Artificial Simulation of Audio Spatialisation: Developing a Binaural System

Brian Carty

Overview

Sound localisation deals with how and why we can locate sound sources in our spatial environment. *Sound spatialisation* defines how sound is distributed in this environment. Several acoustic and psychoacoustic phenomena are involved in sound localisation and spatialisation. The importance of these phenomena becomes apparent when endeavouring to recreate and emulate auditory spatial events using computers.

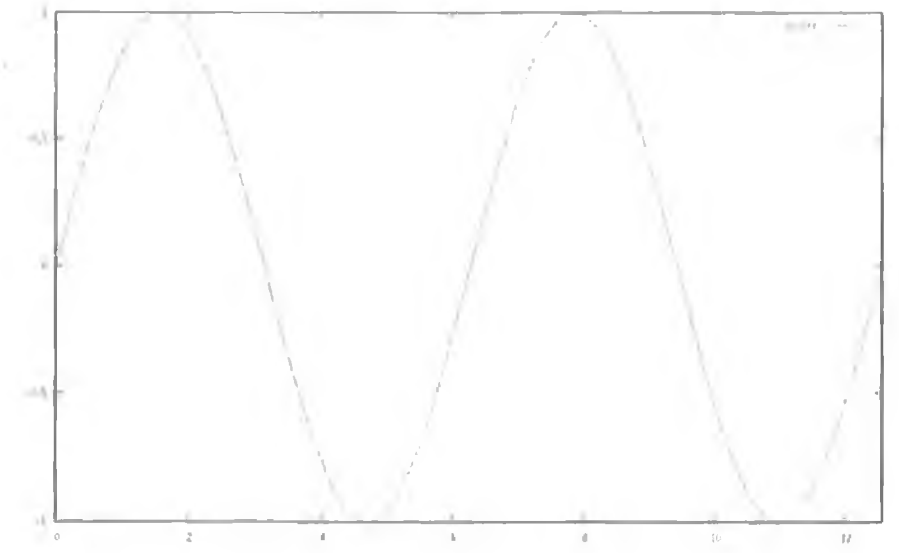
Several systems for auditory spatialisation have been suggested, many of which have provided relatively successful results. This article focuses on binaural processing systems, which aim to accurately model how sound from a particular source location is perceived by our auditory system. Head Related Transfer Functions measure how an arbitrary sound is modified from source to tympanic membrane by the head, outer ear and torso, and are generally utilised in binaural systems. These complex functions can provide the key element of very accurate spatialisation systems, the design of which raises several issues including efficient function measurement, representation and interpolation. An overview of the spatial auditory system is presented, followed by a brief summary of methods currently and traditionally employed to artificially spatialise audio.

The development of a binaural solution to artificial simulation of audio spatialisation is then discussed, with specific mention of some of the major challenges involved. Current solutions to these challenges, implemented in the Csound and Pure Data Computer Music systems are then outlined and reviewed. Finally possible improvements are suggested. The article aims to simplify the topic, addressing it from the point of view of an audience who may not be familiar with the complex digital signal processing algorithms and acoustical and psychoacoustical phenomena involved.

1. Introduction

Sound travels in waves created when a vibrating source disturbs the air. The vibrating source determines the varying characteristics of these waves. The parameters of a simple harmonic motion, for example, result in a sinusoidal/sine wave. The characteristics of this wave include amplitude, frequency (which relates to pitch and can be described as the number of cycles in one given time period or Hertz (Hz)) and phase (the starting point of the wave with reference to a full cycle). Figure 1 illustrates a simple sine wave.

Figure 1. A simple sine wave



Most 'real world' sounds are not this straightforward (sounds like struck tuning forks approach this level of simplicity), and are made up of combinations of simple periodic sounds, with different frequencies, amplitudes and phases. These components can be modelled accurately by sinusoidal functions/sine waves. Figure 1 shows a simple sine wave, with amplitude varying from +1 to -1 and no phase offset. Two cycles of the wave are shown, so if we take one second to play this sound, it will have a frequency of 2 cycles per second/2 Hz. The view of 'real world' sounds as complex signals which can be broken down into simple sine waves with different parameters plays an important role in the area of binaural

artificial sound source spatialisation and indeed many areas of computer music.

Sound waves interact with each other, and the environments in which they exist, in many ways, several of which are relevant to sound localisation. Upon reaching the ears, different component frequencies of the sound wave (which constitute its *spectrum,* and can be thought of as the sinusoidal components mentioned above) trigger different areas of the Basilar Membrane (located in the inner ear), which consequently transmits electric signals to the brain (nerve firings) to be perceived as sound.

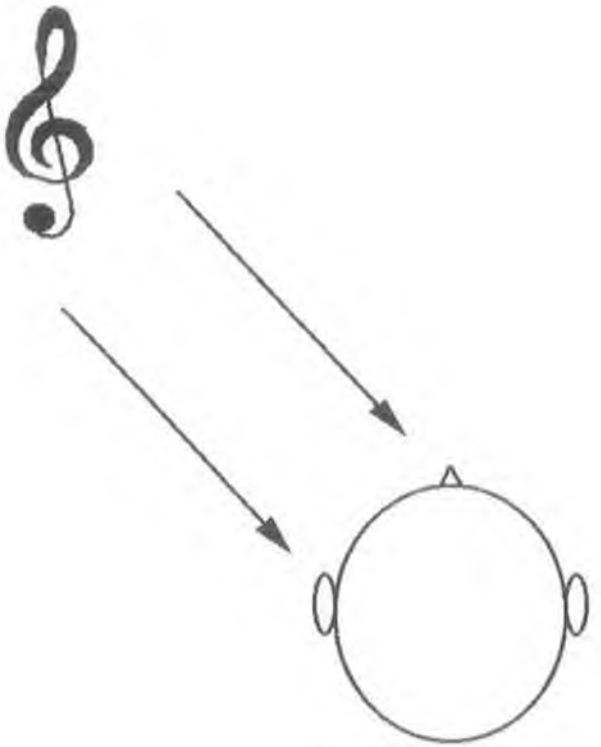
2. Sound Localisation

2.1 Interaural Time Difference

Binaural hearing is the term given to listening with two ears rather than one, and is the main factor involved in sound localisation. The fact that the brain receives an independent signal from each ear allows conclusions to be drawn based on a comparison of the characteristics of each signal.

