Audio Engineering Society

Convention Paper 6800

Presented at the 120th Convention
2006 May 20–23 Paris, France

Usability of 3D-Sound for Navigation in a Constrained Virtual Environment

Antoine Gonot[1,4], Noël Chateau[2], and Marc Emerit[3]

[1] France Télécom R&D, Lannion, 22300, France
antoine.gonot@rd.francetelecom.com

[2] France Télécom R&D, Lannion, 22300, France
noel.chateau@francetelecom.com

[3] France Télécom R&D, Lannion, 22300, France
marc.emerit@francetelecom.com

[4] CNAM, Laboratoire CEDRIC, Paris, 75003, France

ABSTRACT

This paper presents a study on a global evaluation of spatial auditory displays in a constrained virtual environment. Forty subjects had to find nine sound sources in a virtual town, navigating by using spatialized auditory cues that were delivered differently in four different conditions: by a binaural *versus* a stereophonic rendering (through headphones) combined by a contextualized *versus* decontextualized presentation of information. Behavioral data, auto-evaluation of cognitive load and subjective-impression data collected via a questionnaire were recorded. The analysis shows that the binaural-contextualized presentation of auditory cues leads to the best results in terms of usability, cognitive load and subjective evaluation. However, these advantages are only observable after a certain period of acquisition.

1. INTRODUCTION

As a global telecommunications operator, France Telecom is willing to experiment new technological innovations in the hands of customers in order to enhance its user experience. These innovations can modify user experience on several dimensions such as perception (e.g. quality, fidelity of a new coding), cognition (e.g. reduction of cognitive load by a new form of presentation of information) and ease of use (e.g. improvement of usability by a new possibility of interaction).

Sound spatialization is one of the mature technologies that could bring large benefits to customers on these several dimensions. However, although many studies have shown the benefits of spatialization techniques in auditory display, none has been focused on the global evaluation (in terms of perception, cognition and usability) of its advantages as compared to "classic" audio rendering techniques, such as the widespread stereophony.

Spatialization techniques in auditory display can be used, for example, to enhance the feeling of immersion in a Virtual Environment [1] or, thanks to the "cocktail party" effect, to improve intelligibility when multiple sound sources are presented concurrently [2]. However, there are tasks for which spatial sound has an interest for itself, i.e. when information is conveyed by spatial cues rather than by the sound itself. An example of such task is auditory navigation in a virtual environment.

Evaluating the usability of a spatial auditory display is different from classical psychoacoustic experiments. According to Walker and Kramer [3], *first there is "simple" perception […]. Second, there is subtask of parsing the auditory scene into sound sources or streams […]. Finally there is the subtask of associative and cognitive processing*. Then perception is only one stage of a complex process for meaning construction and decision. More generally, the distinction is made on what is called "external validity". For achieving its goal of valid evaluation, a usability test generally requires to simulate a real situation of usage. So it is impossible and often not required to apply by a systematic control of independent and dependant variables. Consequently, in the some cases, qualitative methods are more appropriate than quantitative one.

The usability of an interface is usually determined by different criteria, related to user's behaviour (e.g. learning time, execution speed, number of mistakes, etc.), to user's satisfaction and to the interface's suitability to the task. More precisely *usability is a combination of effectiveness* (accuracy and completeness of goal achieved)*, efficiency* (resources expended in relation to the accuracy and completeness of goal achieved) *and acceptance* (user's subjective level of satisfaction when using the system) *with which a user achieve particular goals in a particular environment* ([5] and [6], quoted in [4]). To develop a global evaluation of spatial auditory displays, the usability evaluation will be completed by subjects' own evaluations on cognitive level (cognitive load) and on perceptive level (impressions).

Non-speech sounds for navigation aid, also called beacons, have already been studied but there are still open questions that our work addresses. Firstly, previous works [3] [7] have only focused on the effect of beacon types (e.g. speech, sonar pulse, burst of broadband noise, pure sine wave, etc.) on localization and navigation. Secondly, experiments have generally been limited to the case of an empty room, without any obstacles, and the problem of navigation has never been addressed globally. It means that the task is generally to go directly from one target to one another, like in [8], and never to find its own way in a complex environment (e.g. a town, either virtual or real). So, it might be considered that Walker and Lindsay's experiment in [8] is rather similar to a psychoacoustic test than a usability one. Finally, except the study on "waypoint capture radius" [9], there is no discussion about the semantic relation between spatial location and the way to represent it by means of spatial auditory cues.

So, it seems necessary to determinate a reference for evaluating the spatialization contribution of an auditory display. Stereo rendering is naturally chosen to be this reference, because it presents fundamental cues for localization (Interaural Intensity and Time Differences) and because most people are familiar with it. Moreover, this evaluation is multi-criteria because the contribution is polymorph and can usability (effectiveness, efficiency and acceptance), cognitive load and perception of quality.

A game environmenrt was selected for support to this study firstly because it's the only mass market application in which interactive 3D sound has a short, but significant history. Even if sound rendering is not developed as much as graphics rendering, home theater systems for personal computers begin to be relatively cheap and common. According to R. Bridgett [10], "the success of the DVD format, for both games and movies, can be viewed as an encouraging and pre-emptive mass consumer move into the surround sound market". However, even if sound for games is still inheriting expectation of motion picture surround, it is really an open and free field for exploring interactive 3D sound in a more abstract way. Secondly, auditory displays are first dealing with usability, and the particular context of entertainment allows us to focus not only on the "effectiveness" dimension but also on the "acceptance" one. Our preconception is that it is more difficult to have a positive judgment on the design of the display if the task itself is boring. Then, it is interesting to design a game-like test, to motivate and engage the tests

subjects. Lokki et al. have also retained this approach for there multimodal experiments in [11].

More specifically we focus on a navigational task in a constrained environment because it allows us to highlight interactions between the spatial sound rendering method (e.g. binaural *vs.* stereo) and the information semantic, i.e. its meaning. According to H Hu and D-L Lee, this meaning can vary from one application to one another. For distance semantic, "some will adopt Euclidian distance, another like finding nearest-restaurant application may view distance as road-network distance, and another like an automatic robot controller may see the distance as the energy consumption along the path" [12]. The opposition between Euclidian distance and road-network distance is particularly relevant to this study, because it can also be transposed to azimuth. What opposes those two semantics is whether or not the constraints of the environmental context (i.e. town configuration) are taken into account. Therefore, the first one is defined as *decontextualized* and the second one as *contextualized*. This contextualization contrast is one of the two independent variables which structure our experimental design. At last, because environmental context is essentially acquired during navigation by means of vision, the manipulation of this variable is also an indirect way to evaluate audiovisual interaction.

