INTERACTIVE 3D SOUND GENERATION

William Fitzpatrick, Sudhanshu Kumar Semwal
Department of Computer Science
University of Colorado, Colorado Springs, 80918, USA
wfitzpat@uccs.edu ssemwal@uccs.edu

ABSTRACT
Real-time 3D sound generation provides an opportunity to create directional sounds in game applications. In addition, they could also be useful for headset based mobile applications for the blind and visually impaired (B/VI) community providing us our primary motivation. This paper describes an implementation towards real-time 3D sound generation. HRTF functions are used to create a 3D experience using headsets. Existing HRTF database is used to composite 3D sound in real-time using TI’s DSP hardware implementing the Finite Impulse Response (FIR) filters. Ear location is estimated using existing facial tracking system. Algorithms implementing parallax correction are explained and future research directions are identified.

KEY WORDS
3D Sound generation, HRTF functions.

1. Introduction

Much interest has been generated in recent times to improve the immersive quality of 3D sound. This interest has arisen because [1,5]. The extent to which the participants feel immersed determines the quality of communication. For example, teleconferencing has evolved from conference calls over the telephone, to one-way videoconferencing, to 3D stereographic video with high-quality stereo sound. At each step of improvement, the remote participants feel an improvement in the quality of the communication, enhancing understanding. An opportunity was provided to create a real-time interaction environment between two geographically separate laboratories [1]. Using the available free software called EVO™, successful sessions were established between two locations. We also created a mobile application incorporating pico-projectors™ using small mobile devices such as android™ phones or a Samsung Galaxy Tablet™ to facilitate teleconferencing. We also started research in 3D sound generation that would provide real-time feedback to all users, especially the people with visual impairment. In this paper, our focus is the real-time 3D sound generation.

2. Sound Localization

The immersive quality of a virtual environment is dependent on the extent to which the participant believes the environment is real. By adding these other senses to the virtual environment, the sense of immersion is enhanced. In particular, if a virtual (visual) object seems to be emitting sound coming from its 3D location, the sense of the reality of the virtual object becomes even more convincing. One way to implement such sound localization is the HRTF function [1,5].

The ability to detect the location of a sound source is called sound localization. In humans, two primary cues are used to localize the source of a sound. The first cue is the relative loudness of a sound as heard in one ear compared to the other. This is called the Interaural Level Difference, or ILD. The reasons that a sound should be louder in one ear than the other are twofold: first, the ear that is closest to the source of sound would hear it louder simply by virtue of being closer, since the sound level (pressure) diminishes in inverse proportion to the distance to the source [5]. However, for distant sounds, this effect (difference in distance) becomes negligible. However, more important than the difference in distance is the fact that the head has a “shadowing effect” – the head partially blocks transmission of the sound to the ear on the opposite side of the head from the sound source. This effect is more pronounced for high frequencies than for low, as the sound waves for low frequencies have a wavelength much larger than the head dimensions, and these waves tend to diffract around the head to reach the opposite ear.

The other primary cue is the time delay difference with which the sound reaches one ear compared to the other, called the Interaural Time Difference, or ITD. For frequencies below 800 Hz, the time delay from one ear to the other is no more than \( \frac{1}{2} \) wavelength of the sound, so it is possible for the brain to evaluate the phase difference for these low frequencies. For frequencies above 1600 Hz, it is no longer possible to use phase (because there is more than one cycle of phase difference), and the brain evaluates group delays instead. This can be thought of as determining the relative delay of the onset of a sound or an amplitude-related feature in the sound envelope. For intermediate frequencies between 800 and 1600 Hz, a combination of the two effects is used. At the high frequency range, distance estimation using ITD is assisted by the ILD as well.

If these two mechanisms were the only ones used to judge the position of a sound source, it would not be possible to do so unambiguously, because at any given
combination of ILD and ITD, the only determination about the location of the sound source which can be made is that the locus lies somewhere on the surface of an imaginary cone, called the cone of confusion.

