Determining Personalized Head Related Transfer Functions using Auralization

Shubham Kumar
Dougherty Valley High School; San Ramon, United States
Email: dh.skumar@students.srvusd.net

Abstract - This project aims to explore the basis of an enhanced way to deliver spatial audio in an observer's horizon using acoustic modeling in place of low-pass filtering, interaural time differences, and simple attenuation functions. The approach described in this paper determines two interpolating functions, used as head related transfer functions ("HRTFs") which provide the illusion of spatial audio. The HRTF is obtained by solving the wave and Helmholtz partial differential equations. The parameters for the aforementioned equations can be adjusted to account for any source and observer orientation in any anechoic space. In addition, the shape of the observer's head can be modeled to create a personalized HRTF, tailored to the specific observer instead of a standard one. The model was able to successfully attenuate different frequency intervals and phase shift the audio signal for each channel to provide sufficient binaural cues to localize sound. Results from this implementation, although obtained through computationally intensive processes, could potentially be extended to virtual reality-based systems which would, in theory, provide for a more realistic audio experience, at a much lower cost.

Key Words – Acoustics, Psychoacoustics, Spatial Audio, Sound Localization, HRTFs.

INTRODUCTION

As virtual reality systems have matured past their infancy stages and serve as a viable medium to deliver immersive experiences, there has been an increase in research and development of such systems. Although much of the recent progress has been focused on enhancing the visual aspect, there is indeed an emphasis on refining the audio experience by using spatial audio and high-fidelity acoustics to provide for more realistic user experience [1, 2]. HRTFs are used in these systems to modify audio and recreate the acoustic filtering of sound as it propagates. When paired with headphones or any binaural audio delivery system, these functions can be used to simulate spatial audio by processing the audio that is delivered to each channel.

Since HRTFs are dependent on the geometry of a person's head, torso, and pinnae, they vary significantly among individuals. Because of the impracticality of determining customized HRTFs with existing experimental approaches, most virtual reality systems use generic or standard HRTFs. Since the aforementioned variations are left unaddressed with generic HRTFs, the user can experience unconvincing spatial audio and localization errors [3, 4]. This motivates the use of more accessible methods to determine HRTFs. The method described in this paper utilizes a numerical, simulation-based approach to deliver spatial audio in an observer's horizon.

BACKGROUND

Neuroscientists have been able to pinpoint the binaural and monaural cues that enable the human brain to accurately localize sound (determining a sound source's azimuth and elevation angles relative to the observer).

I. Horizon Localization

Specifically, they found that the brain determines interaural time differences ("ITDs") and interaural intensity differences ("IIDs") [5], as illustrated in Figure 1, to establish the azimuth of a sound source relative to the observer. ITDs describe the variation in sound arrival time for each ear and exist as a result of the difference in distance between the observer and source. IIDs describe the variation in amplitude and exist as a result of a sound wave attenuating after reaching a boundary such as an observer's head or traversing long distances.

II. Elevation Localization

The elevation of sound relative to the observer is determined by more complex monaural cues, specifically, spectral notches [6], which are closely related to head related impulse responses ("HRIRs"). These responses are dependent on several attributes including head diffraction, reflections from various body parts, and other refractions/diffractions due to the pinna. This phenomenon will not be covered in the scope of this paper, as the focus is to deliver spatial audio in the observer's two-dimensional horizon. However, a three-dimensional rendering of the

user's head could in theory be used in the simulation domain, effectively allowing for the HRTF to accept elevation data as a parameter. This would require including the third spatial dimension, namely elevation, into the models, which may influence the overall performance and introduce new sources of complexity to the process.

