Sensing virtual space: perceptual interaction between acoustic and visual cues in the experience and exploration of a virtual room.

Alexandra Verzier
Graduation Thesis, November 2019
Media Technology MSc program, Leiden University
Supervisors: Edwin van der Heide, Marcello Gómez Maureira
alexandraverzier@gmail.com

Abstract

Human perception is fundamentally multimodal: we combine information from different modalities to interpret our surroundings. A solid body of research probes audiovisual interaction in object perception, resulting in a wide range of perceptual biasses and illusions due to cross-modal effects. The perception of space generated less research, often regarded as a neutral background where events take place. But natural and architectural environments present infinite combinations of dimensions, shapes and materials that are sensed visually as well as aurally. How a space sounds and how a space looks manifests itself through different perceptual qualities, impacting the experience of the environment. However hearing space is rarely considered beyond psychoacoustics or noise-control engineering issues, limiting our understanding of the multisensory perception of space. This exploratory research aims at investigating the representations of space emerging from the senses of audition and vision. Virtual Reality was used to create a dynamic room that participants explored firstly through sound-only environment and then audiovisual environment. Throughout the experiment the dimensions and material of the room are changing. We attempt to understand aural-specific experience of this space, as well as audiovisual integration by introducing a condition of Mismatch between the visual and sonic simulation of the space. Impact in terms of objective and qualitative assessment of the space is reported.

1. Introduction

Space is not a neutral, empty container: some people cannot handle small spaces, while some others feel overwhelmed in open wide spaces. We sometimes feel like we need space to breath and sometimes we need a sense of boundary to face the immensity. This “atmosphere” can affect our perception and experience of the environment.

The perception of space is mainly researched through the visual modality. But hearing is another crucial sense for spatial perception. Firstly, sound has the specificity to connect us to our whole environment, in particular elements inaccessible to the visual field, situated behind our head or occluded. Secondly, each space has its own acoustic imprint depending on its dimensions and the type and repartition of materials that it comprises, which can be sensed. The common behaviours of shouting faced to a mountain to hear an echo or singing underneath bridges show we all have a sensitivity to the way space embraces everyday sounds.

How do hearing and seeing interact in our perception of the atmosphere of spaces ? Studies have shown that multisensory processing can distort our perception, creating sensory illusions or biasses. However, existing experiments have focused on minimal objects or stimuli, and we do not yet know whether this extends to the perception of space.

We aimed at gaining better insights into the specificities of sonic compared to visual perception of space. Moreover, we investigated interaction between both modalities. We designed an experiment where people explore an architectural space in Virtual Reality, firstly only aurally and then audiovisually. Throughout the scene, dimensions and material of the room are changing. In one condition, we created a mismatch between the visual and sonic simulation of the space. We examined how objective and subjective features of the space are perceived by participants, depending on sensory domain and audiovisual congruency of room cues. Finally, short interviews are led to better grasp experiential qualities provided by hearing or seeing the space.

We first review related research in sound space perception, multisensory perception and audiovisual virtual environments. After this context is outlined, we describe more precisely our experiment and results.

2. Background

2.1. Theoretical background

The neglecting of hearing in the perception of space can be apprehended as rooted into historical theorisations of space. Here we briefly describe how conceiving space as static and homogeneous made architecture a discipline essentially linked to the eye, and until recently drastically limited the enquiries made into the relations between sound and space.

In Western science and philosophy, theories of space remained moulded by Euclidean geometry until Einstein's theory of relativity. Following Euclid, space has been defined as static and homogeneous which consequently limited the understanding of perception. Vision has been favoured to gauge space, considered instantaneous and objective, compared to the other senses inscribed in temporality and thus instability. This also extended to architecture, considered exclusively as a spatial, visual art. Traditionally, we would think we cannot hear space, because hearing takes time while space is instantaneously *there.* Juhani Pallasmaa in *The Eyes of the skin* (2005) argues that as a consequence of the hegemony of vision, architecture became throughout the ages assessed by the eye only. Buildings are defined primarily by their exteriority, by the image they will produce when seen from a distance. For a long time, most buildings, even those where sound was critical (for instance concert halls or theatres) had their acoustics left to chance, a mere byproduct of visual design. A change was initiated at the beginning of the 20th century when Wallace Sabin originated Architectural Acoustics by systematically measuring how sounds would resonate in rooms, as a function of room dimensions and materials. Throughout the century, architects became increasingly aware of the sonic dimension of space, as reflected by Sten Eileen Rasmussen's statement in *Experiencing Architecture* (1959, p. 224):

> Most people would probably say that as architecture does not produce sound, it cannot be heard. But neither does it radiate light and yet it can be seen. We see the light it reflects and thereby gain an impression of form and material. In the same way we hear the sounds it reflects and they, too, give us an impression of form and material.

This interest was echoed by composers who gradually incorporated spatiality in their music. In 1969 Alvin Lucier created *I am sitting in a room,* a powerful piece demonstrating the sonic quality of space. He recorded and played back his voice in a room recursively until most of his speech became unintelligible. At every step the room acts as a giant filer, selectively erasing or preserving certain frequencies out of the voice information. Repeating enough time, only the resonant frequencies of the room are still audible. In *Chambers* (1980, p. 35), Lucier comments on this piece: "*Thinking of sounds as measurable wavelengths, instead of as high or low musical notes, has changed my whole idea of music from a metaphor to a fact and, in a real way, has connected me to architecture*".

Hearing space has only been recently considered, connected with an idea of space as heterogeneous, entangled with human perception which necessarily unfolds in time.

2.2. The sonic experience of space

Acoustics are a physical phenomenon caused by the reflections of sound waves by surrounding surfaces. Each space's acoustic environment is unique and can be sensed. We describe here the objective and subjective spatial parameters that humans can perceive through sound.

Sound is constituted by sound waves travelling through a medium. When encountering a surface, a portion of the sound crosses the material, while the rest gets reflected. Each environment has different acoustics, based on the reflective surfaces it comprises, their material and distance from the source.

In architectural spaces, early reflections of the sound can be computed to deduct the position of the walls, ceiling and floor. Reverberation, constituted by later reflections of the sound, is on the other hand indicative of the general volume and materials constituting the space.

Humans can perceive objective features of their environment through sound. But this ability is very disparate among populations; on one end of the spectrum, blind people trained in echolocation can use the reflections of their tongue clicks to perceive the position of objects in space, with an accuracy close to the one found for the peripheral visual field of sighted individuals (Teng, Puri & Whitney, 2011). Normal subjects are able to roughly estimate the size of a room when blindfolded (see Cabrera, 2007 for a review).

Through hearing humans are also sensitive to qualitative features of the space; spaces and their acoustics co-evolved to endow each others with certain affective attributes. In a small reverberant room, such as a bathroom, the sound reflections are confounded with the direct sound, creating a loud, disruptive and overwhelming reverberation. Oppositely, a long delay between the direct sound and early reflections conveys an impression of spaciousness. The clearness of concert halls renders a warm colour, creating a sense of intimacy despite the distance. There, generally two seconds of reverberation time is preferred to create an enveloping sound. In churches, longer and louder reverberation is found, impregnating the space with a "spiritual" character (Blesser, 2007, pp 67-126).

Architectural acoustics enable people to sense objective as well as affective features of the space. Despite being often neglected, it influences spatial experience to an extent yet to be specified.

2.3. Multisensory perception and cross-modal effects

The brain perceives the environment by retrieving and inferring information from different senses. This process gives rise to cross-modal effects, creating biasses and illusions in multimodal perception. We know it applies to the perception of objects — but does it apply to space ? We start by reviewing research in audiovisual interaction and continue with the few studies that considered cross-modal effects in architectural space. Lastly, we examine existing studies in virtual environments providing insights into how audition and vision impact the experience of space.