The experiment described in this paper is a "first person" game-like test, in which the task is to navigate in a simplified virtual town to find the location of ecological sound sources. The two independent variables are: "How is the sound rendered" (Stereo *vs.* Binaural) and "How is indicated the position of the target" ({(Direction, Distance) = polar coordinates of the target} *vs.* {(Direction, Distance) = (Direction, Distance) defined by the shortest path toward the target}). The hypothesis underlying this experience is that usability is correlated with the *precision* of the auditory spatial information, its *richness* and its *coherence* with the visual environment explored. The dependent variables are divided in three groups: one related to the behaviour (length of the path to find the source, time to decide where to go at each crossroad and principal listening direction at each crossroad), one related to cognitive processes (NASA-TLX, memorization of sound sources positions) and one related to acceptance (questionnaire about satisfaction, immersion, perception of spatial location of sound, etc.).

In the following, the virtual environment created for the test and the dependant variables will be described, then, results will be presented and discussed.

2. THE TEST ENVIRONMENT

2.1. Game application

2.1.1. Principle

Nine sound sources are hidden in a town and the goal is to find successively these sources. When the game starts, the first source is presented in front of the listener, with a level corresponding to the minimum distance defined by the distance/attenuation law. A word describing this source is displayed simultaneously on the screen. For example, if the sound source is a church bell ringing, then the subject will see the word “church” on screen. There is no limitation time for listening to the target, because it is not possible to recall it during navigation. When the subject is ready, he/she presses the space bar on the keyboard, and the game begins. Because, there is no visual representation of the sound source, the only way to find it is to guess the direction of it and to go always closer until the source is found and the program presents the next target. All the nine sources are heard during the entire navigation.

2.1.2. Interaction

The player moves a first person camera (i.e. egocentric view) along the town by the aid of the arrow keys of the computer keyboard. In order to avoid the effect of differences in sleight between subjects, the interaction capacities are limited to the minimum. The controls are:

- Key "right arrow" and "left arrow": there keys are used for rotation. The player can let the key pressed to rotate continuously at a constant angular speed of 1.309 rad.s^{-1} (75°.s^{-1}).

- Key "up" is used to go forward. The subject just has to press the key once to go automatically to the next crossroad, even if the first-person camera is not oriented exactly in the direction of a street. The application chooses the closest direction of the camera orientation.

- Key "down" is used to go backward. In the same way the key "up" does, the subject just has to press the key one time to be brought back to the previous crossroad. This key is just a cancellation one, so it is

no more available once the subject has reached the next location.

2.1.3. The network

Because navigation in a town is a succession of choices of directions to take, the 3D model of the road-network has been simplified drastically so navigation is like moving from square to square on a chessboard. So when the subject is on a crossroad, there are maximum four different directions. However, it can happen that two nodes are adjacent (Cf. figure 1 (*)).

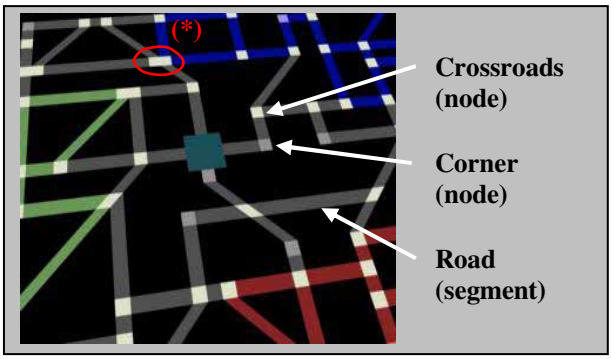

Figure 1 : road-network

Finally, three different zones were created (red, blue and green), in order to facilitate the source location recall during evaluation at the end of the game.

Moreover, the starting point is a bigger node than the other, located in the center of the town. The zones are the only explicit visual landmarks in the environment. Figure 2 shows a top view of the town.

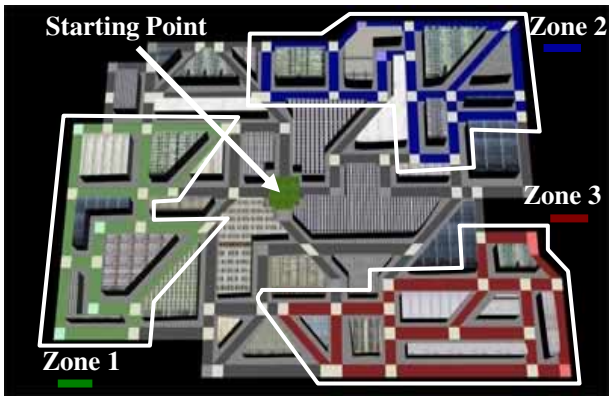

Figure 2 : the town

2.2. Sound rendering

2.2.1. Stereo rendering

The stereo rendering, mode is a model of an ORTF stereo technique. This technique uses two first order cardioid microphones with a spacing of 17 cm between the microphone diagrams, with an 110° angle between the capsules (see Figure 3). The spacing of the microphones emulates the distance between human ears, and the two directional microphones emulate the shadow effect of the human head. This technique provides the two fundamental cues for localization in the horizontal plane: ITD (Interaural Time Difference) and IID (Interaural Intensity Difference).

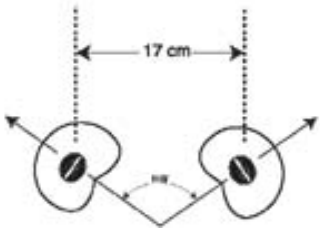

Figure 3 : ORTF stereo technique

2.2.2. Binaural rendering

The binaural rendering is expected to provide exact reproduction of the acoustic field over headphones. The azimuth and elevation of the virtual sound source are controlled by means of two sets $\{H_D,H_G\}$ of filter coefficient from database of HRTFs (Head Related Transfer Functions), according to the specified direction (see Figure 4). In this experiment, HRTF are not individualized.

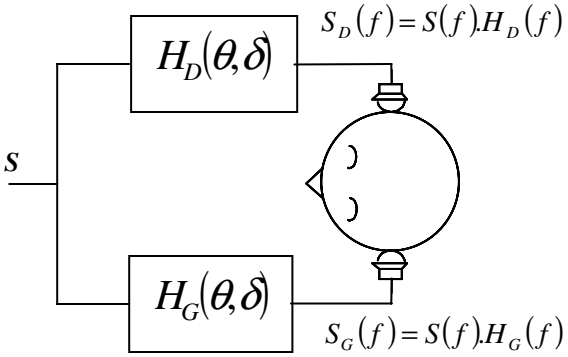

Figure 4 : binaural technique

2.3. Auditory presentation of information

The information presented to the user in order to assist him/her in the navigation task is the distance and the direction of a particular location in the environment. The direction is simply mapped to the azimuth of a sound source and distance to the sound level using the classical inverse square law. So, except final sound rendering (binaural or stereo), the data-to-sound mapping is always the same. Only the data semantics is manipulated. As mentioned above it can be either *contextualized* or *decontextualized*.

