Audio Engineering Society

Convention Paper 10115

Presented at the 145th Convention
2018 October 17–20, New York, NY, USA

Localization of Elevated Virtual Sources Using Four HRTF Datasets

Patrick Flanagan[1], and Juan Simon Calle[2]

[1] *THX Ltd, 1255 Battery St, Suite 100. San Francisco, CA, 94111.*

[2] *THX Ltd, 1255 Battery St, Suite 100. San Francisco, CA, 94111.*

Correspondence should be addressed to Patrick Flanagan (patrick@thx.com)

ABSTRACT

At the core of spatial audio renderers are the HRTF filters that are used to virtually place the sounds in space. There are different ways to calculate these filters, from acoustical measurements to digital calculations using images. In this paper we evaluate the localization of elevated sources using four different HRTF datasets. The datasets used are SADIE (York University), Kemar (MIT), CIPIC (UC Davis) and finally, a personalized dataset that uses an image-capturing technique in which features are extracted from the pinnae. 20 subjects were asked to determine the location of randomly placed sounds by selecting the azimuth and the elevation from where they felt the sound was coming from. It was found that elevation accuracy is better for HRTFs that are located near elevation = 0º. There was a tendency to under-aim and over-aim towards the area between 0º and 20º in elevation. A high impact of elevation in azimuth location was observed in sounds placed above 60º.

INTRODUCTION

The relevance of 3D audio has increased in recent years due to emerging technology that takes advantage of spatial audio. Some of the common applications of 3D audio are virtual reality (VR), augmented reality (AR) and mixed reality (MR). With adoption of new formats such as MPEG-H, in which audio is not restricted to channel-based configurations, spatial audio can be introduced into other areas of the audio industry like broadcast, television and music.

At the core of the 3D technology are the Head Related Transfer Functions (HRTF). These filters are personal to each one of us, as they are impulse responses of our body, shoulders, head, and pinnae. Spatializers and renderers currently available use generalized filters, which sometimes are taken from real people, like in the case of the CIPIC database [1], or from mannequins like KEMAR [2], which is commonly used in research and academia. There is some discussion about the improvement of the subjective experience of 3D audio when using a personalized HRTF. As Begault states [3], the use of custom HRTFs improves the experience by reducing front-back confusion and by improving the perception of elevation.

There are several methods available to personalize HRTF filters. Some of these methods use acoustic techniques in which measurements are taken using impulse responses in controlled environments. Other methods use based on user preference, and some others use image-capturing techniques in which features of the HRTFs are derived from images of the subject’s anatomy.

In this paper, the perception of elevation at different elevation levels was tested using 3 commonly used HRTF databases and a personalized HRTF that was created using an ad hoc image-capturing technique. The three HRTF databases used are the KEMAR dataset created at MIT [2], the CIPIC database created at U.C Davis [1] and the SADIE-KEMAR database created at York University [4]. The personalization method creates a personalized HRTF dataset by analyzing an image of the subject, extracting some feature information, and then cross-referencing the features with a database of pre-determined non-linear transforms.

For this study, 20 subjects were asked to answer 36 localization questions (9 per HRTF dataset), with randomized location and in a randomized order. These localization questions were asked for static sources. Subjects were asked from where they felt the sounds originated, both in elevation and in the azimuth plane, with each one of the two planes displayed separately. Three different broadband percussive sound were used as stimuli for this test. To select the random locations of the positioned files, the azimuth and elevation values were selected from a list of fixed positions. The azimuth plane was divided into 12 different zones of uneven sizes. The elevation plane was divided into two hemispheres with 6 different areas of 20 or 30 degrees each (20 degrees in the areas closer to elevation = 0 degrees). There were no sounds placed below -50 degrees in elevation as some of the datasets had no HRTF information in this area.

The end goal of this paper is to analyze the effects of elevation on the perceived location of virtual cues. The data will be analyzed by looking at the different localization results, divided by elevation areas. By using multiple HRTF databases instead of just one, it is intended to reduce the impact of the specific characteristics of each one of the different capturing techniques.

