Interaction between auditory and visual perceptions on distance estimations in a virtual environment

Mathieu Paquier, Nicolas Côté, Frédéric Devillers, Vincent Koehl

Interaction Between Auditory and Visual Perceptions on Distance Estimations in a Virtual Environment

Mathieu Paquier[1], Nicolas Côté[2], Frédéric Devillers[1], Vincent Koehl[1]

[1] Lab-STICC CNRS UMR 6285, University of Brest, 6 avenue Le Gorgeu - 29238 Brest, France.
mathieu.paquier@univ-brest.fr - vincent.koehl@univ-brest.fr

[2] Institute of Electronics, Microelectronics and Nanotechnology, UMR 8520 CNRS, ISEN,
41, boulevard Vauban - 59046 Lille, France - nicolas.cote@isen-lille.fr

ABSTRACT

Navigation in virtual environments relies on an accurate spatial rendering. A virtual object is localized according to its position in the environment, which is usually defined by the following three coordinates: azimuth, elevation and distance. Even though several studies investigated the perception of auditory and visual cues in azimuth and elevation, little has been made on the distance dimension. This study aims at investigating the way humans estimate visual and auditory egocentric distances of virtual objects. Subjects were asked to estimate the egocentric distance of 2 to 20 m distant objects in three contexts: auditory perception alone, visual one alone, combination of both perceptions (with coherent and incoherent visual and auditory cues). Even though egocentric distance was under-estimated in all contexts, the results showed a higher influence of visual information than auditory information on the perceived distance. Specifically, the bimodal incoherent condition gave perceived distances equivalent to the those in the visual-only condition only when the visual target was closer to the subject than the auditory target.

1. INTRODUCTION

Navigation in our surrounding environment relies on an accurate localization of the objects present in this environment (e.g. furnitures in a room, cars in a street). The position of these objects is defined by the following three coordinates: azimuth (φ), elevation (θ) and distance (ρ). However, distance perception has received relatively less scientific attention than directional localization (i.e. azimuth and elevation). According to Cutting and Vishton [1] our surrounding (perceived) environment can be divided in three regions: the personal space ($\rho < 2$ m), the action space ($2 < \rho < 30$m) and the vista space ($\rho > 30$m).

Humans use mainly auditory and visual perceptions to localize objects. By using either visual [2,3] or auditory [4] cues, egocentric distance is under-estimated by human subjects (the under-estimation degree depends on the target distance and the reporting method). This under-estimation could be a "margin of safety" for possible danger in our surrounding environment [5]. Actually several studies on distance perception showed that object location is under-estimated in action and vista spaces, and accurately or slightly over-estimated in personal space. The literature shows that vision enables more accuracy in object localization than audition, for azimuth and elevation [6], and for distance [7].

Distance perception in real environments has been well studied but few studies have used a virtual environment. In certain studies, similar under-estimations of the target distance have been observed with real and virtual objects, which suggest possible use of auditory and visual displays for distance perception studies. However, several characteristics of these displays can have an influence on distance perception (distance compression [8]). Even though experiments conducted in real environments have more ecological validity, they enable less control on the experimental conditions than experiments that make use of virtual environments (especially conditions with non-coincident visual and auditory cues).

1.1. Distance perception: auditory cues

1.1.1. Intensity

The intensity of the sound source and the variation of this intensity according to the position of the sound source is the primary acoustic cue to distance. Humans are very sensitive to small changes in overall intensity enabling to detect fine distance variation [9]. In free-field condition and for a

distance higher than 1 meter, the intensity of the source decreases according to an inverse-square low, i.e. a 6 dB decay in sound pressure for a doubling of physical distance [10]. Since the intensity is relative to the sound level of the source, this cue requires prior knowledge of the sound source level to provide absolute distance perception. However, it is important to note that the perceived level of the sound source is also influenced by the signal characteristics such as its time- and frequency-components. Therefore, the cue loudness [11] is more pertinent than the cue intensity for distance perception studies.

1.1.2. Direct-to-reverberant energy ratio

A free-field condition is rarely obtained and all other environments include sound reflecting surfaces producing reverberation. In this case overall intensity is highly influenced by the properties of the listening environment (room size, wall and floor materials, furniture) and of the sound source (directivity). Consequently, the perceived auditory distance is influenced by the ratio between the intensity of the direct sound, I_d, and the intensity of the reverberated sound, I_r. Contrary to the intensity cue, the ratio I_d/I_r provides absolute distance information. In addition, the early reflections create an auditory "image" of the listening environment.

Mershon and King [12] showed that distance judgments of listeners are more accurate in a reverberant environment than in an anechoic one. An increase of the reverberation time results in an increase of the perceived egocentric distance [13]. Even though the direct-to-reverberant energy ratio provides absolute distance information, Zahorik et al. [4] demonstrated that this distance cue is less precise than intensity in order to detect a variation of the sound source position.

It is important to note that the diffuse part of I_r is independent of the sound source position. In addition, the diffuse part of the reverberation can introduce a spreading of the sound source that degrades its accurate localization, especially in navigation task [14].

1.1.3. Coloration

In free field and mainly for large distances, a dissipation of the energy in high frequencies is introduced by the viscosity of the air and the thermal conduction from the compressed zones to the expanded ones [15].

A closed listening environment has also an impact on the spectrum according to its dimension and wall absorption.

In near field sound source ($\rho < 1m$), the listener himself introduces an amplification of the low frequencies due to reflections on his head and torso.

However, as the modified spectrum depends on the original sound, the coloration cue provides information on relative distance only, except for the case of a previously known sound source.

1.1.4. Binaural cues

Directional information corresponds to two binaural cues: the Interaural Time Differences (ITD) and the Interaural Level Differences (ILD) introduced by the distance between the two ears and the head geometry. The advantage of these binaural cues is that they provide absolute information of the egocentric distance: by directly comparing the ILD and ITD cues or by analyzing the influence of head rotations on ILD and ITD cues, the listeners are able to quantify the distance of sound sources [16].

For near-field, i.e. distance below 1m, listeners may take advantage of the ILD to estimate the egocentric distance of sound source that are not in the median plane [17] : the ILD depends on the azimuth of the sound source and increases when the sound source approaches the head. For distances above 1m, Shinn-Cunningham [18] showed that the ILD has a reduced influence on the distance perception.

1.1.5. Auditory displays

According to Zhou et al. [19], the use of specific spatial sound rendering systems enables a better immersion in interactive virtual environments than classic mono or stereo systems. Typical auditory displays used in virtual reality applications are Vector Base Amplitude Panning (VBAP) [20], Wave Field Synthesis (WFS) [21] and Binaural rendering [15]. The latter technique makes use of direction- dependent filters called Head-Related Transfer Functions (HRTFs). These filters encode the ILD and ITD and, for each ear, the transfer characteristics of the listener's pinnea, head and torso for a given direction. Begault [22] showed that binaural rendering is convenient for a sound scene reproduction in three dimensions (azimuth, elevation, distance) and provides an accurate immersion of the listener. In addition, he showed that reverberation reduces the in-the-head localization effect when using headphones.

Bronkhorst [23] suggested that differences between egocentric distances perceived in virtual and real acoustic environments are relatively small and, therefore, virtual acoustics can be used for

distance perception experiments. However, the rendering of realistic auditory scenes require a room effect simulation for a correct localization of sound source in azimuth [24] and in distance [25].

1.2. Distance perception: visual cues

1.2.1. Monocular visual cues

The observer is able to extract depth information with one eye only, i.e. from monocular cues. Occlusion or interposition appears in case an object overlaps another. The covered object is judged to be further away than the completely visible object. Occlusion is a major cue for relative distance.

The size of the visual object provide information on the relative distance. A bigger object is judged closer than a similar smaller object. However, as for the intensity of a sound source, without information of the original size of the object, size provides information on relative distance only.

Outside, aerial perspective is a monocular cue for object distance above 100m. This cue introduces an effect similar to the atmospheric attenuation in audition: distant objects are perceived more blurry than nearer objects.

Perspective is an accurate monocular visual cue. This cue means that parallel lines, e.g. railroad tracks, appear to meet in distance.

1.2.3. Binocular visual cues

Proprioceptive cues include two cues: the convergence and the accommodation. The two eyes synchronously converge, i.e. move together, to look at a given place denoted the fixation point. Accommodation is a dynamic process of the visual system which is linked to convergence. Accommodation enables observers to focus on an object by changing the shape of its crystalline lens. The object placed at the fixation point appears sharp whereas the objects placed behind or ahead appear blurry [26]. Both cues provide accurate absolute egocentric cues within the personal space, i.e. below 2 m [1] since for higher distances, the two visual axes are almost parallel.

Binocular vision, i.e. by using our two eyes, enables us to extract depth information from the differences between the image on the left eye and the right one. These differences are induced by the overlapped fields of views of eyes and enable to extract binocular disparity. Binocular disparity provides relative distance information. Visual objects that are in front of or behind the fixation point create a disparity on the two eyes. From this disparity the observer is able to detect the relative position of the objects compared to the fixation point [27].

Several studies reported that binocular vision provides additional distance information to monocular vision in the personal space only [1, 3]. However Palmisano et al. [28] affirmed that binocular disparity is an accurate relative distance cue for near and far objects.

1.2.4 Visual displays

Virtual reality applications make use of specific interactive visual displays, such as Head-Mounted Displays (HMDs) or Cave Automatic Virtual Environments (CAVEs), in order to immerse the observer in the visual scene. These displays are usually based on stereoscopic and real-time rendering. Such stereoscopic visual displays have several limitations [29]. For instance, they yield to an accommodation-convergence mismatch: in real environments accommodation and convergence are linked together whereas by using visual displays observers accommodate on the image plane (the screen). In addition, they provide a restricted field of view and often a low quality of graphics rendering compared to a real environment.

Some studies on visual distance performances showed differences in distance perception between real and virtual visual environments. Willemsen and Gooch [30] and Loomis et al. [7] showed that visual displays introduced a larger under-estimation of the target egocentric distance than real environments. Klein et al. [8] reported that the use of a reduced field of view introduces a compression of target egocentric distance in visual virtual environments. Plumert et al. [33] suggested that distance perception in virtual environments with large-screen visual displays is similar to distance perception in real environments whereas HMDs introduce underestimation of target distances. Interrante et al. [31] suggested that better estimation of the egocentric distance is obtained in case the virtual environment used during the experiment reproduces the test room, i.e. the real environment which surrounds the subject. Nevertheless, in case the virtual environment does not simulate a real environment Murgia and Sharkey [32] indicated that linear perspective cues provide a sufficient amount of information for distance perception. Other studies showed coherent results between real and virtual environments [27, 34].

1.3. Cue combination and non-coincident cues

In natural environments, auditory and visual distance cues are available. Subjects are thus able to combine both cues to localize an object. However, visual and auditory cues do not equally contribute to perceived distance [35]. According to Spence [36], the congruence of the visual and

auditory information appears at two different levels: the cues should be synchronized in time and space and the semantic information provided by the visual and auditory cues should be coherent. In case the spatially disparate visual object and sound source stay within a spatial window (more or less large depending on the direction), the visible target attracts the perceived sound localization. This specific mechanism called « visual-capture effect » or « ventriloquism effect » introduces a significant bias of sound localization in both azimuth and elevation dimensions [37]. Above this integration window humans are not able to create a unique and coherent multimodal object. The visual and auditory cues are thus perceived as incoherent.