2.3.1. *Decontextualized* representation

For this semantic, as illustrated in figure 5 the couple (Direction, Distance) is determined by the polar coordinates of the target.

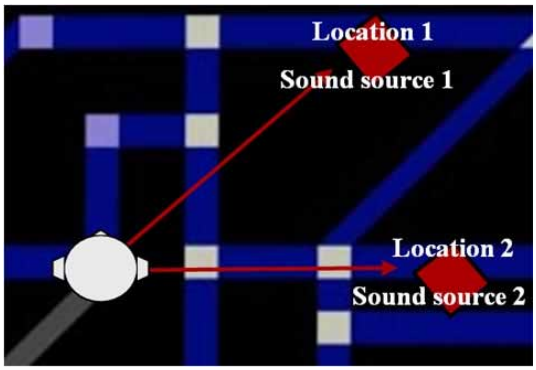

Figure 5 : *decontextualized* representation

- **Navigation task:** the task is similar to navigation with a compass. In this case, the best direction to reach the source is impossible to guess, because obstacles constraint vision to local perception. Then the listener can go the wrong way. He will not be lost, but decision time and the covered distance will not be optimal.

- **Perception subtask:** the topology of the nine sound sources is stable, that is to say, except for camera movement, their relative position does not change over time (from one crossroad to one another). Moreover, sound sources can be in any direction because direction has continuous values. In opposition to the *contextualized* representation, the probability that two sources have the same direction is very low. Then, in general it is easier to discriminate sound sources according to their spatial location. Those two characteristics facilitate the auditory scene perception.

2.3.2. *Contextualized* representation

For this semantic the distance is the length of the path to the location. As illustrated in figure 6, the direction is given by the first node (should it be a corner or crossroad) of the path toward the location.

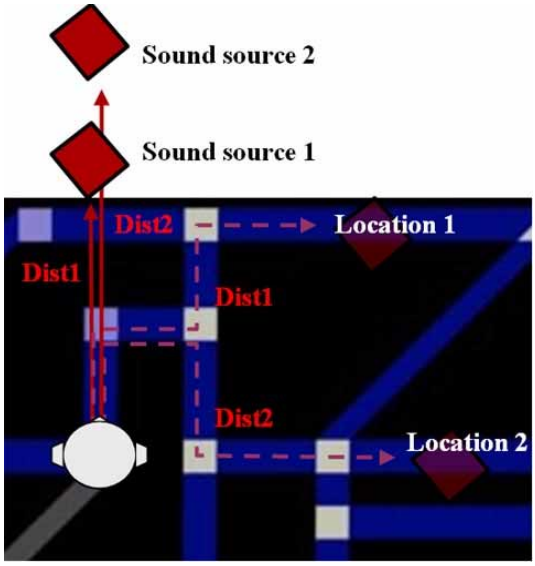

Figure 6 : contextualized representation

- **Navigation task:** the task is easy because the sound sources indicate directly the right way to the location. Then decision time and covered distance should be optimal.

- **Perception subtask:** In opposition to the previous representation, the sound sources direction has discreet values. So, as in figure 6, different sources can share the same direction, which makes the auditory scene perception more difficult. Discrimination must be made on temporal and timbre characteristics.

2.4. Sound source design

The sound sources are every day sounds that could be heard in a town and which can be described easily and without any ambiguity by one or two words. Their description is the following:

- **Fireworks:** loud explosion followed by low level whistle and then a loud crackle.
- **Church:** bell ringing.
- **Guignol Theater:** children yelling.
- **Hospital:** ambulance's siren.
- **Port:** seagull's cry and low level sound of water in background.
- **Roadwork:** jack hammer's sound.
- **Stadium:** crowd singing.
- **Train:** train passing
- **Fanfare:** two bars of a piece for fanfare.

Each sound is 5 seconds long and is played in loop when the player is seeking it. However because hearing simultaneously nine short sounds in loop is extremely cacophonic, we decided to minimize superposition in inserting silences between their occurrences. Then, for each source, we created a second sound file, used when the correspondent sound source was not being sought. As we see in figure 7, it contains two occurrences of the sample separated by two silences, respectively 11 seconds and 15 seconds long. We used two different lengths for silences to limit annoying repetition. So, thanks to an offset of 4 seconds between each sound source, there are never more than six sounds played simultaneously.

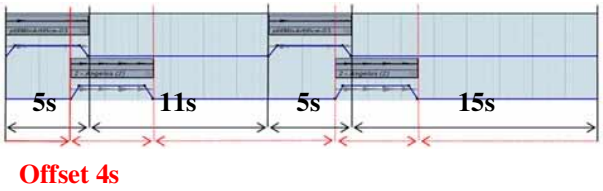

Figure 7 : auditory scene design

3. THE EXPERIMENT

3.1. Task and independent variables

There are two independent variables which concern respectively the rendering mode of sound and the semantic of the data to represent. Each of those variables has two levels:

Binaural ***vs***. Stereo

×

Contextualized ***vs.*** Decontextualized

Then there are four conditions to test:

- "BinCont": binaural rendering for *contextualized* representation of location.
- "BinDecont": binaural rendering for *decontextualized* representation of location.
- "SteCont": stereo rendering for *contextualized* representation of location.
- "SteDecont": stereo rendering for *decontextualized* representation of location.

Those four conditions are distributed to four separate groups of subject according to an inter-group experimental design.

3.2. Dependent variables

The dependent variables are divided into three groups:

Objective evaluation of interaction:

- Mean elapsed duration at crossroads
- Normalized covered distance
- Orientation frequency

Subjective evaluation of workload and memorization

We use the auto-evaluation of the NASA-TLX, which is "a multi-dimensional rating procedure that provides an overload workload score based on a weighted average of ratings on six subscales" [16]: Mental demands; Physical Demands, Temporal Demands, Own Performance, Effort, and Frustration.

Immediately after the subject has found the last source, we also ask to the subject, to report the nine locations one a map.

Impression questionnaire

The questionnaire consists of twelve assertions that the subject can negate (resp. confirm) gradually. The rating is achieved by means of a scale presented as a line divided into 7 intervals anchored by bipolar descriptors (Absolutely/Not at all). The assertions are relative to the following notions: sound quality, easiness of using sound to navigate, engagement, entertaining, immersion, coherence of sound with the environment, easiness of localizing sounds, easiness of the task, general appreciation, 3D sound effect and appreciation of this effect.