1 LITERATURE REVIEW

Human audio localization relies on three main cues, ITD (Interaural Time Difference), IID (Interaural Intensity Difference) and Spectral cues. These cues were first described by Lord Rayleigh [5], who wrote about the importance of IID and ITD cues in localization, especially in the azimuth plane. Studies on the localization of virtually elevated sources using HRTF filters have shown that the effect of the pinna on the sound is crucial to accurately determine the position and of an audio source [6]. In the frontal plane, localization is more accurate than the median plane as elevation increases, due to the correlation with the increased effect of ITD and IID in localization [7]. Other studies show how head rotation can improve the overall localization of elevated sources [8]. Introducing head rotation to the test will generate dynamic cues which could be easier to localize, but as Ikeda, Kim, Ono & Takahashi [9] explain, at the end of the localization task people will be facing the source, and elevation localization in the median plane is reduced as previously stated.

The relationship between the HRTF filters and localization accuracy is crucial. Some studies show that elevated sources are normally perceived in a lower position when there is a discrepancy between the subjects HRTFs and the ones used to render the files in the virtual environment [9]. This perception of elevation is also related to the spectral peaks and notches generated by the pinnae; in some cases reducing the depth of some of the notches generated by the pinnae results in an increase in the perceived location of a source. In relation to those peaks and notches, Iida, Itoh, Itagaki and Morimoto [10] identified whichpeaks and notches were most important in the HRTFs. The first important peak is a frequency band located around 4kHz (3.4 kHz to 4.3 kHz), and appears to be constant with changes in elevation. There are notches located in the higher frequencies from 5.7 kHz to 9.5 kHz and 8.2 kHz to 13.5 kHz, which yield a higher individual variation than the first peak. Based on this, we can assume that achieving personalization of pinnae-related components of an HRTF filter could improve the localization accuracy of elevated sources [10].

2 METHODOLOGY

This experiment is a continuation of a previous study done between NYU and THX Ltd., in which subjects were asked to locate sounds that were spatialized by different 3D audio renderers [11]; because of this, some of the methodology is similar. For this paper, 20 people were tested on the task of localizing virtually placed sources. Individual HRTFs were extracted for each one of the subjects by using an image capturing technique. To create the personalized HRTF, a picture of the ear of each subject was taken. Important characteristics of the ear were extracted from the picture, which were then matched with a database of non-linear transforms. The three other datasets that were used on this test are commonly used in academia and in the industry. These three HRTFs datasets are MIT's KEMAR, York University's SADIE and Subject 3 from the CIPIC database (all these databases are publicly available online). These different HRTFs were labeled from 00-03. Three stimuli were used to test the subject´s localization. The stimuli used were two-second mono drum loops of different styles created in Pro Tools (48 kHz, 24-bit). The drum loops were chosen over other test signals, such as noise, because using natural sources closer resembles a real-world scenario in sound

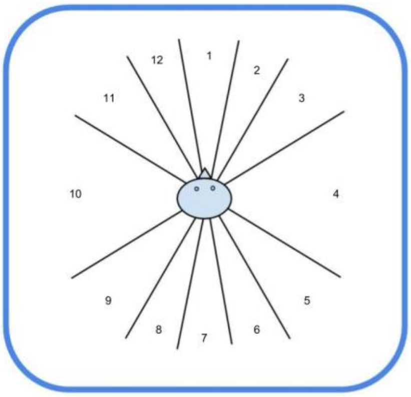

Fig. 1: GUI for azimuth locations

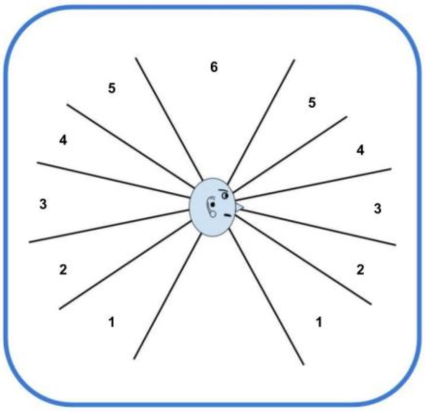

Fig. 2: GUI for elevation location

design. The loops used were broadband in frequency.