One such binaural indication of a sound's spatial characteristics is interaural time difference (itd): the name given to the time it takes a sound to reach one ear after it has first reached the other. As sound travels at a determinable speed, and speed can be defined as distance travelled per unit time, it follows that if a sound source is further from one ear than the other, a delay will occur in the time it takes the sound to reach the further ear
(see Figure 2).

Figure 2. Itd: a source to the left of the listener will take longer to reach the right ear.



Also, the further the sound source is from the lateral centre of the listener's space, the greater this delay will be. Despite the relatively tiny nature of these time disparities, they can provide very accurate localisation cues.

Notwithstanding its accuracy (in favourable conditions), this method of localisation based on interaural time delays is heavily dependent on frequency. Ultimately, localisation using itd breaks down above 1500 Hertz.[1] This limitation is dependent on the characteristics of the source sound however, and itd can play an important role at higher frequencies in some circumstances.

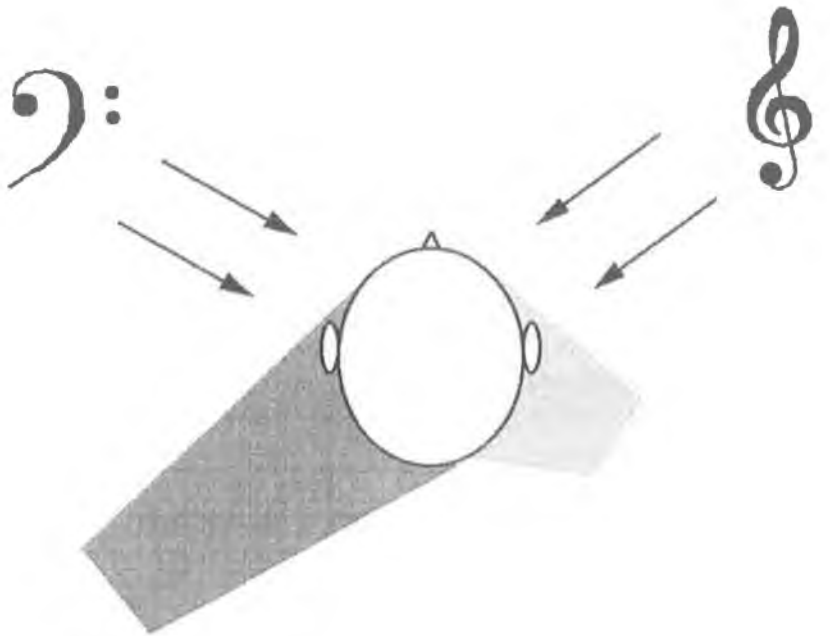
[1] Brian C.J Moore: *An Introduction to the Psychology of Hearing* (London: Elsevier Academic Press, 1977; 5th edn, 2004), 237. Hereafter referred to as Moore: *An Introduction to the Psychology of Hearing*.

2.2 Interaural Intensity Difference

Interaural *intensity* difference (iid) uses varying respective intensities of a signal at each ear to locate source sounds. Interaural intensity difference is based principally on the head (and to a lesser extent the torso) acting as a barrier to sound. Fundamentally, the head obstructs a sound coming from a particular angle off lateral centre, decreasing its intensity at the further ear. This 'acoustic shadow' is analogous to a shadow cast by an obstruction to light.

Again, the system is frequency specific. Obstacles block the path of high frequencies, whereas low frequencies tend to diffract around them. These low frequencies have long wavelengths (the wavelength of a sinusoidal sound can be thought of as the distance it takes to travel one period or cycle of the waveform, and is related to the speed of sound and the frequency of the wave) relative to the diameter of the head and can bend around it, enveloping it within one of their periods. Figure 3 illustrates, using shaded regions, how the head affects sound sources with different frequency content.

Figure 3. Frequency dependence of iid: high frequency sources (represented here by a treble clef) are more affected by the shadowing effect of the head than lower frequencies (represented by a bass clef).



Small wavelengths (i.e. high frequencies), conversely, cannot envelop the listener, as their periods are much shorter. The head consequently acts as a low pass filter, significantly diminishing high frequencies.

Interestingly, intensity differences are used for panning functions on most audio processors, and it is this relative lack of intensity based directionality of low frequencies that leads to bass bin speakers: a single channel dedicated to low frequency signal content, as stereo or surround intensity based effects are superfluous when acting upon these low frequencies.

It is generally accepted that interaural time and intensity differences work together to provide a well-defined spatial image, with itd working best for low frequencies and iid for high. However, it is important to remember that we also rely on information from other

senses for spatial localisation.[2] For example, despite all the intricacies involved in the interaural systems, the listener, in real world circumstances, tends to angle their head towards the source to better localise it, essentially to remove interaural variations.

2.3 Monaural Considerations

It appears from all the above processes that localisation is essentially a binaural system; however, monaural information (independent information from one ear) also plays an important role. The pinna and concha both have a non-linear frequency response over the audible spectrum, altering incoming sounds. Due to the non-uniform shape of the pinna, these alterations vary with sound location. The various folds of the pinna cause tiny time delays, which filter the source sound. So, after the head and the torso, the pinna and concha perform more filtering of the sound, which, after being further transformed by the resonances of the auditory canal, will eventually reach the eardrum.

Interaural time or intensity differences will be the same at the front as those at the same area to the back of the listener. The back of the pinna filters out the high frequency content of the sound; making sounds from the back appear much duller.

The pinna's interaction with the sound is also the main factor involved in localisation in the median plane (where interaural differences are again not helpful). Different source 'elevation values' cause different interactions with the intricate folds of the pinna.

