Spatial audition in a static virtual environment : the role of auditory-visual interaction

Khoa-Van Nguyen[*], Clara Suied[*†], Isabelle Viaud-Delmon[*†], Olivier Warusfel[*]

[*]IRCAM - CNRS UMR 9912
1, place Igor Stravinsky 75004 Paris, France
phone: +33 (0)1 44 78 47 71 , email: khoa-van.nguyen@ircam.fr
www: www.ircam.fr

[†]CNRS UMR 7593 - Hôpital de la Salpêtrière, Paris, France
47, boulevard de l'Hôpital, 75013 Paris, France

Abstract

The integration of the auditory modality in virtual reality environments is known to promote the sensations of immersion and presence. However it is also known from psychophysics studies that auditory-visual interaction obey to complex rules and that multisensory conflicts may disrupt the adhesion of the participant to the presented virtual scene. It is thus important to measure the accuracy of the auditory spatial cues reproduced by the auditory display and their consistency with the spatial visual cues. This study evaluates auditory localization performances under various unimodal and auditory-visual bimodal conditions in a virtual reality (VR) setup using a stereoscopic display and binaural reproduction over headphones in static conditions. The auditory localization performances observed in the present study are in line with those reported in real conditions, suggesting that VR gives rise to consistent auditory and visual spatial cues. These results validate the use of VR for future psychophysics experiments with auditory and visual stimuli. They also emphasize the importance of a spatially accurate auditory and visual rendering for VR setups.

First presented at the Virtual Reality International Conference (VRIC) 2008, extended and revised for JVRB

Keywords: Binaural technology, auditory-visual interaction, auditory localization, virtual reality

1 Introduction

Virtual Environments (VE) enable the user to become an active participant within a virtual world and to give the user a sense of presence in the VE (sense of being in the place depicted by the VE rather than in the physical place where the user is located). The number of sensory modalities through which the user is coupled to the VE is a main factor contributing to the feeling of presence [Loo92] [GW95].

Immersion is also an important factor enhancing presence, and is most often linked to the amount of the sensory receptors (ears, eyes ...) which can be stimulated [SVS05].

Various computer graphics based technologies have been implemented to improve visual immersion. Head Mounted Displays isolate the user from the actual room and CAVE like systems (Cave Automatic Virtual Environment; [CNSD$^+$92]) provide an extended field of view for the virtual environment.

Different virtual auditory displays technologies have been proposed to convey spatial information, such as binaural technologies, Wave Field Synthesis or Ambisonics (see section 2). By providing surrounding cues, spatialised sounds contribute to the participant navigation performances [LGST00], and tremendously improves the presence and the quality of audiovisual simulations in VEs [She92] [GW95] [HB96].

However, without a correct calibration, the auditory and visual rendering components can create multisensory conflicts which may impair the perceived virtual scene or even disrupt the participant adhesion (see [LAVK06] for a preliminary localization experiment in an auditory-visual VE). To contribute to the proper rendering of auditory cues in a VE, we studied auditory localization performances when a participant was immersed in a static auditory-visual VE. We compared the ability to localize auditory, visual, and spatially aligned auditory-visual bimodal stimuli. In addition, we investigated the effect of a spatially disparate visual stimulus on auditory localization judgements. Because of its potential in VR applications, binaural synthesis was chosen for the auditory rendering. The static conditions was mandatory to ensure the proper rendering of the spatial relationship between auditory and visual cues. Once validated, the next step will be to propose complex, dynamic and interactive experiments using VEs.

Overall, the goal of our study was to validate our auditory-visual virtual reality (VR) setup in terms of auditory localization performances, under different auditory and/or visual conditions. Results of this study are important for both the psychophysics and the VR communities. The good agreement of the results obtained here compared with the literature shows that VR setups are valid and reliable to conduct future psychophysics experiments. When auditory and visual cues are not aligned, the auditory localization is shifted toward the visual stimulus (the so-called ventriloquism effect; [HT66]). This observation is crucial for a proper design of auditory-visual VR setups, since accurate localization would be required in future applications.

2 Auditory displays

Over the past few decades, numerous techniques have been introduced for audio spatial rendering. Most simple ones are based on the auditory *summing localization* properties [Bla97], which refers to the perception of a phantom source between two loudspeakers positioned symmetrically with respect to the subject's median plane and fed with identical signals.

This property which is at the basis of stereophonic systems has been further exploited using amplitude panning techniques over a pair or a triad of neighbouring loudspeakers arranged in a circular or spherical array around the listener. For instance, Vector Based Amplitude Panning technique (VBAP) [Pul97] has been used for tridimensional (3D) audio rendering in a VR setup [GLS01].

Other techniques aim at recreating a physically coherent sound field using mathematical formalisms describing wave equation in the reproduction area. Among them, the High Order Ambisonic (HOA) [DNM03] and Wave Field Synthesis (WFS) [BdVV93] have received increasing attention. However, these techniques require a large number of loudspeakers (especially when targeting a 3D reproduction) and strong positioning constraints which may become incompatible with the installation of large video screens. Moreover, although providing acceptable spatial rendering for practical applications, these techniques only reconstruct imperfectly the auditory spatial cues [JLP99] [PH05], especially when the listener is not positioned at the centre of the space, or only in a limited frequency band [SR06]. Hence, in their state of the art, these techniques are not appropriate for conducting psychophysics experiments requiring perfect control of the stimuli, especially of the spatial cues.