For each one of the four HRTF datasets, 9 different sounds were presented to the subject in a random order. Each one of the 9 sounds (3 per stimuli) were rendered by randomly selecting an azimuth and an elevation value from the possible locations displayed in Tables 1 and 2. After selecting the random pair of values, the closest HRTF from the dataset mesh was chosen using the Pythagorean Theorem. No interpolation of filters took place in this test. Overall, each subject answered localization question on 36 different sounds. No sounds between -50 and -90 in elevation or with 90 degrees of elevation were presented to the subjects because some of the HRTF datasets lacked information in this zone. Using a 2D user interface, subjects were prompted to select the azimuth and elevation location from where they perceived the sound. Subjects were able to replay the sound as many times as they wanted. None of the subjects went through a previous training phase and there was no feedback on their performance.

Zone	Azimuth Range	Possible locations
1	350º - 10º	0º
2	10º - 30º	20º
3	30º - 60º	30º,40º
4	60º - 120º	70º,80º,90º,100º,110º
5	120º - 150º	130º,140º
6	150º - 170º	160º
7	170º - 190º	180º
8	190º - 210º	200º
9	210º - 240º	220º,230º
10	240 º - 300º	250º,260º,270º,280º,290º
11	300º - 330º	310º,320º
12	330º - 350º	340º

Table 1: Possible location for azimuth per zone

Zone	Azimuth Range	Possible locations
1	(-60º) - (-30 º)	-50º, -40º
2	(-30º) - (- 10º)	-20º
3	(-10º) - 10º	0º
4	10º - 30º	20º,
5	30º -60º	40º,50º
6	60º - 90º	70º,80º

Table 2: Possible location for elevation per zone

Figure 1 displays the different zones in the azimuth that were available for selection by the subject, with each zone corresponding to the location values listed in Table 1. The azimuth is divided unevenly into zones based on the nonlinearity of the Minimum Audible Angle [12] and the localization blur in different areas around the head. Figure 2 displays the GUI through which the subjects would answer for the elevation of the source. The elevation zones were marked from 1 to 6, with the same nomenclature on each hemisphere. The possible answers ranged from -50 to 80. If the subject selected a zone in one of the hemispheres, the corresponding zone in the other hemisphere would be automatically selected as well. To avoid confusion, no sound was placed in the boundaries of the zones or at 90 degrees of elevation. All the subjects came from different backgrounds, some of whom had experience with 3D audio before, and some of whom had not.

3 RESULTS

20 subjects took part of this test, all of them answering 36 localization questions each. Overall, 720 localization attempts were captured with randomized locations. For these results, we will refer to the zones displayed in Tables 1 and 2 to discuss the different data that was captured. The results section divided into three different subsections. The

first part is dedicated to errors in elevation accuracy, comparing the 6 different elevation zones in which the test was divided. An attempt is part of a zone if the sound was originally played there, regardless of a subject's chosen localization answer. The responses in the absolute elevation accuracy will be taken as a binary answer where the subject is either correct or not. In this case, the percentage of correct answers by elevation zone will be presented. In this same subsection, distribution of error will be analyzed to understand the tendency of the direction (higher or lower) in which the subjects answered per elevation zone. In the second subsection, we will discuss the accuracy in azimuth. This analysis will be made by comparing the overall accuracy on azimuth in each of the elevation zones. Finally, an overall analysis will be displayed in the third subsection, showing the overall accuracy in azimuth and elevation.

3.1 Elevation

Elevation was divided into 6 different zones, with possible positions ranging from -50 to 80 degrees, without selecting any location where there was a line dividing the zone. For this first analysis, answers are treated binary responses, in which the elevation localization was correct or not. Figure 3 displays the percentage of correct answers per zone; the x axis lists the zones from 1 (lowest) to 6 (most elevated). As zones 1, 5 and 6 had more possible locations than zones 2, 3 and 4, we have ensured the effect of the elevation zone was significant enough.