2.3.1. Audiovisual interaction

Multisensory integration research shows that perception functions by combining partial unimodal signals to draw a representation of the environment. Inference between different sensory channels helps the brain to be more efficient in reconstructing the scene probabilistically. Consequently, stimuli close in time, location and semantically congruent are generally processed together. Multisensory integration enables faster and more accurate performances (Spence, 2007).

As a consequence of this functioning, stimuli tend to influence each others across modalities. Indeed, stimuli expected to "go together" can be perceptually distorted as a way to solve a cross-modal conflict. Several multisensory illusions or biasses have been extensively studied: in the flash illusion, additional flashes of light can be perceived instead of one when it is accompanied with several tone beeps. The McGurk effect shows that lip movements influence the way we perceive speech, with for example "DA" heard instead of "BA" when lip movements diverge from the auditory stimulus. A bias in auditory location is often found towards its attributed visual source, an illusion called the ventriloquist effect. Beyond audiovisual interaction, cross-modal illusions can be found in other modalities, such as touch and vision in the Rubber Hand Illusion. Researchers hypothesise that these effects are caused by a cognitive strategy aiming to force stimuli attributed to a common source to align into a coherent percept (Chen & Spence, 2017).

The ongoing paradigm in multisensory research consider information from one or more sensory modalities, generated by one or more objects as the source. Research is rarely done in an ecological context, and stimuli are typically very abstract (such as light flashes and beeps). We do not know yet whether cross-modal effects, and by extension multisensory biasses or illusions, extend to and affect the perception of space.

2.3.2. Multisensory space processing

Only few research investigated multisensory perception of space. It suggests that subjects have general expectations and are sensitive to the relation between sonic and visual properties of the space. Subjects perform well when asked to match an auralised recording to an image of the room it was recorded in (Cabrera, 2007). Furthermore, when asked to rank reverberation levels across sound samples, both architects and non-architects had better performances when presented with congruent pictures of the spaces. Meanwhile, incongruent pictures resulted in a drop of performance in the architect group, but not the non-architect group (Defays et Al., 2015). This might imply that the expertise of architects endow them with stronger expectations based on the visual outline of the room, which affected in turn their perception of the sound.

We still have very restricted knowledge about how different modalities concur to represent the space. As referenced previously, when perceiving objects incongruent information can be resolved by perceptual biasses and illusions, or segregation of the multisensory stimuli. How are conflicting spaces experienced ? Contrarily to objects, stimuli emanating from space do not appear to be segregated: one space is always perceived as one space. As an example, an anechoic room is a very unusual environment — designed to absorb a maximum of the sound energy, it suppresses almost entirely sound reflections. Natural habitats always comport reflective surfaces, and closed spaces generally have reverberation; in an anechoic room, expectations are violated because it is silent, sources sound remote and faded, yet it is enclosed. Barry Blesser described the anechoic chamber as an instance of "spacelessness" (2007, p. 18): without the sound reflections that usually mark - consciously or not - the boundaries of the space, it feels like the sound is being aspirated in a vacuum, and the space fades. Many subjective reports of experiencing this kind of room mention nausea, uneasiness, disorientation, and other affective responses. This could be due to unusual sound energy distribution, but we can also interrogate what is the effect of the mismatch between visual and sonic spatial information.

In Virtual Environments (VE), every spatial property can be composed freely by its designer. This affords combinations not possible in natural environments. Thus, it is interesting to know how VE, which might contradict natural conditions and expectations, are experienced. Moreover, by allowing researchers to tune visual and acoustic simulations of the space, VE can give us more insights about natural multisensory integration of space.

2.3.3. Audiovisual virtual spaces

Interactive Virtual Environments (VE) are constantly improving and can simulate with an increasing efficiency natural visual and auditory perception. Using this medium, we can create a particularly compelling illusion of space, representing an interesting potential for research. However, so far little and heterogeneous research has been led in VE. Moreover, the sense of hearing is rarely considered, as the "virtual world" referred to is most of the time meant as a "visual world" (Viaud Delmond, 2007). In this section, we gather the few research relating to acoustic and/or audiovisual perception of space in VE.

One key area of interest in VE is Presence, defined as the "sense of being there" in the virtual space. Sound increases presence in comparison to silent VE, and there are indications that binaural and spatialised sound cues can also positively impact presence (see Nordahl & Nilsson, 2014 for review). Larsson, Västfjäll & Kleiner (2008) found that in an audio-only VE, the addition of room acoustic cues increased presence ratings (but not realism) compared to anechoic condition. However, in the only experiment so far investigating audiovisual spatial conflict, no relation was found between reported presence and congruency of acoustic and visual room cues (Larsson et Al., 2007). Mismatched and matched auralisation provided similar outcomes, contrasting only with the "no sound" condition. We do not know if it is due to a fidelity issue in the virtual simulation, if it is because realism of the environment does not impact presence, or if audiovisual cross-modal interaction prevent people from perceiving the discrepancy.

A few VR studies considered cross-modal effects in architectural space perception with contradictory results. One recent study found that reverberation judgments were the same in sound-only, visual-confirming and visual-conflicting condition (Schutte, Ewert & Wiegrebe, 2019), implying the visual space did not influence the aural perception of the space. One important limit of the study is that the experiment was preceded by a sound-only training that indicated the grade each reverberation should receive on a scale from 1 to 10. This may have lessened potential cross-modal effects by focusing participants on the aural task. Another recent rigorous study comparing Virtual and Real Environment (RE) draw opposite conclusions (Maempel & Horn, 2018). The authors showed that stimuli features were rated differently based on domain. Audiovisually, sound was perceived as louder and more enveloping than in the auditory domain, and image as darker and more reddish than in the visual domain. Furthermore, when significant differences in features' ratings were found between RE and VE, it was only the case in unimodal scenes. Divergences found between grades made in the auditory or visual domain would disappear in the audiovisual domain. This suggests that the complementary congruent sensory information made participants perceive RE and VE analogously. This was the case even for unimodal features, such as room brightness, reverberance and envelopment. The authors suggest that people cognitively reconstruct the space, compensating for virtualisation, based on information acquired from another stimulus domain and abstract experiential knowledge of rooms.

Although clues exist in favour of cross-modal interaction in architectural space perception, evidences are still scarce, as this topic remains largely under-researched (Maempel, 2017). Moreover, we still have little understanding of space perception via hearing, and the different experiential qualities ears might convey compared to eyes.

3. Experiment

Considering the issues previously outlined we aimed at investigating through an exploratory design the spatial representations emerging from aural and audiovisual domains in interactive Virtual Environment. Participants explored freely a virtual space where acoustic and visual simulations were varied. We paid a close look to the experience of solely hearing the space compared with audiovisual exploration. We sought in the meantime to understand possible interaction between modalities by introducing one audiovisual Mismatch condition. Through questionnaires and semi-conducted interviews, we examined the perception of the subjective and objective features of the space, depending on domain and cross-modal congruency.

3.1. Design

We designed an architectural space to be experienced in Virtual Reality. We took advantage of VR technology to enable free head and body movements within a restricted area. This addresses a drawback of most studies in auditory room size perception where participants are seated and cannot move the head or body, while for spatial hearing movement is paramount.

We wanted to explore the perception of space in a rich cued environment to approximate real-life conditions: sound sources have different spectral characteristics, reflecting different parameters of the space. Additionally, we aimed at selecting sources which would not be typical of one specific space, to prevent participants from relying on the sound semantic type to estimate which space they stood in. Consequently, our scene consists of various trivial events and sounds that could be heard “anywhere”: someone is playing the guitar, a woman is talking over the phone, various ambient and foley sounds are spread around.

