Annual Review of Cybertherapy and Telemedicine
Volume 2 Year 2004 ISSN: 1554-8716
Interactive Media in Training and Therapeutic Intervention

Editors:
Brenda K. Wiederhold, PhD, MBA, BCIA
Giuseppe Riva, PhD, MS, MA

Interactive Media Institute
3-D Sound and Virtual Reality: Applications in Clinical Psychopathology

Isabelle Viaud-Delmon, PhD1, Angeline Seguelas, MS1, Emmanuel Rio, MS2, Roland Jouvent, MD1, Olivier Warusfel, PhD2

1CNRS UMR 7593, Paris, France
2IRCAM CNRS UMR 9912, Paris, France

Abstract: The aim of this study was to provide information about the importance of auditory feedback in a VR system planned for clinical use, and to address the different factors that should be taken into account when building a bimodal virtual environment. We conducted an experiment in which we assessed spatial performances in agoraphobic patients and normal subjects, comparing two kinds of virtual environments, visual alone (Vis) and auditory-visual (AVis), during separate sessions.

Subjects were equipped with a head-mounted display coupled with an electromagnetic sensor system, and immersed in a virtual town in which they could move forward by pressing a mouse button. Subjects had to turn on their own vertical axis in order to change the direction of heading in the virtual town. Their task was to locate different landmarks and become familiar with the town. In the AVIs condition subjects were equipped with the head-mounted display and headphones, which delivered a soundscape updated in real time according to their movement in the virtual town. The sounds were produced through tracked binaural rendering (HRTF).

The two groups of subjects exhibited better scores of presence in the AVIs condition, although patients exhibited more cybersickness symptoms than normal subjects in this condition. While normal subjects preferred the AVIs condition, expressing a better sense of realism, patients did not mention such a preference. Overall, this study might reflect the multisensory integration deficit of anxious patients and underline the need for further research on multimodal VR systems for clinical use.

INTRODUCTION

The special feature of VR as compared to traditional displays is that the environments it provides are places where as many senses as possible are meant to be active. “Multisensory” is a keyword for virtual reality. The number of sensory modalities through which the user is coupled to the virtual environment is a main factor contributing to the feeling of presence. In spite of that, VR technologies rarely integrate the auditory modality, which is the only sense through which we can communicate with the whole space around us.

Psychiatric patients commonly report a hypersensitivity to auditory stimuli, while pure tone audiograms show generally normal hearing. Sound tolerance is influenced by stress and exhaustion, and specific sounds can cause physical pain and irritability.1 There is a correlation between a tolerance of loud noises and anxiety,2 and strong emotional reactions can easily be elicited in reaction to auditory stimuli.3 It is therefore all the more interesting to more realistically integrate the auditory modality when working with anxious patients, and to understand how it can be used for therapeutic purposes.

Several studies have led to the observation that patients with agoraphobia, panic disorder, or space and motion discomfort (SMD) may have a problem with multisensory integration: subjects with symptoms of panic disorder and agoraphobia experience destabilization under conflicting sensory conditions while maintaining upright posture.5,6 Previous studies creating conflicts between vestibular and visual information with a VR setup have attempted to demon-
strate that the abnormal central processes of multisinseory integration in maintaining balance in anxiety disorders was not restricted to balance control. By integrating the auditory modality into a VR setup, it becomes a more general question as to whether multisensory integration in anxiety can be addressed. Without any additional sensory conflict than the one which is inherent to a VR set-up (due to the delay in feedback between action and consequences of actions in the virtual environment), how do anxious patients cope with interactive auditory modality? Would the introduction of the auditory modality generate sensory enhancement or sensory overload?

The study we present here involves technologies, models, and applications linked to the introduction of 3D sound in virtual or augmented reality environments. Auditory augmentation of visual environments is known to improve feelings of presence and immersion. It also appears very promising in terms of creating user-friendly information systems accessible to everyone. To create such an environment and the corresponding content, several concepts and technologies need to be researched, developed, and/or integrated. The introduction of a 3D sound modality also addresses the need for a better understanding of multisensory integration mechanisms. This includes complementary or conflicting perception between the auditory and visual senses, as well as idiothetic cues (cues generated through self-motion, including vestibular and proprioceptive information). These last aspects are important since many applications can now involve user navigation in a virtual or augmented world, or perception-action cues can be provided by interactive devices.

