3-D Sound Rendering System

Anshul Mahajan, Electrical and Computer Engineering, Faculty of Engineering, 300093480
Gursharan Singh, Electrical and Computer Engineering, Faculty of Engineering, 300098491

***Abstract*—3-D sound systems are of great advancement nowadays. Several corporate and university research projects are going on to measure HRTF and utilize it for digital filtering applications. There is a huge role of Digital Signal Processing in Virtual Surround Sound. The generation of 3-D sound requires HRTF. This paper aims at the basics of sound localization to produce a 3-D sound by using some of the filtering concepts of DSP. A sound moving source in horizontal plane as well as vertical plane is implemented using both time and frequency domain. HRIR filter is designed using basic synthetic model of HRTFs. The concepts related to filter design are discussed and mathematical modelling is presented.**

***Index Terms*—3-D Sound, Azimuth, Digital Signal Processing (DSP), Elevation, Head Related Impulse Response (HRIR), Head Related Transfer Function (HRTF), Sound Localization.**

I. INTRODUCTION

THE humans are capable of locating sounds in three dimensions even with the presence of only two ears. To localize the sound source in space, brain and ears work together. This ability in humans is a necessity since vision is not effective in darkness and is limited to viewer's viewing angle whereas sound source can be identified in any direction and in the absence of light as well. [1]

The source localization is done by measuring different cues and then comparing them to produce difference cues. The difference in the time and intensity measured at different ears produced solely or combinedly because of the human anatomy or the auditory processing in the ear canal are captured as impulse responses which are termed as head related impulse response (HRIR). These HRIR's can be used to create 3-D surround sound. [2][3]

The set of HRIR's is called the HRTF (head-related transfer function) which is the Fourier transform of HRIR. The sound propagating is altered by diffraction caused by head, shoulders and torso as it strikes the listener. The size and shape of head, density of head, ear canals, all transform the sound, boosting some frequencies and attenuating others and affect the way it is perceived. A couple of HRTF's for two ears can be utilized to make binaural sound that seems to originate from a specific point. This transfer function is used in some home entertainment products such as two speaker headphones. [1][4]

II. SOUND LOCALIZATION

The process of finding the spatial position of the sound source is called sound localization. In broader sense, sound localization aims at stimulating a specific sound field, including the sounds, listeners and propagation medium and environment. The brain becomes able to find sound source location with the help of differences in intensity, and timing cues. [5][6]

A. Human Ear Localization Theory

The 3-dimensional positions – horizontal, vertical, and the distance are key terms to define localization of static sounds. For moving sounds, one more factor comes into consideration which is velocity of sound in addition to these 3-dimensional position factors. [7]

The horizontal angle (azimuth) is defined by the difference in arrival times between the ears. This difference is created by different amplitude of high frequency sounds and reflections from body parts like torso, shoulders and pinnae. [7] The loss of amplitude, high frequency and the ratio of direct signal to repeated signal contributes to difference cues. [7] In order to find the source of sound our brain, takes help of our head position which alters the intensity and spectral qualities of sound. Such minute differences originated account for interaural cues. [6]

B. ITD and IID

In 1907, Lord Rayleigh studied human head model theory for the sound localization and based on this theory he presented internal clue difference, and called it Duplex Theory. [8]

Human ears are on different spatial coordinates which accounts for differences in the source to ear distance for both ears. Between the two ears, time differences and intensity differences originate which are known as Interaural Time Difference (ITD) and Interaural Intensity Difference (IID) respectively.

In Fig. 1 we observe that without considering the source factor for source S1 or S2, there will be a delay in propagation of sound between the two ears which will produce ITD. Simultaneously, due to shadowing effect because of ears and head, high frequencies will generate IID. [9]

ITD – The auditory system observes ITD from phase delays and group delays at low and high frequencies respectively. In Fig. 1 the sound from the source S1 reaches left ear first and then to the right ear. Many experiments also show that ITD is related to the signal frequency f. [9]

IID – The Interaural Intensity Differences are dependent on frequency and vary with frequencies in the same manner, if

.

frequency increases IID increases and if frequency decreases IID decreases. In Fig. 1 the sound from the source S1 has higher intensity at left ear then the right ear due to the effect of head shadows to the right ear. [9]

