3D SOUND SYNTHESIS FOR HEADPHONES AND SPATIAL AUDIO

DIGITAL SIGNAL PROCESSING EECS 195
RESEARCH PAPER

Rod Jard Paholio
50849727
3/20/05

ABSTRACT

3D sound environments or spatial audio are quickly becoming an integral part of entertainment. It is now possible to watch a movie in a theatre while getting a complete surround sound experience with just two channels of audio. One is able to listen to music in crisp 3D sound just from headphones alone. Video games are more captivating and immersive with 3D headphones. Users will be bombarded with realistic explosions and dialog, all of which are 3D virtual sound representations of the original sound. All this is done on just two channels. The most important aspect of spatial audio is HRTFs or Head Related Transfer Functions. HRTFs are fundamental to 3D sound synthesis. HRTFs provide spatial cues that enable 3D audio synthesis. 3D sound for two channels requires different parts for it to work. Since most people have different shaped heads, HRTFs can be detrimental in synthesizing 3D sound because they provide the necessary data to artificially synthesize spatial audio. The different shapes and sizes of people's heads are important because they affect one's perception of sound as well as the location of that sound. Furthermore, other cues such as reflections from surrounding objects and one's own body can effect how one receives a sound or noise.

INTRODUCTION

Human beings perceive sound in three dimensions. Localization of sound depends on the way the sound waves from the same source differ from each other as they reach the left and right ear. The head, torso, shoulders and the outer ears modify the sound arriving at a person's ears. This modification can be described by a complex response function - the Head Related Transfer Function (HRTF). HRTFs can be used to generate spatial audio as well as sound localization. HRTFs contain all the information about the sound source's location (its direction and distance from the listener). If properly measured and implemented, HRTFs can generate a "virtual acoustic environment". The study of HRTFs is a rapidly growing area with potential uses in virtual environments, auditory displays, entertainment industry, human-computer interface for visually impaired, aircraft warning systems and many others.

BACKGROUND

Spatial Audio

Spatial audio is processed sound that gives the listener a sense of location for a virtual sound source. With headphones, spatial audio should give a sense of a sound that emanated outside of the listener's head. This can be very different from regular recorded stereo, which is usually restricted to a line between the ears when using headphones. Spatial audio works in way that if sound waves that arrive at the listener's eardrums are identical to those of real audio sources at certain positions, the listener will perceive those sounds as emanating from a source at that particular position **[1]**. Since people are born with two ears, people only need two channels of sound to recreate this effect and people can also present this sound with regular headphones.

Individual Differences

An aspect of spatial audio that has frustrated researchers who want to synthesize perceptual cues involved in spatial hearing is taking into consideration individual differences of people's ear sizes, head shapes, location, body volume, etc. Subtle differences in ear physiology equate to individual differences in sound localization that every one of us has learned to "listen through our ears **[7]."** Other aspects of localization that individuals differ in are which cues individuals use to differentiate between sounds that have the same inter-aural time delay and how individuals determine differences in space characteristics and localized sound sources based solely on the nature of the signal's reverberation.

Spatial Audio with Headphones

Headphones are usually designed for most spatial audio systems. This results in limitations for their use. Spatial audio may be limited to applications in which a user already has on some type of headgear, or situations in which advantages of spatial sound far outweigh the inconvenience of wearing headsets. Headphones are mainly used since they fix the geometric relationship between the ears and physical sound sources (headphone drivers). Compared to speakers, headphones eliminate crosstalk between binaural signals. With further signal processing, one can compensate for these effects, allowing spatial audio to be represented over free field speakers. But, in order to compensate for the effects of speakers, the spatial audio system must know the listener's position and orientation with respect to the speakers; meaning that even without headphones, head tracking is still needed. One cannot produce true 3-dimensional spatial sound in any way without head tracking. Still, multi-speaker surround-sound systems are still possible, which may prove useful in many applications depending on the situation **[1].** Furthermore, headphone reproduction is different to loudspeaker reproduction since each ear is sent a single signal from one channel. This showcases the binaural signal situation and allows for the ears to be fed with signals that differ in time by up to the binaural delay, and also differs in amplitude by amounts similar to those differences that result from the shadowing effects of the head. This situation explains the need for microphone recording techniques that use microphones spaced apart by the binaural

distance, and using an object similar to the human head to trick the microphones into producing signals with correct differences **[10]**.