Fortunately, the human hearing mechanism has other features that help to further pinpoint the location of the sound source. The ridges, creases and other structures of the external ear (the pinna) are radially asymmetric, and serve to filter sounds coming from different directions uniquely. To a lesser degree, the head and torso also affect the spectral content, especially to the ear on the opposite side of the head from the sound source. Together, these features color the spectrum of sound in a way that is unique to each possible vector to a sound source. The brain is then able to make use of the spectral coloration to help resolve the ambiguity caused by the cone of confusion. This spectral coloration can be measured, and together with the ILD and ITD comprise what is known as the Head Related Transfer Function, or HRTF.

2. 3D Sound for the B/VI Community

By far the most common problem that has been tackled has been the common practice of using graphical user interfaces (GUI) to provide information to users, a practice which makes finding some way of providing equivalent information to the visually-impaired a necessity. The problem of providing this equivalent information has been solved in many ways – the use of “hearcons” or “earcons” as opposed to icons is the most common solution to this problem [2,3]. There are three barriers – the pixel barrier, the mouse barrier, and the graphics barrier – to the use of GUIs by the blind, and these replacement auditory icons use sound to bypass most of these barriers and provide the symbolic information to the visually impaired user instead [3]. The use of “visual metaphors” is a common way of providing information to users in menus [5]. Subsequently, the use of audio metaphors is an established way of providing quickly understood icon data to the user, and is commonly used by programs today in order to provide information to the visually impaired user.

2.1 3D Sound in Teleconference Environment

One way to enhance 3D immersion is to have a token person at the main meeting site fitted with microphones in such a way as to transmit to remote participants sounds as heard by this one particular person. In order to do this, microphones must be positioned as close as possible to the person’s eardrums so that the full effect of the token’s pinnae, head and torso are encoded into the sound which is to be transmitted to remote sites. Commercially-available headset called – the Echo™ 3D Recording Premium Headset, provides such a facility. Intended to be used with the Livescribe Echo™ Smartpen, this headset has ear-buds, integrated microphones located in close proximity to the ear-buds. This headset has the ability to record, and playback 3D rendering of the sound. This is possible because the wearer’s HRTF is encoded into the recorded sound by virtue of the fact that the microphones, being located very close to the ear canal openings, receive sound which has been filtered in a natural way by the presence of the pinnae, head and torso.

We demonstrated the 3D sound effect by making a recording while wearing the headset while other people were talking during another phase of the demo, and then playing back the sound file to visitors over a standard pair of headphones. Now, strictly speaking, because of the fact that every person has a unique combination of head, torso and pinna topologies, HRTFs are as unique as fingerprints. Fortunately, individual differences in these structures appear to have only a second order effect, and a 3D file recorded with the ear-buds in one person’s ears still had a convincing 3D effect when played back over headphones to another person. In practice, there are problems with having a participant wear a pair of 3D recording ear-buds during a meeting. For instance, if that person were to turn his head, the effect for remote participants would be that the sound source would appear to move when it shouldn’t be moving. One possible alternative would be to use an anthropomorphic substitute for the token at the local meeting. A solution might be a KEMAR (Knowles Electronic Manikin for Acoustic Research). This device is comprised of a head and torso with microphones inserted in the ears which acoustically simulates a human body. KEMARs are fairly expensive; however, there are lower cost alternatives. To take this concept one step further, a pair of webcams might be inserted where the KEMAR’s eyes would be so that to provide a 3D view. In this way, a “virtual seat” at a meeting would exist and is the audio and visual perspective point from where the remote participants would experience the meeting.

3. 3D Sound as Data Visualization

As mentioned above, adding sound to data visualizations can serve to enhance understanding of the data [5]. In order to add sounds that seems to be attached to the 3D visual objects, three things are needed:

1. A method to encode a sound with the participant’s HRTF.
2. A way to transmit this sound to the participants’ ears (i.e., headphones).

