On the Impact of Training HRTF-Based Auralisation

Catarina Mendonça Jorge A. Santos
Escola de Psicologia, Univ. do Minho
Braga
mendonca.catarina@gmail.com
jorge.a.santos@psi.uminho.pt

Guilherme Campos Paulo Dias
DETI/IEETA, Univ. de Aveiro
Aveiro
{guilherme.campos,paulo.dias}@ua.pt

João P. Ferreira
Escola de Engenharia, Univ. do Minho
Guimarães
jpg.ferreira@gmail.com

Abstract
Auralisation with generic, non-individualised, Head Related Transfer Functions (HRTF) is common practice, as obtaining individualised HRTFs poses very serious practical difficulties. It is therefore extremely important to assess to what extent this hinders our 3D sound localisation capabilities. Here, we address this issue from a learning perspective. We carried out a set of experiments to better understand how the 3D virtual source localisation performance of listeners using generic HRTF is influenced by training. In experiment 1, we observed that listeners perform fairly well in terms of azimuth discrimination, but mere exposure to the tests does not cause performance improvement. In experiments 2 and 3 we implemented a short training period on both azimuth (exp2.) and elevation (exp. 3) discrimination. Training involved active learning and feedback and led to significantly better results. We therefore propose that in order to fulfil its perceptual potential, auralisation based on generic HRTF sets should always be preceded by a period of training.

Keywords
Head-Related Transfer Functions, Virtual, Acoustics, Auralisation, Learning, Azimuth, Elevation.

1. INTRODUCTION

Auralisation can significantly increase the feeling of immersion and improve interaction accuracy in virtual environments by providing spatial feedback and complementing visual information.

The assumption underlying earphone-based virtual acoustics is that providing a person with the same binaural stimuli he/she would get in a real environment yields the auditory perception of being in that same environment. Significant technical and scientific efforts have been carried out in recent years to create and perfect such virtual sounds. An auralised sound should provide all the necessary cues to accurately locate its virtual source. For this purpose, it is necessary to simulate room reflections, attenuation effects, interaural time and level differences (ITD and IID, respectively), as well as the shaping produced by the listener's head and pinnae. The interaural differences and other effects of the interaction of a sound wave with torso, head, pinnae (outer ears) and ear canals can be mathematically described by the binaural impulse response for the corresponding source position: the *Head Related Impulse Response (HRIR)*, or, more commonly, its Fourier transform, known as *Head Related Transfer Function* (HRTF). Given that HRTFs depend upon anatomical structures, they differ from person to person. However, given the time and effort involved in obtaining individual HRTFs, most auralisation applications rely on average sets of HRTFs measured on appropriate acoustic research manikins.

It is widely accepted that satisfactory auralisation can be obtained using average HRTFs [Loonis99]. However, there are also several reports of intracranial sound localisation and it is not yet well established how humans adapt to hearing sounds through non-indivualised HRTFs.

[Wenzel93] compares localisation accuracy using external free-field acoustic sources and virtual sounds filtered by non-individualised HRTFs. This study has revealed several front-back and up-down confusions and overall similarity of the results obtained in the two test situations. On the other hand, when listeners are asked to localise complex movements, results are better with a 24-speaker system than with HRTF spatialisation [Ballas01]. Comparing the perception of sounds with individualised against generic HRTFs, the former offer significant improvement [Valjamae04].

In this paper, we address this issue taking into account the influence of learning processes. The adult human brain is able to improve its auditory localisation performance, as connectivity and response properties of the neurons are shaped by experience [King99]. There is also evidence that humans may learn to locate acoustic sources with drastically different ears [King01].

The objective of this study was to assess how training may influence the use of non-individualised HRFT. The experiments were intended to: (1) Understand the baseline localisation accuracy in subjects who had never experienced auralisation with non-individualised HRTF before; (2) Analyse the temporal evolution of localisation accuracy by simple exposure to the test sounds; and (3) Test a brief training model combining active learning and feedback.

2. EXPERIMENT 1

In Experiment 1, we tested the accuracy of inexperienced listeners in localising sounds at fixed elevation and variable azimuth. In 10 consecutive experimental sessions, we analysed the evolution of the subjects' performance as they became gradually more familiarised with the stimuli.

2.1 Method

2.1.1 Subjects

Four naïve and inexperienced young adults participated in the experiment. They all had normal hearing, verified by standard audiometric screening at 500, 750, 1000, 1500 and 2000 Hz. All auditory thresholds were below 10 dB SPL and none had significant interaural sensitivity differences.