While some authors described the ventriloquism effect as a complete capture of the auditory signal by the visual signal [38, 39, 40], Alais and Burr [41] have shown that this effect can be explained by a model of optimal combination of visual and auditory spatial cues, where each modality is weighted by an inverse estimate of its variability. Our ability to make use of visual cues to localize stimuli typically leads to less variability than our ability of using auditory cues only. So when a conflict arises between these modalities, visual information tends to bias responses to auditory stimuli. However if visual stimuli were blurred so that they would become harder to localize, vision could become worse than audition, and the illusion would work in reverse, with sound capturing vision [41].

The visual-capture effect was more studied in azimut and elevation than in distance. About distance, a classic experimental setup consists in placing several speakers at different distances in a row pointing to the subject (who can see only the first loudspeaker), in an anechoic or semi-reverberant room. The results showed that the subject reported that the sound came from the nearest speaker (whatever the real active speaker). Mershon et al. [42] showed that this « proximity image » effect highlighted by Gardner [43] with an anechoïc chamber operates also in semi-reverberant environment. They further indicated that the distance from the sound source can be overestimated or underestimated depending on the position of the visual target.

In a study [44] with a quite similar setup and with a semi-reverberant room, half of the listeners were blindfolded, and other listeners could see the five speakers in the test room. Results showed that the presence of visual cues increased the accuracy of auditory distance perception and lowered the variability of the judgments (and low direct-to-reverberant energy ratio reduced this effect). However, for distances greater than 2 m listeners still significantly underestimate the distance of the sound source (for distances of 4 and 5 m from the source they never exceeded 3 m). It is also worth noting that authors used five speakers lined up, an arrangement that could provide extra

auditory cues from the filtering due to the acoustic shadow that casts the first speaker (i.e. attenuation of high frequencies for the loudspeakers that are hidden behind the first speaker).

For closer auditory sources, in [45], stimuli were either audio-visual 300-ms broadband noise bursts presented synchronously with spatially congruent or incongruent visual stimuli/LEDs, or auditory-only noise bursts. One of 8 speakers (distance 70 cm to 203 cm) presented a stimulus on each trial. During adaptation runs, the auditory-visual stimuli were presented with the visual-component closer or further by 30% than the auditory-component (displacement direction fixed within session). The ventriloquism effect was observed for both visual-closer and visual-further auditory-visual stimuli, with slightly stronger shifts induced by the visual-closer stimuli.

All experiments mentioned above used real environments with fixed loudspeakers as auditory stimuli and light sources as visual stimuli. The use of virtual environments with visual and auditory displays enables new possible experiments for study of multisensory bias (especially with incoherent visual and auditive cues).

1.4. Considerations about head movements and dynamic rendering

Some studies about sound localisation in real environment showed that head movements allow listeners to improve localization performances. Especially they significantly decrease front-back errors [46] (because systematic changes in ILD and ITD occur when the head rotates). However, the utility of these cues for auditory distance perception is doubtful for far-away sources [4]: for targets with an egocentric distance larger than 1m, ILD and ITD do not change with distance, and so the head movements do not improve the distance perception [47]. The interest of head tracking for dynamic audio rendering (usually binaural) in virtual environments was also studied. The increase of localisation performances with a dynamic auditory system is quite low, however the front-back errors decreased, as in real environment [24].

About localization of visual object, it was shown that the angular motion of the image of the object could be used for judging the size and distance of the object [48,49]. Moreover during translatory motion of the head, the images of near objects move more rapidly than those of far objects, relative to the head [50]. Thus, visual dynamic rendering increase the distance perception of visual objects [51].

1.5. Summary

In real environments, the literature shows an under-estimation of the egocentric distance in each visual-only and auditory-only context. When both the visual and auditory informations are available and give coherent informations, the precision of the distance perception of auditory target is enhanced because of the presence of the visual cues. When both the visual and auditory informations were available, but with incoherent cues, some studies indicate a simple visual-capture effect (ventriloquism effect) whereas some other ones evoke a combination of visual and auditory spatial cues, with a weighting by an inverse estimate of the variability of each modality.

In virtual environments, for each modality the distance perception seems to be even more compressed, but the environment itself (especially the amount of cues) could be partially responsible of this compression. The combinaison of both auditory and visual cues could be differently perceived in real and in virtual environments. Especially the presence of incoherent visual and auditory cues in virtual environment were not studied yet.

Taking all these considerations into account, the present study aims at observing, in a virtual environment, the perceived egocentric distance of 2 to 20 m distant objects in three contexts in order to study the interactions between auditory and visual distance perceptions: auditory perception alone, visual one alone, and combination of both perceptions, with coherent and incoherent visual and auditory cues. Several reverberation times and amounts of visual cues will be considered as previous studies showed their influence. For egocentric distances beyond 2m, binaural indices are not useful and the head movements do not modify the perception of distance of audio object. Therefore, dynamic audio rendering was not necessary in order to study the perception of such distances. Moreover, one of the aims of the present study is to compare results obtained in virtual environments to those obtained by past studies in real environments for which the subject was generally not allowed to move the head. So in this study head movements will be not considered and dynamic rendering (audio and visual) will be not used.

2. EXPERIMENTAL SETUP

2.1 Environment and stimuli

The present study focuses on the influence of acoustic and visual cues on the perceived distance. Therefore, the same visual object and auditory source were processed through different conditions. There were four experimental variables: presentation modality (auditory-only, visual-only and bimodal), target distance (2 m, 3 m, 5 m, 10 m, 20 m), room reverberation (« short »: 370 ms or « long » : 860 ms reverberation time) and amount of visual information (« few » or « many »). The 48 test conditions are summarized in Table 1. The present study investigates the distance perception within the action space, i.e. where the binocular and binaural cues are less accurate than in the personal space [1,18]. In addition, the combination of modalities in distance perception were assessed by using 8 conditions with spatially noncoincident auditory and visual cues. For these conditions, the sound source is placed behind or ahead the visual object.

The experiments were carried out in a test room with the light switched off, except a desk light to enable the subjects seeing the keypad. The visual target consisted in a virtual blue loudspeaker of 40×60 cm2. This virtual target takes advantage of the subjects' belief that loudspeakers is an acoustic device which produces sounds. Loudspeaker as a visual target is also one of the most used in studies with real environments.

The visual environment was a virtual room corresponding to the extension of the real test room through the visual display. Figure 1 shows a schematic top-view of the test room including the virtual extension of the room. The visual display was a 2.4 × 1.8 m2 stereoscopic screen with a 1280 × 1024 resolution combined to passive polarized goggles. Throughout the experiment, the subjects were positioned on a chair placed at 2 m in front of the middle of the screen resulting in a 62° horizontal field of view. The visual environments were rendered by the ARéVi library [52], which is based on OpenGL. Two visual environments have been created:

- An environment with few visual cues: walls, ceiling and floor. This visual environment provides mainly linear perspective cues, see figure 2(a).
- An environment with many visual cues: walls, ceiling, floor with a texture, pillars and neon light. This visual environment provides, in addition to the perspective cues, several visual anchors, see figure 2(b).

A stereoscopic rendering technique has been used to provide binocular cues in addition to the monocular information provided by the visual environment. According to Willemsen et al. [51], a fixed inter-pupil distance does not impair subjects' performances in distance estimation task. Therefore, a fixed inter-pupil distance of 6.5 cm was used for the stereoscopic visual rendering.

The auditory source was a speech signal composed of two french sentences, "Le camp d'été s'est passé au bord du fleuve" spoken by a male speaker and "La voiture s'est arrêtée au feu rouge" spoken a female speaker. The sound stimuli were processed by a binaural rendering system and reproduced through Sennheiser HD600 headphones. For this purpose, the auditory source has been convolved with Binaural Room Impulse Responses (BRIRs) at the different distances listed in Table 1. The BRIR accounted for the Head-Related Transfer Functions (HRTFs) as well as the room reverberation. Since the use of non-individualized HRTF does not affect distance estimation accuracy [53], the BRIRs under use have been produced with a dummy head. The BRIRs are composed of two parts:

1. Since the early reflections create an auditory "image" of the listening environment, a specific room simulation technique has been used up to the second order reflections. The Matlab toolbox "Roomsim" [54], which employs the image-source method, simulates the reverberation introduced by the real test room and its virtual prolongation (i.e. 27.1 × 5.5 × 2.60 m3). This script creates BRIRs that simulate a binaural recording through the dummy head Kemar (these HRIR are included in the « Roomsim » package) placed in a virtual room with different possible characteristics in terms of size and wall type. For the experiment, two different sets of absorption coefficients has been used resulting in two different reverberation. In addition, an atmospheric attenuation model has been used.

2. Since the diffuse-field part of the reverberation is independant of the sound position, the diffuse-field part came from a database of real BRIRs [55], referred to as the Aachen Impulse Response (AIR). This diffuse-field part has been used to provide a more realistic interaural cross-correlation than what could have been obtained using a room simulation (because of the numerous reflections). The BRIRs were recorded at the entrance of blocked ear canals of a dummy head (HMS2 Head Acoustics). The BRIRs recorded in a meeting room and a lecture room has been combined to the two sets of absorption coefficients used for the early reflections (RT_{60} = 370 ms and 860 ms, see Table 1).

The two parts were produced at a 44100Hz sampling frequency and combined to provide realistic reverberation conditions. Speech stimuli were then convolved in the time domain with the resulting BRIRs to generate the binaural stimuli. The stimuli were then sent directly to the Lexicon Alpha soundcard and played back over dynamic, open, diffuse-field studio headphones from

Sennheiser (HD 650). No headphone equalization was performed. Both auditory and visual systems were not dynamic i.e. they did not update dynamically to user-head movements.

Table 1. List of the 48 test conditions

Index	Modality	Auditory target	Visual target	RT_{60}	Visual cues
1	Auditory only	2 m	/	Short: 370 ms	/
2	Auditory only	3 m	/	Short: 370 ms	/
3	Auditory only	5 m	/	Short: 370 ms	/
4	Auditory only	10 m	/	Short: 370 ms	/
5	Auditory only	20 m	/	Short: 370 ms	/
6	Auditory only	2 m	/	Long : 860 ms	/
7	Auditory only	3 m	/	Long : 860 ms	/
8	Auditory only	5 m	/	Long : 860 ms	/
9	Auditory only	10 m	/	Long : 860 ms	/
10	Auditory only	20 m	/	Long : 860 ms	/
11	Visual only	/	2 m	/	Few
12	Visual only	/	3 m	/	Few
13	Visual only	/	5 m	/	Few
14	Visual only	/	10 m	/	Few
15	Visual only	/	20 m	/	Few
16	Visual only	/	2 m	/	Many
17	Visual only	/	3 m	/	Many
18	Visual only	/	5 m	/	Many
19	Visual only	/	10 m	/	Many
20	Visual only	/	20 m	/	Many
21	Bimodal coherent	2 m	2 m	Short: 370 ms	Few
22	Bimodal coherent	3 m	3 m	Short: 370 ms	Few
23	Bimodal coherent	5 m	5 m	Short: 370 ms	Few
24	Bimodal coherent	10 m	10 m	Short: 370 ms	Few
25	Bimodal coherent	20 m	20 m	Short: 370 ms	Few
26	Bimodal coherent	2 m	2 m	Long : 860 ms	Many
27	Bimodal coherent	3 m	3 m	Long : 860 ms	Many
28	Bimodal coherent	5 m	5 m	Long : 860 ms	Many
29	Bimodal coherent	10 m	10 m	Long : 860 ms	Many
30	Bimodal coherent	20 m	20 m	Long : 860 ms	Many
31	Bimodal coherent	2 m	2 m	Short: 370 ms	Many
32	Bimodal coherent	3 m	3 m	Short: 370 ms	Many
33	Bimodal coherent	5 m	5 m	Short: 370 ms	Many
34	Bimodal coherent	10 m	10 m	Short: 370 ms	Many
35	Bimodal coherent	20 m	20 m	Short: 370 ms	Many
36	Bimodal coherent	2 m	2 m	Long : 860 ms	Few
37	Bimodal coherent	3 m	3 m	Long : 860 ms	Few
38	Bimodal coherent	5 m	5 m	Long : 860 ms	Few
39	Bimodal coherent	10 m	10 m	Long : 860 ms	Few
40	Bimodal coherent	20 m	20 m	Long : 860 ms	Few
41	Bim non coherent	1 m	2 m	Short: 370 ms	Many
42	Bim non coherent	20 m	5 m	Short: 370 ms	Many
43	Bim non coherent	1 m	5 m	Short: 370 ms	Many
44	Bim non coherent	10 m	5 m	Short: 370 ms	Many
45	Bim non coherent	20 m	10 m	Short: 370 ms	Many
46	Bim non coherent	3 m	10 m	Short: 370 ms	Many
47	Bim non coherent	15 m	10 m	Short: 370 ms	Many
48	Bim non coherent	5 m	20 m	Short: 370 ms	Many

The listening level (measured with calibrated microphones located at the entrance of the blocked ear canals of a Neumann KU100 dummy head) was set to 63 dBSPL at the location of the observer's head for an anechoic speech stimulus placed at an egocentric distance of 2m. Overall, the listening level stayed within [63.8; 58.7] dBSPL for all auditory-only conditions.

