Binaural Rendering of Ambisonics B-format

Stage-I Project Report

Submitted by

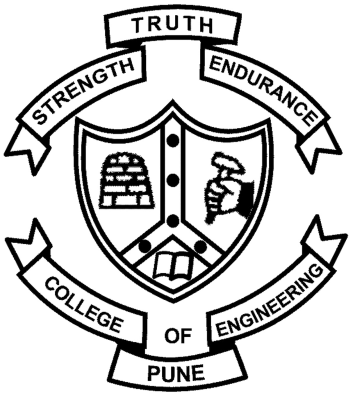
111408016 Anupam Godse

B.Tech Information Technology

Under the guidance of

Dr. V.K. Pachghare

COLLEGE OF ENGINEERING, PUNE



DEPARTMENT OF COMPUTER ENGINEERING AND INFORMATION TECHNOLOGY
COLLEGE OF ENGINEERING, PUNE-5

December, 2017

Contents

Abstract

Recently ambisonics format has gained popularity as directional/spatial audio encoding format for 360 degree videos, virtual reality, etc., with major video distribution platforms such as Youtube and Facebook adopting it for 360 degree videos. One of the most important characteristics of ambisonics is that it does not require the layout of speakers to be predefined for encoding. Rather the encoded representation can be decoded for any given speaker layout, which provides users flexibility to choose any layout of speakers and decode the given ambisonics representation for the same. The first order ambisonics encoding of a sound field requires four channels of audio stream and the directional information (spatialization) can be further improved by going for higher order ambisonics encoding with larger number of channels. Rendering spatial audio requires a large number of speakers (6 or 8 speakers for 5.1 or 7.1 surround respectively) placed in a specific way around the listener. All this hardware setup can be replaced with a headphone and an ambisonics to binaural rendering software. Binaural rendering is based on the concept of creating the effect of a virtual speaker on headphones using Head Related Transfer Function (HRTF). Thus ambisonics to binaural rendering can be achieved by assuming there is a an array of a large number of virtual speakers surrounding the listener and then decoding the ambisonics encoded audio for the speaker array. Next the audio from those speakers can be transformed into corresponding headphone experience by applying Head Related Transfer Functions (HRTFs) on outputs of virtual speakers. The aim of this project is to achieve this goal and give user immersive exeperience of 360 degree video with help of headphones. The target applications for this rendering technique is in virtual reality, 360 degree videos, high end gaming etc.

Chapter 1

Introduction

1.1 About Ambisonics

Ambisonics is a method of recording and reproducing audio in full 360 degree surround. Ambisonics treats an audio scene as a 360 degree sound sphere around center point coming from different directions. Center point is where the microphone is placed while recording, or where the listeners sweet spot is located while rendering. Traditional surround sound technology have several drawbacks. They only work on predefined array of sounds to produce the output sound field is the most important drawback of this technology.By contrast, Ambisonics dosen't render the audio signal for the predefined set of speakers but it can render audio on the fly for any user defined speaker array.

Here are few advantages of Ambisonics over traditional techniques:

1. Ambisonics not only works for static but also for rotating sound field i.e. it works for real time applications. When the sound field rotates the sound tends to jump from one speaker to another when used a traditional approach. Ambisonics uses number of virtual speakers so the transformation is smooth even when the sound field is rotated.

2. Traditional surround sound techniques are front biased but ambisonics distribute the sound evenly in 3D space.

3. Traditional techniques had difficulties in representing sound beyond the horizontal dimension. Whereas, Ambisonics works with the elevation as well and the

effect is more immersive.

1.2 Ambisonics B-format

B-format is widely used format for recording sound field using Ambisonics technique. It has 4 channels: W, X, Y and Z.
W : Omni directional sound pressure.
X : Front-Back direction with respect to the listener
Y : Left-Right direction with respect to the listener
Z : Up-Down direction with respect to the listener

1.3 Recording and Encoding B-format

Recording is done with the help of special sound field microphone. It has one omni-directional microphone (the W channel) and three figure-of-eight microphones (the X, Y and Z channels). It is made up of four cardioid capsules arranged in a tetrahedron, which can be combined as needed to provide the desired polar patterns.

