



Audio Engineering Society

Convention Express Paper 190

Presented at the 156th Convention
2024 June 15-17, Madrid, Spain

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (<http://www.aes.org/e-lib>), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Localization Impact to Experience (LIE) Detector: A Methodology for Assessing Localization of Rendered Spatial Audio

Trinoy Dutta¹ and Wei Shi¹

¹ Dolby Laboratories

Correspondence should be addressed to Trinoy Dutta (trinoy.dutta@gmail.com)

ABSTRACT

This paper introduces a method for assessing the localization of rendered spatial audio using perceptual cues derived from the Interaural Transfer Function (ITF). The method uses a custom window to preprocess the impulse responses of the signals captured by a Head and Torso Simulator (HATS), and computes the Interaural Intensity Difference (IID), Interaural Time Difference (ITD), and Group Delay (GD) of the ITF. The method also applies different weights to the cues based on the source location and compares them to a reference database of HRTFs measured in an anechoic chamber. The method generates plots to show the localization error and the similarity between the reference and the measured cues. The paper evaluates the method on different speaker configurations and rendering algorithms and validates the results with subjective data. This tool represents a first step towards a full audio system development workflow based on simulations or measurements and can be deployed at an early point in the design cycle.

1 Introduction

The popularity of spatial audio is on the rise, with numerous renowned artists producing content in Dolby Atmos and several companies incorporating spatial audio into their products. The spatial audio experience is dependent on a multitude of factors, including the accurate reproduction of spectral cues, accurate localization of sound sources, diffuseness of the source, reverberation, trajectory of moving sources, speakers, and room characteristics, etc. In this study, the renderers used for spatial audio are from Dolby Laboratories.

The evaluation of spatial performance in audio systems is not trivial. Other reproduction attributes such as spectral, dynamics and distortion have well developed workflows to capture, analyze and present information useful to their optimization. These problems may even be reduced to that of meeting a target curve, much to the benefit of those tasked with the fundamental design. In contrast, spatial aspects

are often evaluated with listening tests using human subjects. Statistical methods are usually employed to extract meaningful trends from what is often variable data.

In systems where content creators position audio objects deliberately in a scene that is often accompanied with a visual counterpart, it is vital that the audio reproduction system mirrors the scene. A need therefore exists for an objective means of analyzing perceived spatial parameters. Given the human factors involved in reconciling sound pressures received at the ears with the sensation of sounds being emitted from sources that don't physically exist, the machine interpretation requires careful consideration such that it mimics the human experience.

Abundant literature points to the importance of interaural cues for localizing sound sources [1][2]. The interaural cues that we will focus on this paper are Interaural Intensity Difference (IID), Interaural

Time Difference and phase information from the interaural transfer function.

The *Interaural Transfer Function* (ITF) is the ratio of the Head Related Transfer Functions (HRTFs) at the two ears as shown in Equation 1.

$$ITF = \frac{HRTF_{Left}}{HRTF_{Right}} \quad (1)$$

The *Interaural Intensity Difference* (IID) is the magnitude of ITF as shown in Equation 2.

$$IID = |ITF| \quad (2)$$

The *Interaural Level Difference* (ILD) is defined as the magnitude of the Fourier transform of signal from one ear over the magnitude of the Fourier transform of signal from the other ear as shown in Equation 3.

$$ILD = \frac{|HRTF_{Left}|}{|HRTF_{Right}|} \quad (3)$$

Group Delay (GD) of the ITF is the derivative of the phase component of the ITF as shown in Equation 4.

$$GD(f) = -\frac{1}{2\pi} \left(\frac{d\psi_L}{df} - \frac{d\psi_R}{df} \right) \quad (4)$$

where,

$$\psi_L = \arg[HRTF_{Left}(f)] \quad (5)$$

$$\psi_R = \arg[HRTF_{Right}(f)] \quad (6)$$

The *Interaural Time Difference* (ITD) is the time difference between the signal received at one ear to the other as shown in Equation 7.

$$ITD = \bar{t}_L - \bar{t}_R \quad (7)$$

where,

$$\bar{t}_L = \frac{\int_{-\infty}^{+\infty} t \cdot hrtf_L^2(t) dt}{\int_{-\infty}^{+\infty} hrtf_L^2(t) dt} \quad (8)$$

$$\bar{t}_R = \frac{\int_{-\infty}^{+\infty} t \cdot hrtf_R^2(t) dt}{\int_{-\infty}^{+\infty} hrtf_R^2(t) dt} \quad (9)$$

The IID represents the spectral magnitude difference between the signals at the two ears. The ITD represents the time difference between the signals at the two ears. The ITD can vary with frequency and can influence the perceived location of a sound source. This effect is most pronounced for signals with complex spectral content, where different frequency components can be perceived as coming from different locations. For a more in-depth

understanding of perceptual cues and how to calculate them, readers are encouraged to refer to [1] and [2].

Common literature suggests that IID spectra contain important localization cues in mid and high frequencies [1][2]. ITD data are important at lower frequencies [3]. In this paper we compute ITD as a scalar number of the offset between the two signals filtered at a certain frequency range. GD provides similar info to ITD at various frequencies. The next section provides a background on similar studies. This paper introduces an objective approach to assess the localization of a sound source in a Dolby Atmos environment, taking into account human perception.