In contrast, the binaural rendering over headphones or loudspeakers, coupled with a tracking system, still remains the most convenient and accurate technique since it does not rely on any physical or perceptual approximations and only requires a limited hardware equipment [WWF88] [BLSS00] [DPSB00]. Providing appropriate equalization of the measurement and reproduction chain, binaural synthesis offers accurate 3D reproduction of free-field acoustic conditions. Convolved with any virtual room effect [MFT$^+$99], perceptually designed [Jot99] [Pel00] or derived from a 3D model [LP02], participants may be immersed in the desired virtual sound scene totally uncoupled from the real room [Pel01].

Transaural audio is a method used to deliver binaural signals to the ears of a listener using stereo loudspeakers [CB89] [Gar98]. An elegant solution for auditory-visual virtual environment is described in [LAVK06] where the binaural signals are reproduced via transaural decoding on a limited set of loudspeakers. It requires however a complex real-time update of crosstalk cancellation filters according to the movements of the listener. Moreover, it can hardly prevent from possible perturbations occurring from wall reflections in the room.

According to these reasons, the present study relies on the use of binaural rendering techniques over headphones.

2.1 Binaural rendering

The principle of binaural techniques is to filter any monophonic sound source with the acoustic transfer function measured between a sound source in a given direction and each ear canal of a given subject or a dummy head (Head Related Transfer Function, or HRTF). This binaural transfer function contains all the acoustical cues used by auditory perception to infer a spatial localization judgement and may then endow the monophonic source with the desired 3D direction [Bla97].

A large set of directions all around the head are measured to constitute a database of filters. These filters can be exploited in a real-time application to synthesize dynamically the localization of different sources composing the virtual sound scene.

2.2 HRTF measurement and postprocessing pipeline

The experiments reported in this study were conducted using binaural rendering with individualized HRTFs. The Head Related Impulse Response (HRIR) measurements were made in Ircam's anechoic chamber. The measurements were conducted with blocked ear-canal condition [HM96] using Knowles electret microphones inserted in individual ear-canal moulds.

Over the past few years, Ircam's HRTF database has now reached more than 70 subjects measured in two main acquisition campaigns with two different setups differing mainly by the spatial sampling resolution. In the latter (20 subjects), a total of 651 HRIRs distributed among 11 elevations (±40°, ±25°, ±15°, ±7.5°, 0°, 60°, and 90°) can be collected within 1h and 20 min. The azimuthal resolution is 5° in the horizontal plane and is progressively released according to the elevation. In the former (51 subjects), 181 measurements were performed with an azimuthal resolution of 15°. All subjects participating to the present study belong to the former campaign.

HRIRs were measured at 44100 Hz with a logarithmic swept sine [Far00] which length was 2^{14} (16384) samples. Raw HRIR need to be equalized to compensate for the response of the microphones and loudspeakers. Moreover, accurate binaural rendering would also require an individual compensation of the headphones, done with the same measurement setup as for the recording of HRTFs (i.e. measured within the same session and with the same head and microphones). This is hardly feasible in practice since it would restrict any following VR experiment to use the same headphones exemplar (see [WK05] for comparison of several exemplars of the same headphone model) or to measure them with the same microphones as used for the HRTFs database acquisition session. A more flexible approach, known as "decoupled" technique [BL95] [LVJ98] consists in using a reference sound field (free field or diffuse field) to equalize independently the measured HRTFs and the headphones. According to [LVJ98], diffuse field equalization has been preferred to free field equalization since it is less affected by the peaks and notches of the magnitude response which differ a lot between individuals. Diffuse field equalization is also shown to reduce the interindividual differences [Lar01]. The individual diffuse field equalization filter has been estimated from a weighted sum of the magnitude spectra associated to all different HRTF [Lar01] [LVJ98]. To account for the solid angle represented by each measured direction the weights were approximated using the Voronoi surfaces. Experiments described in the present study were conducted using diffused field equalized headphones Sennheiser HD650.

Last, if a desired spatial direction has not been specifically measured, the corresponding HRTF is interpolated using a spatial weighting of the closest measured HRTFs. The interpolation is conducted separately on the interaural time delay and the minimum phase filters. Langendijk et al [LB00] show that no major localization error are noticed when interpolation is conducted with measurement resolution under 20°.

3 Localization Experiment

The assessment of the spatial rendering accuracy of the VR setup is conducted through the analysis of auditory localization performances using binaural synthesis either in unimodal or bimodal conditions. Except in the visual only condition (see Procedure), participants always performed an *auditory localization task* in which they had to indicate the direction of an auditory stimulus possibly associated to a simultaneous spatially aligned or disparate visual stimulus.

3.1 Materials and Methods

The whole session consisted in four blocks, assessing different experimental conditions. The three first blocks concerned the perceived localization of a stimulus in the auditory modality alone (Block 1), in the visual modality alone (Block 2), or in an auditory-visual spatially aligned condition (Block 3). In the last block (Block 4), we evaluated the influence of a visual stimulus on binaural localization performance, when the auditory and the visual stimuli were spatially disparate. The spatial conditions varied according to a) the direction of the stimulus with respect to the participant's midline and b) the spatial disparity between the visual stimulus and the auditory stimulus. The participants were always asked to report on the perceived position of the auditory stimulus (except in block 2, for the visual only condition).

