Synthesizing 3D Audio (Binaural) from Monaural Moving Sources

Department of Computer Science and Engineering
University of Rajshahi

B.Sc. Engineering Project- Part II (2018)

Session: 2014-15

Name of the Student	Shakil Ahamed
Roll Number	1510876152
Name of the Supervisor	Dr. Md. Ekramul Hamid

Acknowledgment

This work would not have been possible without the contribution of C. Phillip Brown and Richard O. Duda, Fellow, IEEE who had written the original paper I worked on. I like to thank Julius O. Smith, Center for Computer Research in Music and Acoustics (CCRMA) for writing his excellent series on signal processing for audio applications and putting it on public domain.

I especially indebted to my supervisor Dr. Md. Ekramul Hamid, Professor, Computer Science and Engineering Department for providing me guidance and freedom on this project.

I like to thank all of my teachers and mentors from my childhood to this date whose wisdom shaped my thinking and for their amazing works which helped me to create vision of work like this.

Nobody has been more important to me in the pursuit of this project than the members of my family. I would like to thank my parents, whose love and guidance are with me in whatever I pursue. They are the ultimate role models.

Dedication

To the memory of **Johann Sebastian Bach** and **Ayub Bachchu**

Abstract

A structural model is implemented for synthesizing binaural sound from monaural moving sources. The implementation produces well-controlled vertical as well as horizontal effects. This work is based on the structural model of BrownDuda'98, CIPIC HRIR database, spline interpolation of points, Vector Based Amplitude Panning and time-domain description of physics of wave propagation and intensity damping. This implementation is very efficient hence very suitable for real time transformation. Additionally the CIPIC database contain many entry for different subjects which can be easily adopted for person of different head size hence it is very good for personalization. Experimental tests verify the perceptual effectiveness of this implementation.

Index terms - binaural, head-related transfer functions, localization, spatial hearing, 3D sound, virtual auditory space

Table of Content

Chapter 1 – Motivation

Overview

Monaural sound doesn't have any directional clue associated with them. For this reason we can't tell where the sound is coming from. This is a huge problem as we our brain uses this directional clue to locate the sound and instructs our eye to look there. The real world is not monaural, it is already 3D. So we do not face this difficulty in real world. But, if the sound is recorded, or synthesized, then the directional information are usually not added as the existing procedures were not cost effective as well as very difficult. Enough research has been done in this field which makes 3D synthesized sound a reality. In this project we will develop such a synthesizer and evaluate its applicability to reduce some of our disabilities.