2 Background

There is a wealth of research on the perception of spatial audio, including studies on binaural audio and Head Related Transfer Functions (HRTFs). Some research has focused on evaluating binaural renderers in terms of their ability to accurately localize sound sources [4]. Other studies have investigated the effects of training on sound localization and auditory adaptation [5]. Additionally, some research has explored the role of spectral cues in the localization of spatial audio [6].

There are several studies that have used Deep Learning and model-based approaches to measure localization ([7], [8], [9], [10], [11]). Some studies focused on a single active sound source at a time. Some studies make the assumption that the most dominant sound source is the one to be localized, while others model background noise or rely on prior knowledge. Some studies have used HRTF databases collected with different Head and Torso Simulators (HATS) as training data, which can result in an incomplete definition of the relationship between features for a specific head. Many studies reduce the dimensionality of the dataset using techniques such as Principal Component Analysis or a network with a defined standard deviation. A common goal of these studies is to determine the direction of arrival of a sound source, which is similar to locating the physical source of the energy but does not necessarily reflect human perception of the sound source [7]. Object-based audio rendering algorithms like Dolby Atmos take an audio signal and distribute it among multiple speakers to create the perception of a sound event originating from a specific location in space by approximating higher level perceptual cues like ITD and ILD for the desired source position [12]. It is more important to understand if the source was perceived correctly than to ensure the physical source

of the energy. Finally, deep learning and model-based algorithms often have low spatial resolution and are more focused on assigning a rough location to a source in a noisy environment rather than understanding the precise location of a sound source.

Several studies have investigated the comparison of a reference signal to a measured signal in the context of sound localization. Some have focused on binaural localization algorithms in hearing aids, using ITD to estimate localization [13]. Another study examined the precision of spectral magnitude between a reference and captured signal, defining an objective metric using empirical features. This metric is similar to ILD, another factor that contributes to localization. The authors of that study proposed two methods to measure the distance, where distance quantifies the separation or dissimilarity between two objects or probability distributions in a mathematical space, between the reference and captured signals: a tuned regression model and Jensen-Shannon distance [14]. However, both methods heavily rely on the sample size of the data and have limitations. Regression models require tuning and do not generalize, while Jensen-Shannon distance, a popular divergence metric, may not accurately reflect human perception when applied to ILD patterns. A follow-up study proposed the use of ITF as a metric for sound localization. The ITF is defined as the complex ratio of the left and right ears' HRTF Fast Fourier Transform (FFT) coefficients, with the magnitude of the ITF equivalent to the ILD and the phase related to the ITD. The authors suggest using the Jensen-Shannon distance distribution on the ITF output, claiming that it adequately captures both ILD and ITD information [15]. However, it is unclear how the distance is computed using the complex ratio of the FFT coefficients of the two ears' signals. The Jensen-Shannon distance is a symmetrized and smoothed version of the Kullback-Leibler divergence, meaning it does not matter which signal is considered as the reference and which as the comparison. It is a similarity metric, rather than a traditional distance function, and has undefined values for negative numbers. When computing the ratio of the left and right ear's FFT output, negative numbers are commonly encountered. One possible solution is to scale the signals by the global minimum, but this can have adverse effects during probabilistic comparisons and is outside the scope of the study. A subsequent study explored the use of spectral distortion across several Equivalent Rectangular Bandwidth (ERB) bands, calculated using gamma tone filter banks [16]. However, the study provided limited information on

the subjective validation of the method and its generalization across different renderers and playback systems. The study also raised concerns about the method's sensitivity to front-back inversions, the cone of confusion, and its overall robustness in real-life conditions. The study focused exclusively on Extended Reality (XR) devices, where the sound source is close to the ears but not exposed to crosstalk cancellation or other rendering enhancements found in commercial renderers. A rendering algorithm ultimately must balance trade-offs among spectral balance, spatial accuracy, smoothness, and sensitivity to listener placement – all of which can affect how a sound object is perceived by the user.

There is a need to establish a methodology for measuring sound localization that is more reflective of real-life conditions and applicable across different device categories. There is also a need to define the interdependency of the various perceptual cues for localization. This study is focused on sound sources in the horizontal plane only which is defined by the two directional vectors pointing to the front and right (or left) directions.

A few models have been developed for evaluating the similarity of audio using subjective human metrics, though fail to perform at the particular task of using binaural perceptual cues to determine sound localization. ViSQOL, which many sound similarity models are based on, evaluates mono signals by training a model using deep-learning distance functions but cannot differentiate stereo recordings or identify perceptual cues [17]. Ambiqal is limited to single point audio sources, rendered with Ambisonic playback [18]. Stereo InSeNet, while able to analyze and compute similarity of stereo files, does not evaluate binaural audio quality prediction, and doesn't apply weighting to perceptual cues [19]. LIE Detector is built with perceptual cues for localization in mind and identifies sound position based on the binaural and perceptual cues humans rely on. This allows it to identify wide multichannel and dispersed sounds from any source, provided it is captured with a binaural recording.