Participants

Six volunteers (two women; mean age 40.6 years, from 31 to 53; all but one right-handed) participated in the experiment. All were naive with respect to the purpose of the experiment. None of them reported having hearing problems and all reported normal or corrected to normal vision. They provided full prior consent for the experiments. They all had performed the experiment with their individual HRTFs. All the participants had previous experience of psychophysics experiments with individualized HRTFs.

Apparatus

The experiments took place in an acoustically damped and sound proof recording studio with the light switched off. They were conducted in our virtual reality setup. The visual scene was presented on a large 300×225 cm^2 stereoscopic passive screen and was projected with two *F20 SX+ ProjectionDesign* backward projectors, equipped with wide angle lenses and circular polarization filters. Participants wore polarized glasses, and had their head fixed by a chin-rest. The screen provided a ±56° field of view (Figure 1).

Virtual 3D visual scenes were rendered with the *OgreVR* API developed by *INRIA*. *OgreVR* is the combination of the graphics rendering engine *Ogre* [MR02] and the library *VRPN* [HSW$^+$01] which provides a transparent interface between an application program and most of the devices used in VR such as tracking systems, flysticks, joysticks etc. These two libraries are open source projects written in C++.

For the audio setup, binaural stimuli (sampled at 44.1 kHz) were sent directly from the application to the stereo output of a RME Fireface soundcard and played back over *Sennheiser HD650* headphones.

The rendering machine was based on an AMD Dual Core Opteron 275 (2.2 GHz). The machine was equipped with two GeForce 7950 GX2 M500 graphic cards.

The participant reported the perceived localization using a virtual pointer included in the virtual scene and controlled by a mouse. The pointing cursor movements were constrained to the horizontal track. The use of a mouse was preferred to a 3D wand or flystick to limit proprioceptive interactions [See02].

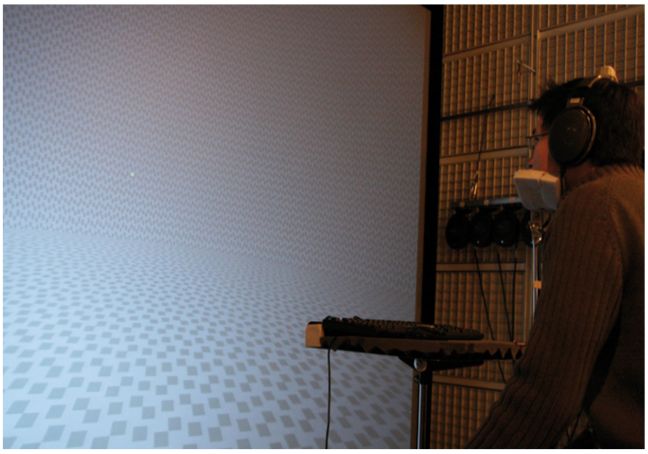

Figure 1: Illustration of the setup used during the experiment. The participant wore headphones and sat in front of the large screen with head fixed by a chin-rest. A mouse fixed on a small tablet allowed to monitor the virtual pointer to report the perceived auditory localization.

Stimuli

The visual stimulus consisted in a virtual orange light spot modulated by a short pulse envelope (50 ms). It was displayed at eye-level and appeared at different fixed locations along a circular segment (radius 500 cm) around the participant. An orange cross positioned at 0° (which was referenced as midline) on the same circle as the visual stimuli was used as a fixation point. To reduce the visual contrast, we added a virtual background scene. Indeed, because the circular polarized glasses present a 15 to 20% visual cross-talk, it is not possible to render a single stereoscopic light spot in a complete dark background without a visible spatial echo effect. This virtual background scene consisted in an 8 m radius cylinder-shaped room centered on the middle of the screen. The wallpaper texture applied to the virtual room boundary was selected for its randomized pattern to avoid obvious visual landmarks.

The sound stimulus was a white noise modulated with a 10 ms onset and offset and a plateau time of 50 ms. It was filtered by individual HRTFs to convey the localization cues associated to different azimuths, at 0° in elevation. As mentioned in section 2.2, diffuse field equalization was applied on the HRTFs but no individual headphone compensation was made. We added a slight synthesized room effect (auditory analog of the virtual room) to facilitate the externalization of the sound (out of head perception). The room effect was made of a binaural diffuse reverberation tail characterized by its reverberation time (150 ms) and its interaural crosscorrelation (iacc(f) modelled after measurement of dummy head in diffuse field). The reverberation parameters were chosen considering the literature on echo detection thresholds as well as fusion thresholds between lead and lag sounds. An initial time delay gap of 5 ms was inserted between the direct sound and the reverberation contribution and an onset cosine envelope of 20 ms was applied to the reverberation tail to avoid fusion with the direct sound and corruption of the perceived sound localization (see [LCYG99] for a review on precedence effect). Furthermore, the ratio between the reverberation and the direct sound level was set to -15 dB. This ratio combined with the reverberation onset time of 20 ms makes the reverberation contribution far below the echo detection curves [Bla97]. A pilot study confirmed that no localization artefacts occurred from the presence of this reverberation contribution. The global level of the sound stimuli was determined using the same sound stimulus design but localized at 0° in azimuth, and set to 65 dB SPL.