3. Approaches to Artificial Simulation of Audio Spatialisation

The main difficulty in designing a system for artificial simulation of audio spatialisation is that the designer is generally attempting to recreate soundscapes defined in three dimensions with a discrete number of sources. Sound, in real world situations, reaches the listener from all locations. Typically, in an artificial situation, a discrete number of sound transmission devices (e.g. loudspeakers) are used.

The intricacies, various systems and system interactions and limitations (for example, the frequency dependence of itd and iid) of the auditory spatialisation system pose various other challenges during the design process. Several approaches to artificial simulation of audio

[2] Moore: *An Introduction to the Psychology of Hearing*, 264.

spatialisation have been suggested, many of which exploit phenomena mentioned in Section 2. Some approaches utilise only one spatial cue, for example intensity panning relies on interaural intensity difference alone. Most only attempt localisation in the lateral plane (in front, behind and to the sides of the listener), omitting the medial plane (not allowing for any auditory information pertaining to the perceived elevation of the sound relative to the listener). Other approaches, such as binaural systems, attempt to provide a much more complete solution. The main approaches are discussed briefly below, followed by a more detailed look at binaural system development.

3.1 'Traditional' Two Channel Systems

The intensity panning system is based on interaural intensity difference and is typical of modern music production techniques. A mono signal can be sent to the right and left channel of a stereo system at different intensity levels depending on the desired source location. If reproduced on loudspeakers, the placement of the speakers is a key factor in creating any spatial illusion. The most effective configuration is generally accepted to describe an equilateral triangle, with the speakers occupying the base points. The optimum listener position (or *sweet spot*) is just behind the point that completes this triangle. Any two channel stereo effects are best appreciated when the listener is as near as possible to this point, and deteriorate progressively as he/she moves away. Recording techniques exist to also include itd in the stereo signal, which can add some clarity to the spatial image.

3.2 'Traditional' Multi Channel Systems

The two channel system mentioned above can be extended to quadraphonic and octophonic systems, which employ four and eight channels respectively. The current 5.1 cinema and home entertainment standard uses comparable techniques, and exhibits similar deficiencies. This setup uses a central, front left, front right, surround left, surround right and low frequency speaker to generate the sound field.

Essentially, these systems are all limited to a small 'sweet spot' for optimal listening. Although accurate recording techniques can be used to spatialise sounds quite well, 'virtual' sound source locations (locations in between loudspeakers) often appear spurious. The three front speakers in a 5.1 setup can provide good localisation in front of listeners, but the

two rear speakers, often duplicated in large cinemas, are generally only used to add ambiance, as they simply cannot provide stable spatial auditory scenes for the whole audience.

Although a well mixed 5.1 film can be an immersive and enjoyable auditory event, the standard is incomplete as a spatial auditory solution. Take, for example, the case whereby a listener is sitting at the rear of a large cinema. In the case of a large cinema, it is typical to duplicate the surround speakers. The listener may now find that he/she is sitting behind duplicated surround channels, completely distorting the spatial image! Also, the above systems can only provide localised sound in the lateral plane (they lack any sense of source elevation).

3.3 More Complete Approaches

3.3.1 Ambisonics

Ambisonics is a unified sound system (with standards for recording, distribution and playback), which results in a level of spatial accuracy that is proportional to the amount of channels employed. It works by capturing the left/right, front/back, and up/down velocity components of the vibration in the air caused by the source(/s), as well as the overall pressure level at the recording point. Shelf filtering is then imposed on the ambisonic signal, increasing the psychoacoustical accuracy of the system. It is only at this point that the number of channels used in reproduction is considered. The signal sent to each channel is derived by passing the outputs from the filters through a complex amplitude matrix. All speakers work together to accurately reproduce the spatial attributes of the source sound. Ambisonics provides convincing spatialisation with a larger sweet spot than more prevalent multi channel approaches, and is based on acoustic and psychoacoustic phenomena.

3.3.2 Wave Field Synthesis

Wave Field Synthesis is a multi-channel approach to sound spatialisation, which uses an array of speakers to recreate the waveform of a sound at a desired location. This technique is based on Huygens' Principle, which states that if every point on a propagating wave serves as the source of a spherical secondary wavelet, then the wave at an arbitrary later time can be described by these wavelets (which have the same speed and frequency as the original wave) at that time. Essentially, several loudspeakers are used to recreate the actual waveform of the desired soundscape using

their own output waveforms, which are designed to sum to the desired spatial event.

Wave Field Synthesis is not limited to a small sweet spot, as the waveform being omitted will appear to be located in the same position in space irrespective of the listener's position. So, for example, in a cinema setting, a source can be virtually localised in (or outside) the room and will appear to be at the same location for all listeners. Wave Field Synthesis does, however, typically require numerous loudspeakers/loudspeaker arrays in practice, and can be thought of as a 'wall of speakers' approach.

3.3.3 Binaural Systems

Binaural techniques aim to model more accurately how a source sound from a particular location is perceived at our ears. Binaural recording techniques involve recreating realistic listening circumstances (sound recorded as it would be heard by a listener, with a dummy head with microphones placed in the ears, for example), with a view to incorporating all sound localisation cues mentioned in Section 2. This clearly has the potential to simulate more accurate spatial images than techniques employing just one binaural phenomenon (for example intensity panning). The remainder of this article will discuss some of the many issues to be considered in designing an artificial binaural spatialisation system.

4 Binaural Systems in More Detail

4.1 Head Related Transfer Functions

Head Related Transfer Functions (HRTFs) are essentially functions that describe how a sound from a specific location is altered from source to tympanic membrane. Binaural recording techniques can capture these functions. For any particular source sound, a pair of transfer functions exist (for the left and right ear) for any location relative to the listener. These functions will encompass all localisation cues mentioned in Section 2.

It is also possible to obtain generalised versions of these functions, based on the assumption from Section 1 that complex sounds can be broken down into sinusoidal sounds of different frequencies and amplitudes. If we record how a system (for example the left or right ear) responds to a signal containing all (audible) frequencies at equal

amplitudes, we know how the system will respond to any sound. This frequency rich sound is known as the *impulse function*, and the result of passing it through a system constitutes the systems *impulse response*. Head Related Transfer Functions, then, can be defined as the impulse responses of the left and right ears to sound at a particular location.

