KATHMANDU UNIVERSITY
SCHOOL OF ENGINEERING
DEPARTMENT OF ELECTRICAL & ELECTRONICS ENGINEERING

PROJECT PROPOSAL

BINAURAL RECORDING

BY:

MUKUL BHATTA(31031)

POLARJ SAPKOTA(31053)

RISHAB RAWAL(31022)

SHREEJAN KARKI (31011)

SUPERVISOR:

MR. MADHAV PRASAD PANDEY

APRIL 2021

ABSTRACT

The way we experience sounds through our ears in everyday life is one of the hardest things to reproduce on a digital system. Early stereo recording techniques managed to localize the angle of the sound sources but not the distance, which is crucial for recording sounds in the way humans hear it. The techniques developed to bypass this limitation are arranged under the term 'Binaural Recording' which take a completely different approach to recording sounds. Modeling acoustic pressure as a function of several variables is currently the most widely employed and refined technique for such recording systems which is the focus of this project.

ACKNOWLEDGEMENT

We would like to thank the whole Department of Electrical & Electronics Engineering for their continued support towards our academic endeavors. We would like to extend our sincere gratitude towards Dr. Samundra Gurung, Mr. Santosh Parajuli, Mr. Anil Lamichhane, Mr. Nawaraj Mahato, Dr. Shailendra Kumar Jha and Dr. Ram Kaji Budathoki who gave valuable advice in regard to choosing an academic project, which ultimately led us to realization of the true essence of an engineering project and choosing this project as the 3rd project of our undergraduate career. We would like thank our project co-ordinator, Dr. Anup Thapa for reminding us of the value of teamwork on a semi-annual basis. Finally, we would like to thank our supervisor Mr. Madhav Prasad Pandey for his enthusiasm towards fundamentally guiding us to maintain an analytical approach towards engineering.

LIST OF FIGURES

ABBREVIATIONS

SN.	Abbreviation	Full Form	First Used in Page
1.	ITD	Inter-Aural Time Difference	6
2.	ILD	Inter-Aural Level Difference	6
3.	DSP	Digital Signal Processor	6
4.	VR	Virtual Reality	6
5.	3D	3-Dimensional	6
6.	UUV	Unmanned Underwater Vehicle	6
7.	HRTF	Head-Related Transfer Function	7
8.	LTI	Linear Time-Invariant	7
9.	VLF	Very-Low Frequency	12
10.	LF	Low Frequency	12
11.	EM	Electromagnetic	12
12.	ADC	Analog-to-Digital Converter	13

Table of Contents

1. BACKGROUND AND INTRODUCTION

Humans not only are capable of hearing sounds but also perceiving it. Our capabilities range from being able to accurately differentiate between different frequencies of sound to determining how loud or diffused a sound is. We can sense time delays between the original and the reflected sound to recognize echoes or reverberations. Using all these pieces of information collectively, we can almost accurately pinpoint the spatial location of the source of sound, which is called **sound localization**. This ability is facilitated by our enormous logical processing capabilities that perform processing tasks on sound such as: direction and frequency selective filtering, amplitude recognition, ITD(Inter-Aural[a] Time-Difference), ILD(Inter-Aural Level-Difference). As evident by ears being one of the main five sensory organs, our hearing system is inherently complex and replication of such a mechanism requires deductions of the process to first principles.

Binaural recording aims to accomplish the same thing an ear can do with the use of an array of microphones, pre-amplifiers and other signals conditioning electronic circuits coupled with signal processing techniques, mainly DSP (Digital Signal Processing). Use-cases of binaural recording techniques generally encompass the audio-industry, film-industry and PC/Console games industry. VR(Virtual Reality) audio spatialization, where binaural sounds recorded at certain conditions are replicated through speakers or headphones to create a virtual 3D(3-Dimensional) soundscape in video games, is one of the major applications of such a recording system. Despite most applications existing largely in the entertainment-industry, sound localization has it's applications in a variety of places. Sound localization technology, which is a byproduct of binaural recording is useful in accurately mapping out surfaces with a high level of accuracy depending on the accuracy of the used localization techniques. These are also used in special underwater communication-devices known as Acoustic Modems which can be used to map underwater surfaces or to establish communication with UUV(Unmanned Underwater Vehicles). Other two common applications are; recording songs in order to create a concert or studio-like experience for the viewer & sampling of sounds for reproduction in movie scores which is famously done for movies by companies such as Dolby Laboratories.