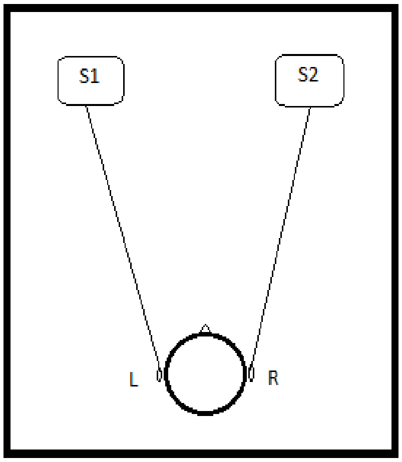

Fig. 1. Duplex Theory

III. HRTF

The process of obtaining HRTF for an individual includes playing analytic signal at a selected position and measuring the impulse response in the vicinity of the ear canal using probe microphones. This data is then stored and fed to DSP filters using simulation stage. Usually, simultaneous measurements are taken from both ears. This operation is performed for all selected positions and data is stored in time domain signal from left to right ears. A binaural impulse response measured is shown in Fig. 2 and its FFT would be equivalent to HRTF. [10]

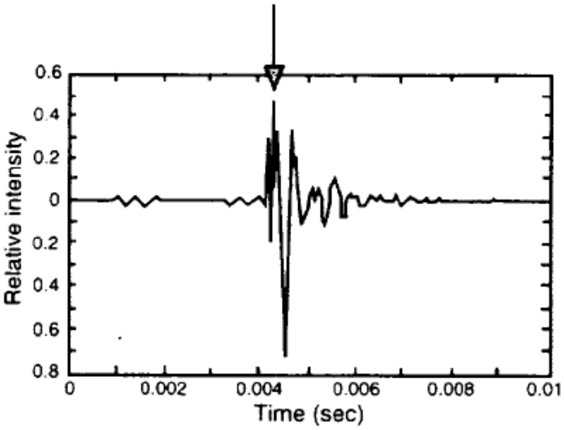

Fig. 2. A binaural impulse response

A. Collecting HRTF Measurements

There is no specific method of measuring HRTF. Research is ongoing for the improvement of the measurement hardware in the terms of signal-to-noise ratio and optimal positioning of the microphones. For the design of 3D sound systems, a database of HRTF measurements is required, from which data can be selected. Since all individuals vary from each other so taking the same HRTF for all applications is impractical. As an example, in 3D sound systems a generalized set of HRTF's that represent most common features of an individual is used. [10]

B. HRTF Equalization

The impulse responses measured are modified in time and frequency domain and then used in DSP. Simpler modifications take part in time domain whereas modifications in frequency domain are not that straightforward. It involves attempts to apply post equalization to HRTF in order to eliminate errors in measurements. For particular architecture of DSP, numerical formatting and conversion is also needed. [10]

IV. MATHEMATICAL MODELLING

We have to implement a 3D sound rendering system which can reproduce the effect of a sound source moving around the head of the listener. To produce a stereo sound file to be played through the headphones we must make use of the binaural cues and head related transfer functions (HRTFs).

This needs to be implemented in two scenarios using .wav file of 20 seconds which are originally recorded at 44.1Khz or 48Khz.

The first scenario being perceiving a sound source to be moving around the head of the listener in a circle in the horizontal plane at ear level from the front of the head; to the left; to the back; to the right; and to the front of the head; thus, following an anti-clock path around the head.

The second scenario being perceiving a sound source to be moving in a vertical plane around the head of the listener from the left of the head; to the top of the head (above); to the right of the head; thus, sound moving on a half-circle.

Both scenarios need to be implemented using some filtering methods. The first filtering method is 'Time domain convolution equation for FIR causal filters' which needs to be implemented using both the scenarios discussed above. This constitutes to the 'Part I' of the project. The second filtering method is 'Frequency domain FIR filtering using the "overlap-add" block processing method' which also needs to be implemented using both the scenarios discussed above. This constitutes to the 'Part II' of the project. In the third method, 'HRIR filters need to be designed from a basic synthetic model of HRTFs' which needs to be implemented using first scenario only. This constitutes to the 'Part III' of the project.

