Audio Engineering Society

Convention Paper

Presented at the 144th Convention
2018 May 23 – 26, Milan, Italy

# Head Related Transfer Function Uniqueness for Predicting Spatial Localization Error

Eric C. Deng[1], Paul Chyz[1], and Chris Kyriakakis[1]

[1] *University of Southern California, Ming Hsieh Department of Electrical Engineering*

Correspondence should be addressed to Eric C. Deng (denge@usc.edu)

## ABSTRACT

In this work we seek to explore quantifiable, acoustic phenomena that may help immersive audio content developers better select sound source positions because, when done properly, sound source positions can be actively chosen to be more salient when desired, or more disorienting when appropriate. We started with an exploratory study with 4 participants that varied frequency, amplitude, and type (white noise versus pure tones) of test tones along different angles on the azimuth plane at 0° elevation. That set of experiments had results that led us to focus on exploring head-related transfer function (HRTF) uniqueness and how it relates to a person's ability to localize sound sources across different elevation ($\theta$) and azimuth plane angle ($\phi$). We then ran an extended study with 21 participants with 49 test tones varied over different $\theta$ and $\phi$, collecting 49 datapoints per participant, totaling in *1029 measurements*. Using the results of this larger experiment, we focused on two primary types of error that are useful in both advising design decisions for immersive content designers but also to support the underlying acoustic phenomena we accredit with creating the patterns we observe in practice. Our results showed both practical patterns that designers can use as guidelines for content generation and strong indicators for supporting our hypothesis that relates HRTF uniqueness to localization accuracy. In this work, we introduce our experimental designs, the results of both studies, and implications for future work in the form of immersive audio design guidelines and a experimentally-grounded heuristic for combining the left and right ear HRTFs in binaural hearing applications.

## 1 Introduction

When designed incorrectly, immersive audio content can be confusing, disorienting, and uncomfortable. As immersive audio applications become more and more prevalent [1], it is becoming increasingly important to have grounded design guidelines for content producers to improve the overall experience for users. One such design space that can greatly benefit from a set of data-driven design guidelines is the positioning of sound sources in these immersive settings, and our research is focused on quantifiably generating a set of design suggestions. The specific facet of immersive audio design that we focus on is the placement of sound sources [2]. The perceived locations of sound sources can act as landmark features in space; for example, a slight, constant buzz in a corner of a space can act as a robust point of reference when navigating a space. Although these sounds can be helpful for a user's localization ability, a discontinuous sound that originates from the same physical location can also be perceived as coming from different locations depending on the *acoustics of the space*, the *features of the sound sample*, and the *actual position of the sound source* [3]. If the perceived locations of these sounds are hard to locate or inconsistent, the user experience can be highly confusing and the system will be lower-performing overall.

Humans localize sound based on how sound is transformed by certain physical features such as the size and shape of the head, ears, pinna, and ear canals [4, 3, 5]. During development, typically-developing children learn through experience what a sound of a specific frequency, or band of frequency, should sound like from different source positions by using a combination of sensory-perceptual processes. Once these expected responses are learned, people then use these learned *head-related transfer functions* (HRTFs) as references for localizing sound–essentially "looking up" the position of a heard sound by comparing the observed frequency response for each ear to their known HRTF for that frequency (or frequency band) and estimating the location of the sound [4] These HRTFs are unique to each ear for every individual and are constantly being updated through observation–every time that sounds are heard and localized by non-auditory senses (often vision), the new observation is then used to update the existing HRTF and theoretically improve its accuracy.

Although these HRTFs are adaptive and get increasingly accurate over time, certain acoustic properties of sound and the human form limit an individual's ability to localize different sounds, as the frequency response of a sound of any frequency within the audible range at different positions is relatively limited. Because of this limited range, for every sound, there are always going to be sets of positions of sound sources that have similar, if not identical, expected frequency response. We expect these similar sets of values to be confusing to users and to be predictive of localization error when other senses are removed. In this work we perform two experiments that explore this relationship–first, an exploratory validation study that explored multiple dimensions of sound in a 2-dimensional localization task and second, an in-depth study focused on how frequency is related to error in a 3D localization task. In the next few sections we will present our study design, experimental setup, and analysis for both experiments as well as discuss impact on real-world applications and potential future work.

## 2 Background

This work aims to explore and validate a functional relationship between acoustic properties of audio signals and an individual's ability to localize the source of those signals. We build off a few primary bodies of work exploring *immersive audio*, *psychoacoustics* and, of course, *human sound localization*. In this section we will discuss the relevant outcomes from these research fields as well as an exploratory study we conducted that acted as a foundation for our primary work.