[a] Aural, latin for Ear has been widely used in literature pertaining to localization & spatialization of sound

2. PROBLEM DEFINITION

Recording sound binaurally firstly requires that it be localized in space. Before sound can be localized, it needs to be recorded. Recording will be simultaneously done by two microphones which replicate our two ears. The main approach to be taken in this project to localize sounds is to generate a corresponding HRTF (Head-Related Transfer Function). As inherent by the name, HRTFs are LTI (Linear-Time Invariant) systems and are therefore accessible to work with. HRTFs are distance selective filters that depend on a variety of other factors like environmental noise, reverberation, geometry of the recording room, among others. The function of HRTFs in our system is to characterize the pinna, which plays a key role in localizing sound in the human auditory system.

Successful build of a robust HRTF will provide us with a filter unsurprisingly known as a HRTF filter which will act as the main sub-system besides the microphone array itself. The received signal can now be processed with a variety of techniques in order to obtain desired results such as removing noise or extracting ambient sounds from the recording. Unfortunately, typical micro-controller units are not as capable of DSP, so dedicated DSP chips or a single-board computer needs to be used to accomplish those tasks with relative ease and speed. There exist microcontrollers from famous manufacturers that are capable of DSP which are planned to be bypassed as interfacing external memory with them presents yet another challenge which doesn't seem feasible given the allocated duration of the project.

3. LITERATURE REVIEW

Localization, most-widely and most successfully is done by determination of several variables as a function of Acoustic Pressure. This function of acoustic pressure is known as an HRTF, which has gained much fame in audio. HRTFs can be modelled as the following five-variable functions:

$$H_L = H_L(r, \theta, \varphi, \omega, \alpha) = P_L(r, \theta, \varphi, \omega, \alpha) / P_0(r, \omega)$$
$$H_R = H_R(r, \theta, \varphi, \omega, \alpha) = P_R(r, \theta, \varphi, \omega, \alpha) / P_0(r, \omega)$$

Where H_L & H_R represent respective HRTFs of the left and right ear, P_L & P_R represent amplitude of sound pressure waves at the entrance of both left and right ear canals, P_0 is the amplitude of the pressure wave at the centre of the head with the listener removed, L is left ear, R is right ear, r is distance between source and centre of the head, θ is source angular position, φ is elevation angle, ω is angular velocity, α is equivalent dimension of the head.

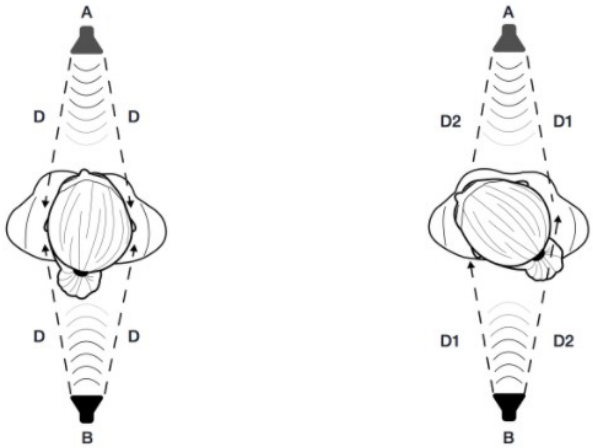

Fig 1: Front/Back ambiguity and it's resolving as observed from top of the listener

[Source: Oculus VR Audio Design Guide]