Zone 3, containing only sounds with zero elevation, received the highest percentage of correct answers at 37.28%. The next highest percentage of correct answers is zone 4, with 29.8% correct, followed by zone 2 with 24.5% correct. Zone 5 follows with 14.41% of correct answers, and finally zones 1 and 6 with 10.59 and 7.06% respectively.

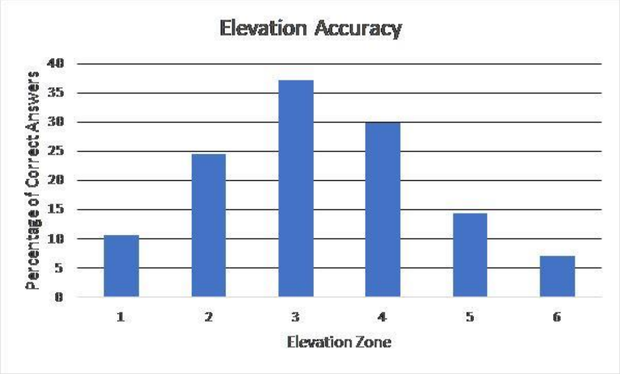

Fig. 3: Elevation Accuracy by elevation zone

The error of distribution was analyzed to determine tendencies in direction. The percentage of trials by zone that were over- and under-aimed by one zone of distance is shown in Figure 4. For example, if the sound originated in zone 3 but the subject answered zone 4, this is counted as an over-aimed attempt in zone 3. In Table 3, only those attempts that were off by one zone are considered. The values in Table 3 are represented graphically in Figure 4.

Zone	Over-aimed	Under-aimed
1	29.31%	-
2	31.58%	8.77%
3	27.11%	16.95%
4	21.05%	24.56%
5	5.41%	31.53%
6	-	14.73%

Table 3: Percentages of over-aimed and under-aimed attempts by elevation zone

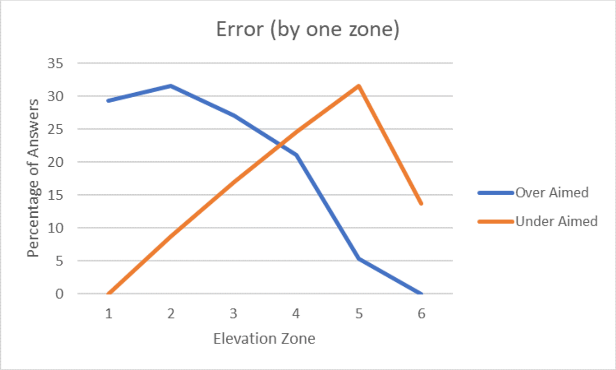

Fig. 4: Distribution of error. Percentage of attempts over-aimed or under-aimed by elevation zone

3.2 Azimuth

The localization in azimuth is different than the localization for elevation in virtual cues, as it is dependent on parameters like IIDs and ITDs [5]. This subsection of the results shows the impact of elevation in the azimuth localization. In Figure 5, the accuracy percentages in the azimuth location for each of the elevation zones are shown. The darker bars represent the percentage of correct answers in the azimuth without front-back confusion. The lighter bars show the percentage of correct answers with front-back correction. A front-back correction is done when the subject answers in the opposite hemisphere in the azimuth plane but in the correct opposing zone. As an example, if the sound originated in zone 2, but the subject answered zone 6, that is a clear front-back confusion and is taken as a correct answer.

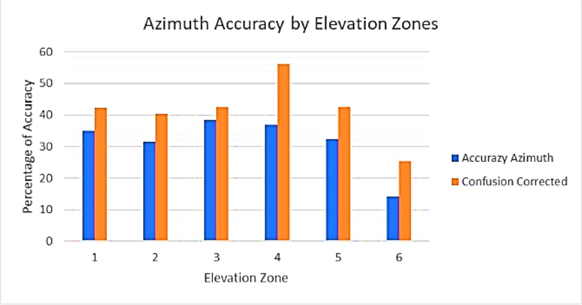

Fig. 5: Azimuth Accuracy by elevation zone

Table 4 shows some of the overall results from the study. These results are relevant when they are compared with other studies in localization. For these overall results, the answers were considered binary, where the answer either fits the condition or not. "Azimuth accuracy" is the percentage of answers where the azimuth zone was correct. "Azimuth Accuracy F/B corrected" is the correct percentage of answers when correcting the front-back confusion. Elevation accuracy is the percentage of correct answers in the azimuth. This table demonstrates that some of the zones had double the possibilities of being selected randomly in a dataset, and is why the overall elevation accuracy can't be taken by adding all the attempts and determining which ones were correct, mostly because the least accurate zones correspond to the highest-numbered. The mean "Elevation Accuracy" was taken by averaging the percentages in Figure 3.

