Audio Engineering Society

Conference Paper 26

Presented at the Conference on
Immersive and Interactive Audio
2019 March 27 – 29, York, UK

Analysis and training of human sound localization behavior with VR application

Yun-Han Wu[1] and Agnieszka Roginska[1]

[1]*New York University*

Correspondence should be addressed to Yun-Han Wu (natalie.yh.wu@gmail.com)

ABSTRACT

A VR training application is built in this paper to help improve human sound localization performance on generic head-related transfer function. Subjects go through 4 different phases in the experiment, tutorial, pre-test, training and post-test, in which he or she is instructed to trigger a sound stimuli and report the perceived location by rotating their head to face the direction. The data captured automatically during each trial of the experiment includes the correct and reported position of the stimuli, reaction time and the head rotation at each 50ms. The analysis results show that there is a statistically significant improvement on subjects performance.

1 Introduction

Recent advances in audio technology allow us to start reproducing 3-dimensional acoustic environment more and more accurately. Binaural audio has been in the field for a while to deliver spatial audio through headphones and appears to be most suitable for simulating human listening experience in real world scenarios. There are several main factors which decide the quality of binaural audio [1], and the accuracy of spatial sound localization is significantly crucial among them, because it differentiates this new audio format from traditional surround or stereo sound.

However, due to the difficulty of capturing individualized head-related transfer function (HRTF) on a per user basis, generalized HRTFs are commonly used in various types of application these days. It is worth noting that game industry has been researching different methods to improve the quality of spatial audio in VR experiences in the past few years. The main stream of all these researches focused more on making progress on the selection or generation of HRTF. For example, developing a procedural HRTF rendering system or an efficient algorithm to search for the one specific set of HRTF in the database which best fits each user.

On the other hand, this paper proposed a method to solve the degradation of human localization performance resulted from the use of generalized HRTF by building a VR application in Unity to train the subject's localization ability. To be more specific, a training application designed for human sound localization is developed and evaluated in this project. By building a virtual environment in Unity where subjects are surrounded by spatialized sound stimuli, we can access subject's localization ability by asking them to localize sound stimuli while capturing their responses and behavior.

Similar to how humans have been constantly learning to localize sound since we were born, this paper assumed that people can be actively trained to establish the relationship between a sound source being perceived and its corresponding physical location, but in a relatively faster pace. By saying that the application aims to help people improve their localization performance under a specific condition, not only accuracy but also reaction speed and subject's movement pattern will be put into consideration to represent different aspect of the improvement.

2 Literature Review

2.1 Sound Localization Behavior

Human sound localization behavior has been studied in the past few decades thoroughly, and a lot of phenomena are observed and discussed by researchers. First of all, there is a significant difference in the localization blur for directional hearing. An experiment done in [2] examined human's localization blur, defined as the smallest angle difference needed for the person to detect a position-change of the sound source, from 0 to 360 degree on the azimuth plane and -90 [front] to 90 [behind] degree on the median plane. It is concluded that human performs best around 0 degree azimuth [front] and gets worse around 180 degree [back]. Localization blur reaches its maximum value around 90 degree [side], where it attains between 3 to 10 times of the angle in the front.

Another common phenomenon found in various kinds of sound localization experiment is that subjects often recognize sound stimuli to be at the symmetric position respect to the axis of ears, which is generally called Front/Back or Back/Front confusion [3]. The cause of this type of error is that when a sound source is located at the opposite positions relative to the axis of ears, for example, 30 degree and 150 degree azimuth, Interaural Time difference (ITD) and Interaural Level Difference (ILD) for both location will be the same. As a result, subjects have to depend solely on spectral cues to recognize the difference between two sounds, which is relatively hard and requires experience to achieve a better performance.

2.2 Previous Work

Previous researches have proposed various approach to solve the issues caused by the usage of generic HRTF. It is proved in some literatures that alternate HRTF is capable of providing as good of a spatial image as the individualized ones can. By asking subjects to select the datasets which gives a good spatial impression based on certain criteria, such as externalization, elevation discrimination and front/back discrimination, It is found that some sets of HRTF get picked as often as individually measured ones[4]. Another method was proposed to code a particular elevation onto sound stimuli which differs in spectral content in order to represent the height of virtual sound source in an acoustic image[5].

