Evaluation of the Sonification Protocol of an Artificial Vision System for the Visually Impaired

Pablo Revuelta Sanz
Electronic Technology
Carlos III University of Madrid
Leganés, Spain
Email: prevuelt {at} ing.uc3m.es

Belén Ruiz Mezcua
Computer Science
Carlos III University of Madrid
Leganés, Spain

José M. Sánchez Pena
Electronic Technology
Carlos III University of Madrid
Leganés, Spain

Bruce N. Walker
Sonification Lab
GeorgiaTech
Atlanta, U.S.A.

Abstract— **In this study we present the results of evaluating the sonification protocol of a new assistive product aiming to help the visually impaired in perceiving their surroundings through sounds organized in different cognitive profiles. The evaluation was carried out with 17 sighted and 11 visually impaired participants. The experiment was designed over both virtual and real environments and divided into 4 virtual reality based tests and one real life test. Finally, four participants became experts by means of longer and deeper trainings and then participated in a focus group at the end of the process. Both quantitative and qualitative results showed that the proposed system is able to effectively represent the spatial configuration of objects through sounds. However, important limitations have been found in the sample used (some important demographic characteristics are intercorrelated, impeding segregated analysis), the usability of the most complex profile, and even the special difficulties faced by totally blind participants relative to the sighted and low vision ones.**

Keywords-component; formatting; style; styling; insert (key words)

I. Introduction

We call “artificial vision” the processing and transmission of visual information into non-visual formats. This processing is very useful for people who temporally or permanently cannot receive the visual information from their surroundings. Likewise, due to the large bandwidth of the auditory system, the use of sounds is one of the most used ways to represent the visual world. Moreover, there is strong evidence of the benefits of auditory displays to transmit visual information to the blind [1].

“Sonification” is the way we translate data into sounds. We can find many types of sonification, such as text-to-speech programs (converting text into audible speech), color readers (color into synthetic voice), Geiger counters (radioactivity into clicks), acoustic radars or MIDI synthesizers, etc. It has also been widely used in the assistive technology field to substitute visual information and thus specially oriented to the visually impaired.

Technology has been applied to mobility since the 60’s and 70’s [2;3]. Focusing specially in the image processing based Assistive Products (AP), we can find, among others, the Sonic Pathfinder [4], Tyflos [5], Echolocation [6], vOICe [7], FIU Project [8], 3-D Space Perceptor [9], NAVI [10], SVETA [8;11;12], AudioMan [13], CASBliP [14], EAV [15;16], 3-D Support System [17], Brigham Project [18], the Optophone [19] or the Cross-Modal ETA [20]. Some of the latests advances in this filed can be found in [21-24]. These systems use different strategies to provide the relevant information to the users, mainly tactile and auditive. For a review of them, see [25;26].

Among these proposals, we find an important problem: the sonification uses non-redundant transformations of spatial information into sounds, which make is harder to be understood.

II. Materials and Methods

In this paper, we discuss the evaluation of a redundant sonification protocol described in [27], for its utility in an artificial vision system for the blind. The sonification used is a variation of the point mapping, as described in [28], height is codified as frequency, horizontality as binaural loudness. The volume, again, is related to the brightness. Another example can be found in [29].

Figure 1 shows the block diagram of the system in which these sonification rules will run.

The cameras used were a couple of low-cost USB webcams [30] with a resolution of 320×240 pixels at 30 fps. Take into account that the visual cone is 90º width (in vertical and horizontal axis). The programming language for the image processing was the OpenCV library running with an ANSI C program. The sonification was implemented with the MIDI (GM2) protocol.

Although the complete system presents 7 different profiles of sonification, we tested only the 4 more complex ones (those useful for artificial vision). The complete set of sounds used in this evaluation in relation to the three dimensions of space (from the user’s point of view) is summarized in Table I.

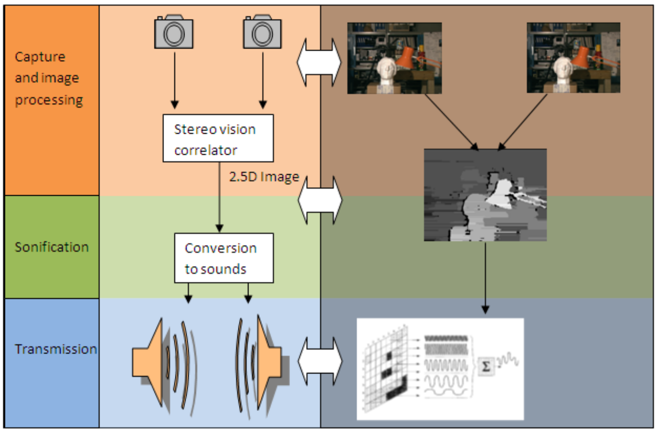

Figure 1. Block diagram of the system under test.

TABLE I. SONIFICATION DETAILS FOR EACH PROFILE COMPLEXITY LEVEL

Profile	Depth	Vertical	Horizontal
3	Loudness+ Low Pass Filter (LPF)	2 pitches in 1 note and 2 different octaves	8 columns, Stereo + vibrato
4	Loudness+ LPF	4 pitches in 1 note and 2 different octaves	8 columns, Stereo + vibrato
5	Loudness+ LPF	8 pitches in 2 notes and 4 different octaves	8 columns, Stereo + vibrato
6	Loudness+ LPF	16 pitches in 4 notes and 4 different octaves	8 columns, Stereo + vibrato

Figure Labels: Use 8 point Tim

As it can be seen, the different profiles implement increasing complexity in the vertical representation of the scene, using the double of pitches of the previous level. The sounds are organized as follows:

- The brightness (the depth) is correlated with the loudness in a range of 0-127. A Low Pass Filter increases the sharpness of the sound proportionally with the loudness.
- The lateralization is performed by differences in the loudness and the time of each sound, as it is described since the early psychoacoustic studies [31]. To avoid ambiguities, a vibrato is applied to lateral points: the closer to the side (i.e., more lateral) a point, the deeper the vibrato. With a 90º visual cone, the discretization in 8 columns gives an accuracy of 5.6º, below the 6º azimutal error found by some researchers, as [32].
- The vertical axis is represented by means of harmonic musical notes (which perform the CMaj7m chord when all the height levels are excited). However, some simpler profiles have also been proposed, this last one being the most complex. In this maximum level, 16 notes are used for height (the CMaj7m chord in 4 octaves). In a 90^0 visual cone, each row represents 5.625^0. A harmonic chord allows the user to perceive music, instead of unpleasant noise. There are around 20dB between the response of the hearing system for the lower tone (C2, 65Hz) and the higher one (Bb5, 932Hz) according to the sensitivity curves. Given that the MIDI protocol is designed to play these two frequencies at the same perceived loudness, we made no additional compensations.
- The subjacent idea is that redundant codes for each axis may enhance the perception and understanding of the scene in front of the user.
- In validating the sonification protocol for use in an artificial vision system for the visually impaired, we had the following hypothesis:
- H1: The sonification protocol helps in representing basic structures in the space.
- H2: The higher the profile complexity level, the more detailed the perceived representation.
- H3: There are factors related to previous user experience modulating the usability and performance of the system.
    - H3.1 Previous user experience with computers helps in the understanding of a new sonification protocol.
    - H3.2: Educational level is positively correlated with the ease of learning a new sonification protocol.
    - H3.3: Age is inversely correlated with the utility of the protocol as artificial vision.
    - H3.4: Visually impaired (VI) people encounter more problems in understanding new sonifications.
    - H3.5: Longer training leads to higher performances.
    - There are, thus, five a priori independent factors to be analyzed: use of computer, educational level, age, visual impairment and training length.