3 Methodology

LIE Detector uses a reference database of HRTFs. Those HRTFs should be captured in an anechoic chamber using physical loudspeakers placed at the source locations. These reference HRTFs capture the perceptual cues generated by a real source placed at those locations. Using a renderer of choice, one can

place audio objects at different locations and render a scene. The same HATS as the reference database can be placed at the user location. The perceptual cues recorded by the HATS can then be compared to the reference database to check for similarity. The above approach is the same taken by the authors on these studies [14, 15, 16]. The main difference is how the data is pre-processed and the way the different cues are weighted during the similarity comparison. Fig 1 shows the algorithm diagram. In this study, ground truth is the intended sound source location.

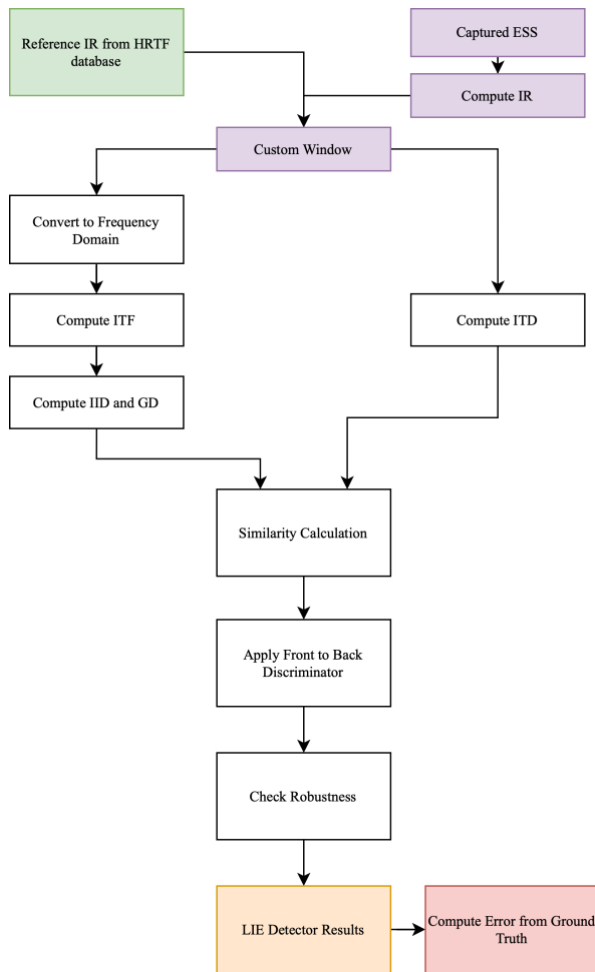


Figure 1. The LIE Detector algorithm

In this study, an Exponential Sine Sweep (ESS) is used as the stimulus. An impulse response is then computed using the Farina method from the captured HATS signal [20]. The ESS signal is 2.5 s long and has frequency content from 20 Hz to 20 kHz. The Impulse Responses are then windowed using a custom window shaped as shown in Fig 2.

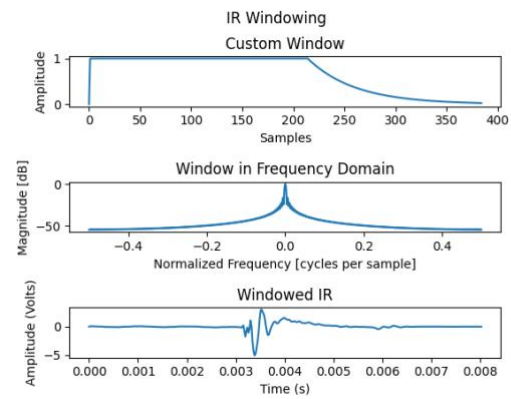


Figure 2. The top plot shows the shape of the window. The bottom plot shows the effects on an Impulse Response

Windowing of the IR:

The window has a cosine start followed by ones and ends with an exponential decay. Various studies have used gamma tone filter [16, 21]. The significance of using those for localization vs. windowing the data is not very well studied. It is well known that the early reflection and the late reverberation has an impact on localizing sounds. The authors observed that applying the custom window to the impulse responses preserves just enough of the energy that shapes the frequency response. Discussing the effects of the window and its relation to precedence effect is beyond the scope of the paper. Internal research has shown that this window shape is general enough that it can be used in normal rooms (i.e. someone's living room) or even in closed space (i.e. the cabin of a car). Fig 3 shows the IR and the Frequency Response of a measurement in a car cabin. In the top plot, a rectangular window was applied. In the bottom plot, the custom window was applied. The windowed IR shows how the window preserves the early reflection energy. The Frequency Response after applying the custom window is smooth and preserves the spectral behavior of the scene. The authors have also tried frequency-dependent windowing [22] but recommend the custom window as it preserves the energy that is essential for localizing sound.