4.2 HRTF Based Localisation

The process of simulating an auditory location using HRTFs can be summarised thus:

Record the response of the left and right ear to the impulse function, to learn how each ear will treat sound at all frequencies for the desired point in space.

Analyse the frequency content of the sound you wish to spatialise (calculate what frequencies/simple sinusoidal waves are present in the input sound). Impose the HRTF for the left and right ears on the sound (boost or attenuate and delay the frequencies contained in the input in accordance with how the ear treats the appropriate frequencies).

To reiterate: find out how the ears treat all frequencies for your desired location, and treat the frequencies contained in your input sound in the same way. This is done using a process called *convolution*, which imposes the characteristics of the impulse response onto another signal. A typical example of convolution is a reverberation application. The impulse response of a room can be captured in a similar way to the impulse response of the left or right ear. When convolved with an arbitrary signal, the room impulse response imposes the reverberant characteristics of the room onto the signal. Similarly, the impulse responses of the left and right ear to an impulse originating from a specific spatial location imposed on an input signal alters the signal such that it appears to originate from the same place. The input signal is typically a mono, non-localised signal, for example a recorded piece of speech.

It is important to note that binaurally generated two channel (left ear and right ear) signals should be reproduced on headphones. Loudspeaker reproduction significantly diminishes spatialisation accuracy, which is intuitive for several reasons:

If the signals are played back on headphones, the HRTF derived left/right ear signal can be sent to the left/right headphone speaker exclusively. However, if relayed through two loudspeakers, the signal

from *both* speakers will reach *both* ears i.e. the right ear will not only hear the right ear HRTF processed signal, but also, shortly afterwards (assuming an optimal two speaker setup, as described in Section 3.1) the left ear HRTF will be heard. Loudspeaker playback also introduces environmental sonic artefacts, such as the reverberation effects of the room. Finally, as the effects of the pinna have already been considered by the HRTF processing, loudspeaker playback reintroduces the pinna, thus pinna filtering (and any other filtering between source and ear canal) will occur again!

4.3 Towards a Complete Binaural HRTF Based Audio Spatialisation System

The theory of HRTF based filtering of mono signals to artificially recreate spatial locations has been presented above. Several other factors have to be considered in a complete system, however, some of which are outlined below.

4.3.1 Cognitive Issues

Perhaps the most pertinent of these issues is the inherent limitations of artificial spatialisation. For example, the listener must be familiar with the original sound spectrum to make any judgement on how the pinna is filtering it due to its spatial location. Even in the binaural system, although spectra at the two ears will be different, knowledge of the inherent spectrum of the sound is necessary, in the median plane, for example, where both ears will receive almost identical signals. If the source is moving, comparisons can indeed be made with preceding spectral information to further define its spatial image, however, it appears that only slow changes in location can be successfully followed.[3]

4.3.2 HRTF Measurements

The process of HRTF localisation outlined in Section 4.2 describes the localisation of a source sound to one specific area of space. When other locations are required, however, the relevant HRTF data is needed. The measurement process is a cumbersome and timely one, and can be intrusive if using human subjects (with small microphones in their auditory canals) as opposed to mannequins. Therefore a fixed amount of

[3] Moore: *An Introduction to the Psychology of Hearing*, 261.

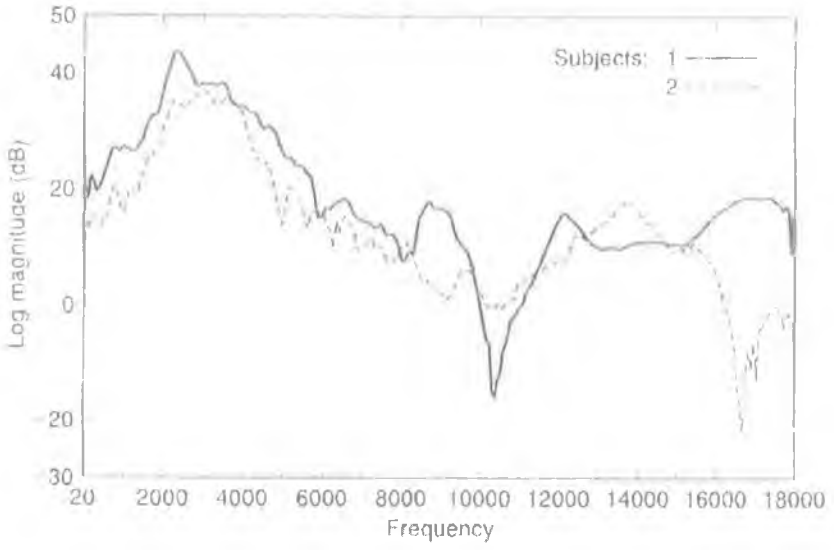
points are typically recorded and stored. However, if a location is required that has not been measured, or if a sound is required to move smoothly from one location to another, some kind of averaging or interpolation must be attempted (see Section 4.3.4).

4.3.3 The Individual Nature of HRTFs

As the physiology of everyone's ears is different, HRTFs vary considerably from subject to subject. The complex folds and structures on each pinna, and varying torso shapes give rise to different transformations of sounds as they travel from source to eardrum. An example of differences in individual HRTF measurements is given in Figure 4, which illustrates the HRTFs of two individuals when the source is at 0 degrees elevation and centred. This is a frequency domain representation, which is described in more detail in Section 4.3.4.1, and can be thought of as simply illustrating the combination of simple sine waves in the represented sound.

Ideally, listeners should use binaural systems customised to their ears. However, certain consistencies can be observed and generalised/*non-individualised* HRTFs, recorded with a dummy head and torso model or a specific subject, are frequently used to remove the necessity for measurements to be taken for each individual user.