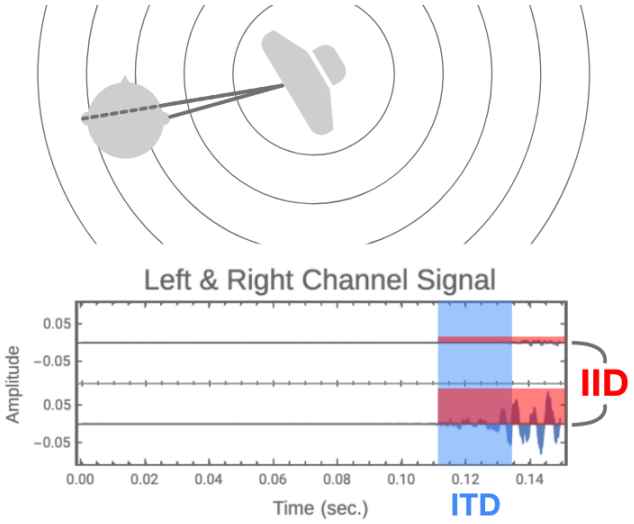

FIGURE 1: Certain arrangements of a sound source and observer yield subtle differences in the sound wave arrival time and amplitude for each ear. The signal for the left ear is described by the audio plot above and the signal for the right ear is described by the audio plot below. The areas shaded in blue and red on the graph highlight the ITD and IID respectively. Notice the delayed start and lower amplitude for the left ear signal relative to the right ear signal.

However, it's worth noting that these cues may not be completely unique for every source and observer arrangement, yielding the same signal produced for both ears despite the orientation being different. This results in a phenomenon known as the cone of confusion [7]. One can easily circumvent this problem by simply moving their head and take note of changes that may occur in the signal for each ear.

RELATED WORK

Since the brain relies on relatively simple parameters to localize sound, there has been significant progress over recent years to create standard HRTFs for spatial audio [8]. These HRTFs alter an audio signal to provide the illusion of spatial sound by processing two parameters, namely, the frequency of sound and the spatial location of the source relative to the observer.

However, many of the studies required investment in expensive resources including state-of-the-art anechoic chambers, high-quality speakers, and microphones to collect readings experimentally. Typically, these speakers are arranged spherically around a subject who has a pair of microphones inserted into their ear canals. These microphones collect readings of various tones which are then analyzed to design an HRTF. The measurement process takes up to one hour and the computational processes can take days. Additionally, this approach uses standard dummy heads to generalize the geometry of a human head, which yields a "one size fits all" HRTF. Hence, it fails to address the details that typically vary depending on the shape of an individual's head and ears. Although one investigation has shown that one can adapt to a new hearing profile, it doesn't make for the most realistic experience [9] since it takes time to adjust and often fails to capture the variances in perception from person to person. As a result, localization errors are produced and the generated audio is no longer capable of providing sufficient cues to enable the listener to correctly localize sound.

The process of creating personalized HRTFs for every single individual experimentally is infeasible. It would require the use of expensive acoustic equipment and computational resources which are generally inaccessible to the average consumer. Hence, the proposed method utilizes acoustic modeling to simulate the traditional measurement-based process. Ideally, the only equipment required would be a modern smartphone camera to recreate a three-dimensional rendering of a user's head. LIDAR mapping is another option that could be used, especially since there is an increasing number of handheld devices that integrate it into their optical systems. By simulating the experiment with a deterministic system, the costs and infrastructure associated with creating a custom HRTF would effectively be bypassed.

ACOUSTIC MODELING

Pressure changes in a medium, caused by wave propagation, are typically described using the wave equation in the time domain and the Helmholtz equation in the frequency domain. These differential equations enable us to probe pressure data at any spatial location and at any point in time or at any frequency in order to obtain attenuation coefficients. Typically HRTFs accept a spatial location and frequency as parameters, but examining wave propagation in the time domain helps visualize some of the behaviors that aren't immediately clear when visualized in the frequency domain. Moreover, the data visualizations in the time domain helped fine-tune some of the parameters that helped eradicate inconsistencies observed in the simulations and results.

I. Approach Overview

By initializing a sound source to emit the original audio signal or a relevant frequency and the head of an observer in the anechoic region as depicted in Figure 2, the pressure at the location of each ear can be probed to define the attenuation coefficients. Additionally, the phase shift can also be determined through a similar process, either through the model or a simple delay, based on the difference in distance from both ears. However, the latter is less computationally intensive.

II. *Defining Equations*

Variable symbols, definitions, and units are used as described in Table 1.

TABLE 1: Nomenclature as used in the following equations and expressions.

