A Customizable, Headphone-based Spatial Audio Core for Generalized Use Within Third-party Programs

Written by: Carissa Ouelette, Sean Cunningham, Ben Mahlman, Matt Holland
Engineering Undergraduates, University of Victoria

Abstract

The study of binaural synthesis is fast becoming the topic of much research and exploration in the expanding digital world. In the real world, sounds that are processed by the brain provide a wide range of information about it's sound environment such as direction, elevation, and motion. Standard headphone audio-digital outputs lack such information, but through binaural synthesis the construction of a realistic 3D sound environment is possible. This project explores the underlying concepts necessary for the development of a core application capable of digitally reproducing binaural sounds comparable to real life sound environments. Based on the underlying concepts, a core application was developed that, given an input sound file, direction, elevation, and motile information, a binaural sound file is produced. The core application was developed with the intention of versatile integration with future interactive applications. With the core application as this project's starting point, future works will investigate possible methods to calibrate the virtual 3D sound experience. From there, the core algorithms and an appropriate calibration method can be integrated within a vast array of possible applications.

Intro

The Desire for Real World Experiences

The rapid advancement of technology has provided the world with widely available interactive media which have provided mediums of information exchange between the analog "real" world and the ever-expanding digital world. Recent technological developments such as digital TVs, the internet, personal computers, iPods, and smartphones have played key roles in the exponential rise in the exchange of digital information. As an example, the last century has given rise to the development of new technology designed to replicate depth perception: the third visual dimension attributed to human *binocular vision*. (Binocular means "having" or "relating to" two eyes). This third dimension technology synthesizes an output homologous to the way humans see in the real world. The commercialization of 3D films in the entertainment industry has clearly demonstrated the value in closing the gap between recognizably synthetic experiences, and the analog experiences that make up human perception of the real world.

The Uprising of Virtual 3D Audio

In the last two decades, there has been a significant uprising of research and development in the field of Virtual 3D Audio, or synthetic sound vision. Especially now, on the heels of visual 3D technologies, synthetic 3D sound boasts yet another desirable feature in a step towards a fully immersive digital experience. Similar to binocular vision, having two separate ears and a brain to process the left and right signals constitutes what is called *binaural hearing.*

A person's auditory processing centers, which are found in the temporal lobes of the brain, are responsible for one's sense of spacial awareness. It is easy to take for granted the information derived from sound environments since the brain automatically processes and integrates the variations in sound that the left and right ear receive. The intricate structures of the human ear sense minute variations in air pressure produced by sound waves, which instantaneously activate auditory neural pathways of the central nervous system where the sensory input is processed. Overall, this process detects

analog variations in sound amplitude and frequency (or pitch) and converts them to bioelectric signals. Perceived amplitude can assist in discerning distance and directionality from the subject, and pitch can assist in determining the motility of an object (due to the Doppler Effect); though their functions are not strictly limited to these assertions. Another key component to sound spatialization lies within the brain's ability to discern the minute variations in a sound wave's arrival time at the left and right ear. For example, if a sound arrives at the left ear slightly before the right ear, the brain processes this information as a sound that is coming from the left. Variations of this time delay provide a very sensitive perception of the direction or elevation of a sound in relation to one's self. The brain combines and compiles the received amplitude, pitch, and time delay information and constructs a "sound image" which relays a large amount of useful information regarding one's environment.

In the age of information, the ability to simulate the manner in which sound enters the human ear canal boasts a wide array of exciting applications. As mentioned above, binaural hearing enables the perception of positional, referential, and motile information about the a person's sound environment. The method simulating this effect, and the focus of this project, is called *binaural synthesis. Binaural* meaning having or relating to two ears, and *Synthesis* meaning to recombine elements to create something new. With regards to 3D audio, the intention is to digitally create binaural sounds in an effort to simulate auditory spatialization. With algorithms that can accurately reproduce a 3D spatial awareness, various applications can be developed to take advantage of the information present in the sounds. The topic of binaural synthesis is not a new topic and is currently being explored in video game applications, audio teleconferencing applications, binaural hearing aids, audio cognition tests, and cinematic applications.