Similar to the method this paper proposed, the following researches focused on solving the problem by putting subjects through some kind of training and can be roughly separated into two categories. The paper in the first group aims to observe how human adapt to a new set of HRTF by actively changing the shape of a person's pinna before conducting the training process, while the experiments in the second group were all using headphone and looked at how the negative influence of generalized HRTF can be reduced through training.

In order to understand how human can relearn sound localization with new cues, an experiment is done by applying mold onto subjects' outer year, and their adaptation process were observed and recorded. The results showed that localization accuracy steadily improved over time in all subjects, and most of the subjects managed to perform reasonably accurate localization after 30 days. Moreover, it is proved that the adaptation to another set of pinna does't influence subject's ability to localize using the original cues[6].

In terms of the second category mentioned in previous paragraph, a systematic review for several experiments on training human localization abilities was done and their experimental designs were compared respectively, test procedure and results in detail. By concluding all of the experiments included in this review, the author pointed out that white pulse Gaussian noise with 100-300ms duration provides the best spatial recognition, and the sound resolution is better if the acoustic signal used in the training phase is periodic and of long continuous duration[7].

Along all the related researches, the importance of visual feedback in sound localization training is examined and discussed[8]. Inside the paper, the results of 3D sound localization assessment before and after the training session, in which visual-feedback was provided to help participants learn the accurate position of each sound object, are compared. It was concluded that this method can reduce localization error caused by the use of non-individual HRTFs, and the effect is proved to last for several days.

Another paper which focused on spatial sound localization in AR environment use a slightly different approach for training subjects[9]. Subjects were given 5 attempts to localize each sound stimuli attached on a visual object. It was observed that they navigated according to the sound first. And after the approximate location was found, they started to depend on visual cues for pinpointing the exact position of the stimuli. This again proves the importance of visual feedback in sound localization training.

3 Methods

3.1 Experimental design

The final goal of this paper is to improve human sound localization performance on generic HRTF using a VR training application built in Unity. A first-person game environment is built to access the localization ability of the subjects, and a training process is designed on top of the same environment to help subjects learn the correct position of a perceived sound stimuli. The default set of HRTF in Unity audio engine is used throughout the experiment.

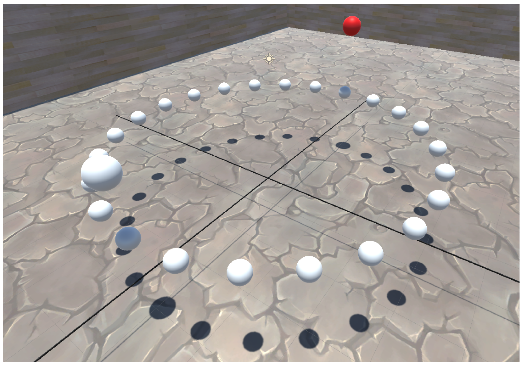

Fig. 1: The scene created in Unity for the application

The application contains 4 different phases, which are the tutorial, pre-test, training and post-test sections. The subjects are placed in a virtual environment through out the whole experiment. Twenty four spheres surrounding the subjects along a fixed circle placed on the horizontal plane serve as the indicators of possible spawn point of sound stimuli and also as an object that the subjects can interact with, which its full function will be explained in detail later. They are placed at eye level and in fifteen degree intervals. On a higher level, two spheres are positioned on top of the 0 and 180 degree spheres, so that the subjects can identify their straight front and straight back easily.

The tutorial part provides users a chance to get familiar with the virtual acoustic environment and the process of going through each trial. A 5-second musical signal sampled at 44100Hz is triggered 8 times at random position chosen from the 24 possible locations around the subject.