3.3. Test procedure

Apparatus

The game has been designed with Virtools and is running on a personal computer, using a 19" cathodic video screen and a Sony MDR-CD1700 Digital Reference closed Headphone as displays.

Participants

Forty subjects participated to the experiment and were distributed in four groups of ten, one per experimental condition.

Half of them were researchers of the France Telecom R&D laboratory (not necessarily sound experts). The other part was recruited outside. The subjects are between 15 and 45 years old, they do not have any known auditory disability and they regularly use personal computer at home or at their office. There are 65% of men (resp. 35% of women) and 79% of right-handed (resp. 21% of left-handed). They were asked how frequently they play to video game letting them the choice between the four answers: (1) "Never", (2) "Few", (3) "Often", and (4) "Very often". The results indicate that, in general, the subject played infrequently to video games (2^{nd} answer).

Instructions

First, participants had to read a three pages instruction presenting:

- The principle of the game, i.e. what the task was.
- The controls, i.e. how to use the interface
- The different steps of the post-evaluation: reporting the sources location on a map, a two-part simplified TLX evaluation and 12 general questions about their experience.
- The definition of the 6 factors used for the determination of the overall workload score, directly translated from NASA-TLX's guide [16].
- The definition of the terms "engagement" and "immersion", used for the impression questionnaire.

When one participant had finished to read the instructions, the experimenter looked over those with him/her and responded to any questions to insure everything was understood correctly.

Training

A smaller environment has been designed for the training phase, in which a unique sound source ("train") has been located. Figure 8 shows a top view of the environment. It presents all the characteristics of the test environment except zones: crossroads, corners, adjacent corners and oblique roads.

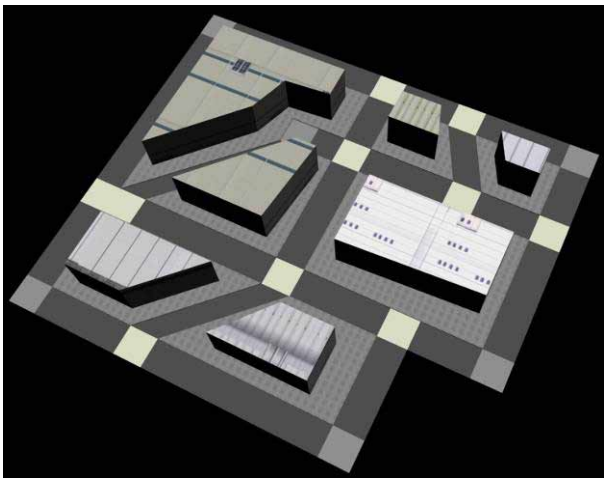

Figure 8 : training environment

No subject encountered problems during training and all of them quickly found the sound source.

The test

Three sources were randomly positioned in each of the three zones. The locations of the sound sources were the

same for every subject. Only the order in which they had to be sought changed. Subjects played the game three times in three consecutive sessions in order to observe acquisition effects. For each session, the sequence of the sound sources and their locations were the same. Figure 9-10-11 show the location of sound sources in each zone:

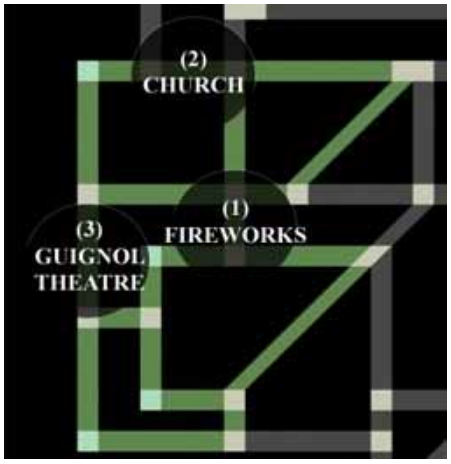

Figure 9 : Source location in Zone 1

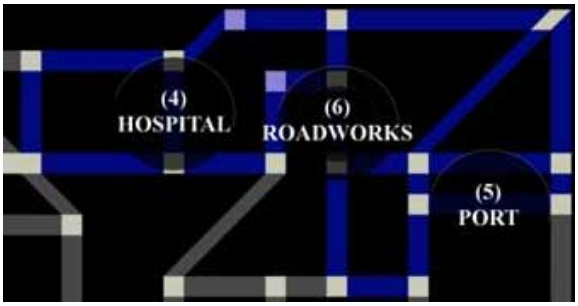

Figure 10 : Source location in Zone 2

Figure 11 : Source location in Zone 3

Evaluation

Subjects had to carry out the evaluation at the end of each of the three sessions, just after they had found the last sound source.

- First the map of the environment was displayed on the screen. The three different zones and their color were indicated. Using the mouse, the subject had to position the nine sources on the map, represented by white cubes.
- Secondly, the subject achieved the rating part of the NASA TLX.
- Thirdly, he/she achieved the weighting part of the NASA TLX.
- At last, he/she had to fill in the questionnaire.

4. RESULTS AND DISCUSSIONS

An ANOVA was conducted on the objective dependent variables (mean elapsed duration spent at crossroads, normalized covered distance and orientation frequency). For each group, there are 90 realizations of each variable (9 sources × 10 subjects)

However, for the subjective workload and the scales of the questionaire, data were collected only once per session, so there are only 10 realizations of the variables. It is not enough for applying the classical ANOVA model. In this case, the statistical analyses were conducted with non-parametric versions of the ANOVA. The Kruskall-Wallis test was used for inter-group comparisons (i.e. effects of condition) and the Friedman test for intra-group comparisons (i.e. effect of acquisition)

Next, the following convention will be respected:

- **Condition 1 or "BinCont":** binaural rendering for *contextualized* representation of location.
- **Condition 2 or "BinDecont"**: binaural rendering for *decontextualized* representation of location.
- **Condition 3 or "SteCont":** stereo rendering for *contextualized* representation of location.

- **Condition 4 or "SteDecont":** stereo rendering for *decontextualized* representation of location.

4.1. Mean elapsed duration at crossroads

4.1.1. Results

Figure 12 shows the elapsed duration in seconds for the four conditions.

The ANOVA reveals a significant effect of the condition on the elapsed duration, F(3,356) = 19.054, $p<0.001$.

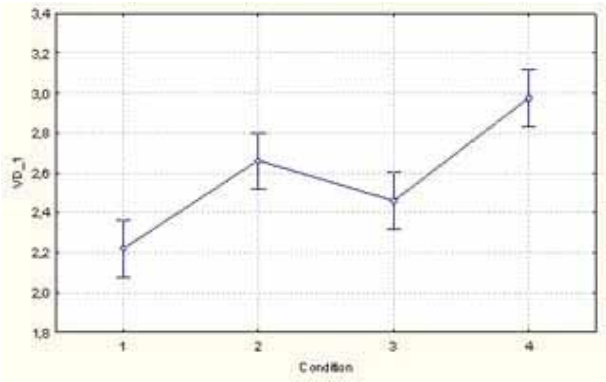

Figure 12 : Effect of condition on the mean elapsed duration at crossroads

Figure 12 shows that:

- For identical spatial sound rendering, the *contextualized* representation of location allows better performances than the *decontextualized* representation.