Optimizing Spatialization Perception

In order to innovate and expand the quality of information exchange offered through binaural synthesis, the topic of calibrating the experience to the unique user must eventually be addressed. The manner in which any one individual perceives a 3D sound image is entirely unique, much like a fingerprint. The unique anatomical structures of the ear involved in receiving sound effects how the brain processes the information. Since no two individuals can be characterized to hear the exact same way, it is then possible to say that a synthesized audio spatialization can be optimized, in some manner, to cater to the unique user. Finding a suitable calibration methodology to match the application will also be a challenge. This could range from the utilization of generalized ear shapes, to the exact modelling of a individual's ear. The effectiveness of calibration techniques might be determined by the demand for accuracy given the application. For example, a game might require a less accurate calibration method than the design of a binaural hearing aid.

The Synthesis Initiative

This report will begin with an overview of the project's approach and initiatives for developing and demonstrating binaural synthesis. This will be followed by a survey on existing 3D Audio literature that pertains to the foundations on which the core audio processing application has been developed. This will be rounded off with a detailed description of the developed core application that is capable of creating uncalibrated binaural sounds. Finally, future initiatives pertaining to the calibration and possible applications of core algorithms will be discussed.

Approach

The goals of the 3D Stereo Navigation team is to develop the algorithms necessary to

digitally create binaural sounds that effectively simulate real world sound immersion. This project's method of approach can be summarized in the following three points:

1. Create a core application to digitally synthesize a 3D spatial awareness with discernible directionality and elevation.
2. Explore and develop various calibration methodologies to optimize the experience for physiologically unique users.
3. Create an interactive application(s) that utilizes 1), and 2) to demonstrate accurate 3D sound images.

The following topics of discussion in this report will address the initiative mentioned in the first point. The overall goal is to create an algorithm that is generic enough to accept a sound file and its respective localization information as inputs with the 3D binaural sound as the output. This core application can then be integrated within various other applications.

The second and third points regarding the development of calibration methods and applications will be the focus of this project's future works. Through the exploration of the intricacies of the anatomically unique structures involved in sound reception, a calibrated user experience will enhance the accuracy of the spatialization techniques developed in the core application. As will be discussed in the future applications section, it will be beneficial to explore multiple calibration methodologies. Finding an appropriate calibration method for the specific application will be key in this endeavour.

With a highly versatile core application and an appropriate method of calibration, an interactive application will be developed to demonstrate an immersive 3D audio experience.

Literature survey "What is 3D audio?"

Through the continuous progression of both digital signal processing techniques and available hardware capabilities, software can now be developed to digitally reproduce sounds containing the localization cues necessary for one to perceive directionality. Discussed in Blauert's book [4] and summarized within [1], the auditory cues used by humans may be broken down into the seven individual categories of:

1) the interaural time difference (ITD)
2) the interaural level difference (ILD)
3) spectral cues resulting from the shape of each individual outer ear, or pinna
4) torso reflection and diffraction cues
5) the ratio of direct to reverberant energy due to the inner ear canal length and size
6) cue changes induced by voluntary motion of the individual's head
7) familiarity with the sound.

Various combinations of these cues may play an important role in aiding a person in localizing the origin of a sound. It should also be noted that this list represents the relative strength of these seven cues have in aiding a person in descending order. Although all cues would be necessary for the peak sound reproduction of a source at a specific location, the stronger cues will essentially override weaker, conflicting cues. If the conflicts are too great, however, the end result will be confusing for the listener, resulting in either an indeterminate or incorrect location.