Binaural Recording

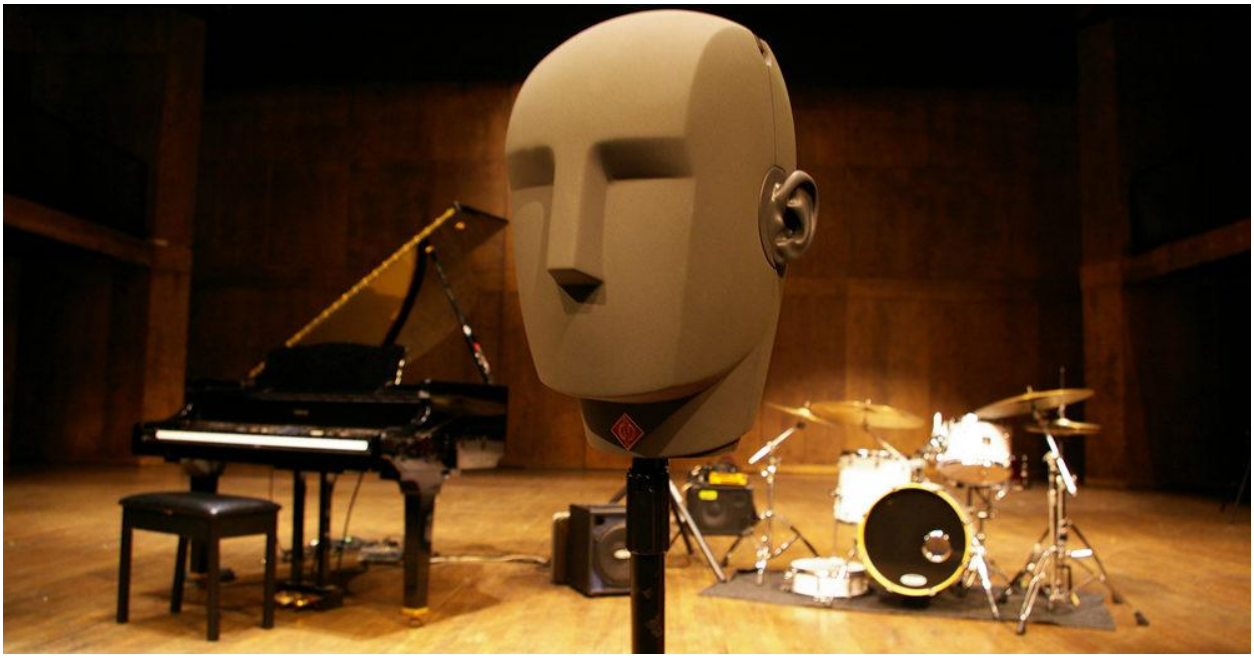

Fig: Binaural Recording Microphone

Binaural recording is a method of recording sound that uses two microphones, arranged with the intent to create a 3-D stereo sound sensation for the listener of actually being in the room with the performers or instruments. This effect is often created using a technique known as "dummy head recording", wherein a mannequin head is outfitted with a microphone in each ear. Binaural recording is intended for replay using headphones and will not translate properly over stereo speakers. This idea of a three dimensional or "internal" form of sound has also translated into useful advancement of technology in many things such as stethoscopes creating "in-head" acoustics and IMAX movies being able to create a three dimensional acoustic experience.

Sonar

Fig: Sonar Enabled Hand

Sonar (originally an acronym for sound navigation ranging) is a technique that uses sound propagation (usually underwater, as in submarine navigation) to navigate, communicate with or detect objects on or under the surface of the water, such as other vessels. Two types of technology share the name "sonar": passive sonar is essentially listening for the sound made by vessels; active sonar is emitting pulses of sounds and listening for echoes. Sonar may be used as a means of acoustic location and of measurement of the echo characteristics of "targets" in the water. Acoustic location in air was used before the introduction of radar. Sonar may also be used in air for robot navigation, and SODAR (an upward-looking in-air sonar) is used for atmospheric investigations. The term sonar is also used for the equipment used to generate and receive the sound. The acoustic frequencies used in sonar systems vary from very low (infrasonic) to extremely high (ultrasonic). The study of underwater sound is known as underwater acoustics or hydro acoustics.

Computer Music

A sound wave is a 1 dimensional signal. Music consists of various sound wave with different frequencies. The field of computer music analyze sound wave to find various information and synthesize sound that carries some certain characteristics The field of synthesis is so rich that we can generate millions of such sounds in real times.

3D Sound for 360 Vision

We can use advances of computer music in our favor. The blind persons can only detect objects having high reflection coefficient. We who are blessed with vision can also see only 120 degree. With a head mounted 360 camera, we can capture all the objects around us. It will create a visual chaos if we feed the picture through our eyes. Instead we can synthesize sound for important object and feed that through our ear. The problem is, real world is 3D where synthesized sound is not. With traditional methods we can know which objects exists around us without location information. This work is an attempt to synthesize sound that has location information associated with them.

Chapter 2 – Background Study

Interaural Time Difference

When a sound wave is generated, it spread out in all directions from it source. When it enters our inner ear, we perceive the sound. Before it enters our inner ear our head, outer ear (pinna) and body together transform it. The transformation occur differently for different horizontal location, vertical elevation and frequency. Our brain detects the transformation differences and can detect both intensity and location of the sound [1].

