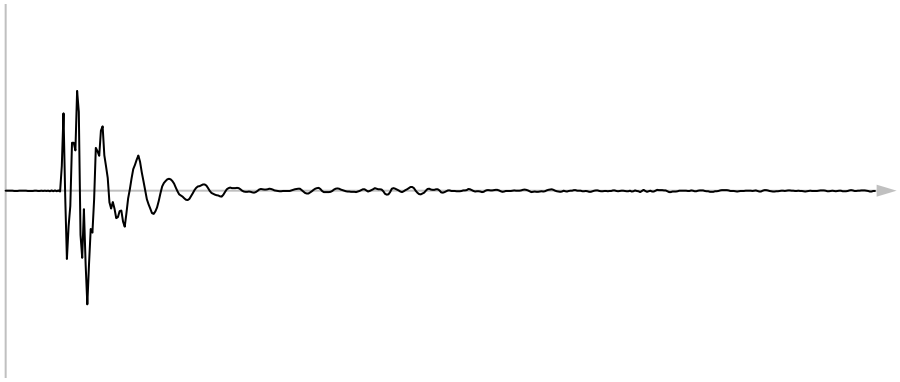


Fig. 6: Part a). An arbitrary impulse response read directly from the Gardner and Martin database (Note that it would actually be equalised before this processing was applied to it.)

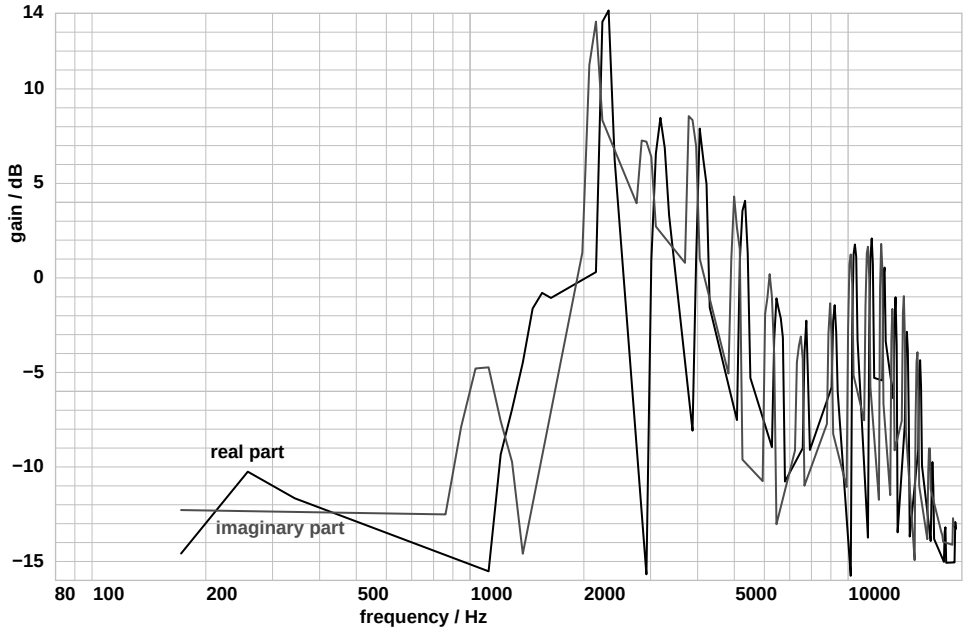


Fig. 6: Part b). The real and imaginary parts of this impulse in the frequency domain

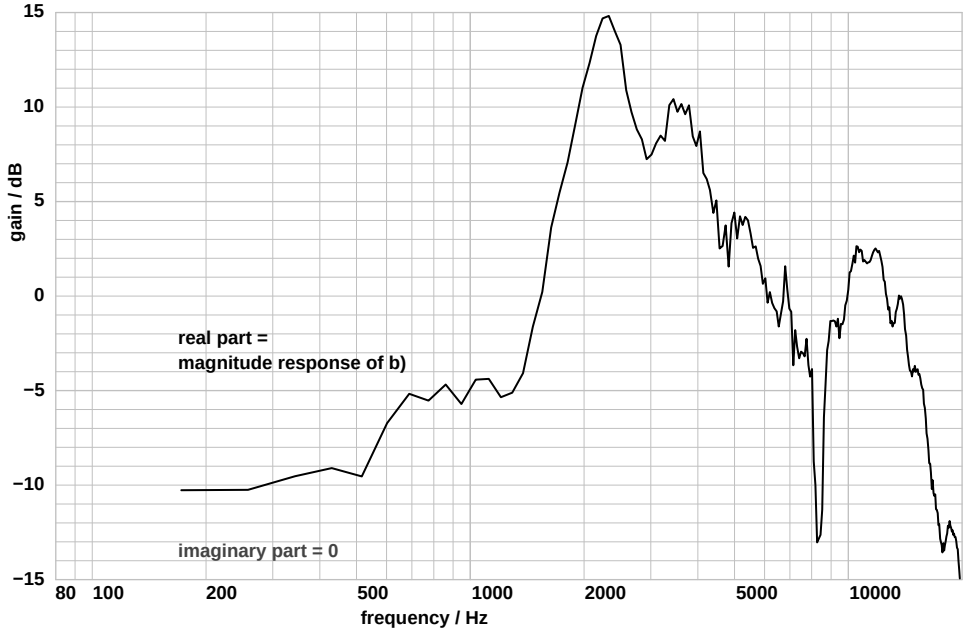


Fig. 6: Part c). The frequency response altered so that magnitude response is identical to b), but the phase shift is uniformly zero

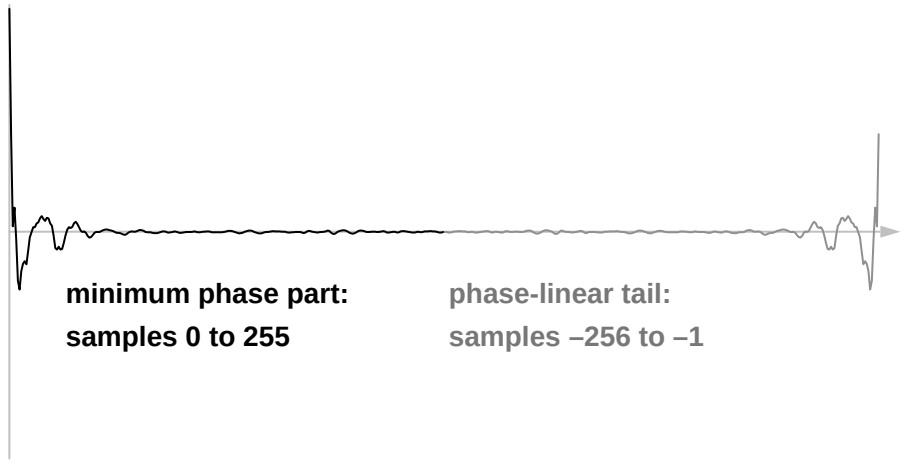


Fig. 6: Part d). The altered frequency response transformed back into the time domain

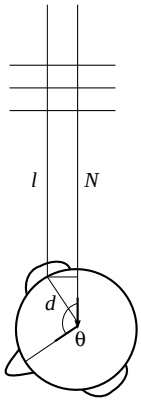
2.1.1.3 RE-INTRODUCTION OF TIME DELAY

The delay for each impulse response is calculated using a simple formula (based on [Savioja et al 1999: 690]), which is derived in *Fig. 7*. The parameter *N* is set to 25 samples: this proved to be a large enough value for the sample delay never to cross zero, whilst remaining small enough to keep the simulator compact in terms of memory requirements.

N = Nominal distance
l = length of signal path to ear
d = radius of head (typically 0.1m)
θ = azimuth of head relative to sound source in radians
ψ = angle of elevation

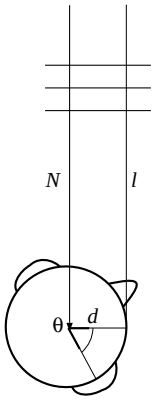
a) Sound approaches ear from near side

assuming a plane incident wave,
$l = N - d\cos(\theta - 90)$
$= N - d\sin\theta$



b) Sound approaches ear from far side

assuming a plane incident wave taking the shortest possible path,
$l = N + d\theta$



adding a simple cosine elevation dependency,
sample delay = l × [sampling frequency] × cos Ψ / [speed of sound]

Fig. 7: Derivation of relative time delay (in samples) against azimuth and elevation angle

It has already been mentioned that the accuracy with which a subject may localise sound is 3° at the finest. Using the formula above, this translates as an interaural delay of 15μs, or approximately 0.7 samples at 44.1kHz. Assuming, therefore, that the ear is able to detect such small time differences, it is clear that it is not satisfactory simply to round the delay to an integer number of samples: the unit delay must somehow be subdivided. This was proved by a non-working early attempt to interpolate the database using only multiples of the unit delay.

This is achieved by venturing again into the frequency domain using a discrete Fourier transform. A delay can be introduced into the transformed data by manipulating the phase response of each frequency component using a formula derived from first principles:

$\varphi = 2\pi f T$ radian

where φ = phase shift;
f = frequency in Hertz;
T = constant time delay.

A fixed delay in the time domain can therefore be seen in the frequency domain as a phase shift which is directly proportional to frequency. This may be translated empirically into digital signal theory. When the Nyquist frequency component (at $f_s / 2$) is shifted in phase by π radian and the other frequency domain values are scaled linearly around this, with zero phase shift at zero frequency, the delay will be exactly one sample.

The phase of a particular component in a 512-point transform delayed by a fractional part δ of the unit delay, is therefore:

$\varphi = \pi \delta f / 256$

Using this law to adjust the phase of the weighted and combined data before converting it back using an inverse discrete Fourier transform causes the phase-linear impulse response to be delayed by the appropriate fraction of a sample. This can then be used, as before, from the origin to the halfway-point, as a minimum-phase impulse response.

This procedure may be enacted quite simply on a minimum phase transfer function:

$M = \Re(f)$ because $\Im(f) = 0$

The new values then become:

$\Re(f) = M \cos \varphi$

$\Im(f) = M \sin \varphi$

2.1.1.4 INTERPOLATION METHOD

Because the database pre-processing algorithms are completed before the simulator is assembled, time constraints are not a significant issue. The time that the interpolation algorithm takes is therefore unimportant, so it is beneficial to select an interpolation algorithm which favours quality of output over time of execution. A number of suitable algorithms are demonstrated in Hartung et al [1999]. The procedure of interpolation by inverse distance weighting was used, whereby the four nearest impulse responses are combined, weighted according to the reciprocal of their great circle distance from the output point. This algorithm takes considerable time to compute a single output, as a large number of floating-point operations are required in order to produce each output response. The pre-processor, programmed in a mixture of BBC BASIC V and machine code and running on a 233MHz StrongARM processor, compiles the simulation database in approximately an hour and a half.

2.1.1.5 REDUCTION OF IMPULSE RESPONSE LENGTH

Now that the interpolated impulse response has been obtained, it is necessary to truncate it to a usable length. It has already been stated that an 11.6ms impulse response is too long to be practical: this is the main reason to reduce its length. It is also desirable to shorten the impulse responses so that they occupy less memory.

The first way to reduce the impulse response data is to remove its leading pause. Conveying an interaural delay by setting a certain number of leading samples in the impulse response to zero will work, but this is a wasteful use of storage and processing resources. It is far more efficient to store the unit delay as a single number, and then to store with this the undelayed impulse response. When the response is convolved with the audio, the program may re-introduce this delay by referencing the audio data a number of samples further back: it does not waste processing time by having to multiply a large number of samples by zero to achieve the same effect.