1.4 Decoding

The decoders job is to produce loudspeaker signals that create a good illusion of the required directional sound field.[5] The Ambisonics format can be rendered on any speaker layout. It needs to be rendered on a virtual speaker layout (any) and then the output of each speaker needs to be filtered with Head Related Transfer Functions (HRTFs) (the HRIR(l) and HRIR(r)) from taking the HRIRs stored in the CIPIC HRTF Database to generate 2 channel (Left and Right) output i.e the Headphone output.

1.5 Binaural Rendering

Binaural rendering is converting the output of speakers to headphone output (binaural-left and right) by applying Head Related Transfer Functions (HRTFs)

1.6 HRIR and HRTF

HRIR (Head Related Impulse Response) Humans detect the sound source by taking derived cues from one ear and by comparing cues from both the ears. The cues have two differences one is time difference and another one is the intensity difference between cues of both ears. The sound source interaction with the human body modify the original sound before it enters the ear. These modifications can be portrayed with the help of the HRIR's, the head-related impulse response, which locates the source location. HRIRs help to convert the sound so that it appears to the user to be played at the desired location. They are used to generate virtual surround sound. The HRTF is the Fourier transform of HRIR. HRTFs for left and right ear describe the filtering of a sound source (x(t)) before it is perceived at the left and right ears as xL(t) and xR(t), respectively.

HRTF (Head Related Transfer function) is a response that characterizes how an ear

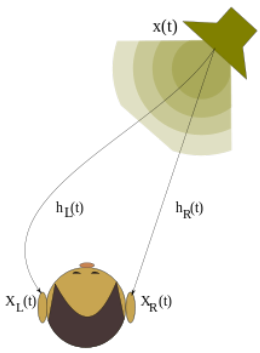


Figure 1.1: HRTF filtering effect

receives a sound from a point in space. A pair of HRTFs for two ears can be used to synthesize a binaural sound that seems to come from a particular point in space. It is a transfer function, describing how a sound from a specific point will arrive at the ear (generally at the outer end of the auditory canal).

1.7 CIPIC HRTF Database

The CIPIC HRTF Database is a public-domain database of high-spatial-resolution HRTF measurements for 45 different subjects. The database includes 2,500 measurements of HRIR for each subject. 25 different interaural-polar azimuths and 50 different interaural-polar elevations were considered for taking the measurements for each subject. [8]

1.8 System Overview

The following figure (figure 1.2) gives the system overview. It takes 4 channels as input (W, X, Y and Z) and generated a 2 channel i.e. binaural headphone output.

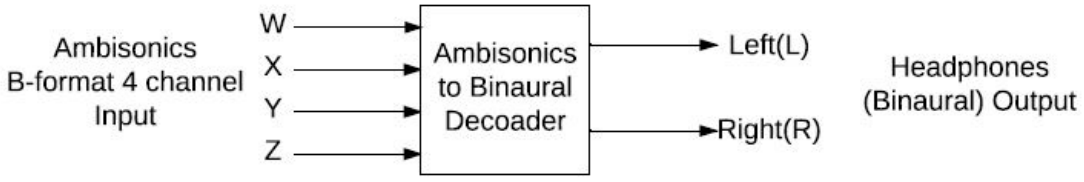


Figure 1.2: Ambisonics to Binaural Decoder System Overview

Inputs: 4 Channels
W : Omni directional sound pressure.
X : Front-Back direction with respect to the listener
Y : Left-Right direction with respect to the listener
Z : Up-Down direction with respect to the listener
Output: 2 Channels
Left(l) : Left headphone output
Right(r) : Right headphone output