A. Part I

As mentioned before this part will use the 'Time domain convolution equation for FIR causal filters' to implement 3-D sound rendering. This needs to be done using an input .wav file which is originally recorded at 44.1Khz or 48Khz. The audio file used is of 30 seconds, but the number of samples taken is 882000 at 44100Hz. So, this represents that only 20 seconds of audio file being used. For HRTF, a dummy head HRTF database is used which is made available by the 'Center for Image Processing and Integrated Computing (CIPIC)'. The database used contains a.) horizontal-plane KEMAR HRIR data for large pinnae which is stored in two arrays of size 200 x 72, for the left ear as 'left', and for the right ear as 'right', b.)

frontal-plane KEMAR HRIR data for large pinnae which is stored in two arrays of size 200 x 99, for the left ear as 'left', and for the right ear as 'right'.

The audio file data needs to be convolved with the HRIR data in the time domain but before that audio file data needs to be divided into samples of 12250 samples per angle of 5 degrees and the HRIR data needed is given as 200 samples per angle. In simpler terms, first 12250 samples of the audio file data need to be convolved with the first column of the HRIR database. The next 12250 samples of audio file data need to be convolved with the second column of the HRIR database and the so on and the loop continuous like this saving the result of each convolutions in an array. This process is done for both the left HRIR database and the right HRIR database. Both the results of convolution are then stacked together to produce the moving sound source perception.

This can be further described in few simpler steps:

1. Take 882000 samples from audio signal, because the audio file is of 30 seconds.
2. Extract the left and right array from mat file.
3. Do this step for 'i' from 0 to 71:
    a) Take ith column of the left data
    b) Take audio data of size 12250, starting from i*12250 to (i+1) *12250.
    c) Do convolution of left data with audio data.
    d) Truncate convolved data of size 12250 from position 99 to 12348
    e) Append the convolved data with recent convolved data. If there is no recent convolved data, it will append it starting from 0th location.
4. Do the same substeps for right data as in step 3.
5. Pass the both convolved data sets through low pass filter to remove noise.
6. Stack the both filtered data sets vertically and make a .wav file.

The same needs to be done for the audio file in scenario 2 in which sound source perceives to move in the frontal-plane. The steps are same, but some changes are with audio data size and with 'i' range. The range for 'i' is from 0 to 98, and audio data size is 8909. Truncated convolved is from position 99 to 9007.

B. Part II

As mentioned before this part will use the 'Frequency domain FIR filtering using the "overlap and add" block processing method' to implement 3-D sound rendering. As discussed in Part I same audio file data and HRIR database is used with same number of samples. The difference being in the frequency domain is that zero padding needs to be done to the HRIR data and audio file data prior to their multiplication.

The length of the output signal is one less than the sum of both data lengths (12250+200-1). So,

1. The HRIR data (200) gets padded with 12249 zeroes and the Fast Fourier Transfer (FFT) is taken of this new padded data. This data set is now swapped.
2. The audio file data (12250) gets padded with 199 zeroes and the Fast Fourier Transfer (FFT) is taken of this new padded data. This data set is now swapped.

Now both these new data sets are multiplied to obtain an output data set. The resulting data set is again swapped to reconstruct the original data length. To convert the output which is in frequency domain into time domain, Inverse Fast Fourier Transfer (IFFT) is taken of the output.

For easier processing we use overlap-add method. This method provides smaller segments of long signals by breaking them into parts. When the filtered sections are overlapped and added to construct the output, the procedure is referred to as overlap-add method. This method constructs filtered output from filtered sections.

In DSP, overlap-add method is a main technique which consists of dividing the signal first and then processing each divided component separately and then recombining the signal to get the final signal.

As per in our program we used overlap-add method on the output obtained by the IFFT. Zero padding is done. The output is overridden with this new zero padded output. This can be represented in the Fig. 3.