The most natural audio technique for virtual reality applications is the binaural rendering on headphones that relies on the use of HRTFs (HRTF refers to Head-Related Transfer Function, which is a set of filters measured on an individual or artificial head and used to reproduce all the directional cues involved in auditory localization). This technique still needs some studies to overcome its implementation cost and individual adaptation limitations (a fully convincing spatial rendering requires the use of HRTFs measured on the listener's head).

Incorporating real-time updated 3D sound with virtual reality technologies therefore addresses several practical issues. If there is a consensus on the fact that presence is improved by 3D sound, little is known about how an auditory virtual environment (VE) should be designed so that it does not interfere with the visual VE. We thus conducted a study to provide information on the importance of auditory feedback in a VR system planned for clinical use, as well as information about the different factors which should be taken into account to build a multimodal VE (sense of realism, presence, and coherence between the visual and auditory VE).

If agoraphobic patients are effectively more sensitive to sensory conflicts than normal subjects, multisensory feedback in VR could represent a challenge for them. However, since presence during a bimodal stimulation should be higher, this might provide an interesting way to both convey supplementary spatial information and engage patients in a task. We conducted a study in which we compared navigation performances in a virtual town in two immersive conditions: visual alone (Vis) and auditory-visual (AVis). We intended to test the emotional and behavioral reaction of patients sensitive to space and of normal subjects in order to develop new procedures and find an integrated approach to work with visual and auditory stimuli in VR.

**MATERIAL AND METHODS**

**Design**

All subjects included in the comparative study took part in two sessions of virtual navigation. The order of sessions was counterbalanced so that the same number of subjects began the trial with AVIs and Vis conditions. Sessions were performed at least one week apart. After each session, subjects had to complete several questionnaires and two memory tests related to their experience. In the first test they were presented with a survey view of the virtual town and had to locate the different landmarks they were asked to find during navigation. In the second test they were submitted to a two-choice forced recognition task, during which they were presented with 10 pairs of snapshots and were required to chose between a view taken in the
virtual town they had experienced and a view taken in another town.

The measures taken during the navigation were the number of landmarks found (score out of 11) and the time spent in the virtual town. The measures taken after navigation included the number of correctly localized landmarks on the survey view of the virtual town (score out of 12) and the number of correct answers to the two-choice forced recognition task (score out of 10). Participants were debriefed after each session but were not informed about the content of the following session. At the end of the second session and after the debriefing they were informed of the differences between the two conditions if they had not noticed them.

**Procedure**

Subjects were equipped with a head-mounted display coupled with an electromagnetic sensor system. They were immersed in the virtual town in which they could move forward by pressing a mouse button. Subjects had to turn on their own vertical axis in order to change their direction of heading in the virtual town. Their task was to locate different landmarks (movie theater, swing, bus stops) and become familiar with the town. In the AVIs condition subjects were equipped with the head-mounted display and headphones, which delivered a soundscape updated in real time according to their movement through the virtual town. The sounds were produced through tracked binaural rendering (non-individual HRTF) and were dependent upon the subject's movements.

**Subjects**

Patients were individuals suffering from agoraphobia as their main complaint. Seventeen patients were referred to us for the study at the local day hospital (12 females, 5 males). Two of

![Figure 1. Survey view of the Visual Environment in Vis and AVIs conditions. The subject's task is to find the movie theater, then the swing, then count how many bus stops are in this virtual town. The subject stops navigating when he/she thinks that he/she is familiar with the town and that he/she has localized all the targets.](image-url)
them were excluded from the following analysis since their fear of empty spaces was so severe that they could not accomplish the task. Five patients stopped the protocol after the first session (3 females, 2 males). The remaining 10 patients completed the project (7 females, 3 males). Nine control subjects (7 females, 2 males) were included in the study. The mean ages of the anxious and control samples were 35.3 (SD 10.2) and 32.7 years (SD 10.6), respectively. The control participants were not afflicted with any mental disorders. A semi-structured interview based on the Mini International Neuropsychiatric Interview was administered to all the participants to ensure that they met these criteria.

Questionnaires and Interview Measures

The state portion of the STAI was used to measure the anxiety levels upon arrival at the laboratory and after completion of the experiment. A 22-item cybersickness scale was used to assess the level of discomfort after exposure to VR. It was comprised of a list of symptoms and sensations associated with autonomic arousal (nausea, sweating, pounding heart, etc.), vestibular symptoms (dizziness, fainting, etc.), respiratory symptoms (feeling short of breath, etc.) and could also be used to estimate signs of somatisation (tendency to complain of a large number of diverse symptoms). Items were rated on a scale from 0 to 4 (absent, weak, moderate, strong). The presence questionnaire from the I-group was presented after completion of the experiment.