There an extensive body of work exploring fundamental theory and limitations of immersive audio [1], techniques for generating effective immersive audio content [6, 7, 8], and the importance of high-quality immersive audio [9]. Kyriakakis [1] studied limitations and considerations of real-world applications of immersive audio systems with the goal of understanding constraints of systems to "synthesize, manipulate, and render sound fields". Their work in acoustical considerations and human listening characteristics is of particular interest to us as our research is focused on exploring specifically these components of the localization experience. Georgiou and Kyriakakis [8] developed an abstracted approach for generating location-matched HRTF filter for manipulating immersive audio signals by segmenting the continuous space of potential source locations into a computationally-constrained discrete set that exploits similarities across different HRTFs (by angle). Mouchtaris et al. [7] examined key signal processing factors in spatial sound rendering and Kyriakakis [1] discussed limitations to immersive audio systems in the realm of physiological signal processing.

In order to localize the source of audio signals, humans without aural impairment use a process called binaural hearing. When a sound reaches a person's ears, there are inherent differences between the sounds perceived by the left and right eardrum fur to physical differences in the shape of a person's head, chest, and ears [4]. As the sound approaches the ear, phase and amplitude differences, also known as Interaural Time Difference (ITD) and Interaural Level Difference (ILD) respectively, account for course localization of sound–*head shadows* that occur when sound sources are on either side of a person's head make it obvious to the listener which side of the head the sound is on and phase changes eliminates some possible locations based on how out of phase the sounds are. ITD and ILD are effective models for lateral localization but due to the cone model effect, produce sets of seemingly identical sets of sound sources. But humans have a much more accurate sense of hearing, which is explained by the *pinna filtering effect theory*.

This theory is focused on how the unique, complex, and asymmetrical shape of the human pinna affects sound waves, reflecting and directing waves in a special pattern and generating a frequency response within the ear related to audio sources. These sources are then localized by different auditory nerves and interpreted using *head-related transfer functions* (HRTF). These HRTFs are unique to each person and act as a lookup table for relative frequency content in each ear for a sound coming from any position in the sphere around a person's head [10]. As sound propagates around the chest, head, and pinnae, the reflection and absorption that takes places uniquely for each ear imposes two different sets of frequency responses on the original sound. The brain then recognizes these differences in frequency content between the ears and compares those differences to learned sets of HRTFs to best estimate the location of the sound source [11].

Because HRTFs are inherently specific to individuals it is infeasible to model HRTFs for every participant in every audio experiment, the HRTF database was created by the CIPIC International Lab in collaboration with the MIT Media Lab, and the University of Oldenberg [12]. This database is useful as a reference for the general population and resulted in the KEMAR head and torso simulator for audio research, a now-standard test dummy used for hearing-related audio research and has physical characteristics (including pinnae) that are considered to be accurate representations of an "average person".

In order to bridge the two fields of acoustics and psychoacoustics within the context of localization, we first ran a general preliminary data collection to find interesting patterns to further explore. Existing work already investigated the role of the pinnae in localization ability, but had not yet addressed any of the acoustic and psychoacoustic variables regarding the sound source and environment. These additional variables lead to our focus on some important acoustic and psychoacoustic qualities and their impact on localization. Acoustically, this experiment focused on the differences between normal modes and more directional frequencies, as well as any relationship between the amplitude of the sound source, HRTFs, and localization accuracy. In the following section we will introduce this exploratory study that led to our primary work on HRTF uniqueness and localization accuracy.

### 2.1 Exploratory Study

This experiment set out to explore the effects of frequency, bandwidth, and amplitude on sound localization ability [5]. We hypothesized that there would be a directly proportional relationship between frequency and localization accuracy, while there would be a single-peaked distribution of error versus amplitude. We also expected to see better localization with wider bandwidths compared to pure tones, and that there would be a correlation between localization accuracy and uniqueness of HRTF values at various angles. We conducted a between-by-within study with a randomized set of sound source angles, using sine waves and white noise at different frequencies and amplitudes. We found that our expectation of a directly proportional relationship between frequency and accuracy was correct, with the exception of some front-back error due to initial reflections at high frequencies. Our hypothesis about the distribution of error versus amplitude was correct in the mid to high frequencies, but inverted in the low frequencies. Finally, we confirmed that higher bandwidth leads to more accurate localization, and we found strong correlation between HRTF uniqueness and localization accuracy.

#### 2.1.1 Study Design

In an effort to generally explore how features of sound can affect a person's ability to localize point sources we ran an preliminary study in which we varied (1) *frequency*, (2) *amplitude*, and (3) *bandwidth* of test sounds and played those test sounds at different angles along the azimuth plane. With 48 total conditions (8 frequency conditions, 3 amplitude conditions, and 2 noise conditions (white noise versus pure tones)), our experiment sought to reveal general patterns in 4 participants between these features and people's localization accuracy.