Sound localization Cues

Accurately synthesizing spatial sound would add to the immersiveness of virtual environments. We notice sound in our everyday lives, but we don't understand that they provide us with cues about our natural environment. Sound localization, though, is a more complex human process. One must first understand how humans hear and localize sound in order to artificially spatialize it. To help locate the position in space of a sound source, humans use auditory localization cues. There are several sources of localization cues, such as: interaural time difference, head shadow, pinna response, shoulder echo, head motion, early echo response, reverberation, and vision to name a few. **[11]** The first four are static and the rest dynamic. They are referred to as dynamic since they involve the movement of the individual's body, affecting how sound enters and reacts with the ear. The most important of these cues is Interaural time difference (ITD).

Interaural time difference

John Strutt, who is well known as Lord Rayleigh, was one of the first pioneers in spatial audio research. Rayleigh developed a Duplex Theory about 100 years ago. His theory established a model for estimating a source's spatial location by two primary binaural cues: the interaural time difference or delay (ITD) and interaural level differences (ILD), the latter is also known as interaural intensity difference (IID) **[8]**.

Interaural time difference is described as the time delay between sounds arriving at the left and right ears. This is a primary localization cue for interpreting the lateral position of a sound source. The interaural time delay of sound sources that are directly in front or behind a subject are approximately zero, while sound sources to the far left or right are around 0.63ms **[11]**. The frequency and the linear distance of a sound source are factored into the ITD as well. ITD comes from the difference in distance between a sound source and the two ears. Since sound travels at a constant velocity, this distance difference equates to a time difference. A sound wave traveling from a sound source located on your left side will reach your left ear before it reaches your right ear. By the time the sound source moves towards the front of one's head, the interaural time difference will drop to zero when the sound source is centered between one's ears. The ITD Maximum depends on the width of one's head and on the speed of sound in the listener's environment. If we take the speed of sound to be 330m/s, and the distance between your ears to be 15cm, then the maximum interaural time difference will be about 0.45ms. **[2]**

The ILD can be defined as the level difference that is generated between both ears by the sound. Most synthesizer users are familiar with IID, which comes from the fact that, due to the shadowing of the sound wave by the head, a sound coming from a source located to one side of the head will have a higher intensity, or be louder, at the ear nearest the sound source. Therefore, one can create the illusion of a sound source emanating from one side of the head merely by adjusting the relative level of the sound that are fed to two separate speakers or a pair of headphones. **[2]**

There is a simple method to derive the interaural time delay which is depicted in **Fig. 3**, which shows a plane view of a conceptual head, with left ear and right ear receiving a sound signal from a distant source at azimuth angle (θ) or about +45deg as shown in diagram. Imagine a sound wave from a distant source that hits a spherical head with a radius **r** from a direction specified by the azimuth angle (θ). When wavefront A-B arrives at the right ear, it can be seen that there is a path length of (a+b) that has to travel before it finally arrives at the left ear. With a symmetrical configuration, the **b** section is equal to the distance from the head center to wavefront A-B, therefore: **b = r*sin (θ)**. It should be clear that the section represents a proportion of the circumference, subtended by (θ) **[8].** Visually, the path length (a+b) is given by **Fig. 1:**

$$(a+b) = (\frac{\theta}{360})2\pi r + rsin\theta$$

Fig. 1

The ITD can provide a major cue for azimuth localization, and is a core part of any HRTF model. If the head is approximated by a sphere of radius **a**, the ITD for an infinitely distant source can be computed by the formula in **Fig. 2**:

$$\mathrm{ITD} = \frac{a}{c}(\theta + \sin\theta)$$

Fig. 2
Equation for Interaural Time Delay
http://ieeexplore.ieee.org/xpl/abs_free.jsp?arNumber=759855