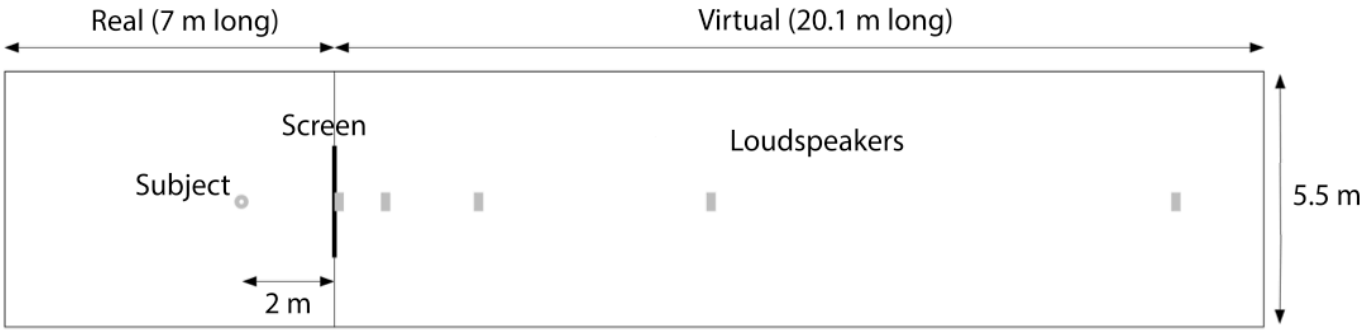

Figure 1. Dimensions of the test room and its virtual extension. The points corresponds to the position of the subject and the visual object.

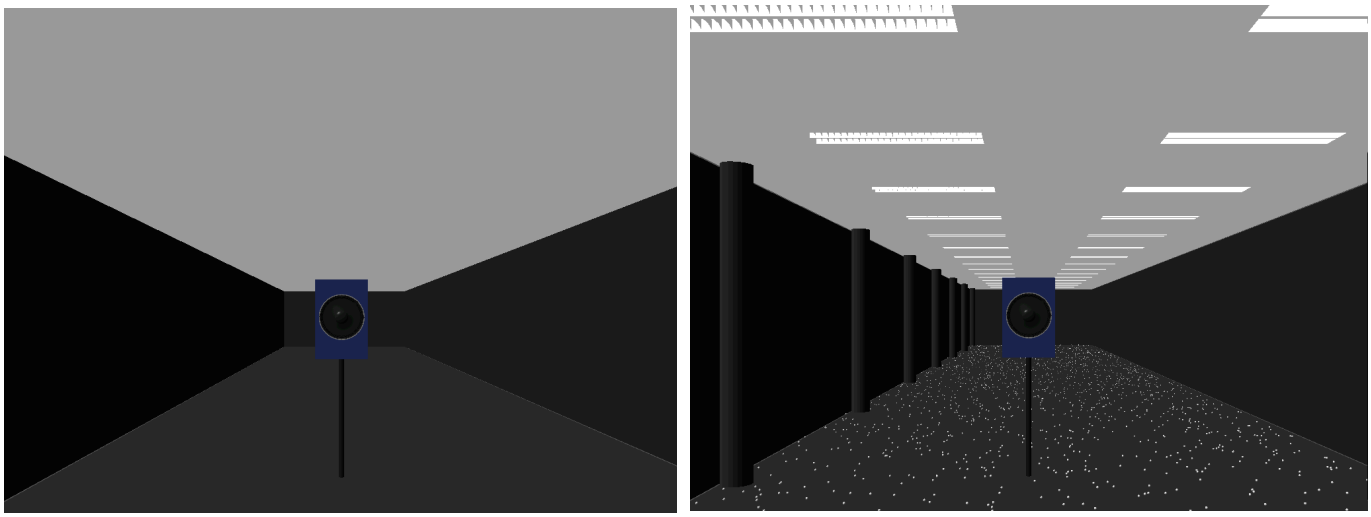

Figure 2. Visual environment created for the experiment. On the left (a) with few visual cues. On the right (b) with many visual cues.

2.2 Procedure

Several procedures can be used for the estimation of target distances. Examples of such procedures are: verbal report based on familiar units (e.g. feet or meters) [56], ratio scale between multiple conditions or multiple objects [2] or directed action such as blind-walking task [7]. This latter procedure takes advantage of subjects proprioceptive sense to provide more accurate object localization: the subjects movements induce dynamic distance cues that provide more information on the object position. In a triangulated blind-walking task, subjects first see/hear the object placed in front of them and then are asked to walk without vision/audition to a specific point indicated by the experimenter. Then, still without vision/audition, the subject is asked to turn and face the target position. The new subject's position and angle provides an estimated egocentric distance of the visual/auditory object.

Da Silva [2] showed that the response procedures have an influence on the perceived values. For instance, directed action enables more accurate distance estimations than verbal reports [8]. However, this procedure provides directional cues while our study focuses on distance perception only. Therefore, a static judgment task procedure has been used in the present study: the subjects were asked to report the egocentric distance of static visual object and/or sound source on a measurement scale. In this case, the subjects were provided with the distance cues only. After presentation of the auditory and/or visual stimuli, subjects were asked to report their egocentric distance judgments by using a keypad (with one digit after the decimal point). Subjects were limited in time (12 s) to enter their judgments. This time-window has been employed to force the subjects to provide spontaneous distance values.

All participants were first provided with a written description of the experimental task. After reading the instructions, an experimenter presented an equivalent verbal description of the task. During the experiment, the written instructions (related to the distance judgments and pauses) were displayed on the screen. The experiment consisted of two sessions:

1. During the first session the subjects judged the auditory- and visual-only conditions. The conditions were assessed in blocks, half of the subjects starting with the auditory block and half with the visual block. During the auditory block, the screen displayed a uniform gray similar to the wall color.

2. During the second session the subjects judged the three auditory-visual blocks. The last block included the 8 spatially disparate auditory and visual cues conditions.

The two sessions were separated in time by at least 36 hours. The participants had 8 and 4 training trials in the first and the second sessions, respectively, including the nearest (2 m) and most distant (20 m) conditions. The data from the training trials were discarded in the following analysis. In each session, all combinations of target distance and reverberation and/or amount of visual information were presented in random order, with four repetitions per condition. Since the visual and auditory rendering were not modified according to the position of the subjects, they were asked not to move their head during the test.

A total of 24 subjects participated in the experiment. They were recruited from the European Center for Virtual Reality (mainly students and researchers). Participants were naive with respect to the purpose of the experiment. They had normal or corrected to normal vision and reported no auditory impairments. Except for 15 trials out of 4 608, the subjects were always able to provide a score within the 12s time-window.

3. RESULTS

The relative error E between the perceived distance ρ_{per} and the real distance of the target ρ_{tar} were calculated using the following equation:

$$E = (\rho_{per} - \rho_{tar}) / \rho_{tar} \quad \text{(Eq. 1)}$$

Then, repeated measure analyses of variance were realized. The Greenhouse-Geisser correction was applied when the Mauchly test showed that the sphericity assumption was violated.

During the first session (unimodal conditions), the presentation order of the two blocks, auditory- and visual-only conditions, has been randomized: half of the subjects started with the auditory block and half with the visual block. So the block order was analyzed as between-subjects factors in the following analyses. Neither simple effect nor interaction involving this factor was found, whatever the condition.

In this study the relative error appeared to be almost always negative (i.e. the distance was be under-estimated by subjects), and in the rest of the paper « a large absolute value of the relative error, with a negative magnitude » will be simply denoted « large relative error ».

3.1 Auditory-only condition

Figure 3 indicates the relative error E as a function of the target distance, for the auditory-only condition, for each of the two reverberation times (short: 370ms, long: 860ms).

The relative error was analyzed using repeated measures analysis of variance with target distance (5 levels), reverberation time (2 levels), and repetition (4 levels) as within-subjects factors and block order (2 levels) as between-subjects factors.

The analysis (see Table 2) revealed a significant effect of target distance ($F(4,92)=23.627$; $p<0.0001$ with Greenhouse-Geisser correction) : the more the target distance, the more the relative error, nevertheless the Bonferroni post-hoc test indicated that only the target distance 20m gave errors significantly higher than those with the other target distances ($p<0.0001$). The Anova revealed an effect of the reverberation time ($F(1,23)=16.077$; $p=0.001$ with Greenhouse-Geisser correction) : the error was lower when the reverberation time was long. Finally the Anova indicated a significant interaction between target distance and reverberation time ($F(4,92)=11.277$; $p<0.0001$ with Greenhouse-Geisser correction). The Bonferroni post-hoc showed that for short reverberation time, only a 20m target distance gave error higher to those with other target, but with a long reverberation

time, both 10m and 20m target distance gave errors different to those with other target distances ($p<0.016$ in the less significant case).

Overall, with a long reverberation time, the relative error was low for a target distance up to 5 m, and increase significantly above this distance. In contrary, with a short reverberation time, the relative error was always large, whatever the target distance (and even increased significantly for target distance above 10m). This finding agrees with Bronkhorst and Houtgast [57], who showed that perceived distance increases as the direct-to-reverberant energy ratio decreases.

Previous research indicates that, for the auditory modality, the distance to a sound source is overestimated when the source is located closer than 2 m, while it is substantial and progressively underestimated for greater distances [4]. Moreover Cabrera et al. [58, 59] have shown that auditory distance perception and auditory room-size perception were related. Judgments of apparent source distance made by participants seem to be more accurate in a reverberant than in an anechoic environment [12, 18, 60]. Our results agree with these studies.

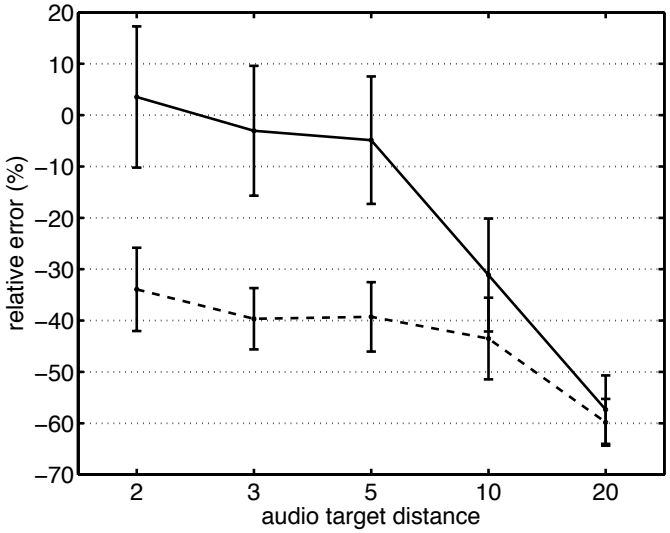

Figure 3. Relative errors (means and 95% confidence intervals) as a function of auditory target distance and reverberation time (solid line: long RT60 - 860 ms; dashed line : short RT60 - 370 ms).