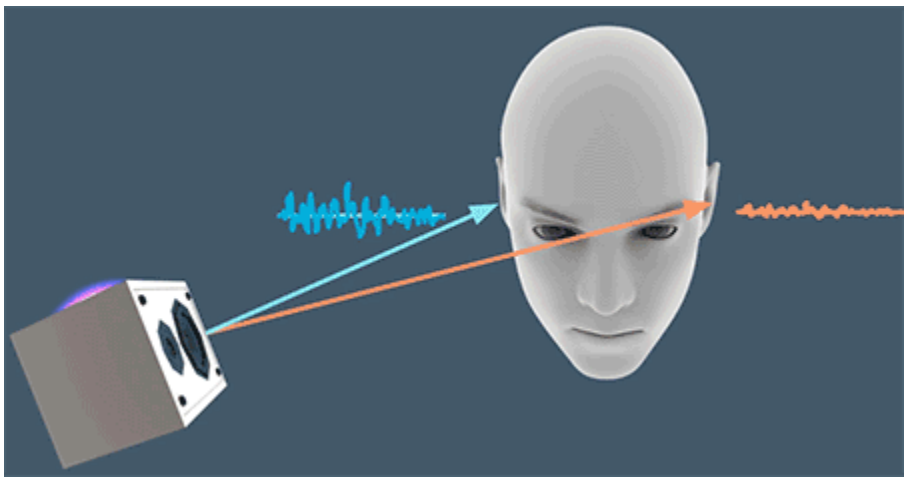

Fig: Interaural Time Difference

For different horizontal locations, the same sound wave arrives at different ear at different time. Our brain uses this little time difference clue to figure out the horizontal location of the sound source. This can be modeled very easily by thinking the head as a sphere and calculating the time differences of the sound for both ear coming from a certain direction. The sphere size depends on the head size of the person. For sphere with radius a and speed of sound in the medium c, for angle θ, the time difference is calculated by [2],

$$\Delta T(\theta) = \begin{cases} -\frac{a}{c}\cos\theta & \text{if } 0 \le |\theta| < \frac{\pi}{2} \\ \frac{a}{c}\left(|\theta| - \frac{\pi}{2}\right) & \text{if } \frac{\pi}{2} \le |\theta| < \pi \end{cases}.$$