A. Methodology

The sonification strategy was tested in three different ways: (1) with a virtual reality environment (VRE), in order to evaluate the sonification itself (with no interference of the image processing system); (2) in a real environment (RE); and (3) with four experts (with longer training) in both VRE and RE tests.

Participants. In the VRE test, 17 undergraduate and graduate students from a technical university in the USA, plus 11 clients and employees from the Center for the Visually Impaired (CVI) of Atlanta (Georgia) participated. The sample included 11 males and 17 females, with a mean age of 33.46 years (range 18-62). Among them, 13 were sighted, 10 had low vision, and 5 were completely blind. All reported normal or corrected-to-normal hearing. All the blind and low vision subjects had previously participated in other sonification experiments in the same laboratory. The same group of people participated in the RE experiments. The experts were four students of 22 years of age (3 female, 1 male). All of them reported normal vision and hearing.

Apparatus VRE. The evaluation of the sonification protocol was first done using a virtual reality environment, directly sonified and transmitted to the participants through earphones.

The virtual environment was developed with the Unity3D engine (http://unity3d.com/), connected through the IServer program to the InterSense InertiaCube2 head tracker (http://www.intersense-.com/pages/18/11/). The Unity3D rendered images are sent to the sonification program, written in C and connected to the V-Stack (http://www.steinberg.net/en/support/unsupported_products/vstack.html) and Edirol HQ Hyper Canvas Synth (http://www.roland.com/products/en/HQ-GM2/). This synthesizer allows processing General MIDI 2 signals (http://www.midi.org/techspecs/gm.php) produced by the sonification program, whose correlative sounds are transmitted to the user by means of a pair of earphones.

Procedure VRE. Participants were briefly trained in one single level, which was assigned to them randomly (but according to their visual status, divided as sighted, low vision, and blind, to cover all the cases as uniformly as possible). This training was done over static images (available over the online test of the protocol at http://163.117.201.122/validacion_ATAD_cerrada/encuesta2.html, please refer to this link for sonification examples) for around 5 minutes. After this step, they passed through 7 training scenes (Figure 2) that were designed in Unity3D.

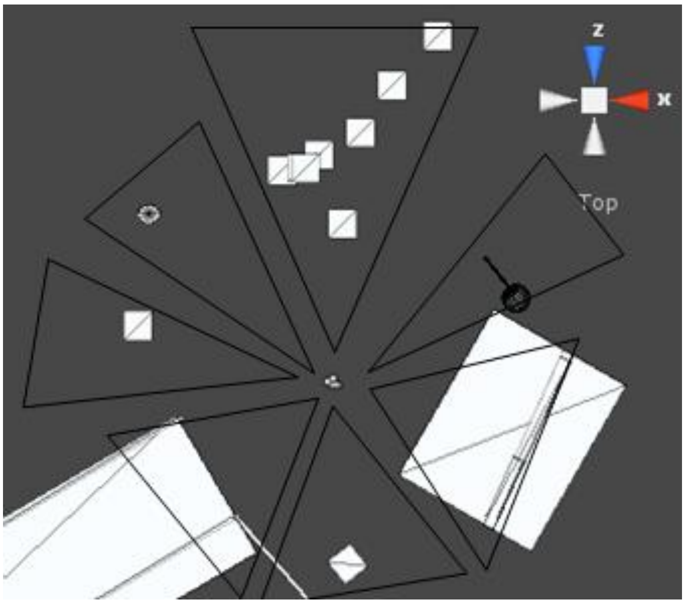

Figure 2. Unity3D training environment with the 7 scenes surrounding the user, static, in the center.

The VRE training consisted of 7 scenes with different objects, some of them static and some others performing periodic movements in different axis. The participant had the opportunity of facing (clockwise order from the upper cone in Figure 2) a stack of boxes, a pendulum, an open door, a box moving horizontally near a wall, a corridor, a box moving on a path like an infinity symbol (closer to and farther from the participant), and a column. The experimenter verbally described each scene and when the subjects were sighted, they were allowed to see the screen being sonified.

They decided whenever they wanted to progress to the next scene saying “next scene”.

The pointing direction of the VR avatar was controlled by the head tracker, which only detects rotational movements. Thus, the user was able to freely move the head to look at different parts of the scene whenever they wanted to. In every case, they couldn’t see the screen after the training and the only feedback of the virtual reality was provided through the earphones.

This step had no time limit. Although this, all of the participants completed it in less than 20 minutes.

Four testing scenes were also designed in Unity3D, and they are shown in Figure 3.

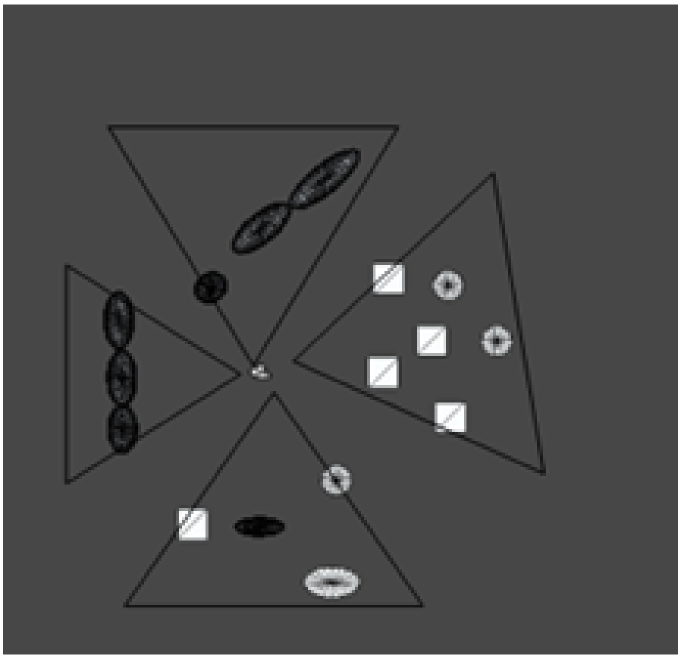

Figure 3. Four testing scenes around the participant avatar, located in the center. Top view.

Starting from the left, the scenes were composed of:

- Scene 1: three balls at three different heights, same distance (located in a 3x3 grid from the user point of view).
- Scene 2: three balls at three different heights, three different distance (located in a 3x3 grid from the user point of view) and the farther one repetitively moving from the bottom height to the middle height and back to the bottom.

- Scene 3: 6 boxes and balls at three heights and distances (located in a 3x3 grid from the user point of view).
- Scene 4: four objects in the same horizontal line, the left one slightly changing its position in sudden movements every 10 seconds.

For the first three scenes, participants were asked to indicate where the objects were located in the 3x3 grid (up, middle, down and left, center and right). They did not know the actual number of objects that were present in each test. The aim of this test was to identify differences in the horizontal position and height in the first scene, and these two parameters combined with the distance in the second one. The third one presented a more complex composition of positions, trying to find masking effects over the further objects. In the fourth scene, the moving object was used to study the ease of perceiving relatively rapid changes in the scene. Participants received the instruction “organize the objects in terms of distance and indentify which one is moving every ten seconds”.

The participants had to take their time to decide where were the objects, reporting found positions in the grid, for example “first row, middle column”. Whenever they though no more objects remained to be found, they had to say “next combination”.

All the VRE was scaled in the same way. This is done through the same correlation between distance and brightness (through the “fog” function of the Unity3D). With this stability, it is irrelevant whether the avatar of the user and the objects are smaller or bigger, since their relation remains stable along the experiment. Thus, no metric was given thinking it wouldn’t provide extra information.