Figure 4 shows the relationship between the target distance, ρ_{tar}, and the perceived egocentric distance, ρ_{per} (not the relative error) for the auditory-only conditions. The dotted line represents ideal performance. This figure confirms the large under-estimation of the target distances.

Many experiments on auditory distance perception yield a compressive function for the perceived versus actual distance [4]:

$$\rho_{per}' = k \cdot \rho_{tar}^{a} \quad \text{(Eq. 2)}$$

where ρ_{tar} is the physical source distance and ρ_{per}' an estimate of the perceived distance with fit parameters k and a.

By using the perceived and the target egocentric distance values, the k and a coefficients in Eq. 2 were estimated in a least-square sense. By using these two coefficients perceived egocentric distances are estimated and then compared to the subjects' judgments. For this purpose, the root mean square error, σ, was calculated for each reverberation time. The estimated coefficients, $0.96.\rho_{tar}^{0.72}$ $(\sigma = 0.29)$ for the short reverberation time and $1.83.\rho_{tar}^{0.53}$ $(\sigma = 0.40)$ for the long reverberation time, show a compression of the target distance ($a < 1$) and higher errors for the short RT_{60} (370 ms) than for the long RT_{60} (860 ms). Since Zahorik [44] obtained a similar compression (a = 0.66) with a real environment, the virtual auditory environment used in the experiment seems to provide enough auditory cues to simulate the position of a sound source in distance.

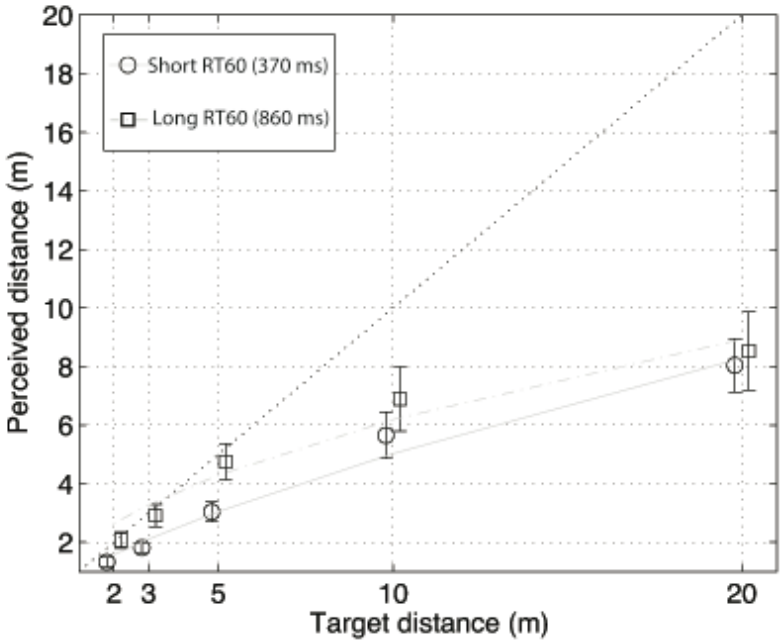

Figure 4. Relationship between the target distance, ρ_{tar}, and the perceived egocentric distance, ρ_{per} (means and 95% confidence intervals) for the auditory-only conditions.

One common problem with auditory displays simulated using an headphone is that auditory events are perceived within the head when compared against real environments where auditory events are

externalised. This impression is enhanced when the headphone is not equalized, when head movements are prohibited, and when the room effect is poor. If the room effect is enhanced, for example if the reverberation time arises, the "within the head" impression decrease. The fact that the under-estimation of distance was mainly observed when the reverberation time was short could let think that the "within the head" impression and the distance under-estimation are linked. However, several studies about distance perception of audio objects, with other protocols and methods, also showed that apparent source distance reported by participants seem to be more accurate in a reverberant than in an anechoic environment [12, 18, 60].

3.2 Visual-only condition

Figure 5 indicates the relative error as a function of the target distance, for the visual-only condition, for each of the two types of visual cues (few or many visual cues). This figure shows that whatever the degree of visual cues, there is a floor for target distances 2m and 3m, another floor for target distances 10m and 20m, and a large decrease between them. However, the distance was always under-estimated.

The relative error was analyzed using repeated measures analysis of variance with target distance (5 levels), amount of visual cues (2 levels: few or many visual cues), and repetition (4 levels) as within-subjects factors and block order as between-subjects factors.

The analysis (see Table 2) revealed a significant effect of target distance ($F(4,92)=6.121$; $p=0.006$ with Greenhouse-Geisser correction), and the Bonferroni post-hoc test indicated that only the target distance 20m gave errors significantly higher than those with the target distances 3m ($p=0.002$) and 5m ($p=0.047$).

The Anova showed no simple effect of the amount of visual cues, but indicated a significant interaction target distance / visual cues ($F(4,92)=2.92$; $p=0.039$ with Greenhouse-Geisser correction). The Bonferroni post-hoc showed that the relative error for target distances 10m and 20m was surprisingly larger when the visual cues were numerous (respectively $p=0.008$ and $p=0.01$).

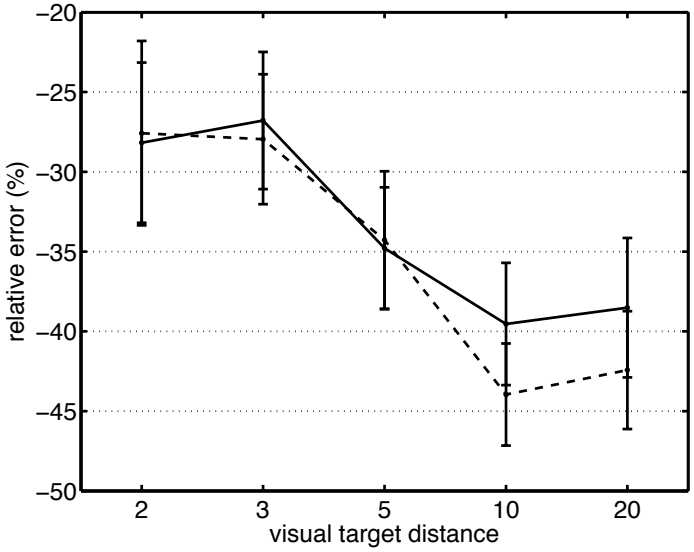

Figure 5. Relative errors (means and 95% confidence intervals) as a function of visual target distance and visual cues (solid line : few visual cues; dashed line : many visual cues).

Figure 6 shows the relationship between the target distance, ρ_{tar}, and the perceived egocentric distance, ρ_{per}, for the visual-only conditions. The dotted line represents ideal performance. The interpolation of Eq. 2 in a least-square sense estimated coefficients of $0.70.\rho_{tar}^{0.96}$ ($\sigma = 0.16$) and $0.71.\rho_{tar}^{0.93}$ ($\sigma = 0.24$) for the restricted and full visual cues, respectively (gray curves on figure). These values and figure 6 show that target distance is not really compressed ($a > 0.93$: quasi-linear) but still under-estimated ($k < 1$). The prediction error, σ, is reduced compared to the auditory modality and higher for the full visual cues than for the restricted visual cues. This latter result confirms that the distance perception is worse in case that numerous anchors are included in the visual environment.

The under-estimations obtained in this experiment show that the used visual environments do not accurately simulate real environments. A linear relationship between the perceived egocentric distance and the target distance of a real visual object has been observed by Loomis et al. [7]. But it has been reported that visual distance perception is quite accurate for targets up to 20 m away if full visual cue conditions are available to the observer [7, 61, 62, 63]. When the distant visual cues are restricted, visual perception of distance becomes less accurate [3]. In this study, the relationship between perceived and target distances were quasi-linear, but under-estimated. It is possible that even when the visual cues were considered as « many », these were still a too poor visual content.

Moreover, with head-mounted displays (HMD), perceived visual distances have been observed to be systematically under-estimated [64]. Virtual reality systems based on large immersive screens

were thought to offer a better distance perception [33]. But studies focusing on visual distance perception in virtual environments rendered by large immersive screens have found that visual distances were underestimated using these systems, exactly as in HMD systems [8, 65, 66, 67, 68]. Our results agree with that.

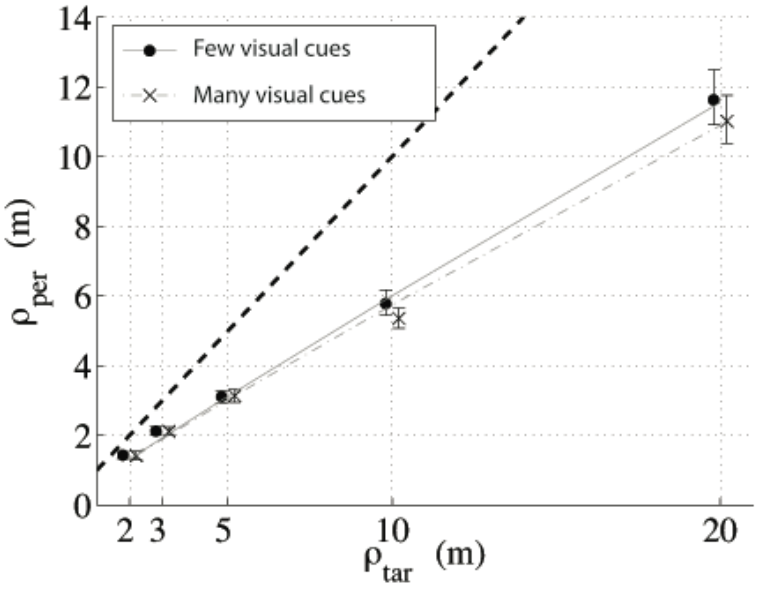

Figure 6: Relationship between the target distance, ρ_{tar}, and the perceived egocentric distance, ρ_{per} (means and 95% confidence intervals) for the visual-only conditions.

3.3. Bi-modal condition, coherent auditory and visual targets

The solid line of Figure 7 indicates the relative error as a function of the target distance for the bimodal condition (with coherent visual and auditory targets), all types of visual cues and all reverberation times confounded.

The relative error in bimodal condition was analyzed using repeated measures analysis of variance with target distance (5 levels), scenery (2 levels : few or many visual cues), reverberation time (2 levels : short or long), and repetition (4 levels) as within-subjects factors and block order as between-subjects factors.

The analysis (see Table 2) revealed a significant effect of target distance ($F(4,88)=5.478$; $p=0.011$ with Greenhouse-Geisser correction), and the Bonferroni post-hoc test indicated that the

target distance 3m gave errors significantly smaller than those with the target distances 5 (p =0.009), 10 (p=0.001) and 20 (p=0.038), and that the target distance 5m gave errors significantly smaller than those with the 10m target distance (p=0.007). The Anova did not highlight any other significant effect or interaction.