Fig: Rayleigh's Spherical Model ITD approximation

Interaural Level Difference

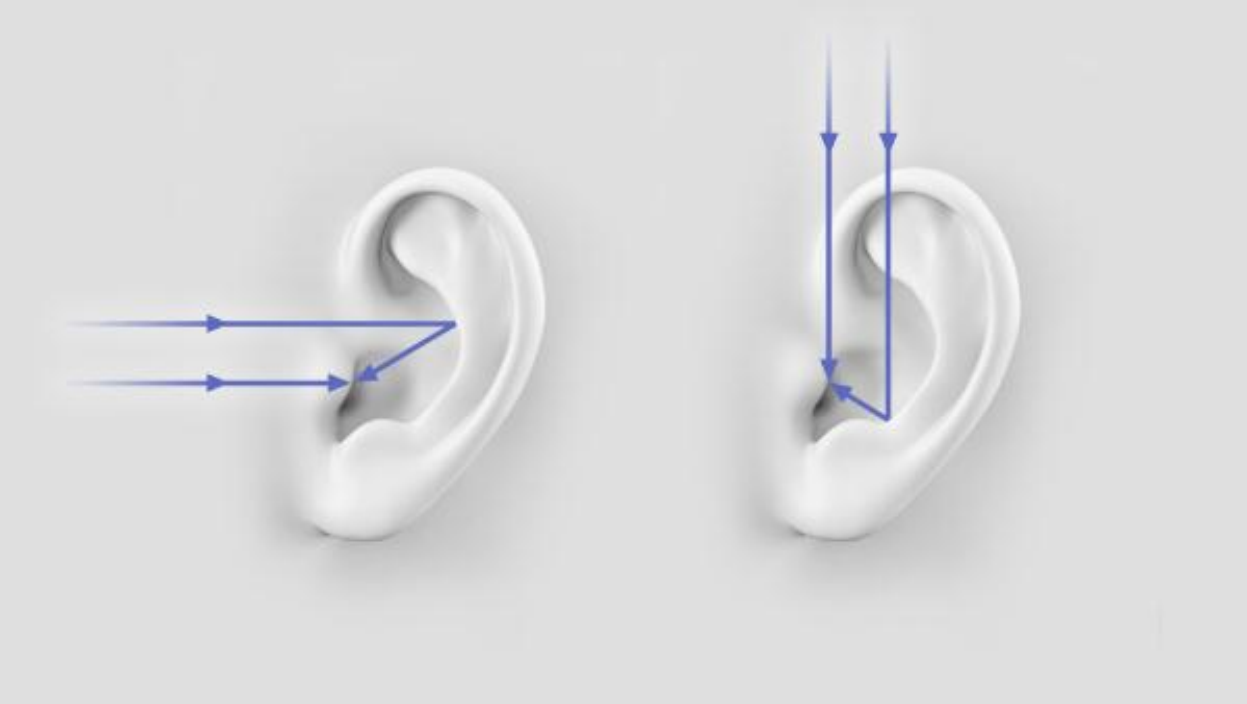

Fig: Interaural Level Difference

The ear is a unique feature of human. It is designed in such a way that it processes differently elevated sound differently. It has great effect on the elevation of a sound sources. There are various model designed to take this into account [4].

Head Shadow Filter

The head blocks and diffracts different frequency differently. This phenomena is known as head-shadow. The head shadow allow us to add some elevation information to the sound. The diffraction is very complex method. But there is a simple approximation which is very close to the actual result. It's also based on the spherical model. For radian frequency w and w0 = w/c, the head shadow filter is given by [2],

$$H_{HS}(\omega, \theta) = \frac{1 + j\frac{\alpha\omega}{2\omega_0}}{1 + j\frac{\omega}{2\omega_0}}, \quad 0 \leq \alpha(\theta) \leq 2$$

$$\alpha(\theta) = \left(1 + \frac{\alpha_{\min}}{2}\right) + \left(1 - \frac{\alpha_{\min}}{2}\right)\cos\left(\frac{\theta}{\theta_{\min}} 180°\right)$$

Fig: Head shadow filter

There are also various other methods exist for calculating ITD and ILD effects. One such method is to record a sound from the inner part of ones ear for different frequency and add then calculate the transfer function by deconvolution [2][3].

CIPIC HRIR Database

The CIPIC HRTF Database is a public-domain database of high-spatial-resolution HRTF measurements for 45 different subjects, including the KEMAR mannequin with both small and large pinnae.

The database includes 2,500 measurements of head-related impulse responses for each subject. These “standard” measurements were recorded at 25 different interaural-polar azimuths and 50 different interaural-polar elevations. Additional “special” measurements of the KEMAR manikin were made for the frontal and horizontal planes.

This can be used as replacement for all the ITD, ILD and Head Shadow filters. The database is in the public domain and each HRIR is only 200 samples long making it a very desirable FIR filter.

Ambisonics

Fig: Ambisonics

To produce 3D sound movie theatre uses many loudspeaker is different directions. Sound that should be coming from a certain direction are played in the corresponding speaker [5]. That creates an illusion of being in the scene instead of being a 3rd party observer.