According to Lord Rayleigh's duplex theory [1], the ITD and ILD cues are considered to be the primary source of sound localization on the azimuth for humans. Within this same theory, it has been shown that the priority of these two cues swaps at a crossover frequency of approximately 1.5 kHz. The ITD of a binaural sound takes precedence in the lower frequencies, where head shadowing and other distortions of the sound are quite weak. In higher frequencies, however, this phase difference between each of the two channels (left and right) becomes very difficult for our brains to process. As such, the ILD begins to take precedence [1]. As shown by this theory,

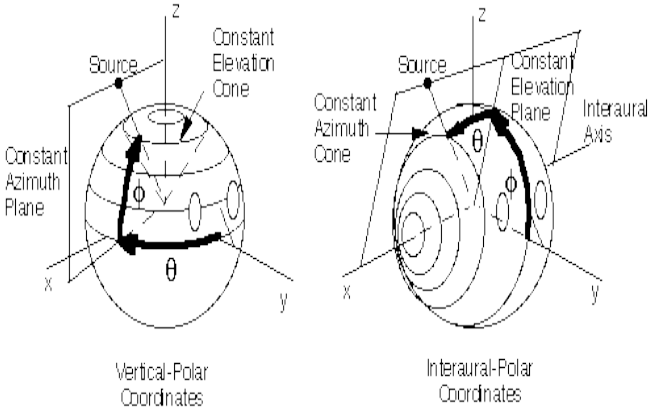
the direction of sounds may be adjusted using additional delays between the left and right channels.

From this, one is able to synthesize the basic directionality of a sound by altering only the delay between, and attenuation of, the left and right channels of a stereo or mono-stereo sound. Because this level of computation far exceeds the realism of the synthesized sound, however, this technique is not typically used to create true 3D sounds. Unlike the cues for the azimuth of a sound source, the elevation cues are not nearly as mathematically easy to represent and adjust. The primary source of elevation cues stems from the outer pinna size and shape of a particular person. Because of this, it becomes very difficult to simply modify a sound single to simulate these changes in elevation. Instead, these principles of changing the azimuth are typically used for small adjustments to a predetermined transfer function. Measured impulse responses may be instead used to apply the delay and attenuation of sound from a known direction, however, must first be measured for a specific subject (at a specific location) before they are able to be utilized.

These impulse responses may be measured on individual subjects and represented at locations in either the time or frequency domain, with exception to those induced by head movement and familiarity of a person with a sound. These two representations are referred to as the head-related impulse response (or HRIR) and the head-related transfer function (or HRTF), respectively. Typically, however, these representations will also not simulate any reverberant energy within the ear canal as this may be induced for each listener during playback.

The HRIR for a single subject at a specific location may obtained by measuring the ratio of the sound pressure of the noise within the subjects ear canal (while the sound is being played in the required location) to the sound pressure obtained at the head-center with no listener present. Once this data has been obtained, the fourier transfer function of it may be taken to acquire the HRTF. Many groups, such as the Central Image Processing and Integrated Computing department at the University of California (or CIPIC) [3], have simplified this process by creating databases of these HRTFs for others to use. As the pinna size and shape directly relate to the measured response, databases will typically contain either individual HRTFs measured from multiple subjects (as in the CIPIC database) or a single HRTF measured from an anatomically average mannequin in place of a living subject [5].

To accurately specify any location of a sound relative to a listener, a coordinate system is needed in which all angles around, above, and below the listener may be represented. Two slightly modified versions of spherical coordinate system were selected to accomplish this, although any 3D coordinate system would suffice, to allow the accurate modelling of a human's head shape [3]. Known as the vertical-polar ([Fig.1], left) and interaural-polar ([Fig.1], right) coordinate systems, each of these two representations offer their own advantages within different scenarios. The vertical-polar coordinate system is commonly used when elevation (Φ) is either locked or infrequently changed. Computationally, this allows for less resources to be used while changing the azimuth (θ) of the perceived sound. Once the alteration of



[Fig 1] Demonstrating Differences in Spherical Coordinate Methods

Image sourced from:
http://interface.cipic.ucdavis.edu/sound/tutorial/psych.html

the elevation becomes necessary, however, the interaural-polar coordinate system's becomes far more advantageous. As can be seen within Figure 1, when the azimuth is held constant for the vertical-polar technique and elevation is held constant for the interaural-polar technique, two different planes will manifest. The slight angle present in the plane of the interaural-polar technique allows for significantly more elevations and azimuths to be attained [3].