Visual and Auditory Stimuli

The 3D visual environment was based on a 3D model developed by Sense8 Corp (San Rafael, CA). It was composed of noticeable landmarks, several streets and alleys, and was rendered using Virtuools Dev 2.5 (Virtuools SA, Paris, France). This software has limited features in terms of auditory design, which consists mainly of stereo-rendered sources, even though specific plug-ins may be also implemented. In order to allow the maximum flexibility with regards to sound design, we used the Spat~ sound rendering engine and the ListenSpace auditory scene authoring tool; both developed at Ircam. A specific network interface was developed in order to connect Virtuools with ListenSpace and Spat~, enabling the transmission of position coordinates and rotation angles of the user's head in the virtual world. The sounds were produced through tracked binaural rendering (HRTFs) and were dependent upon the subject's movements.

The model used for the audio environment was almost static; sound sources and their activation zones did not move. Only the movement and position of the subject drove audio events and their spatialization according to their relative coordinates. Two types of auditory elements were used; binaurally rendered monophonic sources and ambisonics sound scenes (Figure 2). Binaurally rendered monophonic sources were put at precise locations in the scene. The ListenSpace environment made it possible to program the position of these sources, and to define the small activation areas where the sources could be heard. Ambisonics sound scenes, linked to large activation areas, were added to the soundscape. Ambisonics is a four-channel audio format that embodies spatial information of a sound scene according to the three directions of space (left/right, front/back and up/down), thus allowing full immersion of the listener inside an auditory environment. The ambisonics sound scenes were recorded using the dedicated ST 250 Soundfield microphone in an urban environment, which worked well with the visual context of the experiment. These sound scenes were decoded in real time for reproduction over headphones according to the listener's position in the virtual space. The large activation areas covered the whole town, so that the subject was either at the center of one sound scene or in a cross-fade region between two sound scenes. The cross-fade mechanism was tuned to ensure smooth transitions between the four sound scenes.

Virtual Reality Set-Up Specifications

We used a V8 head-mounted display (Virtual Research Systems, Santa Clara, CA). The LCD displays had a monocular field of view of 48° by 36°, with an array of 640x480 (true VGA) color triads (pixels), refreshed at 60 frames per second. The subject's head orientation was measured by an electromagnetic sensor system (Fastrak Polhemus) which has an update rate of 120 Hz. The image generator (2.4GHz Pentium IV with 512 megabytes of RAM and an NVIDIA
Quadro4 750 XGL graphics card) took the head angular position information from the tracker and sent the corresponding image to the display and to ListenSpace (Pentium IV 2.4GHz), which calculated the position of the sound sources with respect to the head angular position information and sent them to the Spat~ (Mac 1GHz), which generated the sound. The Mac was equipped with a Hammerfall DSP system. Sennheiser HD570 circum-aural open headphones were used in the AVis condition.

**Statistical Analysis**

To assess the effect of auditory modality in the procedure, repeated measures ANOVAs (2 x 2) were performed on the different scores as the dependent variable, with condition (Vis and AVIs) as a within-subjects factor and with group (patients and control) as between-subjects factor. The effect of VR session on state anxiety was evaluated with repeated measures ANOVAs (2 x 2 x 2). To check for a potential presentation order of conditions effect on the different variables, repeated measures ANOVAs (2 x 2) were performed on the different scores as the dependent variable, with presentation order of conditions (Vis first and AVIs first) as a between-subjects factor and with condition (Vis and AVIs) as a within-subjects factor. Non-parametric statistical tests were used when needed.

Figure 2. Auditory environment in AVIs condition. The sounds are played according to the computed position and distance of the subject with respect to the source when he/she enters an activation area. Large activation areas: ambisonic sounds scenes (four channel audio format) recorded in a urban environment. Small activation areas: binaurally rendered monophonic sources.
RESULTS

Of the 17 recruited patients, two females had to be excluded because of strong emotional reactions. Interestingly, the protocol served as a therapy session for the two of them, who eventually managed to perform part of the navigation task at the end of the second session (with 2 landmarks found). Five patients completed only one session. Two did an AVis condition (2 females) and 3 did a Vis condition (1 female, 2 males). Only the 10 patients who completed two sessions were included in the comparative analysis (see Table 1).