The next way to reduce the amount of data required is by cutting off the impulse response at a certain length. Huopaniemi and Zacharov [1999] successfully truncated head-related impulse responses to 48 coefficients each. Huopaniemi and Zacharov [1999: 222] suggest that there is no disadvantage in cropping the impulse response using a rectangular window, as a head-related transfer function contains no sharp notches and no discontinuities. The effect of progressively harsh rectangular truncation upon the frequency response of the resulting filter can be seen in *Fig. 8, page 24.* In my database pre-processing program, impulse responses are truncated to 48 samples.

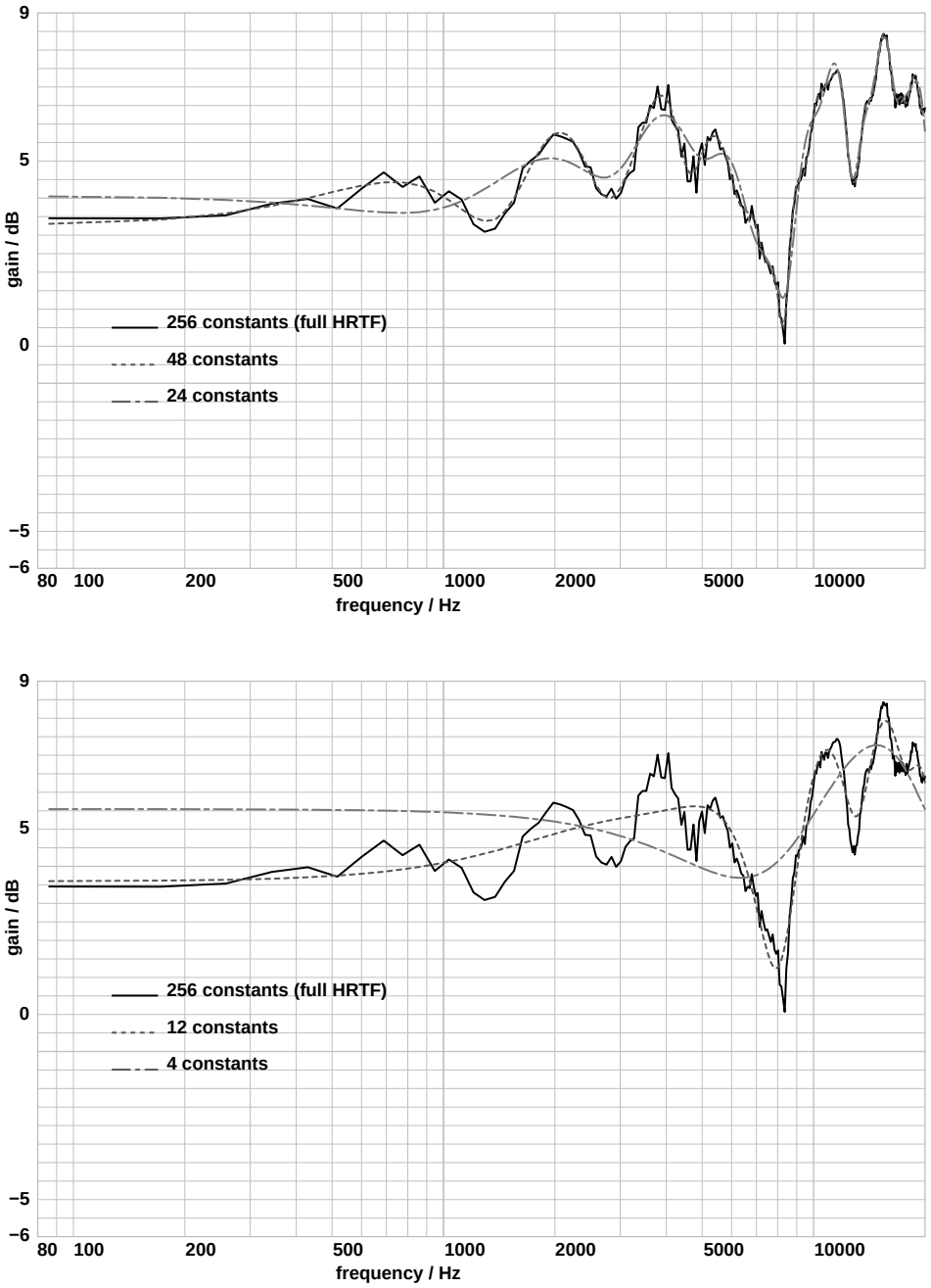


Fig.8: Equalised, phase-corrected HRTF at 30° towards the ear and 0° elevation, using energy-corrected rectangular truncation

For the sake of mathematical correctness (although it is not a strict psychoacoustic necessity), it was decided to build an energy-correcting algorithm into the truncation routine. This measures the total energy of the truncated part of the response. This is proportional to the sum of the squares of the sample values; the digital equivalent of the equation

$$P \propto V^2 \qquad \Rightarrow E \propto \int V^2 \, dt$$

so in the digital domain,

$$E \propto \sum V^2$$

This value is compared with the total energy in the whole impulse response. Each sample in the truncated part of the response is then treated in the following way:

[Sample value] = [Old sample value] ×
√([Total impulse energy] / [Energy of truncated part of impulse]

The truncated response has now been corrected to possess the same energy as the full-length impulse response. This did not make much difference to the values stored in the database: typically, the responses were amplified by a value between 0.5dB and 1.5dB. This has been included because interaural level differences are known to play a role in sound localisation: it is best to keep the simulation as precise as possible.

A further increase in computational efficiency is gained in the program by storing two 16-bit impulse responses alongside each other, packaged in 32-bit words. The impulse response for angle $(360° - \theta)$ is stored with each impulse response for angle θ up to 180°: the correct impulse response for the left and right ears at any particular angle can therefore be retrieved from the database simultaneously, with little extra processing power and no extra space demanded.

2.1.1.6 EXTENT OF THE PROCESSED DATABASE

Fig. 9, page 26 illustrates the resolution of the original and interpolated databases; the other statistics are shown below.

	MIT database	Interpolated database
Data points	710	5258
Median plane resolution / °	10	5
Maximum horizontal plane resolution / °	3	1
Horizontal resolution at 60° elevation / °	10	5
Memory occupied per impulse / bytes	1024	96
Total memory occupied / kilobytes	710	503.2

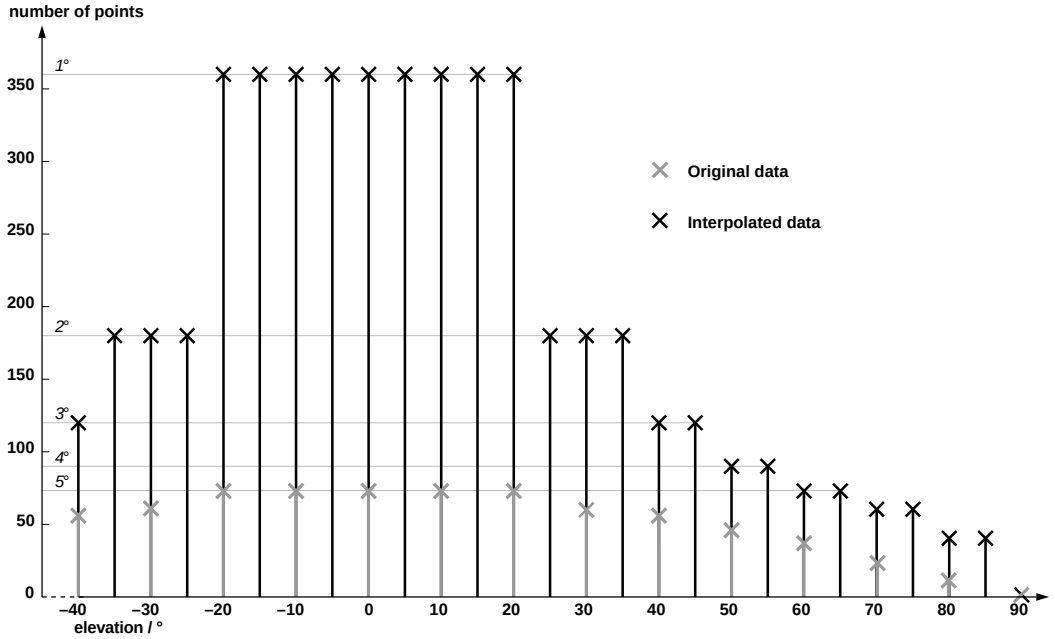


Fig. 9: Comparison of the number of interpolated points against the number of original points

A database has been created which has significantly reduced the memory and processor requirements required for data retrieval and manipulation, which has a flat average frequency response, and which possesses a spatial resolution significantly finer than the resolution of the original database. This has been achieved with only slight data quality impairment. With the individual variations in head-related transfer functions sometimes being very pronounced [Møller 1999], this is not a disadvantage: the head-related transfer functions will be no less correct for a real listener than they were before the process of truncation.

The simulation program will now be able to draw on a database which has adequate resolution to provide time difference, amplitude difference and spectral cues: an exhaustive list of the static binaural and monaural localisation cues is described by Wightman and Kistler [1992: 2]. These cues are stored at a high enough resolution to allow them to be varied synchronously with information supplied from a head tracker, thereby providing the dynamic cues necessary for above/below and front/back discrimination.

2.2 DISTANCE CUES

Ideally, distance cues would be subjected to the following restrictions:

- They must colour the simulated sound as little as possible;
- They should not demand so much processor time that the rest of the processing is unworkable.

It was decided immediately, however, that a small amount of simulated surround reverberation should be added. It is suggested that this is extremely helpful in externalising audio:

> *The addition of barely-audible reverberation 'pushes' the virtual source away from the listener.* [Robinson and Greenfield 1998: 4]

There is no shortage of papers which concur [Begault 1991: 10; von Békésy 1960: 302–304; Mershon 1979: 320], and it is also well-known that decreasing the correlation between the signals at either ear, which would be helped by the addition of some early reflections, is an aid to externalising sound [Sakamoto et al 1976].

The fact that this will inevitably colour the audio by introducing room modes is sometimes perceived as a disadvantage. It is preferable, however, to have slightly coloured sound than to have an anechoic simulation, which is unpleasant to listen to [Persterer 1991: 5]; especially when it is remembered that real acoustic environments, and particularly small rooms, possess room modes. They will improve the veridicality of the simulation.

Another important reason for including reverberation is to counteract listening fatigue (documented in [Watkinson 1998: 161]): a problem caused by listening in acoustically

dead environments, where the unnatural experience of hearing sound coming only from the direction of the loudspeakers, with no enveloping room reflections, tires the listener's hearing mechanism after a period of time. Watkinson puts his case strongly:

> *[Poor off-axis response in many loudspeakers] has led to the well-established myth that reflections are bad and that extensive treatment to make a room dead is necessary for good monitoring. This approach has no psychoacoustic basis* … [1998: 162]

This statement reinforces the body of evidence which suggests that artificial reverberation enhances headphone listening. The implementation of early reflections in the simulator is covered in detail in §2.2.2.

It was not deemed necessary to include air absorption in the simulation, which affects sound over large distances, because the distances involved in a rectangular room simulation are comparatively small. Interaural level differences, which are subtle and affect sound only over short distances, were not included because the distances over which they are most effective are greater than the dimensions of the simulated rooms.

A reverberant tail, to complement the early reflections, has also been omitted. This would be too demanding on the processor, and it was decided to assume that a small number of early reflections would provide all the envelopment necessary to avoid listening fatigue and to provide a sense of distance from the loudspeaker.

Apparent source width is also not an issue, as the simulation deals with loudspeakers which are ideal point sources: image width is an illusion which will be created explicitly by the interaction of the two sources.