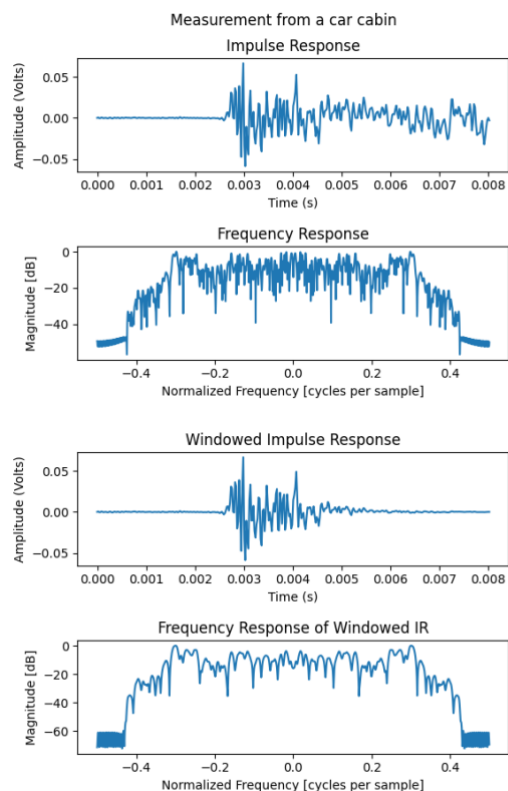


Figure 3. Shows the effect of applying the custom window in a car cabin. The top plot shows the IR and its Frequency Response with a rectangular window. The bottom plot shows the same after applying the custom window. It preserves the early reflection and the spectral characteristics.

After applying the custom window, the IRs of the two ear signals are transformed to the frequency domain by Fourier transform. Instead of using an ERB band, the algorithm opts for a Fourier transform to check the similarity with a database at a later stage. The resolution of the FFT of the two ears play a significant role in this process. If the FFT resolution is too high, it may look like a noise signal, but if it's too low, it averages out the spectral cues necessary to judge localization in a scene. In this study a few milliseconds capture at a 48k sampling rate is taken, centered around the ipsilateral ears' impulse. The authors have tried various smoothing functions and settled on using a Gaussian filter to smooth the IID in the log domain, and GD response in the linear domain. Fig 4 shows a few plots with the smoothing filter applied.

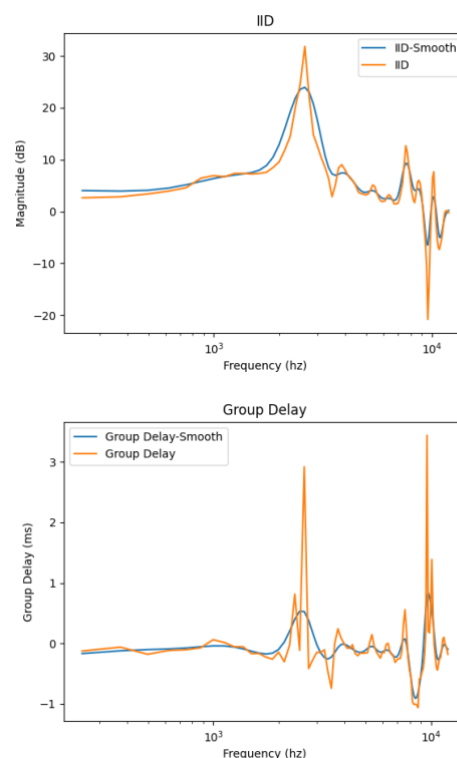


Figure 4. Shows the Gaussian smoothing filter applied to IID and GD.

Many studies have relied on using a minimum phase representation of GD of an ITF. The authors have found that using minimum phase smooths the GD to a much higher degree such that it no longer provides enough resolution to match to the database. Fig 5 shows the GD with and without a minimum phase transformation.

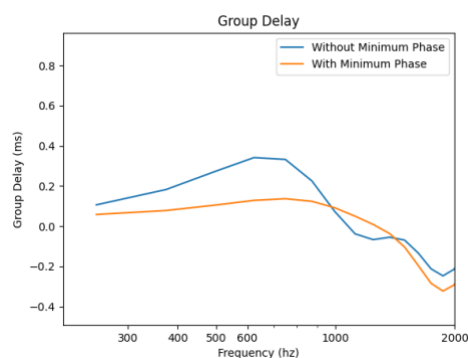


Figure 5. Shows minimum phase representation along with the linear GD.

The perceptual cues at various frequency ranges have varying levels of importance in localizing sound sources. IID at lower frequencies has important cues

for depth perception and is important especially if the sound source is close to the listener [1]. IID at mid frequency bands are known to contribute to the spectral naturalness of a scene and is an important cue for perceiving sources in the left or right zone of the listener [2]. GD below 600 Hz is critical to separate sources in the left to right zone but not that important for perceiving sources in the front or back stages [2]. It is well known that human perception relies on different cues for localizing sources at different locations. Readers are encouraged to read [1] [2] for more details. To mimic that behavior in LIE Detector, we use the perceptual cues shown in Table 1.