Once the directionality of a sound has been synthesized, a realistic 3D sound environment may be established by also accounting for the distance of the sound source. Though there are many different techniques to achieve this, with varying amounts of accuracy, a simple method dependent on only the distance from the point source was chosen. This technique is simply performed by using a factor of $1/(1+r_2)$, where r is representative of the distance from the listener to the source. By using this factor, a simple method of attenuating the sound may be used to reasonably represent a change in the sound source's distance.

Core application

The current generation of the 3D-Audio synthesis core program provides two unique functions – the output of a stationary sound at a specified distance and elevation, and the output of a moving sound traveling along a straight path specified by start and finish points. To simplify the program for this initial stage of development, the sound locations were restricted to the anterior and posterior half planes defined by elevations of 0° and 180° respectively.

Initial development of the core program was done using MATLAB. There are a number of built in functions such as filter(), wavread(), and wavplay() that simplified the handling of the wav files and the utilization of the CIPC HRTFs. The user interacts with the core either though the MATLAB command line interface by calling one of the two sound synthesis functions or by creating an m-file script that automatically steps through a series of points and movements.

The synthesis of the audio cues needed to perceive directionality is accomplished using the CIPIC HRTF database. The CIPIC HRTF database contains HRTFs for 86 different subjects. Each HRTF contains 2500 unique HRIRs (1250 each for the right and left ear). There are likely a small number of subjects in the CIPIC database for which the HRTF works well for a unique user. Through trial and error a user can select a subject that closely matches their own perception of sound directionality. The main task of the core program is to determine when to use each unique HRIR within the HRTF to synthesize the desired directionality of the sound.

The HRTFs do not account for the distance of the sound source from the listener. The HRTFs were produced by taking measurements in a one meter radius around the subject's head. Without adjusting for distance, all sounds produced using the HRTFs are perceived to be originating within this one meter radius.

To simulate the interaural amplitude difference needed for distance perception the core attenuates the sound by a factor of (R_2+1) where R is measured in meters. The plus one is required to ensure volume levels are not increased to the point of clipping for distances closer than one meter. Both the left and right channels are attenuated individually based upon the distance of the sound source from each respective ear. A standard interaural distance of 0.17m [11] is used. Future versions may also implement filtering to simulate the attenuation of high frequencies in air.

The audio core references sound locations using a simple x-y Cartesian coordinate system with the listener centered at the origin. Positive x values are to the right of the listener, and positive y values are in front. The software converts the rectangular coordinates into the azimuth and elevation values used by the HRTF interaural-polar coordinate system for simple selection of the correct HRIRs.

Producing a stationary source requires

a sound clip recorded as mono wav file and a desired location within the rectangular coordinate system from which to play the sound. The program reads in the wav file and loads the selected HRTF. The software calculates the azimuth and elevation and selects the correct HRIRs from the provided subject HRTF for both the right and left ears. It then calculates the distance between the sound source and each ear. The wav file is filtered using the right and left ear HRIRs after which the distance attenuation is applied to each channel. The two one dimensional arrays of filtered data are then combined into a single two dimensional array. This can be played through the system sound card using or written to an output file using built in MATLAB commands.

The vector of a moving sound source is determined by the start and end coordinates along with a speed at which it traverses the path. As with the stationary source, the program loads the desired wav file while also determining the sample frequency.

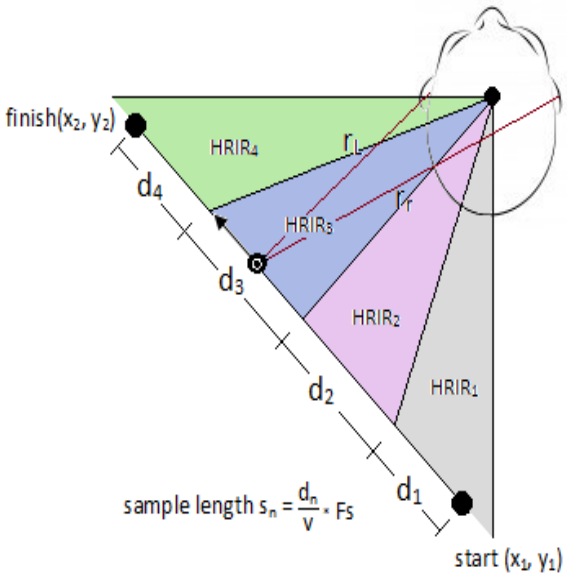
The output wav is built by filtering sections of the audio file with the corresponding HRIRs as it travels past the listener. As the sound source can only move in a straight line, each sound file will be broken into a maximum of 25 sections (corresponding to 180° of rotation on the azimuth). Beginning at the start point azimuth, the program determines the point on the path that intersects with the azimuth corresponding to the HRIR that will be used for the second section. The distance between the two points is then used to determine the sample length (in samples) of the first section by multiplying the time required to traverse the section by the sampling frequency.