Procedure

Participants sat in a chair at a distance of 1 m from the screen with their head orientation and height maintained in a straight-ahead direction by a chin-rest. Each trial began with an orange fixation cross positioned at 0°. After 500 ms, the stimulus (auditory, visual, or bimodal) was displayed. The fixation cross was turned off 300 ms after the stimulus offset, and a visual cursor was turned on, appearing first at 0°. The participant could move then this cursor to the perceived location of the target. The visual pointing cursor consisted in a continuous orange light spot moving along the same circular segment as the visual stimuli, and was slaved to the lateral movements of the mouse. The cursor was turned off after the participant had validated his localization judgement by pressing the mouse key (auditory localization judgement for all blocks except block 2). The response mode we used was designed to minimize interference with any other modality than auditory or visual, as opposed to a localization task with hand-pointing or head pointing that also involves proprioceptive or vestibular interaction. The next trial began automatically 1000 ms after.

Participants were instructed that both visual and auditory stimuli would come from front. Therefore, the front/back confusion that often appears with static binaural rendering was minimized and none of the subject reported any front-back confusion.

In the three first blocks, localization performance was investigated with 7 different directions of target stimuli distributed in the frontal hemifield. The target stimuli could be located directly in front of the participant (0°), in the left hemispace (stimulus direction: -30°, -20°, -10°) or in the right hemispace (stimulus position: 10°, 20°, 30°). Each of the 7 target stimulus directions was repeated 8 times. Thus, each block consisted of a total of 56 trials presented in a pseudorandom order. Each block lasted about 5 min with a pause allowed between blocks.

In the fourth block (in which auditory and visual stimuli were not always spatially aligned), only five of the directions of the auditory target were tested: in front of the participant (0°), in the left hemispace (stimulus direction: -30°, -10°), or in the right hemispace (stimulus direction: 10°, 30°). Seven different values of spatial alignement between the visual stimulus and the auditory target were tested: the visual stimulus could be either to the left of the auditory target (-15°, -10°, -5°), to the right of the auditory target (5°, 10°, 15°), or spatially aligned with the auditory target (0°). As a control condition, the auditory target was also presented alone. The different spatial directions of the auditory target and auditory-visual spatial disparities were fully crossed to build the 40 resulting conditions (5 auditory stimulus directions × 8 visuo-auditory spatial disparities). Each condition was repeated 6 times. The 240 resulting trials were randomly presented. This fourth block lasted around 20 min.

Analysis

To compare the perceived localization through the auditory, the visual, and the conjonction of auditory and visual modalities, the first three blocks were analysed together. Two measures were chosen. Firstly, we calculated the average signed localization error (SE). It is defined as the difference between the perceived localization and the actual stimulus direction. A negative localization error means that the perceived localization is to the left with respect to the actual stimulus direction. A positive localization error means that the perceived localization is to the right with respect to the actual stimulus direction. The associated standard deviation (SD) provides an estimate of the precision of a given sensory modality: the smaller the values, the more reliable the localization percept is (in terms of reproducibility and robustness). Secondly, the absolute localization error (AE) for a given direction was computed. It is defined as the absolute difference between the perceived localization and the actual stimulus direction. It describes the global accuracy of the localization.

To identify between-conditions differences in the SE as well as in the AE, repeated-measures analysis of variance (ANOVA) were conducted with the stimulus angle (-30°, -20°, -10°, 0°, 10°, 20°, 30°) and the modality (auditory only, visual only and bimodal aligned) as within-subjects factors.

In addition, a regression analysis was performed on the SE to evaluate possible under or overestimation of the actual target stimulus in each sensory modality. A linear fit of the SE was computed to describe its general tendency. The statistical validity of this fit was estimated with the Spearman rank correlation (see [See02] for a similar method).

For the fourth block (with spatially disparate auditory and visual stimuli), the goal was to evaluate the biasing effect of vision on auditory localization performances. For this purpose, SE and SD were calculated for each auditory stimulus angles and auditory-visual spatial disparities. A negative SE occurring with a negative auditory-visual spatial disparity means that the visual cue attracted the auditory localization. In the same way, a positive SE occurring with a positive auditory-visual spatial disparity means that the visual cue attracted the auditory localization. An ANOVA on the SE was thus performed with the auditory stimulus angle (-30°, -10°, 0°, 10°, 30°) and the auditory-visual spatial disparity (-15°, -10°, -5°, 0°, 5°, 10°, 15° or A, auditory only condition) as within-subjects factors. The AE was not analysed since it can not exhibit the notion of visual attraction on auditory localization.

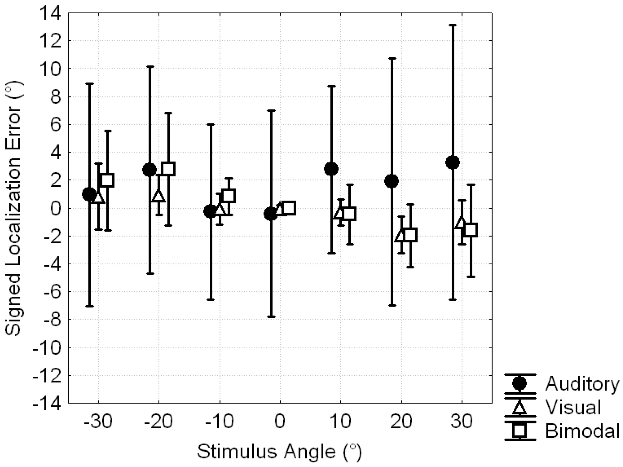

Figure 2: Average signed localization error and standard deviation for each stimulus angle and in each modality condition (block 1, 2 and 3), collapsed across all subjects and repetitions. Localization performances between the three blocks are consistent: for the three conditions, SE is relatively small.