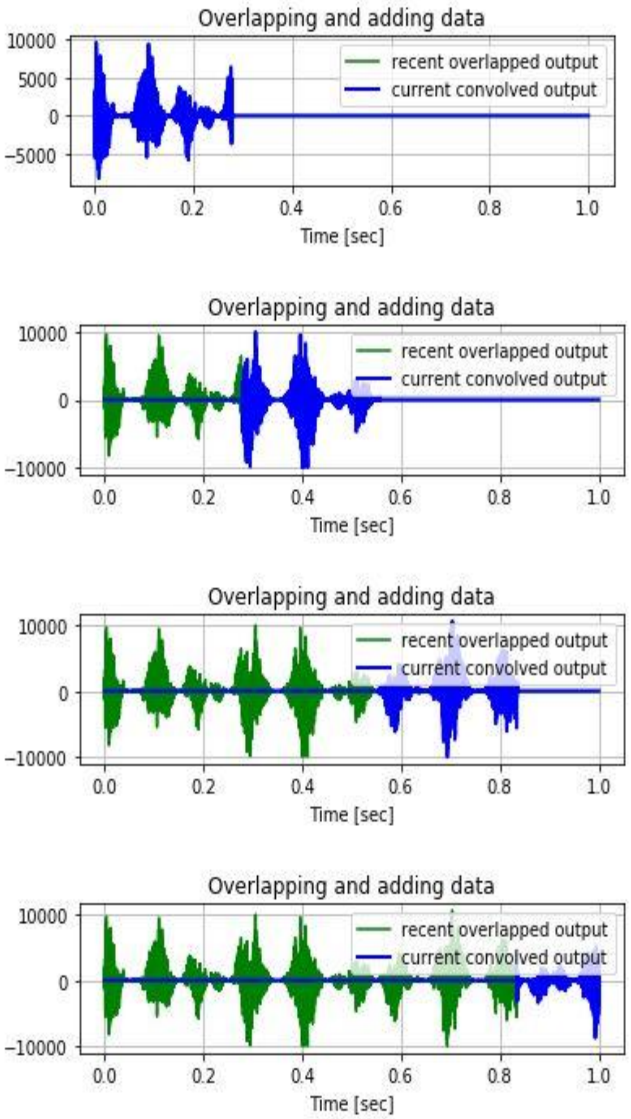

Fig. 3. The output data showing overlap-add properties which is observed for 1 second.

This process is done for both the left HRIR database and the right HRIR database. Both outputs are then stacked together to obtain the moving sound source perception.

The same needs to be done for the scenario 2 in which sound source perceives to move in the frontal-plane.

This can be further described in few simpler steps:

1. Take 882000 samples from audio signal, because the audio file is of 30 seconds.
2. Extract the left and right array from mat file.
3. Do this step for ‘i’ from 0 to 71:
   a) Take i^{th} column of the left data
   b) Zero pad the left data with 12249 zeros appended to make it of size 12449.
   c) Do fast fourier transfer of zero padded left data and then swap one half with the other
   d) Take audio data of size 12250, starting from i*12250 to (i+1) * 12250.
   e) Zero pad the audio data with 199 zeros appended to make it of size 12449.
   f) Do fast fourier transfer of zero padded audio data and then swap one half with the other
   g) Multiply zero padded left data with zero padded audio data.
   h) Swap one half with the other resulted from multiplication, and then do the inverse fast fourier transfer. This is convolved data.
   i) Take the recently convolved data and add it with above convolved data while taking care of position of both data sets as it is shown in fig. 3.
4. Do the same substeps for right data as in step 3.
5. Pass the both convolved data sets through low pass filter to remove noise.
6. Stack the both filtered data sets vertically and make a .wav file.

The same needs to be done for the audio file in scenario 2 in which sound source perceives to move in the frontal-plane. The steps are same, but some changes are with audio data size and with ’i’ range. The range for ‘i’ is from 0 to 98, and audio data size is 8909. Truncated convolved is from position 99 to 9007.