Figure 4. The individual nature of HRTFs[4]



It is worth bearing in mind that when using a non-individualised data set, a user is essentially listening through a different set of ears i.e. the ears of the mannequin/person used in the data recording process. This leads to some difficulties with localisation, specifically increasing *reversals*: errors in front/back and up/down judgements. If a source is 30 degrees behind a listener to the right (150 degrees off centre), for example, the listener may perceive the source to be 30 degrees in front (60 degrees off centre). This implies that some of the complex monaural cues to localisation caused mainly by the non linear frequency response of the pinna are misrepresented. Perhaps, with exposure, a user may get accustomed to listening with these new ears and these errors may be reduced. Results from Wenzel's *et al.* study of non-individualised HRTFs suggest that although non-individualised HRTFs are certainly a useful tool for binaural simulation, they result in a distortion of the spectral

[4] From F. Rumsey and T. McCormick.: *Sound and Recording: An Introduction* (Oxford: Focal, 1992; 4th edn, 2002), 38.

characteristics used in front/back and elevation resolution when compared to listening in the free field.[5]

4.3.4 HRTF Interpolation

4.3.4.1 The Frequency Domain

The topic of HRTF interpolation is a complex and multi faceted one. Much work has been done in the area, particularly in the relatively new discipline of digital signal processing.

The representation of sound illustrated in Figure 1 can be thought of as a *time domain* representation. The sine wave shows how the vibrating source changes air pressure at the recording point over time. This is a perfectly intuitive way to represent sound. However, we can also represent sound as the sum of simple sinusoidal waves, as discussed in Section 1. We can use the *frequency domain* to represent this sum, and can think of frequency domain representation as showing the *frequency content* of a sound: the frequencies of the simple sinusoids present in the sound, their relative amplitudes and phases. Figure 4 shows a frequency domain representation, illustrating the levels of the component frequencies in the signal. We are now dealing with frequency values on the x axis, and level/magnitude/amplitude values on the y axis: how much of each sinusoidal frequency makes up this sound. It is also important to note that as we are no longer in the time domain, this frequency domain representation can be thought of as a *snapshot* of the sound to be represented. The representation changes over frequency instead of time.

The mathematical tool used to generate this representation is called the Fourier Transform, and is very useful in the case of binaural processing, as we are dealing with the intricate frequency alterations made by the HRTF of each ear.

4.3.4.2 Minimum Phase

HRTF interpolation can be thought of as taking the two (or more for increased accuracy) nearest HRTF representations to a non-measured point in between, and deriving a new HRTF by averaging the known values with greater relative weighting(/s) on the nearer known point(/s).

---

[5] E.M. Wenzel, M. Arruda, D.J. Kistler and F.L. Wightman: 'Localization using Non-individualized Head Related Transfer Functions' *Journal of the Acoustical Society of America* 94/1 (7/1993), 111–123.

Interpolation can be performed in the time or frequency domain. Essentially, the difference between time and frequency domain interpolation techniques is that time domain methods typically calculate a new 'in between' value using actual *sample* (the smallest unit of digital audio data) values, as opposed to magnitude/level values of each simple sinusoidal frequency in the two (or more) nearest measured HRTFs typically used in the frequency domain. Frequency domain interpolation can give more accurate results.[6] However, there is one major factor still to be considered.

Two of the three main parameters of the simple sine waves discussed in Section 1 are accounted for in magnitude interpolation: frequency and amplitude. However, phase also plays an important role. Phase values are closely related to itd, so are important in the localisation process. The linear interpolation of phase values, which are calculated in the Fourier Transform process, is, however, flawed. Phase, unlike magnitude, is a periodic quantity, measured in fractions of a full cycle. Uncertainty arises when trying to interpolate phases, as a phase value can be +/- any amount of full cycles. For example, if trying to derive a point equidistant from two points with phase 10 and 50 degrees respectively, the phase appears to be 30 degrees. However, the 50 degree point may also be a full cycle ahead of the 10 degree point, therefore representing a cumulative phase of 410 degrees, leading to an interpolated phase of 210 degrees. The significance of phase information and the auditory systems limitations in responding to changes in phase information thus comes into question, and has been investigated in depth.[7]

HRTFs have also been scrutinised more closely to determine whether they can be further simplified and how the auditory system responds to such simplifications. Results appear to point to the fact that HRTFs approach a property called minimum phase, which essentially allows a pair of HRTFs (for the left and right ears) to be broken down into three parts: a minimum phase representation of each empirical HRTF pair (left and right ear), and an interaural delay. The key to these

---

[6] K. Hartung, J. Braasch and S.J. Sterbing: 'Comparison of Different Methods for the Interpolation of Head Related Transfer Functions' *Audio Engineering Society 16th International Conference: Spatial Sound Reproduction* (3/1999), 319–329.

[7] For example, A. Kulkarni, S.K. Isabelle and H.S. Colburn: 'Sensitivity of Human Subjects to Head-Related Transfer-Function Phase Spectra' *Journal of the Acoustical Society of America* 105/5 (5/1999), 2821–2840.

minimum phase representations is that phase values for each component frequency can be derived from the corresponding magnitude values. This is a unique and, in this case, extremely useful property of minimum phase systems.

The process of interpolation thus involves analysing each HRTF pair to find the relevant interaural delay and reducing them to minimum phase representations. The minimum phase values and delay can then be linearly interpolated. The new HRTF pair is thus defined as two minimum phase functions and an interaural delay.

This method employs complex digital signal processing, is quite computationally expensive, and involves approximations that can introduce errors. Other, less complex solutions may be considered bearing in mind the limitations of the auditory system and the inherent difficulties involved with all binaural systems. Specifically, the author is currently developing two magnitude interpolation systems, one using phase truncation, the other a geometric phase model with frequency dependent scaling.

5. Current Computer Music Solutions to HRTF Based Binaural Processing

5.1 The Csound HRTFER Opcode

5.1.1 Csound

Csound is a text based audio processing environment. Users can generate complex audio events by creating simple, often very short text files. Csound is open source, maintained fervently by a group of competent developers, and is used worldwide as a stable, valuable audio tool.