Our space has dynamic features. While the nature of the scene suggests it should take place continuously within one space, along the experiment the room enlarges and changes material. Aurally, early reflections strength and delay and late reverberation are varied in real-time in order to simulate the space transformations. In the audiovisual scenes the room transformations are also visually simulated. We created one Mismatch condition where a conflict is introduced between modalities, the visual room being larger than the acoustic room.

We created two scenes to be able to examine both spatial changes separately. In **scene_1** the dimensions of the room change, and in **scene_2** the main material of the room changes. Each scene lasts approximatively 3 minutes, split between 1:30 minutes after and before the transformation of the space. The scenes are experienced firstly only aurally and then audiovisually, following a similar assumption as Maempel & Horn (2018): the perception of the sound environment should not affect subsequent estimates made once the visual modality is available.

We can define the three following rooms simulated to be visually and acoustically congruent:

1. **Room 1**: A small room made of wooden tiles on the walls and parquet.
2. **Room 2**: A large room made of wooden tiles on the walls and parquet.
3. **Room 3**: A large room made of concrete on the walls and parquet on the floor.

Additionally, for the Mismatch condition, the following visual room was created:

4. **Room V+1**: A room of medium size made of wooden tiles on the walls and parquet.

The characteristics of the rooms are summarised in **table 1.**

Table 1: Characteristics of the virtual rooms. Width, length and height are expressed in meter and volume in meter cube. The enlargement factor is expressed compared to room 1.

	Width	**Length**	**Height**	**Volume**	**Enlargement factor**	**Dominant Material**
Room 1	7.2	10	3.7	268	-	Wood
Room V+1	14.2	21.6	4.0	1230	5	Wood
Room 2	20.6	31.3	4.2	2708	10	Wood
Room 3	20.6	31.3	4.2	2708	10	Concrete

3.2. Experimental procedure

Each participant is assigned to one group, Match or Mismatch. They witness the two scenes in aural and audiovisual domains, resulting in four small experiments in total. We constructed the following experiments, where A and V denote respectively the Auditory and Visual modality, and 1, 2, 3 denote the room we are using, V+1 referring to the mismatched room. The arrows denote the transformation of the space, from one room to the other:

1. **Experiment 1**: A_{scene1} = A1 → A2
2. **Experiment 2a**: AV_{scene1} = AV1 → AV2
3. **Experiment 2b**: $AV{+}_{scene1}$ = AV+1 → AV2
4. **Experiment 3**: A_{scene2} = A2 → A3
5. **Experiment 4**: AV_{scene2} = AV2 → AV3

Participants in the Match group are tested through experiments **1**, **2a**, **3** and **4**, while the Mismatch group is assigned to **1**, **2b**, **3** and **4**. After each experiment, qualitative and objective assessment of the space is surveyed through questionnaires. Participants report (when relevant) their estimation of the room dimensions, room change of size factor and material. They also provide subjective ratings of the sound quality, the graphic quality and matching between sound and visuals. Spatial presence is assessed after each experiment using the spatial presence component of the I-group Presence Questionnaire (Schubert, Friedmann, & Regenbrecht, 2001). In order to make the questionnaire suitable to our aural experiment, one item (SP2) was slightly modified: the question "I felt I was just perceiving pictures" was changed to "I felt I was just seeing pictures and/or hearing recordings". At the end of the session, a brief semi-conducted interview is recorded to collect impressions from the participants. Through this more qualitative aspects of the relations between the sonic and visual environment are investigated.

3.3. Technical specifications

Unreal Engine 4 was used to create the virtual environment, where we integrated the audio middleware Wwise to generate the sound spatialisation and room acoustic model. The auditory stimuli were for the most part recorded in the anechoic chamber of Delft, Netherlands with an Audio-Technica AT-4040 microphone.

The acoustic simulation of the space is done in Wwise. The early reflections are generated based on an Image-Source method through the Reflect plugin. The late reverberation of each space designed is based on convolution reverberation. We selected Impulse Responses (IR) belonging to spaces that roughly correspond to the visual spaces we are creating. The IR for the wooden room is sampled from a recording studio of 455 meter squares, made mostly from wood.

It was stretched slightly up or down to fit our smaller and larger wooden room. The IR for the concrete room is sampled from a parking lot with concrete as a main material. It was slightly stretched down as our virtual space is slightly smaller. Reverberation times of each room were adjusted based on Sabine formula, using each room's dimensions and materials as variables. Thus, although there is no exact mapping between the visual and sonic representation of the space we ensured coherence between our different rooms. Finally, Head-Related Transfer Function is applied to generate binaural rendering of the scene.

Visually, geometrical meshes and materials of the rooms were based on the engine's native assets. Two characters were created in the software Fuse CC and animated with Mixamo. Additional custom 3D models and animations were created in Autodesk Maya.

The final VR experience uses a Vive Pro virtual reality headset. The experimental room had dimensions of 3 x 5.28 x 2.5 meters, and participants could move within a play area of about 8 meter squares. It was delimited by virtual red lines displayed on the floor to prevent participants from bumping into the walls of the real room. During the audio-only experiments, the Head Mounted Display was entirely black, with the exception of the red lines indicating the play area.

4. Results

Participants were recruited on a voluntary basis and consent for use of their data and recording of the semi-conducted interview was obtained. We tested 16 participants (50% female, M_{age}= 29.13, *SD* = 8.41) blinded to their condition for the whole experiment, excluding for the end interview. No participant reported work experience or education in architecture, while 37.5% indicated they had education in music, and 18.75% reported education in acoustics. All participants except one (93.75%) had already experienced Virtual Reality.

4.1. Questionnaires

In this section we describe the results obtained through analysis of the questionnaires. We start by reporting results related to the objective features of the room. We then pursue with the qualitative assessment of the virtual space, by examining quality grades and presence scores.

4.1.1. Objective room features (room dimensions & material)

After the two auditory scenes, the very first item of the questionnaire enquired if participants were able to detect the transformation of the space by hearing alone. They were simply instructed to report if they perceived the space to change throughout the scene, and what change they heard. As we wanted to know whether those changes would be spontaneously noticed, we did not provide any suggestions to participants and they were free to write anything.

Firstly we examine results relative to the dimensions estimates of the virtual room.

In **experiment 1***,* 75% of participants reported some kind of change in the space after being exposed to the sonic environment, whether right or wrong. 68.75% of participants spontaneously reported a change of dimensions. 43.75 % of them indicated that the room got bigger, while one participant thought the room got smaller, and three participants (18.75%) indicated they perceived the room changed dimensions but did not indicate in which direction. One participant described the acoustic quality of the space (saying "it got more reverberant") without specifying which space physical change it implied. Although they were clearly instructed to focus on the space, 25% participants did not refer to the space and commented instead the events of the scene, for instance describing the movements of the sound sources.

On the next part of the questionnaire, it was disclosed to them that the room changed size without specifying if it shrank or expanded. At this stage, they had to report their estimate of the first and second room dimensions, and they were prompted to report the same estimates for both rooms in case they did not perceive the space to change. Interestingly, with this extra information only three respondents made a mistake, with one participant indicating that the room diminished in size and two indicating it stayed the same. The remaining participants (81.25 %) all reported that the room increased in size.

After each experiment participants reported their estimate of the room dimensions. They tended to underestimate the size whether in auditory or audiovisual domain, although this error was slightly greater aurally. For the first room, we separate the estimate of the Match and Mismatch group as they were confronted to a different visual room. We ran paired two-tailed t-test with a significance level of 0.05 to test for differences between aural and audiovisual features results.

In the Match group, aural and audiovisual size estimates for **Room 1** were similar. No statically significant difference between domains was found. These values are visible in **Figure 1**.

In the Mismatch group, aural estimates for the first room were similar to the Match group. However, in the AV scene, the estimates increased to be closer to the incongruent visual simulation. Width and length estimates were significantly lower in the auditory domain ($p < 0.05$), with height only falling outside the confidence interval ($p = 0.11$). This is coherent, as height was only slightly different between aural and visual simulation. Results are visible in **Figure 2**.