The figure above presents a problem that is resolved easily by the human ear. It shows two sound source, one at the back & one at the front, equidistant from the listener. Even when nonlinearity of the environment is taken into consideration, our ears will have a hard time localizing the sources of sound, because sound arrives at both ears simultaneously. How does the listener determine where one source is located? This perplexing problem has a simple solution, if the listener simply turns their head horizontally, it creates a time-difference between the arrival time of different sources

which then simplifies the problem and sound is localized easily. This simple action of turning the head to differentiate between the location of two ambiguous sources is executed using multiple microphones i.e. a microphone array. The microphone array is arranged such that each microphone sits at different variable heights and a fixed planar distance from each other. This allows only one diagonal plane of symmetry to exist in the whole 3D space where the ambiguity presented by Figure 1. occurs and if it were to occur, the array can be controlled such that whenever the ITD approaches zero, the microphones change their height by a small amount, rotating the plane of symmetry to a convenient location. Ambiguities as such of the above example are even more easily resolved when the HRTFs depend on variables other than the ITD as well. Further ambiguities such as motion parallax and loudness are also resolved by HRTFs.

With these basics in mind, the following other techniques are use to localize sound precisely:

1. Loudness - The louder the sound, the nearer it seems, the lesser the loudness the further we assume it to be. Example: music playing with low volume in the same room is perceived to be playing in another room or another house
2. Initial time delay - Interval between the direct sound and it's first reflection
3. Ration of direct sound to reverberant sound - When direct sound/reverb sound is small, the sound is assumed to be closer. When it is bigger, the sound is perceived to be farther.
4. Motion parallax - Sounds with small ITD are perceived to be nearer. Example of a loud superbike passing by a road 300m away seems to be nearer to us than it actually is. Another example is the airplane taking off/landing near an airport
5. HF Attenuation - HF sounds attenuate faster

How to create HRTFs?

Huge HRTF datasets are available online for use in prototyping. Creation of HRTFs require a mold in the shape of a human ear and a mannequin head closely resembling human head. After these tasks, obtaining a HRTF is a matter of making precise measurements in a controlled environment, interpolating, passing test signals, iterating and perfecting.

Devices well isolated from environmental effects prevent interfering HRTFs from affecting the sound experience. Bluetooth communication is not recommended while using listening devices as latency of up to 500ms is introduced.

Environmental Modeling

The 'shoebox' model can be used for environmental modeling. This model places the test subject inside a room to allow modeling of environmental effects caused by virtue of geometry such as reverb, early reflections and such. Environmental Modeling techniques for VR as the basis of precise HRTFs can be implemented in the following ways:

- Using volumetric source-based modeling
- Introducing the Doppler effect by increasing the freq. as sound source approaches and decrease the freq. as it leaves
- Delaying the time of arrival of the sound can seem intuitive in some cases such as the case when recording a thunder. We don't immediately hear the sound of the thunder after observing it. But popular media has made it the intuitive to most people that even long distance events are immediately audible.
- Avoid head-locked audio. Most music is recorded in stereo mixes and even VR applications use stereo playback devices but stereo doesn't spatialize the sound. It pans sound to a specific frequency so that only one ear perceives it well. But imagine the same thing in a game. **If you hear birds chirping from a certain direction, when you turn your head, you expect the bird's chirp to stay in the same location, not rotate along with your head.** This kills immersion. So mixing with ambisonics (spherical 3D sound field) gives better results compared to stereo when accurate spatialization of the localization is required.
- Make sure to keep latency under 100ms at all times.
- Other effects such as low-pass filtering can emulated sounds heard in an underwater setting as HFs attenuate rather quickly underwater

4. OBJECTIVES

1 Find optimal arrangement of a mic array.

2 Create a low-reverb, low-echo chamber.

3 Develop a HRTF filter.

4 Condition the received audio signals with signal processing software tools.

5 Store the recorded sounds.

6 Develop an audio processor to extract ambient sound from the recorded sound.

7 Encode the final sounds into compressed formats via third-party applications for demonstration purposes.

5. SIGNIFICANCE OF STUDY

The big-picture aim is creating a binaural ambient sound extractor, the first of it's kind as far as current literature is concerned. Ambient sound extraction finds itself useful in applications that aim to create virtual experiences as close to real-life as possible. Binaural ambient sound extractors make it possible to create high quality audio sampling systems. The main method to execute this lies within the primary objective i.e., locating the spatial position of sounds which then allows for a freedom of choice on which sounds to neglect and which to accept. Such a function opens up a large array of possibilities in setting up sound systems in movie halls, home-theatre systems as well as household stereo setups. It also gives VR experiences (which is the main targeted application) an added sense of realism. Some far reaching consequences might exist in the field of underwater modeling, which tried to employ VLF(Very-Low Frequency) and LF(Low Frequency) transmission techniques in the past. But since VLF and LF EM(Electro-magnetic) waves have huge wavelengths i.e. in the range of kilometers, the infrastructure required proved to be infeasible, therefore the idea was withdrawn. But, the same application with sound waves requires much less infrastructure and is therefore a feasible solution for modeling underwater surfaces. Besides this, binaural recording has implications in music, specifically in studio recording applications, as binaural methods provide far better quality compared to stereo sound mixing methods.