Sound Wave Propagation

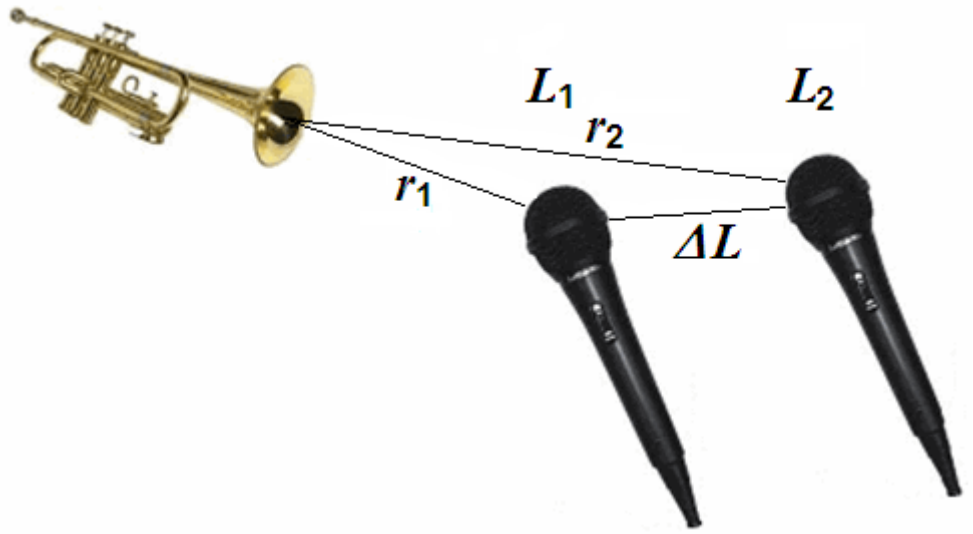

Fig: Sound Damping for Distance

As sound waves propagate through Air, the wave loses intensity. The loss is 6 dB for every time the distance doubles.

Sound level L and Distance r

$$L_2 = L_1 - \left|20 \cdot \log\left(\frac{r_1}{r_2}\right)\right| \qquad L_2 = L_1 - \left|10 \cdot \log\left(\frac{r_1}{r_2}\right)^2\right|$$

$$r_2 = r_1 \cdot 10^{\left(\frac{|L_1 - L_2|}{20}\right)} \qquad r_1 = \frac{r_2}{10^{\left(\frac{|L_1 - L_2|}{20}\right)}}$$

Fig: Level of damping of sound wave

3D Spline Interpolation

A cubic spline is a spline constructed of piecewise third-order polynomials which pass through a set of m control points. The second derivative of each polynomial is commonly set to zero at the endpoints, since this provides a boundary condition that completes the system of m-2 equations. This produces a so-called "natural" cubic spline and leads to a simple tridiagonal system which can be solved easily to give the coefficients of the polynomials. However, this choice is not the only one possible, and other boundary conditions can be used instead.

Chapter 3 – Requirement Analysis

The CIPIC database is in the public domain. So, it's available for free for everyone. The output requires a binaural headphone to be used. Other thing needed is a DSP enabled computational device.

The time requirement is a bit high as there are lots of material to be reviewed, but we can manage it in 6 month period easily.

The extended project requires some hardware manufacturing, but for this work, we are safe with hardware that we won at home.

Chapter 4 – Methodology

Moving Source Position

The moving source path can be defined with some key points. At various time of the sound wave, we define where the source was by defining azimuth, elevation and distance relative to the listener.

The missing points can be easily found through interpolation. We can use the 3D spline interpolation for smooth transition of the source from point to point.

CIPIC HRIR Subject 21

The subject is a large pinnae person which approximate many peoples. The choice is not hard since this can be easily changeable and adaptable for people with different head size. For this particular demonstration, we used subject 21.

Frame by Frame Analysis

As the audio sources moves, we need to filter it with different HRIR values. So, we can't use the whole audio to be processed at the same time. Hence we can process the audio chunk by chunk.

We can use Blackman window to select the chunk with 50% overlap which will ensure no power or spectral leakage.

VBAP HRIR Interpolation

The HRIR is measured at 2500 locations. Hence, we can't use CIPIC HRIR directly to filter all the sound positions.