Similarity Calculation:

Comparing the test signal to the reference database requires a distance metric or metrics. Various studies have used Jensen-Shannon distance which treats every element of the distribution equally and checks for mutual information between the two distributions. Perceptual cues are frequency dependent data and using a distance function modeled after a probabilistic distribution might be adding complexity to the problem. Humans are also more tolerant to mismatch in perceptual cues at higher frequency ranges over lower frequency ranges. Table 1 shows the distance functions used. Equations 10 to 12 shows the mathematical definition of the distance functions.

Attribute	Named	Freq Range	Distance metrics
IID	iid_low	low	Euclidean
IID	iid_mid	mid	Cosine
IID	iid_high	high	Cosine
GD	gd_low	low	Cosine
GD	gd_mid	mid	Euclidean
ITD	itd	high	Mean Squared

Table 1: Distance metrics used in LIE Detector

$$d_{\text{Euclidean}}(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (10)$$

$$d_{\text{Cosine}}(\mathbf{p}, \mathbf{q}) = 1 - \frac{\mathbf{p} \cdot \mathbf{q}}{\|\mathbf{p}\| \|\mathbf{q}\|} = 1 - \frac{\sum_{i=1}^n p_i q_i}{\sqrt{\sum_{i=1}^n p_i^2} \sqrt{\sum_{i=1}^n q_i^2}} \quad (11)$$

$$\text{MSE}(\mathbf{p}, \mathbf{q}) = \frac{1}{n} \sum_{i=1}^n (p_i - q_i)^2 \quad (12)$$

where $\mathbf{p}$ and $\mathbf{q}$ are attributes of n -dimensional vectors.

To account for the fact that human perception uses a combination of perceptual cues to localize a sound source, we propose weighting the various cues based

on the location of the sound source as shown in Fig 6. The x-axis shows the azimuth angle of the sound source, and the y-axis shows the weight of each cue.

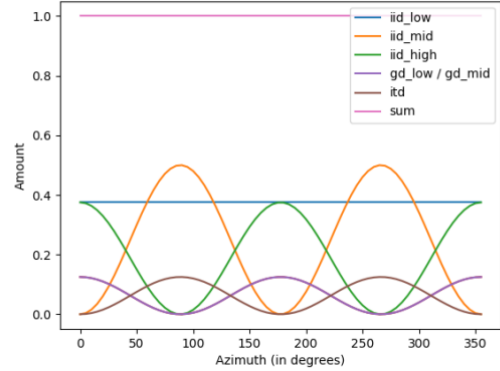


Figure 6. Shows the weights to be applied to the various cues based on location of the sound source.

Front to back discrimination:

The perceptual cues look similar for angles that are in the cone of confusion [1]. This problem is especially pronounced when the sound source is in the front or the back stages. Fig 7 shows the distance plots for a few scenes. The azimuth angles are plotted in the x-axis and the distance values in the y-axis. A distance value of 0 means a match and a value of 1 indicates no similarity. The blue line, labeled one minimum, shows the case where the distance function has a clear minimum at azimuth 190°. This happens when the attributes used in the distance calculation has a clear minimum. The orange line, labeled two minima, shows a case where the algorithm finds two possible solutions. The curve shows a dip at 45° and at 135° which are at the cone of confusion. Sources placed at those azimuth angles produce similar perceptual cues which result in similar distance values. Similarly, for the green plot, labeled two minima, the algorithm identifies 205° and 335° as possible matches. These are also complimentary angles and produce similar perceptual cues resulting in similar distance values. It is important to understand the tolerance of when to consider a minimum as a possible match. Humans are known to use micro head movements to disambiguate the scene. To address that, in LIE Detector there is an additional step where we look at features due to sources in the front vs. the back of the HATS.

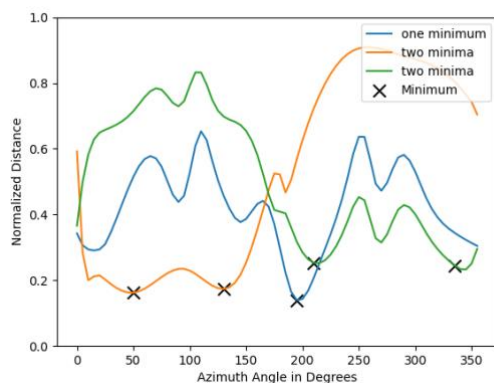


Figure 7. Shows the distance calculation for a few sound sources. The x-axis shows the azimuth angle in degrees and the y-axis shows the distance where 0 signifies a match and a value of one signifies a mismatch.

When a source is in the front stage vs. the rear stage, we see different behavior in the frequency responses of the two ears as shown in Fig 8. The curves do not overlap when the source is in the front stage but overlap around 5 kHz for sources in the rear stage. A separate study will explore the reason behind that and its significance to human perception.

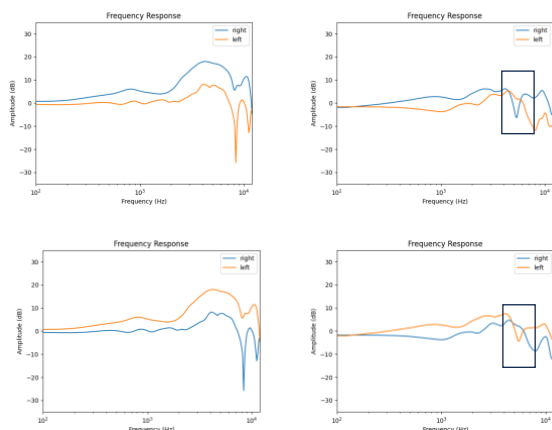


Figure 8. Shows the frequency response of the HATS for sound sources at azimuth 25° (top left), 155° (top right), 205° (bottom left) and 335° (bottom right). The curves do not overlap when the source is in the front stage but overlap around 5 kHz for sources in the rear stage as highlighted with the box.