2.1.2 Stimuli

The stimuli consisted of pink noise sounds. Pink noise is characterized by its power density being inversely proportional to the frequency. This provides equal energy in all octave bands, and therefore the human auditory system (which behaves analogously) perceives approximately equal magnitude in each frequency. The sounds were auralised at 8 different azimuths: 0º (front), 180º (back), 90º (left and right), (45º left and right), and 135º (left and right). They had constant elevation (0º) and distance (1m).

In the auralisation process, the original sound is convolved with the HRTF pair corresponding to the current source position. The resulting pair of signals – for the left and the right ear – is then reproduced through earphones. These HRTFs were recorded using a KEMAR dummy head microphone at the Massachusetts Institute of Technology [Gardner94]. Sounds were reproduced with a Realtec Intel 8280 IBA sound card, and presented through a set of Etymotics ER-4B MicroPro in-ear earphones.

2.1.3 Procedure

All sounds were presented pseudo-randomly for 3 seconds with a 1 second interstimulus interval. There were 10 blocks of 10 stimulus repetitions each, and each block had a 5,3 minute duration. Participants were told to indicate the perceived sound source location for each stimulus. The answers were recorded by selecting, on a touch screen, one of the eight possible stimulus positions.

2.2 Results and Discussion

The average percentage of correct answers for each stimulus position is presented in figure 1. As there were 8 possible answers, random answers would result in 12,5% correct answers. Thus, all results were well above chance.

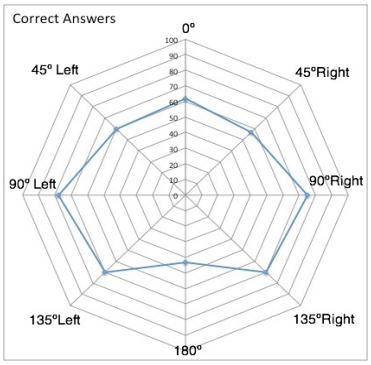

Figure 1: Percentage of correct answers for each stimulus azimuth.

As [Wenzel93] had observed, there were several front-back confusions that account for the lower accuracy at 0º, 180º, 45º and 135º. Indeed, the left and right 90º sounds were the most accurately located, with a correct response rate of 78%. The average accuracy of all azimuth localisation was significantly above chance (65%), but no ceiling performances were observed.

Analyzing the average participant's performance along time (Figure 2), we see that in spite of small fluctuations, accuracy remained largely constant.

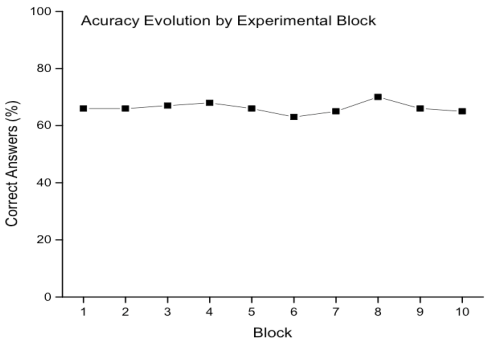

Figure 2: Average performance evolution through time

Our results reveal that naïve participants are able to discriminate sounds at several azimuths well above chance, but without ceiling performances. Throughout the exposure blocks, their accuracy does not evolve, leading to the conclusion that simple exposure is not enough for significant localisation improvement in short periods of time.

Taking these conclusions into account, a second experiment was developed where, in the same amount of time, listeners were trained to discriminate the localisation of several sounds.

3. EXPERIMENTS 2 AND 3

In experiment 2, we tested the participants' accuracy in localising sounds at several azimuths before and after a

short training program. In this training, we selected only a small portion of sounds and trained them through active learning and response feedback. In experiment 3, the same methodology was used, in an elevation discrimination task.

3.1 Method

3.1.1 Subjects

Four young adults served as participants. None of them had any previous experience with virtual sounds. They all had normal hearing, tested with a standard audiometric screening, as described in experiment 1.

3.1.2 Stimuli

As in experiment 1, all stimuli consisted of pink noise sounds, auralised with the same algorithms and software.

In experiment 2, the stimuli varied in azimuth, keeping elevation (0º) and distance (1m) fixed. Azimuths ranged from the front of the subjects head to their right ear, spaced at 6º intervals (6º left, 0º, 6º right – 96 º right). In experiment 3, the stimuli varied in elevation, but not in azimuth (0º) or distance (1m). They ranged from the front of the listeners' head to the top in 10º intervals (0º - 90º).

All sounds had a 3 second duration, with 1 second intervals between them.

3.1.3 Procedure

Both experiment 2 and 3 started with a pre-test. In the pre-test, all sounds were presented pseudo-randomly 4 times. Participants had to indicate, on a continuum displayed on a touch screen, the point in space where they expected the sound source to be.