Presence

The two-way ANOVA with condition as a within-subjects factor and with group as between-subjects factor on presence scores indicated a main effect of condition (F(1.17)=7.3, p=0.01). Presence scores were higher in the AVis condition in both groups of subjects (Table 1). The analysis of variance with condition as a within-subjects factor and with presentation order of conditions as a between-subjects factor indicated only a main effect of condition (F(1.17)=9.29, p<0.01). However, the interaction between presentation order and condition was marginally significant (F(1.17)=4.1, p=0.06). Indeed, presence scores increased during the second session only in Vis first presentation order. In AVis first presentation order, presence scores decreased during the second session (for Vis first, Vis=39.6, SD=15.2, AVis=41.7, SD=18.7; for AVis first, AVis=41, SD=15, Vis=30.7, SD=14.7). This observation is in agreement with the finding that auditory modality improves the sense of presence, since removing it has the opposite effect.

Cybersickness

The ANOVA with condition as a within-subjects factor and with group as a between-subjects factor on cybersickness scores showed an interaction between the factors group and condition (F(1.17)=10.6, p<0.01). Cybersickness scores significantly increased in AVis condition in patients group. It is unlikely that these scores represent signs of somatisation since there is no difference between the two groups of subjects in the Vis condition. The analysis of variance with condition as a within-subjects factor and with presentation order of conditions as a between-subjects factor indicated only a main effect of condition (F(1.17)=6.1, p<0.05).

State Anxiety Levels

The two groups differed in all measures of state anxiety (Table 2). A three-way analysis of variance with two repeated measures on condition and the two state anxiety scores (before and after the VR session) was performed between the two groups (2 x 2 x 2). The analysis indicated a main effect of group (F(1.17)=12.9, p<0.01) but there was no interaction between the different factors, suggesting that state anxi-

<table>
<thead>
<tr>
<th>Measure</th>
<th>Vis condition</th>
<th>AVis condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control group</td>
<td>Patients group</td>
</tr>
<tr>
<td></td>
<td>(n=9)</td>
<td>(n=10)</td>
</tr>
<tr>
<td>Time in VR (in sec)</td>
<td>474 (206)</td>
<td>509 (194)</td>
</tr>
<tr>
<td>Landmarks found during navigation</td>
<td>9.7 (2.9)</td>
<td>8.6 (2.9)</td>
</tr>
<tr>
<td>(Max=11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-choice forced recognition</td>
<td>6.9 (2.0)</td>
<td>7.7 (1.4)a</td>
</tr>
<tr>
<td>(Max=10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correctly localized landmarks</td>
<td>7.3 (2.6)b</td>
<td>5.0 (3.3)b</td>
</tr>
<tr>
<td>(Max=12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cybersickness</td>
<td>5.2 (5.3)</td>
<td>6 (7.8)</td>
</tr>
<tr>
<td>Presence</td>
<td>32.4 (13.1)</td>
<td>38 (17.2)</td>
</tr>
</tbody>
</table>

Table 1. Means and (SD) of the scores to the different measures related to VR according to the group and the condition.
ety levels before and after the VR session were not different for any of the groups.

Navigation in Vis and Avis Condition

No difference between the groups or between the conditions was found in the two measures taken during navigation (time spent in VR and number of landmarks found). The analysis of variance with condition as a within-subjects factor and with presentation order of conditions as a between-subjects factor indicated an interaction between presentation order of conditions and condition on time spent in VR (F(1.17)=11.6, p<0.01). Indeed, time spent in VR always decreased during the second session, but time spent in VR during the first session was much shorter in the case of AVis first (for Vis first, Vis=573, SD=149 sec, AVis=438 sec, SD=162; for AVis first, AVis=462 sec, SD=220, Vis=403 sec, SD=209). Since time spent in VR is a measure of time to find the landmarks and to feel familiarity with the town, this indicates that AVis condition does facilitate more efficient spatial exploration.

Correctly Localized Landmarks (After Navigation)

No difference between the groups or between the conditions was found. The analysis of variance with condition as a within-subjects factor and with presentation order of conditions as a between-subjects factor indicated an interaction between presentation order and condition (F(1.17)=12.9, p<0.01). Performance always increased for the second session, but the increase is higher when the first session was in AVIs condition (for Vis first, Vis=5.1, SD=3.3, AVis=6.5, SD=3.1; for AVis first, AVis=5.3, SD=3.3, Vis=7.2, SD=2.7).