Symbol	Definition	Units
X	Position vector	$[m]$
t	Time	$[s]$
ρ	Density of a medium	$[kg/m^3]$
c	Speed of sound in a medium	$[m/s]$
p	Sound pressure	$[Pa]$
Ω	Simulation domain	$[m]$
ω	Sound wave angular frequency	$[rad/s]$
$Z_{\square}$	Specific Impedance	$[Pa * s/m]$
Z_b	Boundary Impedance	$[Pa * s/m]$
F	Dipole Source	$[1/s^2]$
Q	Monopole Source	$[1/s^2]$

The partial differential equations ("PDEs") used to model sound propagation and pressure distribution inside the space are the wave equation and the time-independent variant of the wave equation, the Helmholtz equation [10, 11]. The equations are defined as follows:

Neumann boundary conditions were used to dictate how sound waves interacted with walls and the observer's head. The head used an impedance boundary condition and the walls used an absorption boundary condition; each defined as follows respectively:

$$\nabla \cdot \left(-\frac{1}{\rho}\left(\nabla p(\boldsymbol{X}) + F\right)\right) - \frac{\omega^2}{\rho\, c^2}\, p(\boldsymbol{X}) = Q$$

$$\frac{1}{\rho\, c^2}\,\frac{\partial^2 p(t,\, \boldsymbol{X})}{\partial t^2} + \nabla \cdot \left(-\frac{1}{\rho}\left(\nabla p(t, \boldsymbol{X}) + F\right)\right) = Q$$

$$\hat{\boldsymbol{n}} \cdot \left(\frac{1}{\rho}\nabla p(\boldsymbol{X})\right) = -\frac{i\,\omega}{\rho\, c}\, p(\boldsymbol{X}) - \frac{R(r)}{\rho}\, p(\boldsymbol{X})$$

$$\hat{\boldsymbol{n}} \cdot \left(\frac{1}{\rho}\nabla p(\boldsymbol{X})\right) = -i\,\omega\,\frac{p(\boldsymbol{X})}{Z_b}$$

III. *Solving the PDEs*

The PDEs were solved using discretization using the finite element method for easy implementation on digital computers. The results returned were in the form of interpolating functions. The Helmholtz PDE was also parametric, with the parameter being frequency.

IV. *Visualizing and Interpreting Results*

By plotting the interpolating functions, many of the expected behaviors become apparent, as shown in Figure 2 (time domain) and Figure 3 (frequency domain). Sound propagation in the time domain occurs as anticipated and effects such as head shadowing are visible in the pressure distribution in the frequency domain. By solving the PDEs with discretization, the data is no longer continuous and hence yields certain anomalies in the data; however, these anomalies aren't too noticeable when it is applied to the audio.

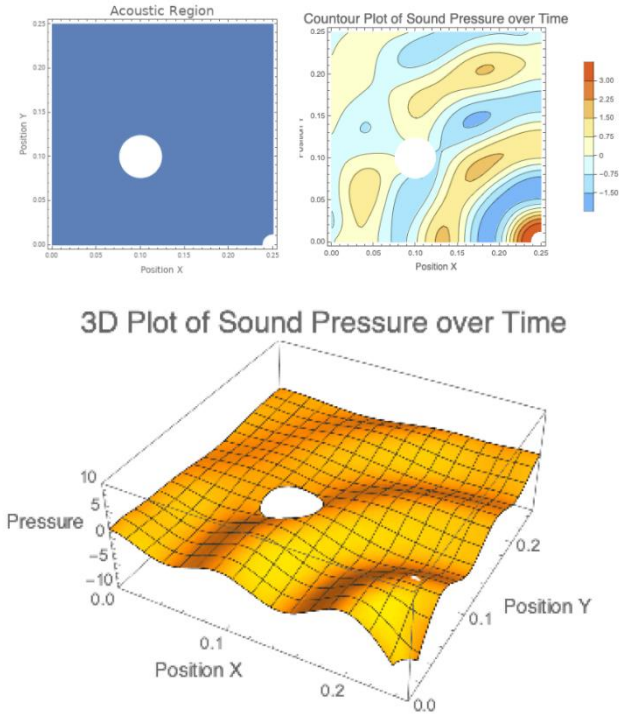

FIGURE 2: Graphs are identified one through three starting at the top left clockwise. Graph one shows the anechoic region with the large white circle representing the observer's head and the small white quarter circle at the bottom right representing the sound source, defined by a Neumann radiation boundary. A circle was used to model the head and ears instead of a more representative shape because of the lack of depth-sensing instruments. However, the models could indeed be adjusted to accept a more accurate geometric representation. Graph two and three display the pressure distribution at a single point in time as a contour and three-dimensional plot respectively; generated using the wave equation in the time domain.