The Ambisonics suggests a different alternative. We can treat these 2500 locations as 2500 speakers around our head. To play sound on a certain location, we can turn 4 speakers around these location with different gain factor that sums to 1.

The gain can easily approximate with spherical distance of the virtual speakers and by using 1/(1+r^2) to calculate gain ratio for the speakers.

FIR Filtering

The DSP enables system should have some efficient way to filter the current frame with the corresponding interpolated HRIR.

Sound Energy Damping

Some mathematical calculations shows that sound intensity changes with I * r1 / r1 formula. As CIPIC is measured from 1m, the damp would be I/r.

Chapter 5 – Implementation

Moving Source Position

For each sound source, create a key frame information database

```
% the keyframes contains information about location about some
% specific times. here, the entire length is divided in range [0-1]
% azimuth and elevation is in degree with interaural-polar coordinate
% system. distance is in meter.
% the missing points will be inter/extrapolate
keypoint = [0 .1 .2 .3 .4 .6 .7 .9 1];
azimuth_key = [65 -25 46 53 -36 62 53 -23 12];
elevation_key = [230 90 230 90 230 90 230 90 230];
distance_key = [2 3 4 10 5 6 3 2 5];
```

This data will be interpolated to create a continuous location information

```
% the frame will be played through a speaker which is placed at
% the middle location of the frame. we now want to know the point
% when it occurs.
frame_point = (i+frame_size/2)/sample_count;

% now find the actual location based on the frame point
[azimuth, elevation, distance] = interpolateLocation(keypoint,
azimuth_key,...
     elevation_key, distance_key, frame_point, 'spline');

% the interpolated location doesn't uses interaural polar coordinate.
% also, interpolated distance can be zero even negative. hence, a
% transformation is applied to get there principle values.
[azimuth, elevation, distance] = normalizeLocation(azimuth, elevation,
distance);
```