3.2 Results

Unimodal and aligned bimodal conditions (Blocks 1-3)

The repeated-measure ANOVA on the SE did not reveal any significant effect of the stimulus angle (F(6,30) = 1.27, p = 0.3), of the sensory modality (F(2,10) = 1.47, p = 0.25), nor of the interaction between both (F(12,60) = 0.99, p = 0.45). Overall, this shows that localization performances were accurate across all conditions (auditory, visual, and auditory-visual) and all directions (stimulus angles). Localization performances were good, as suggested by relatively small SE (Figure 2). The VR setup provided a coherent spatial rendering across the visual and the auditory modalities.

The second repeated-measure ANOVA performed on the AE revealed a main effect of the stimulus angle (F(2,10) = 34.1, p < 0.00005) and a main effect of the sensory modality (F(6,30) = 8.6, p < 0.0001). There was no significant interaction between the stimulus angle and modality (F(12,60) = 1.17, p = 0.3).

Concerning the effect of the sensory modality, post-hoc analysis (Tukey HSD) revealed that the AE was significantly higher in the auditory alone condition (block 1) compared to the visual alone (p < 0.0005) or the bimodal condition (p < 0.0005). There was no difference between the visual and the bimodal condi-

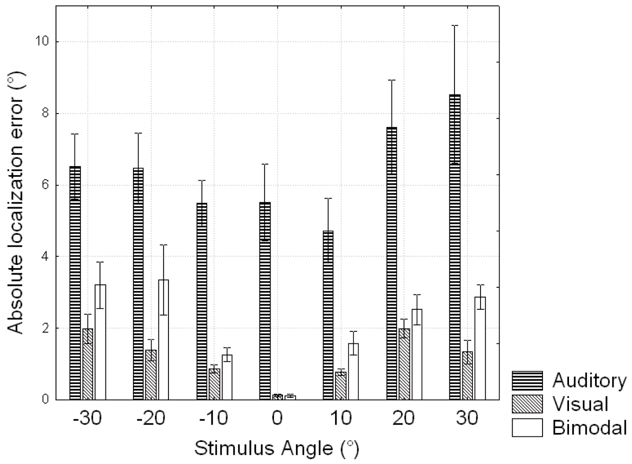

Figure 3: Average absolute localization error for each stimulus angle and in each modality condition, collapsed across all subjects and repetitions. Error bars represent one standard error of the mean.

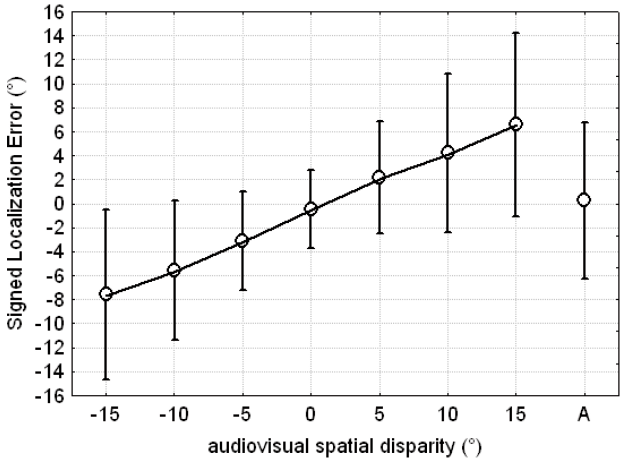

Figure 4: Average signed localization error and standard deviation, according to the disparity between the auditory stimulus and the visual cue, collapsed across all subjects, repetitions, hemispaces and auditory stimulus direction. In abscissa, the *A* label refers to the auditory only control conditions embedded in block 4.

tions (p = 0.4), thus showing a dominance of visual cues in the localization task. Globally, auditory localization with auditory cues only was less accurate than localization under visual or bimodal conditions (Figure 3).

Concerning the stimulus angle, post-hoc analysis (Tukey HSD) revealed that the AE was significantly higher for peripheral stimulus locations (±20°, ±30°) than for central ones (0°, ±10°) ($p < 0.1$).

Regression analyses performed on the SE showed an underestimation of the stimulus angles in the visual as well as in the bimodal condition, but no underestimation or overestimation were found in the auditory condition. In the auditory only condition (block 1), the linear fit of the SE gave a linear coefficient of 0.03 (general tendency of +3%) and an additive constant of +1.5, thus showing a slight shift to the right of the midline. However, this fit was not significant according to the Spearman rank correlation ($r_s = 0.5$, p = 0.3). In the visual only condition (block 2), the linear coefficient was of -0.05 (underestimation tendency of -5%) and the additive constant was of -0.3. This estimation was significative ($r_s = -0.89$, $p < 0.001$). In the bimodal aligned condition, the linear fit gave a linear coefficient of -0.07 (underestimation tendency of -7%) and an additive constant of 0.2. This estimation was significative ($r_s = -0.92$, $p < 0.05$).