In this case θ is the azimuth angle and **c** is the speed of sound or 340m/s. This formula is restricted to angular frequencies greater than **a/c**, and corresponds to the differences at times when the sound first arrives. A surface of constant ITD is a cone of revolution about the interaural axis. This revolution is called the "cone of confusion." **[3]** Neither the ITD nor the ILD (Interaural Level Difference) changes as a point is moved around a cone of confusion in the case of an ideal spherical head. Many localization errors (such as front/back errors and up/down errors) can result in mis-location on the cone of confusion; this can be attributed to the ITD and the ILD being the primary cues for source localization. But in reality, the ITD does vary around a cone of confusion; this shows that the ITD is a function of elevation as well as azimuth. This elevation dependence comes about because, although the distance from the source to the ipsilateral ear (the ear that is visible from the source) is constant, the length of the shortest path from the source to the contralateral ear (the ear that is not visible from the source) changes with elevation **[3]**. These changes can be attributed to the non-spherical shape of the head and that ears are not positioned across a diameter, but displaced behind and below the center of the head.

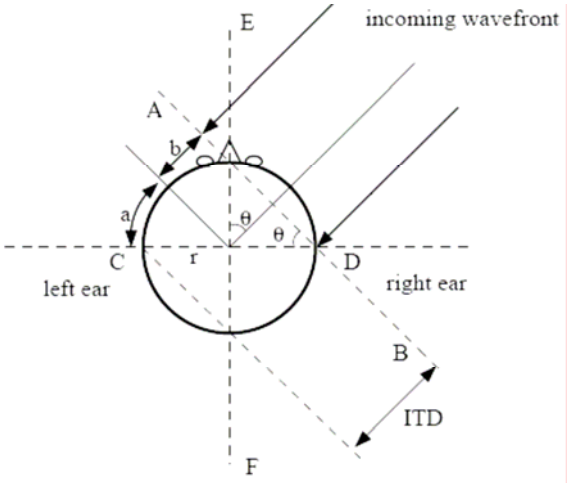

Fig. 3: Plane view showing interaural time delay (ITD) across the head and interaural level difference (ILD)
Source: http://disi.eit.uni-kl.de/arbeiten/mingli/hrtfmodelling_thesis.pdf

Reverberation

Real world sounds are combinations of original sound sources and their reflections from surfaces in the world such as floors, walls, tables, etc. **[11]** In enclosed environments the sound that reaches your ears from a sound source comes not just from the direct path from the source to your ear, but also through indirect pathways that correspond to the reflection of the source sound waves off different objects in the environment. Reflections can give addition cues to localization because the timing and frequency response characteristics of the secondary sounds depend on the location of the sound source relative to the reflecting objects, and on the location of those objects to the listener. **[2]**

Other Spatialization Cues

Sound spatialization cues are not limited to just ITD and Reverberation. There are some cues that can affect how individuals perceive sound in spatial situations. All of these cues, which I am about to discuss, contribute in some way or another to the ability to spatially locate sounds in 3D space. In order to provide accurate sound immersion, 3D sound synthesis needs to deal with these cues. This is difficult since researchers do not understand how the brain translates signals it gets from the ear or understand the characteristics that cause sound to be perceived in 3D space.

Head shadow

This term describes sounds that go around or through the head for the sound to reach an ear. The head is one of the reasons for attenuation or reduced amplitude of overall intensity. The head also gives a filtering effect. Filtering effects of head shadows can cause perception problems with direction and linear distances of sound sources.

Pinna response

This describes the effects of the external ear, or pinna, on sound. The pinna filters high frequencies in a way that affects the perceived lateral position, or azimuth, and elevation

of sound sources. Pinna "filter" responses are dependent on overall direction of sound sources.

Shoulder echo
Frequencies within 1 and 3kHz are reflected from the upper torso of the human body. This reflection produces echoes that ears perceive as a time delay, which is dependent on the elevation of sound sources. Reflectivity of a sound is dependent on the frequency; though some sources don't reflect as strongly as others.

Head Motion
This describes the movement of the head in determining a location of a sound source, which is an integral part of human hearing. As the frequency of a sound source increases, head movements occur more often. This is attributed to high frequencies since they don't bend around objects as much.

Vision
Vision is important since it enables us to quickly locate the physical location of a sound and confirm the direction that we perceive the sound to be coming from **[11]**.