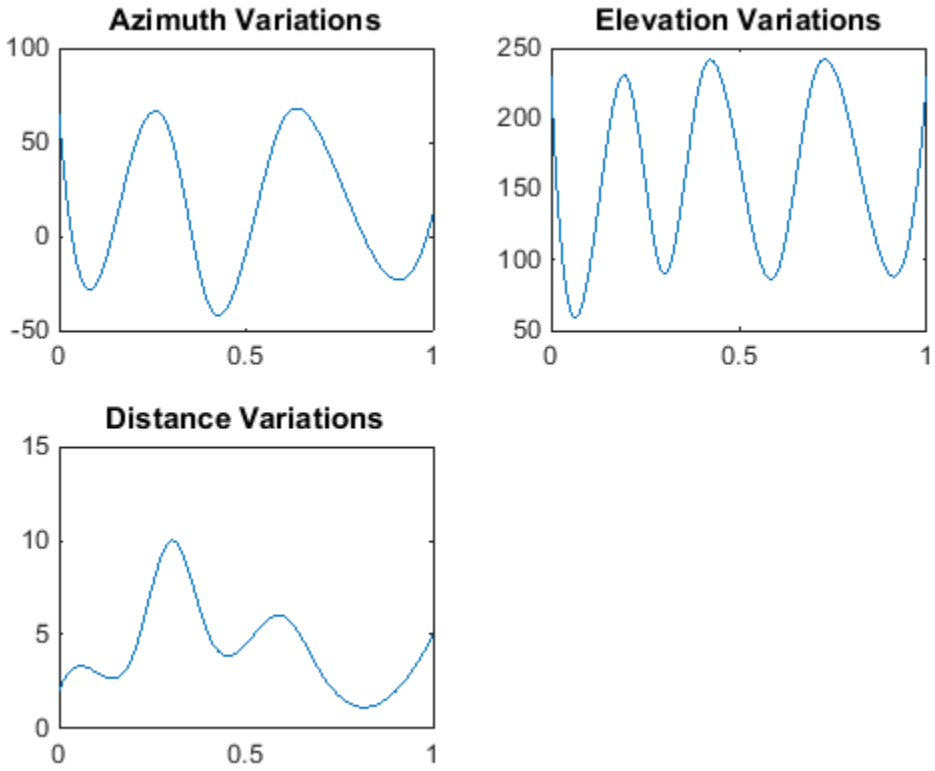

Fig: Interpolated key frame

CIPIC HRIR Subject 21

The CIPIC comes in Matlab .mat format files. Loading this data is easy. The azimuth is measured at unevenly spaced interval. The elevation is measured at 360/64 interval but at 50 points. HRIR also comes with several other structures which we don't need for the current application, hence released.

```
% load head related impulse response data
% this implementation uses CIPIC hrir database for hrir data,
% subject 165 and 21 was used for small and large pianae
% there are 25x50 location where HRIR was measured.
% the missing data will be interpolated
load 'hrir\large.mat';
azimuth_record = [-80 -65 -55 -45:5:45 55 65 80];
elevation_record = -45+5.625*(0:49);
distance_record = 1;
hrir_fs = 44100;

% there are some other measurement given, we dont need them.
clear OnL OnR name ITD;
```

Frame by Frame Analysis

We have tried several frame size, 0.3s was a clear winner in terms of informal hearing test. It reduces auditory artifact and still being local enough.

```
% frame by frame analysis
…
% the duration is critical. less frame duration means more spatial
% resolution. but, smaller the duration, larger the energy leakage.
% experimenting with few values show .3 is a good choice.
frame_duration = .3;
frame_size = fs*frame_duration;
```

The frame uses 50% overlap to prevent power loss as Blackman window drops at zero at the end and start of the frame.

```
% now the actual frame by frame filtering is done. to avoid spectral
% leakage we used blackman window with 50% overlap.
for i=1:floor(frame_size/2):sample_count
   j = min(i+frame_size-1, sample_count);

   % extract frame data and apply window
   frame = mono(i:j) .* blackman(j-i+1);
   …
end
```

VBAP HRIR interpolation

We divided the sphere into several quads. Then calculated where the current location exists. Then calculated gain for each vertex of that quad.

```
% the CIPIC HRIR is measured in 25*50 fixed locations. But we want to
% play the sound at some arbritary point. I used Vector Based Amplitude
% Panning technique. First, the sphere is divided into quads, then the
% quad where the point is located is found.
[ai, aj, ei, ej] = findQuad(azimuth, elevation, azimuth_record,
elevation_record);
```

The function findQuad is defined as

```
function [ai, aj, ei, ej] = findQuad(az, el, az_r, el_r)
% find four nearest points of azimuth and elevation

    % calculate azimuth locations
    if az <= -80
       ai = 1;
       aj = 1;
    elseif az >= 80
       ai = 25;
       aj = 25;
```

```
    else
        for i=1:24
           if az_r(i) <= az && az_r(i+1) >= az
              ai = i;
              aj = i+1;
           end
        end
    end

    % calculate elevation locations
    if el > 230.625
       ei = 1;
       ej = 50;
    else
        for i=1:49
           if el_r(i) <= el && el_r(i+1) >= el
              ei = i;
              ej = i+1;
           end
        end
    end
end
```

Now, calculate the VBAP gain

```
function [ai, aj, ei, ej] = findQuad(az, el, az_r, el_r)
% find four nearest points of azimuth and elevation
    % calculate azimuth locations
    if az <= -80
       ai = 1;
       aj = 1;
    elseif az >= 80
       ai = 25;
       aj = 25;
    else
        for i=1:24
           if az_r(i) <= az && az_r(i+1) >= az
              ai = i;
              aj = i+1;
           end
        end
    end

    % calculate elevation locations
    if el > 230.625
       ei = 1;
       ej = 50;
    else
        for i=1:49
           if el_r(i) <= el && el_r(i+1) >= el
              ei = i;
              ej = i+1;
           end
        end
    end
end
```