Chapter 2

Literature Survey

2.1 Traditional approach

Michael Gerzon has criticized the traditional surround sound approaches and also has given the criteria for the design of the surround sound systems [1]. The traditional quadraphonic systems never gave the optimum results. The aim of these systems were to duplicate the effect of 'original' 4 track tape, but they failed to do so [3]. Peter Fellgett said that the existing techniques were inadequate in number of ways like they restricted to fixed number of speakers and the production needs of 4 channels to be available [3]. Moreover, these technique rely on encoding the speaker channel information which can be rendered on predefined speaker layout only - 5.1, 7.1 - otherwise doesn't give the intended effect. In addition, the traditional surround techniques were limited to horizontal plane excluding the height attribute. These techniques are only suitable when the image is stable and doesn't suite well for real time applications due to audio scene rotations and the output jumps from one speaker to another as there is fixed discrete predefined speaker layout. These existing approaches resulted in poor conditions even under ideal surroundings [3]. They suffered from 'hole in the middle' effect and if the situation is less ideal it becomes unusable. For example when the room is non-square or when the listener is not at the sweet spot [5]. The use of the 4th channel always degraded the localization quality, the mentioned 'hole in the middle' effect. Thus only 3 channels were recommended and the use of 4th channel was still a question. This led to the addition of the periphonic (height) information [3]. While traditional technique of surround sound had its limitations and disadvantages, Ambisonics, developed in the

early 1970s by Peter Fellget [3] and Michael Gerzon [4] is a way of recording and reproducing surround sound in both horizontal and vertical surround, which gave more immersive experience to the listener and provided full upward compatibility to any number of loudspeakers in the user defined configuration. The traditional approaches failed to give the intended immersive audio effects, they required significantly higher numebr of channels to improve the sound quality, they required the speaker layout to be predefined and needed the listener to be present at only a particular position. These were the disadvantages of the traditional approach.

2.2 An Ambisonics Approach

Monophonic reproduction merely provided information about direction and distance only. Then the stereo added explicit information for front sector not more than 60 degree in width [3]. Apart from this, various techniques were suggested by using more loudspeakers, more channels, extending the directional information beyond 60 degree. As these are seperate ways, Ambisonics aimed to combine these as an integrated whole [2]. To record, to convey and to regenerate the accurate and repeatable surround sound with the perfect directional effect was the main aim of the ambisonics technology [3]. It is the technology for surround sound which aims not to make the loudspeakers audible as seperate sources of sound [1, 2]. Ambisonics technique can be used with any more number of loudspeakers with reasonable configuration thus providing for full upward compatibility. Moreover, it is not limited to any number of channels, the more the number of channels the higher is the directional resolution [5]. The technique is based on a precise and unambiguous specification of how the encoding should handle directionality in contrast with quadraphonic approach which handled only 4 directional signals [5]. It defines encoding such that all the directionals are equally covered in contrast to the traditional techniques [5].

Why ambisonics is good because it covers 360 degree information of sound with limited number of channels. 4 channels (first order Ambisonics) can be rendered on 4 or more speakers with user defined speaker layout. Ambisonics, in contrast to traditional surround sound techniques, can create a smooth, continuous and stable sound field even when the sound field rotates and this is because it is not predefined for any particular speaker layout, thus suitable for real-time applications.

2.3 Virtual speakers array

Convincing binaural sound reproduction requires to filter the sound sources with the HRTFs. Moreover incorporation head-tracking further improvements in localization. Also increasing large number of virtual sound sources helps [6]. Using the finite number of speakers gives good approximation of the original sound field over a finite area.
Here is how to do the rendering over N virtual speakers:
If P is the vector denoting input to the sources, 1st order ambisonics B format is given as:
B = C * P
Now, as we already have B i.e. the Ambisonics B-format (W, X, Y, Z channels) and we need to regenerate P. There P can be calculated as:
P = pinv(C) * B;
Here C is the encoding matrix generated from the speaker configuration i.e. by considering azimuth and elevation of each speaker (each column represents one speaker) and pinv is the pseudo inverse.
Thus P matrix has the mono output signal for each loudspeaker.
Now the rendering of these signals over virtual speaker array is done. [6 7 10]

2.4 Development of Binaural System

After rendering over the virtual speaker array HRTFs are used to filter these signals of each speaker and converting mono to left and right signals for each speakers. When these all signals are superimposed we get a single left and single right (Binaural) headphone output signals. Further using head tracking to take into account the head rotation will further improve the effects [6 7 10].

Chapter 3

Problem Statement

3.1 Problem Definition

Design a system which takes ambisonics B-format (4 channels) as an input and generates output for headphones i.e. binaural output (2 channels). For real time applications, it should take information from the head tracking devices and generate the output accordingly.