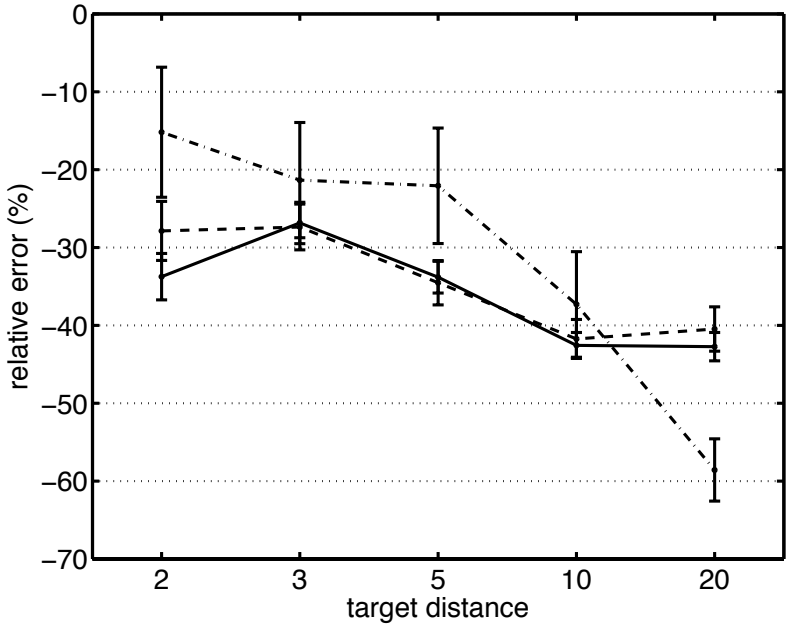

Figure 7. Relative errors (means and 95% confidence intervals) as a function of target distance for auditory-only (dashed-dotted line), visual-only (dotted line), and bimodal with coherent visual and auditory targets (solid line).

Table 2: Significant factors for modalities Auditory-only, Visual-only, and Bimodal (with coherent targets).

MODALITY	SIGNIFICANT FACTORS	F value	p value
AUDITORY-ONLY	Target distance	F(4,92)=23.627	p<0.0001
	Reverb. time	F(1,23)=16.077	p=0.001
	Target distance * Reverb. Time	F(4,92)=11.277	p<0.0001
VISUAL-ONLY	Target distance	F(4,92)=6.121	p=0.006
	Target distance * Visual cues	F(4,92)=2.92	p=0.039
BIMODAL - COHERENT TARGETS	Target distance	F(4,88)=5.478	p=0.011

Figure 8 shows the relationship between the target distance ρ_{tar} and the averaged perceived egocentric distance, ρ_{per}, for the bimodal conditions with coherent visual and auditory targets. Error

bars correspond to the 95% confidence intervals. The dotted line represents ideal performance. This figure shows similar results to the visual estimations: the target source distances are under-estimated for both visual environments and both reverberation time, and the variability is lower than for the auditory-only conditions.

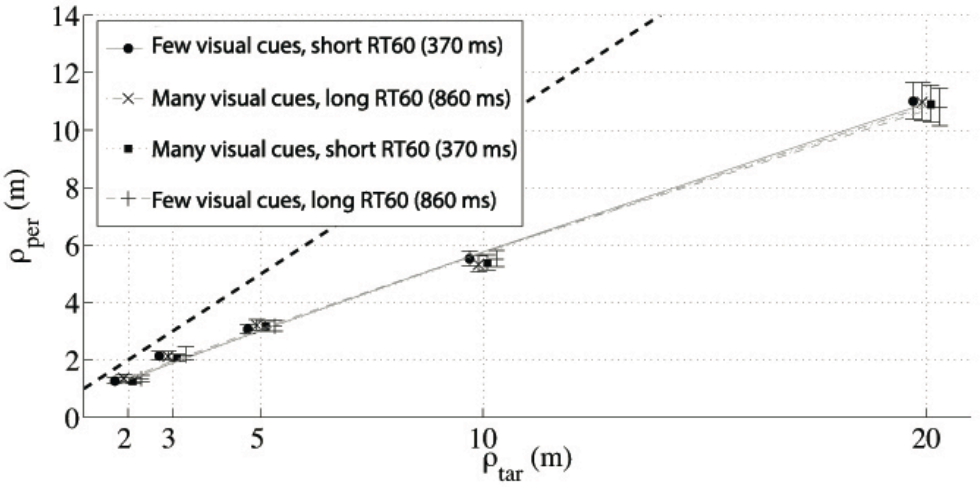

Figure 8: Relationship between the target distance, ρ_{tar}, and the perceived egocentric distance, ρ_{per} (means and 95% confidence intervals), for the bimodal conditions with coherent auditory and visual cues.

The interpolation of Eq. 2 in a least-square sense shows estimated coefficients of:

- $0.69.\rho_{tar}^{0.92}$ $(\sigma = 0.18)$ for the restricted visual cues and short reverberation time,
- $0.69.\rho_{tar}^{0.92}$ $(\sigma = 0.26)$ for the full visual cues and long reverberation time,
- $0.68.\rho_{tar}^{0.92}$ $(\sigma = 0.21)$ for the full visual cues and long reverberation time,
- $0.73.\rho_{tar}^{0.89}$ $(\sigma = 0.17)$ for the restricted visual cues and short reverberation time.

The coefficients show a quasi-linear relationship between the perceived and the target egocentric distances (gray curves on figure).

3.4. Comparison between auditory-only, visual-only, and (coherent) bimodal condition

3.4.1. Auditory-only vs. Visual only

The comparison between figures 3 and 5 indicated that the relationship between the target and perceived distances is different between visual-only and auditory-only cases : in auditory case, the relative error was very dependent of the reverberation time : with long time reverberation, the error went from zero (for close targets) to -60% errors (for far targets). For short time reveb, the error went from -35% to -60%. In visual case, whatever the decor, the errors were lower and ranged between -27% and -43%. Moreover the variance was lower for visual case than for auditory case (especially with long reverberation time).

It is generally agreed that the distance to an object is perceived more accurately and with less variability in the visual than in the auditory modality [7, 39, 41]. In the present experiment, the under-estimation of visual target distances is certainly due to the virtual-reality system [8, 65, 66, 67, 68].

3.4.2. Coherent Bimodal condition vs. Auditory-only and Visual only

When looking at the figure 7, all reverberation times and all types of visual cues confounded, the variances in bimodal condition seems to be very close to those in the visual-only condition. The figure 9 indicates relative errors (means and 95% confidence intervals) as a function of target distance for auditory-only (dotted black line), visual-only (solid black line), and bimodal with coherent visual and auditory targets (grey solid line). The four sub-figures indicate the different conditions (short or long reverberation depending on the first or the second line ; few or many visual cues depending on the first or the second column). The two dotted black lines are identical for the two sub-figures on a same line (short or long reverberation time), and the two solid black lines are identical for the two sub-figures on a same colums (few or many visual cues). On the contrary, the grey line, indicating the bimodal condition, is different in each sub-figure. For example, compare the relative errors between auditory-only condition with short reverberation and visual-only condition with few visual cues is not of interest, but the comparison of the bimodal condition with short reverberation and few visual cues with each of these unimodal conditions could indicate if the bimodal condition is closer to auditory- or visual-only condition.

The grey line is always very close to the solid black line (visual only condition). For the long reverberation condition, the relative error for auditory-only is low for close targets, and large for 20m target, but the bimodal shape follows the visual-only one.

For each combination of reverberation time and amount of visual cues, the relative error was analyzed using repeated measures analysis of variance with modality (3 levels: visual only, auditory only, or bimodal coherent), target distance (5 levels), and repetition (4 levels) as within-subjects factors.

Short reverberation time, few or many visual cues: These two conditions gave quite similar results: no difference was oberved between visual-only, auditory-only, and bimodal conditions. However the Anova highlighted two exceptions (see Table 3) showed by the interaction between target distance and modality ($F(8,184)=2.837$; $p=0.039$ with few visual cues, and $F(8,184)=2.78$; $p=0.047$ with many visual cues): at the target distance 20m, auditory-only condition gave larger errors than bimodal conditions ($p<0.0001$ according to Bonferroni post-hoc test), and at the target distance 3m and only with many visual cues, auditory-only condition gave larger errors than bimodal conditions ($p=0.019$ according to Bonferroni post-hoc test).

Long reverberation time, few or many visual cues: For the case with few visual cues, the modalities (visual-only, auditory-only, and bimodal) were not statistically different. For the case many visual cues, only modalities visual-only and auditory-only gave significant different relative errors ($F(8,184)=15.074$; $p<0.0001$ according to the Anova, and $p=0.047$ according to Bonferroni post-hoc).

Whatever the amount of visual cues (see Table 3), the interaction between target distance and modality highlighted by the Anova ($F(8,184)=16.137$; $p<0.0001$ with few visual cues, and $F(8,184)=15.074$; $p<0.0001$ with many visual cues) indicated three stages :

- for target distances 2m, 3m , and 5m, auditory-only condition gave errors lower than those for bimodal conditions ($p<0.015$ for the less significant case according to Bonferroni post-hoc).

- for target distance 10m, auditory-only curve is crossing the other curves, and no significant difference were oberved.

- for target distances 20m, auditory-only condition gave errors higher than those for bimodal conditions ($p=0.028$ according to Bonferroni post-hoc).

In sum, with long reverberation time, bimodal and visual-only modalities gave statistically similar relative errors whatever the target distance, and except for the target distance 10m, bimodal and auditory-only modalities gave always different relative errors. So our results suggest that distance estimation accuracy in bimodal coherent condition is completely determined by visual cues, and that the addition of auditory cues has no influence.

Moreover both reverberation times provide the same bimodal distance estimations whereas reverberation time has an influence on the auditory-only conditions. The restricted influence of

reverberation time on the bimodal conditions can be considered as a visual capture effect of the perceived auditory distance.

Zahorik [43] showed that the presence of visual cues increased the accuracy of auditory distance perception and lowered the variability of the judgments. However, for distances greater than 2 m listeners still significantly underestimated the distance of the sound source : for distances of 4 and 5 m from the source they never exceeded 3 m. These data give respectively relative errors (4-3)/4=25% and (5-3)/5=40%. When looking at the figure 9, the grey solid lines (bimodal coherent targets) indicate similar values. Thus results of the present study in bimodal condition in virtual environment are in agreement with Zahorik results [43] in real environment. Thus, using virtual visual environments makes participants underestimate distances (§ 3.2.), but this did not affect the auditory-visual interaction.