For the second room of **scene_1**, we analyse both groups together as after the space enlarges it aligns to the same dimensions among both conditions. We obtained similar mean values between domains, although more clear differences emerged regarding the median values. We also observed considerably higher sample Standard Deviation for the aural modality, which can be partially explained by the few respondents that dit not hear the room growing larger and reported estimates closer to the first room dimensions. Nonetheless, no statistically significant difference emerged through the two-tail t-test. These results are summed up in **Figure 3**.

During **scene_2** the virtual room did not change dimensions. This was disclosed to participants in the second part of the questionnaire but they were still prompted to report one estimate. Again no significant difference appeared between domains. Aurally, compared to the

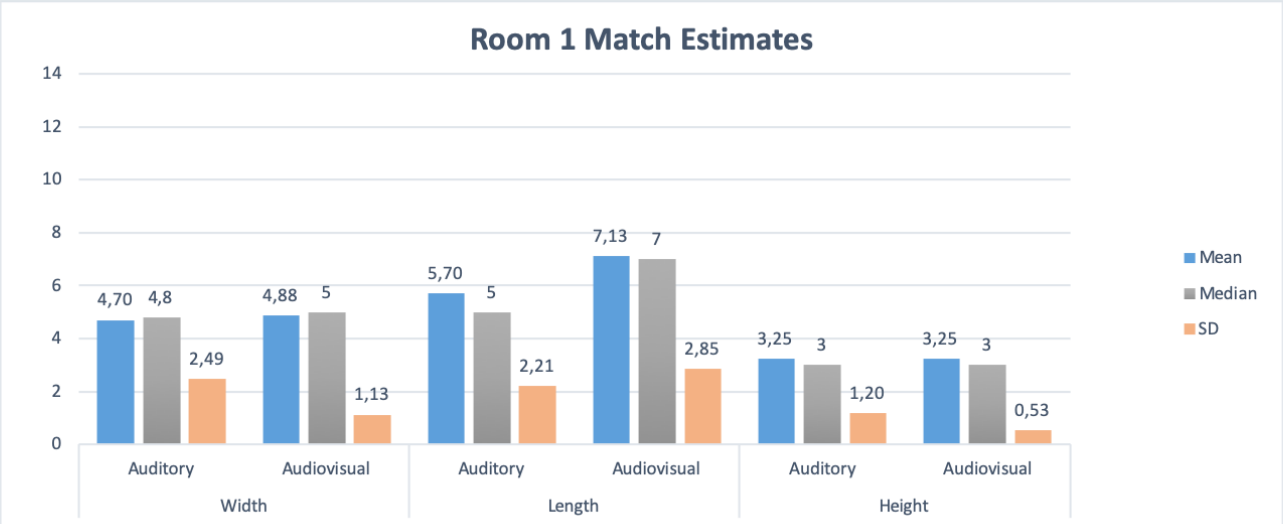

Figure 1 - Dimension estimates in meter of Room 1 in the scene 1, Match group

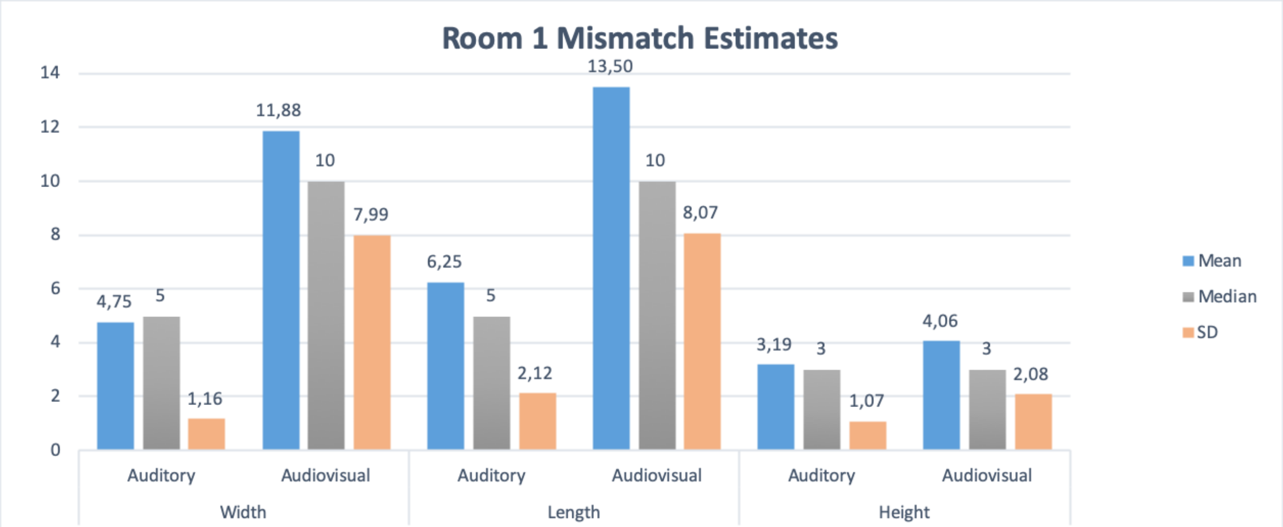

Figure 2 - Dimension estimates in meter of Room V+ 1 in scene 1, Mismatch group

aural estimates for **Room 2**, the mean estimates increased for the room width and length, while it decreased for height. This difference was however not significant. Interestingly, we expected the sample Standard Deviation to lower and the data to be less skewed, as participants had access to a visual reference for the room in the previous step. Moreover, they should then all be aware they are standing in a large room. The SD indeed decreased for estimates of width and height, but it actually increased for the length. Furthermore, it remained considerably higher than the SD of the AV estimates. Contrarily to what we could expect there seemed to be only a slight effect of the visual simulation of the room, and participants relied on their ears rather than on their memory to estimate room size. Results are visible on **Figure 4**.

As estimates between domains significantly differ only when a mismatch between the auditory and visual space is created, we confirm existing research indicating that subjects can have a rough idea of the room size by listening to it. Moreover, it seems that when the visual modality is available people rely mostly on their eyes to report room dimensions, as estimates in