C. Part III

As mentioned before HRIR filters need to be designed in this part. In principle, it should be possible to figure out the HRTFs by explaining the wave equations, taking into considerations the effects generated by head, torso and pinnae. More than 100 years prior, Lord Rayleigh got low frequency approximation by getting a definite solution of diffraction of wave by rigid sphere problem. [11] Also with these solutions it becomes evident that a.) around 1Khz, the IID effects created by head shadow start and, b.) the variation of IID with azimuth is sinusoidal and rather complex with frequency. [12] More advancements to simplify this solution lead to a simple ray-tracing model which further gives ITD formula

$$\Delta T = 2\left(\frac{\alpha}{c}\right)\sin\theta$$

where, α = head radius, c = speed of sound. However, a better fit to the experiment can be given by using the formula, [13]

$$\Delta T = \left(\frac{\alpha}{c}\right)(\theta + \sin\theta)$$

A successful simpler model to obtain desired fit to Rayleigh’s solution was obtained:

$$H_R(\omega,\theta) = \frac{1 + j2\alpha\omega\tau}{1 + j\omega\tau} e^{-j\omega T_R}$$

$$H_L(\omega,\theta) = \frac{1+j2(1-\alpha)\omega\tau}{1+j\omega\tau} e^{-j\omega T_L}$$

where $\alpha = \frac{1}{2}(1 + \sin\theta)$, $\tau = \frac{1}{2}\left(\frac{\alpha}{c}\right)$, $T_R = (1-\alpha)\tau$ and $T_L = \alpha\tau$. Below frequencies of 2Khz, this model works well with Rayleigh’s solution. This filter is capable of producing synthetic moving binaural sounds in the azimuth plain but unfortunately not in the elevation plain. [4] This can be further described in few simpler steps:

1. Make two 2D arrays from the above equations, one representing the HRIR for left ear and other for right ear.
2. The output after implementing equation will then be convolved with audio files. These can be convolved in the same manner as it has been in time and frequency domain for azimuth.

V. Results

In time domain, the output sound rotates accurately. A little amount of noise can be heard. In frequency domain, by using the overlap and method, noise has been reduced when comparing to time domain results. However, while doing it using spherical head model, great amount of noise can be heard. After implementing low-pass filter, the noise has been reduced up to great extent.

VI. Conclusion

It can be seen that if enough data is provided of HRIR database, the output can be noise free. In addition to that, less computational cost could be achieved by doing it in frequency domain and results are better as well. A better spherical head model is needed to get smooth transition of sound. According to some researches the model works well for 2Khz signals. So, firstly down-sampling the audio signal and after convolving, up-sampling it might give better results. At the same time, the output audio produced will be smooth. Therefore, if the audio is above 2Khz, the audio signal should be convolved with HRIR database. Otherwise, the spherical head model can be considered.

References

[1] *Daniel Starch (1908), Perimetry of the localization of sound. State University of Iowa. p. 35 ff.*

[2] Begault, D.R. (1994), 3D sound for virtual reality and multimedia, AP Professional.

[3] So, R.H.Y., Leung, N.M., Braasch, J. and Leung, K.L. (2006) A low cost, Non-individualized surround sound system based upon head-related transfer functions. An Ergonomics study and prototype development, Applied Ergonomics, 37. pp. 695–706.

[4] Modeling Head Related Transfer Functions, Richard O. Duda, Department of Electrical Engineering. San Jose State University. San Jose. CA.

[5] Blauert, J, Spatial hearing: the psychophysics of human sound localization, MIT Press, Cambridge, Massachusetts (1983).

[6] Thompson, Daniel M. Understanding Audio: Getting the Most out of Your Project or Professional Recording Studio. Boston, MA: Berklee, 2005. Print.

[7] Roads, Curtis, The Computer Music Tutorial, Cambridge, MA: MIT, 2007, Print.

[8] Rayleigh L. XII. On our perception of sound direction[J]. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1907.

[9] Zhou X. Virtual reality technique[J]. Telecommunications Science, 1996.

[10] Durand R Begault, "Implementing 3-D Sound Systems, Sources, and Signal Processing" in *3-D Sound for Virtual Reality and Multimedia*, *Ames Research Center*, *Moffett Field, California*, August 2000, pp. 95–153.

[11] Rayleigh, J. W. S., The Theory of Sound (Macmillan, London, 1877), second edition republished by Dover Publications, NY.

[12] Kuhn, G. F., "A caustics and Measurements Pertaining to Directional Hearing," in Directional Hearing, W. A. Yost and G. Gourevitch, Eds., pp. 3-25.

[13] Mills, A. W., "Auditory Localization," in Foundations of Modern Auditory Theory, Vol. II (J. V. Tobias, Ed.), pp. 303-348 (Academic Press, NY, 1972).