Blindfolded, each participant was then asked to point to the location of a sound source from one of these 48 conditions. We placed a swiveling office chair 2.5 feet offset from the center of the room endwise and centered lengthwise. We then placed our speaker, an M-Audio BX5, 5 feet away from the user, also centered lengthwise. This centered our speaker-participant pair endwise and was our effort to minimize the effects of reflections from the back wall.

During the experiment, we chose to rotate the user rather than reposition the speaker for logistical reasons.

Moving the speaker would have slowed the experimentation process and would likely be inconsistent. Instead we randomly generated 48 integer values between 1 and 12 (each value corresponding to the location on a clock face) and paired those with our random ordering of the sounds we previously generated.

We had one experimenter on a computer playing the sounds while another experimenter rotated the chair to the appropriate location, communicating using hand signals. The experimenter rotating the chair was instructed to stay directly behind the user at all times as to not hint to the participant where the speaker is coming from or where they were previously facing and rotate the participant at least one full rotation between two different sounds. This minimized the risk of having participants make guesses based, at least partially, on their previous guesses.

The participants were blindfolded and given a laser pointer. After the sounds were played, the participant pointed in the direction of the sound they heard, while the experimenter standing behind the participant marked their indicated guess on the speaker table. We initially started by marking the wall where the laser hit, but given constraints in the number of experimenters and difficulty in measuring angle to a point over 15 feet away, we decided to go with this approach instead. We chose to work on a single plane for similar reasons but recognized that exploration of effects in 3D would be very interesting to do.

We ran through a random ordering of all the sine wave sounds with random participant positions and then ran through a random ordering of all white noise samples with another set of random participant positions. All participant guesses were labeled with pre-made tape markers on the table and then collected afterwards using the iPhone compass app with one experimenter standing at the location of the chair and collecting the absolute position of each marker and comparing that to the actual position of the speaker.

#### 2.1.2 Results and Analysis

The results of that preliminary study supported most of our hypotheses–we found that larger bandwidths led to better localization, amplitude made no significant impact on the localization ability for a sound of set frequency from the same location, and that there was a weak negative relationship between frequency and localization accuracy. Most importantly, we saw strong correlation between HRTF uniqueness and localization accuracy. At a given frequency, the HRTF relies on the uniqueness of each angle's amplitude to accurately localize sound. In the event that two source positions for the HRTF of a frequency had similar HRTF amplitudes, there is an increased likelihood of localization confusion. There are distinct peaks in error at the angles where each frequency has the least unique amplitude levels, such as at 30 and 330 degrees. Additionally, there is a clear dip in localization error around 160 degrees where HRTF amplitudes are at their most unique values. These patterns can be seen as strong indicators for causation and are the drivers for our primary experiment.

## 3 Methods

The key result from this exploratory study and data collection that we wanted to further explore was the correlated, and potential causal, relationship between the uniqueness of the HRTF to an individual's ability to locate a sound source. We hypothesize that **the more indistinguishable the frequency response, the more difficult a sound will be to localize** and in the next sections we will discuss our main experiment focused on exploring this relationship.

### 3.1 Study Design

We conducted a between-by-within study with **21** participants, 14 male and 7 female, all of whom were undergraduate university students in an on-campus room treated specifically for audio-related research. Each participant was told that they would be hearing test sounds and to point at where they thought the test sound came from. They were told that the test sound could be coming from anywhere around them as to not give them information about the limited positions from which the sounds could come from and that they were being recorded and the vector formed by their wrist and fingertip were being used as their guess. Then the participants were given a blindfold and then ushered into the room and into the chair in the center of the speaker rig (Figure 1, 3).

Each participant was subject to 49 test sounds, varied over *center frequency*, *elevation*, and $\phi$. In order to select the center frequencies for the test sounds, we visualized the HRTFs of the KEMAR head [13], for both

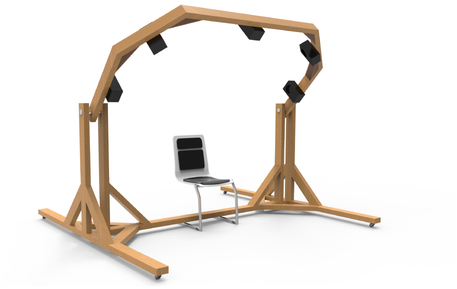

**Fig. 1:** Render of Speaker Rig with Speakers Mounted

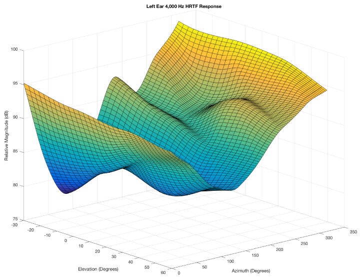

**Fig. 2:** HRTF for Left Ear at 4kHz over $\theta$ and $\phi$

ears, and chose frequencies that spanned the audible range for which the HRTF seemed unique for *both* the left and right ears (Figure 2).