After these experiments, they were asked to complete a survey about the subjective perception of the training and the tests parts, as well as some demographic questions.

Apparatus RE. The setup of the RE test consisted of a table (1×1m2 with lines drawn on it dividing it into a 3×3 grid) supported objects in different spatial combinations, as shown in figure 4. The objects included a plastic cup, a spray bottle, a camera cover, and a balloon.

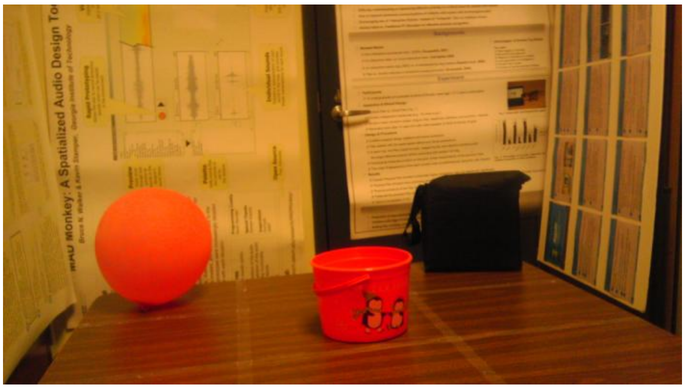

Figure 4. Example of configuration of objects on the table.

A computer running the stereovision algorithm used to build the depth map of the scene and the previously used sonification program. A pair of webcams [30], with 90^0 of field of view, attached to a helmet, captured the scene which was transmitted through two USB cables to be processed. The produced sound was transmitted to the user through a pair of earphones.

Procedure RE. The participant, blindfolded if needed (sighted or low vision cases), was sitting in front of the table, at 20 cm from the edge, wearing the real system, and had to report where the objects were whenever the experimenter said “go ahead”. Participants did not know the actual number of objects on the table. Whenever they thought they had found all of them, they could say “change” to ask the experimenter to move on to the next configuration. Nine combinations of two or three objects were used in total. Before starting the test, a short training was done for 3 minutes: the experimenter explained there were some simple combinations of objects in different parts of the table, and participants were asked to report the row and column of the objects (e.g., “there is an object in the third row, left column”). No time limit was established for this training or test.

Apparatus Experts. The test with the experts used the same VRE and RE hardware and software.

Procedure Experts. The four students were trained between 5 and 6 hours (depending of the time they took for some exercises) in mobility and artificial vision tasks. The training included twice the complete VRE test, plus walking in the testing room, eyes opened, to learn to correlate the sound with the real objects, and a mobility test in the same room (with a different configuration of obstacles than that of the eyes-open step). The artificial vision test consisted of guessing the pose of a person in front of them, at 1 m distance, kneeling, sitting or standing up, in 9 cases. They were explained the three poses proposed, with one example of each one, and then they had to guess the pose. Whenever they made a decision, they were asked to look up at the ceiling while the experimenter changed his pose. After 5 seconds, participants had to explore and report the new pose.

After this, they were asked to complete a survey and participated in a focus group qualitatively discussing the experience, pros and cons of the real system.

III. Results

A. *Preprocessing of data and dependent variables*

Initial analysis of the data gathered from the demographic questions of the survey raised the evidence of non-independency of some of the factors exposed in the previous section. The factors are coded as follows:

- AGE: the age in years the day of the test.
- VI: three ordered values: sighted (1), low vision (2) and blind (3).

- EDU: four values: elementary school (1), high school (2), some college (3) and college degree or higher (4).
- COMP: five ordered values about the use of computers: never (1), rarely (2), once a week (3), once a day (4) and many times a day (5).

Table II shows the one way ANOVA mean comparison in terms of visual impairment and the other demographic descriptors.

TABLE II. MEANS COMPARISON OF AGE, GENDER, EDU AND COMP AGAINST VI

Variable	Sighted	Low Vision	Blind
EDU	3.46	2.8	2.8
AGE	21.92	38.3	53.2
GENDER	1.31	1.6	1.2
COMP	4.92	4.5	3.75

More in detail, we found the following Pearson correlation matrix, shown in Table III.

These results are consequences of the specific characteristics of participants from both the university and the CVI pools. The first group is composed mostly by sighted individuals, aged between 18 and 26 years old, using the computer many times a day and with some college as minimum educational level.

The CVI participants were a mean of 51.9 years old (range 36 - 64 years), with an average educational level of 2.7 and a computer use of 3.82. The university participants were a mean of 21.5 years old (range 18 - 26), educational level of 3.35 and computer use of 4.94.

Use of computer, age, and visual impairment are highly correlated and, thus, the analysis will be done over the use of computer, since it is the most descriptive variable (the age is quite variable and the visual impairment is so narrow with only 3 different values).

TABLE III. PEARSON CORRELATION INDEX AND P-VALUE

Pearson correlation	**AGE**	**VI**	**EDU**	**COMP**
AGE	1	0.752** (*p*<0.001)	-0.458* (*p*=0.014)	-0.708** (*e*<0.001)
VI		1	-0.474* (*p*=0.011)	0.637** (*p*<0.001)
EDU			1	-0.527* (*p*=0.004)
COMP				1

B. *VRE Scene 1*

The participants reported the position of the objects, correctly localizing an average of 1.71 of them (SD = 0.854) over 3 and a false positives average of 1.39 (SD = 1.197). No significant correlations were found when comparing these results with the educational level, the use of computer, the age or the profile complexity level. However, significant differences were found when comparing the number of false positives with the visual impairment (one way ANOVA: F(2,25)=3.728, p=0.028). Figure 5 shows this result.

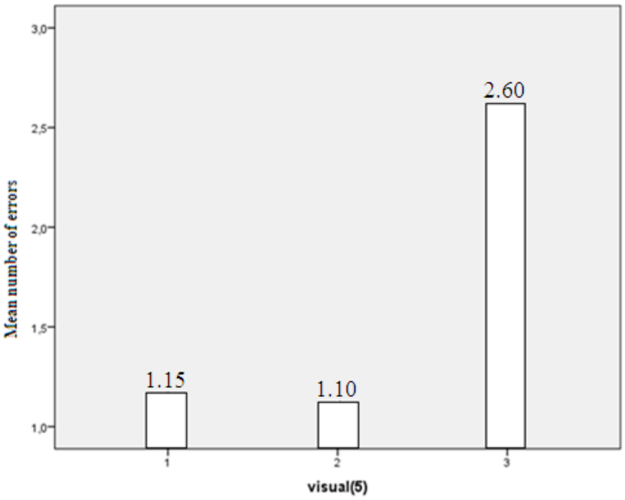

Figure 5. False positives against visual impairment.

Given that for each grid the participants had to say whether there was an object or not, the chance level, in this 3×3 grid, is 9/2 = 4.5. We can appreciate that even the blind group achieved detection rates below this level.

The difference of means between errors and correct detections and educational level was not significant (F(2,25)=0.648, p=0.532 for the false positives, F(2,25)=1.311, p=.287 for the correct detections).

Notice that every time a participant localized an object in the wrong place (for example, displaced one cell in some direction), two errors were added: one due to the false positive in the empty cell (in which the object was perceived) and a false negative in the original cell (in which the object was actually placed). Thus, we measure the worst case.

C. *VRE Scene 2*

In the second scene, the number of correctly detected objects presented a mean of 1.29 (SD = 0.763) and the same mean for the false positives (1.29; SD = 1.197). No significant differences were found comparing these two variables with the different factors already discussed. However, the number of false positives in this test and that of the previous one presented a significant Pearson correlation (r=0.497, p=.007).