The gain can be calculated using

```
function gain = calculateVBAP(point, quad)
% calculate Vector Based Aplitude Panning gain
    gain = zeros(1, 4);
    for i=1:4
       gain(i) = 1000/(1+arcDistance(point, quad(i, :))^2);
    end
    gain = gain/sum(gain);
end
```

The gain values are gain linearly combined with the HRIR values to interpolate the location specific HRIR.

```
% each speakers HRIR is them linearly combined with their corresponing
% gain. This creates a composite HRIR both for left and right ear.
left_h = interpolateHRIR(hrir_l, gain, [ai aj ei ej]);
right_h = interpolateHRIR(hrir_r, gain, [ai aj ei ej]);
```

The method is simple,

```
function hrir = interpolateHRIR(hrir_in, gain, p)
% apply VBAP gain to combine the HRIR inputs
    hrir = zeros(200, 1);
    hrir = hrir + gain(1) * squeeze(hrir_in(p(1), p(3), :));
    hrir = hrir + gain(2) * squeeze(hrir_in(p(1), p(4), :));
    hrir = hrir + gain(3) * squeeze(hrir_in(p(2), p(3), :));
    hrir = hrir + gain(4) * squeeze(hrir_in(p(2), p(4), :));
end
```

FIR Filtering

The filtering is done via built in function,

```
% here the actual filtering wents.
left_out = filter(left_h, 1, frame);
right_out = filter(right_h, 1, frame);
```

Sound Energy Damping

There are various models exists for sound attenuation form distance. Here we ignored air humidity, temperature, viscosity and other factors. That given us simple model,

```
% sound loses enery as it travels. here the actual damping is done. this
% is a simplified technique. In actual case damping is dependent on
% frequency, air humidity etc.
left_out = damp(left_out, distance);
right_out = damp(right_out, distance);

function namp = damp(amp, dis)

% sound loses enery as spread through medium.
    namp = amp / dis;
end
```

Chapter 6 – Testing

Informal testing was on several subject without changing the HRIR for subject 21. All reported that they can observe azimuth changes at degrees resolution.

There are some confusion for small elevation changes. The elevation is clearly detectable from front, top, back and some other locations. The resolution is not very high. This is due to the fact that we didn't changes out HRIR to match person with different pinnae size. Also, we move out head all the time to detect elevation in real life. This project didn't implements head movement as it will require special motion enabled device.

The test for distance is also clearly visible as the amplitude goes up and down. To get better resolution, a better filter is to be used.

Chapter 7 – Discussion and Conclusion

It was a great experience to work with a project that taught me so many different things. DSP was a poorly understood idea during my course. By implementing this, I clearly saw the need to window and some other features of DSP which seemed useless.

The idea of the extended project is also very practical which will help a lot of people. With the help of this project, I am now more motivated to conquer bigger things.

References

[1] J. P. Blauert, Spatial Hearing, rev. ed. Cambridge, MA: MIT Press, 1997
[2] G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane,", J. Acoust. Soc. Amer.,vol. 62, pp. 157–167, July 1977.
[3] D. Hammershøi and H. Møller, "Sound transmission to and within the human ear canal," J. Acoust. Soc. Amer.,vol. 100, pp. 408–427, July 1996.
[4] D. Wright, J. H. Hebrank, and B. Wilson, "Pinna reflections as cues for localization," J. Acoust. Soc. Amer.,vol. 56, pp. 957–962, 1974.
[5] Producing 3D Audio in Ambisonics, Frank, Matthias & Zotter, Franz & Sontacchi, Alois, 2015
[6] A Structural Model for Binaural Sound Synthesis, C. Phillip Brown and Richard O. Duda, Fellow, IEEE

Shakil Ahamed, Roll: 1510876152 Computer Science and Engineering, University of Rajshahi	Dr. Md. Ekramul Hamid, M.Sc. (Raj), M.C.S (India), Ph.D. (Japan), Professor, Department of Computer Science and Engineering, University of Rajshahi