- For an identical representation of location, the binaural rendering allows a better performance than the stereo rendering.

Moreover, the ANOVA also reveals a significant effect of acquisition, F(2,712) = 279.95, $p < 0.001$. Figure 13 shows the elapsed duration in seconds for the three sessions.

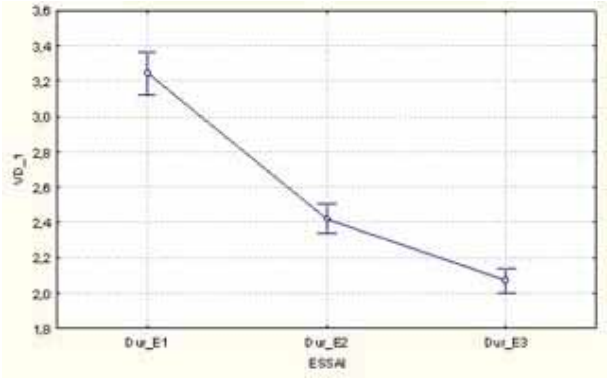

Figure 13 : Effect of acquisition on the mean elapsed duration at crossroads.

4.1.2. Discussion

When the representation of location is *decontextualized*, there is often an obstacle between the sound source and the listener. So he/she had to resolve an ambiguity (e.g. to choose the nearest direction), which adds extra time for decision, relatively to *contextualized* representation.

Moreover, the superiority of binaural rendering over stereo rendering is an evident result in appearance, but it is not. Unless the binaural rendering was expected to provide more precise reproduction of the acoustic field than stereo rendering, the difference could have been negligible. Actually, the HRTFs were not individualized and, in the context of active perception (i.e. when the listener moves his/her head), the stereo rendering already presents all the necessary cues for localization in the horizontal plane (Interaural Time and Intensity differences).

4.2. Normalized covered distance

4.2.1. Results

Concerning the covered distance, the ANOVA reveals a significant global effect of the condition, F(3,356) = 3.0817, $p < 0.001$.

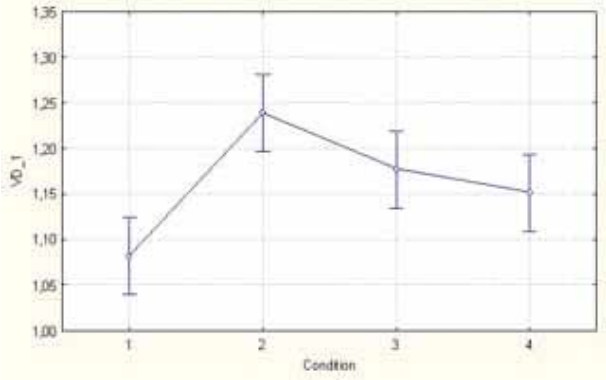

Figure 14 : Effect of condition on the normalized covered distance

It can be noticed that the mean normalized distance is close to the optimal distance (=1) whatever the condition is.

However, the ANOVA does not reveal any effect of acquisition, $F(2,72) = 1.9106$, $p = 0.14875$.

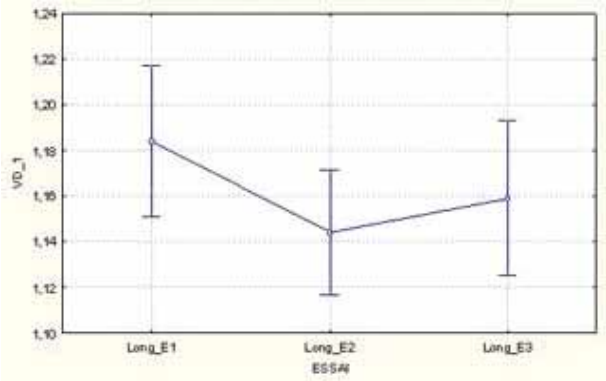

Figure 15 : Effect of acquisition on the normalized covered distance

4.2.2. Discussion

The first interpretation for this high level of performance is that the task was probably too easy. However for making it more difficult or increasing differences between conditions, it should have been necessary to increase the chances to be lost in the environment. The design of a network which satisfies this condition, could have led to a singular configuration that would not have been necessarily interesting for this study. Worth, maybe, the contextualized representation, would have lost its relevance, in becoming too trivial (i.e. in facilitated too much the task relatively to the *decontextualized* representation).

Maybe the most valuable hypothesis is that *decontextualized* representation did provide enough information for the navigation task. However, this is only valid for the particular context of the study: Navigation for entertainment (low constraint on time and distance) in a virtual environment (no constraint on energy consumption) and with a standard network (no "trap").

At last, when comparing acquisition for this factor and for the previous one, it seems that the subjects needed less time to decide which way to go, although decisions were not better. In fact, the game instructions told subjects to find the sources as quickly as possible. Then, decision time seems to have been more critical for subjects than the distance covered. This could be due to the absence of a real experience of moving. Without "physical" immersion, distance could be just an abstract notion (without consequences for the task). Moreover, because of the easiness the task, it is also likely that finding a shorter path was either impossible or simply a waste of time. The first hypothesis seems to be confirmed by the values of the mean normalized distance, and the second by the rate of the factor performance in the TLX evaluation. Actually, whatever the condition, the perceived performance has a very high rate (Cf. § 4.4).

4.3. Orientation frequency

4.3.1. Data filtering

This dependant variable measures the mean number of stops for a particular azimuth of the target. If θ_L is the absolute angle of the listener and θ_T the absolute angle of the target, the variable observed is $\theta = \theta_L - \theta_T$. That is, if $\theta > 0$, the target is on the right of listener, and respectively to the left.

When the subject wanted to rotate, he/she did sometimes a succession of brief keystrokes instead of letting the key pressed. Then the log presents a lot of brief arrests that must be filtered out, because they do not correspond to listening directions.

The distribution of this duration is present figure 16.