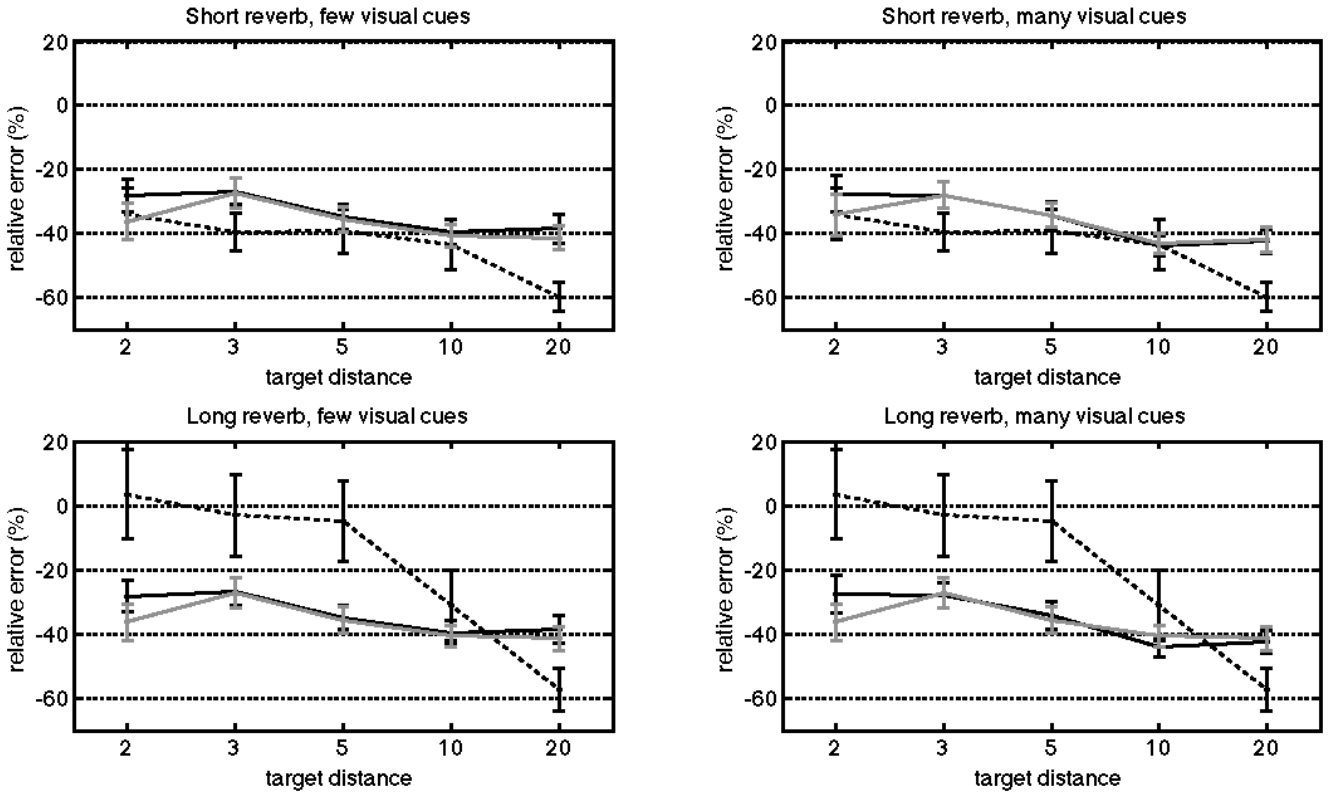

Figure 9. Relative errors (means and 95% confidence intervals) as a function of target distance for auditory-only (dotted black line), visual-only (solid black line), and bimodal with coherent visual and auditory targets (grey solid line). The four sub-figures indicate the different conditions: short (RT_{60}=370 ms) or long (RT_{60}=860 ms) reverberation depending on the first or the second line; few or many visual cues depending on the first or the second column.

Table 3: Significant factors when comparing Coherent Bimodal condition vs. Auditory-only and Visual only.

RT_{60}	VISUAL CUES	SIGNIFICANT FACTORS	F value	p value
SHORT (370 ms)	FEW CUES	Target distance * Modality	F(8,184)=2.837	p=0.039
	MANY CUES	Target distance * Modality	F(8,184)=2.78	p=0.047
LONG (860 ms)	MANY CUES	Target distance	F(8,184)=15.074	p<0.0001
	FEW CUES	Target distance * Modality	F(8,184)=16.137	p<0.0001
	MANY CUES	Target distance * Modality	F(8,184)=15.074	p<0.0001

3.5. Bi-modal condition, incoherent auditory and visual targets

The test conditions presented in section 3.1 include eight bimodal conditions with spatially disparate visual and auditory distances. These are defined by the target distances of the visual stimulus $\rho_{tar,V}$, and the auditory stimulus $\rho_{tar,A}$, see the first column of Table 4. For these eight conditions the stimuli were placed in the visual environment with many visual cues and the room effect corresponding to the short reverberation time RT_{60} = 370 ms has been used. Concerning visual-only, auditory-only, and bimodal with coherent cues conditions, the results were expressed as relative error (in percent). This relative error was based on the target distance (Eq. 1). In bimodal condition with different visual and auditory targets, the calculation of relative error would not make sense. So in this paragraph the results will be simply expressed as perceived distance (in meters).

3.5.1. Comparison between bimodal incoherent condition [$\rho_{tar,V}=i$, $\rho_{tar,A}=j$] and visual-only [$\rho_{tar,V}=i$] and auditory-only [$\rho_{tar,A}=j$] conditions.

On the figure 10, the crosses with errorbars (respectively means and 95% confidence intervals when considering all subjects and repetitions) indicate the perceived distances for the bimodal incoherent condition with visual target distance Vi (or $\rho_{tar,V}=i$) and auditory target distance Aj (or $\rho_{tar,A}=j$). « V5 A20 » means visual target at 5 meters and auditory target at 20 meters. The circles with errorbars (means and 95% confidence intervals) indicate the perceived distances for the visual-only condition with visual target distance Vi (i.e. the visual target in the incoherent condition). The circles without errorbar indicate the visual target distances themselve (for example 2 meters for the target V2). The squares with errorbars (means and 95% confidence intervals) indicate the perceived distances for the auditory-only condition with auditory target distance Aj (i.e. the distance of the

auditory target in the incoherent condition). The auditory-only conditions A1 and A15 have no corresponding perceived distance because the auditory targets at 1m and 15m were not tested. The squares without errorbar indicate the auditory target distances themselve (for example 20 meters for the target A20).

Table 4: Bimodal incoherent visual and auditory targets (column 1), which were compared to visual-only condition with the same visual target (column 2), auditory-only condition with the same auditory target (column 3), bimodal coherent conditions with the same visual (column 4) or auditory target (column 5). The asterisk marks indicate that for the concerned targets, the perceived distance was significantly not different from the one in the correspondant bimodal incoherent condition (cell on the first column on the same row).

INCOHERENT BIMODAL CONDITION	**VISUAL-ONLY CONDITION**	**AUDITORY-ONLY CONDITION**	**COHERENT BIMODAL CONDITION with target = visual target of the incoherent condition.**	**COHERENT BIMODAL CONDITION with target = auditive target of the incoherent condition.**
V2 A1 Visual target: 2m Auditory target: 1m	V2 Visual target: 2m	Target 1m non tested	V2 A2 * Visual target: 2m Auditory target: 2m	Target 1m non tested
V5 A20 Visual target: 5m Auditory target: 20m	V5 * Visual target: 5m Auditory target: 20m	A20 Auditory target: 20m	V5 A5 * Visual target: 5m Auditory target: 5m	V20 A20 Visual target: 20m Auditory target: 20m
V5 A1 Visual target: 5m Auditory target: 1m	V5 Visual target: 5m	Target 1m non tested	V5 A5 Visual target: 5m Auditory target: 5m	Target 1m non tested
V5 A10 Visual target: 5m Auditory target: 10m	V5 * Visual target: 5m	A10 Auditory target: 10m	V5 A5 * Visual target: 5m Auditory target: 5m	V10 A10 Visual target: 10m Auditory target: 10m
V10 A20 Visual target: 10m Auditory target: 20m	V10 * Visual target: 10m	A20 Auditory target: 20m	V10 A10 * Visual target: 10m Auditory target: 10m	V20 A20 Visual target: 20m Auditory target: 20m
V10 A3 Visual target: 10m Auditory target: 3m	V10 Visual target: 10m	A3 Auditory target: 3m	V10 A10 * Visual target: 10m Auditory target: 10m	V3 A3 Visual target: 3m Auditory target: 3m
V10 A15 Visual target: 10m Auditory target: 15m	V10 * Visual target: 10m	Target 15 m non tested	V10 A10 * Visual target: 10m Auditory target: 10m	Target 15m non tested
V20 A5 Visual target: 20m Auditory target: 5m	V20 Visual target: 20m	A5 Auditory target: 5m	V20 A20 * Visual target: 20m Auditory target: 20m	V5 A5 Visual target: 5m Auditory target: 5m

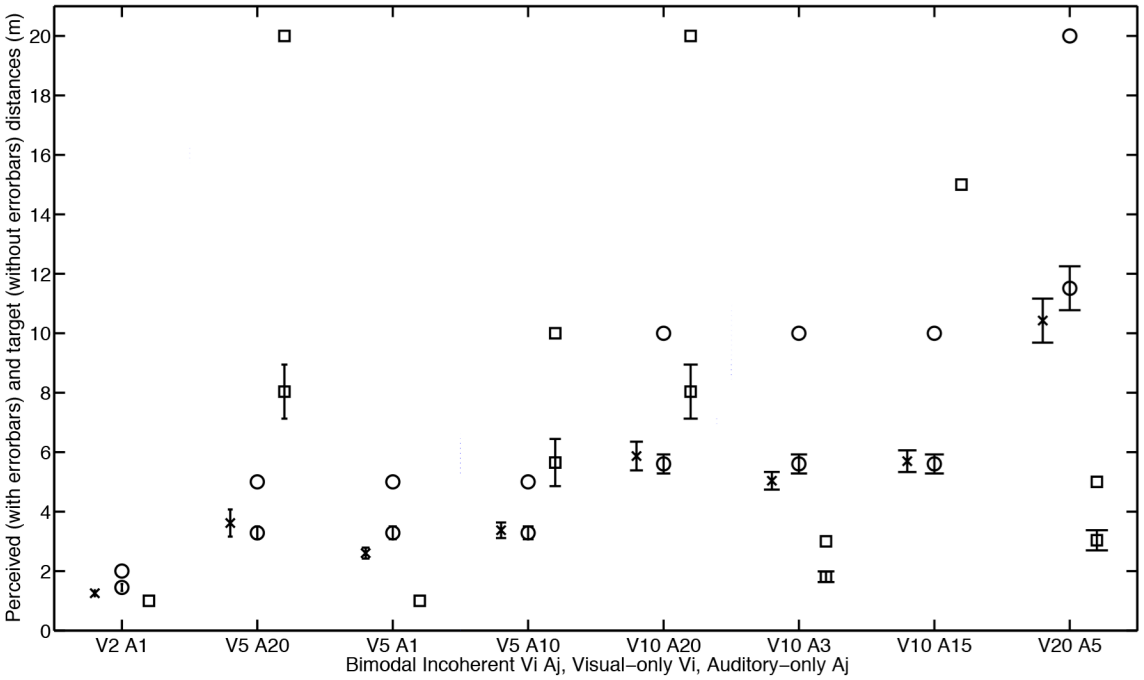

Figure 10. Perceived distance (means and 95% confidence intervals among listeners and repetitions) for bimodal condition Vi Aj with incoherent visual and auditory targets (crosses with errorbars), visual–only condition Vi with the same visual target (circles with errorbars) and auditory-only condition Aj with the same auditory target (squares with errorbars). The circles and squares without errorbars indicate respectively the visual and auditory target distances themselve.

The examination of figure 10 shows that the incoherent condition is always closer to the visual-only condition with same visual target than to the auditory-only condition with the same auditory target. The Bonferroni post-hoc test associated to the one-way Anova indicated that the bimodal incoherent conditions gave perceived distances systematically different to the ones given by the auditory-only modality with the same auditory targets. Bimodal conditions V2 A1, V5 A1, V10 A3, and V20 A5 gave also perceived distances significantly different to the ones given by the visual-only modality with the same visual targets. On the contrary, bimodal conditions V5 A20, V5 A10, V10 A20 and V10 A15 gave statistically similar perceived distances than visual-only modality with the same visual targets. These similar results are indicated by asterisk marks in table 4. The bimodal conditions that are equivalent to the visual-only ones are those with a visual target closer to the subject than the auditory target. In this case, the visual attraction was complete and the attraction by the auditory target was nonexistent. On the contrary, when auditory target was closer to the subject than the visual target, the auditory target slightly attracted the percept, in consequence the perceived distance in bimodal condition was placed between the ones for visual-only and auditory-only conditions (and was statistically different to them).