The core of the system is based on opcodes: processing units designed to perform specific tasks. For example, the *oscil* opcode generates an oscillating signal. As discussed in Section 1, several of these oscil opcodes, generating oscillating sinusoidal signals can be combined to create more complex sounds, as 'real world' sounds can be thought of as combinations of sinusoidal sound waves. The name oscil comes from the fact that sound is created by a vibrating/*oscil*lating source. The oscil will have parameters, as mentioned in Section 1, including amplitude, frequency and phase. Thus the Csound opcode oscil accurately recreates how a vibrating source generates sound. The final parameter this opcode can take is a data table which is essentially used to define the way the

source oscillates (a sinusoidal oscillation for a simple output, a more complex and irregular oscillation for a more intricate output).

Thus, the text

```
aout    oscil   10000, 440,  1,      0
```

will be understood by Csound as an oscillator with an amplitude value of 10,000, a frequency value of 440, reading its oscillation pattern from data table number 1 and with initial phase offset of 0, outputting as aout.

Various combinations of these opcodes can be used to create instruments. Csound uses the musical analogy of an orchestra of instruments playing a score. So the Csound instruments are defined in the orchestra part of the text file used to control Csound, and the score part of the text file 'plays' these instruments.
For example if the oscillator above constituted instrument number 1, the score text statement

```
i 1 0 2
```

will be understood by Csound as a command to play instrument number 1, start at zero seconds and play for 2 seconds.

The user has complete control over the instruments used and how they are played in the score. To give a complete example: perhaps a user has acquired data describing the combination of sinusoidal waves required to synthesise a brass instrument. In Csound, the required number of oscil opcodes, reading from sinusoidal oscillator data tables, with the required individual amplitude, frequency and phase values could constitute an instrument. This instrument can be played in the score, and with a little extra text, the score can also control the parameters (amplitude etc.) of the individual components or the entire instrument output. Perhaps the user may want the sound to have a natural rise in amplitude, then a brief decay, followed by a sustained period and concluding with a release/fadeout. In this case, a *linen* opcode would be useful. In this way, the user can create original instruments of a given complexity to suit their needs.

There are various ways to render audio using Csound. The package comes with graphical interfaces to open the text file(/s) where

the instruments and score are defined, or the user can simply run Csound from the command line. Audio can be rendered in real time, or to a sound file.

Although there are in excess of 1300 opcodes available to the user, a simple opcode development API is provided, allowing developers to add their own opcodes through the C/C++ programming languages. The above is meant as a very brief introduction to the functionality and capabilities of Csound as a computer music language.[8]

5.1.2 HRTFer

Csound provides an opcode to spatialise a mono input sound using binaural HRTF based techniques called HRTFer. The opcode uses a generalised set of HRTF measurements and convolves the input with the desired location in space.

The HRTFer opcode uses a comprehensive set of dummy head HRTFs available from the MIT Media Lab.[9] A KEMAR (Knowles Electronics Mannequin for Acoustics Research) dummy head and torso was used to take the measurements. This dummy represents 'a median individual in the human adult population'[10] based on meticulous measurement and design.

A carefully prepared set of 128 point/sample HRTFs was derived from the initial measurements. The high degree of accuracy required necessitates painstaking measurement and data preparation procedures. For example, the non linear frequency response of the speaker used to play the impulse must be considered, as well as other similar details.

It is important to appreciate that this data constitutes a non-individualised HRTF data set, and so using it may result in some slight distortion of localisation ability (see Section 4.3.3), but having a reliable, accurate data set allows a user to achieve instant, relatively accurate results without the difficulties of measuring their own set of HRTFs. HRTFer uses this data set to spatialise the desired source sound. As

---

[8] For more information on using Csound, see www.csounds.com.

[9] W.G. Gardner and K.D. Martin: 'HRTF Measurements of a KEMAR Dummy Head Microphone' (Massachusetts Institute of Technology: 1994) (http://sound.media.mit.edu/KEMAR.html, accessed 7/ 2007).

[10] M.D. Burkhard and R.M. Sachs: 'Anthropometric Manikin for Acoustic Research' *Journal of the Acoustical Society of America* 58/1 (7/1975), 214–222.

Csound is open source software, one can get an insight into how this is achieved by examining the source code for the opcode.

The opcode reads audio in from a file or real time input and convolves it, block by block (a block here representing a fixed number of samples), with the relevant HRTF data. The data, described above, is stored in one large file, HRTF by HRTF. This data must be carefully ordered and read accordingly. When a specific HRTF is required, it is read, transformed into the frequency domain and convolved with the Fourier representation of the current input block.

Csound HRTFer code:

```
aleft, aright HRTFer ainput, angle, elevation, "HRTFcompact"
```

The opcode has two outputs, aleft and aright, takes input audio, an angle value, an elevation value (which can both change to allow dynamic sources) and the data file name as its parameters.

This system provides accurate spatialisation for static locations which correspond exactly to HRTF measured points. However, if a static point is required that has not been measured, the system simply chooses the nearest point. This obviously leads to inaccuracies. However, considering the density of this data set these inaccuracies may be tolerable for certain situations. For central elevations, there is a measurement for every five degree increment. As the data set approaches the vertical extremities of the measured space, the HRTFs become more sparse, leading to greater inaccuracies.

This approach causes more significant errors when a specific trajectory is desired for a source. A dynamic rather than static source will skip from one nearest measured point to the next along the user defined trajectory. This staggered movement causes irregularities in the output, manifesting themselves as 'clicks', an undesirable result. The original authors of this opcode suggested a fade out of the old convolution result and a fade in of the new to reduce this noise. This addition does reduce the severity of this noise. However, in the most recent version of Csound (5.06) these crossfades have been disabled as they cause dropouts in the output, leading to worse irregularities, which are assumed to be caused by an error in the source code. In tests performed by the author these crossfades, when implemented, reduce the irregularities to a degree

depending on the frequency content of the source. For example, the irregularities are still audible in a source with a relatively small number of sinusoidal components, whereas a richer source can mask the irregularities more successfully.