6. PROJECT METHODOLOGY

1 Thoroughly read through and document contents of accessible literature.
   - Search for a well documented VR company website and go through available articles.
   - Go through review papers on the topic and summarize.
   - Search for similar projects and research papers on the topic, read through and summarize.
   - Determine the best possible approach considering available tools and resources.
   - Study about common architectures, bit resolution & sampling rates needed to decide the right ADC (Analog-to-Digital Converter).
2 Develop the system block diagram.
3 Purchase mannequin dummy head
4 Determine the HRTF of the dummy in a minimum-noise environment.
   - Model multiple iterations of required HRTF in a recording studio.
   - Build a small chamber with uniform shape and smooth geometry.
   - Use basic soundproofing/reverb cancellation techniques & materials for the chamber.
   - Calculate optimal location for placement of a high quality speaker within the chamber.
   - Troubleshoot with test signals for the best impulse response
5 Process & condition the localized audio signals with MATLAB or Python's DSP tools
   - Fetch the recorded sounds directly to a laptop in .wav or other uncompressed lossless formats during testing phases.
   - Design required digital filters, mixers, etc and troubleshoot to make mic array capable of extraction of certain sounds, such as isolating and recording only the ambient sounds.
   - Implement the completed signal conditioning algorithms on a single board computer.
6 Design a dummy-replica of the human head with mic array as the ear.
   - Authorize & gain permission for data collection.
   - Collect inter-aural length data from willing volunteers.
   - Organize data and determine the best fit excluding outliers.
   - Compare the collected data with HRTF datasets from laboratories around the world which includes the IRCAM Listen [1], MIT KEMAR [2] & CPIC HRTF [3] databases and document the shortcomings & advantages of the collected data.

- Design at least two dummy heads according to the most feasible specifications covering two largest groups of people using non-reflective acoustic materials.

7 Store the final recording in compressed form.

- Store the final recording in lossless .wav or .flac format.
- Write a script to run a third-party conversion program to compress the .wav file to .mp3 or .aac formats.