3.2 Outcomes

1. The system is expected to work with minimum latency i.e. it should generate binaural output without any time delay, this ensures that it works for real time applications.
2. The headphones output should clearly able to distinguish between sounds from different directions.
3. It should consider head rotations and generate the output accordingly to get the best effect.

Chapter 4

System Design

The following figure (figure 4.1) explains the System Architecture Diagram. The system requires the ambisonics B-format file as input and then it requires virtual speaker array configuration so that the output can be rendered on the array of speakers. Later this mono output of each speaker is filtered with HRIR-l and HRIR-r filters to convert this into left and right output for headphones. This left and right filtered output of each speaker is now superimposed to get the final Binaural (headphones) output. The filtering is aided with the CIPIC HRTF Database which is a database having HRTFs for different subjects for different azimuths and elevations.

Explaning the basic blocks of the System Architecture Diagram:

1. **Generating the Decoder matrix:** The speker configuration (the azimuth and evevation for each speaker) will be taken as input and this function will generate a decoder matrix.

2. **Rendering output to virtual speaker array:** This function will take the 4 channels (W, X, Y and Z) and the decoder matrix as inputs and generate a mono output for each speaker of the speaker array.

3. **The Cipic HRTF Database:** This is the Database which has the HRIR pairs(left and right) for the range of azimuth and elevation pairs for each of the speakers.

4. **Adder:** This is a simple adder which will add all the outputs from each HRIR-l and HRIR-r filters and generate a single left and right final binaural audio.

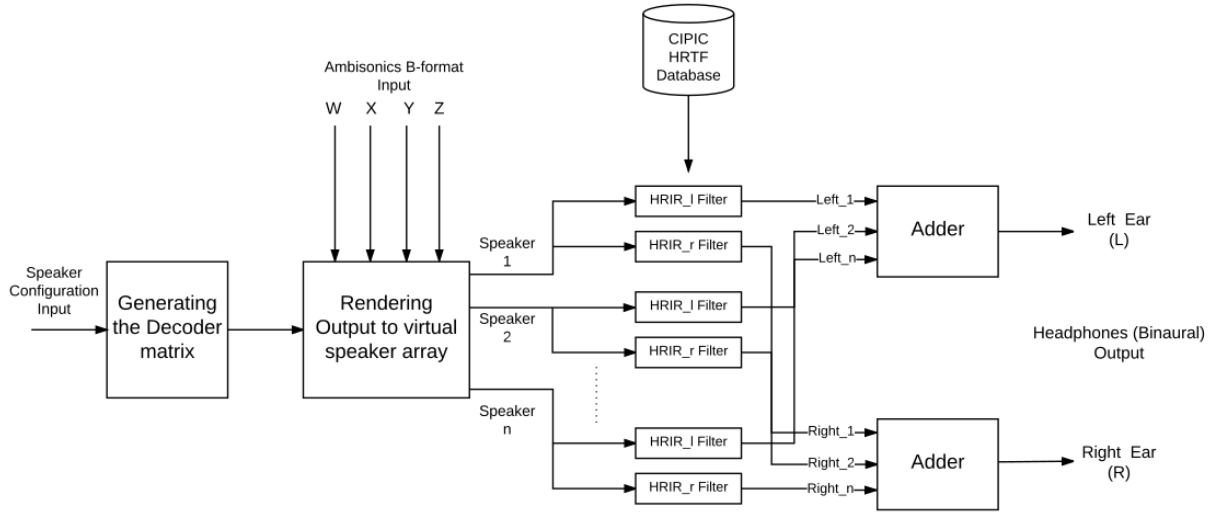


Figure 4.1: Ambisonics to Binaural Decoder System Architecture

Chapter 5

Requirements Specification

5.1 Functional Requirements

1. The system must take an ambisonics B-format (4 channel) file as input and give 2 channel output.
2. The head-tracking device inputs should be taken into consideration for real time applications and output should be generated accordingly.
3. It should work for 360 degree videos, real-time high-end gaming and other Virtual Reality applications.

5.2 Non-Functional Requirements

1. **Usability:** The system should be easy to integrate with any Virtual reality application.
2. **Performance:** The system should be efficient to provide the output by reducing the number of computations.
3. **Latency:** The latency must be very low to be suitable for real time applications.