Once the length of the first section has been determined the program filters the corresponding number of samples from the beginning of the wav using the HRIRs related to the start position azimuth. Similar to the stationary source, the filtered output is then attenuated by dividing each array by R_2+1. In this case R is taken to be the distance from the middle of the sound path section to the respective ear. The left and right outputs are then appended to an empty two dimensional array, and all subsequent filter outputs are appended to the end of this array.

The program continues along the sound path by traversing to the next azimuth and computing the next sample length using the distance to the intersection of the path and the next HRIR azimuth. This process continues until the HRIR corresponding to the finish point has been reached. Figure [2] illustrates a sound source traveling on a path from behind the listener to the left. The diagram has been simplified for illustration purposes, as this path would likely pass through 11-12 different HRIR sections in the actual implementation.

The final output array contains all the individual segments placed back to back resulting in a sound source that traverses the specified path with a smooth transition between the different HRIR filters.

The core program has been tested by the development team with a number of different wav files and has performed successfully in its limited implementation. Provided the user has identified an appropriate HRTF subject, the sounds are being perceived by the testers at approximately the intended distance and direction. However, the use of an



Fig[2] - Determining the sample length and source distance

inappropriate test subject routinely results in the sound being perceived opposite from the intended half plane (e.g. a sound intended to originate from in front of the listener is perceived behind), so the selection of the correct HRTF is critical.

The moving sounds produced by the core smoothly transition between HRIRs as they move past the listener. The speed of the sound source has been varied between 0.5 m/s and 5.0 m/s over source paths up to 20m in length. When simulating a slow speed or long path the user must ensure that a wav file of suitable length has been selected to avoid the program attempting to index the sample beyond the length of the array. The sample length must be longer than the time required to traverse the path at the given speed.

When it is fully developed the 3D Audio core will be a powerful tool for use in a wide variety of areas. Further development related directly the the core will focus on implementing the entire range of elevations available in the HRTF, outputting multiple sound sources simultaneously, and performing the audio synthesis in real time.

Future Works

With the core 3D audio application nearing completion the direction of the project is now turned towards developing simpler and more effective methods of calibrating the system for individual users and generating applications based on the core program.

As mentioned previously in the report, there is a strong need to calibrate the project to each individual user's unique physiology in order to enhance its performance. However, the complexity and accuracy of calibration will be highly dependent on the end user's application of the project. Take for example, a visually impaired user who uses a mobile navigation system that utilizes 3D audio to navigate around and perform daily tasks. The sensitive nature of this user would need a very high level of accuracy in order to ensure the safety of their guidance. To provide the most accurate results, it is likely that any given application would only need to be initially, so it would pose an inconvenience. Alternatively, those people who might be using 3D spatial awareness in a medium designed solely for entertainment, such as mobile applications, video games, or home theaters, may not have the available time to invest in calibrating the system to a high degree of accuracy. Oftentimes it is not the same user using the device all the time and, as such, a simpler but less precise method of calibration may be more appropriate. A more time consuming calibration period would yield more personalized and accurate results, though it would see that high precision is less necessary in a cases such presented by the entertainment industry. With such a varying number of uses for 3D spatial awareness, multiple calibration methodologies corresponding to the user's need for speed or accuracy of calibration may need to be generated. Determining the user's needs and developing methods of calibration based on these needs is one of the three core goals of this project that will be addressed in the near future.