In the end we ended up selecting center frequencies of 250Hz, 1kHz, 2kHz, 4kHz, and 10kHz and made 200Hz-wide test tones with those center frequencies using Audacity. This bandwidth was selected by observing results from the previous study that showed, as expected, that the wider the band, the easier a sound is to localize. 200Hz bandwidth, or 100Hz in each direction, did not result in drastic differences in HRTF for all five center frequencies like 300Hz did but was wide enough to have an audible difference for the lower center frequencies.

In order to achieve our desired 49 data-points per participant, we had to select 10 sound source positions for 4 test tones and 9 for one. To do so, we visualized the HRTF over phi and elevation for each ear for each center frequency. This would allow us to see peaks and valleys and estimate which positions, according to our hypothesis, would be easier or harder to localize. We went through each plot and selected 5-10 sound source locations that we thought would be hard to localize and 5-10 sound sources that we thought would be easy to localize. We then cross-referenced these sets for the two ears for each frequency and created 10 test positions per frequency and 9 for 250 Hz from the overlapping sets. This way, the source locations that we expect to be difficult to localize would be difficult for both ears and the source locations we expect to be easily localizable would be expected to be localizable for both ears. We also made an effort to distribute these positions somewhat evenly over phi and elevation as we knew that we would be doing surface regressions over phi and elevation for much of the analysis and an even distribution would help combat potential aliasing and over-interpolation. There are cases in which the location can be expected to be localizable only for one ear but those results would be less definitive and we wanted to focus on more definitive locations.

After the 49 test sounds, defined by *center frequency*, *elevation*, and *phi*, were selected, we made a slide deck with each sound displayed that would help us run the experiment as well as make it easy to shuffle ordering between participants in an effort to negate ordering effects. In the following section we will further discuss our experimental setup and data collection methods.

### 3.2 Experimental Setup

This experiment was carried out in a large, on-campus room treated specifically for audio-related research and teaching with a noise floor of 30.5 dB SPL and Schroeder frequency of 85.84 Hz. We constructed a speaker rig that held 5 speakers (Audyssey Lower East Side Media Speakers) that pivoted around the center of the rig (Figure 1) and placed an adjustable, swiveling office chair at the center of the rig for participants to sit in. The adjustability of the chair height allowed up to line up each participant as to center their head in the middle of the rig.

The speaker rig was designed to pivot in elevation and for feasibility and symmetry, we chose to limit the elevation of the *rig* to *-30, 0, 30, and 60*. On the rig we mounted 5 speakers, labeled *A through E*, 45 degrees apart. Because of the geometry of the rig, this resulted in the following speaker positions, in *phi* and elevation, for the shown rig elevation and speaker IDs in Table 1.

**Table 1:** Speaker Elevation ($\theta$) and Speaker Azimuth Plane Angle ($\phi$) for Speaker and Rig Elevation

| **Rig Elevation** | **A ($\theta$)** | **A ($\phi$)** | **B ($\theta$)** | **B ($\phi$)** | **C ($\theta$)** | **C ($\phi$)** | **D ($\theta$)** | **D ($\phi$)** | **E ($\theta$)** | **E ($\phi$)** |
|---|---|---|---|---|---|---|---|---|---|---|
| -30 | 0 | -90 | -22.817 | -21.158 | -25 | 0 | -22.817 | 21.158 | 0 | 90 |
| 0 | 0 | -90 | 0 | -48.66 | 0 | 0 | 0 | 48.66 | 0 | 90 |
| 30 | 0 | -90 | 22.817 | -21.158 | 25 | 0 | 22.817 | 21.158 | 0 | 90 |
| 60 | 0 | -90 | 56.829 | -36.295 | 60 | 0 | 56.829 | 36.295 | 0 | 90 |
| 120 | 0 | -90 | 56.829 | -126.295 | 60 | 0 | 56.829 | 126.295 | 0 | 90 |
| 150 | 0 | -90 | 22.817 | -111.158 | 25 | 0 | 22.817 | 111.158 | 0 | 90 |
| 180 | 0 | -90 | 0 | -138.66 | 0 | 0 | 0 | 138.66 | 0 | 90 |
| 210 | 0 | -90 | -22.817 | -111.158 | -25 | 0 | -22.817 | 111.158 | 0 | 90 |

The speakers had independent inputs via 3.5mm audio connectors and had their gains set to 50% for consistency before each set of experiments. To switch between the speakers being used to play the test tones throughout the experiment, we had to switch which plug was being connected to the output of the computer and in the process, static would play through the speaker that the previous sound was played out of as well as through the speaker for the next sound. In order to combat this potential signal that could affect participant responses, we played our test tones through a Focusrite Scarlett 18i8 Audio Interface with two outputs, one adjusted to play output with the proper gain and the other with gain turned to 0. At any one point we connected the four unused speakers to a splitter and into the 0 gain output and the fifth speaker into the second output from the interface. Before each test tone, we adjust the audio jacks appropriately (, i.e. connect to the splitter or not) and plug the splitter and the fifth connector back into the audio interface at the same time, causing all the speakers to generate a static noise at the same time.