Although not statistically significant, a relation can be between the number of errors and the profile complexity level used, as shown in Figure 6.

The correct detections did not follow the same pattern.

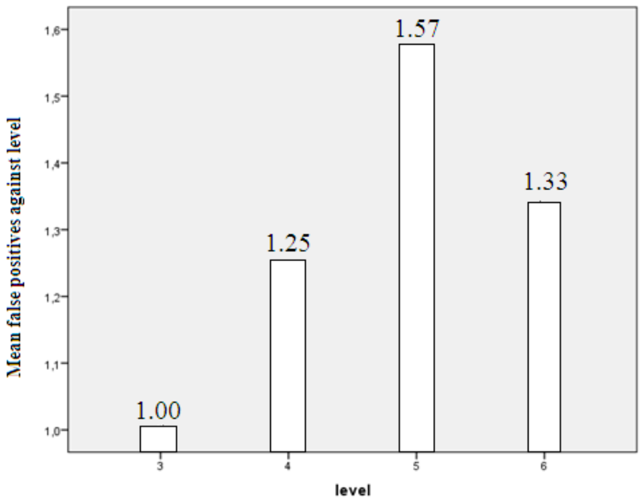

Figure 6. Mean of the number of correctly detected objects in terms of the profile complexity level.

D. VRE Scene 3

The third test produced different results. On one hand, the average number of errors was 0.5 (SD = 0.745) and the mean of the correctly detected objects was 1.96 (SD = 1.17). Given that there were 6 objects, only 32.6% of the objects were detected.

Marginally significant Pearson correlations (r=0.328, p=0.088) were found when correlating the correct detections and the profile complexity level. The result of the one way ANOVA analysis was not significant (F(3,24)=1.906, p=0.156) even though a tendency can be appreciated in Figure 7.

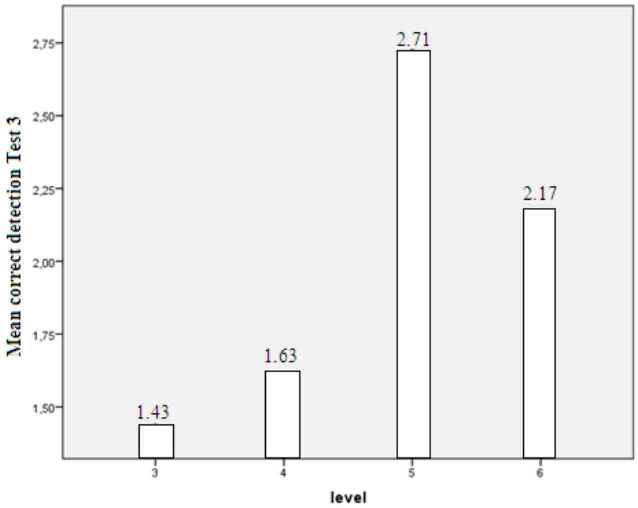

Figure 7. Correct detections in test 3 against the profile complexity level.

E. VRE Scene 4

The average number of correctly detected objects in the fourth scene was 1.79 (SD = 1.134) and that of false positives 0.79 (SD = 1.101). Only two participants reported to have found more than 3 objects. Once again, no significant results were found when calculating the Pearson correlations between pairs of variables and factors.

Other relevant results found were the mean of the number of detected objects in terms of the profile complexity level can be seen in Figure 8.

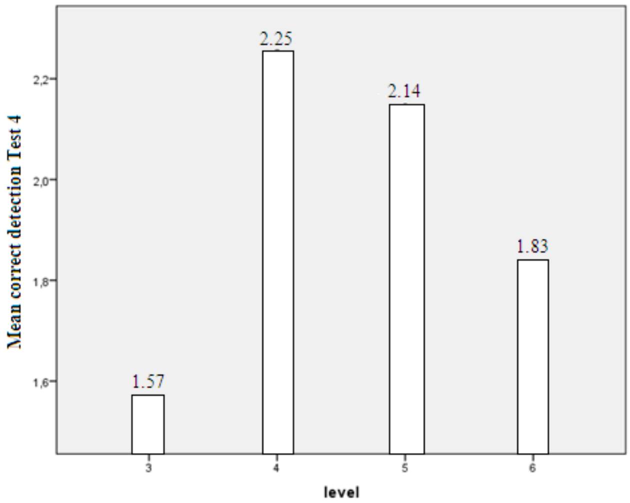

Figure 8. Mean of the number of correctly detected objects in terms of the profile complexity level.

Not reaching conventional levels of statistical significance, but nevertheless presenting a clear pattern, the relation between the same variable and the educational level can be appreciated in Figure 9.

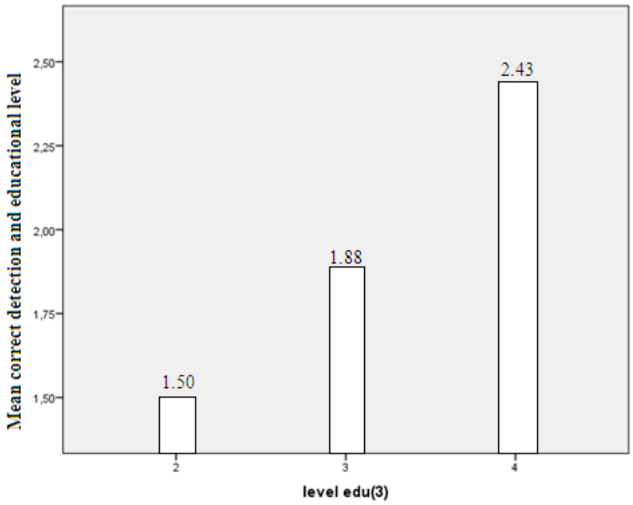

Figure 9. Mean of the number of correctly detected objects against the educational level.

In this test participants were asked to identify the moving object in the horizontal plane. The mean of correct identification index (ranged 0 –no detection- and 1 –correct detection) of the moving object was 0.79 (0.69 in the sighted, 0.90 in the low vision and 0.80 in the blind group). The average number of objects correctly ordered in terms of distance was 1.79 (1.62 for the sighted, 2.0 for the low vision and 1.8 for the blind groups).

F. Real Environment

The real environment experiment consisted, as mentioned, of 9 sequential trials, locating two or three objects in a 3x3 grid on a real table at 20 cm from the closer edge of the table after a short training to get the reference of the pointing directions of the cameras. Twenty-seven people participated of this experiment (data from one participant were deleted, because the participant initially reported himself as totally blind, but later admitted he was able to partially see some of the objects). The participants had to locate in the 3 rows and 3 columns the objects. Each time they located an object away from its real position, two errors were marked (one because of the false positive—locate an object were it is not—and another one because of the false negative—no localization of an object were it actually was), as done in the first VRE scene.

The average number of errors was 33.69% (ranged between 18.05 and 51.39; SD = 7.58%, chance level = 50%).

Marginally significant Pearson correlation was found between the time required to complete the test and the visual impairment (r=0.345, p=0.078). This relation is shown in Figure 10.