3.5.2. Comparison between bimodal incoherent condition [$\rho_{tar,V}$=i, $\rho_{tar,A}$=j] and bimodal coherent condition [$\rho_{tar,V}$=i, $\rho_{tar,A}$=i] or [$\rho_{tar,V}$=j, $\rho_{tar,A}$=j]

On the figure 11, the crosses with errorbars indicate the perceived distances for the bimodal incoherent condition with visual target distance Vi (or $\rho_{tar,V}$=*i*) and auditory target distance Aj (or $\rho_{tar,A}$=*j*). « V5 A20 » (the second bimodal incoherent condition) means visual target at 5 meters and auditory target at 20 meters. The circles with errorbars indicate the perceived distances for the bimodal coherent condition with visual target Vi and auditory target distance Ai (i.e. the distance of the visual target in the incoherent condition, for example V5 A5). The squares with errorbars indicate the perceived distances for the bimodal coherent condition with visual target Vj and auditory target distance Aj (i.e. the distance of the auditory target in the incoherent condition, for example V20A20). The bimodal coherent conditions V1 A1 and V15 A15 have no corresponding perceived distance because the auditory targets at 1m and 15m were not tested. As in the figure 10, the circles and squares without errorbar indicate respectively the visual and auditory target distances themselve.

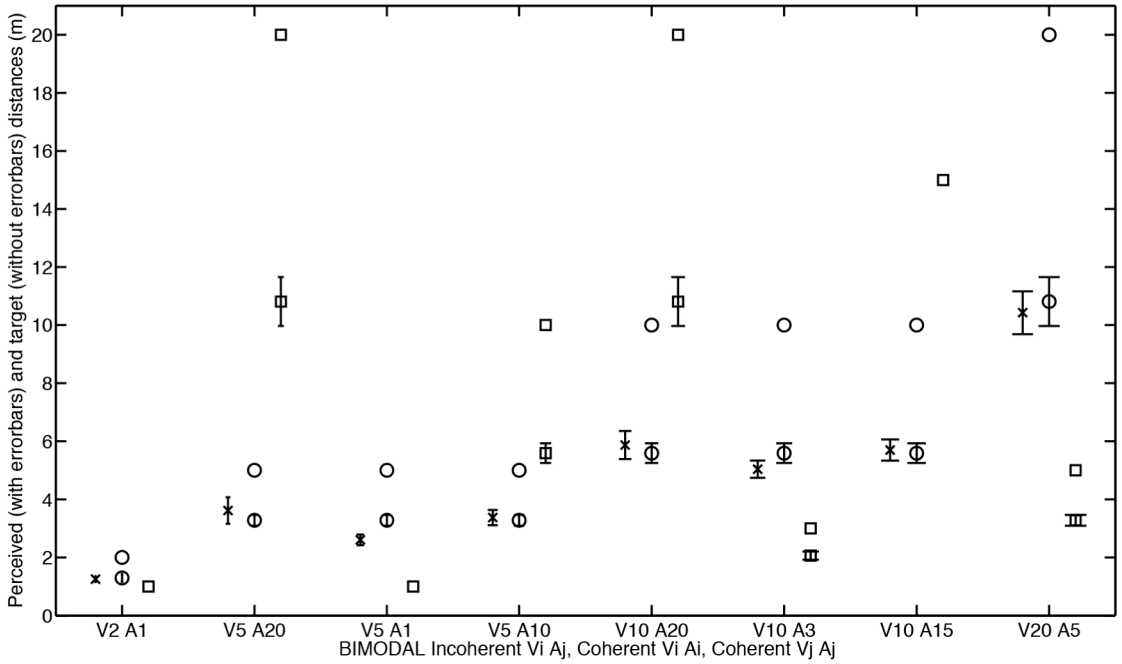

Figure 11. Perceived distance (means and 95% confidence intervals among listeners and repetitions) for bimodal condition Vi Aj with incoherent visual and auditory targets (crosses with errorbars), bimodal coherent condition Vi Ai with the same visual target (circles with errorbars) and bimodal coherent condition Vj Aj with the same auditory target (squares with errorbars). The circles and squares without errorbars indicate respectively the visual and auditory target distances themselve.

The examination of figure 11 shows that the incoherent condition is always closer to the coherent condition with same visual target than to the coherent condition with the same auditory target. The Bonferroni post-hoc test associated to the Anova indicated that the bimodal incoherent conditions gave perceived distances systematically different to the ones given by the bimodal coherent conditions with the same auditory targets. On the contrary, bimodal incoherent conditions (except for the V5 A1 one) gave statistically similar perceived distances than bimodal coherent conditions with the same visual targets. This finding means that for most targets, the perceived distance for bimodal incoherent condition Vi Aj is the same than the one for bimodal coherent condition Vi Ai (with the same visual target).

This result globally agrees with studies (realized in real life and not specific to distance perception) describing the ventriloquism effect as a capture of the auditory signal by the visual signal [38, 39, 40]. Nevertheless, in real life, the visual-capture effect would have given perceptual distance with quite low under-estimations. Here, the bimodal incoherent condition gave a systematical large under-estimation. These under-estimations were also obtained in visual-only condition and in bimodal coherent condition. So the visual capture is effective, and the under-estimation is certainly due to the virtual reality system [8, 65, 66, 67, 68].

3.5.3. Influence of the offset between auditory and visual targets

In 3.5.1., the nature (visual vs. auditory) of the closer target had an influence on the visual capture of the incoherent bimodal stimulus : the visual capture was effective only when visual target was closer to the subject than the auditory target. So for each couple of targets the offset $\Delta\rho$ between the auditory and visual target distances was calculated:

$$\Delta\rho = \rho_{tar,A} - \rho_{tar,V} \qquad \text{(Eq. 3)}$$

Then the difference between perceived distance in bimodal incoherent condition Vi Aj and visual-only condition with the same visual target Vi (this distance represents representing the visual capture) was also calculated. The correlation between this difference and the offset was low: $r(768)=0.264$; $p<0.0001$ according to Pearson test. This lack of offset effect could mean that the difference between auditory and visual target distances is not of sense without considering the distances themselve.

3.5.4. A model of distance perception in bimodal incoherent condition

Alais and Burr [41], in a study about azimuth, explained that our ability to make use of visual cues to localize stimuli typically leads to less variability than our ability of using auditory cues, and proposed a model of optimal combination of visual and auditory spatial cues, where each modality is weighted by an inverse estimate of its variability :

$$S_{AV} = w_V S_V + w_A S_A \quad \text{(Eq. 4)}$$

where S_{AV} is the estimation of the percept in bimodal incoherent condition, S_A the percept in auditory-only condition, S_V the percept in visual-only condition, and w_V and w_A are the relative weights for each modality, inversely proportional to their variances:

$$w_A = \frac{1/\sigma_V^2}{1/\sigma_A^2+1/\sigma_V^2} = \frac{\sigma_V^2}{\sigma_A^2+\sigma_V^2} \quad \text{(Eq. 5)}$$

We applied this model to the perceived distances obtained in visual-only and auditory-only conditions, and calculated an estimation of perceived distance for each bimodal incoherent condition. Then we observed the correlation between the estimations and the experimental perceived distances in bimodal incoherent condition. This correlation is substantial (but not really high): $r(120)=0.709$; $p<0.0001$. Anyway it confirms the consistency for the distance estimation of the combination of visual and auditory cues weighted by an inverse estimate of their variability. It is worth noting to the consistence of the prediction was higher in the study of Alais and Burr [41], but it concerned the ventriloquism effect in azimuth (not in distance), moreover authors used six subjects with numerous trials whereas we used 24 subjects with only 4 repetitions.

4. CONCLUSION

This study used virtual environments in order to analyze the egocentric distance perception for targets placed between 2 and 20 meters. It focused on the comparison of uni-modal (visual or auditory) and bimodal (visual and auditory) coherent and non-coherent conditions.

The compression of distances for auditory-only condition agreed with previous studies realized in real life. The results for visual-only condition showed an under-estimation of distances (without compression). This under-estimation is larger than those obtained in real life studies, but was already higlighted in studies in virtual environments.

With long reverberation time, the “bimodal coherent conditions” (with coherent visual and auditory targets) and visual-only conditions gave statistically similar relative errors whatever the target distance, and except for the target distance 10m, bimodal coherent condition and auditory-only modalities gave always different relative errors. So our results suggest that distance estimation accuracy in bimodal coherent condition is completely determined by visual cues, and that the addition of auditory cues has no influence.

The « bimodal incoherent conditions » (with incoherent auditory and visual targets) gave perceived distances systematically different to the ones given by the « bimodal coherent conditions » (with coherent auditory and visual targets) with the same auditory targets. On the contrary, bimodal incoherent conditions gave similar perceived distances than bimodal coherent conditions with the same visual targets.

The bimodal incoherent conditions gave perceived distances systematically different to the ones given by the auditory-only modality with the same auditory targets. The bimodal incoherent conditions gave perceived distances equivalent to the visual-only ones when the visual target was closer to the subject than the auditory target. In this case, the visual attraction was complete and the attraction by the auditory target was nonexistent. On the contrary, when auditory target was closer to the subject than the visual target, the auditory target slightly attracted the percept, in consequence the perceived distance in bimodal incoherent condition was placed between the ones for visual-only and auditory-only conditions (and was statistically different to them).

Globally the results showed a higher influence of visual information than auditory information on the perceived distance.

ACKNOWLEDGEMENT

The authors would like to thank all test subjects. This work is partly supported by the French National Research Agency (ANR) as a part of the EDISON 3D project (ANR-13-CORD-0008-02) and by the Finistère General Council (29), France.

REFERENCES

[1] J. Cutting and P. Vishton. Perception of Space and Motion, chapter Perceiving Layout and Knowing Distances: The Integration, Relative Potency, and Contextual Use of Different Infor- mation about Depth, pages 69–117. Academic Press, New-York, USA, 1995.

[2] J. Da Silva. Scales for Perceived Egocentric Distance in a Large Open Field: Comparison of Three Psychophysical Methods. The Am. J. of Psychology, 98(1):119–144, 1985.

[3] J. Philbeck and J. Loomis. Comparison of two indicators of perceived egocentric distance under full-cue and reduced-cue conditions. Journal of Experimental Psychology: Human Perception and Performance, 23(1):72–85, 1997.

[4] P. Zahorik, D. Brungart, and A. Bronkhorst. Auditory Distance Perception in Humans: A Summary of Past and Present Research. Acta Acust. united with Ac., 91(3):409–420, 2005.

[5] A. Ghazanfar, J. Neuhoff, and N. Logothetis. Auditory Looming Perception in Rhesus Monkeys. Proc. of the National Academy of Sciences of the United States of America, 99(24):15755– 15757, November 2002.

[6] B. Shelton and C. Searle. The Influence of Vision on the Absolute Identification of Sound-Source Position. Perception & Psychophysics, 28(6): 589–596, 1980.

[7] J. M. Loomis, R. L. Klatzky, J. W. Philbeck, and R. G. Golledge. Assessing Auditory Distance Perception Using Perceptually Directed Action. Perception & Psychophysics, 60(6):966–980, 1998.

[8] E. Klein, J. Swan, G. Schmidt, M. Livingston, and O. Staadt. Measurement protocols for medium-field distance perception in large-screen immersive displays. In IEEE Virtual Reality Conference, VR'2009, pages 107–113, Lousiana, USA, March 14–18 2009.

[9] T. Z. Strybel and D. R. Perrott. Discrimination of relative distance in the auditory modality: The success and failure of the loudness discrimination hypothesis. The Journal of the Acoustical Society of America, 76(1):318–320, 1984.