Using the selected test sounds and their respective locations, we generated a slide deck that listed *center frequency*, *speaker ID (A-E)*, *elevation*, and *participation orientation (front or back)*. This deck allowed us to efficiently run our experiment as well as easily mix up the ordering between participants. For each test sound, we followed this protocol:

1. Play test tone and wait for participant response
2. Adjust 3.5mm audio jack–isolating the audio jack of the speaker for the next test tone and plugging in the others into the splitter
3. Unplug and plug in the splitter and the output speaker jack into the interface
4. Rotate participant at least one full rotation until properly oriented
5. Pivot speaker rig to appropriate elevation

### 3.3 Data Collection and Annotation

Following this protocol, we conducted our experiment with 21 participants and recorded their responses using 3 RGB cameras (GoPro HERO 6 Black) for post-hoc annotation. This allowed for efficient experimentation as well as robust annotation. In this section we will discuss our data collection and annotation methods.

In our experimental protocol, we asked participants to point and hold for a second in the direction that they think the sound came from. We also told them that their guess would be based on the vector formed by their wrist and the tip of their index finger so that they could most accurately present their guess. In order to record their guesses, we set up 3 GoPro cameras recording video and audio, one facing the right side of the participant, one facing the left side of the person, and one facing straight at the person from the front (Figure 3). This video was then annotated by the experimenters for the elevation ($\theta$) and azimuth plane angle ($\phi$) for each guess as well as for the correct answer, based on the slide deck ordering for each participant.

## 4 Results

After annotating all the videos, we ended up with 1029 unique datapoints from 21 participants. Building off the results of our previous work, we focused on finding patterns between a person's ability to localize sound sources, acoustic features of sound, and head-related transfer functions.

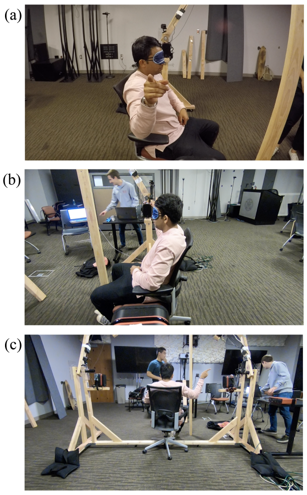

**Fig. 3:** Sample camera views

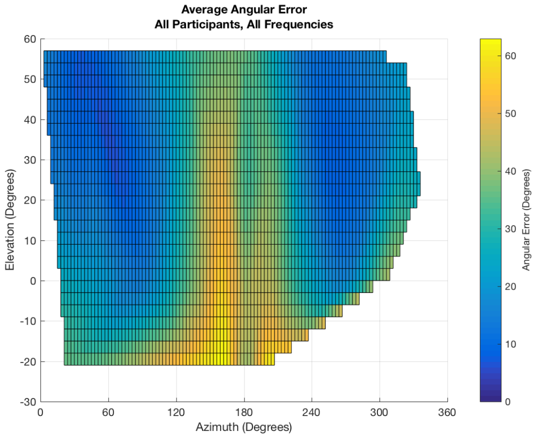

**Fig. 4:** Aggregated Angular Error over Elevation $\theta$ and Azimuth Plane Angle $\phi$

In this context, *localization ability* is not a well-defined concept. For our study, when a participant responds to a test tone they heard, is the *angular difference* between their guess and the actual sound source or the difference between the *expected frequency response of a test tone from the guess location* and the *expected frequency response of a test tone from the actual location* a better metric for accuracy? We argue that both have merit in being representative of an individual's localization ability, just in different ways.

### 4.1 Angular Error

*Angular error* is defined as the absolute angle between the guessed position vector and the position vector of the speaker (Table 1). In the context of the pinna filtering effect theory, angular error is a manifestation of both a person's ability to distinguish differences in frequency response as well as the uniqueness of the expected frequency response of a sound from the actual location based on the HRTF. Although angular error is also significantly impacted by the Interaural Time and Level Differences, it is still critically important to consider for practical purposes.