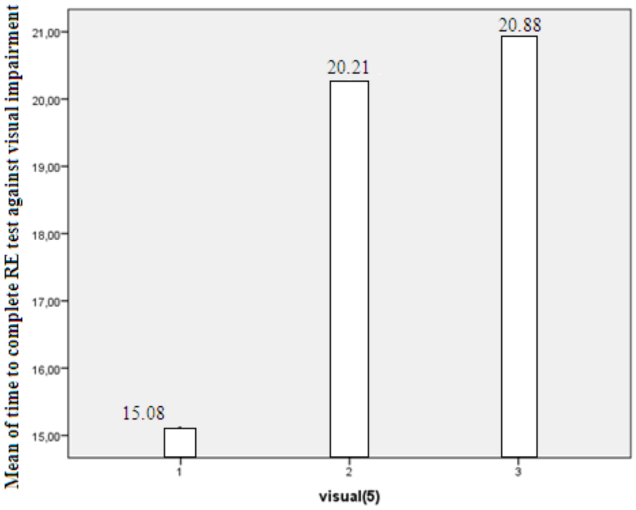

Figure 10. Time to complete the RE test against the visual impairment.

Attending at each cell, the share of the total errors is not equally distributed all around the table. Figure 11 shows the percentage of errors from each cell.

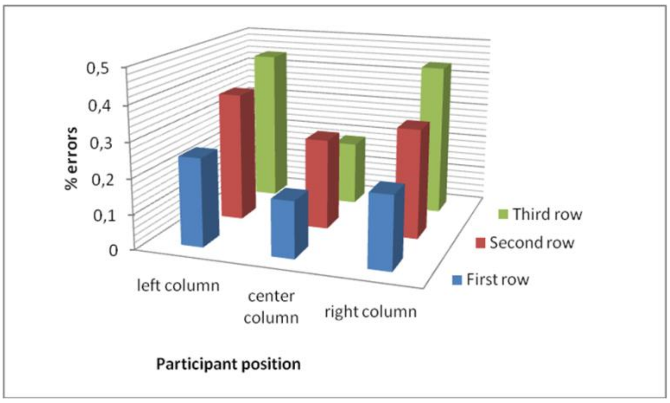

Figure 11. Errors in the RE test, spatially distributed in the table grid.

These errors cannot be separated in false positives and wrong positioning given the design of the experiment and, as said before, this figure represents the worst case.

G. Experts' results in the repeated tests

As explained, the experts repeated the RE test in three sessions, as check points to measure the improvement of their skills in artificial vision. Figure 12 shows the progression for each one of them, after following the training described in Table IV. Tests are marked in bold.

TABLE IV. PROCEDURE FOR EXPERT SESSIONS.

Step	**Time**
Online training	15 min
VR training and testing	**1h30**
Table test 1	**30 min**
VR training and testing	**1h30**
Table test 2	**30 min**
Free play with objects	20 min
Table test 3	**30 min**
Walk around in a known room	25 min
Walk around in an unknown room	**30 min**

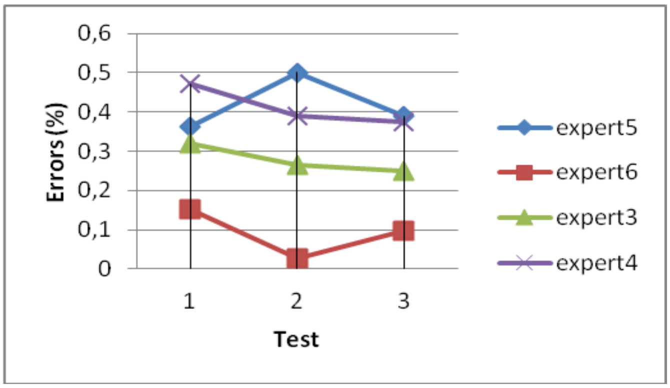

Figure 12. Progression of errors in the RE table in the three different moments for each "expertX" ("X" indicates the profile complexity level of each expert).

H. Experts' results in the pose estimation test

Nine static poses (in random order, but keeping the order in each test) were performed, three times per pose, by the experimenter in front of the experts, who were all blindfolded and at 1 m distance. Figure 13 summarizes the correct and erroneous estimations of the poses.

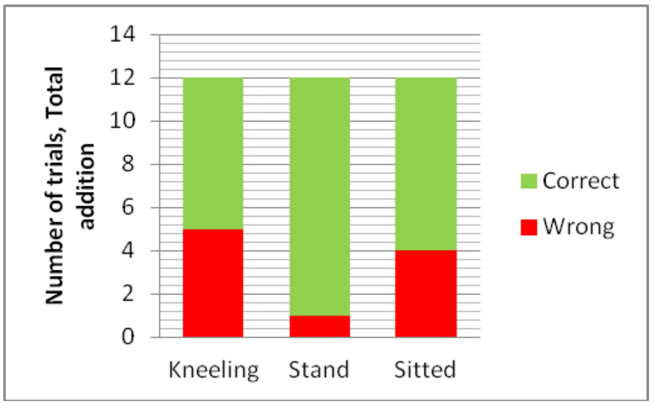

Figure 13. Addition of correct pose estimations (in green) and incorrect estimations (in red) for all experts.

I. Focus group

A one-hour discussion between the experimenter and the four experts took place to analyze subjective and qualitative aspects of the system and the perception of the whole process.

When discussing the sonification protocol itself, some of them pointed out the difficulties due to the fact that the same object, when it is tall, produces different pitches. Thus, taller objects were the hardest to be understood. However, some others remarked that they got used to this problem and they felt easier every additional time they did the test. Also, some problems were found when trying to identify the moving object in scene 4 of the VRE test.

Regarding the real system, two main problems were marked:

- The position of the eyes (under the helmet and, thus, the cameras) doesn't match with that of the cameras, so the head (the only physical reference of the participant) is not pointing where the cameras are. This displacement caused some errors and difficulties in the RE test, and the position of objects is confused because of this. For the blind participants, this could interfere with the natural reference taken by the users via touching the table.
- The real system has some errors (the cameras have auto-exposure functions that put into troubles the stereovision algorithm) and it is more confusing than the VR system. This problem is critical when pointing to non-textured surfaces, where the stereovision algorithm encounters more problems. However, it was quite easy for them to know if there was something there, one of them said.

Two main strategies to know what is in front of them were discussed: the so called "scanning" (when the participant just focused on one single tone -that of the middle in general- and tried to find the objects with this single sound), and the "holistic" strategy, where the user tried to figure out the whole scene by attending to the complete combination of sounds. The first one was used by all of them in the table test, they remarked. However, in the VRE, some of them said they had used the holistic approach to understand the scenes. Another one said that in the RE test, s/he used the holistic approach first, and then started to scan the table for more accurate perceptions.

When discussing the limitations of the system, they also found it difficult to know what was below their knees when standing up. In this line, the pose estimation test was easier, some of them said, however they agreed that distinguishing between kneeling and sitting was not that easy. One of them said this test was hard. The edges of the table presented important problems and it was hard, they said, to know whether there was an object or it was empty in these cells. Likewise, one of them pointed out the confusing effect of the vibrato when many objects are in the scene.

As a general evaluation, though, "the mapping (sonification protocol) is fine," one of them said. However, the noise and the mismatch between the cameras and the eyes remained the most important problem. Another important complaint was the number of sounds in level 6; so many, they said, it became confusing. In the same way, another one of them suggested the necessity of cutting some information before sonifying, since there is a lot of redundant or irrelevant information that only adds confusion to the perception.

A final problem noted was the shortness of the training, and they agreed that with longer trainings the usability of the system would increase dramatically.

They agreed that it was easy to feel there was something there, overall when it was a wall and all the tones were excited.

J. Final Evaluations

Final questions about the global process were asked to the participants, among which the tiredness of the global process, considerations about the length of the training, feelings of safeness and use of the white cane or the guide dog.

Marginally significant Pearson correlation was found between the educational level and the perception of the tiredness of the whole process (r=0.330, p=0.086). This result is shown in Figure 14.