Characteristic	Value
Azimuth Accuracy	30.15%
Azimuth Accuracy F/B corrected	40.46%
Elevation Accuracy	20.06%

Table 4: Overall Data

4 DISCUSSION

In this study it was intended to reduce the effect of the capturing technique used for the HRTFs. Because of this, some of the most common spatializers available were used. To minimize the effect of the dataset, the study was not limited to generalized HRTFs, instead opting to introduce a personalized HRTF..

As seen in Table 3, the elevation accuracy was better as the HRTF location is closer to zero degrees of elevation. The zone with the highest accuracy was zone 3, which only used HRTFs with elevation zero. For this zone the accuracy was around 37.28 %, which is still lower than normal localization in the azimuth plane [11]. The zone that followed in accuracy was zone 4, which only had HRTFs with elevation of 20 degrees, followed by zone 2, which only had values of -20 degrees of elevation. This supports the results shown by Ikeda, Kim, Ono & Takahashi [9], in which elevated sources were localized lower than what they were located.

In Figure 4, the percentage of answers that were under-aimed or over-aimed for each one of the different elevation zones is presented. It can be observed that zone 1 has zero under-aimed answers, since there are no zones below zone 1 (down to $^{-}$60º). It can also be observed that zone 6 has zero over-aimed zones, as there is no zone above it (zone 6 extends to 90º). It was decided to do graph errors by one zone because graphing all the over-aimed attempts would be less statistically relevant, as there is a higher number of possible zones above or below the edge zones.

In Figure 4, the darker line represents the number of attempts per zone that were over-aimed by one zone. For example, over-aimed zone 2 answers show the amount of times that subjects selected zone 3. The lighter line represents the under-aimed number of attempts per zone. There is a clear tendency to under-aim and over-aim towards 0-20º. The crossing of the two lines is around zone 4, which show a response bias towards 20º. These results resemble the conclusion drawn by Ashby, Mason and Brookes [13] in which they state that there is a response bias where sounds played from speakers below 30º were reported above their actual location and sounds above 30º were reported below their location.

In Figures 3 and 4 it can be observed that the only zones in which the accurate number of attempts was higher than the error by one zone were zone 3 and zone 4. The only data point that shows a different result out of the expected behavior was zone 6, where the under-aimed percentage is lower than the same value in zone 5. The distribution of error for this zone shows that a high percentage of the attempts were under-aimed by two zones, which means that for these high elevation values, the tendency to “feel” the sound coming from a lower elevation can exacerbate the error. To check this hypothesis, a study where data is placed between -90º and -60º in elevation should be conducted.

Azimuth accuracy was variable across the different elevation zones, being relatively poor in zone 6 (highly elevated sources). This is understandable, as the circumference of the circle around the user in which the sound can be placed is smaller as it reaches the vertical edges of the sphere, which reduces zone size, making it harder to recognize where in the azimuth the sound originates. This variance may also be caused by reduced ITD and IID as elevation increases [7]. For all zones between -60 and 60, the azimuth accuracy was between 30% and 40%, which is consistent with the previous study [11] in which different renderers where tested on azimuth localization, with most of them in the same range. After correcting front-back, most of the zones increased their azimuth accuracy higher than 40%.

Finally, the overall results show that azimuth accuracy was 30.15% and the elevation accuracy was 20.06% throughout the collection of 720 attempts. It was expected that azimuth accuracy was higher than elevation accuracy. After performing front-back correction, the overall amount of attempts yielded an accuracy of 40.46%.