A person's ability to distinguish differences in frequency response effects the number of *possible* sound source positions based on a given HRTF (, i.e. an individual who has hearing sensitive enough to perceive frequency amplitude differences above 5 dB will have fewer "options" to choose from than an individual with sensitivity above 10 dB, as seen in Figure 2). The uniqueness of the HRTF comes hand-in-hand with a person's hearing sensitivity as well. Depending on the perceived frequency response, or rather the perceived range of possible frequency responses, the more unique the range within the HRTF, the easier it is for the person to accurately localize the sound source. The overall shape of the HRTF also affects how the *angular error* manifests. For instance with 4kHz noise (Figure 2), the expected relative magnitudes are similar for sources at ($\phi \approx 45^\circ$, $\theta \approx 45^\circ$) and ($\phi \approx 45^\circ$, $\theta \approx -45^\circ$), which are $90^\circ$ apart, and the expected relative magnitudes are similar for sources at ($\phi \approx 0^\circ$, $\theta \approx 0^\circ$) and ($\phi \approx 180^\circ$, $\theta \approx 0^\circ$), which are $180^\circ$ apart. Even for the same individual, who inherently has the same hearing sensitivity, an "error" in localizing a source at ($\phi \approx 45^\circ$, $\theta \approx 45^\circ$) will likely result in an angular error $\approx 90^\circ$ whereas an "error" in localizing a source at ($\phi \approx 0^\circ$, $\theta \approx 0^\circ$) will likely result in an angular error $\approx 180^\circ$.

Although angular error is not effective in representing a person's localization *capability*, it is still important in the practice of designing immersive audio applications. We are fundamentally interested in examining any sort of error in this context is because, in immersive audio applications, incorrectly localized sound sources can lead to nausea, confusion, and general discomfort. Angular error can also be used as a design technique for immersive environments in which you desire an "eery", "ghostly", or "disorienting" experiences for the user.

From our experiments, we plotted angular error over $\phi$ and $\theta$ and found that for a large majority of participants for all test tones, the distribution along the

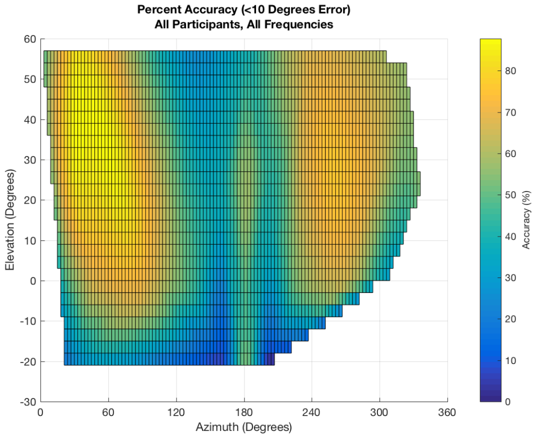

**Fig. 5:** Aggregated Percentage Error over Elevation $\theta$ and Azimuth Angle $\phi$

azimuth plane angle peaked at $\phi \approx 180^\circ$ with an approximately normal distribution along the elevation centered around $\theta \approx 30^\circ$ (Figure 4). The plots also show smaller peaks for elevation values at the bounds of the values that we explored in our experiment (, i.e. $\theta \approx -25^\circ$ and $60^\circ$), suggesting that people are more prone to mixing up sound source positions above or below the azimuth plane than mixing sound sources further from the median plane. This makes sense because of both the features of HRTF for all these center frequencies but also because of ITD and ILD, especially effects from head shadows. Regardless, these patterns can be helpful tools for immersive audio designers. Depending on whether they are designing audio cues to be easily recognizable and salient, for instance as signals for fighter pilots or in a first-person shooter game, or confusing and disorienting like in a haunted house or ghostly virtual reality experience, these designers can make advised design decisions based on these patterns.

One method we found helpful in further visualizing and understanding angular error over $\phi$ and $\theta$ is by classifying results into *correct* and *incorrect* guesses, which we defined by a threshold of 10°. While Figure 4, the plot of absolute angular error, is effective in representing *how incorrect* one can expect a user to be in localizing a sound from that position, Figure 5 is a simpler representation of whether or not the user will be *correct* in their guesses and may be more helpful when designing salient or disorienting sounds.

The angular error plots for different center frequencies *within* participants were not significantly different from one another as people still tended to make fewer, larger errors along the median plane but more, smaller errors at higher values on the azimuth plane.

### 4.2 Intensity Error

The second type of error we measured is *intensity error* or the difference between the expected relative magnitude of a certain sound from the guessed position and the expected relative magnitude of that same sound from the actual position of the source, both based on the HRTF from the KEMAR dummy. As the acoustic phenomena we are trying to observe is grounded in a person's ability to perceive and interpret differences in frequency responses in their ears to localize sounds, this intensity error metric is key to testing our hypothesis.