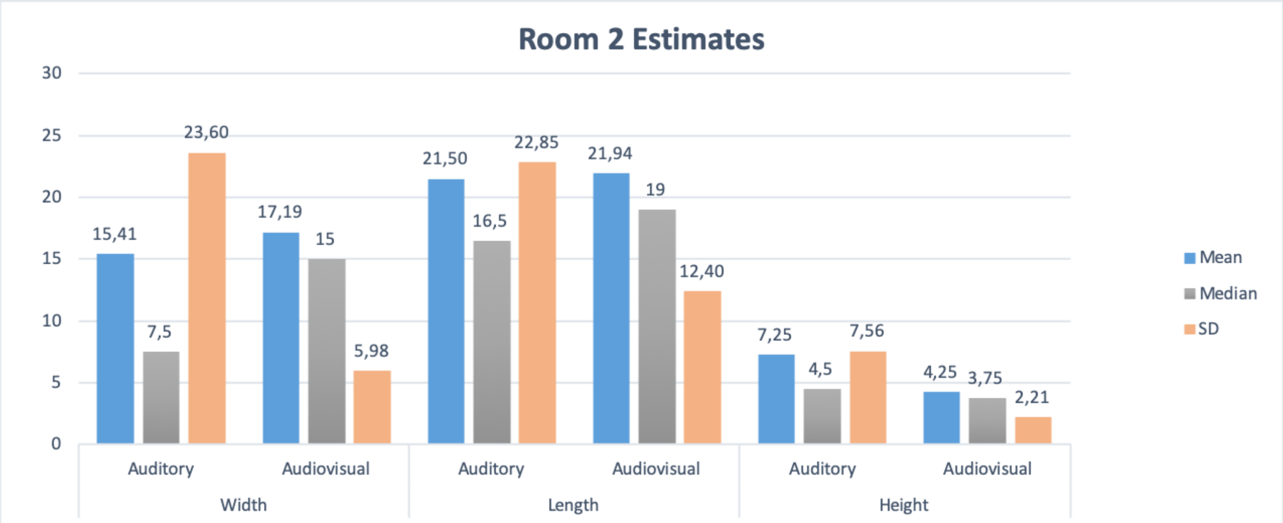

Figure 3 - Dimension estimates in meter of Room 2 in scene 1, all groups

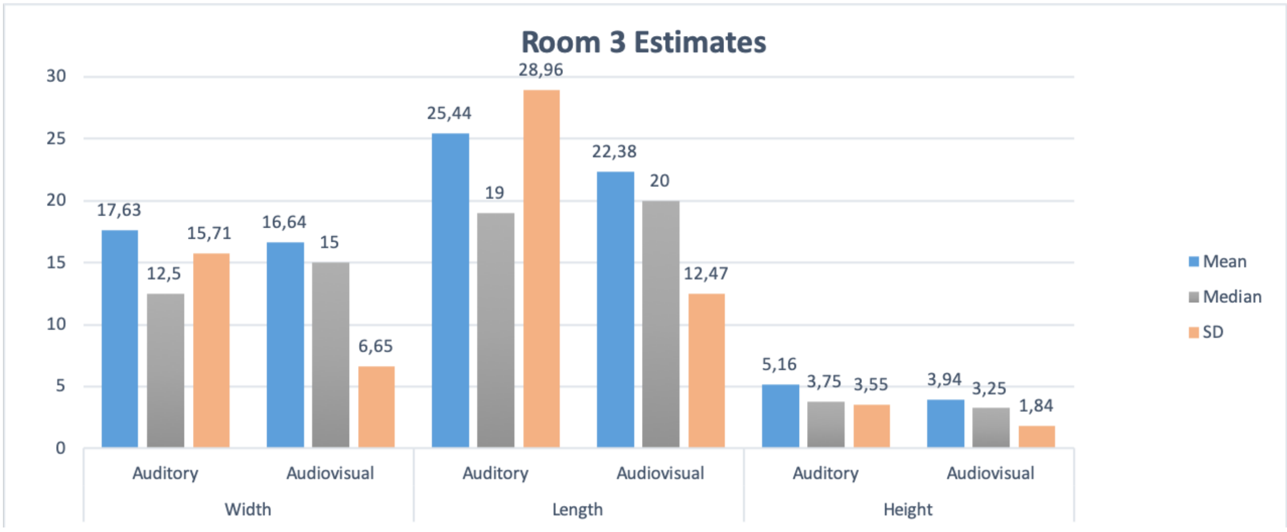

Figure 4 - Dimension estimates in meter of Room 3 in scene 2, all groups

the Mismatch condition are closer to what is visually simulated. However, it does not put aside the possibility of an AV interaction to estimate room size — a more rigorous setup and larger sample would be needed to find a potential effect.

During **scene_1**, participants were asked to report their estimate of the change of size factor between the first and second room — they could indicate that the room grew or shrank, or write 1 if they did not perceive the change. No significant difference between domains emerged in the statistical analysis. In the Match group, the factor reported by participant was lower aurally than in the AV domain but not significantly so. In the Mismatch group, aurally the size changed by a factor of 10 while it visually changed by a factor of 5 only. Although the mean aural estimate was larger ($M_{A|mis}$ = 4.17) than the audiovisual estimate ($M_{AV|mis}$ = 1.79) it fell slightly outside our confidence interval (p = 0.07). Remarkably, one participant in the Mismatch group did notice the change in the aural domain, but not in the AV domain, commenting she was "*focused on the*

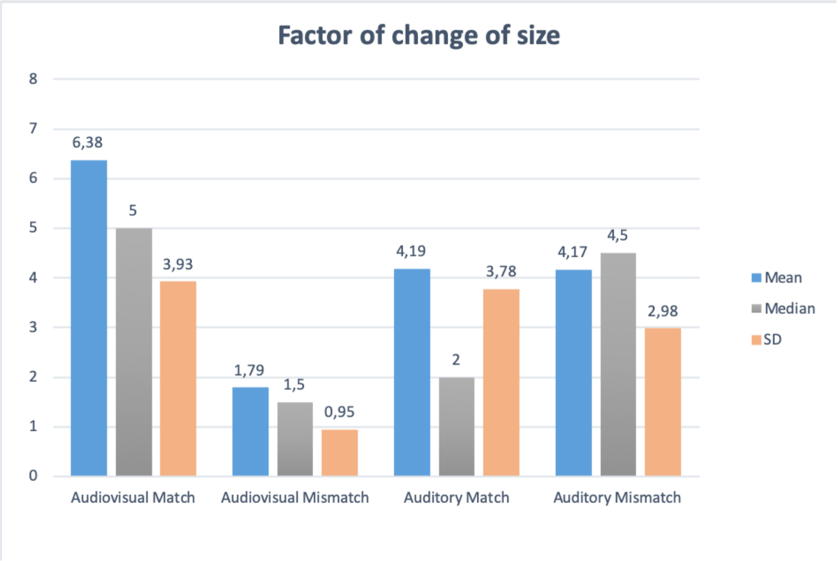

Figure 5 - Estimate of the enlargement factor, per group and domain

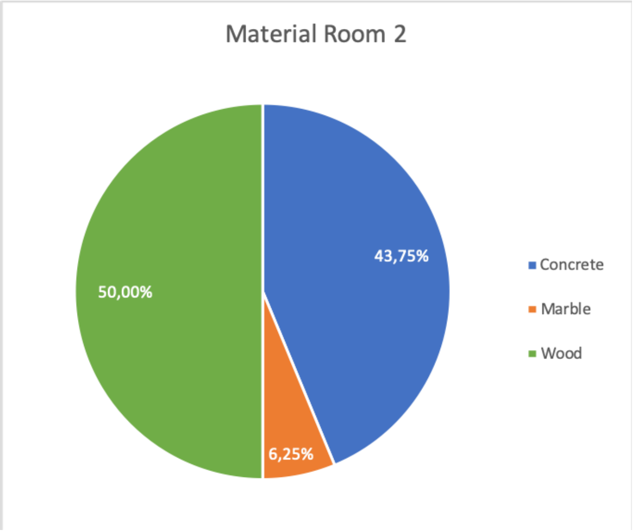

Figure 6 - Aural estimate of material, second virtual room in scene 2

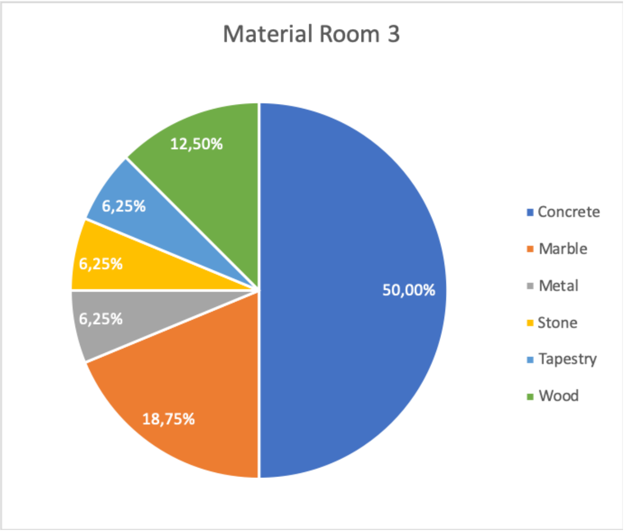

Figure 7 - Aural estimate of material, third virtual room in scene 2

characters and what they were doing". Results are visible in **Figure 5**.