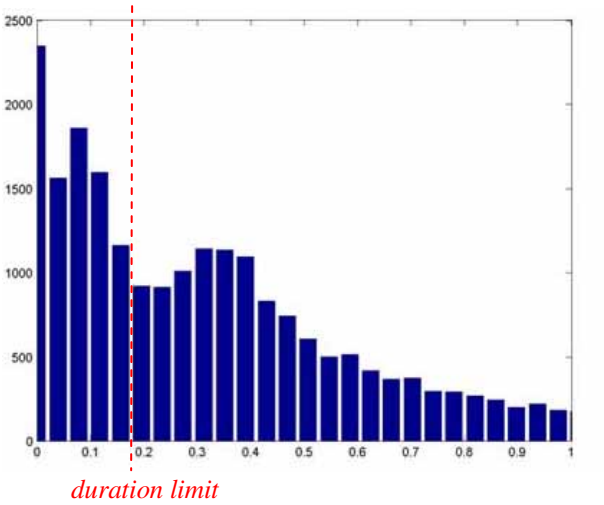

Figure 16 : distribution of the stop duration. The X-axis corresponds to duration in second and Y-axis to the frequency of this duration

According to the distribution, it seems that the influence of brief keystrokes behavior begins below approximately 175 ms (between 150ms and 200ms).

Moreover, the interesting durations for the perception subtask could be one of the two following:

- *The minimum elapsed duration for sound localization* – according to Hofman and Van Opstal [14], it is very brief, just a few milliseconds, so it is not a constraint here. Effectively, the frame rate of the game corresponds appreciatively to a 20 ms period (50 f/s).

- *The minimum elapsed duration for reacting to a sound and decide if it corresponds to the direction to be followed* - according to [15], the mean simple reaction time is about 160 ms

Then, according to the distribution and the mean reaction time to a sound, 160 ms seems to be the best compromise for the minimum duration limit.

4.3.2. Results

In order to analyze a fine listening behavior, the angular interval [0 360°] was sampled in sixteen sectors of 22.5° each, with sectors equally spaced and centered around 0° and +/- 90°. Figure 17 shows the center of each angular sector and the correspondent index in the histogram:

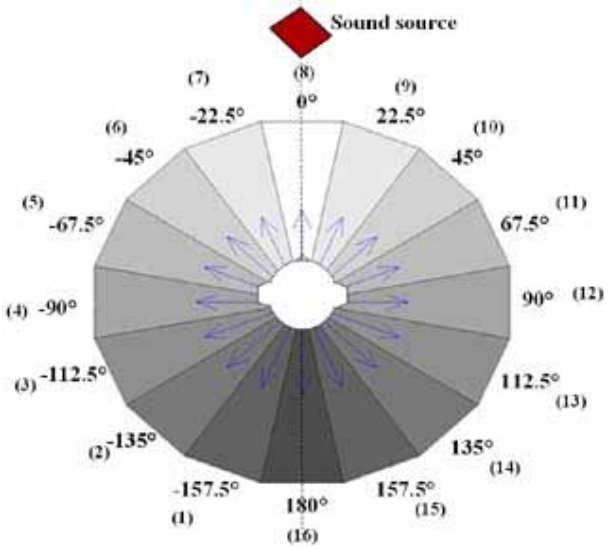

Figure 17 : sampling of the angular interval for orientation frequency distribution

Figure 18 shows the mean orientation frequency distribution for *contextualized* representation of location. The scale is the same for all six distributions and the doted circles represent the lines of equal frequency.

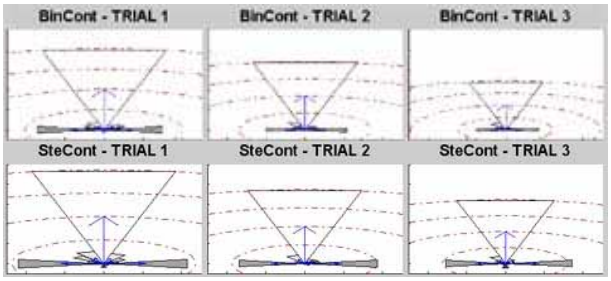

Figure 18 : mean orientation frequency for *contextualized* representation of location - sixteen sectors histogram

Likewise, figure 19 shows the mean orientation frequency distribution for *decontextualized* representation of location. Once again, the scale is the same for all six distributions and the doted circles represent the lines of equal frequency.

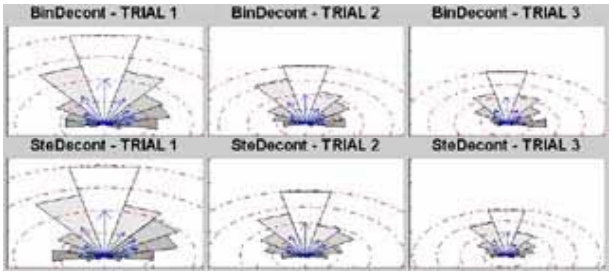

Figure 19 : mean orientation frequency for *decontextualized* representation of location - sixteen sectors histogram

The ANOVA confirms what can be observed on the previous scheme. First, as can be seen figure 20, there is a strong effect of the sector, $F(15,5340) = 975.00$, $p < 0.001$.

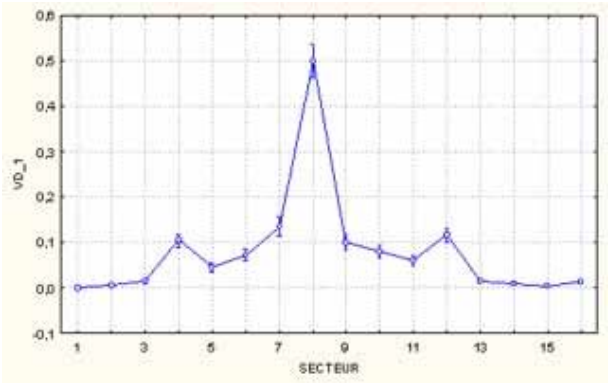

Figure 20 : effect of sector on mean orientation frequency - sixteen sectors histogram

As expected, there were two main azimuths used for localizing sound: frontal (0°) and lateral (+/- 90°). The frontal position is more frequent (around 0,5 stop per node) than lateral ones (around 0,1 stop per node).

For studying the cross-effect of sector and condition, the four conditions have been reduced into two: *contextualized* representation (condition 1) and *decontextualized* representation (condition 2). The ANOVA shows a significant effect of this interaction: $F(15,5370) = 184.10$, $p < 0.001$, and the post hoc Tuckey HSD tests are significant between every sector, except between sector 5 and 6.

The two shapes for the histogram shown in figure 21 are considered to be significantly different:

- *Contextualized (condition 1, thick line):* the global shape is the same than previously, with the same two main azimuths.

- *Decontextualized (condition 2, dotted line):* the shape is more smoothed than the previous one. There is only one main azimuth, 0°.