2.2.1 DISTANCE PERCEPTION — THE CRAVEN HYPOTHESIS

Gerzon [1992] states the Craven hypothesis, and introduces evidence to support it. The hypothesis states that the brain is able to ascertain the distance of a sound source from the listener by considering early reflections.

When a sound wave propagates, it obeys the inverse distance pressure law: its sound pressure is proportional to the reciprocal of the distance it has travelled. A reflection from a boundary will have travelled further than the direct sound, and therefore possesses a

sound pressure relative to the original signal:

$p = d / d'$

where p is the sound pressure;

d is the distance which the direct sound has travelled;

d' is the distance travelled by the reflection.

The delay between the direct sound and its reflection is also a function of source and image distance:

$t = (d' - d) / c$

where t is the time delay between the direct sound and its reflection reaching the listener;

c is the speed of sound.

By combining these two equations, d' may be eliminated:

$d = t\,c\,/\,(1 - p)$

According to the Craven hypothesis, the brain can use this formula to approximate source distance solely by assessing the relationship of time and amplitude of a number of early reflections with respect to the direct sound. This is true even though the formula is only approximate for room reflections, owing to the energy absorbed by the boundaries.

2.2.2 IMPLEMENTATION OF EARLY REFLECTIONS

A reverberation simulation program was designed, called ReverbCalc. This operates on a two-dimensional model of a rectangular room, whose basic parameters can be adjusted by modifying a short text file *(Fig. 10, page 31)*. The program uses the image-source method [Jot et al 1995; Allen and Berkley 1979; Lerhnert and Blauert, 1992: 264] to calculate the path length and angle of incidence of each reflection. From this it can derive the attenuation owing to distance travelled and surfaces encountered, and the delay, in terms of milliseconds and samples, relative to the direct sound. The program also lists the surfaces which each reflection has encountered.

A number of decisions were made based upon psychoacoustic principles: these are summarised below.

a) A two-dimensional simulation was used. The floor and ceiling of the room are therefore anechoic. This simplification is based on two assumptions: that height information is not required to achieve a sense of auditory envelopment, and that only a small number of reflections are required to give the simulated loudspeakers a sense of distance. Rubak [1991] suggests that a convincing simulation can be achieved using only four early reflections. An early experiment conducted with the simulator, which attempted to introduce one floor reflection and one ceiling reflection, showed that this was not enough to provide a sense of distance. This approach was also rejected because it would fail to provide a sense of envelopment.

b) The front wall (the wall behind the loudspeakers) was also considered to be anechoic. Spatial information is already presented in this sector of the listening room by the loudspeakers: it was decided that adding reflections here would only muddy the sound and make small room simulations too ‘live’. Implementing virtual loudspeakers here for extra early reflections would not be a prudent use of computing power which is more urgently needed to represent early reflections in the remaining 300 degrees of the horizontal plane.

c) Psychoacoustic literature [Blauert 1989; Hartmann 1997; Moore 1989: 208] suggests that any sound arriving 40ms or more after the direct sound (or even earlier for sources of a transient nature) will be perceived as a discrete echo. As the purpose of these reflections is to lend a sense of envelopment and depth to the simulation without altering the nature of the programme material or compromising the ‘quality’ of the audio passing through it, these later reflections are not included in the simulation.

Taking these assumptions into account, there are only a small number of early reflections from each loudspeaker which are valid for simulation, nine of which were chosen. Four additional anechoic point sources around the head were then chosen to convey the acoustics of the virtual listening room. These are distributed fairly evenly around the listener, and are referred to as Left 75, Right 75, Left 160 and Right 160 *(Fig. 11, page 32)*. Nine reflections from each loudspeaker were used in the simulation because they were approximately coincident with these ambient points. The capital letters correspond

to the names given to each early reflection (see footnote at *Fig. 12, page 32*):

Left A, Right O, Left OA, Right AO, Left AOA, Right OAO	→ Left 75 source
Right A, Left O, Right OA, Left AO, Right AOA, Left OAO	→ Right 75 source
Left B, Left AB, Right OB	→ Left 160 source
Right B, Right AB, Left OB	→ Right 160 source

Two alternative reverberation models based on this system were used in listening experiments: see §3.2.2.

```
# ReverbCalc file
# ------------------------------
# Domestic listening environment
# Near field loudspeakers
# ------------------------------
#
SRATE:       44100
ORDERS:      8
SPEED_SOUND: 326
ROOM_WIDTH:  2.8
ROOM_LENGTH: 3.55
SPEAKERS_IN: 0.75
LISTENER_IN: 1.8
BACK_ATTEN:  0.9
SIDE_ATTEN:  0.9
```

Fig. 10a: An input file for the simulation. BACK_ATTEN and SIDE_ATTEN are constants which are multiplied by the amplitude of a signal each time it is reflected from a side wall or the back wall.

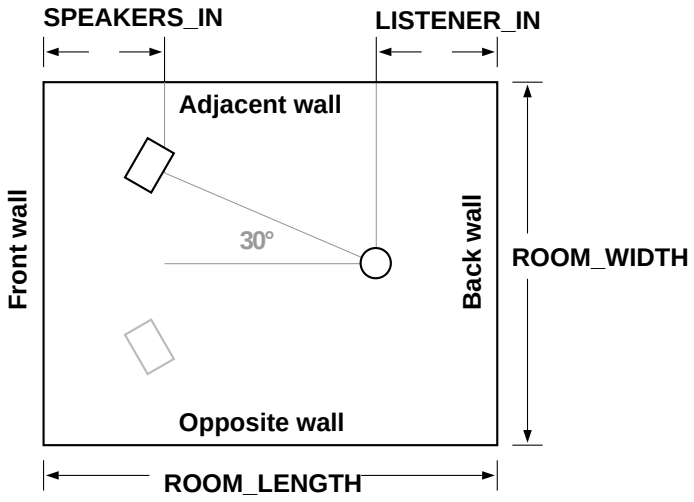


Fig. 10b: Listening room illustrating the parameters used by ReverbCalc.

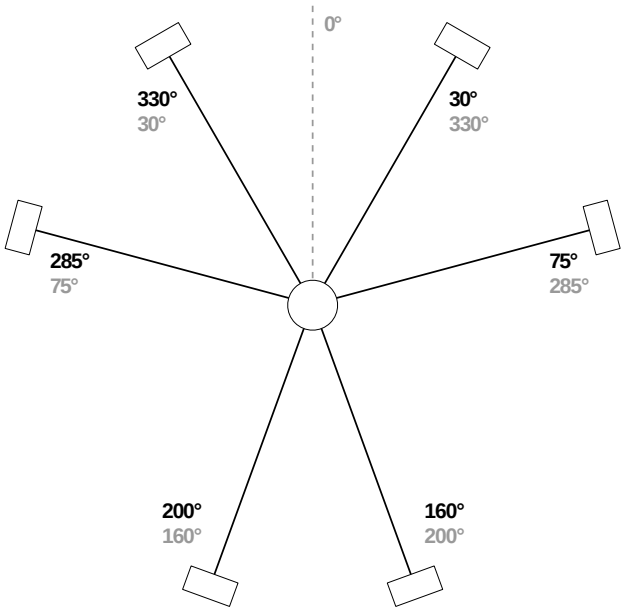


Fig. 11: The six virtual angular positions for the main stereo and ambient audio.

Sample rate: 44100

	Distance			Angle	Attenuation		
	Metres	mSecs	Samples	Degrees	Rel.1	Rel.32768	dB
Absolute:							
Direct	**1.155**	**3.54**	**156.2**	**30.0**	**1.0000**	**32768**	**+0.00**
Relative:							
A	**1.283**	**3.93**	**173.5**	**65.8**	**0.4264**	**13972.1**	**-7.39**
O	**2.368**	**7.26**	**320.3**	**286.5**	**0.2950**	**9668.0**	**-10.59**
B	**3.481**	**10.68**	**470.9**	**172.8**	**0.2242**	**7345.3**	**-12.98**
AB	**3.954**	**12.13**	**534.9**	**154.2**	**0.1831**	**5999.1**	**-14.74**
OA	**3.967**	**12.17**	**536.6**	**281.3**	**0.1826**	**5984.5**	**-14.76**
OB	**4.552**	**13.96**	**615.8**	**216.3**	**0.1639**	**5370.6**	**-15.70**
AO	**5.103**	**15.65**	**690.3**	**80.8**	**0.1495**	**4897.6**	**-16.50**
OAB	5.656	17.35	765.1	227.5	0.1236	4049.9	-18.15
AOB	6.547	20.08	885.7	126.7	0.1093	3581.4	-19.22
AOA	**6.732**	**20.65**	**910.6**	**82.7**	**0.1067**	**3497.6**	**-19.42**
OAO	**7.878**	**24.17**	**1065.7**	**276.4**	**0.0932**	**3053.7**	**-20.60**
AOAB	7.920	24.30	1071.4	120.5	0.0835	2735.6	-21.56
OAOB	8.933	27.40	1208.4	242.9	0.0751	2461.0	-22.48
OAOA	9.515	29.19	1287.1	275.4	0.0710	2326.7	-22.96
OAOAB	10.421	31.97	1409.7	246.6	0.0589	1930.1	-24.59
AOAO	10.665	32.71	1442.7	85.1	0.0641	2100.3	-23.85
AOAOB	11.489	35.24	1554.2	111.3	0.0539	1767.1	-25.35
AOAOA	12.305	37.75	1664.6	85.7	0.0507	1659.9	-25.90
AOAOAB	13.034	39.98	1763.2	108.9	0.0432	1417.2	-27.27

Fig. 12: ReverbCalc's output file, showing all valid early reflections for the domestic near-field listening environment. The reflections which were accepted for the simulation are shown in bold face. (Note: 'B' indicates a back wall reflection; 'O' indicates a reflection from the side wall opposite the loudspeaker; 'A' indicates a reflection from the wall adjacent to the loudspeaker).

3 AURALISE: A LISTENING ROOM SIMULATOR

Real-time processing requirements for the simulator have already been decided implicitly by limiting the HRTF filters to 48 coefficients, and by placing four extra point sources around the listener to simulate a listening environment, into which a total of eighteen simulated reflections will be fed. A flowchart depicting the way in which the program must work is shown in *Fig. 13, page 34.* This amount of real-time calculation is far from trivial: as soon as the implementation was finished, it was realised that a number of compromises would have to be built in as well. This includes being able to remove the Left 75 and Right 75 sources from the simulation (avoiding the need to calculate reflections and to convolve the audio for these points), and to be able to change the number of HRTF filter taps for the loudspeakers and ambient points separately. Adding these features meant that the program could run comfortably in real time without causing the audio to glitch, and the HRTF filter quality of both categories of sources could be balanced against processing power.