HEAD RELATED TRANSFER FUNCTIONS (HRTF)

The head-related transfer function for 3D sound synthesis

In synthesizing accurate 3D sound, attempts to model the human acoustic system have taken binaural recordings a step further by recording sounds with tiny probe microphones in the ears of a real person. These recordings are then compared with original sounds to compute a person's head-related transfer function (HRTF). The HRTF is a linear function that is based on the sound source's position and takes into account many of the cues humans used to localize sounds. The HRTF is then used to develop pairs of finite impulse response (FIR) filters for specific sound positions; each sound position requires two filters, one for the left ear, and one for the right. Thus, to place a sound at a certain position in virtual space, the set of FIR filters that correspond to the position is applied to the incoming sound, producing spatial sound. **[11]**

Sound source spatialization in virtual acoustic environments using headphones requires the filtering of the sound streams with HRTFs. The HRTFs capture both, the frequency and time domain aspects of the listening cues to a given sound position. Some tests involve generic HRTFs using the KEMAR **[6]** database. But, the use of non-individualized transfer functions can lead to degradation of localization accuracy, and an increase in the following errors: localization error, localization blur, externalization error, and cone of confusion. **[9]**

HRTFs are combinations of spatialization cues that enable researchers to artificially synthesize spatial audio. HRTF data can be acquired through different mediums. A well-known research project on HRTF involved students at MIT and their Knowles Electronic Manikin for Acoustic Research (KEMAR) dummy. KEMAR is an anthropomorphic

manikin whose dimensions were designed to be similar to an average human. Molds of human pinna were used in the research as mini microphone "pinna." They acquired HRTF measurements by inserting miniature microphones into the ear canals of the mannequin. While a loudspeaker is playing a measurement signal, microphones record the sounds. The recorded signals are then processed with a computer to derive pairs of HRTFs (for the left and right ears) that correspond to the sound source location. Each HRTF, which typically consist of several hundred numbers, describe the time delay, amplitude, and tonal transformation for the particular sound source location to the left or right ear of the subject. This procedure is repeated for many locations of the sound source relative to the head, which results in a database of hundreds of HRTFs that describe the sound transformation characteristics of a particular head for a given degree level. **[5]** The measurements were made in MIT's anechoic chamber. This data is very extensive and describes ITD and responses for different elevation angles. The data represents responses from sounds heard from different locations of a room. According to the researchers, the sound was sampled at a 44.1KHz frequency and about 710 different locations of the room were used **[6]**.

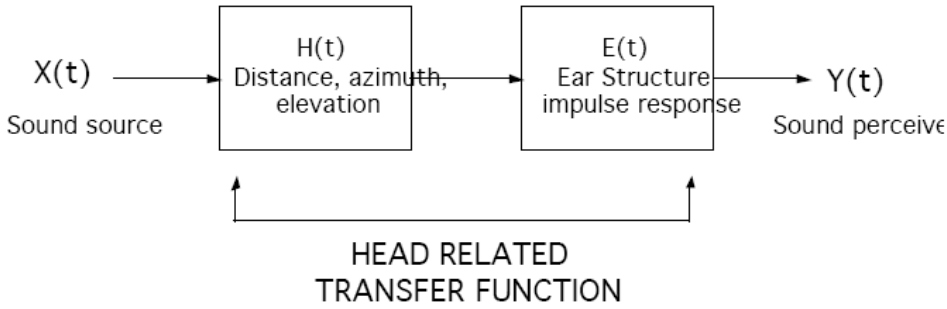

Fig. 4
Source: http://www.ee.columbia.edu/~dpwe/e4810/projects/

To be able to generate 3D sound, an HRTF is implemented. An HRTF is the Fourier transform of the impulse response from the source of the sound to the human eardrum. For example, to generate a sound that feels like the source is coming from the right side of the ear, we need a specific HRTF of the human ear's impulse response to sound coming from the right area. We want the image that HRTF created so we can use that data to synthesize 3D sound. Because the HRTF is from the source of the sound to the eardrum, **Fig. 4**, it is a function of frequency, azimuth angle and elevation which is the path that sounds use to travel to the ear - right or left, up or down, near or far, as well as the function of the pinna structure (how sound is collected and reflected into the ear drum).