System Block Diagram

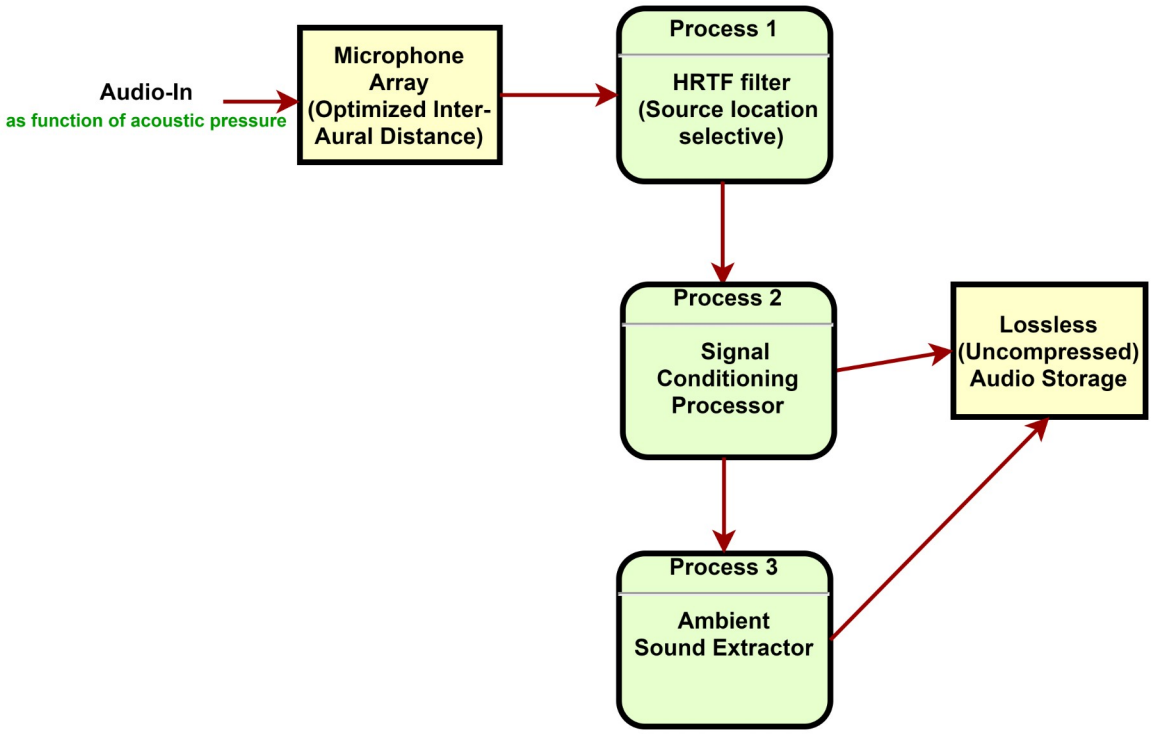

Fig 2: Binaural Recording Block Diagram Representation

Project Specifications

Materials & Components:

- Acoustic Energy Absorbing Material
- ADCs
- A Single-Board Computer: Asus Tinkerboard, Raspberry Pi, Arduino Mega 2560 etc.
- Two or more Microphones with good frequency response for the range 50Hz - 18kHz
- Rubber Molding tools and materials
- Mannequin Head

Min. dimensions of environmental modeling box i.e. the 'Shoe-box model': 1m x 0.75m x 0.5m

Budget

Microphones - Rs. 2000

Mannequin Head - Rs. 100

Acoustic Materials - Rs. 400

Other materials - Rs. 600

Single Board Computer - Rs. 2000

Total Estimated Budget: Rs. 5100

Note: All costs are approximations at best

Limitations

Most of the obvious limitations are expected to be brought forward by the damping capabilities of the inexpensive acoustic materials planned to be used. Other subtle limitations include, the ready-made mannequin head's dimensions not resembling a real person's head with the best accuracy even after modifications. Microphones to be used also are considerably inexpensive compared to studio quality condenser microphones. As such, the system won't be able to process sounds with frequencies above 18kHz. Prototype processing on a computer does not present any limitations although the final system implementation on single board computers might present some unforeseen challenges.

Project Timeline

Tasks	Mar/Apr	May/Jun	Jul/Aug	Sep/Oct	Nov/Dec	Jan/Feb
Literature Review						
Proposal Submission						
Prototype Design						
Mid-Term Report						
Final System Synthesis						
Final Report and Demo						

Completed Task

Remaining Task

7. EXPECTED RESULTS

The system is expected to be able to localize a sound field and appropriately record it. When played on a stereo speaker system, the recorded sounds are expected to spatialize accurately if the speakers' positions are configured according to the record's specifications in a geometrically uniform room. Any level of robust spatialization is not expected. Attempts at robustness might be undertaken in the future.

8. REFERENCES

1 Listen HRTF Database, L'Ircam, Institut de Recherche et Coordination Acoustique/Musique, Sept. 2002. [Online]. Available: http://recherche.ircam.fr/equipes/salles/listen/

2 B. Gardner, K. Martin, "HRTF Measurements of a KEMAR Dummy-Head Microphone ", MIT Media Lab, Cambridge, Massachusetts, USA, Technical Report #280, May 1994, [Online]. Available: https://sound.media.mit.edu/resources/KEMAR.html

3 V. R. Algazi, R. O. Duda, D. M. Thompson and C. Avendano, “The CIPIC HRTF Database,” Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics, pp. 99-102, Mohonk Mountain House, New Paltz, NY, Oct. 21-24, 2001, [Online]. Available: https://www.ece.ucdavis.edu/cipic/spatial-sound/hrtf-data/