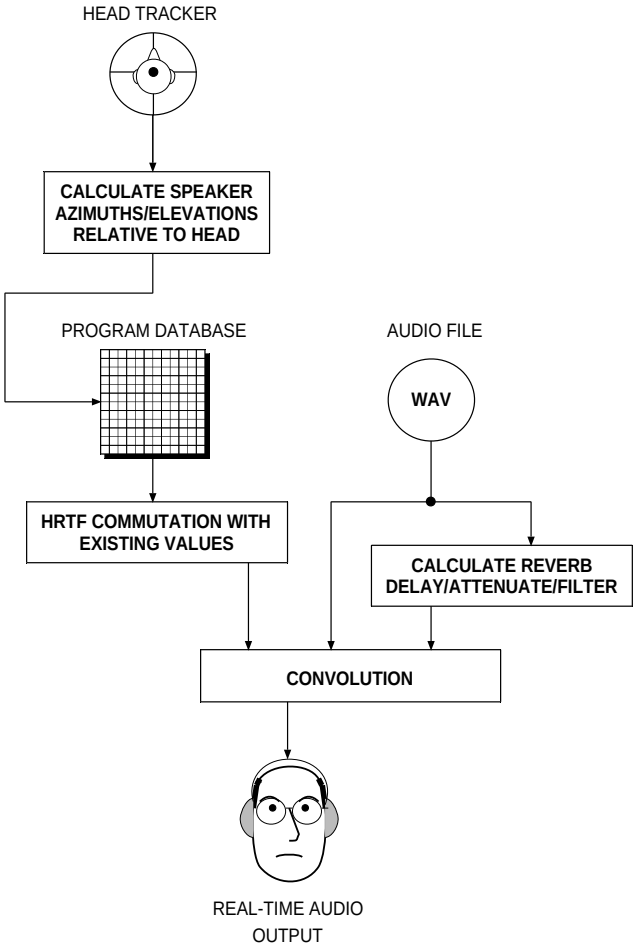


Fig. 13: Real-time processing in the simulator

3.1 HANDLING AUDIO FILES

The program will only accept stereo WAV files with a sampling frequency of 44.1kHz and a word length of 16 bits. The handling routine is implemented (referring to [Baharav 1996] and [Cross 1997] on details of the WAV file format) by loading data into one of two separate areas of memory (buffers). The computer can work on one whilst the other is updated by loading the next section of an audio file from disc. Owing to the nature of convolution and reflection calculation, there must be some interplay between the buffers: the computer needs information from the recent past to perform these operations. Therefore, a small chunk of data from the end of one buffer is copied to an area just before the next buffer so that there is always enough past sample data to perform this processing.

The double-buffering technique is illustrated in *Fig, 14, page 37.* Owing to the way in which it only loads up small portions of the sample at once, the program can run using a comparatively small amount of RAM, which is ideal when the design brief of the project specifies an affordable listening room simulation.

	Memory occupied
Audio buffers, one for each point	6 × 256kb
Large history buffers for the main audio	2 × 9kb
Small history buffers for the ambient points	4 × 0.75kb
HRTF database	503kb
Buffers to contain retrieved HRTFs	24 × 0.2kb
Object code	8.5kb
Total	2073kb (2.02Mb)

3.2 REAL-TIME DSP

3.2.1 CONVOLUTION

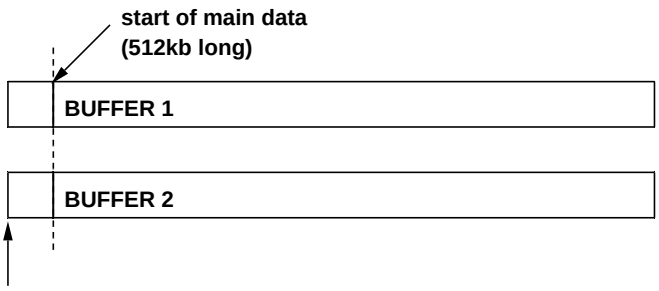
The process of audio convolution is well-documented in digital signal processing textbooks [Ifeachor and Jervis 1993; Gardner and Martin 1994: KEMAR Q18], so it is not necessary to present a quantitative description of the process here.

Convolution is useful because it filters audio quickly and simply by using the time-domain impulse response of a FIR filter. The number of arithmetic operations required to perform this process is directly proportional to the impulse response length, so that the digital signal processing power of the computer can be measured coarsely in terms of the number of filter taps it can convolve in real-time.

The 233MHz StrongARM processor upon which this software was developed can convolve up to 180 orders in real-time when the reverberation processing and commutation are working. This means that the maximum workable impulse response lengths for a six-point (stereo with four-point surround) system are:

Main points:	25 orders each
Ambient points:	10 orders each
	25 × 2 [points] × 2 [ears] = 100
	10 × 4 [points] × 2 [ears] = 80
Total:	176

An informal aural examination suggests that although no serious quality impairment is audible when running such low-order convolution, a detectable (but very small) amount of image smearing occurs at the main points. This performance ceiling, however, is an inevitable by-product of using a microcomputer rather than a digital signal processor for performing DSP operations.



Initial state	Step 1	Step 2	Step 3
First buffer has been filled with three seconds of audio. First buffer's contents are played out. Second buffer is empty.	52ms of audio is copied from the end of the first buffer into the second buffer's history section. This is necessary for convolution and reverberation calculations.	The next three seconds of audio are loaded into the rest of the second buffer, while the contents of the first buffer are played out.	The first buffer has been exhausted after three seconds.

Step 4	Step 5	Step 6	Step 7
The computer instantly begins to play out the contents of the second buffer.	52ms of audio is copied from the end of the second buffer into the first buffer's history section.	The next three seconds of audio are loaded into the rest of the first buffer, while the contents of the second buffer are played out.	This is similar to the original state. The contents of the first buffer are played out, and the sequence is repeated from step 2.

Fig. 14: Double-buffering technique used to read audio from disk in short sections

3.2.2 GENERATION OF REFLECTION DATA

Audio data to be reproduced at the four ambient points is obtained by delaying and attenuating the left and right loudspeaker audio, according to the parameters obtained from ReverbCalc. Each reflection is then routed to one of four extra audio buffers. The buffer used for each reflection is the one whose simulated angle most correctly approximates the calculated direction of the reflection.

To improve the realism of the simulation, each delayed reflection is also filtered to approximate a surface reflection, although this is achieved with a highly-simplified FIR filter structure. Three two-tap FIR filters were used initially to represent the reflections: a different filter for each simulated order of early reflection. The impulse responses and corresponding frequency responses are reproduced in *Fig. 15, page 39.*

It was apparent from feedback from the first six listening experiment results (§4) that many listeners were perturbed by colouration of the processed sound with the ambient loudspeakers on, particularly at high frequencies. Consequently, it was decided to modify the simulator before the second set of tests, firstly to remove some of the high-frequency colouration, and secondly to approximate the behaviour of real surfaces more convincingly.

Instead of the eighteen two-tap FIR comb filters which had previously been used, the reverberation model was changed to fourteen three-tap FIR filters, containing both a comb element and a low-pass element *(Fig. 16, page 39)*: this approximates more closely the Bode plots of reflecting surfaces reproduced in Lehrnert and Blauert [1992: 275] and Savioja et al [1999: 685]. Six reflections (three from each loudspeaker) had to be disregarded owing to the amount of extra data processing required in the formation of ambient point data: these were reflections AO, OAO and OB (see *Fig. 12, page 32*).

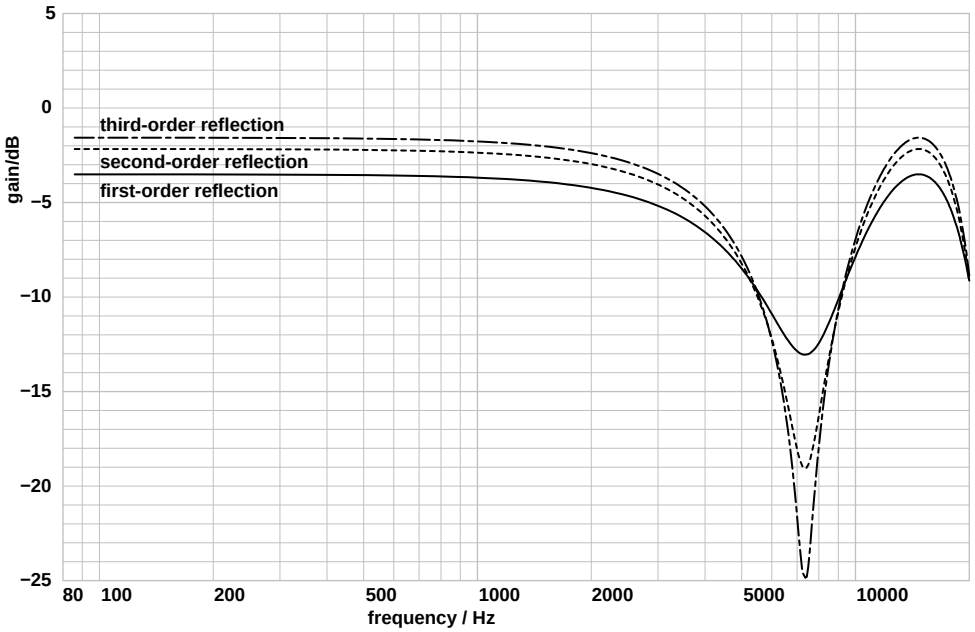


Fig. 15: Two-tap FIR filter frequency response for first six listening tests.

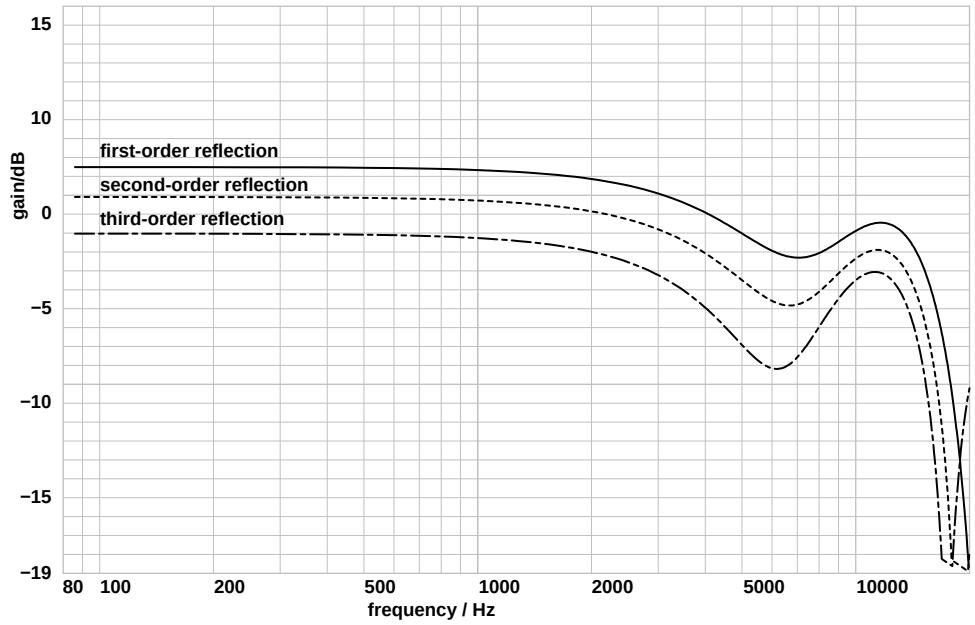


Fig. 16: Three-tap FIR filter frequency response for remaining listening tests.

3.2.3 COMMUTATION OF HEAD-RELATED IMPULSE RESPONSES

At an early stage, it was discovered that changing the filter coefficients during run time, even when they were changed only to adjacent angles, produced audible clicks in the audio. It was therefore necessary to add some commutation to the simulator. Using an IIR filter style geometry, which transforms the impulse response coefficients exponentially from one to the other over a number of samples, succeeded in attenuating the clicks, but they were still audible and annoying.

The successful solution, which is employed within the program, works by sliding the old impulse response samples into the new ones linearly, by adding or subtracting two from each impulse sample every time an audio sample is calculated. The remainder is tested for and added if necessary whenever the filters are changed. Audible clicking no longer occurs, but the penalty is an added, indeterminable slowness of reaction to the change of head angle.

3.3 HEAD TRACKING

The computer is required to interface with the head tracker at least every 85ms [Horbach et al 1999: 5]. The simulator is informed by another program when the head angles need to be re-calculated and the filter banks changed. This permits the use of alternative head tracker driver software.