5.3 Hardware Requirements

1. **RAM:** It should have minimum 2 GB RAM.
2. **Hard Disk:** It should have minimum 40 GB of free space.
3. **Head tracking device:** The user should have a head tracking device to connect to be system to give head rotation inputs for real time applications.
4. **Graphics Card:** System should have a sufficient graphics card to support the required Virtual Reality Application.

5.4 Software Requirements

1. **C++ Platform**
2. **MATLAB:** 5x or higher version
3. **CIPIC HRTF Database**

Conclusion

It is understood that indeed the ambisonics has many advantages over the traditional approaches. It can also be used for the real-time applications by applying the appropriate rotations over the matrices. It gives the better audio effects compared to the previously used approaches and thats why the technology is adopted by Facebook, Google and many other companies which work in Virtual Reality area. It has wide range of applications in 360-degree videos, high-end gaming and other virtual reality applications. Combining ambisonics technology with the virtual ambisonics approach to generate the binaural output has advantages of eliminating the hardware demand for loudspeakers and it can be used for mobile applications too.

Plan of Execution

1. Firstly, the function will be developed which will take input as virtual speaker configuration and will generate a decoding matrix from azimuth and elevation of each speaker.

2. Further, this decoding matrix will be used to convert the 4 channels of input ambisonics file into mono output for each speaker.

3. Now the HRIR filters needs to be extracted from the CIPIC HRTF database using MATLAB and that needs to be used to convert the mono audio output of each speaker into binaural output.

4. These, all signals need to be added and then the final left and right signal will be generated, which will basically be the signal for headphones(i.e the final binaural output).

5. This will work on static surroundings. To make it work for dynamic surroundings i.e for real time applications for ex. gaming, the input from head tracking device will be taken and the rotation matrix will be applied on the outputs to get the real-time immersive effect.

The following table (table 5.1) shows the expected timeline for the project.

Table 5.1: Project Timeline

Stage	**Expected Completion**
Finalizing Project Topic	Mid September
Literature Survey	Mid October
System Architecture	Mid November
Proof of Concept for static applications	November End
Working System for Static Applications	December End
Proof of Concept for Real-time Applications	Mid-January
Working System for Real-time Applications	Mid February
System Optimizations	February End

References

1. Michael Gerzon ‘Surround Sound Psychoacoustics’ Wireless World, December 1974, pp 483-6, www.ai.sri.com/ajh/ambisonics/wireless-world-gerzon-12-1974.pdf

2. Michael Gerzon ‘What’s Wrong with Quadraphonics?’ Studio Sound May 1974

3. Peter Fellgett. ’Ambisonics Part one: General system description’, August 1970, pp338-342

4. Gerzon, M.A. ’Ambisonics Part two: Studio Techniques’, STUDIO SOUND, August 1975. pp24 - 30, https://www.michaelgerzonphotos.org.uk/articles/Ambisonics

5. Gerzon M.A. ’Ambisonics in multichannel broadcasting and video.’ J. Audio Eng. Soc. 1985, 33 (11) pp.859-8711549-4950

6. Noisternig M., Musil T., Sontacchi A., Holdrich R. A 3D Real Time Rendering Engine for Binaural Sound Reproduction, Proceedings of the 2003 International Conference on Auditory Display, Boston, MA, USA, 6-9 july, 2003

7. Noisternig M., Musil T., Sontacchi A. Robert Holdrich, 3D Binaural Sound Reproduction using a Virtual Ambisonic Approach, VECIMS 2003 - lntemafional Symposium on Virtual Environments, Human∼Compuler Interfaces, and Measurement Systems Lugann, Switzerland, 27-29 July 2003, pp 174-178, IEEE, Washington (2003)

8. Cedric Yue, Teun de Planque, 3-D Ambisonics experience for Virtual Reality

9. Algazi V.R., Duda R.O., Thompson D.M., Avendano C. “The ClPlC HRTF Database”, in Proc IEEE Workhop on Applicalions of Six. Proc lo Audio and Elecfrmcou∼ficpsp. 9-102, NY, 2001

10. Noisternig M., Musil T., Sontacchi A., Holdrich R. A 3D Ambisonic based Binaural Sound Reproduction System, 2003