The idea behind this study was to test the different set of HRTFs and compare the elevation accuracy between them. The amount of data taken did not show important trends in between HRTF datasets to draw relevant conclusions. Front-back confusion by zones was analyzed in this study to determine whether there is a describable behavior, but the lack of data points made it very difficult to draw conclusive statements.

5 CONCLUSIONS

Four different sets of HRTFs were tested in a localization task. Subjects were asked to locate a virtually rendered source in space by submitting both the azimuth and the elevation of the source. The HRTFs used were the KEMAR, CIPIC, SADIE and a personalized that was created for each subject by using image-capturing techniques. 720 data points were captured from 20 subjects. Subjects had a mixed background expertise in 3D audio.

The analysis shows that in a localization task that involves virtually elevated sources, subjects had a higher accuracy with sources placed closer to elevation 0º. This thesis was supported by examining the distribution of error, specifically analyzing the number of attempts that were over-aimed or under-aimed. It was shown that attempts tend to be under-aimed when they were in the upper hemisphere (higher than 0º-20º) and localization was over-aimed in the lower hemisphere (lower than 0º -20 º).

Azimuth localization for elevated sources tends to be less accurate, impacted by decreased ITD and IID as the elevation increases [2][7], but in this experiment, there was not a substantial difference in the localization in azimuth by elevation zones, except for sounds located between 60º and 90 º. This might be the case because there is no consistency with the HRTFs IID and ITD across the experiment.

6 FUTURE WORK

As the goal of this study is to test localization with commonly used HRTFs, a comparative study with a larger number of subjects could show the difference between HRTFs. To reduce the impact of IID and ITD, an elevation test with fixed azimuth locations could show the importance of personalization techniques that affect the higher spectrum of the frequencies, to observe the effect in elevation localization. Overall, the improvement of the quality of HRTF datasets will help the spatial audio content creators to deliver high quality products.

7 REFERENCES

[1] Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). “The CIPIC HRTF database”. In *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on the* (pp. 99-102). IEEE.

[2] Gardner, B., & Martin, K. (1994). “HRTF measurements of a KEMAR dummy – head microphone”. Massachusetts Institute of Technology, 280 (280), 1 -7.

[3] D. R. Begault, E. M. Wenzel, and R. Anderson, (2001). "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source" Journal of the Audio Engineering Society 49.10

[4] https://www.york.ac.uk/sadie-project

[5] L. Rayleigh, (1907). "On Our Perception of Sound Direction," Philosoph. Mag., vol. 13

[6] Blauert, J. (1969-70). Sound Localization in the Median Plane, Acustica, Volume 22, pp. 205-213.

[7] Barbour, James L. Elevation perception: Phantom images in the vertical hemi-sphere. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society, 2003

[8] [12] Wightman, F. L., & Kistler, D. J. (1999). Resolution of frontback ambiguity in spatial hearing by listener and source movement. Journal of the Acoustical Society of America, 105(5), 2841–2853

[9] Ikeda, M., Kim, S., Ono, Y., & Takahashi, A. (2010, October). Investigating Listeners' Localization of Virtually Elevated Sound Sources. In *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*. Audio Engineering Society.

[10] K. Iida, M. Itoh, A. Itagaki, and M. Morimoto, (2007). “Median Plane Localization Using a Parametric Model of the Head-Related Transfer Function Based on Spectral Cues,” Applied Acoustics, vol. 68, no. 8, pp. 835–850 .

[11] Reardon, G., Genovese, A., Zalles, G., Flanagan, P., & Roginska, A. (2018, May). Evaluation of Binaural Renderers: Localization. In *Audio Engineering Society Convention 144*. Audio Engineering Society.

[12] Perrott, D. R. and Saberi, K. (1990). “Minimum audible, angle thresholds for sources varying in both elevation and azimuth,” The Journal of the Acoustical Society of America, 87(4), pp. 1728–1731.

[13] Ashby, T., Mason, R., & Brookes, T. (2014, April). Elevation Localization Response Accuracy on Vertical Planes of Differing Azimuth. In *Audio Engineering Society Convention 136*. Audio Engineering Society.