3.3.1 COMMENTS ON THE CHOICE OF HEAD TRACKER

The simulator was built around a General Reality (now i-Reality) CyberTrack-II head tracker. It was decided that this was the most suitable piece of equipment for the experiment owing to its relatively low cost and small size. Its specifications are well within the boundaries of this experiment, the main issues being the delay between the head angle being read by the head tracker and its arrival at the computer (specified as 18ms) and accuracy of tracking (±1°) [General Reality 1996].

Its low price compared with other digital head trackers (such the Polhemus FastTrak, which is frequently used in audio experiments [McKeag and McGrath 1997b; Horbach et al 1999]) results in a number of compromises.

The first is its relatively large size. The more expensive head trackers, including FastTrak,

are sourced: the head tracker comprises a transmitter which is fixed to the computer and a passive, cordless receiver which is attached to the head. The CyberTrack-II, however, is sourceless. This means that box of electronics containing an electronic compass and two inclinometers, accompanied by a trailing RS232 cable, needs to be attached to the headphones. The box is lightweight, but sufficiently cumbersome to require listening experiments to be conducted with a head-band around the headphones to keep them stable.

The second compromise is that the CyberTrack-II provides only three degrees of freedom: it will track the listener's head movements in terms of yaw (rotation), pitch (tipping) and roll (pivoting). (The terms in brackets are the ones used in Thurlow et al [1967]).

Polhemus offer six degrees of freedom: in addition to the angles of rotation, it has limited ability to track the spatial co-ordinates of the listener's head. At this stage, the spatial co-ordinates would almost certainly furnish too much information to be processed in real time. The potential usefulness of a head tracker permitting six degrees of freedom is discussed in §5.2.

3.3.2 THE CYBERTRACK-II DRIVER

Because the CyberTrack-II is not supplied with driver software that is compatible with Acorn computers, it was necessary to write an Acorn compatible driver. It is a relatively simple piece of software, and interfaces with the computer's RS232 port, providing three basic features which can be used by another program:

- Initialising (resetting) the head tracker;
- Reading and formatting data in terms of the three angles of rotation;
- Arithmetically zeroing the head-tracker to an arbitrary position.

There was a single important design challenge when writing this software, and it resides in the simplicity of the head tracker's implementation of RS232 serial protocol. A 'three-pin' version of the system is used, in which there is no hardware hand-shaking, simply data transmit and receive lines and a common ground. Because the simulator takes a relatively long time to process data, the computer can stay in interrupt mode and lose any data that is sent by the head-tracker to the computer. This corrupts received data.

There is no way, as there is with RTS/CTS RS232 protocol [Horowitz and Hill, 1989], to stop the sender passing data, and no way for the sender to tell whether or not the other device is in a position to receive it.

The only way around this problem is to re-enable interrupts during the sound processing routine so that the serial port can take priority over the sound system; this method is used in the simulator. However, it has the intrinsic disadvantage that other system interrupts can then stop the processor from working, and the routine already occupies a large proportion of processor time. It was found, for example, that loading audio data from the IDE hard disc, which is an interrupt-heavy operation, diverted so much processor time away from the sound system that the sound system would fail whenever it happened. Fortunately, a SCSI hard disc became available: this works using DMA (direct memory access) and therefore interrupts the processor minimally. The drawback still persists, however, that the simulator cannot be guaranteed a proportion of processor time during this operation when it has to work in tandem with the CyberTrack-II. Occasionally, when the simulation is run with a high number of filter coefficients, this can cause an audible glitch in the ambient points every time the buffer is changed.

3.3.3 PROCESSING THE HEAD TRACKER DATA

Each of the two loudspeakers and the four ambient points are represented as a source at a fixed distance and a given angle from the head. Taking such a point and rotating it appropriately through yaw, pitch, and roll to obtain a point in terms of azimuth and elevation is not a trivial undertaking. Not only do the points themselves rotate in space, but also the axes of subsequent rotation. The mathematics have been reproduced in Appendix A.

4 EVALUATION OF THE SYSTEM

In order to assess the veridicality of this simulation, it was necessary to compile an objective test in which a number of its features were evaluated. The program *!Test* along with the answer paper in Appendix C were presented to each individual. A set of quality 'hi-fi' headphones, the Vivanco SR750, was used.

The listening test was divided into three sections: each section was devised to test a different aspect of the simulation. In each test, the filter resolution was set throughout to 19 coefficients for the loudspeaker points, and 7 coefficients for the ambient points. These were the highest usable values. If higher values were used, the reverberation calculations would not be finished in time which would cause audible glitches, and the increased sluggishness of the head-tracking would become perceptible.

4.1 SUBJECTIVE EVALUATION

The first section is designed to test Martin Thomas's discovery, under informal conditions, that listeners prefer listening to sound passed through a stereo-to-binaural converter over listening to unconverted sound. All subjects in the revised test were expert listeners. This reflects the target market for a commercial head-tracked auralisation system (see §5.2).

Subjects were played five extracts, and provided with an 'A/B' button (as an icon on the screen) to enable them to switch the processing on or off. The order was randomised so that in some extracts, A was unprocessed and B was processed; in other extracts this was reversed. To prevent the test results from being biased artificially, care was taken not to inform the subjects about the nature of the differences between the two signals, or the type of processing involved. A space was provided so that they could comment on the reasons for each choice, with instructions on the screen asking the subjects to concentrate on stereo image and frequency response. The programme material was chosen to represent a broad variety of musical styles and recording techniques (see Appendix B).

The 'stereo-to-binaural' conversion set the simulator to a simulation with four ambient points, with the reverberation characteristics of a 'reference near-field' environment, to avoid excessive room mode colouration (*Fig. 17, page 44).* In order not to give anything away, head tracking was turned off and the simulation stayed fixed at 0° yaw, pitch, and

roll.

```
# ReverbCalc file
# -------------------------------
# Reference environment
# Near field loudspeakers
# -------------------------------
#
SRATE:       44100
ORDERS:      3
SPEED_SOUND: 326
ROOM_WIDTH:  4.7
ROOM_LENGTH: 5.2
SPEAKERS_IN: 1.47
LISTENER_IN: 2.73
BACK_ATTEN:  0.87
SIDE_ATTEN:  0.75
```

Fig. 17: ReverbCalc input file for the 'Reference near-field' environment

The results, collected from twelve subjects, were unexpected and consistent. In all but one case, nine of the twelve subjects favoured the unprocessed sound. Many subjects criticised the lack of bass response in the processed extracts: this can be attributed to two factors:

- The presence of room modes in the reverberation synthesis;
- The need to compromise simulation quality by truncating filter coefficients caused extra bass roll-off.

A brief personal confirmation revealed that the definition and sound quality of the bass, particularly the bass guitar in the Steve Reich extract, is restored when the simulation is anechoic and 48 filter coefficients are used (this is possible when only two points are simulated). Many listeners also preferred the un-natural envelopment of unprocessed audio compared with the simulated listening environment. The minority who did not, made comments favouring the externalisation effect.

A few contradictory comments were noted between the answer papers, although these all referred to clarity versus width of the stereo image: some preferred the greater width of the headphone image, and others (especially in the classical extracts) preferred the accuracy of the processed image. A few notable comments from each category follow.

"Wider, more spacious, more bass [in the unprocessed extract]."
"[The unprocessed extract] sounded slightly muffled. The image of [the processed

extract] seems clearer."
"[The processed extract] sounds quite 'thin', but [the unprocessed extract] suffers from a slight 'hole in the middle' and is closer to the head."
(Ravel orchestral extract)

"[The processed extract] sounds muddy and weird"
"[The processed extract] felt too mono"
"I get the impression in all of these that there is some trade-off between high frequency clarity and width of image"
(Squarepusher extract)

"[In the unprocessed extract] panning is very effective"
"More defined positioning and envelopment [in the unprocessed extract]"
"[The processed extract] has the better frequency response]"
(Steve Reich extract)

"[The unprocessed extract] sounds like you're in the piano; [the processed extract] like you're looking at the piano."
"[The unprocessed extract] is tiresome to listen to, but it is louder. [Subject preferred the processed sound.]"
(Flute and piano extract)

In the remaining example, 'Planetary Citizen', six favoured processed sound and six favoured unprocessed sound. Those who preferred the processed sound commented that the panning of sources in this extract was too extreme: it was generally felt that the processed sound was more spacious. The unprocessed track was considered disconcerting because the subjects seemed to be positioned in the centre of the band; the loudspeaker simulation placed the band in front of them.

This phenomenon, interestingly, was not seen as often in the less artificial Steve Reich extract, which featured panned, undistorted electric guitars (some fully left or fully right) with a general stereo reverberation. It can be inferred that the lack of recorded ambience in 'Planetary Citizen' made its panning seem more extreme and. in some cases,

unpleasant.

Most listeners remarked at the end of the test that if they had heard the head-tracked examples, they would have reconsidered these choices. It is possible that, because only expert listeners were employed, many were put off by listening to an image which they were not used to hearing through headphones, and that explaining the processing beforehand would have influenced the results in favour of the loudspeaker simulation. A situation where people would favour a particular choice out of a sense of loyalty is the one which I had made an effort to avoid.

4.2 LOCALISATION

The second section of the test was designed to gauge the effectiveness of head-tracking on each subject's ability to discern between sources located in front and behind. Six examples were presented: three to the front and three behind the listeners. Subjects were instructed that they were free to move their heads, and that this may or may not change the headphone sound in each extract. The Ravel orchestral music was used for each extract.

In one pair of extracts, the head tracking was turned off and the only cues which the listener obtained were from the HRTFs. When the loudspeakers are rotated via 180 degrees, they are also left-right reversed with respect to the listener. With hindsight, it is possible that some listeners based their decisions on this left-right reversal, but the results suggest that this did not generally happen.

In the other four extracts, head tracking was turned on. In one pair of these, an anechoic listening environment was simulated. The other pair included reverberation from the 'Reference mid-field' environment *(Fig.18, page 47).* This part of the test was intended to determine whether or not the addition of ambient points around the listeners would confuse their ability to discriminate between front and rear stimuli. The environment was chosen for its relatively large reverberant content.

```
# ReverbCalc file
# ———————————————
# Reference environment
# Mid field loudspeakers
# ———————————————
#
SRATE:       44100
ORDERS:      3
SPEED_SOUND: 326
ROOM_WIDTH:  4.7
ROOM_LENGTH: 5.2
SPEAKERS_IN: 1.69
LISTENER_IN: 2.16
BACK_ATTEN:  0.87
SIDE_ATTEN:  0.75
```

Fig. 18: ReverbCalc input file for the 'Reference mid-field' environment

Results were again consistent. They have been presented in *Fig. 19*. (Note that this test only features eleven of the twelve subjects owing to some incorrect settings on the pilot test. This pilot test has been included in the other two sections).

Head tracking	Ambience	Stimulus	Correct responses	Incorrect responses
OFF	OFF	FRONT	6	5
		REAR	6	5
ON	OFF	FRONT	11	0
		REAR	10	1
ON	ON	FRONT	9	2
		REAR	10	1

Fig. 19: Localisation test results

The results show conclusively that the addition of head tracking to the simulation increases the subjects' ability to discriminate between sources to the front and sources to the rear. The incorrect responses can almost certainly be attributed to inexperience with the system: more than one subject reported that at first, they were reluctant to make any significant head movements in the test. The single unanimous correct response (front stimulus, head tracking on, ambience off) was obtained in the last extract of the section, by which time subjects were used to listening to the head-tracked audio. The results also show that reverberation may have the potential to confuse some listeners, but a larger number of subjects would need to be examined before any conclusive statement can be

made. Front/back localisation results are still far better than those obtained without head-tracking.