After the pre-test, participants engaged in a training period. In experiment 2, the trained sounds corresponded to azimuths 0º, 21º, 45º, 66º and 90º. In experiment 3, the sounds were at elevations of 0º, 50º and 90º. The training conformed to the following steps:

- *Active learning*: Participants were presented with a sound player where they could hear the training sounds at their will. To select the sounds, there were several buttons on a screen, and each button clearly displayed which sound it triggered (the sound's position). The buttons were themselves arranged in the screen according to their respective position in space. For example, if the listener chose a button on the top of the screen and afterwards another button below it, he/she would hear a first sound on the top and the second sound at a lower elevation. They were informed that they had 5 minutes to learn the sounds and that afterwards they would be tested.
- *Passive Feedback*: After the 5 minutes of active learning, participants heard the training sounds and had to point their location. After each trial, they were told the correct response. Therefore, if they gave the wrong answer, they would be able to learn the correct one. The passive feedback period continued until participants could answer correctly in 80 percent of the trials (after 5 consecutive repetitions of all stimuli with at least 20 correct answers in experiment 2 and 12 correct answer in experiment 3).

After the training period, subjects performed the post-test: a testing session equal to the pre-test, to assess the discrimination differences.

3.2 Results and Discussion

3.2.1 Experiment 2

The discrimination results of experiment 2 are presented in figure 3. The accuracy was measured as the average of the differences between the stimulus position and the response position, in azimuth degrees. The dashed line corresponds to the average error of a person responding randomly.

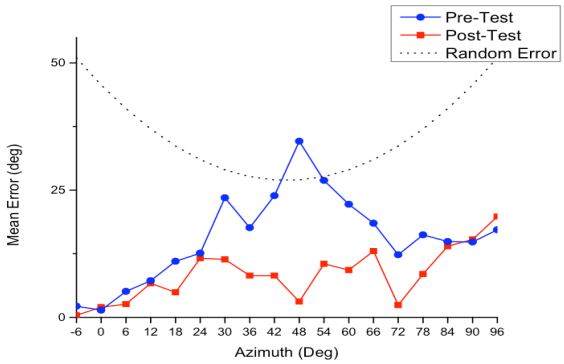

Figure 3: Average angle error in pre and post-test for azimuth

Analysing the pre-test (red) curve, we observe that azimuth discrimination is easier for frontal stimuli, where average errors are below 5 degrees. These results might be explained by the fact that there were no stimuli located at the back of the head, and therefore all front-back confusions were prevented. Similarly to the results of experiment 1, listeners were also fairly precise in identifying sound positions when these were presented laterally. On the other hand, sounds were most difficult to locate in the intermediate azimuths, between 40º and 50º. For these sounds, pre-test localisation errors were maximal. A short analysis of response accuracy along time revealed that listeners were as accurate in the beginning of the test session as in the end, confirming that simple auditory contact does not provide performance enhancement.

The training sessions were very successful for all participants. All took less than 30 minutes and in average, they lasted for 22 minutes.

The post-test results (blue curve) revealed a large error reduction (7.23º in average). This difference was statistically significant in a paired samples T-test ($t_{(287)}$=14.94, p≤0.001). This reduction was most expressive in the intermediate azimuths, where the average error decreased 20 degrees. Analysing the trained azimuths (0º, 21º, 45º, 66º, 90º), we observe that performance enhancement was substantial not only for these stimuli, but also for others, not trained. As an example, the best error reduction was obtained with the 48º azimuth, a non-trained stimulus. On the other side, the 90º azimuth, a trained one, revealed similar results in both sessions. These findings allow us to conclude that the trained discrimination abilities for some stimuli positions are generalized to other, non-trained, auditory positions.

3.2.2 Experiment 3

Figure 4 displays the elevation discrimination results from the pre and post-test sessions.

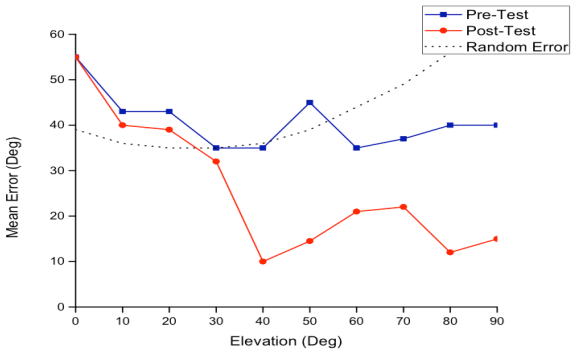

Figure 4: Average angle error in pre and post-test for elevation

Overall, participants were less accurate estimating a sound position in elevation than in azimuth. In the pre-test session, the average error was 40.8º, similar to random error. In this session, results were worst in the frontal stimuli (55º average error), but there were no large differences in acuity among all sound elevations. Again, performance was similar in the beginning and in the end of the session, confirming the absence of learning effects during exposure without feedback.