In **scene_2**, the virtual room changed material. We examined if participants could hear this transformation, and then asked them to report the material they heard.

In the audio-only experiment, 81.25% reported a transformation but only five participants (31.25%) spontaneously indicated a change of material. Furthermore, among these participants two of them indicated that the room changed material *and* dimensions. Many respondents were apparently expecting the dimensions to change again: 25% participants focused exclusively on the size, saying the room either got smaller or bigger, or for one participant that they "*did not perceive a change of dimensions*". Finally, 31.25% participants described a change in acoustic quality but did not seem able to point out the change of space it implied. For example, they would say that the space got more reverberant or that the sound surrounded them more. However, in the second part of the questionnaire they were told that the material changed, and prompted to report the same material for both rooms if they did not perceive it. At this stage only three participants (18.75%) indicated twice the same material, while the rest all reported two different ones.

We again did not suggest any option. In the AV domain, all participants but one indicated the right information, going from wood to concrete. In the auditory domain, 50% participants indicated wood as a first material, while the others wrote concrete or marble. This can seem surprising as we were starting with the exact same room as in the preceding experiment, where they could visually check what was the material. But as marble and concrete were the other materials indicated, and as a few participants declared during the interviews that they did not remember which material came first, we can hypothesise that some participants failed to remember how the first room of the scene sounded. As for the second material, 50% participants indicated "concrete", while marble was the second most frequent answer with 18.75% and stone obtained 6.25% , which can be considered closely "sounding" materials. Answers for the aural domain are visible in **Figure 6** and **7**.

4.1.2. Subjective features (presence, quality ratings)

After each scene in VR, part of the questionnaire filled by participants was focused on spatial presence. It was evaluated by the homonymous subset of the I-Presence Questionnaire. Each item is rated on a bipolar Likert scale from -3 to 3, resulting in a total presence score comprised between -15 and 15.

Presence tended to be higher in the AV domain ($M_{AV|scene1}$ = 9, $M_{AV|scene2}$ = 7.88) than aurally ($M_{A|scene1}$ = 6.88, $M_{A|scene2}$ = 6.5). Taking both groups together, this difference was not significant. However, analysis intra-group revealed two contradictory tendencies: in the Match group, presence scores were equivalent between the aural and AV domains. Conversely in the Mismatch group, the AV presence was significantly higher than aurally ($p < 0.01$) for both **scene_1** and **scene_2**. This cannot be a consequence of the Mismatch condition as the first aural scene was the exact same for both group, but rather reflects that although assigned randomly both samples exhibited levels of presence differently sensitive to the sensory domain.

We wanted to know if a mismatch between the aural and visual room simulation would affect presence score. Therefore, we compared spatial presence of the Match and Mismatch group for the audiovisual **scene_1** through two-sample t-test. Even though the presence in the Mismatch group ($M_{AV|Mis}$ = 7.75, $SD = 4.53$) was lower than in the Match group ($M_{AV|M}$ = 10.25, $SD =$ 3.88), it was not significant ($p > 0.05$). The conflict we introduced between sonic and visual simulation of the space did not seem to affect levels of presence. We must remain careful, as the inconsistency observed in domain-dependent presence scores between groups show that both samples might have been too small to find meaningful results between groups.

Quality grades were also comprised between -3 and 3. In **scene_1**, "AV matching" was graded lower in the Mismatch condition ($M_{AV|M}$ = 1.625, $M_{AV|Mis}$= 0.125) but this was not significant. Grades for audio quality and graphic quality were neither sensitive to domain or AV congruency.

4.2. *Semi-conducted interviews*

As our design was exploratory in nature, we decided to lead semi-conducted interviews at the end of the experimental session to have more qualitative insights about subjects' experience of the virtual space. Our main objective was to get a general idea of the sense of space that the audio-only exploration conveyed, how it compares to the AV exploration, as well as finding out more about how the sonic and the visual simulation of the space interacted and were perceived by participants.

4.2.1. *Aural exploration of the virtual room*

Almost all participants pointed out a feeling of disorientation during the first auditory scene. For most, it was the first aspect spontaneously evoked. Having not seen the virtual room yet, it proved difficult to them to build a clear representation of the scene. They would also compare the first to the second audio-only experiment, implying that once they saw and had an idea of the space they could orient themselves a lot more.

> The first one was very confusing and abstract; I was trying to interpret a lot more […] In the second audio condition I was way less disoriented. And I could kind of interpret the sounds way more by knowing where they were located.

> Of course I could picture the objects, based on sound sources. But I had still a very hard time picturing the space and it was making me feel lost. I imagined it was a similar room size. It was more the position of objects in space that was confusing.

The moving sources were more disorienting, as for example the female character which was walking around the virtual room.

> I wasn’t sure if it was me that was moving, if the room was moving, or just the girl.

Nonetheless, for a few participants the sound was enough to create a clear sense of space. One strategy they mentioned would be to mentally separate which sources were moving or not, and use the still sources to orient themselves:

> The guitar player was still so it was like a marker, I would always use it to know where I was, and if I was turning.

We wanted to know where disorientation stemmed from and asked participants if they found it diffficult to perceive the direction and distance of the sound sources. This was the case only for a couple of them. Those could for instance struggle to distinguish between front and back, or distinguish close from far objects. It is possible that for those participants misperception of the sources affected in turns their assessment of the space. Notably, in the first audio-only experiment several participants said they perceived the second space as a long and narrow room, but only one participant referred explicitly to one sound source to justify this representation.

> I thought maybe [the room] would be more of a hallway. I didn’t get [the female character] was walking around in a circle, I thought she was just going from one side to the other.

Most participants did not experience difficulty to locate specific sources but were confused to a greater extent by the disposition of elements in relation to each others. Furthermore, as we saw in the questionnaires relatively high accuracy was reached in terms of perceiving the room growing, which was confirmed in the interviews where most participants asserted to be able to interpret the acoustics. To them, what was the most unclear was the general “mapping” and understanding of the scene.

What appeared from the interviews is that the situation we created might have caused this confusion; as we stated, our intention was to select sound sources which did not evoke one specific space by their nature (say, railway sounds in a train-station) as a way to compel participants to rely on acoustic cues and not sound type to build a representation of the environment. This lack of context placed participants in a “blurry space”, somewhere that felt real yet they could not make sense of. In comparison with the second aural scene, which was considered more legible, the first one altered their perceptual experience: they felt lost, confused, sometimes even lost track of their own movements and position. Although we limited the influence of the sound sources, the drawback of our method is that it caused a few participants to think there was no relation between the cues whatsoever.

> I didn’t understand first that the footsteps were from the woman, or that the man, at some point we hear him briefly, that it was him playing guitar. Only after I saw it I understood. So first I understood every sound as different source.

> I did not have a context, I just thought, these are a collection of things that are not related. I did not imagine there were two people, or even someone playing guitar.

> I thought it was hard to picture why the guitar was there because it was an empty room and… The wind chimes and… I couldn’t picture any space where all those things would occur at the same time.

It is surprising that some participants did not perceive the relations between the sounds, even when some were very obvious (sounds emanating from a common source). For them, it seems that spatialised sound was not enough to create a real sense of space. Therefore, they interpreted auditory stimuli more as a sort of composition rather than as a space surrounding them. Moreover, it seemed to us that certain participants ignored or classified some sounds as external from the scene when they would compromise the coherence of the scenario they consciously formed. For instance, one of these participants thought he heard scissors at the beginning (which was in fact the sound of a lighter) and then imagined to be in a barber shop. Then, he described listening to the rest as extra compositional elements on top of the scene, without depicting the sources he heard as real objects in the space.