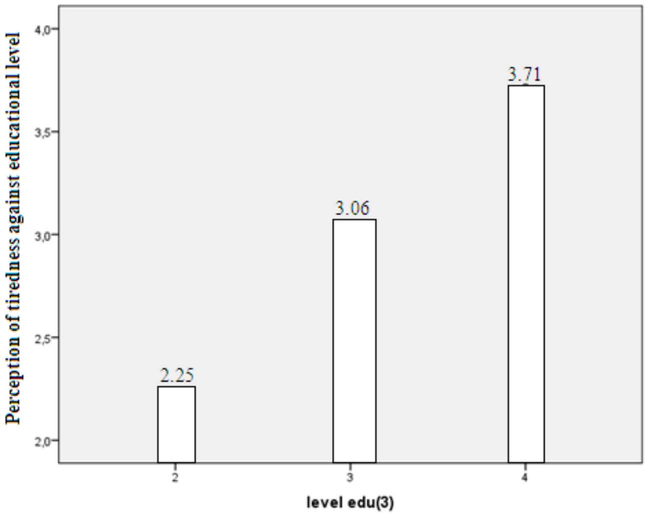

Figure 14. Perception of tiredness (ranged between 1 –not tiring at all- and 5 –very tiring-) against the educational level.

Another interesting result is the distribution of the same perception against the level, shown in Figure 15.

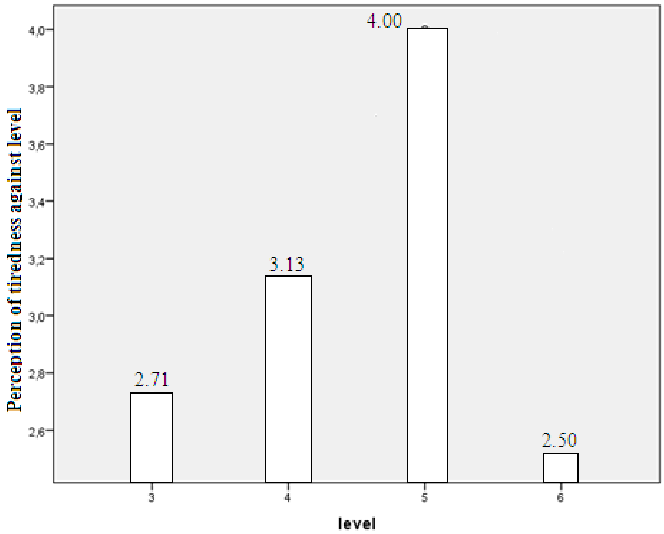

Figure 15. Perception of tiredness (ranged between 1=not tiring at all, to 5=very tiring) against the profile complexity level.

An unexpected result comes from the comparison of the use of computer and the perception of safeness (marginally significant with a one way ANOVA, F(3,24)=2.707, p=.068), shown in Figure 16.

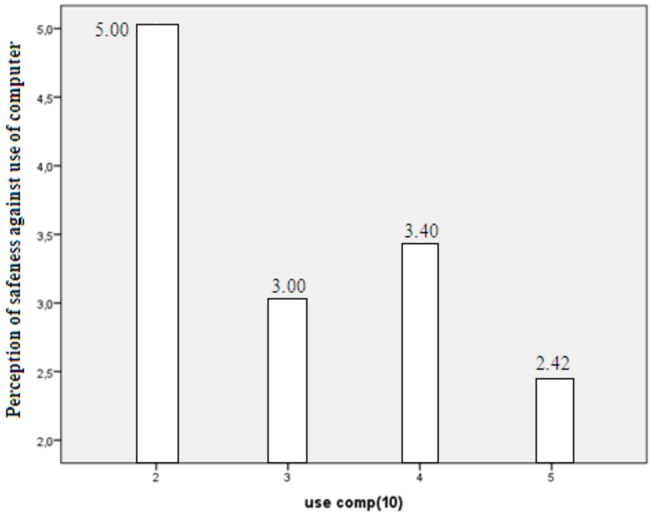

Figure 16. Perception of safeness (ranged between 1=not safe at all, to 5=very safe) against the use of computer.

Another marginally significant result (one way ANOVA, F(2,11)=3.378, p=0.072) is the relation between the intention of keep using the white cane or the dog guide against the visual impairment (with a positive Pearson correlation of 0.585 and p=0.028).

The lower correlation was found between the visual impairment and the perception of tiredness (r=0.030, p=0.878).

IV. Discussion

First of all, we should discuss the specific composition of the participant pool available for this experiment. The high correlation found between visual impairment and other theoretically independent variables such as use of computer, educational level or age is due to the composition of the clients and staff of the CVI, compared to the average student of the university (much younger and using computers daily). This is an important point to be discussed: the American society, in which this experiment was carried out, present strong differences in the access to public health systems and, thus, it is not surprising that blindness may be related with social class (as shown in [33]). Moreover, Wilson *et al.* [34] found these differences (as well as a correlation between incomes and level of education) only relevant among the whites, and nonsignificant among Hispanics and African Americans. It was an unsolved problem, during the study, to find homogeneous sample among ethnical groups, educational level, use of computer, etc. Nonetheless, as it was shown, some important results have been found, in terms of limits in the accuracy perceived, the utility for all the groups for some basic tasks related with objects detection, etc. We will discuss them in detail in this section.

Regarding the sample size, 28 participants (and 4 experts) are not a very big set of subjects, and the quantitative data obtained should be contrasted with larger experiments. This one must be taken as a preliminary study of the usability and efficiency of the system proposed in [27]. No important limits were found, so the evaluation must be kept open for further experiments.

The first hypothesis states that the sonification serves to represent spatial configurations of objects. We found in the VRE tests that the number of correctly identified objects was higher than the number of incorrect identifications (1.71 out of 3 versus 1.39 out of 3, in each of the first and second tests; 1.96 versus 0.5 in the third test; and 1.79 versus 0.79 in the fourth). Similar or even better results were found in the RE tests (66% accuracy). It can be concluded, with some caveats related to distance localization, that the sonification protocol was successful at conveying the location of objects in the real world. The caveats are details related to the subtleties of the sonification design, the limits in the human auditory system (as it applies to H2), and the autocorrelations within the set of H3 hypotheses.

Regarding the first two parameters, we can consider the reduction in the detection of objects between the first and second scene of the VRE test. As seen in Figure 3, the first scene presents objects at the same distance, whereas the second one presents the objects at three different distances from the observer. This reduced the correct detection rate in the second test, because of the limits of auditory perception in the human auditory system.

The hypothesis H2, proposing a relationship between the profile complexity level and the efficiencies of the tests (and, maybe, the perception of usability by the users) could be tested independently of the demographic characteristics of the participants pool, assigning levels to each visual category. We found what may seem to be two apparently contradictory results, shown in Figures 6 and 7. In the first result, the average number of false positives (in the second scene) increases until profile complexity level 5, and then decreases for the sixth (and most complex) profile. In the second result, the number of correct detections (in the third scene) increases until the profile level 5, and then decreases for the sixth level. The second result seems to be easier to explain: the higher the sonification profile's complexity, the more information available to the user. This may cross a threshold for auditory information presentation, which could lead to a decrease in performance in the last (most complex) profile level. This is also shown in Figure 11 with the errors in the RE test and in the focus group discussion about the problems with this level. However, Figure 6 (relating the errors in the VRE test, second scene), Figure 8 (correct detections in the fourth scene of the VRE test), and Figure 15 (subjective perception of tiredness of the process) show the opposite results. When evaluating the efficiency of the VRE system in the fourth scene, a simple interpretation can be done: since the height in this test was not relevant (and this was part of the instructions given to the participants), and only horizontal and distance information mattered, the higher the profile complexity level, the more redundant information is being provided for the user and, thus, the more problems s/he may encounter to understand the horizontal position of the objects. In the first case (i.e., errors in scene 1 of the VRE test) we can explain this result recalling one comment of the experts during the focus group: it was difficult to understand the scene when the same object produced more than one tone. In that case (and remembering that we are representing false positives), the higher profile complexity levels seems to be related (although not reaching statistical significance) with the problems in understanding the number of objects. This is not contradictory with the previous argument, because higher levels of complexity allow users to more easily understand the vertical combination of objects (with the marked limits) but, at the same time, may produce the perception of "phantom objects" in addition to the real objects.