The results also show, surprisingly, that there was no tendency in listeners to envisage both stimuli coming from behind them when the head tracking was turned off. In addition to this, listeners are equally adept at correctly identifying sources behind them when the head tracking is turned on as they are at locating them correctly in front.

4.3 APPARENT DISTANCE PERCEPTION

Three different extracts (Ravel, Steve Reich, and Squarepusher) were played three times, each time with a different reverberation characteristic. One anechoic and two different reverberant simulations were used. Early reflections were modelled according to the approximate proportions of the room in which the tests were conducted (2.8 by 3.55 metres), with the loudspeakers placed either 1m or 1.5m from the listener in the room. Extracts were again played in a randomised order. For each extract, subjects were asked to place a mark on a line on the answer paper, corresponding to source distance. The line started inside the head and finished at 2.5 metres from the listener, with five markings at convenient distances. The results are shown in *Fig. 20, page 50.*

Many subjects commented on the difficulty they had forming answers for this section of the test. Owing to the disparity between the spectra of the simulated reflections and real room reflections (see §3.2.2), it would be tempting to disregard the first set of results in absolute terms. They are still important, however, because they reveal some interesting traits:

1 An overwhelming majority of listeners placed the first extract (Steve Reich) at the verge of the head, in spite of the fact that a one metre simulation was used. This could be attributed to unfamiliarity with the system. The second extract (Ravel), which is anechoic but contains a large amount of natural ambience, was externalised more successfully.

2 Most listeners externalised the anechoic source (Steve Reich, extract 8) at the end of the test. The wide scattering of readings towards the end of the test suggests that the subjects had become so used to the system that all sources were becoming externalised. It may suggest that reverberation is only needed to convey the suggestion

of distance initially, and that, when it is removed, the listener still perceives a similar sense of distance even though the stimulus has been taken away.

3 The anechoic simulation played as extract 4 of the test (Squarepusher) was localised at the verge of the head by a large number of listeners.

4 Although test subjects could generally not quantify absolute distance (this may be down to the shortcomings of the first early reflection simulation), there is a correlation between simulated and perceived distance in the middle extracts. In the later extracts, answers were distributed fairly evenly along the distance line. This reinforces the suggestion that the listeners were inexperienced to start with, and perhaps so used to the sensation of externalised sound sources by the end of the test that either their hearing mechanism was exhausted, or they did not notice the removal of the reverberation stimulus.

5 There were only three reported cases of ‘in-head-localisation’ in the distance perception experiment, in over one hundred seperate stimuli. Many listeners commented, however, that even though sources were successfully externalised, there were still some in-head artefacts present in the audio: this could be caused by a number of factors, such as the use of non-individual HRTFs, insufficient resolution of HRTFs, too few early reflections, insufficient accuracy of the ambience simulation, or simply the sensation of listening through unequalised headphones. This is an area which would require further investigation.

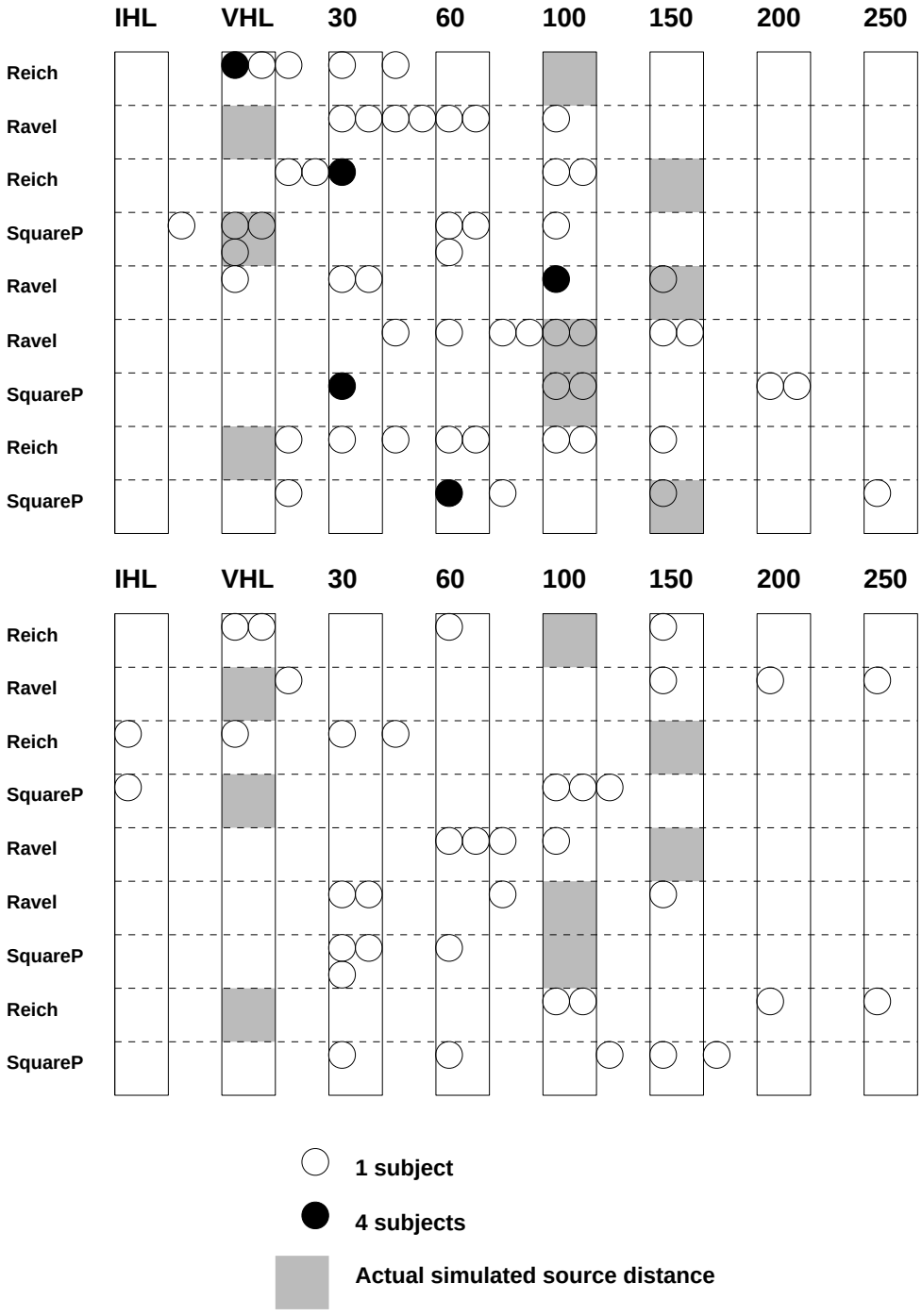


Fig. 20: Collated results from Section III of the test. *Top: first reverberation model*
Bottom: refined reverberation model.

The numbers heading each column are the marked distances, in centimetres, along the distance line. 'IHL' stands for 'In-Head Localisation'; 'VHL' stands for 'Verge-of-Head Localisation'.
Any marks put at a point between the main markings on distance lines have been placed between the numbered columns.

5 CONCLUSION

The development and evaluation of this simulator has explored a number of aspects of loudspeaker stereo to binaural conversion. It is possible to design a high-quality head-tracked system on a single conventional microprocessor with relatively small amounts of memory by current standards. This simulator is unique because, in spite of these constraints, it eliminates after a few minutes' use the common problems of front/back ambiguity and in-head localisation. It achieves this not by using a database of individualised head-related transfer functions, but by employing head tracked dynamic cues. The simulation does not require a large amount of computation resources in order to store and process the data necessary to make the simulation realistic.

It has been necessary to make a number of compromises. The number of filter coefficients was limited severely by processing constraints if early reflections were simulated. Although this did not upset the directional illusion for any test subject, the image of each projected point lost definition when a small number of coefficients was used. This should not be disconcerting when simulating the early reflections, as ideally they should not come from easily locatable point sources. The simulated loudspeakers, however, appear to smear the image very slightly. These compromises could also be a possible cause of the in-head component suggested by a minority of test subjects.

Another compromise was the limit to the number of early reflections used: only six could be modelled before the practical processing limit was reached. It would have been interesting to investigate the addition of point sources at Left 105° and Right 105° degrees and some floor and ceiling reflections to increase the illusion of envelopment, and to change the number of early reflections simulated to see whether or not this enhanced the simulation. A novel way of simulating many virtual point sources without increasing the processing burden appreciably, by using the Ambisonic B-format in an intermediate stage, is suggested by McKeag and McGrath [1997a].

There would be little point in investing considerable time and energy developing technology like this if there was no area of the audio industry where it could be useful. Discussion of the technology's potential is divided into three categories (§5.1–5.3), addressing the consumer and professional markets, and the research field.

5.1 VIABILITY AS A CONSUMER PRODUCT

Although the additional hardware required to create a head-tracked auralisation system is expensive, attempts are constantly being made to apply simplified versions of this technology to consumer products. The latest manifestation of these attempts is Dolby Headphone [Dolby Laboratories, 1999], which claims to simulate any one of three listening environments through headphones using a similar but un-tracked system, built onto a single integrated circuit. However, the listening tests within this project (§4.1) demonstrate that people who have not been introduced to this system may find the headphone image unfavourable, especially if they do not understand the processing which is being applied to the audio. If Dolby Headphone becomes a common inclusion in consumer products, manufacturers would be foolish not to include a way for the user to turn off the processing.

The cost of building a product which compares with the demonstrated system, and particularly the cost of a sourced head tracker, is unreasonably high for the domestic market. It seems unlikely that many people would buy one in preference to a high-quality set of loudspeakers of equivalent cost. Domestic virtual reality technology, with the exception of transaural processing, has not yet become affordable or commonplace. If it ever does, manufacturing technology will be ready for a far more sophisticated system.

Finally, the main problem with a headphone-based system is that it cannot be used by more than one person at once: it it useless for a social environment as it would prevent conversation, and each individual would need his or her own system. It can never realistically be seen as an outright replacement for hi-fi loudspeakers.

5.2 VIABILITY AS A PROFESSIONAL PRODUCT

Although most professional environments already possess high-quality monitoring and listening rooms, it is expected that the loudspeaker simulator would find use chiefly in these environments. An enhanced version of this prototype could serve two valuable purposes in this field: firstly, the ability to simulate different environments would be an ideal way of gaining an impression of the sound of a mix in a number of different rooms. A multi-channel version intended for film use could simulate a large cinema, either empty or full of people, or a small home cinema, for example. The realism of the simulation, however, may depend considerably upon the design of headphones.

The system would also be very useful for making location recordings without the aid of a mobile studio. If a loudspeaker simulation with a sourced head tracker was built into a small rack, or even into the mixing console, then it really could behave as a loudspeaker replacement. The fact that such a system could only be listened to only by one person at a time would not be as important an issue as it is in the domestic situation. If the performers and the recording engineer had to share the same space, isolation between the two would be far better than it would be if normal monitoring loudspeakers had to be used. If the simulation can be re-written on a modern digital signal processor, it might be realistic enough for loudspeakers to become superfluous for small mobile recordings. This would reduce significantly the volume and weight of equipment which needs to be taken to the recording venue, and hence the time taken to unload, set up and load the apparatus again. If a surround mix is required, this saves even more time which would otherwise be invested connecting and calibrating each loudspeaker.

If a real product could be based upon this software, it would be wise to make it a multi-channel (5-channel at the very least) version to be aimed at the film sound and mobile recording sectors of the audio industry. This would have to use a sourced head tracker for two reasons:

- A sourceless head tracker, as used in the listening tests, is too heavy to be attached comfortably either to the head or to the headphones. Receivers for sourced head trackers, however, are small, lightweight and eompletely unobtrusive when they are are attached to a pair of headphones.
- The six degrees of freedom (three degrees of rotation and three dimensions of spatial

displacement) permitted by devices such as the sourced Polhemus IsoTrak would be essential in the professional field: recording engineers will always listen to the recording off-axis to test the stability of the stereo image, and might be dissuaded from purchasing a system which does not allow them to do this.