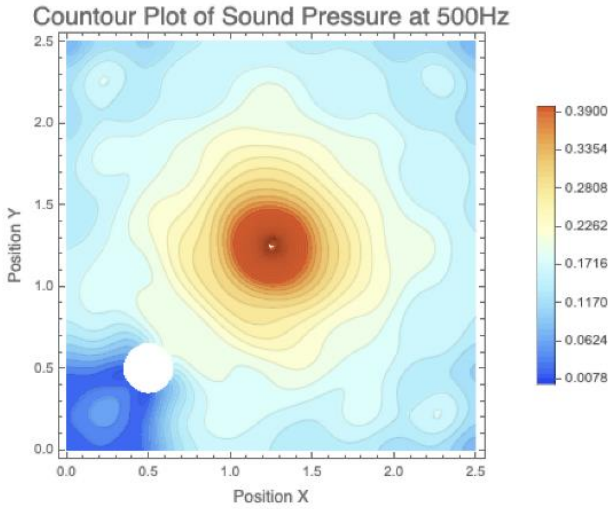

FIGURE 3: This contour plot depicts the pressure distribution in the anechoic region with a monopole sound source emitting a frequency of 500Hz and the head of an observer; generated by solving the Helmholtz equation in the frequency domain. Head shadowing is evident from the low pressure behind the head and the contour matches the predicted distribution.

AUDIO PROCESSING

Having determined an HRTF that accepts a frequency and a spatial location as parameters, the next step would be applying the functions to an audio input. The most effective way to do so is by using a Fast Fourier Transform and an Inverse Fast Fourier Transform pair along with a window function to fill in any discontinuities that appear while processing.

I. Fourier Transform

Decomposing the original audio into a sum of sinusoids using a Discrete Fourier Transform gives access to frequency intervals which can be attenuated using the parametric function obtained by solving the Helmholtz equation. Additionally, applying the Fourier Transform allows for informative visualizations using the spectrogram. The spectrogram helps visualize the changes in the amplitude for each frequency interval and the quality of the generated audio.

II. Attenuating Frequency Intervals

The spatial location of each ear is passed into the parametric function in addition to a frequency parameter. To expedite processing time, frequencies were inputted at intervals of 500Hz yielding a total of ten interpolating functions. While this didn't produce the most accurate results, it was the most computationally reasonable process. Obtaining an interpolating function from the parametric was a rather time-consuming operation.

Each interpolating function returns an attenuation coefficient for its specific frequency interval and is convolved with corresponding frequencies in the original audio data. Once each frequency interval has been attenuated successfully, the generated audio is recreated using an Inverse Fast Fourier Transform.

III. Hann Smoothing Window

Through this process, several artifacts are introduced into the generated audio and there is substantial data loss. To alleviate this problem, a Hanning (Hann) window was used to smooth discontinuities in the audio, preserve frequency resolution and reduce spectral leakage. Upon close examination of the spectrogram plots of the audio before and after applying the function, it is revealed that each frequency interval has been attenuated as expected, shown in Figure 4.

Other window functions including Blackman, Gaussian, and Poisson were used to experiment with different audio files. However, the Hann window was able to generalize to all tested audio inputs the best and outperformed the others in terms of quality.

RESULTS

As seen in Figure 4, attenuation appeared to have been performed as expected. When compared to spectrograms of experimentally determined HRTFs, such as MIT KEMAR, Figure 5, the results were similar, but there are a few notable shortcomings. The most prominent was the low-quality audio output after processing.

One reason could be the significant data loss even after applying a Hann window. While some discontinuities were eradicated, the overall audio still experienced a noticeable loss in quality. Additionally, the introduction of artifacts in certain frequency intervals still persisted. The investigation presented in this paper was conducted in its entirety using the Wolfram Language in Mathematica [12], which is known for being a high-level language and may have been the cause of quality loss. The aforementioned issues would likely be resolved if the entire process was translated to a lower level language with proprietary algorithm implementations. Not only would this make the entire process run at a reduced time complexity, but it would also enable more flexible experimentation since there would be no restriction to exclusively use the built-in functions offered by the Wolfram Function Repository.