In addition to that check, we compare the RMS level of the high-frequency content to mimic the micro head movement behavior of humans. Subjective analysis performed with a group of listeners (shown in section 6) provides the necessary information to tune the algorithm on when to consider minima as a possible match.

Robustness:

The algorithm is designed to be insensitive to the source of the sound (i.e. speaker characteristics, tuning, directivity pattern etc.), the path characteristics (i.e. reverberation of the room, early reflections etc.), and receiver characteristics (mismatch in HATS model between the reference and the measurement dataset). This methodology has been tried in a variety of use cases and has shown good correlation with human feedback. The details of how to tune the model parameters for maximum robustness to real life will be the subject of a future study.

4 Experimental Testing

In an acoustically treated room, 10 Genelec 8010A speakers were arranged in a square configuration at ear-height. Sound objects were played at an interval of 5° azimuth locations, using different speaker configurations and different object spatialization rendering – standard 7.0 and 5.0 systems using object-audio rendering, a 5.0 system with Atmos rendering, and a stereo system with Atmos rendering. Fig 9 shows the layouts. A HATS (B&K Model Type 4100) was connected through an RME Fireface 802. Sine sweeps (20 Hz- 20 kHz) were rendered at every angle 5 degrees apart in the horizontal plane, and then played through each speaker configuration. The playback of the files and the recording of the HATS signals were done using a python script. The HRTF database used is from Korea Advanced Institute of Science and Technology (KAIST) which was acquired in an anechoic chamber using Genelec 8010A and a B&K Type 4100 HATS [23].

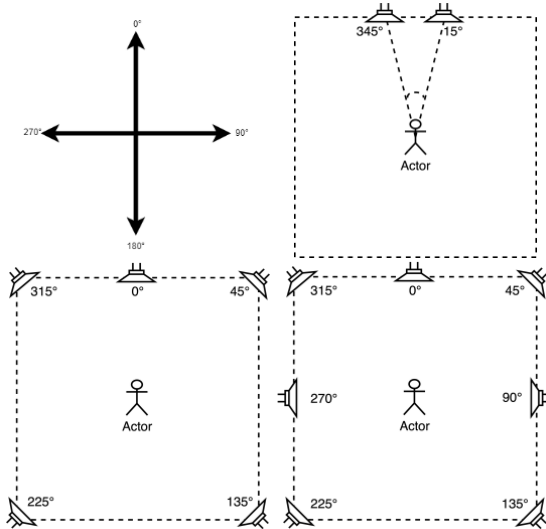


Figure 9. Shows the layouts. (top left) Coordinate system, (top right) Stereo setup, (bottom left) 5.0 object-audio renderer and Dolby Atmos renderer with 5 speakers and (bottom right) 7.0 object-audio renderer.

5 Experimental Results

The section below shows the results for the 4 layouts. LIE Detector provides a graphical map of localization error that is easy to understand. Localization error is defined as the smallest angle between the actual object location and its predicted output, regardless of the direction (clockwise or counterclockwise). The color bar shows the errors in localization in a linear scale. In Fig 10, the user is at the center of the plot. The hexagons in the outer circle show the speaker locations. The color, seen in spoke-like patterns, indicates the localization accuracy. Darkest blue represents the best spatial fidelity. These align with the physical speaker locations. The worst case is when an object appears to be 180° opposite of its intended location. Such a condition will be shown with a maroon spoke.

Top left shows the LIE Detector results for a 7.0 speaker layout with an object-audio renderer. The plot mainly has darker shades of blue which indicates that the renderer is able reproduce sound objects around the user. Top right is a plot for the 5.0 speaker layout with an object-audio renderer. The biggest error (shown in cyan) is directly to the sides of the listener where the gap between speakers is largest. As the object pans from a left to a left surround it will pass through a zone of poor localization accuracy. The bottom left plot shows the results for the Dolby Atmos renderer with 5 speakers. It has mostly darker shades of blue which indicates that the renderer is able to reproduce sound objects around the user. A

Dolby Atmos renderer is different from a 5.0 object-audio renderer because it has many additional features that contribute to a more perceptually accurate scene reproduction. The plot on bottom right shows a stereo setup. The map shows that the renderer can reproduce an accurate front stage but is not able to place sound objects intended for the side and the rear stage which is shown by maroon to yellow spoke-like patterns.