Spatially disparate conditions (Block 4)

Figure 4 shows the results for the fourth block. The SE, collapsed across all subjects, repetitions and auditory stimulus directions, is given according to the spatial disparity between the auditory and the visual stimuli.

The repeated-measure ANOVA performed on the SE revealed a main effect of the auditory-visual spatial disparity (F(7,35) = 14.9, $p < 0.0001$) but no effect of the auditory stimulus angle (F(4,20) = 0.2, p = 0.92), nor of the interaction between the auditory stimulus angle and the auditory-visual spatial disparity (F(28,140) = 1.33, p = 0.14). Post-hoc analysis (Tukey HSD) revealed that the effect of the spatial disparity was due to significant differences ($p < 0.006$) between two stimuli when their disparities differed by at least 15°. For example, the SE in the bimodal aligned condition was not significantly different from the SE of the stimuli with ±5° or ±10° of auditory-visual disparity, but became significantly different from the SE of the stimuli with ±15° of auditory-visual disparity. In summary, for all auditory stimulus angles, the perceived auditory localization was shifted toward the spatially disparate visual stimulus.

4 Discussion

This study investigated auditory localization performances under various unimodal and bimodal conditions in a static auditory-visual VE. The results showed that 1) our auditory-visual VR setup renders reliable auditory and visual spatial cues and 2) spatially disparate visual cues influence the auditory localization.

We measured the accuracy of the auditory localization in the context of a VR setup using binaural rendering over headphones. The localization task of the present study, i.e. using static sources and fixed head conditions, is challenging for binaural synthesis since it is subject to intracranial perception and front-back confusion [WK99]. A direct comparison of our results with reference studies dedicated to sound localization in real or binaural rendering conditions is critical since these differ noticeably on the range of the tested source directions (they include generally elevation and back hemifield directions), sound stimuli and reporting method (verbal judgement [WJ89] [WAKW93], head pointing [MM90], 3D pointing on a sphere [GGE+95], etc.). Localization performances obtained in the auditory only condition (block 1) are in agreement with auditory localization studies using a protocol similar to ours. In an experiment investigating auditory localization of real loudspeakers spread from -50° to 50° in the horizontal plane and using a laser pointing task, Seeber [See02] found a median and upper (lower) quartile values of the SE of 1.6° and 1.7° (-1.7°). In the present study the median and upper (lower) quartile values are 1.5° and 5.5° (-4.5°). Both studies show similar shift to the right (see section 3.2 regression analysis). In contrast, the dispersion of the responses is higher in our study. This is probably linked to the use of binaural rendering which increases the localization uncertainty and is also subject to large performance differences among the participants [WJ89]. However, in the study of Hairston et al. [HWV+03], based on a very similar protocol but using real loudspeakers, the standard deviations of the SE are about 8°, thus slightly larger than those observed in the present study (Figure 2).

We also investigated the consistency of the spatial rendering across the visual and auditory modalities provided by the VR setup. Comparing the three first blocks of the experiment, it was shown that subjective localization performances were coherent between visual and auditory modalities when considered in unimodal or spatially aligned conditions. In real conditions, obtaining coherent subjective localization across modalities is obvious since visual and sound devices can be easily positioned at the same physical location. However, in a VE this needs careful calibration of both auditory and visual rendering algorithms as well as their associated display setup.

This study also showed the role of auditory-visual interaction. Vision contributed to clear improvements in auditory localization accuracy when auditory and visual stimuli were spatially aligned. The auditory localization accuracy in the spatially aligned bimodal condition (block 3) was significantly improved compared with the auditory alone condition (block 1). The localization accuracy was similar in the visual only condition and in the bimodal condition. In addition, the slight underestimation of the stimulus angle noticed in the visual only condition (block 2), was observed as well in the spatially aligned bimodal condition (block 3) but not in the auditory only condition. This would suggest that this underestimation may be a strictly visual phenomenon or could have been hidden within the larger deviations of the auditory only condition.

When the auditory and visual stimuli were spatially disparate (block 4) the auditory localization was systematically shifted toward the visual stimulus (spatial ventriloquism effect [HT66] [BR81]). These results are consistent with previous psychophysical studies related to auditory-visual interaction (for instance [HWV+03] [LEG01]).

5 Conclusion

We tested a VR setup, using a stereoscopic display and binaural rendering under auditory, visual and spatially aligned or disparate bimodal conditions. This study showed that the auditory-visual VE was able to deliver correct auditory and visual spatial cues in a static situation and with generic stimuli. The auditory localization performances obtained with our setup are comparable with the ones obtained in real conditions. The present study also confirmed the role of auditory-visual interaction on auditory localization. These results validate the use of VR to conduct psychophysics experiments and highlight the need of a precise alignment between audio and visual rendering channels. The extension of this study with dynamic, interactive VE, and realistic stimuli seems a particularly interesting topic for future investigations.

6 Acknowledgments

This research was funded by the EU funded IST project (FET Open) N° IST-014891-2, http://www.crossmod.org.

References

[BdVV93] Augustinus J. Berkhout, Diemer de Vries, and Peter Vogel, *Acoustic control by wave field synthesis*, The Journal of the Acoustical Society of America **93** (1993), no. 5, 2764–2778, ISSN 0001-4966.

[BL95] Jens Blauert and Hilmar Lehnert, *Binaural technology and virtual reality*, Proc. 2nd International Conf. on Acoustics and Musical Research, 1995, pp. 3–10.