Training sessions were faster than those of experiment 1, as there were only 3 trained elevations. On average, they took 17 minutes. Only one subject did not evolve as expected. After 10 minutes testing, this subject was still making excessive mistakes, and was allowed a second learning phase, after which the 80 percent accuracy was rapidly achieved.

The post-test results were better than those of the pre-test for all subjects. This difference was significant in a paired samples T-test ($t_{(159)}$=16.678, p≤0.001) The average error decreased 14.75 degrees, more than in experiment 2. The training effect was most expressive for the upper stimuli, namely at 80º, 40º and 50º elevations. Among these stimuli, the only trained one was at 50º. On the other hand, sounds at 0º elevation, a trained stimulus, revealed no decrease in the post-test session. Similarly to what was found in experiment 2, training was highly effective and well generalized to other stimuli.

4. FINAL DISCUSSION

In this paper we intended to analyze the accuracy of listeners in locating virtual sound sources generated with non-individualised HRFTs. We also intended to analyze the evolution of this ability along short periods of time.

In experiment 1, we assessed the acuity of the subjects in azimuth discrimination in the course of 10 testing blocks where stimuli were successively heard but no feedback was provided. The results were well above chance, but no learning effect was detected. Indeed, [King01] had reported auditory learning with severely altered ears, by mere exposure to the new sounds, but such learning took 19 days to be complete. Such a lengthy training period is impractical for most virtual sound applications.

In experiments 2 (variable azimuth) and 3 (variable elevation), we aimed at performance improvement through a training program, which combined active learning and testing with feedback. Both experiments showed there was significant performance improvement in localisation after training. This improvement was not restricted to the trained sound positions, but generalized to other source locations as well.

We conclude that in binaural auralisation using generic HRTF, it is possible to improve significantly the auditory performance of a naïve subject in a short period of time. However, this could not be achieved by mere exposure to the auralised sounds. Given the poor accuracy levels observed before training, we argue that virtual sounds with non-individualised HRTFs should only be used after short learning sessions. We propose such sessions might involve active learning, feedback and training small samples of sounds, as the new hearing abilities generalise to untrained locations.

5. ACKNOWLEDGEMENTS

This project was partially supported by the Portuguese Foundation for Science and Technology (SFRH/BD/36345/2007 and PTDC/TRA/67859/2006).

6. REFERENCES

[Ballas01] J. A. Ballas, H. Fouad, D. Brock, J. Stroup. The effect of auditory rendering on perceived movement: Loudspeaker movement and HRTF. *Proceedings of the International Conference on Auditory Dysplay,* 2001, Finland

[Gardner94] B. Gardner, K. Martin. HRTF Measurements of a KEMAR Dummy-Head Microphone. url: http://sound.media.mit.edu/resources/KEMAR.html visited - June 2010.

[King99] A. J. King. Auditory perception: Does practice make perfect? *Current Biology*, 9, 143.146

[King01] A. J. King, J. W. H. Schnupp, T. P. Doubell. The space of ears to come: Dynamic coding of auditory space. *Trends in Cognitive Sciences*, 5, 261-260.

[Loonis99] J. M. Loonis, R. L. Klatzky, R. G. Golledge. Auditory distance perception in real, virtual and mixed environments. In Y. Ohta, H. Tamura (Eds.) *Mixed Reality: Merging Real and Virtual Worlds*, Tokio: Ohmsha.

[Valjamae04] A. Valjamae, P. Larson, D. Vastfjall, M Kleiner. Auditory pressure, individualized Head-Related Transfer Function, and illusory ego-motion in virtual environments. *Proceedings of the Seventh Annual Workshop in Presence*, 2004, Spain

[Loonis99] J. M. Loonis, R. L. Klatzky, R. G. Golledge. Auditory distance perception in real, virtual and mixed environments. In Y. Ohta, H. Tamura (Eds.) *Mixed Reality: Merging Real and Virtual Worlds*, Tokio: Ohmsha.

[Wenzel93] E. M. Wenzel, M. Arruda, D. J. Kistler, F. L. Wightman. Localisation using nonindividualized Head-Related Transfer Functions. *Journal of the Acoustical Society of America*, 94, 111-123.