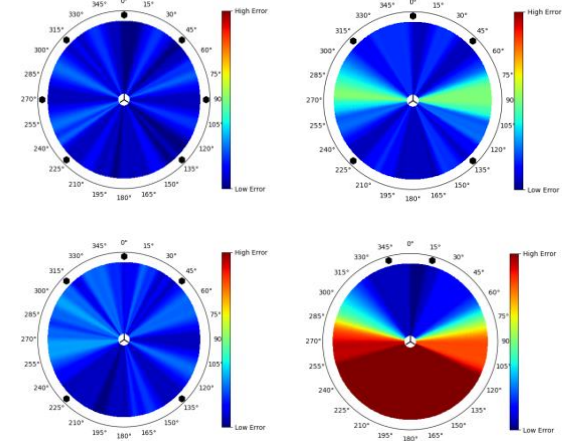


Figure 10. LIE Detector results. (top left) 7.0 object-audio renderer, (top right) 5.0 object-audio renderer, (bottom left) Dolby Atmos renderer with 5 speakers and (bottom right) Stereo setup.

6 Subjective Validation First Steps

A subjective evaluation was performed in an effort to compare perceptual data against the objective model. A future paper will go in-depth on the subjective validation. In this paper, we just show a preliminary result for 5.0 speaker layout with object-audio renderer. To perform this assessment, we developed a modified version of the High-Dimensional Objective Assessment of Spatial Audio Quality [24] designed to run in a closed virtual reality system. In the experiment, bursts of broadband pink noise were played over loudspeakers at varying azimuths while subjects were oriented forward. Subjects would assess the azimuthal location of the target relative to their listening position and mark the perceived location using the VR controller. Subjects were also enabled to indicate the diffuseness of the sound from within the VR game. It's important to note that subjects were only able to hear the sound while they were facing in the forward direction to control for listener position and orientation. The azimuth and diffusion data were recorded for processing and analysis.

After data collection was completed, a HATS was then positioned in the listening position. ESS from varying azimuth angles was then recorded using the same speaker configurations in the experimental setup. LIE detector then localized the sound, which was then compared to the localization azimuths from human subject data. Fig 11 shows the comparison between the subjective feedback and the LIE Detector predictions for a 5.0 speaker layout using an object-audio renderer. The human feedback represents the 5th to 95th percentile of the total responses. Fig 12 shows the LIE Detector results vs the mean subjective data for the 5.0 speaker layout with object-audio renderer.

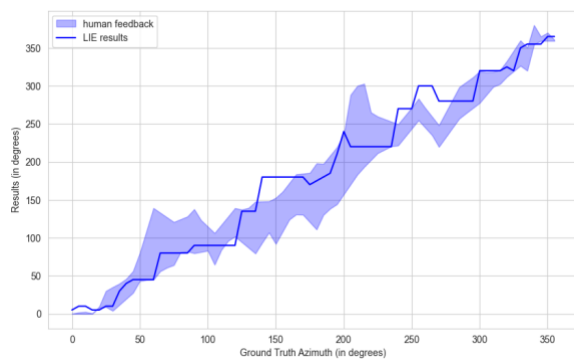


Figure 11. The human feedback which is compared with the LIE Detector prediction shown in blue.

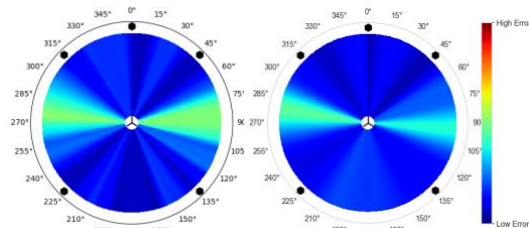


Figure 12. Comparison of LIE Detector results with human feedback. LIE Detector result (left), mean subjective data (right).

7 Discussion

Localization is a highly subjective matter and is only one aspect of a spatial experience. The ultimate evaluation of the localization potential of an audio system should be done through critical listening studies. The information from LIE detector has been shown to be consistent with perceived localization of both real and phantom sources. Throughout the development of this methodology, the authors have used averaged subjective feedback from a group of participants to tune the algorithm such that its outputs more closely match the subjective results. As the dataset grows, the algorithm will become an even

better generalized representation of average human spatial perception.

LIE Detector provides repeatable and valuable insights into spatial audio localization as an adjunct to human data. This should allow users of LIE Detector to easily validate a setup, benchmark a system, optimize a rendering algorithm, and more effectively close the loop between the intent of the content creator in creating a dynamic 3D acoustic scene, and the ability to reproduce this accurately with specific and scaled loudspeaker configurations that exist in a variety of audio products. Furthermore, this capability opens the door to early hardware design decision-making and a measurable objective means of benchmarking spatial performance.

8 Conclusions

A methodology has been presented that can be used to measure the localization of a sound source in the horizontal plane. The use of a custom window applied to the IRs improves performance in various real-life scenarios. To predict localization performance, the system uses perceptual cues derived from measured ITFs, and smoothing of these cues plays an important role when comparing them to the reference dataset. Adjusting the weights of various cues in the distance calculation as a function of the source location allows tuning the algorithm to improve conformance with human subjective measurements. The authors present an easy-to-understand plot that shows how the algorithm's localization error correlates with human subjective feedback.