[Bla97] Jens Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, Mit Press, 1997, ISBN 0262024136.

[BLSS00] Jens Blauert, Hilmar Lehnert, Jorg Sahrhage, and Holger Strauss, *An Interactive Virtual-Environment Generator for Psychoacoustic Research. I: Architecture and Implementation*, Acta Acustica united with Acustica **86** (2000), no. 1, 94–102, ISSN 1610-1928.

[BR81] Paul Bertelson and Monique Radeau, *Cross-modal bias and perceptual fusion with auditory-visual spatial discordance*, Percept Psychophys **29** (1981), no. 6, 578–84, ISSN 0031-5117.

[CB89] Duane H. Cooper and Jerald L. Bauck, *Prospects for transaural recording*, Journal of the Audio Engineering Society **37** (1989), no. 1/2, 3–19, ISSN 0004-7554.

[CNSD$^+$92] Carolina Cruz-Neira, Daniel J. Sandin, Thomas A. DeFanti, Robert V. Kenyon, and John C. Hart, *The CAVE: audio visual experience automatic virtual environment*, Communications of the ACM **35** (1992), no. 6, 64–72, ISSN 0001-0782.

[DNM03] Jérôme Daniel, Rozen Nicol, and Sébastien Moreau, *Further Investigations of High Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging*, 114th AES Convention, Amsterdam (2003), no. 5788, 22–25.

[DPSB00] Thomas Djelani, Christoph Pörschmann, Jorg Sahrhage, and Jens Blauert, *An Interactive Virtual-Environment Generator for Psychoacoustic Research II: Collection of Head-Related Impulse Responses and Evaluation of Auditory Localization*, Acta Acustica united with Acustica **86** (2000), no. 6, 1046–1053, ISSN 1610-1928.

[Far00] Angelo Farina, *Simultaneous measurement of impulse response and distortion with a swept-sine technique*, 108 th AES Convention (2000), no. 5093, ISSN 1054-7460.

[Gar98] William G. Gardner, *3-D Audio Using Loudspeakers*, Kluwer Academic Publishers, 1998, ISBN 0792381564.

[GGE$^+$95] Robert H. Gilkey, Michael D. Good, Mark A. Ericson, John Brinkman, and John M. Stewart, *A Pointing Technique For Rapidly Collecting Localization Responses In Auditory Research*, Behavior research methods, instruments & computers **27** (1995), no. 1, 1–11, ISSN 0743-3808.

[GLS01] Matti Gröhn, Tapio Lokki, and Lauri Savioja, *Using binaural hearing for localization in multimodal virtual environments*, Proceedings of the 17th International Congress of Acoustics (ICA 2001 **4** (2001).

[GW95] Robert H. Gilkey and Janet M. Weisenberger, *The sense of presence for the suddenly-deafened adult: Implications for virtual environments*, Presence: Teleoperators & Virtual Environments **4** (1995), no. 4, 357–363, ISSN 1054-7460.

[HB96] Claudia Hendrix and Woodrow Barfield, *The sense of presence within auditory virtual environments*, Presence: Teleoperators & Virtual Environments **5** (1996), no. 3, 290–301, ISSN 1054-7460.

[HM96] Dorte Hammershøi and Henrik Møller, *Sound transmission to and within the human ear canal*, The Journal of the Acoustical Society of America **100** (1996), no. 1, 408–427, ISSN 0001-4966.

[HSW+01] T.C. Hudson, A. Seeger, H. Weber, J. Juliano, and A.T. Helser, *VRPN: a device-independent, network-transparent VR peripheral system*, Proceedings of the ACM symposium on Virtual reality software and technology (2001), 55–61.

[HT66] Ian P. Howard and William B. Templeton, *Human Spatial Orientation*, John Wiley and Sons Ltd, 1966.

[HWV+03] W. David Hairston, Mark T. Wallace, J. Vaughan, B. E. Stein, J. L. Norris, and J. A Schirillo, *Visual Localization Ability Influences Cross-Modal Bias*, Journal of Cognitive Neuroscience **15** (2003), no. 1, 20–29, ISSN 0898-929X.

[JLP99] Jean-Marc Jot, Véronique Larcher, and Jean-Marie Pernaux, *A Comparative Study of 3-D Audio Encoding and Rendering Techniques*, 16 th International conference of the Audio Engineering Society (1999), no. 16-025.

[Jot99] Jean-Marc Jot, *Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces*, Multimedia Systems **7** (1999), no. 1, 55–69, ISSN 0942-4962.

[Lar01] Véronique Larcher, *Techniques de spatialisation des sons pour la réalité virtuelle*, Ph.D. thesis, 2001.

[LAVK06] Tobia Lentz, Ingo Assenmacher, Michael Vorländer, and Torsten Kuhlen, *Precise Near-to-Head Acoustics with Binaural Synthesis*, Journal of Virtual Reality and Broadcasting **3** (2006), no. 2, ISSN 1860-2037.

[LB00] Erno H. A. Langendijk and Aldebert W. Bronkhorst, *Fidelity of three-dimensional-sound reproduction using a virtual auditory display*, The Journal of the Acoustical Society of America **107** (2000), no. 1, 528–537, ISSN 0001-4966.