Another consequence of abruptly changing these complex filters (HRTFs) as a source travels along a defined trajectory is the sudden perceptual change in the output, which can be detrimental even in frequency rich sources. For example, in a trajectory going from 50 degrees above the listener to directly in front the source will appear to jump downwards every 10 degrees, as this is the measurement increment. Clearly, this opcode could benefit from the addition of interpolation between measured points.

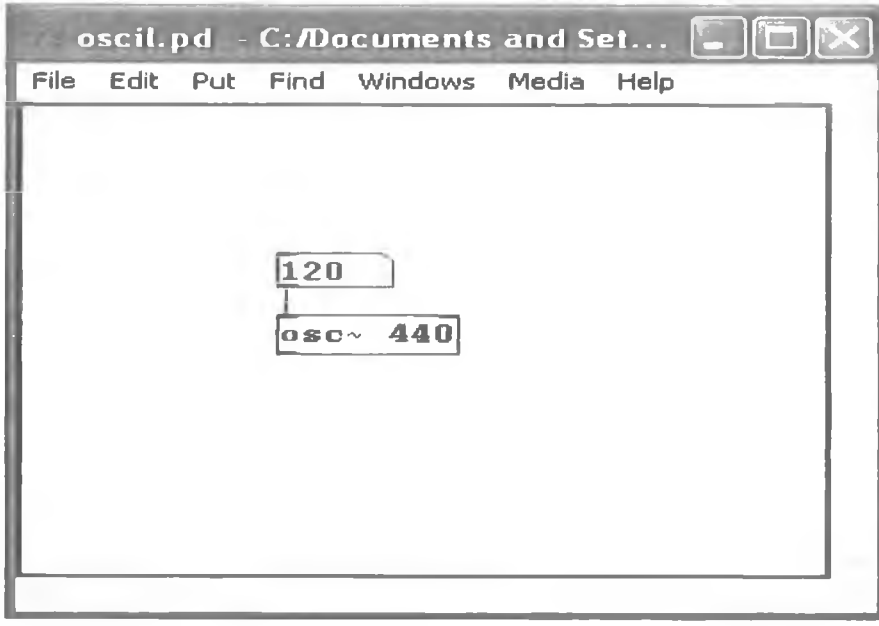
5.2 The Pure Data Earplug~ External

5.2.1 Pure Data

Pure Data (Pd), like Csound, is a programming environment for audio. However, unlike Csound, Pd is a graphical language. Users work with a *canvas/patch* instead of a text file. The core of the language consists of object boxes, much like Csound's opcodes. These objects can be connected using patchchords, which also connect parameters to objects. To use the same example as the introduction to Csound above, the patch illustrated in Figure 5 shows an osc~ object, which creates an oscillating signal. The frequency of this signal is initialised by the number beside the name of the object, but can be altered using a number box attached to the left *inlet* of the object using a patchchord. The right *inlet* can be used to reset the phase. Pd has two modes, edit mode, whereby the user can add/alter objects/connections, and run mode, whereby, for example, the user can change the value in number boxes. Much like Csound, users can extend Pd by creating new objects in the C programming language as externals.[11]

[11] For more information on Pd, see http://crca.ucsd.edu/~msp/Pd documentation/index.htm.

Figure 5. A Simple Pure Data Patch



5.2.2 Earplug~

One such external is earplug~, which uses a similar, if more advanced approach to binaural HRTF localisation. As with Csound, Pd is open source software and so the source code can be viewed. However, as earplug~ is an external, it is not part of the basic download. Xiang *et al.*[12] describe the processing details of the object and provide the object and its sourcecode for download.

Earplug~ also uses HRTF data from the MIT dataset. It adds an interpolation algorithm to HRTFer's process. The four nearest measured HRTFs to the desired point are read, and a fractional amount of each (calculated relative to the desired point) is added for each sample in the HRTF to derive a new interpolated HRTF. This new HRTF is created for each processing block. Furthermore, for each block, the previous

[12] P. Xiang, D. Camargo and M. Puckette: 'Experiments on Spatial Gestures in Binaural Sound Display' *11th International Conference on Auditory Display* (7/2005), 1–4.

interpolated HRTF is stored, and a similar interpolation is done between the current and previous HRTF. Essentially, the new value is faded in and the old faded out over the course of the processing block. This algorithm is performed in the time domain. The potential for inaccuracies in HRTFs resulting from time domain interpolation has been noted by Wenzel *et al.*, who also suggests that minimum phase representation (not implemented in earplug~) or a sufficiently dense data set may remove/minimise these errors.[13] Hartung *et al.* report that frequency as opposed to time domain interpolation provides more successful results in experiments involving more accurate interpolation methods than the four–point linear interpolation used here, suggesting that frequency domain interpolation is superior.[14] In this study, HRTF data was also prepared in a similar manner to a minimum phase representation (essentially time shifted to avoid the inaccuracies described by Wenzel), which is, again, not implemented in earplug~.

Performing the convolution operation in the time domain is also significantly slower than using an optimised algorithm to transform both the input and the HRTF into the frequency domain, where the convolution operation is a great deal more efficient. The topic of computation time becomes pertinent when the user wishes to, for example, spatialise several sources using several earplug~ objects (or indeed other objects). A frequency domain solution allows for many more objects running simultaneously.