[10] P. Coleman. An Analysis of Cues to Auditory Depth Perception in Free Space. Psychological Bulletin, 60(3):302–315, 1963.

[11] E. Zwicker and H. Fastl. Psychoacoustics: Facts and models. Springer, DE–Berlin, 1st edition, 1990.

[12] D. Mershon and L. King. Intensity and Reverberation as Factors in the Auditory Perception of Egocentric Distance. Perception & Psychophysics, 18:409–415, 1975.

[13] R. Butler, E. Levy, and W. Neff. Apparent Distance of Sounds Recorded in Echoic and Anechoic Chambers. Journal of Experimental Psychology: Human Perception and Perfor- mance, 6(4):745–750, 1980. doi: 10.1037/0096- 1523.6.4.745.

[14] T. Lokki, M. Grohn, L. Savioja, and T. Takala. A Case Study of Auditory Navigation in Virtual Acoustic Environments. In Proc. of the Int. Conf. on Auditory Display (ICAD). Citeseer, 2000.

[15] J. Blauert. Spatial hearing: the psychophysics of human sound localization. The MIT Press, 1997.

[16] P. Coleman. Dual Rôle of Frequency Spectrum in Determination of Auditory Distance. The Journal of the Acoustical Society of America, 44 (2):631–632, 1968.

[17] D. Brungart, N. Durlach, and W. Rabinowitz. Auditory Localization of Nearby Sources. II. Localization of a Broadband Source. The Journal of the Acoustical Society of America, 106(4): 1956–1968, Oct. 1999.

[18] B. Shinn-Cunningham. Distance Cues for Virtual Auditory Space. In Proc. of the IEEE Pacific-Rim Conf. on Multimedia, pages 227– 230, Sydney, Australia, 2000. Citeseer.

[19] Z. Zhou, A. Cheok, X. Yang, and Y. Qiu. An Experimental Study on the Role of Software Synthesized 3D Sound in Augmented Reality Environments. Interacting with Computers, 16(5): 989–1016, 2004. ISSN 0953-5438.

[20] V. Pulkki. Virtual Sound Source Positioning Us- ing Vector Base Amplitude Panning. J. of the Aud. Eng. Soc., 45(6):456–466, 1997.

[21] A. Berkhout, D. de Vries, and P. Vogel. Acoustic Control by Wave Field Synthesis. The J. of the Acoustical Society of America, 93(5):2764–2778, 1993.

[22] D. Begault. Perceptual Effects of Synthetic Reverberation on Three-Dimensional Audio Systems. J. of the Aud. Eng. Soc., 40:895–904, 1992. ISSN 0004-7554.

[23] A. Bronkhorst. Localization of Real and Virtual Sound Sources. The J. of the Acous. Soc. of Am., 98(5):2542–2553, Nov. 1995.

[24] D. Begault, E. Wenzel, and M. Anderson. Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source. J. of the Aud. Eng. Soc., 49(10):904–916, 2001.

[25] N. Kopco, S. Santarelli, V. Best, and B. Shinn-Cunningham. Simulating Distance Cues in Virtual Reverberant Environments. In 19th INTERNATIONAL CONGRESS ON ACOUSTICS, MADRID, 2-7 SEPTEMBER 2007.

[26] M. Matlin and H. Foley. Sensation and Perception. Allyn and Bacon Boston, MA, Needhman Heights, Mass., 4th edition, 1997.

[27] R. Patterson. Human Factors of 3-D Displays. J. of the Society for Information Display, 15(11): 861–871, 2007.

[28] S. Palmisano, B. Gillam, D. Govan, R. Allison, and J. Harris. Stereoscopic perception of real depths at large distances. Journal of Vision, 10 (6):1–16, 2010.

[29] J. Wann, S. Rushton, and M. Mon-Williams. Natural Problems for Stereoscopic Depth Perception in Virtual Environments. Vision Research, 35(19):2731–2736, 1995.

[30] P. Willemsen and A. Gooch. Perceived Egocentric Distances in Real, Image-Based, and Traditional Virtual Environments. In Proc of the IEEE Virtual Reality Conf., pages 275–276, 2002.

[31] V. Interrante, B. Ries, and A. L. Distance Perception in Immersive Virtual Environments, Revisited. In IEEE Virtual Reality Conference (VR 2006), pages 3–10, 25–29 March 2006.

[32] A. Murgia and P. Sharkey. Estimation of Distances in Virtual Environments Using Size Constancy. The International J. of Virtual Reality, 1(8):67–74, 2009.

[33] J. Plumert, J. Kearney, J. Cremer, and K. Recker. Distance Perception in Real and Virtual Environments. ACM Transactions on Ap- plied Perception (TAP), 2(3):216–233, 2005.

[34] C. Ziemer, J. Plumert, J. Cremer, and J. Kearney. Making distance judgments in real and virtual environments: Does order make a difference? In Proc. of the 3rd Symposium on Applied Perception in Graphics and Visualization, pages 153–153. ACM, 2006.

[35] M. Landy, L. Maloney, E. Johnston, and M. Young. Measurement and Modeling of Depth Cue Combination: In Defense of Weak Fusion. Vision Research, 35(3):389–412, 1995.

[36] C. Spence. Audiovisual multisensory integration. Acoustical science and technology, 28(2): 61–70, 2007.

[37] J. Lewald, W. Ehrenstein, and R. Guski. Spatio-Temporal Constraints for Auditory-Visual Integration. Behavioural Brain Research, 121(1–2): 69–79, 2001.

[38] Pick H L, Warren D H, Hay J C, Sensory conflict in judgements of spatial direction, Perception & Psychophysics, 6, 203-205, 1969.

[39] Bertelson P, Radeau M, Cross-modal bias and perceptual fusion with auditory-visual spatial discordance, Perception & Psychophysics 29, 578-584, 1981.

[40] Warren D. H., Welch R. B., McCarthy T. J., The role of visual auditory « compellingness » in the ventriloquism effect: implications for transitivity among the spatial senses'' Perception & Psychophysics, 30, 557-564, 1981.

[41] Alais D., Burr D.,The ventriloquist effect results from near-optimal bimodal integration, Current Biology, 14, 257-262, 2004.

[42] Mershon D. H., Desaulniers D. H., Amerson T. L., Kiefer S. A., Visual capture in auditory distance perception: proximity image effect reconsidered, Journal of Auditory Research, 20, 129-136,1980.

[43] Gardner, M. B. (1968). Proximity image effect in sound localization. The Journal of the Acoustical Society of America, 43(1), 163-163.

[44] P. Zahorik. Estimating Sound Source Distance With and Without Vision. Optometry & Vision Science, 78(5):270–275, 2001.

[45] Hládek, Ľ., Le Dantec, C. C., Kopčo, N., & Seitz, A, Ventriloquism effect and aftereffect in the distance dimension, Proceedings of Meetings on Acoustics. Vol. 19. No. 1. Acoustical Society of America, 2013.

[46] Iwaya, Y., Suzuki, Y., & Kimura, D. (2003). Effects of head movement on front-back error in sound localization. Acoustical science and technology, 24(5), 322-324.

[47] Simpson, W. E., & Stanton, L. D. (1973). Head movement does not facilitate perception of the distance of a source of sound. The American journal of psychology, 151-159.

[48] Watson, D. G., & Humphreys, G. W. (1999). Segmentation on the basis of linear and local rotational motion: Motion grouping in visual search. Journal of Experimental Psychology: Human Perception and Performance, 25(1), 70.

[49] Ono, M. E., Rivest, J., & Ono, H. (1986). Depth perception as a function of motion parallax and absolute-distance information. Journal of Experimental Psychology: Human Perception and Performance, 12(3), 331.

[50] Medendorp, W. P., Tweed, D. B., & Crawford, J. D. (2003). Motion parallax is computed in the updating of human spatial memory. The Journal of neuroscience, 23(22), 8135-8142.

[51] P. Willemsen, A. Gooch, W. Thompson, and S. Creem-Regehr. Effects of Stereo Viewing Conditions on Distance Perception in Virtual Environments. Presence: Teleoperators & Virtual Environments, 17(1):91–101, 2008.

[52] ARéVi. Atelier de Réalité Virtuelle (Virtual Reality Toolkit), Retrieved 31 March 2011. URL http://svn.cerv.fr/trac/AReVi.

[53] P. Zahorik. Auditory Display of Sound Source Distance. In Proc. Int. Conf. on Auditory Display, pages 326–332. Citeseer, 2002.

[54] D. Campbell, K. Palomaki, and G. Brown. Roomsim, a Matlab Simulation of Shoebox Room Acoustics for use in Teaching and Research. Computing and Information Systems, 9 (3):48–51, 2005.

[55] M. Jeub, M. Schafer, and P. Vary. A Binaural Room Impulse Response Database for the Evaluation of Dereverberation Algorithms. In 16th Int. Conf. on Digital Signal Proc., pages 1–5, 2009.

[56] W. Gogel and J. Tietz. Absolute motion parallax and the specific distance tendency. Attention, Perception, & Psychophysics, 13(2):284– 292, 1973.

[57] A. Bronkhorst and T. Houtgast. Auditory Distance Perception in Rooms. Nature, 397(6719): 517–520, February 1999.

[58] Cabrera D., Jeongb D., Kwake H. J., Kimd J. Y., Auditory room size perception for modeled and measured rooms, Internoiseo Congress and Exposition on Noise Control Engineering, Rio de Janeiro, Brazil, 2005.

[59] Cabrera D., Azzali A., Capra A., Farina A., Martignon P., Perceived room size and source distance in five simulated concert auditoria'' paper presented at the 12th International Congress on Sound and Vibration, Lisbon, Portugal, 2006.

[60] Nielsen S H, Auditory distance perception in different rooms, Journal of the Audio Engineering Society, 41, 755-770, 1993.

[61] Thomson J. A., Is continuous visual monitoring necessary in visually guided locomotion?, Journal of Experimental Psychology: Human Perception and Performance, 9, 427- 443, 1983.

[62] Elliott D., Continuous visual information may be important after all: A failure to replicate Thomson (1983), Journal of Experimental Psychology: Human Perception and Performance, 12, 388-391, 1986.

[63] Fukusima S. S., Loomis J. M., DaSilva J. A., Visual perception of egocentric distance as assessed by triangulation, Journal of Experimental Psychology: Human Perception and Performance 23, 86-100, 1997.

[64] Interrante, V., Ries, B., Lindquist, J., Kaeding, M., & Anderson, L. Elucidating factors that can facilitate veridical spatial perception in immersive virtual environments. Presence: Teleoperators and Virtual Environments, 17(2), 176-198, 2008.

[65] Armbrüster, C., Wolter, M., Kuhlen, T., Spijkers, W., & Fimm, B., Depth perception in virtual reality: distance estimations in peri-and extrapersonal space. Cyberpsychology & Behavior, 11(1), 9-15, 2008.

[66] Naceri, A., Chellali, R., Dionnet, F., & Toma, S. , Depth perception within virtual environments: a comparative study between wide screen stereoscopic displays and head mounted devices. In Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns, COMPUTATIONWORLD'09. Computation World: (pp. 460-466), 2009.

[67] Grechkin, T. Y., Nguyen, T. D., Plumert, J. M., Cremer, J. F., & Kearney, J. K., How does presentation method and measurement protocol affect distance estimation in real and virtual environments?. ACM Transactions on Applied Perception (TAP), 7(4), 26, 2010.

[68] Alexandrova, I. V., Teneva, P. T., de la Rosa, S., Kloos, U., Bülthoff, H. H., & Mohler, B. J. , Egocentric distance judgments in a large screen display immersive virtual environment. In Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization (pp. 57-60). ACM, 2010.