On the other hand, the sound stimuli used in the testing part of this experiment is a 800 ms white pulse Gaussian noise sampled at 44100Hz. The pre-test and post-test are essentially the same, except that the results of the frontier is treated as a reference to see if each subjects produce a better results in the latter. During the test, subjects will go through 48 trials in each section, which is derived from the two rounds of twenty-four sound stimuli played from all the possible locations. First, the subjects have to navigate their gaze to the front sphere at higher level and pushed a button on the Oculus touch controller to trigger a sound stimuli. Then, they will be asked to localize and report the perceived position of the sound source as quick and accurate as possible by turning around to face the target and pressing the same button while highlighting the sphere with their gaze.

Finally, the training section is the most crucial part of the research, which puts subject through 48 trials of task similar to the pre- and post-test. However, there are two major differences regarding the experimental design of this section. First, the subjects will receive visual feedback for each triggered sound stimuli, so they can learn to identify the connection between the perceived sound characteristic and the correct location of the sound source.

Another important feature designed for the training section is the stimuli randomization algorithm. Based on what we know about human sound localization behavior, most of the errors found in human localization behavior can be categorized into two groups, localization blur and front/back (or back/front) confusion. People make the first type of error when they identify the sound stimuli to be at roughly the same direction as its correct location but is off by several degrees, while the latter happens when people perceive a sound source to be located at the opposite positions of its correct location relative to the axis of the ears. For the purpose of this paper, rather than randomizing the location of sound stimuli completely, two different kinds of training sessions aim to assist subjects with specific difficulty in sound localization with distinct approaches.

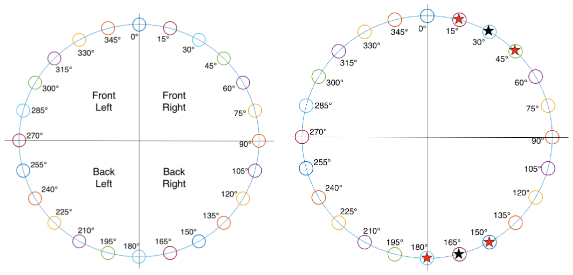

Fig. 2: Representations of two stimuli randomization algorithms

In order to provide the subjects extensive training on front/back confusion, all the stimuli are randomly paired as one at the ventral side and the other at the dorsal side of the human body, and certain pair of stimuli are also located at the same side of the Sagittal plane. (Figure 2 left)

On the other hand, stimuli are presented in pairs and are adjacent to each other in the localization blur focus training. For instance, if the first stimuli in a pair is at the 30 azimuth degree position, the next one in the same pair should be at either 15 or 45 azimuth degree position. Two examples are shown on the right of figure 2.

3.2 Data capturing

A series of data capturing actions is performed after the sound stimuli has been triggered to extract necessary information, which helps access subjects' localization ability and later analyze their behaviors. All of the data gathered from each trial during the capturing phase is recorded in a certain order, so it can be efficiently analyzed later in the process.

In the pre- and post-test phase, the four sets of data collected in each trial are as follows : 1) The correct location of sound stimuli - The position of each sound stimuli can be recorded as an integer in the range of 0 and 360. 2) The reported location of sound stimuli 3) Reaction time - The amount of time it takes for the subject to report their answer after the sound stimuli is triggered. 4) The y rotation value in every 50ms - The azimuth degree of the direction where the subject is facing captured every 50ms.

The test is conducted in the research lab located on the 6th floor of the education building at New York University. The devices used include a 13 inches Alienware laptop, an Oculus Rift VR set and a pair of Sennheiser semi-closed HD650 Headphone. 17 subjects, whose age ranges from 22 to 32, participated in the experiment.

4 Results

4.1 Data preparation and filtering

Data captured in this experiment includes a lot of information and need to be organized before analysis. Moreover, there are a lot of different ways to interpret "an improvement" on sound localization ability. As a result, a particular part of the data needed for certain purpose of statistical analysis is extracted and presented in the following section.