We found many patterns of results that lend support to the general hypothesis H3 (see Figures 5, 8, 10 and 12 about objective results; 14, 15 and 16 about subjective perceptions, with their correlative statistic data). The connection between previous experience and the performance achieved in this kind of tests was found in almost every single test. Moreover, it also influences the subjective perception of the test itself.

However, due to the specific demographic composition of the subject pool, we did not find statistically significant differences between the COMP, AGE, EDU and VI factors. Even if age, use of computer or visual impairment were correlated in our dataset, sometimes we can find significant results by focusing our attention on some specific factor. More in detail, the higher the familiarity with computers (and the higher the educational level), the better the results. Figure 10 is consistent with the H3.1 hypothesis. People not used to manage computers may find extra difficulties when getting involved in experiments with virtual reality. Figure 9 shows a clear (but again not statistically reliable) relationship between educational level and the number of detected objects. This is also consistent with the H3.2 hypothesis. The reason behind could be that higher educational levels allow better comprehension of the problem and experiment proposed, and faster understanding of the context in which they had to behave. Thus, the final performance has been found to be higher when the educational level increases. H3.3 hypothesis states a dependency of the age with the results. Given the strong correlation of age and visual impairment (0.752, $p<0.001$) and use of computer (-0.708, $p<0.001$) we cannot discuss this variable independently. However, in reference [35], for example, we can find the dependency of the age with the flexibility of the brain, which is consistent with our hypothesis and results of use of computer. When evaluating the H3.4 hypothesis, we found unexpected results. Data represented in Figure 5 shows the low vision group as the best one when interpreting the sonification protocol and the hypothesis of an ordered decrease of performances from sighted to blind people was not supported.

H3.5 hypothesis, due to the small amount of experts performing a longer training, can only be qualitatively evaluated. In Figure 12, we see different tendencies of the different experts. Three of them increased their performance from the first to the last test and two of them had some inverse tendency at some point (increasing the number of errors in some consecutive tests). We should point out that 5 or 6 hours of training for an artificial vision system is still a very short one, given that the reconfiguration of the brain exploiting the cross-modal plasticity needs much more time to appear (see [35] for more details).

The pose estimation test shows some of the potential applications of the system. Blind people often report it is problematic to find free seats in classrooms, bars, or restaurants [36]. We found an accurate perception of the pose of a person in front of the participant. Actually, most of the errors, with one single exception amongst all 36 tests, were due to the confusion between kneeling and sitting (which is not that relevant when looking for a free seat).

The experts allowed us to reach a new understanding of the usability and efficiency of the system, since longer trainings permitted to them to get more familiar with the sonification and the specific problems of the real system. The main problems were found in the presence of visual noise in the real system (this makes it harder to produce reliable sonifications); and in the complexity of profile level 6. Regarding the first issue, we found correlations between the complaints about the noise of the real system (overall the problems to identify the objects or absence of them in the sides of the table) and the spatial distribution of errors in the RE test, Figure 11. Somehow related to this problem, the mismatch of the directions of eyes and cameras seems to be a critical problem which will be solved in the next prototype. Turning now to the second issue of profile over complexity, some different aspects converge: on the one hand, this problem can be solved, as proposed, by cutting out some information (the less relevant data) to increase the usability. On the second issue, we have to differentiate between redundant information and irrelevant information. The main complains of the users were related with the second one, since the tones (16) to represent the height, are not redundant, but somehow perceived as irrelevant, given that this level of accuracy is not necessary to identify objects in the tests (as seen in figures 7 and 8) and can be, actually, counterproductive (see figure 15). Finally, longer trainings can lead to easier and more intuitive understandings of the sonification.

The scanning strategy, although providing less information at any given moment, produces a more accurate localization, at least in the first moments of use of the new product. This approach will be used, in the future, with lower profile complexity levels, to avoid irrelevant information. However, and once again, we think that longer trainings will help users develop an intuitive and holistic use of the system.

Regarding the general evaluations, we did find some unexpected results, such as the direct correlation between educational level and the feeling of tiredness after the tests. People with higher educational level should be more prepared for intellectual and sometimes boring tasks, but they manifested the higher rates of tiredness (Pearson correlation of r=0.33 and p=0.086, Figure 14). This could be explained by the lower levels of criticism usually linked with lower educational levels. The comparison of the feeling of safety with the use of computers, gives us another unexpected result (F(3,24)=2.707 and p=0.068, Figure 16): the group with better results (with higher rate of computer use) felt less safe than those with lower performances.

Finally, as expected, blind people want to keep using the white cane or the guide dog even if they could use this system (4.8 out of 5, as the average response of 3.0 of the sighted group and 4.33 of the low vision group). Given that blind people typically encounter more dangers in the middle height than in the bottom part [36], and the comment of the experts about the higher ease of detecting middle and higher obstacles, the system can increase the safety in the travels of this collective.

The main limitations of this study (the demographic composition of the participant subgroups and the size of the sample) should be the first tasks to be revisit in the future, with the goal of obtaining statistically stronger data to describe the utility, efficiency, and usability of the system.

Longer training, as pointed out by the experts, should be tested to evaluate the real potentials of the system, and more real life tests should also be designed.

V. Conclusions

The sonification can help visually impaired people to perceive information about the spatial configuration of their surroundings. However, this translation must be trained.

In this study we have tested a new assistive product over a set of 28 people, sighted, with low vision and blind, trying to find its limitations and strengths. The main problem of our experiment is the limited size of the participants' pool. Another weakness of the experiment is the biased sample, with two main groups in terms of age, blindness and cultural class. However, we found important advantages in the proposed system, with high degree of accuracy with virtually no training. Users were able to detect tiny objects in a table via sonification, and follow walls avoiding obstacles in a virtual reality system. Finally, the detection of people (and their different poses) in front of them in real environments was generally perceived as affordable. The research, finally, needs to deepen with larger and more homogeneous samples.

Acknowledgment

We would like to acknowledge the collaboration of the Center for the Visually Impaired of Atlanta, GA, for their support and helpful contacts. We would like to thanks, as well, the GeorgiaTech and the Carlos III University for their facilities and economical support to carry out with this study.

References

[1] B. N. Walker, A. Nance, and J. Lindsay, "SPEARCONS: Speech-based Earcons Improve Navigation Performance in Auditory Menus,"

Proceedings of the 12th International Conference on Auditory Display (ICAD 2006). pp.63-68, 2006.

[2] D. Bolgiano and E. J. Meeks, "A laser cane for the blind." IEEE Journal of Quantum Electronic vol. 3 no. 6, p.268. 1967.

[3] A. D. Heyes, "Auditory Information and the mobile,", U. of Nottingham, England., 1979.

[4] T. Heyes, "The domain of the sonic pathfinder and an increasing number of other things." From http://www.sonicpathfinder.org vol. 4.15. 2004.

[5] N. G. Bourbakis and D. Kavraki, "An intelligent assistant for navigation of visually impaired people," 2nd Annual Ieee International Symposium on Bioinformatics and Bioengineering, Proceedings. pp.230-235, 2001.