A mechanism for tracking the participant’s head movements to adjust the virtual sound source’s apparent location relative to the participants’ head position and location is needed. Without the head tracking, creating the illusion that a virtual sound source is fixed in space would not be possible.
3.1 HRTF Encoding Hardware

Implementation of the HRTF in real time is a time consuming process. It is most easily implemented in Digital Signal Processing (DSP) hardware as a finite impulse response (FIR) filter with 200 taps. The tap weights are derived from HRTFs measured from actual people, databases of which are available online (http://interface.cipic.ucdavis.edu/). These datasets are taken from dozens of people with measurements taken for each ear at roughly 5° increments in azimuth and elevation angle. The CIPIC database consists of HRTFs recorded at each of 25 different azimuth and 50 elevation combinations, providing 1250 HRTFs per record.

For convenience, the coordinate system used for the sound source locations is a modified radial coordinate system – the radial axis passes sideways through the centers of the ears, its first coordinate (elevation) ranges from -90° (directly below), to 0° (directly in front), to +90° (directly above), to +180° (directly behind), to +270° (directly below). The other coordinate (azimuth) ranges from +90° (straight left), to 0° (medial plane), to -90° (straight right). The FIR filter tap weights derived from the HRTF database must be updated on the fly as the user’s head changes position.

The hardware selected for our implementation of the HRTF FIR is the Texas Instruments OMAP-L138 processor, a combination of an ARM9 processor and a C6748 floating point DSP processor. It has a touchscreen, many connectivity options, and the ARM9 can be programmed with any number of operating systems, including CE6.0, Linux or Android. The board can also be used without an operating system as a standalone DSP. It is small and portable, and can be battery-powered for portable use. Possible software development environments are numerous. Texas Instruments’ included tool, Code Composer Studio is very versatile. Figure 1 shows a twin-HRTF 256-tap FIR filter implementation using C6Flo, a Code Composer Studio plug-in. Another possibility is software developed by Educational DSP, LLC – winDSK8. This software has a GUI for easy learning and also provides a way to directly interface with MATLAB (Figure 2).

3.2 Head tracking

Transmission of sound is a relatively straightforward task and uses headphones. Headphones using wireless technology (i.e., infrared transmission) are fairly common. For tracking, we integrated faceAPI [4,5] with the 3D sound implementation [4,5]. Localized sounds are located in real-time using our implementation in realtime [5].

3.3 Parallax Correction for 3D Sound

The CIPIC HRTF database, compiled at U.C. Davis (http://interface.cipic.ucdavis.edu/), was generated by using a sound source at a fixed distance of 1 meter from the subject’s head. Attempts to simulate a source at some distance other than the reference distance by simply adjusting the sound volume does not work well as the sound didn’t seem to move closer or farther away, it simply sounded louder or quieter. It turns out the problem was caused by parallax error. If the sound source is at the reference distance of 1 meter, the FIR coefficients can be directly taken from the database entry corresponding to an azimuth angle of 0°. However, if the source is located much closer at point B, say 0.3 meters, the FIR coefficients for each ear must be taken from different azimuth value entries. In order to determine which set of FIR coefficients to use, one must calculate where the “sound ray” going from the ear to the sound locus intersects with the reference sphere. Equation of the sphere, simultaneously with the equation of the line, needs to be solved. Also, we need to find out the proportional distance to the sound locus compared to the distance to the reference sphere so that the inverse-distance law can be applied to the sound amplitude. Both of these can be found in a straightforward manner if we use a parametric representation for the equation of the line. These azimuth and elevation values can be calculated by finding the intersection of the unit sphere with the ray originating from the two points representing the left and right ear, and through sound source location, say point B. There could be two intersections in front and two intersections at the back representing the situation when the sound source B is in front, or at the back. The intersection points were found using the formulation of a vector and unit sphere in Cartesian coordinate space, which can then be converted to inter-aural polar coordinate system. This allows us to access the appropriate FIR coefficients from the HRTF database. Once the elevation and azimuth values are known, the values used to index into the HRTF database are selected by picking the closest angular value available in the database. Since the elevation values are evenly spaced by 5.625°, the closest value can be found by using a simple linear function, and then rounding. However, the azimuth values are not evenly spaced in the database, so a cubic function, found using curve-fitting technique, rounded the value to nearest integer azimuth value available. Finally, transformation was developed to adjust the apparent position of the sound relative to the listener’s head to accommodate head movement.