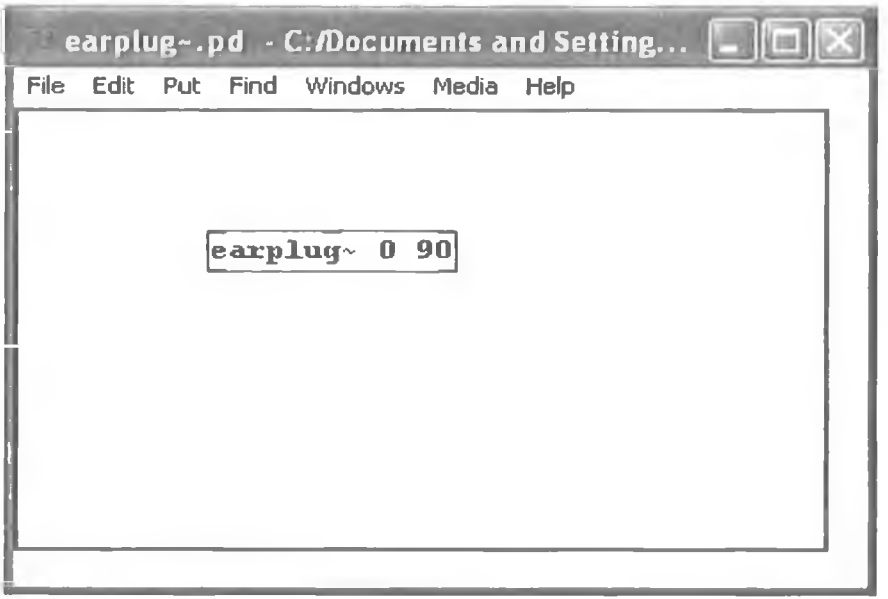
The screenshot in Figure 6 shows the Earplug~ object, which takes angle and elevation as initialisation arguments and centre and right inlets, respectively. The left inlet takes a mono input and the outlets output the processed left and right signals.

---

[13] E.M. Wenzel and S.H. Foster: 'Perceptual Consequences of Interpolating Head Related Transfer Functions during Spatial Synthesis' *Institute of Electrical and Electronics Engineers Workshop on Applications of Signal Processing to Audio and Acoustics* (10/1993), 102–105.

[14] Hartung *et al.*, *Comparison of Different Methods for the Interpolation of Head Related Transfer Functions.*

Figure 6. The Pure Data Earplug~ Object



The earplug~ object, although a more complete solution than the Csound HRTFer, could be greatly improved by performing interpolation and convolution in the frequency domain.

6. Future Work

The author is currently completing development of HRTF based binaural localisation systems for Csound, Pure Data and stand alone applications. The improvements suggested in this article are being implemented by means of a minimum phase system and two original magnitude interpolation systems, one based on phase truncation, the other on a system which models phase on a geometric model with frequency dependent scaling. It is hoped that these systems will be used as the core of an original loudspeaker auralisation tool.

7. Conclusion

This article has presented an overview of the spatial auditory system, followed by a brief summary of methods currently and traditionally employed to artificially spatialise audio. Based on these insights, the

development of a binaural solution to artificial simulation of audio spatialisation is outlined. The topic is dealt with in an intuitive way derived from and with constant reference to Fourier representation of complex sounds as sums of simple sinusoids. An overview and critique of current solutions in the computer music domain is presented and improvements suggested.

Acknowledgements
This work is supported by the Irish Research Council for Science, Engineering and Technology: funded by the National Development Plan and NUI Maynooth.

Select Bibliography

Primary Sources:

Begault, Durand R.: *3-D Sound for Virtual Reality and Multimedia* (London: AP Professional, 1994).

Burkhard, M.D. and Sachs, R. M.: 'Anthropometric Manikin for Acoustic Research' *Journal of the Acoustical Society of America* 58/1 (7/1975), 214–222.

Gardner, W. G. and Martin, K. D.: 'HRTF Measurements of a KEMAR Dummy Head Microphone' (Massachusetts Institute of Technology: 1994) (http://sound.media.mit.edu/KEMAR.html, accessed 7/2007).

Hartung, K., Braasch, J. and Sterbing, S.J.: 'Comparison of Different Methods for the Interpolation of Head Related Transfer Functions' *Audio Engineering Society 16th International Conference: Spatial Sound Reproduction* (3/1999), 319–329.

Kulkarni, A., Isabelle, S.K. and Colburn, H.S.: 'Sensitivity of Human Subjects to Head-Related Transfer-Function Phase Spectra' *Journal of the Acoustical Society of America* 105/5 (5/1999), 2821–2840.

Moore, Brian C.J.: *An Introduction to the Psychology of Hearing* (London: Elsevier Academic Press, 1977; 5th edn, 2004).

Oppenheim A. and Schafer, R: *Discrete-Time Signal Processing* (New Jersey: Prentice Hall, 1989; 2nd edn, 1999).

Rumsey, F. and McCormick, T.: *Sound and Recording: An Introduction* (Oxford: Focal, 1992; 4th edn, 2002).

Wenzel, E.M., Arruda, M., Kistler, D.J. and Wightman, F.L.: 'Localization using Non-individualized Head Related Transfer Functions' *Journal of the Acoustical Society of America* 94/1 (7/1993), 111–123.

Wenzel, E.M. and Foster, S.H.: 'Perceptual Consequences of Interpolating Head Related Transfer Functions during Spatial Synthesis' *Institute of Electrical and Electronics Engineers Workshop on Applications of Signal Processing to Audio and Acoustics* (10/1993), 102–105.

Xiang, P., Camargo, D. and Puckette, M.: 'Experiments on Spatial Gestures in Binaural Sound Display' *11th International Conference on Auditory Display* (7/2005), 1–4.

http://www.csounds.com (accessed 7/2007).

http://crca.ucsd.edu/~msp/Pd_documentation/index.htm (accessed 7/2007).

Secondary Sources:

Blauert, Jens: *Spatial Hearing: The Psychophysics of Human Sound Localization* (London: MIT Press, 1997).

Gelfand, Stanley A.: *Hearing: An Introduction to Psychological and Physiological Acoustics* (New York: Marcel Dekker, 1981; 3rd edn, 1998).

Howard, David M. and Angus, J: *Acoustics and Psychoacoustics* (Oxford: Focal Press, 1996; 3rd edn, 2006).

Kulkarni, A., Isabelle, S.K. and Colburn, H.S.: 'On the Minimum-phase Approximation of Head-related Transfer Functions' *Institute of Electrical and Electronics Engineers Workshop on Applications of Signal Processing to Audio and Acoustics* (10/1995), 84–87.

Moore, F. Richard: *Elements of Computer Music* (New Jersey: Prentice-Hall, 1990).