[6] T. Ifukube, T. Sasaki, and C. Peng, "A Blind Mobility Aid Modeled After Echolocation of Bats," IEEE Transactions on Biomedical Engineering, vol. 38, no. 5. pp.461-465, 1991.

[7] P. B. L. Meijer, "An Experimental System for Auditory Image Representations," IEEE Transactions on Biomedical Engineering, vol. 39, no. 2. pp.112-121, 1992.

[8] D. Aguerrevere, M. Choudhury, and A. Barreto, "Portable 3D sound / sonar navigation system for blind individuals." 2nd LACCEI Int.Latin Amer.Caribbean Conf.Eng.Technol. pp. 2-4. 2004.

[9] E. Milios, B. Kapralos, A. Kopinska et al., "Sonification of range information for 3-D space perception." IEEE Transactions on Neural Systems and Rehabilitation Engineering vol. 11 no. 4, pp. 416-421. 2003.

[10] G. Sainarayanan, R. Nagarajan, and S. Yaacob, "Fuzzy image processing scheme for autonomous navigation of human blind." Applied Soft Computing vol. 7 no. 1, pp. 257-264. 2007.

[11] G. Balakrishnan, G. Sainarayanan, R. Nagarajan et al., "Fuzzy matching scheme for stereo vision based electronic travel aid," Tencon 2005 - 2005 Ieee Region 10 Conference, Vols 1-5. pp.1142-1145, 2006.

[12] G. Balakrishnan, G. Sainarayanan, R. Nagarajan et al., "Stereo Image to Stereo Sound Methods for Vision Based ETA." 1st International Conference on Computers, Communications and Signal Processing with Special Track on Biomedical Engineering, CCSP 2005, Kuala Lumpur , pp. 193-196. 2005.

[13] J. Xu and Z. Fang, "AudioMan: Design and Implementation of Electronic Travel Aid." Journal of Image and Graphics vol. 12 no. 7, pp. 1249-1253. 2007.

[14] D. Castro Toledo, S. Morillas, T. Magal et al., "3D Environment Representation through Acoustic Images. Auditory Learning in Multimedia Systems." Proceedings of Concurrent Developments in Technology-Assisted Education , pp. 735-740. 2006.

[15] J. Gonzalez-Mora, A. Rodriguez-Hernandez, E. Burunat et al., "Seeing the world by hearing: Virtual Acoustic Space (VAS) a new space perception system for blind people," International Conference on Information & Communication Technologies: from Theory to Applications (IEEE Cat.No.06EX1220C). pp.6-ROM, 2006.

[16] L. F. Rodríguez Ramos and J. L. González Mora, "Creación de un espacio acústico virtual de aplicación médica en personas ciegas o deficientes visuales." From: www.iac.es/proyect/eavi/documentos/EXPBEAV_25v1.DOC. 1997.

[17] Y. Kawai and F. Tomita, "A Support System for Visually Impaired Persons Using Acoustic Interface - Recognition of 3-D Spatial Information." HCI International vol. 1, pp. 203-207. 2001.

[18] D. J. Lee, J. D. Anderson, and J. K. Archibald, "Hardware Implementation of a Spline-Based Genetic Algorithm for Embedded Stereo Vision Sensor Providing Real-Time Visual Guidance to the Visually Impaired." EURASIP Journal on Advances in Signal Processing vol. 2008 no. Jan., pp. 1-10. 2008. Hindawi Publishing Corp. New York, NY, United States.

[19] M. Capp and Ph. Picton, "The Optophone: An Electronic Blind Aid." Engineering Science and education Journal vol. 9 no. 2, pp. 137-143. 2000.

[20] F. Fontana, A. Fusiello, M. Gobbi et al., "A Cross-Modal Electronic Travel Aid Device." Mobile HCI 2002, Lecture Notes on Computer Science vol. 2411, pp. 393-397. 2002.

[21] I Lengua, L Dunai, G Peris Fajarnes, B Defez, "Navigation device for blind people based on time-of-flight technology", Dyna rev.fac.nac.minas vol.80 no.179 May/June 2013, pp. 33-41. 2013. Medellín.

[22] Y. Tian, X. Yang, Ch. Yi, A. Arditi, "Toward a computer vision-based wayfinding aid for blind persons to access unfamiliar indoor environments", Machine Vision and Applications. April 2013, Volume 24, Issue 3, pp. 521-535.

[23] C. Vincent, F. Routhier, V. Martel, M.-È. Mottard, F. Dumont, L. Côté, and D. Cloutier, "Field testing of two electronic mobility aid devices for persons who are deaf-blind", Disability and Rehabilitation: Assistive Technology, Ahead of Print : pp. 1-7, (doi: 10.3109/17483107.2013.825929).

[24] G. Bologna, J. D. Gomez, Th. Pun, "Vision Substitution Experiments with See ColOr", Natural and Artificial Models in Computation and Biology, LNCS, Vol. 7930, pp 83-93. 2013.

[25] L. Hakobyan, J. Lumsden, D. O'Sullivan, H. Bartlett, "Mobile assistive technologies for the visually impaired", Survey of Ophthalmology. Vol. 58, Issue 6 , pp. 513-528 , November 2013.

[26] P. Revuelta Sanz, B. Ruiz Mezcua, and J. M. Sánchez Pena, "ICTs for Orientation and Mobility for Blind People. A State of the Art." ICTs for Healthcare and Social Services: Developments and Applications. Isabel Maria Miranda and Maria Manuela Cruz-Cunha, eds. 2011.

[27] P. Revuelta Sanz, B. Ruiz Mezcua, and J. M. Sánchez Pena, "A Sonification Proposal for Safe Travels of Blind People." Proceedings of the 18th International Conference on Auditory Display (ICAD 2012) , pp. 233-234. 2012. Atlanta, GA.

[28] R. M. Fish, "Audio Display for Blind," IEEE Transactions on Biomedical Engineering, vol. 23, no. 2. pp. 144-154, 1976.

[29] P. B. L. Meijer, "An Experimental System for Auditory Image Representations," IEEE Transactions on Biomedical Engineering, vol. 39, no. 2. pp.112-121, 1992.

[30] ICECAT, "NGS NETCam300." http://icecat.es/p/-ngs/netcam300/webcams8436001305400netcam3003943-712.html 2013.

[31] L. Rayleigh, "On our perception of sound direction," Philos.Mag., vol. 13. pp.214-232, 1907.

[32] M. Pec, M. Bujacz, P. Strumillo et al., "Individual HRTF Measurements for Accurate Obstacle Sonification in an Electronic Travel Aid for The Blind." International Conference on Signals and Electronic Systems (ICSES 2008) , pp. 235-238. 2008.

[33] Frick KD, Gower EW, Kempen JH, Wolff JL. "Economic impact of visual impairment and blindness in the United States", Arch. Ophthalmol. pp. 544–50, 2007.

[34] Wilson CJ, Rust G, Levine R, Alema-Mensah E. "Disparities in vision impairment Among adults in the United States". Ethn Dis. 2008;18(Suppl 2):S242–6.

[35] D. Bavelier and H. J. Neville, "Cross-modal plasticity: where and how?," Nature Reviews Neuroscience, vol. 3, no. 443. pp.452, 2002.

[36] P. Revuelta Sanz, B. Ruiz Mezcua, and J. M. Sánchez Pena, "Users and Experts Regarding Orientation and Mobility Assistive Technology for the Blinds: a Sight from the Other Side." Proceedings of the AIRTech Int.Conference 2011. pp. 3-4. 2011.