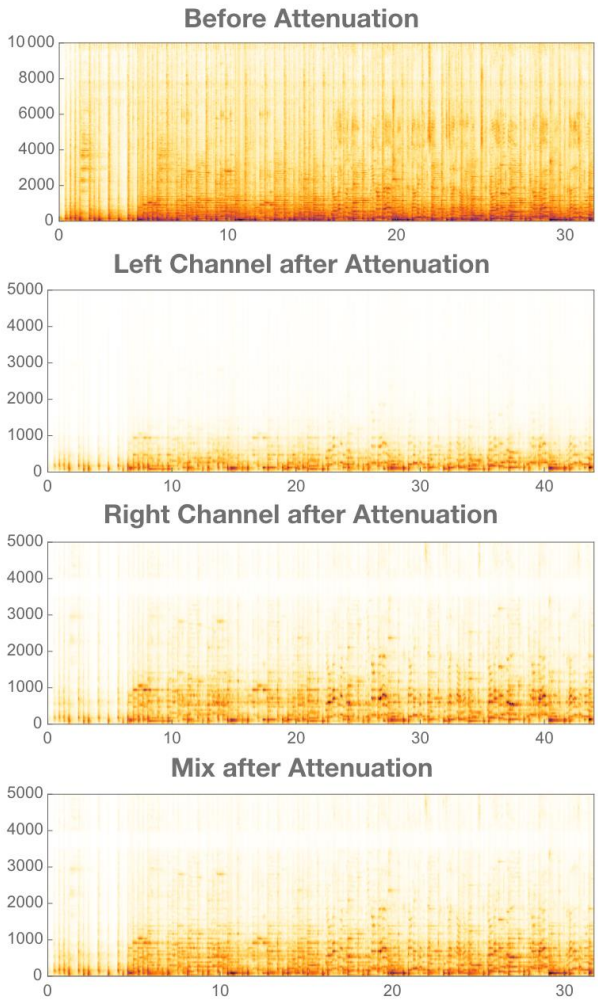

FIGURE 4: Spectrograms are identified one through four starting at the top. Spectrogram one depicts the original audio with data for nearly all audible frequencies. Spectrograms two and three show frequency data for both the left and right channels. Since the top end of the spectrogram became lighter, it is suggested that the higher frequencies are attenuated at a higher magnitude in comparison to the lower frequencies, as predicted. Moreover, the left channel experienced a more significant loss in higher frequencies in comparison to the right channel due to the orientation of the sound source and observer. The sound had to propagate through the observer's head to reach the left ear, hence preserving only the lower frequency sound. Spectrogram four z the final mix after combining both channels. Notice the change in scale along the frequency axis for the first spectrogram and the rest; The generated audio has no data for frequencies above 5000Hz. This is most likely attributed to a loss in quality after processing since experimentally measured HRTFs possessed data for those frequencies.

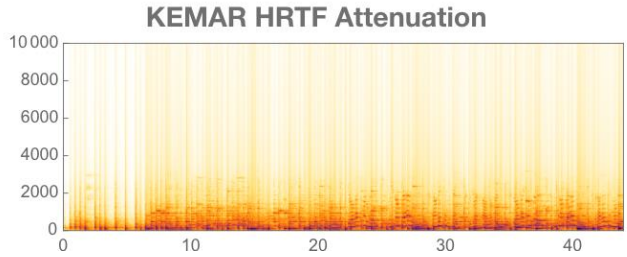

FIGURE 5: Spectrogram result after applying the KEMAR HRTF to the same audio input. Frequency data was possessed, although at lower amplitudes for up to 10000Hz, resulting in a fuller sounding result. Applying an audio normalization to the audio generated in this paper yields a similar amplitude distribution, however there is a noticeable loss in frequency.

Frequency interval ranges could have also been decreased to provide for a more immersive experience but would have come at the expense of processing time. This tweak would have required solving for more than the ten interpolating functions from the parametric function obtained from the Helmholtz equation. Hence, it wasn't feasible with the Wolfram Language. However, as mentioned earlier, lower language implementation may alleviate this problem. Note that all code from this project is stored in Github [13] and the preliminary work is documented in the Wolfram Community [14].