This directs us to an unexpected element that appeared in the aural exploration, which is the extent to which participants would rely on other, non-acoustic cues to picture the space they were standing in. Indeed, even though we strove to carefully design the experiment to limit this effect, subjects would still be influenced by semantic associations. This was the case even when the cues they relied on were minimal or contradictory.

> First because I heard the lighter, so I thought maybe about fire, and the wind chimes, I thought we were maybe outside. These sounds reminded me of a festival I recently went too, and people were outside their tents smoking cigarettes, and there were wind chimes all around. But then there was no wind and the acoustics were as if I was indoor so I was confused. Then the girl started talking about yoga so I immediately thought of the room where I once did yoga.
>
> In the very first audio experiment, I heard scissors and stuff so I imagined I was in a barber shop.

Moreover, we realised that these associations did not only come from the sound sources themselves, but from a variety of elements present or not during the experiment.

> When the room changed I did not feel that much that it was the same room getting bigger, more that it turned into a tunnel. […] Earlier today I walked underneath a tunnel so it might have influenced my impression.
>
> [In the second audio-only experiment] I also thought it must be concrete rather than marble because of how the space looked in the previous step, with this big industrial door and this emptiness, it made me think directly of a kind of industrial site.
>
> [In the second audio-only experiment] It was suddenly very reverberant and I felt immersed in the sound, I stopped moving. Suddenly I thought I was outside […]. (*But you answered that the material of the room was concrete?*) Yes, I knew it was more echo-y, like in a cave. But the song reminded me of one I heard at the Piano Sous Les Arbres[1], so even though it should have been inside, I felt like I was outside, like I was again at this place.

Lastly, one participant was influenced by the red lines on the floor, which only represented the play area. Although the role of these lines was clearly stated in the experiment instructions, she thought it also represented walls in the virtual world, and this impacted how she perceived the auditory space.

> First I thought I was in the entrance of an apartment. So I thought the woman was far away and maybe a little off, so I thought she was descending […]. The red square I thought it was walls and I thought she was coming behind. And then when the dimensions of the room changed I thought maybe she entered another room behind. So I thought there were really thin walls. Maybe inside a big apartment, but I was in the hall where I could easily hear different rooms and people.

4.2.2. Interaction between aural and visual information

Another matter of interest was the interaction between both modalities during the AV

[1] Festival of classical music in France, taking place outside in the nature

exploration of the room. To this end, we interrogated participants from both the Match and Mismatch group about how the visual rendering of the scene compared to their expectations built from the preceding perception of the sound space.

We gathered some clues in favour of audiovisual interaction. Most respondents said the visual and sonic space seemed to go together. In some cases, even participants that had built a strong and different expectation from the auditory scene said they did not feel a discrepancy in the AV scene. For them, the sound and visual would blend together and create a coherent whole. The following quote is from a participant in the Mismatch group.

> When I would see the space kind of everything would come back more into proportions for me, so I would associate the dimensions as being equal to each other, so it made sense as soon as I would see the visual part of it. But that's the thing [the change of dimensions] seemed very drastic before seeing the visual, but when you see the visual it kinds of blends into it, so… It made sense.

Moreover, among the few participants that experienced a difficulty to situate the sound sources, the incoherences they noted tended to be alleviated by the visuals in something analogous to the ventriloquist effect. This effect was however limited.

> The footsteps at the beginning I thought it was just random footsteps noises. It didn't feel like someone was walking around me. Although when I saw the video it did look like it was corresponding…

> The small sounds sometimes felt off compared to the big sounds. So for example, the wind chime, it sounded very very close. And same with the lighter, it sounded closer to the ear than it actually was […] *(Did the visual space change anything about how you perceived the sounds ?)* With the visuals, for the lighter yeah it made sense, then it sounded normal, and it was where it should be. But the wind chime still sounded very close.

However, a few respondents perceived a conflict between aural and visual representations of the space. Those respondents formed a strong spatial percept from hearing the scene, and when the visual space contradicted what they previously pictured, they would find it unconvincing. The following quotes are from subjects in the Mismatch group, referring to **scene_1**.

> From how the sound changed, I imagined the room was twice as big, or something like that. And in fact, it was I don't know, maybe 0.1 % as big or something like that. So that was strange. And it might be the only thing that, I thought, there is a mismatch. (*Once you saw it did you still feel something was wrong ?*) Yeah. Yes. Because the sound changes quite a bit. I'm not sure about the acoustic properties of the room so maybe it is supposed to be like that.

> I also thought that the walls were way more absorbing of the sound, so when I saw [the first room] I thought this room is so big, it would create a very different sound. I expected it to look different, the room was bigger than I imagined, I thought it would be more reverberant if the walls were that flat.

Interestingly, some people evoked rather the second scene when interrogated about mismatch. One individual that did not hear the change to concrete was still firmly set on his judgement:

> If it lined up with my estimations that are totally auditory, then it felt like natural. If it did not align with my expectations, then it would feel off. The estimation of the size was alright, it's just the material I did not get […]. And even when the material of the room [visually] changed to concrete, I was still not convinced.

Sometimes a sense of mismatch would emerge from a transformation experienced as more dramatic aurally than visually. Some participants expected the change of space to be more radical, and the AV scene would underwhelm the effect they heard.

> The change was way more dramatic aurally than visually. I thought it became like a big church, and instead it was simply the texture of the walls that changed. I expected the ceiling to change too.
>
> For me there was more of a dissonance in [the second] one… Still it made sense. Seeing it together it was not like I thought there was a real mismatch, more that I thought it was not as good as the thing I pictured in my head.

Disparities between more "sound-based" or "visually-based" respondents also reflected the differences we previously observed in presence score. Indeed, a small majority of respondents explained they felt more present in the visual scenes: it would help them to put things into perspective and into place, and was necessary to create a "true" sense of space for them.

> Everything made sense. For example, this metallic sound, I could see it was because [the male character] moved the chair, and it made total sense […]. Presence, for sure, was higher with visuals. Because you associate these moving sounds with something, objects. Still it was not high quality images, but it was enough to make sense of the scene. Emotionally I don't know. But it makes you feel more present in the space.

Our second group, which demonstrated significantly lower presence aurally, might have randomly selected more visually-based people. Oppositely, for some other respondents the AV scenes would not be as immersive or presence-inducing. They cited notably the graphic quality to be too low and the human avatars to look too fake to justify their experience. Furthermore, as we evoked before, respondents that built strong aural expectations were sometimes underwhelmed by the corresponding visual scene. They were able to rely on the sound to construct a real enough spatial representation that they could "fill" with their imagination, which would be more powerful than the visual simulation we offered. These two tendencies would sometimes be present within a same individual, as shown by the reflection of this respondent:

> In the first [audiovisual] one, maybe because I was so disoriented just before, I thought oh my god this is magical. I completely understood what was going on and it was way more immersive. It made everything click in a way that it made it more convincing. The second one, because I was really sure of this parking lot situation, when the material changed from wood to concrete I understood what was going on but it didn't feel as real as the sound itself.

5. Discussion

One overall observation that could be made after examining our results is that experience of the space diverged in the auditory domain but converged in the audiovisual domain.

In our background, we emphasised that sound space perception is indirect and heterogeneous compared to visual perception, resulting in very disparate abilities, from echolocators to individuals fully unaware of spaces' sonic qualities. Our results strengthen this claim: while the audiovisual space was consistently perceived among respondents, wider disparities emerged when hearing alone was available, reflected in large sample dispersion. The interviews further unveiled how profound these disparities were; some individuals were unable to hear any change in the sound, let alone the room. Most participants were able to "guess" the transformations by reflecting back on what they heard, while a few of them could pick up on the change of space right away and experienced it as more striking aurally than visually.

> I knew what to look for in the visual scene. I remembered that after the guitar started to play, a

> couple notes in, that’s when the sound changes. So I was just looking around waiting for the change to happen at this moment.