Table 2. State anxiety levels at the beginning of and after completing a session

<table>
<thead>
<tr>
<th></th>
<th>Control Group (n=9)</th>
<th>Patient Group (n=10)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Vis condition</td>
<td>AVIs condition</td>
</tr>
<tr>
<td>Time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before the session</td>
<td>23.7 2.7</td>
<td>23.4 2.7</td>
</tr>
<tr>
<td>At the end of session</td>
<td>22.7 4.2</td>
<td>24.4 5.1</td>
</tr>
</tbody>
</table>

Two-Choice Forced Recognition (After Navigation)

No difference between the groups or between the conditions was found with either analysis of variance. However, the interaction between group and condition was marginally significant (F(1.17)=3.04, p=0.09). While normal subjects’ performance tended to increase in AVIs condition, patients’ performance tended to decrease in comparison with performance in Vis condition. No presentation order effect could explain this observation.

DISCUSSION

In the present experiment, we compared navigation in a visual VE with navigation in an auditory-visual VE in two samples of subjects. As expected, presence scores were significantly higher in AVIs condition. Presentation order of conditions was not at stake in this result, since the order of sessions was counterbalanced across subjects. However, analysis of the effect of presentation order pointed to a beneficial effect of auditory modality: time spent in VR to find the landmarks was shorter in AVIs first order, and the increase in the number of correctly localized landmarks at the end of the second session was higher in AVIs first order.

The two groups behaved differently with regard to the two conditions. While normal subjects did not exhibit more signs of cybersickness in AVIs condition, the level of discomfort of anxious patients was significantly higher in this condition. In addition, anxious patients had poorer performances at a visual recognition test in AVIs condition. These results are in agreement with the hypothesis of a multisensory integration deficit in this population. The attempt to continuously adjust the relative weighing of auditory, visual, and idiothetic information may have
caused an attention load which prevented allocation of attention resources to the VE.

In the setup we used, the imperfect mapping between the motor outflow and the multiple sensory feedbacks (movement of the head and its visual and auditory feedbacks) could be the cause of the increased symptoms of cybersickness in the AVIs condition. It exhibited the significance of mastering the delay between sound and images updating so that no supplementary conflict is introduced.

The present experiment confirmed the importance of 3D audio for the construction of a virtual space. Control subjects said that the experience was more compelling when 3D auditory information was delivered during the virtual navigation, while several patients reported that the two worlds (auditory and visual) could not fulfill a sense of realism when presented together. The visual world we were using was composed of rich textures attempting to model a realistic urban environment. The auditory world was mainly composed of ambisonic sounds recorded in a city, which corresponded to highly textured sounds. In spite of the equivalent richness of both channels, anxious patients tended to perceive them separately. If this mode of perception can be linked to a sensory overload originating from their high sensitivity to multisensory information, a question remains at the semantic level with regards to the sensation of coherence between the visual scene and the auditory scene.\cite{14} Focusing on this issue, we are currently conducting an experiment in which the visual environment is purely symbolic. Assuming that patients would not have a dual mode of perception in a symbolic VE, it might be possible to unravel primitives that might be sufficient to elicit emotional reactions, presence and rehabilitation.

Computationally, it is currently easier to achieve high resolution and realism in an auditory VE than in a visual VE. In an attempt to address this issue, we conducted a trial on four patients who were asked to navigate blindfolded in an auditory only VE and were surprised to observe that time of immersion tremendously increased (patients were willing to explore the VE as long as possible) while navigation was efficient (all auditory landmarks were found). Furthermore, realism was judged as very high and patients produced an accurate graphic reproduction of layout of auditory landmarks. This condition seems promising for research in therapeutic methods in which VR should not limit its aim to copying reality, but should invent new ways to engage the immerged subject. Applied research with virtual sound has been performed in the last decade in order to allow the visually impaired to develop more accurate and extensive knowledge of spatial layout.\cite{15,16} Hopefully this kind of VR will provide therapeutic benefit for all kinds of populations.

**Acknowledgements**

We are grateful to Olivier Delerue for providing ListenSpace, the software which enabled the authoring of soundscapes, and Guillaume Vandernoot for his work on HRTFs measurements. We thank Ludvine Sarlat, Feryel Znaïdi, Antoine Pelissolo, and Johana Santos for their help with the patients. This study was supported by the French program “Cognition et traitement de l’information” from the CNRS, grant CTI 01-54.

**REFERENCES**

8. Viaud-Delmon I, Berthoz A, Jouvent R.


Submitted: January 7, 2004
Accepted: May 24, 2004

CONTACT

Isabelle Viaud-Delmon
CNRS UMR 7593, Pavillon Clérambault
Hôpital de la Salpêtrière
47 boulevard de l’Hôpital
75013 Paris, France
Ph: +33 1 44 23 07 50
Fax: +33 1 53 79 07 70
E-mail: ivd@ext.jussieu.fr