The only foreseeable problem with such a simulation is that it may sound too idealised: there will be no loudspeaker crossover distortion, no cone break-up, minimal interference from the simulated room acoustic, and no displacement of loudspeakers from their 'correct' angles. Loudspeakers represented by the system are point sources; real loudspeakers have a physical width. The influence of these subtleties on the veridicality of the simulation may require more investigation. The only way to overcome some of these would be to include them in the system as options, which increases its complexity, perhaps unnecessarily, from the point of view of the operator. Increasing the complexity of the device would also make it more expensive and more difficult to implement.

5.3 VIABILITY AS A RESEARCH TOOL

The viability of a loudspeaker auralisation system for use in academic and industrial research was discussed in detail by Horbach et al [1999: 9]. Important potential uses are stated including the blind comparison of different loudspeaker surround formats, and the development of listening environments, so that many institutions might be able to compare their research in a standard simulated room. This could either simulate an entirely hypothetical environment, with idealised loudspeakers and enough reflection simulation only to provide out-of-head localisation, or could be based on the parameters of real loudspeakers in a real room. The latter would require more processing, but could be developed and implemented at a fraction of the cost of building a listening room. There would be a number of issues to resolve, including the standardisation of headphones, to allow equalisation for an exact room response, and the establishment of a standard listening level. The simulation may even be expected to model exactly a real listening room. Its capacity, and indeed the capacity of any similar system for doing this realistically is a subject which would require further investigation.

6 GLOSSARY

auralisation The act of passing audio into a simulated listening environment, and the development of scientific principles behind the acoustic modelling of this environment.

DSP Digital Signal Processor: a type of microprocessor which is engineered specifically to manipulate and process data streams in real time.

DTF Directional Transfer Function: a single HRTF equalised by dividing the average response of a full set of HRTFs.

FIR Finite Impulse Response: a class of digital filter design

glitch A transient noise reulting from a defect in a digital audio system.

hand-shaking A type of communication between a computer and its peripherals where either device has the ability to prevent the other from transmitting data which it is too busy to process.

head tracker An electronic device which can gather instantaneous data regarding the angular, and sometimes spatial, position of a listener's head and relay this data to a computer.

HRIR Head-Related Impulse Response: an HRTF specified in the time domain.

HRTF Head-Related Transfer Function: a frequency domain representation of the effect of the head and the outer ear on a sound source which occurs at a given angle and distance from the head.

IIR Infinite Impulse Response: a class of digital filter design.

interrupt mode The mode of operation which a microprocessor enters when an event happens which requires the immediate attention of a particular piece of software. This could be, for example, a byte arriving at the serial port or the sound buffer needing to be filled. These events are called interrupts. Further interrupts will normally be ignored when the computer is already in interrupt mode.

ITD Interaural Time Difference: the time interval between an sound reaching one ear and it reaching the other.

OETF Own-Ear Transfer Function: a measurement obtained by recording the head-related impulse response of an individual; intended for inclusion in a system to which only this individual will be listening.

run time The period during which the computer is executing the auralisation program.

virtual reality An environment which has been generated with the aid of a computer, which provides correct visual, auditory, or tactile feedback, and with which a user can interact realistically.

All other terms are either explained within the text, or are recognised audio terminology.

7 REFERENCES

Allen, Jont, and Berkley, David, 1979: 'Image method for efficiently simulating small room acoustics' *J. Acoustical Society of America*, Vol. 65, No.4, pp943–950

Baharav, Roy, and Dror, Joel, 1996: *WAV File Format*
http://www.eng.tau.ac.il/~acmidi/format.htm

Bauer, B., 1961: 'Stereophonic Earphones and Binaural Loudspeakers' *J. Aud. Eng. Soc.* Vol. 9, No. 2, pp148–151

Begault, Durand, 1991: 'Perceptual Effects of Synthetic Reverberation on 3-D Audio Systems' *91st AES convention*, preprint number 3212

von Békésy, G, 1960: *Experiments in Hearing* (New York: McGraw-Hill)

Blauert, Jens, 1997: *Spatial Hearing — The psychophysics of human sound localisation* (Cambridge MA: MIT Press)

Cross, Don, 1997: *Reading and writing WAV files*
http://www.intersrv.com/~dcross/wavio.html

Gardner, Bill, and Martin, Keith, 1994: *HRTF Measurements of a KEMAR Dummy-Head Microphone*, ftp://sound.media.mit.edu/pub/Data/KEMAR/hrtfdoc.txt

Goodyer, Tim, 1997: *Surround Monitoring*
http://www.prostudio.com/studiosound/aug97/t_surround.html

Gerzon, Michael, 1992: 'The Design of Distance Panpots' *92nd AES Convention*, preprint number 3302

Hartmann, William Morris, 1997: 'Listening in a Room and the Precedence Effect' in ed. Gilkey, Robert, and Anderson, Timothy: *Binaural and Spatial Hearing in Real and Virtual Environments*, (New Jersey: Laurence Ellbaum Associates) pp191-210

Hartung, Klaus; Braasch, Jonas, and Sterbing, Susanne, 1999: 'Comparison of Different Methods for the Interpolation of Head-Related Transfer Functions' in *Proceedings of the*

AES Sixteenth International Conference: Spatial Sound Reproduction, pp319−329

Horbach, Ulrich; Karamustafaoglu, Attila; Pellegrini, Renato; Mackensen, Philip, and Theile, Gunter, 1999: 'Design and Applications of a Data-based Auralisation System for Surround Sound' *106th AES Convention*, preprint number 4976

Horowitz, Paul and Hill, Winfield, 1989: *The Art of Electronics* (Cambridge: Cambridge University Press)

Huopaniemi, Jyri, and Zacharov, Nick, 1999: 'Objective and Subjective Evaluation of Head-Related Transfer Function Filter Design', *J. Aud. Eng. Soc*, Vol. 47, No. 4, pp218–239

Ifeachor, Emmanuel and Jervis, Barrie, 1993: *Digital Signal Processing: A Practical Approach* (Harlow: Addison-Wesley)

Inanaga, Kiyofumi; Yamada, Yuji, and Koizumi, Hiroshi, 1995: 'Headphone system with Out-of-Head Localisation Applying Dynamic HRTF (Head-Related Transfer Function)' *98th AES Convention*, preprint number 4011

Jot, Jean-Marc; Larcher, Veronique, and Warusfel, Oliver, 1995: 'Digital signal processing issues in the context of binaural and transaural stereophony' *98th AES Convention*, preprint number 3980

Kistler, Doris, and Wightman, Frederic, 1992: 'A Model of Head-Related Transfer Functions Based on Principal Components Analysis and Minimum-phase Reconstruction' *J. Acoustical Society of America*, Vol. 91, pp1637−1647

Lehrnert, Hilmar, and Blauert, Jens, 1992: 'Principles of Binaural Room Simulation' *Applied Acoustics*, Vol. 36, pp259–291

McKeag, Adam, and McGrath, David, 1997a: 'Sound field format to binaural decoder with head tracking' *6th Australian Regional AES Convention*, preprint 4302

McKeag, Adam, and McGrath, David, 1997b: 'Using auralisation techniques to render 5.1 surround to binaural and transaural playback' *102nd AES Convention*, preprint 4458

Mershon, Donald, and Bowers, John, 1979: 'Absolute and Relative Cues for the Auditory Perception of Egocentric Distance' *Perception*, Vol. 8, pp311–322

Møller, Henrik; Hammershøi, Dorte; Jensen, Clemen Boje, and Sørensen, Michael Friis, 1999: 'Evaluation of Artificial Heads in Listening Tests' *J. Aud. Eng. Soc*, Vol. 47, No. 3, pp83–99

Moore, Brian, 1989: *An Introduction to the Psychology of Hearing* (London: Academic Press)

Persterer, Alexander, 1991: 'Binaural simulaton of an 'Ideal Control Room' for Headphones Reproduction' 90th AES Convention, preprint number 3062

Plenge, G, 1974: 'On the difference between localisation and lateralisation' *J. Acoustical Society of America,* Vol. 56, No.3, pp944–951

Robinson, David, and Greenfield, Richard, 1998: 'A binaural simulation which renders out-of-head localisation with low-cost digital signal processing of head related transfer functions and pseudo reverberation' *104th AES Convention*, preprint number 4723

Rubak, Per, 1991: 'Headphone signal processing system for out-of-head localisation', *90th AES Convention*, preprint number 3063

Sakamoto, Naraji; Gotoh, Toshiyuki, and Kimura, Yoichi, 1976: 'On Out-of-Head Localisation in Headphone Listening' *J. Audio Eng. Soc.* Vol.24, No. 9, pp710–716

Savioja, Lauri; Huopaniemi, Jyri; Lokki, Tapio, and Väänänen, Riitta, 1999: 'Creating Interactive Virtual Acoustic Environments' *J. Audio Eng. Soc.* Vol. 47, No. 9, pp675–705

Thomas, Martin, 1977: 'Improving the Stereo Headphone Sound Image' *J. Aud. Eng. Soc.* Vol.25, No. 7/8. pp474–478

Thurlow, Willard; Mangels, John, and Runge, Philip, 1967: 'Head Movements During Sound Localisation' *J. Acoustical Society of America*, Volume 42, p489-493

Travis, Chris: 'Virtual Reality Perspective on Headphone Audio' *100th AES Convention*, preprint number 4354

Wallach, Hans, 1939: ‘On Sound Localisation’ *J. Acoustical Society of America*, Volume 10, pp270−274

Watkinson, John, 1998: *The Art of Sound Reproduction* (Oxford: Focal Press) pp159−162, p207.

Wightman, Frederic, and Kistler, Doris, 1997: ‘Factors affecting the Relative Salience of Sound Localisation Cues’ in ed. Gilkey, Robert, and Anderson, Timothy: *Binaural and Spatial Hearing in Real and Virtual Environments* (New Jersey: Laurence Ellbaum Associates) pp1-23

General Reality, 1996: *CyberTrack II CT-3.2 Sourceless Head Tracker Developer Manual* http://www.genreality.com/Manuals/CT-II_Developer_Manual.pdf

Dolby Laboratories 1999: *Dolby Headphone*, http://www.dolby.com/headphone/

8 BIBLIOGRAPHY

Møller, Henrik, 1992: Fundamentals of binaural technology, Applied Acoustics, vol.36 (Barking: Elsevier Science Publishers Ltd)

1992: *Risc OS 3 Programmer's Reference Manual*, Vols.1–4 (Cambridge: Acorn Computers Limited)

1995: *Risc OS 3 Programmer's Reference Manual*, Vol.5a (Cambridge: Acorn Computers Limited)

A MATHEMATICAL DERIVATION OF POINT ROTATION

Yaw, pitch and roll, as described by the CyberTrack-II are commutative: they may each be applied to a point around the head in any order to achieve the same result, because the axes of rotation are also affected by each transformation.

The algorithm which is used in the simulator applies each rotation individually to the three unit vectors; this is used to determine the location of the six simulated points, firstly in Cartesian co-ordinates, and then in terms of azimuth and elevation. Rotations are applied in the program in the order presented below (§B.1–B.4), and by using the formulae described.

The following variables are used:

x, y, z stable axes: these do not change with each rotation.