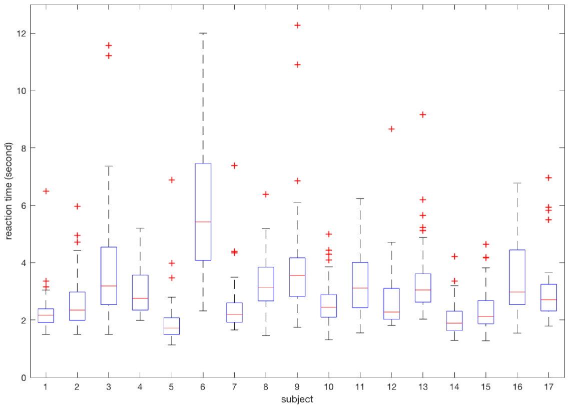

Fig. 3: Distribution of each subjects reaction time data in pre-test

The outliers identified using each subject's reaction time for each trial are excluded from the whole dataset, and the criteria is two standard deviations away from the mean. The reason that only reaction time but not error angle is chosen as the criteria is that the subjects were instructed to react as quickly as possible and the length of the signal is very short. If the subjects spent

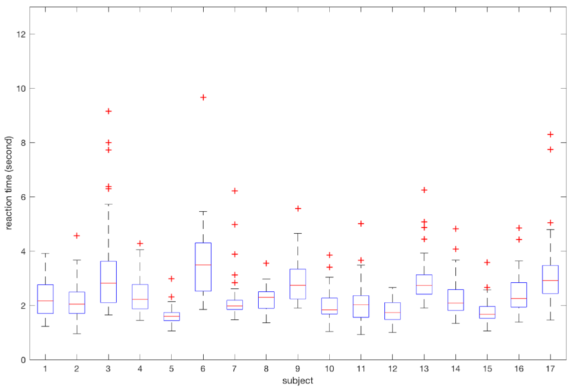

Fig. 4: Distribution of each subjects reaction time data in post-test

too much time on one trial, that specific data failed to represent user's immediate reaction. In the end, 71 attempts were removed from 1564 attempts, which are resulted from the 46 trials for each of the 17 subjects during both pre-test and post-test. The reaction time distribution in pre-test and post-test is presented in Fig. 3 and Fig. 4 respectively.

4.2 Overall performance

In this section, a one-way ANOVA test is performed on 1) the reaction time data in pre-test and post-test 2) the error angle data from the pre-test and post-test to determine whether there is a difference between the localization performance before and after the training.

$$f(x) = \begin{cases} x, & \text{if } x < 90 \\ 180 - x, & \text{if } 90 \leq x < 180 \\ x - 180, & \text{if } 180 \leq x < 270 \\ 360 - x, & \text{otherwise} \end{cases}$$

Fig. 5: The conditional equation for converting the error angle (x) from its original number to the correct representation.

First of all, the error angle data needs to be processed before analysis. Originally, all of them are a value in the range of 0 to 360, since each number is calculated as the absolute difference between the angle of the correct answer and the subject's response. In order to make these numbers reflect the degree of error truthfully, they are re-calculated to a number ranging between 0 and 90 according to the equation shown above. (Fig.5)

As mentioned in the literature review section, front/back or back/front confusion is a common error that happens when people perform sound localization task using generic HRTF. On the other hand, it should also be kept in mind that the error angles are symmetrical on the left and right side of the subject. As a result, the error is converted from a number ranging from 0 to 360 to 0 to 90, where 0 happens at the straight front and back of the subjects when they are facing the target, while 90 happens at the straight right hand and left hand side. Although this calculation method neglect the presence of front/back confusion error, it will be taken care of later in the results section.

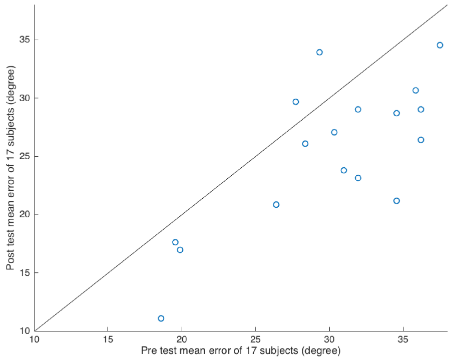

Fig. 6: Scatter plot for each subject's mean of error angle during pre-test and post-test.

	Pre-test	Post-test
Error angle	29.92	24.88
Reaction time	2.80	2.27

Table 1: Mean value of all subjects' error angle and reaction in pre-test and post-test

	Error Angle	Response Time
p-value	0.0321	0.0226

Table 2: Paired t-test results between pre- and post-test.