3. Implementation of 3D Sound Algorithm

Once the main loop is entered, it loops indefinitely until the Quit button is pressed. The procedures performed within the main loop are in [5], summarized below:

1. Retrieve the positions of the x, y and z sliders.
2. Get head pose data from the face-tracking engine.
3. Convert the position of the sound source from world coordinates to head coordinates.
4. Perform the geometric calculations (parametric sound ray, and reference sphere calculations described above) and determine azimuth and elevation indices to be used for extracting the appropriate HRTF filter coefficients for each ear, as well as the distance-dependent values for the gain. Because the inverse-distance law can cause extremely high volume levels for sound positions very close to the ears, the gain values are intentionally limited to a certain level.

5. Extract the impulse response (200 coefficients) for each ear and multiply each coefficient by the gain values.

6. Write the modified coefficient values to the memory addresses on the DSP board, which has previously been assigned to hold these variables.

7. The winDSK8 library function, called C8X_CONTROL_WriteFloat(), is used to send these values to the DSP board via the serial port. Since the PC did not have any physical serial ports, a USB to serial port converter was used. Serial port number and baud rate is set by the C8X_CONTROL__InitializeHostInterface() function.

8. On the DSP board, four baud rates are possible, which are set on the board by DIP switches S2-23: 115200, 230400, 460800 and 921600. The baud rate settings on the DSP board and in the PC program must match. Experimentally, 460800 baud was determined to be the maximum speed supported by the hardware (PC, USB-serial converter) in our case.

9. At the end of the loop there is a Sleep() function used to minimize unnecessary CPU activity, since updating the coefficients multiple times during the same video frame would be redundant.

There is a large gap in the HRTF database for sound positions in a wedge-shaped region beneath the subject due to physical limitations of the test apparatus. Because of this “out of bounds” region, there was a noticeable jump during our testing when the sound source moved from front to back beneath the horizontal. In our limited testing, using the headset, the sound appear to come from a 3D location when the head was moved as participant put their hand behind the head or in front indicating where the sound was coming from.

Also, in general, during our testing we found out that the HRTF function itself is fairly unreliable in the general downward direction, due to the fact that the torso plays a much more important role in the case where the sound is coming from below. There are few reliable and repeatable sound-influencing features on the torso compared to the pinnae, which have very well defined geometric features in fixed positions relative to the eardrums. The effects of the torso on the HRTF are very much dependent on the head orientation relative to the torso. Consequently, the brain doesn’t concern itself much with fixing the precise location of a sound source coming from a downward direction; rather, the head is turned downward and then uses the reliable mid-elevation HRTF functions to obtain a more precise simulation.

4. Conclusions and Future Work

Our motivation for developing 3D sound was for teleconferencing and 3D-location based sound applications. In addition to the above, HRTF hardware could be used to explicitly place 3D sound objects in virtual environments. The real-time and interactive implementation using DSP approach for the sound localization is both effective and inexpensive. The addition of head tracking capability dramatically enhances the believability of the virtual 3D sound source. The DSP processor can be dedicated to performing the computationally intensive task of implementing the two 200th order FIR filters, freeing the CPU to perform head tracking and other functions. Although the system is somewhat portable as is, it could be compacted down to truly portable size to embed in a stereo headset in future. More studies and formal testing needs to be done before the system can be deployed. The possibilities for creative development in this area are endless, especially for novel applications for community with blind and visual impairment.

Acknowledgement(s)

Ms. Bonnie Snyder, Technology Consultant for the Blind and Visually Impaired in the State of Colorado, and Dr. Mark Wickert are gratefully acknowledged for thier motivation and guidance.

References

Figure 1: Code Composer Studio GUI

Figure 2. winDSK8 implementation of 256-tap FIR.