[LCYG99] Ruth Y. Litovsky, H. Steven Colburn, William A. Yost, and Sandra J. Guzman, *The precedence effect*, The Journal of the Acoustical Society of America **106** (1999), 1633–1654, ISSN 0001-4966.

[LEG01] Jörg Lewald, Walter H. Ehrenstein, and Rainer Guski, *Spatio-temporal constraints for auditory-visual integration*, Behavioral Brain Research **121** (2001), 69–79, ISSN 0166-4328.

[LGST00] Tapio Lokki, M. Grohn, Lauri Savioja, and Tapio Takala, *A case study of auditory navigation in virtual acoustic environments*, Proceedings of Intl. Conf. on Auditory Display (ICAD2000) (2000), Paper session 6 : 3d sound and immersive environments.

[Loo92] Jack M. Loomis, *Distal attribution and presence*, Presence: Teleoperators and Virtual Environments **1** (1992), no. 1, 113–119, ISSN 1054-7460.

[LP02] Tapio Lokki and Ville Pulkki, *Evaluation of geometry-based parametric auralization*, 22nd International Conference: Virtual, Synthetic, and Entertainment Audio (June 2002) (2002), no. 000247.

[LVJ98] Véronique Larcher, Guillaume Vandernoot, and Jean-Marc Jot, *Equalization methods in binaural technology*, Preprints of the 105 th AES Convention (1998), no. 4858, 26–29.

[MFT+99] Philip Mackensen, Uwe Felderhof, Gunther Theile, Ulrich Horbach, and Renato Pellegrini, *Binaural room scanning - A new tool for acoustic and psychoacoustic research*, The Journal of the Acoustical Society of America **105** (1999), 1343, ISSN 0001-4966.

[MM90] James C. Makous and John C. Middlebrooks, *Two-dimensional sound localization by human listeners*, The Journal of the Acoustical Society of America **87** (1990), no. 5, 2188–2200, ISSN 0001-4966.

[MR02] Iain Milne and Glenn Rowe, *OGRE-3D Program Visualization for C++'*, Proceedings of the 3rd Annual LTSN-ICS Conference (2002), http://www.ogre3d.org/.

[Pel00] Renato S. Pellegrini, *Perception-Based Room Rendering for Auditory Scenes*, Preprints-Audio Engineering Society (2000).

[Pel01] Renato S. Pellegrini, *Quality assessment of auditory virtual environments*, Proceedings of the 2001 International Conference on Auditory Displays, Espoo, Finland, 2001, pp. 161–168.

[PH05] Ville Pulkki and Toni Hirvonen, *Localization of virtual sources in multichannel audio reproduction*, Speech and Audio Processing, IEEE Transactions on **13** (2005), no. 1, 105–119, ISSN 1063-6676.

[Pul97] Ville Pulkki, *Virtual sound source positioning using vector base amplitude panning*, Journal of the Audio Engineering Society **45** (1997), no. 6, 456–466, ISSN 0004-7554.

[See02] Bernhard U. Seeber, *A New Method for Localization Studies*, Acta Acustica united with Acustica **88** (2002), no. 3, 446–450, ISSN 1610-1928.

[She92] Thomas B. Sheridan, *Musings on telepresence and virtual presence*, Presence: Teleoperators and Virtual Environments **1** (1992), no. 1, 120–126, ISSN 1054-7460.

[SR06] Sascha Spors and Rudolf Rabenstein, *Spatial Aliasing Artifacts Produced by Linear and Circular Loudspeaker Arrays used for Wave Field Synthesis*, 120th AES Convention (2006).

[SVS05] Maria V. Sanchez-Vives and Mel Slater, *From presence to consciousness through virtual reality*, Nat Rev Neurosci **6** (2005), no. 4, 332–9, ISSN 1471-003X.

[WAKW93] Elizabeth M. Wenzel, Marianne Arruda, Doris J. Kistler, and Frederic L. Wightman, *Localization using nonindividualized head-related transfer functions*, The Journal of the Acoustical Society of America **94** (1993), 111, ISSN 0001-4966.

[WJ89] Frederic L. Wightman and Doris J.Kistler, *Headphone simulation of free-field listening. II: Psychophysical validation*, The Journal of the Acoustical Society of America **85** (1989), 868, ISSN 0001-4966.

[WK99] Frederic L. Wightman and Doris J. Kistler, *Resolution of front–back ambiguity in spatial hearing by listener and source movement*, The Journal of the Acoustical Society of America **105** (1999), 2841, ISSN 0001-4966.

[WK05] Frederic L. Wightman and Doris J. Kistler, *Measurement and Validation of Human HRTFs for Use in Hearing Research*, Acta Acustica united with Acustica **91** (2005), no. 3, 429–439, ISSN 1610-1928.

[WWF88] Elizabeth M. Wenzel, Frederic L. Wightman, and Scott H. Foster, *A virtual display system for conveying three-dimensional acoustic information*, Human Factors Society, Annual Meeting, 32nd, Anaheim, CA, Oct. 24-28, 1988, Proceedings, vol. 1, 1988.

Citation

Khoa-Van Nguyen, Clara Suied, Isabelle Viaud-Delmon, and Olivier Warusfel, *Spatial audition in a static virtual environment: the role of auditory-visual interaction*, Journal of Virtual Reality and Broadcasting, 6(2009), no. 5, March 2009, urn:nbn:de:0009-6-17640, ISSN 1860-2037.