FUTURE WORK

This project, although far from perfect, serves as a strong underlying foundation for future improvement. The proposed method produced reassuring results and provided a clear path for future work. As suggested in the results, one of the next steps would include transcribing the Wolfram Language code into a slightly lower level language like Python that would still offer built-in methods and frameworks which would streamline the development process, while still allowing for flexibility in experimentation and customizability.

Additionally, introducing the third spatial dimension into the model and comparing the monaural spectral cues it produces with experimentally measured cues could be interesting. Another possible extension would be to create three-dimensional renderings of an actual head using LIDAR or another form of depth-sensing to truly produce a custom HRTF. Most importantly, however, there is a need for a concrete metric to quantify the performance and quality of results. As of now, the comparison consisted of looking at the phase shift for each channel and the amplitude differences. If the project is implemented to simulate a three-dimensional space, the current form of analyzing results won't be as effective since it would be difficult to closely examine the monaural cues through a spectrogram.

It may also be informative to investigate some of the behaviors of the model in an environment with obstacles and different types of boundaries. This may help gauge the accuracy and performance of the model. Another possible extension could include utilizing an accelerometer on the user's headset to determine their position relative to other objects in the simulated room and make changes to the audio in real-time. These ideas are indeed far-fetched, but may be possible to implement if the quality of the audio is improved and the third spatial dimension is incorporated into the existing model.

ACKNOWLEDGEMENTS

I would like to express my gratitude for Wolfram Research and their undeniably helpful staff who've managed to provide effective mentorship throughout the entire development of this project, especially amidst these uncertain times. The Wolfram summer program enabled me to put forward my best effort and ultimately produce some of my highest quality work. I would also like to thank my family and close friends for their continued support and commitment to keep me motivated.

REFERENCES

[1] D. R. Begault. 3-D Sound for Virtual Reality and Multi media. Academic Press, 2000.

[2] R. D. Shilling and B. Shinn-Cunningham. Virtual auditory displays. Technical report, DTIC Document, 2000.

[3] P. Larsson, A. Väljamäe, D. Västfjäll, A. Tajadura-Jiménez, and M. Kleiner. Auditory-induced presence in mixed reality environments and related technology. In The Engineering of Mixed Reality Systems, pages 143–163. Springer, 2010

[4] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. The Journal of the Acoustical Society of America, 94:111, 1993.

[5] Jeffress, L. A. (1948). A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, 41(1), 35-39. doi:10.1037/h0061495

[6] Butler, R. A., & Humanski, R. A. (1992). Localization of sound in the vertical plane with and without high-frequency spectral cues. *Perception & Psychophysics*, *51*(2), 182-186. doi:10.3758/bf03212242

[7] Carlile, S. (2011). Psychoacoustics. In Hermann, T., Hunt, A., Neuhoff, G., editors, The Sonification Handbook, chapter 3, pages 41-61. Logos Publishing House, Berlin, Germany.

[8] Kapralos, B., Jenkin, M. & Milios, E. (2003). Auditory Perception and Spatial (3D) Auditory Systems, chapter 3, pages 70-76.

[9] Hofman, P. M., Riswick, J. G., & Opstal, A. J. (1998). Relearning sound localization with new ears. *Nature Neuroscience*, 1(5), 417-421. doi:10.1038/1633

[10] Wolfram Research, Inc, Acoustics in the Time Domain. (n.d.). Retrieved July 14, 2020, from https://reference.wolfram.com/language/PDEModels/tutorial/Acoustics/AcousticsTimeDomain.html

[11] Wolfram Research, Inc, Acoustics in the Frequency Domain. (n.d.). Retrieved July 14, 2020, from https://reference.wolfram.com/language/PDEModels/tutorial/Acoustics/AcousticsFrequencyDomain.html

[12] Wolfram Research, Inc., Mathematica, Version 12.1, Champaign, IL (2020).

[13] Kumar, S. (2020). Wolfram Spatial Audio. *Zenodo*. https://doi.org/10.5281/zenodo.3987494

[14] Kumar, S. (2020). *[WSC20] Simulate Audio with Different Arrangements of Speakers*. Online Technical Discussion Groups Wolfram Community. https://community.wolfram.com/groups/-/m/t/2034599.