3D audio systems work by implementing the process of natural hearing, basically reproducing sound localization cues at the ears of the listener. This can be done by using a pair of measured HRTFs as a specification for a pair of digital audio filters (equalizers). When a sound signal is processed by the digital filters and listened to over headphones, the sound localization cues for each ear are reproduced, and the listener should perceive the sound at the location specified by the HRTFs. This process is called binaural

synthesis (binaural signals are defined as the signals at the ears of a listener) **[5]**. This binaural synthesis process is shown in **Fig. 5**.

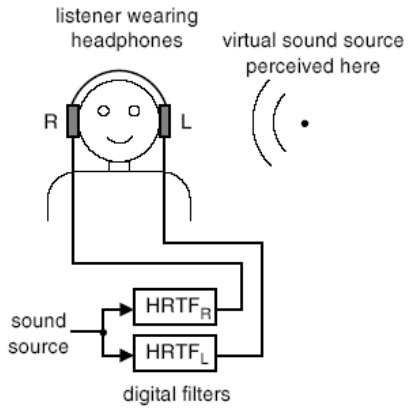

Fig. 5 Binaural Synthesis with HRTFs
http://www.harmony-central.com/Computer/Programming/3d-audio.pdf

IMPLEMENTATION WITH MAX/MSP

3D SOUND SYNTHESIS

The idea is that the farther the distance the lower the intensity and vice versa. Also, there should be a delay from when the right ear hears the sound as compared to when the left hear perceives it. From here I should apply an FIR filter. At first, I designed my patch to implement azimuth angles and distances, and then I realized I could synthesize 3D Sound just by using simple panning, **Fig. 6**, with reverberation from a reverberation object, **Fig. 8**. I used the same idea from a previous panning assignment from Professor Dobrian's EECS 195 class. From there I applied the same idea for the distance cue. But instead of changing the value for left and right channels, I just subtracted from a constant of 1 the value I received from the user interface and multiplied that with a sound file and then put that into the panning delay, this is shown in **Fig. 9**. The connection with the sound file is shown in the bottom left hand corner while the subtraction is shown on the top right hand corner of the figure.

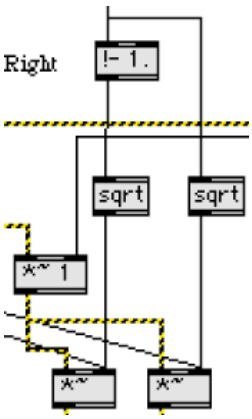

Fig. 6
Panning

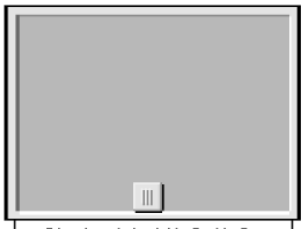

Fig. 7
User Interface

Fig. 8
Reverberation Object

Max/MSP Reverberation Object from: http://www.akustische-kunst.org/maxmsp/

It is basically impossible to implement HRTF into Max/MSP since there are too many data points to consider. Instead of using HRTFs from the KEMAR database as data to be implemented into Max, I decided to use the concept of how HRTFs are used in binaural sound spatialization. There were key equations I had to implement and translate into Max. But nothing will sound like true 3D sound unless I actually used those specific HRTF data. Since most of our lobes are different in a sense that we hear sounds differently, generalizing HRTFs and going with that concept would be the best scenario in implementing spatialized 3D sound in Max. Since I had previous problems with Max, and being that I am using the generalized concept of HRTF, it would be best to just make one 3D binaural sound from a stereo sound file. I would then make the sound as if it was coming from the front right or front left side of the users view.

RESULTS: 3D SOUND SYNTHESIS

The results were quite adequate. I didn't fail, since the patch worked. But was not completely successful. I had many problems in synthesizing 3D sound with Max/MSP. One problem I had was creating a patch that would synthesize in 3D planes. I used only X and Y coordinates to place my sound source, giving me a 2D plane. Still, I was able to achieve some sort of spatialization of a sound source through panning and implementing freeverb~ into the patch. To me, it felt like the 3D sound was being synthesized, but in reality, it just seemed as though my patch was just "tricking" itself into creating a spatialized sound through sound and object manipulations in the patch. I didn't use any equations implementing ITD, HRTF, etc., or from anything I researched. I would say this was due to the panning delay which worked and I just went from there. I thought, if it worked for left and right, why not front and away. The patch sort of worked.