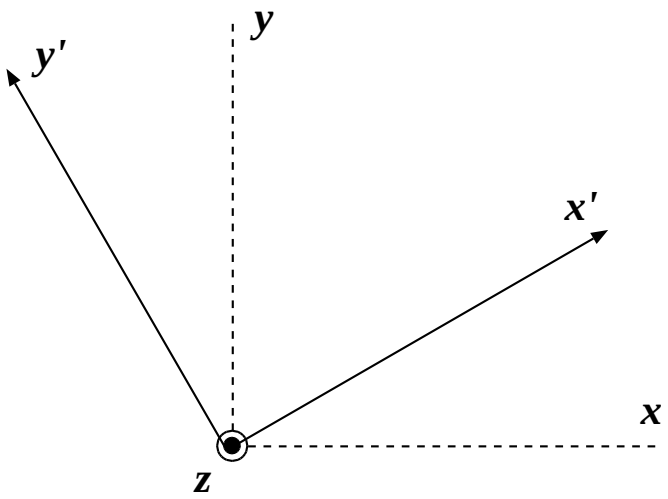
The following are all reference vectors of unit length:

x', y', z' rotating axes for the first rotation.
x'', y'', z'' rotating axes for the second rotation.
x''', y''', z''' rotating axes for the third rotation.

$x'(x)$ and similar terminology
the x-axis projection of the unit vector x'

θ angle of yaw rotation
ϕ angle of roll rotation
ψ angle of pitch rotation

A.1 ROTATION BY YAW (ROTATION)



The first rotation, around the *z*-axis, is the easiest. Because the *z*-axis remains constant, it is necessary only to rotate the *x*- and *y*-axes as if they were located in two-dimensional space. The unit vectors can be transformed using standard trigonometry:

$x'(x) = \cos\theta$

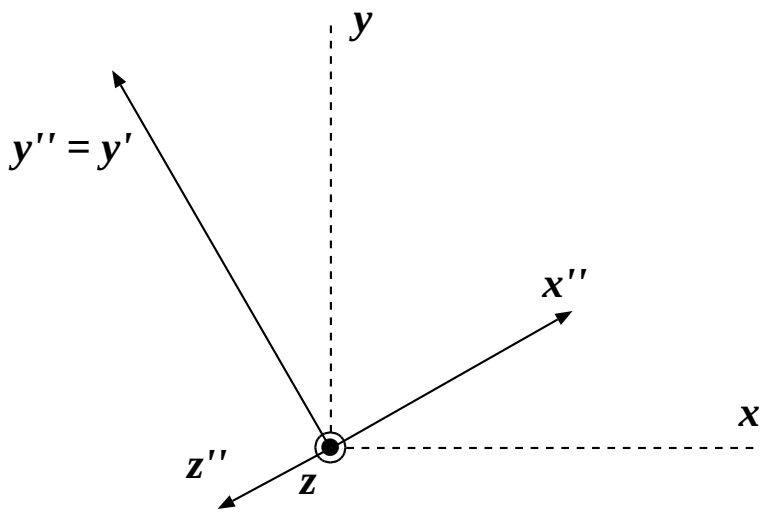
$x'(y) = \sin\theta$

$y'(x) = -\sin\theta$

$y'(y) = \cos\theta$

$x'(z) = y'(z) = z'(x) = z'(y) = 0;\ z'(z) = 1$

A.2 ROTATION BY ROLL (PIVOTING)



With roll rotation, the new *y'*-axis remains constant, and the *x'*- and *z'*- axes are rotated around it. Because the starting points are now rotated with respect to the original *x*, *y* and *z*, this is not as simple as the yaw rotation formula.

It is useful, however, to bear in mind that the projection of *x''* and *z''* (the new rotations) onto the *x*-*y* plane, as in the diagram, are exactly opposed, and that the gradient of *x''* on this projection is identical to the gradient of *x'*, so that:

$z''(x) = -x'(x) \sin \phi$

$z''(y) = -x'(y) \sin \phi$

$z''(z) = \cos \phi$

$x''(x) = x'(x) \cos \phi$

$x''(y) = x'(y) \cos \phi$

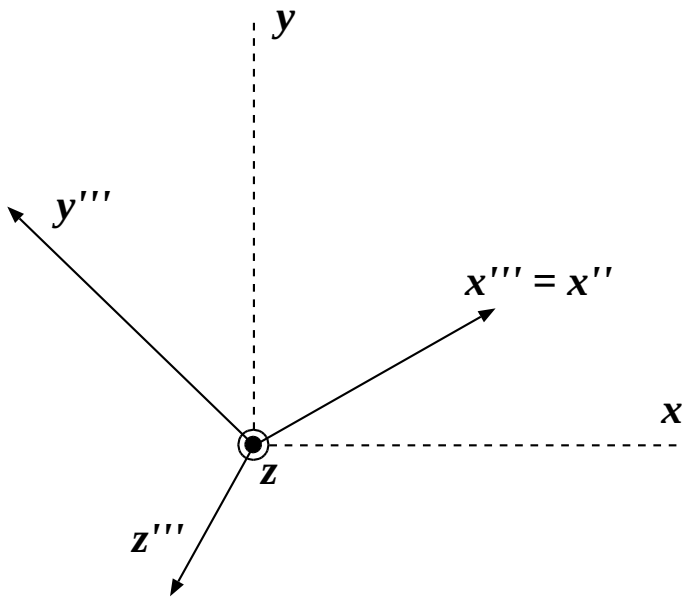
$x''(z) = \sin \phi$

$y''(x) = y'(x)$

$y''(y) = y'(y)$

$y''(z) = y'(z) = 0$

A.3 ROTATION BY PITCH (TIPPING)



In the final rotation, the *x''*-vector stays constant whilst the *y''*- and *z''*- vectors are rotated about it. The simplest way of solving this equation is to remember that the transformed axes must still lie perpendicular to the *x''*-vector, and therefore be on the *y''*-*z''* plane.

The program uses a method of rotating the point by interpolating between the *y''* and *z''* axes. This is simplified mathematically by the fact that pitch data from the head tracker varies only between ±45°.

A point linearly between the *z''* and *y''* points will have the following equation:

$$p(x) = y''(x) + k\,(z''(x) - y''(x))$$
$$p(y) = y''(x) + k\,(z''(y) - y''(y))$$
$$p(z) = y''(z) + k\,(z''(z) - y''(z))$$

k defines the distance of a point along this line, where $p = y''$ when $k = 0$, and $p = z''$ when k = 1. When the rotation is performed with positive co-ordinates, the rotated vector *z'''* will pass through a point on this line (*Fig. 21)*. The precise position where this occurs may be derived using the sine rule:

$$\sqrt{2}\,k / \sin\psi = 1 / \sin(135° - \psi)$$

$$\Rightarrow \quad k = \sin\psi / (\sqrt{2} \times \sin(135° - \psi))$$

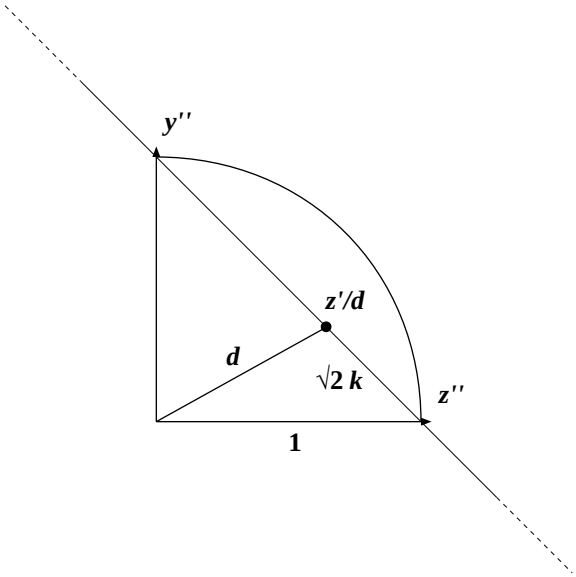


Fig. 21: Trigonometry used in pitch rotation

Now that the correct bearing for *z'''* has been determined, it must be scaled to unit magnitude. This is done, again using the sine rule:

$$z'''(x) = p(x) / d$$
$$z'''(y) = p(y) / d$$
$$z'''(z) = p(z) / d$$

$$d / \sin 45° = 1 / \sin(135 - \psi)$$

$$\Rightarrow \quad d = 1 / (\sqrt{2} \sin(135 - \psi)$$

Each point can therefore be successfully rotated: the *y''* vector can be rotated by re-defining *p(x,y,z)* in terms of *y''(x,y,z)* and − *z''(x,y,z)*.

These equations have no solutions for values greater than ±45°: at $\psi = -45°$, the derived point *z'''* runs parallel with the line between the two axes, so they will never meet. In mathematical terms, the denominator will go to zero. *y'''* is undefined for the same reason if $\psi = 45°$.

A.4 POINT ROTATION

The three reference axes have successfully been rotated through yaw, roll and pitch. It is now a relatively simple task to discover the coordinates of any point *s(x,y)*, specified as untransformed Cartesian co-ordinates, at unit distance from the origin in the horizontal plane. The point is given in terms of an azimuth angle α, and this is transformed using the *k* and *d* system described in §A.3.

In order to permit 360° specification of points, the computer algorithm divides the horizontal plane into four quadrants:

when $0° < \alpha < 90°$:

$$p(x,y,z) = x'(x,y,z) + k \times (y'(x,y,z) - x'(x,y,z))$$

$$\psi = \alpha$$

when $90° < \alpha < 180°$:

$$p(x,y,z) = y'(x,y,z) + k \times (- x'(x,y,z) - y'(x,y,z))$$

$$\psi = \alpha - 90°$$

when $-90° < \alpha < 0°$:

$$p(x,y,z) = - y'(x,y,z) + k \times (x'(x,y,z) + y'(x,y,z))$$

$$\psi = \alpha + 90°$$

when $-180° < \alpha < -90°$:

$$p(x,y,z) = - x'(x,y,z) + k \times (x'(x,y,z) - y'(x,y,z))$$

$$\psi = \alpha + 180°$$

The transformed point *s'(x,y,z)* is in Cartesian coordinates, and these can be transformed into azimuth and elevation using the following formulae:

azimuth = arctan (*s'(y)* / *s'(x)*)

elevation = arcsin (*s'(z)*)

B EXTRACTS USED IN THE LISTENING TESTS

Extract 1: Maurice Ravel
from ‘Le Tombeau de Couperin’
II. Forlane
Deutsche Grammophon Classikon 439 414-2, 1986

Well-defined chamber orchestral recording of a short movement, mostly quiet dynamics.

Extract 2: Mahavishnu Orchestra / John McLaughlin
‘Inner Worlds’ album
Track 8: Planetary Citizen
Columbia, COL 476905 2, 1975, 1994

1970s Jazz-Funk fusion. Artificial plate reverb, closely-recorded and and extremely-panned electric guitars, dry drum mix.

Extract 3: Squarepusher
‘Burningn’n Tree’ album
Track 1
Warp Records WARPCD 53/SPY 7, 1997

Drum ‘n’ bass and Jazz. Electroacoustic and synthesised instruments; sampled drum track. Moderately flat spectrum and a very compressed, very close sound. Extremely narrow stereo image.

Extract 4: Steve Reich
from ‘Electric Counterpoint’
III. Fast
Elektra/Nonesuch 7559-79176-2, 1989

Six electric guitars and two electric bass guitars. No added distortion. Closely-miked amplifiers with artificial panning, and a general superimposed hall-type reverberation.

Extract 5: Hamilton Harty
'In Ireland'
Performed by Michael Jefferies and Chris Warner
Portfolio recording by Ben Supper
University of Surrey, 2000

Flute and piano recording of a piece of Twentieth Century classical music. Hypercardioid cross-pair and two spaced outriggers, well-defined stereo image, no artificial reverberation, but a large amount of natural ambience.

C THE LISTENING TEST PAPER