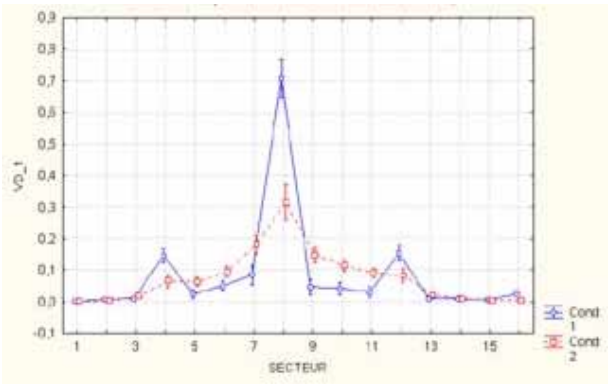

Figure 21 : cross-effect of sector and grouped condition on mean orientation frequency - sixteen sectors histogram

4.3.3. Discussion

Several studies have shown that allowing a listener to move his/her head can improve localization ability and decrease the number of reversals ([17], [18] and [19] quoted in [16]). In order to highlight a typical behaviour, D; Begault, took as an example, the informal observation of the household cat exposed to an unfamiliar sound source: "first, it spreads its ears outward to receive the widest possible arc of sound, moving the head while maximizing interaural differences; and then, once the head is turned toward the source, the ears come up straight to fine-tune frontal" [16]. Such behaviour has been partially observed in figure 20, showing the global effect of angular sector on mean orientation frequency.

However, the study of the cross-effect of angular sector and condition (figure 21) reveals something more: there are two different behaviours, depending, on the a priori knowledge on source location:

- For *contextualized* conditions, the listeners knew, when looking around them, the few possible azimuths for the sound source; the sound was coming from one of the roads around them. In this case they generally stopped rotating when the source was in

frontal and lateral position. It is likely that, with visual knowledge, static localization was enough. Actually, the movements of listeners' head seem automatic and effective (only to useful orientation), as if it was only a simple confirmation of the sound source azimuth.

- On the contrary, for the *decontextualized* condition, the distribution looks like a Gaussian one with the mean at 0°. Even if subjects tried to turn their head toward the source, the lateral position (+/- 90°) is no more distinctive. Such a behaviour looks like fumbling for the frontal position of the sound source, for more accurate localization. Then, other azimuths seem to be just localization errors or steps before reaching the right azimuth. Unfortunately, because any information on the temporality of angular sectors (neither instant nor duration) was not available by now, it is impossible to confirm such an hypothesis.

4.4. Subjective workload

4.4.1. Results

Only for the third session, the Kruskall-Wallis test reveals a significant effect of the condition on overall workload, H (3, N= 40) =12.579, $p < 0.01$. Figure 22 shows that:

- For identical sound rendering, the *contextualized* representation (condition 1 and condition 3) requires less workload than the *decontextualized* representation (condition 2 and condition 4).

- For identical representation of location, the binaural rendering (condition 1 and condition 2) requires less workload than the stereo rendering (condition 3 and condition 4).

Moreover, the dispersion of the value for the "BinCont" (condition 1) is less important, than for the other conditions.

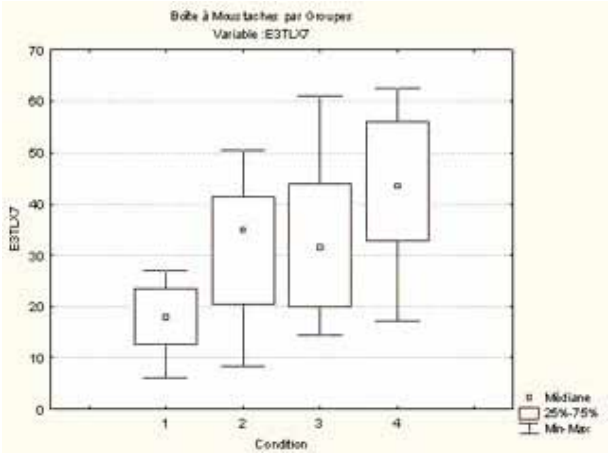

Figure 22 : effect of condition on overall workload, for the third session

Finally, for the condition "SteDecont", the Friedman test reveals a significant effect of the session on the global workload, Chi² (N = 10, dl = 2) = 7.8, $p < 0,05$.

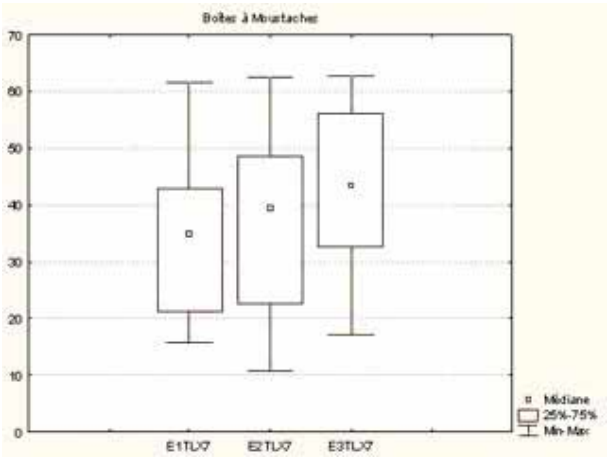

Figure 23 : effect of the session on global worload, for "SteDecont" condition

Surprisingly, the subjective workload increases with the sessions.

At last, in order to echo back the discussion in section 4.2.2, the next figure shows the perceived performance for the first session. It reveals a relatively high rate (~70/100) for all conditions:

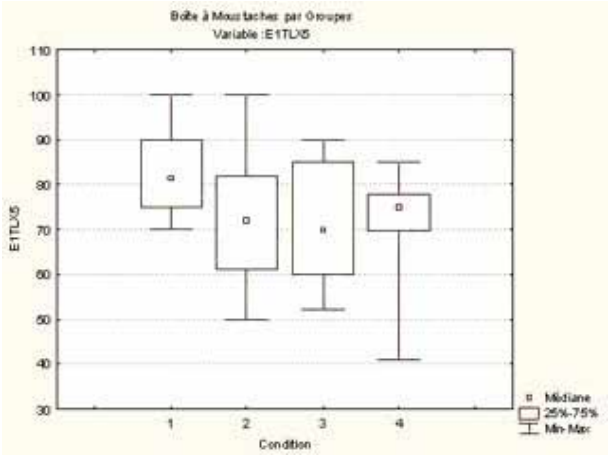

Figure 24 : effect of the condition on the perceived performance, for the first session

4.4.2. Discussion

There are only little significant effects observed on all six factors of the TLX, or in the global workload score. It seems there is no strong agreement between subjects. In general, the dispersion is important, except for the "BinCont" condition in the third session. Maybe, this rendering technique needs a certain exposition to be appreciated.