The mean of error angles for each subject during the pre-test and the post-test are presented in Figure 6. It can be observed that most of the data points fall on the right side of the graphic, which shows the tendency that subjects have a lower error in post-test than in pre-test. On the other hand, the mean value of all subjects' error angle and reaction time are presented in Figure1.

The results of the paired t-test for both error angle and reaction time in the pre-test and post-test are presented in Figure2. It can be observed that the training has a significant effect on subject's sound localization performance judging from both aspects at the p<.05 level.

4.3 Categorized response

Sound localization performance is a relatively complicated concept as explained in the previous section of this thesis. As a result, looking at the values of reaction time and error angle alone is not sufficient to pinpoint how the training process help people perform sound localization better. In this part, every test attempt is categorized into either of the four following groups : 1) correct 2) localization blur error 3) front/back or back/front confusion error 4) neither, using the following criteria. (Fig.7)

$$\text{group} = \begin{cases} \text{correct,} & \text{if } error = 0 \\ \text{blur,} & \text{if } 0 <\mid error \mid \leq 45 \\ \text{f/b or b/f,} & \text{if } 135 \leq \mid error \mid \leq 225 \\ \text{neither,} & \text{otherwise} \end{cases}$$

Fig. 7: Criteria for categorizing each attempt into one of the four groups

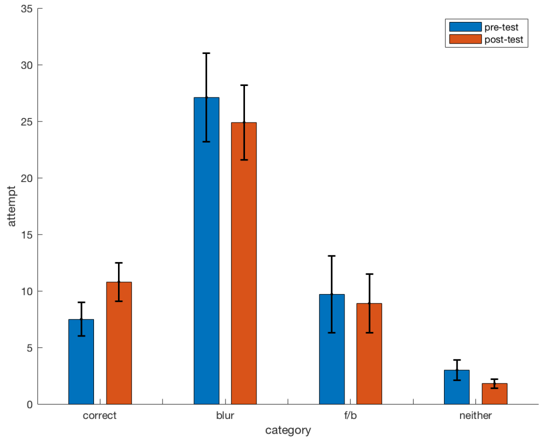

Fig. 8: The mean number of each type of responses made by the subjects in pre- and post-test

After putting all the responses into its corresponding category, we can take a closer look at how exactly the number of each type of response the subjects made changes. Moreover, the mean number of attempts categorized into each group across subjects during pre- and post-test is calculated and compared. The results is shown in Figure 8. The blue bar is the pre-test data and the yellow one is the post-test data, and a general tendency can be observed on this box plot. The correct responses increased from pre-test to post-test, while the other three have decreased.

	Pre-test	Post-test
Blur	26.20	25.5
Front/Back	24.74	23.50

Table 3: The mean error angle for localization blur and front/back confusion

Several extra tests are conducted on the values of error angle calculated for localization blur and front/back (back/front) confusion two response categories to find out whether subject's performances have improved specifically on either of them. A preliminary analysis based on scatter plot demonstrates no clear tendency for improvement on performance. Despite that there is a slight drop in the mean error value (table 2), the paired t-test results turns out to be insignificant.

5 Discussion

The first thing presented in the results part is an overall examination on the mean error angle and reaction time in pre-test and post-test. Since the most important purpose of this thesis is to evaluate the effectiveness of the designed application, a statistically significant decrease on both subject's error angle and reaction time can definitely support the hypothesis that this application improves people's localization performance.

If we take a look at the mean value for all subjects on the error angle and reaction time, we can see that the frontier and the latter went from 29.92 down to 24.88 and 2.80 down to 2.27 respectively. The scatter plot for each subject's mean error angle during pre-test and post-test also clearly demonstrate that 15 out of 17 subjects have lower mean error angle after training. In addition, the paired t-test is used to approve that the true mean difference between the paired samples is significant. The p-value for them are 0.0321 and 0.0226, which are both smaller than the critical value of 0.05. In conclusion, it can be stated that judging from the overall performance, this application does help train subjects to localize sound better in the virtual environment.