Unlike angular error, intensity error is a more isolated metric that is focused on an individual's ability to perceive intensity differences in sound and then "translating" those perceptions into sound locations. The way people localize sound is by comparing the perceived frequency response in each ear to the HRTFs of each ear–in the first study we only compared error to HRTF uniqueness of each ear (isolated) as it was preliminary and a proof of concept. But uniqueness is not well defined in the context of binaural localization–there is no heuristic for defining the overall value of the HRTF for *both ears* given a sound and position. In our analysis, we explore two potential options for binaural HRTF and seek to find a representative heuristic for future work.

Compared to angular error, intensity error has a more meaningful notion of *error magnitude*. Like we previously discussed, the magnitude of angular error is closely tied to the actual sound and position being tested–some sound sources from certain positions are more likely to be confused with sound sources at positions 90° off while others with positions 180° off. But the magnitude of intensity error is a continuous representation of a person's ability to recognize differences in intensity and therefore need not have a "correct" versus "incorrect" determination. Regardless of the angular difference between the guess and the actual positions, intensity error will still be isolated. For instance if the position of the speaker for a specific sound is correlated a sharp peak on the HRTF plot, 5 degrees can be a large difference in relative magnitude compared to a position that is located on a flatter portion of

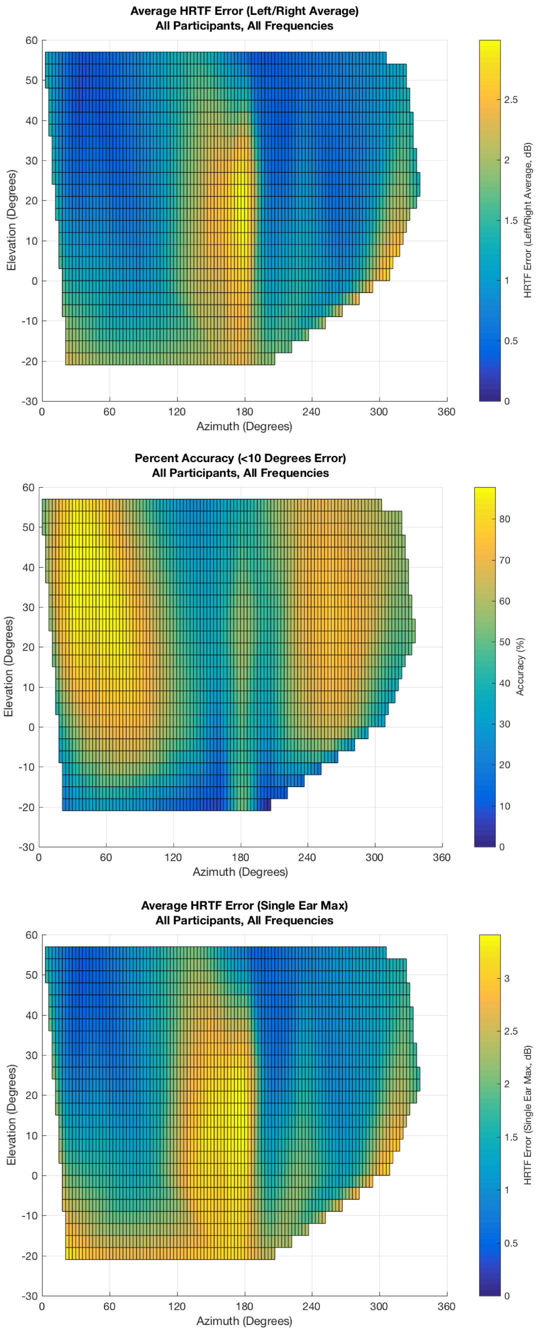

**Fig. 6:** Heuristic Comparison for Intensity Error–(a) Left-Right Ear Intensity Error Average (b) Error Rate (c) Left-Right Ear Intensity Max

the HRTF in which a 5 degree angular difference may map to a relative magnitude very comparable to that of the actual source position.

The two heuristics we want to explore for combining the HRTF errors for the left and right ears are–(1) *Left and Right Average (LRA)*, the average of the intensity errors of the left and right ears, and (2) *Left and Right Max (LRM)*, the larger of the two intensity errors for the left and right ears. In order to see which of these two heuristics was most representative of localization ability, we compare the results to error rate (with the previous definition of correct and incorrect–a 10° threshold), as seen in Figure 6. Based on our data, it clearly seems like the **average of the intensity errors for the left and right ear** is better correlated to a person's ability to localize sound sources, which we represent with their error rate.