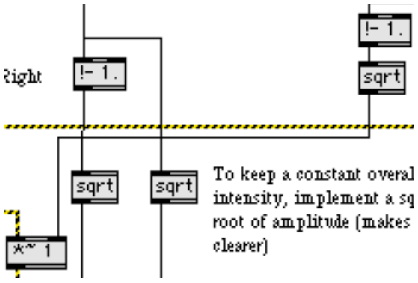

Fig. 9
Distance

DISCUSSION:

I didn't take into account elevation, as in the 3D field. I also didn't take into account interaural time delay, which is an important aspect of sound spatialization. I believe, though sparse in content, the patch works in creating 3D sound synthesis. Though it may look like Max/MSP manipulation since little if any ITD or any equations where used other than the expressions, it does seem to do what it was intended to do. Implementing the user interface, **Fig. 7**, was a convenient choice. Using the block interface I was able to use the data values that came from it and implemented it with the panning. One difficulty I had with synthesizing 3D sound was how to synthesize sound resonating from the back of the head. What kinds of cues will I have to know or what HRTFs would I need to know to accomplish that?

REFERENCES

[1] Burgess, David; Mynatt, Elizabeth; Lee, Mark. "Spatial Sound - A system for synthetically spatializing sound sources" Georgia Tech University, Georgia. http://www.cc.gatech.edu/gvu/multimedia/spatsound/spatsound.html

[2] Clark, James J., "Advanced Programming Techniques for Modular Synthesizers." McGill University. Montreal, Quebec, Canada: 2003. http://www.cim.mcgill.ca/~clark/nordmodularbook/nm_spatialization.html

[3] Duda, R. O. "An Adaptable Ellipsoidal Head Model for the Interaural Time Difference." Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings, 1999 IEEE International Conference, Phoenix, AZ: 1999. <http://ieeexplore.ieee.org/xpl/abs_free.jsp?arNumber=759855>

[4] Fuller, Eric; Omodunbi, Bankole; Sheng-Lung, Lee. "Synthesizing 3D Sound and Sound Localization." Columbia University: 2003. http://www.ee.columbia.edu/~dpwe/e4810/projects/

[5] Gardner, William G. Harmony-Central. 15 March 1999. Wave Arts, Inc. 3 March 2005 <http://www.harmony-central.com/Computer/Programming/3d-audio.pdf>

[6] Gardner B., Martin, K.: "HRTF Measurements of a KEMAR Dummy Head_Microphone," http://sound.media.mit.edu/KEMAR.html, 1994.

[7] Holm, Frode; Kouznetsov, Alex and Pope, Stephen T. "ATON Report 2003: A representation and Infrastructure for Flexible Sound Spatialization." University of California at Santa Barbara: June 2000. <http://www.create.ucsb.edu/ATON/00.06/SpSo.1c.pdf>.

[8] Li, Ming. "Implementation of a model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction." Masters Thesis, University of Kaiserslautern: 2003. http://disi.eit.uni-kl.de/arbeiten/mingli/hrtfmodelling_thesis.pdf

[9] Noisternig, Markus; Musil, Thomas; Sontacchi, Alois; Holdrich, Robert. "A 3D Real Time Rendering Engine for Binaural." 2003 International Conference on Auditory Display, Boston, MA: 2003. <http://www.icad.org/websiteV2.0/Conferences/ICAD2003/paper/26%20Noisternig.pdf>

[10] Rumsey, Francis. "Spatial Audio," Oxford; Boston; Focal Press, 2001

[11] Tonnesen, Cindy; Steinmetz, Joe. "3D Sound Synthesis." Human Interface Technology Laboratory, Washington University: 1 Sept. 1993. http://www.hitl.washington.edu/scivw/EVE/I.B.1.3DSoundSynthesis.html