The second part of the data analysis focus on grouping the responses according to certain criteria and taking a closer look at whether subjects' improvements vary depends on the type of error they make. The initial observation made on the mean numbers of the four types of responses before and after the training turn out to match with our hypothesis, with the number of attempts in correct category going up and the rest going down, which represents a better localization performance.

Since the training session is conducted as two parts, localization blur and front/back (or back/front) training, operated with distinct algorithms, the error values of these two types of responses were analyzed separately. Looking at the results of paired t-test, both of the p-value 0.5329 and 0.6731 are noticeably larger than the critical value 0.05 demonstrate the fact that the true mean difference between the paired samples is zero.

In the first paragraph, we identified that there is a significant improvement on subjects' overall performance. However, when the responses are categorized into four different types, the results turned out to be insignificant. The first possible explanation for this situations is that the pre-processing procedure for the error value is different in these two kinds of analysis, so it ranges from 0 to 90 under one condition and 0 to 45 under the other. When subjects are trained to locate such a short sound stimuli, they start from learning to identify the rough direction from which it's coming. As a result, obvious decrease in error values, such as from 75 degrees to 45 degrees, can be seen in the overall performance analysis.

On the other hand, the example mentioned above results in a completely different outcome in the other type of analysis. The improvement will reflect on the numbers of various type of responses rather than the value of the error angle itself.

6 Conclusions

In conclusion, the VR sound localization application is proved to be effective based on the results of several different analysis. The decrease of overall error angle, reaction time and the improvement on numbers of each type of error subjects make indicates a better performance in general. However, there are still some aspects that this application needs to be improved. It is important to think about how to refine the design of this application so it has a better effects on helping subjects to resolve localization blur and front/back confusion specifically rather than just the sound localization in general, considering that this is one of the main purpose of designing this application in the first place.

Another thing to keep in mind is that the head movement data has not been analyzed yet. The rotation degree captured in every 50ms is a relatively large set of data, and the nature of these data also makes it hard to perform statistical analysis on them. As a result, a future work will be focusing on whether the head movement onset time influence and the observations made based on the head movement trajectory, which provides us more insight regarding the influence of head motion on human sound localization behavior and whether the training changes the pattern of subject's head movement.

Finally, the data capturing system of this application can be very useful for projects with topics related to HRTF and sound localization. The framework built in it allows the application to gather each subject's data automatically while he or she is doing the localization task. To think in a larger scale, the possibility of incorporating this app with HRTF databases and machine learning concepts to build a HRTF selection program can also be explored. Although there are still a lot of researches that need to done before coming up with a concrete plan, this application can served as a handy tool in such a big project.

References

[1] Rumsey, F., "Spatial Audio : Binaural challenges," in *AES 55th International Conference on Spatial Audio*, 2014.

[2] Haustein, B. G. and Schirmer, W., "A measuring apparatus for the investigation of the faculty of directional localisation," *Hochfrequenztech. u. Elektroakustik*, 79, pp. 96–101, 1970.

[3] Fisher, H. G. and Freedman, S. J., "The role of the pinna in auditory localization," *International Journal of Auditory*, 8, pp. 15–26, 1968.

[4] A. Roginska, T. S. S. and Wakefield, G. H., "Stimulus-dependent HRTF preference," in *129th Audio Engineering Society Convention*, 2010.

[5] R. Susnik, J. S. and Tomazic, S., "Coding of Elevation in Acoustic Image of Space," in *the 2005 convention of Australian Acoustical Society*, 2005.

[6] P.M. Hofman, J. V. R. and Opstal, A. V., “Relearning sound localization with new ears,” *Nature Neuroscience*, 1(5), 1998.

[7] O. Balan, A. M., F. Moldoveanu and Asavei, V., “Experiments on training the localization abilities: A systematic review,” in *The 10th International Scientific Conference*, 2014.

[8] P. Zahorik, K. W., C. Tam and Bangayan, P., “Localization accuracy in 3D sound displays: The role of visual-feedback and training,” in *University of California, CA*, 2003.

[9] J. Sodnik, R. G., S. Tomazic and Duenser, A., “Spatial sound localization in an Augmented reality environment,” in *Proceedings of the 2006 Australasian Computer-Human Interaction Conference*, 2006.