9 Acknowledgement

The authors acknowledge the help, guidance, and support from our team members – Doug Replogle, Douglas Mandell, Lucas Orr, Samuel Goldberg, Kelvin Griffiths, James May, Jerome Halley, Karen Stackpole and David Carr. The authors would like to acknowledge the help and support from Alan Seefeldt on Dolby technologies. Subjective evaluation of LIE Detector was performed by Nathan Swedlow, Audrey Howard, Evan Gitterman, and Dan Darcy.

References

- [1] Zhong X., Xie B., *Head Related Transfer Function and Virtual Auditory Display*, InTech, doi: 10.5772/56907 (2014).
- [2] Gilkey R., Anderson T.R., *Binaural and Spatial hearing in real and virtual environments* (1997).

-
- [3] Breebaart J., "Comparison of Interaural Intensity Differences Evoked by Real and Phantom Sources," *Journal of the Audio Engineering Society* vol.61, no.11, pp. 850-859 (2013).
- [4] Reardon G., Genovese A., Zalles G., Flanagan, P., & Roginska A., "Evaluation of Binaural Renderers: Localization," in *Audio Engineering Society Convention 144*, Audio Engineering Society (2018).
- [5] Martin F.L., "Sound localization training and auditory adaptation: a review," in *Audio Engineering Society Convention 150*, Audio Engineering Society (2021).
- [6] Langendijk E.H.A., *Spectral Cues of Spatial Hearing*, Proefschrift Technische Universiteit Delft, S.n. (2002).
- [7] Yang Q., and Zheng Y., "DeepEar: Sound Localization with Binaural Microphones," in *IEEE Transactions on Mobile Computing*, vol. 23, no. 1, pp. 359-375, doi: 10.1109/TMC.2022.3222821 (2024).
- [8] Braun S. and Tashev I., "Acoustic Localization Using Spatial Probability in Noisy and Reverberant Environments," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2019, pp. 353-357, doi: 10.1109/WASPAA.2019.8937163 (2019).
- [9] Ma N., Gonzalez J. A. and Brown G. J., "Robust Binaural Localization of a Target Sound Source by Combining Spectral Source Models and Deep Neural Networks," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2122-2131, doi: 10.1109/TASLP.2018.2855960 (2018).
- [10] Riedel, S., Frank, M. and Zotter, F., "The effect of temporal and directional density on listener envelopment", *Journal of the Audio Engineering Society* vol.71, no.7/8, pp. 455-467 (2023).
- [11] O'Dwyer H., Bolan F., "HRTF Clustering for Robust Training of a DNN for Sound Source Localization", *Journal of the Audio Engineering Society* vol.70, no.12, pp. 1015-1026 (2022).
- [12] Roginska A., & Geluso P. (Eds.), *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio (1st ed.)*, Routledge, <https://doi.org/10.4324/9781315707525> (2017).
- [13] Courtois G., Marmaroli P., Lissek H., Oesch Y., Balande W., "Implementation of a binaural localization algorithm in hearing aids: specification and achievable solutions," in *Audio Engineering Society Convention 136*, Audio Engineering Society (2014).
- [14] Crawford S., Audfray R., Jot J., "Quantifying HRTF Spectral Magnitude Precision in Spatial Computing Applications," in *AES International Conference on Audio for Virtual and Augmented Reality*, Audio Engineering Society (2020).
- [15] Mathew J.D., Januszkiewicz L., Pensko M., Cortes D., Hertensteiner, M., Audfray R., "Quantifying Localization Potential using Interaural Transfer Function," in *Audio Engineering Society Convention 151*, Audio Engineering Society (2021).
- [16] Mathew J.D., Januszkiewicz L., Pensko M., Cortes D., Hertensteiner, M., "Measuring Auditory Localization Potential on XR Devices," in *AES International Conference on Audio for Virtual and Augmented Reality*, Audio Engineering Society (2022).
- [17] Hines, A., Skoglund, J., Kokaram, A. C., & Harte, N., "ViSQOL: an objective speech quality model", *EURASIP Journal on Audio, Speech, and Music Processing* 2015, no.1, pp. 1-18 (2015).
- [18] Narbutt M., Allen A., Skoglund J., Chinen M. and Hines A., "AMBIQUAL - a full reference objective quality metric for ambisonic spatial audio," *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, Cagliari, Italy, 2018, pp. 1-6, doi: 10.1109/QoMEX.2018.8463408 (2018).
- [19] Biswas, A. and Jiang, G. "Stereo InSE-NET: Stereo Audio Quality Predictor Transfer Learned from Mono InSE-NET," in *Audio Engineering Society Convention 153*, Audio Engineering Society (2022).
- [20] Farina A., "Impulse Response Measurements", *23rd Nordic Sound Symposium*, (2007).
- [21] Slaney M., "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank" (PDF). *Apple Computer Technical Report #35*. (1993)
- [22] Karjalainen M. and Paatero T., "Frequency-dependent signal windowing," *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, New Platz, NY, USA, pp. 35-38, doi: 10.1109/ASPAA.2001.969536 (2001).
- [23] Lee G., Choi S., Ko B., Park Y., "HRTF Measurement for Accurate Sound Localization Cues" (2022).
-