Similar to the results seen in angular error, there were no significant differences within users for different test tones. Most participants' intensity error plots were relatively flat (Figure 7), which made sense given the measurement. Regardless of where the sound originates, we expect a person's hearing sensitivity to be consistent for intensity differences. This flat pattern shows strong support for our hypothesis that the more "unique" an expected magnitude difference on the HRTF, the more likely a person is going to correctly localize the sound because regardless of where the sound came from, people guessed sound sources with similar expected magnitude differences. By this logic, the fewer options for "similar expected magnitude differences" will result in higher localization accuracy. Because of this, immersive content designers can now rely on analyzing HRTFs to advise their decisions in selecting sound source positions for salience. This is helpful as it extends past the limited dataset and patterns we found with angular error in our experiment and demonstrates probably generalizability of this theory to other sounds.

## 5 Conclusion

In this research we explored the underlying phenomena that helps people localize sound sources with a focus on HRTF uniqueness as a result of a preliminary study. Our hypothesis was that the more unique the value of the relative magnitude of a sound source at a certain position, based on the HRTF for that sound, the easier it would be to localize. To test this, we varied $\theta$ and $\phi$ for 5 test tones with 21 participants and analyzed their performance in the context of two error metrics–*angular error* and *intensity error*. Angular error is an important consideration in the practice of immersive audio design. We are interested in observing how people make mistakes across $\theta$ and $\phi$ because difficult sources to localize can lead to errors which then lead to discomfort and disorientation. Thus angular error is a practical measure of both *where* and by *how much* people tend to make localization errors and the pattern we found was that people tended to make **fewer, larger errors along the median plane** but **more, smaller errors at higher values on the azimuth plane**. Based on the plot generated from the 1029 datapoints (Figure 6), designers can make advised decisions on where to place sound sources depending on whether they want to confuse the user or make the auditory cue as salient as possible. Intensity error is representative of the underlying phenomena we are trying to observe. As HRTFs are unique to each ear, we first had to figure out a good heuristic for combining the HRTF values from the left and right ears in bianural applications such as ours. By comparing two obvious options for this heuristic, LRA and LRM, by their correlation with error rate, it became clear that averaging the left and right HRTF value was a better approach (LRA). Regardless of where the actual sound source is, we hypothesized that people would choose guess locations with similar intensity errors, which we saw in our results manifested in a flat plot of HRTF error (LRA) over all of $\theta$ and $\phi$ (Figure 5). So not only did we find that averaging left and right ear HRTF values was an effective approach in the context of predicting localization error, our results support the underlying phenomena and give us reason to believe that the patterns we observe in the angular error from the constrained set of test tones can be generalized to other sounds as well.

## References

[1] Kyriakakis, C., “Fundamental and technological limitations of immersive audio systems,” *Proceedings of the IEEE*, 86(5), pp. 941–951, 1998.

[2] Bharitkar, S. and Kyriakakis, C., *Immersive audio signal processing*, Springer Science & Business Media, 2008.

[3] Blauert, J., *Spatial hearing: the psychophysics of human sound localization*, MIT press, 1997.

[4] Batteau, D. W., “The role of the pinna in human localization,” *Proceedings of the Royal Society of London B: Biological Sciences*, 168(1011), pp. 158–180, 1967.

[5] Middlebrooks, J. C. and Green, D. M., “Sound localization by human listeners,” *Annual review of psychology*, 42(1), pp. 135–159, 1991.

[6] Miner, N. E. and Caudell, T. P., “Using wavelets to synthesize stochastic-based sounds for immersive virtual environments,” Georgia Institute of Technology, 1997.

[7] Mouchtaris, A., Reveliotis, P., and Kyriakakis, C., “Inverse filter design for immersive audio rendering over loudspeakers,” *IEEE Transactions on Multimedia*, 2(2), pp. 77–87, 2000.

[8] Georgiou, P. and Kyriakakis, C., “Modeling of head related transfer functions for immersive audio using a state-space approach,” in *Signals, Systems, and Computers, 1999. Conference Record of the Thirty-Third Asilomar Conference on*, volume 1, pp. 720–724, IEEE, 1999.

[9] Whitton, J. P., Hancock, K. E., and Polley, D. B., “Immersive audiomotor game play enhances neural and perceptual salience of weak signals in noise,” *Proceedings of the National Academy of Sciences*, 111(25), pp. E2606–E2615, 2014.

[10] Beranek, L. L. and Ver, I. L., “Noise and vibration control engineering-principles and applications,” *Noise and vibration control engineering-Principles and applications John Wiley & Sons, Inc., 814 p.*, 1992.

[11] Algazi, V. R., Avendano, C., and Duda, R. O., “Elevation localization and head-related transfer function analysis at low frequencies,” *The Journal of the Acoustical Society of America*, 109(3), pp. 1110–1122, 2001.

[12] Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C., “The cipic hrtf database,” in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pp. 99–102, IEEE, 2001.

[13] Gardner, W. G. and Martin, K. D., “HRTF measurements of a KEMAR,” *The Journal of the Acoustical Society of America*, 97(6), pp. 3907–3908, 1995.