Once an effective method of analyzing a user's needs and then calibrating the project towards the individual user has been developed, the next step will be to develop interactive applications which utilize and demonstrate the power of 3D spatial awareness. While the realm of possibilities for such applications is near infinite, the project will primarily be focused on those which are able to exhibit the effectiveness of the aforementioned calibration methods. Developing applications for the purpose of learning assistance has been one proposal, as it would serve the dual purpose of demonstrating the validity of the project while at the same time aiding in the education of others. At the very basic level these learning assistance applications would have the users identifying the direction in which a specific sound has been projected from, but this idea could be delivered in the form of small games

or adventures in order to provide amusement while learning.

Another proposed style of a demonstration application could involve projecting several simultaneous sounds, whether they be musical instruments, some variation of sound effect, etc., which appear to the listener to be moving. The users would then be able to modify their relative position to the sound projections, as well as the actual projected sounds within the 3D space. In this case, depending on the speed of the moving sound sources, the doppler effect may need to be taken into account and compensated for. While this would add additional sophistication to the project, it might not be necessary for other applications. While these are only two examples of potential demonstration applications, as stated before, the possibilities of an effective demonstration are nearly endless.

Conclusion

With such a high demand for life-like digital experiences and the mainstream availability of personal computing devices, the feasiblity of introducing virtual 3D sound to the expanding digital world is strong. The developed core algorithms at the heart of this project have successfully achieved the synthesis of binaural sounds, which set the foundation for this initiative's future works. The versatile nature of the core application enables further expansion and growth within the field of virtual 3D audio production. From here, a thorough exploration of possible calibration techniques will aim to optimize the 3D sound experience. Partnering an appropriate calibration method with the core application will enable the development of interactive 3D audio demonstrations and a vast array of user applications.

Acknowledgements

We would like to deeply thank our advisers in this project, Dr. Fayez Gebali and Dr. Haytham El Miligi, for all their guidance and support throughout this term.

References

[1] V. R. Algazi and R. O. Duda,"Headphone-based spatial sound," IEEE Signal Processing Magazine, Vol. 28, No. 1, pp. 33-42,Jan. 2011.
[2] V. R. Algazi and R. O. Duda, "Immersive spatial sound for mobile multimedia," ISM 2005 (Proc. Seventh IEEE International Symposium on Multimedia), pp. 739-746, Irvine, CA (Dec. 2005).
[3] V. R. Algazi, R. O. Duda, D. M. Thompson, C. Avendano, "The CIPIC HRTF database, "in WASSAP '01 (2001 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk Mountain House, New Paltz, NY, Oct. 2001).
[4] J. P. Blauert, ~1997!. Spatial Hearing ~revised edition! ~MIT, Cambridge, MA!, pp. 50–93, 373–374.
[5] W. G. Gardner and K. D. Martin,"HRTF measurements of a KEMAR dummy head microphone", MIT Media Lab Perceptual Computing Technical Report #280, 1994.
[6] "Spatial Sound." CIPIC Interface Laboratory.University of California, Davis. 23 Aug. 2005 <http://interface.cipic.ucdavis.edu>.
[7] L. Gaye, "Design of a 3D Sound System for Headphones." Master's degree thesis in Electroacoustics, TMH-KTH, Drottning Kristinasväg 31, Stockholm, Sweden, 2001.
[8] R. Nishimura, P. Mokhtari, H. Takemoto, and H. Kato, "Headphone Calibration for 3D-Audio Listening," in 3rd International Universal Communication Symposium, Tokyo, Japan, 2009, pp. 205-210
[9] K. J. Faller II, A. Barreto, N. Gupta, and N. Rishe, "Time and Frequency Decomposition of Head-Related Impulse Responses for the Development of Customizable Spatial Audio Models", Department of Electrical and Computer Engineering, University of Bridgeport, Bridgeport, CT, <http://dsplab.eng.fiu.edu/DSP/Publications/Pubfiles/pub152.pdf>
[10] I.S. Pardo, "Spatial Audio for the Mobile User", Master's degree thesis, Drottning Kristinasväg 31, Stockholm, Sweden, 2005
[11] "Size of a Human Skull", website, <http://www.dimensionsinfo.com/size-of-a-human-skull/>