When comparing figure 22 and figure 12, it seems that the condition has the same effect on the overall workload, than on the mean duration elapsed at crossroads. This likely illustrates the fact that facilitating decision (i.e. reduce the decision time) reduces the subjective workload in the same way.

The only significant effect of the session concerns stereo rendering. Surprisingly, while the task is supposed to be easier, the overall workload increases for "SteDecont" condition. This condition is supposed to be the worst of all: the spatial sound rendering and the representation of location are the poorest and the more common. Then the subject for this condition could have felt more quickly bored about repeating the same task three times.

4.5. Impression

4.5.1. Results

As for the TLX evaluation, there is no strong agreement between subjects. The effect of condition is significant for just few factors:

- ***"Easiness for localizing sound"***: the Kruskall-Wallis test was found to be significant for session 2, $H(3, N= 40) =10.737$, $p < 0.05$, and session 3, $H(3, N= 40) = 11.300$, $p < 0.05$. In both case the "BinCont" (condition 1) obtains the best score.

- ***"Easiness of the task":*** the Kruskall-Wallis test was found to be significant for session 1, $H (3, N= 40) = 9.041$, $p < 0.01$. Once again, the "BinCont" (condition 1) obtains the best score.

Finally, the effect of the session influences more criteria. The Friedman test reveals a significant effect on:

- ***"Sound quality"*** for "BinCont" condition, $Chi^2(N = 10, dl = 2) = 8.000$, $p < 0.05$

- ***"Feeling of immersion"*** for "BinCont" condition, $Chi^2(N = 10, dl = 2) = 6.421$, $p < 0.05$.

- **"Coherence of sound with the environment"** for "BinCont", $Chi^2 (N = 10, dl = 2) = 6.000$, $p < 0.05$, and "SteDecont", $Chi^2 (N = 10, dl = 2) = 6.500$, $p < 0.05$

- ***"Utility of visual information"*** for "SteCont" condition, $Chi^2 (N = 10, dl = 2) = 8.375$, $p < 0.05$

4.5.2. Discussion

Each time there is a significant effect of acquisition, it is for a more positive impression: sound quality is better, the feeling of immersion is greater, the coherence of sound with the environment increases and the visual information is less useful. Moreover, it seems that the "BinCont" condition is more often involved than the others in significant effects of either the condition or the session factors. This reinforces the hypothesis that this condition is the more interesting one.

5. CONCLUSION

This paper presents a global approach for investigating not only usability but also perception and subjective-impression of spatialized non-speech sounds. However, the significant results concern rather usability and perception.

Regarding usability, this study has first shown that the binaural rendering allowed shorter decision time than stereo rendering. Moreover, the evaluation of the NASA-TLX has underlined a correlation between decision time and workload. HRTFs was not individualized and dynamic localization decreased the binaural rendering superiority relatively to the stereo. However, the binaural was still the most usable rendering mode.

Concerning perception, the listening behaviour has not revealed significant difference between stereo and binaural rendering. The main difference was observed in the case of a comparison of the location representation mode. For the *contextualized* one, the distribution of the *orientation frequency* has shown two main azimuths are used for localization, corresponding respectively to frontal and lateral position of sound source. This localization behaviour is subject to the influence of multimodal interaction defines by the *contextualized* representation. There is no such effect for the *decontextualized* representation.

This study has allowed to particularly highlighting the effect of multimodal perception on sound source localization. Another result has appeared about the left ear-dominance. Works are underway to deeply analyse these.

6. REFERENCES

[1] Larsson P., Vätfjäll D., Kleiner M., (2001), *Ecological Acoustics and the multimodal perception of rooms: Real and Unreal experiences of auditory-visual virtual environments.* Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 20-August 1, 2001

[2] Arons B., (1992), *A Review of the Cocktail Party Effect.* Journal of America Voice I/O Society.

[3] Walker, B. N., Lindsay, J., *Auditory navigation Performance is Affected by Waypoint Capture Radius,* Proceedings of the 2004 International Conference on Auditory Display, Sydney, Australia, July 6-9, 2004

[4] Pellegreni R. S., (2001), *Quality Assessment of Auditory Virtual Environment,* Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 20-August 1, 2001

[5] Gleiss N. (1992), *Usability – Concept and Evaluation*. Tele 2/1992

[6] ISO 9241-11 (1998), *Ergonomic requirements for office work with visual display terminal*s (VDTs) -- Part 11: Guidance on usability, International Standards Organization.

[7] Tran T.V., Letowski T., Abouchacra K.S., (2000), *Evaluation of Acoustic Beacon Characteristics for Navigation Tasks*, Ergonomics, 43 (6), 807-827.

[8] Walker, B. N., Lindsay, J. (2003), *Effect of Beacon Sounds on Navigation Performance in a Virtual Reality Environment*. Proceedings of the 2003 International Conference on Auditory Display, Boston, MA (6-9 July) pp 204-207.

[9] Walker B. N., Kramer G., (2004). Auditory display. In J. Neuhoff (Ed.), *Ecological psychoacoustics*. New York: Academic Press. Ecological Psychoacoustics and Auditory Displays: Hearing, Grouping and Meaning Making.

[10] Bridgett R., *Hollywood Sound: Part Two*, www.gamasutra.com, September 30, 2005

[11] Lokki T., Gröhn M., *Navigating with Auditory Cues in a Virtual Environment,* Multimedia, IEEE, Volume 12, Issue 2 April-June 2005, Page(s): 80-86.

[12] Hu H., Lee D-L, *Semantic location Modeling for Location Navigation in mobile environment*, Mobile Data Management 2004, 52-61

[13] NASA Human Performance Research Group (1987). Task Load Index (NASATLX) *v1.0 computerised version*. NASA Ames Research Centre.

[14] Hofman P.M., Van Opstal A.J. (1998). *Spectro-temporal factors in two-dimensional human sound localization.* Journal of the Acoustical Society of America, 103 (5), 2634-2648.

[15] Brebner J. T., A. T. Welford. 1980. *Introduction: an historical background sketch*. In A. T. Welford (Ed.), *Reaction Times.* Academic Press, New York, pp. 1-23.

[16] Begault D. R.,*.3-D Sound for Virtual Reality and Multimedia* Cambridge, MA: Academic Press Professional, 2004.

[17] Wallach, H. (1940). *The role of head movements and vestibular and visual cues in sound localization*. Journal of Experimental Psychology, 27, 339–368.

[18] Thurlow, W. R., and Runge, P. S. (1967). *Effects of induced head movements on localization of direct sound.* Journal of the Acoustical Society of America, 42, 480–487.

[19] Thurlow, W. R., Mangels, J. W., and Runge, P. S. (1967). *Head movements during sound localization*. Journal of the Acoustical Society of America, 42, 489–493.