Divergence in aural experiences were also apparent in presence scores. We found an unexpected difference between auditory and audiovisual spatial presence only in one group; although it indicates our sample might have been too small to find meaningful results between groups, it nonetheless emphasises that hearing alone seems unable to trigger spatial presence in some individuals. Interestingly, a few respondents reported the opposite effect. They felt more present in the aural scenes as they could fill the empty visual space with their imagination. Furthermore, we witnessed less frequent but more “intense” expressions of presence in the aural scenes: losing track of time, having more trouble coming back to reality… One participant that imagined she was outside said she felt the temperature of the room dropping. Presence in the AV scenes was overall slightly higher but also constant among individuals.

The aural scenes granted more space to participants to construct a subjective representation, resulting in a broad range of spaces and imageries evoked by respondents. The space was felt as less stable, evolving as they would gain more cues throughout the scene. Remarkably, one participant said she associated the space with the type of sound source; for example, the guitar would set for her an intimate and cosy, small space, while the woman voice would bring her back to the space that was really simulated. She felt like switching from one space to another depending on which source would resonate.

The visual modality enabled participants to disambiguate an abstract situation — the sentence we maybe heard the most throughout all the interviews was “it made sense”, when describing the AV experiments. Furthermore, we noticed through the vocabulary employed the extent of the “truth” value attributed to the eye: participants “thought the space to be” a certain way when they would hear it, but “realised the space to be” another way when seeing it. “I saw” would be a synonym of “the space was that way”, and “I heard” a synonym of “I imagined it to be that way”.

Audition and vision brought about a very different experientiality of the space. Hearing stimulated subjective associations and delimited the space in a poetic way, while vision made different subjectivities converge to a common space.

We still need more research to understand to which extent both modalities interact. As we have seen, although room size estimates made in the aural and audiovisual domains were significantly different in the Mismatch condition, proving the mismatch was somehow perceived, this did not significantly affect any qualitative features. The interviews showed that the mismatch impacted people differently, either making expectations vanish or deterring presence and immersion. Future research should examine more in details incongruent room cues. For instance, it can be tested how far can we accentuate the AV conflict before it gets noticed. Moreover, if a very large mismatch is introduced between sonic and visual simulation of the space, it would be interesting to see if this affects spatial presence — one possibility is that virtuality of the space entails a relaxation of reality-based expectations which enables participants to accept the contradictory space as real.

6. Conclusion

We explored phenomenon of auditory and visual perception of space. Through our research, we underlined that acoustics are rarely taken into account in spatial perception, and aimed at gaining a better understanding of hearing architectural space. Additionally we examined how the relations between the aural and visual simulation of the space affected participants.

We found, confirming other studies, that by hearing people can roughly estimate the size of the space. Indeed, when visual and sonic space were congruent, no significant difference appeared between aural and audiovisual estimates related to the room dimensions. The perception of the material of the room caused more difficulty, as only one participant reported the wrong material audiovisually while we found 50% accuracy aurally.

In the analysis, the only significant difference that emerged between domains was in the Mismatch condition, where the aural space was perceived as smaller than the audiovisual space. Even though evidenced through dimensions estimates, mismatch between aural and visual simulation of the room did not affect presence or quality grades, even the item "audiovisual matching". This could indicate cross-modal effects which blended visual and sonic space together, alleviating audiovisual conflict, but more research is needed to clarify this result.

During the interviews, we found that mismatch differently affected respondents. It was consciously noticed only by individuals that built strong expectations and sense of space from their aural explorations, which were contradicted or lessened by the visual representations. For others, visual and aural simulation would make sense together, and form a compelling whole.

We observed that auditory spatial awareness greatly varies from one individual to another, resulting in large sample dispersion in the estimates made in the aural domain. This is also reflected by the unexpected result we found in spatial presence, which was significantly lower aurally only in one group. Considering the aforementioned and echoing disparities, there might be interesting relations between auditory spatial awareness, spatial presence and audiovisual integration to unveil in the future.

Reviewing all discussions with participants, we showed that the sonic and visual experience of the space engage different experiential qualities: vision has a truth value, and can make clear an ambiguous phenomenon, creating a baseline where experiences converge. Hearing, although more volatile and unstable, creates a "blurry space" which convokes imagination and subjectivity to a larger degree. This calls for more consideration of hearing space in VR and multisensory research. Moreover, evidences for audiovisual interaction are still frail, needing more investigations in the topic.

References

Blesser, B., & Salter, L. (2007). *Spaces speak, are you listening?: Experiencing aural architecture*. Cambridge, Massachussetts: M.I.T. Press.

Cabrera, D. (2007). Control of perceived room size using simple binaural stimuli. In *Proceedings of the 13th International Conference on Auditory Display*. Montréal, Canada. https://www.researchgate.net/publication/228384747_Control_of_perceived_room_size_using_simple_binaural_technology

Chen, Y., & Spence, C. (2017). Assessing the Role of the 'Unity Assumption' on Multisensory Integration: A Review. *Frontiers in Psychology, 8*. doi:10.3389/fpsyg.2017.00445

Defays, A., Safin, S., Billon, A., Decaestecker, C., Warzée, N., Leclercq, P., & Nyssen, A. (2015). Bimodal Interaction: The Role of Visual Information in Performing Acoustic Assessment in Architecture. *The Ergonomics Open Journal, 7*(1), 13-20. doi:10.2174/1875934301407010013

Larsson, P., Västfjäll, D., & Kleiner, M. (2008). Effects of auditory information consistency and room acoustic cues on presence in virtual environments. *Acoustical Science and Technology, 29*(2), 191-194. doi:10.1250/ast.29.191

Larsson, P., Västfjäll, D., Olsson, P., & Kleiner, M. (2007). When What You Hear is What You See: Presence and Auditory-Visual Integration in Virtual Environments. *Presence 2007* (pp. 11-18).

Lucier, A., & Simon, D. (1980). *Chambers*. Middletown, Conn: Wesleyan University Press.

Maempel, H. (2017). Audio-Visual Room Perception. *Scientia*. https://www.scientia.global/wp-content/uploads/2017/05/hans-joachim-maempel.pdf

Maempel, H., & Horn, M. (2018). Audiovisual perception of real and virtual rooms. *Journal of Virtual Reality and Broadcasting,* (14), 5th ser. doi:10.20385/1860-2037/14.2017.5

Nordahl, R., & Nilsson, N. C. (2017). *The Oxford handbook of interactive audio* (K. Collins, B. Kapralos, & H. Tessler, Authors). New York, NY: Oxford University Press.

Pallasmaa, J. (2005). *The Eyes of the skin*. Chichester: John Wiley & Sons Ldt.

Rasmussen, S. E. (1962, first edition: 1959). *Experiencing architecture*. Cambridge: M.I.T. Press.

Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The Experience of Presence: Factor Analytic Insights. *Presence: Teleoperators and Virtual Environments*, *10*(3), 266–281. doi: 10.1162/105474601300343603

Schutte, M., Ewert, S. D., & Wiegrebe, L. (2019). The percept of reverberation is not affected by visual room impression in virtual environments. *The Journal of the Acoustical Society of America, 145*(3). doi: 10.1121/1.5093642

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Science and Technology, 28*(2), 61-70. doi:10.1250/ast.28.61

Teng, S., Puri, A., & Whitney, D. (2011). Ultrafine spatial acuity of blind expert human echolocators. *Experimental Brain Research, 216*(4), 483-488. doi:10.1007/s00221-011-2951-1

Viaud-Delmon, I. (2007). Corps, action et cognition : la réalité virtuelle au défi des sciences cognitives. *Intellectica. Revue De L'Association Pour La Recherche Cognitive*, *45*(1), 37–58. doi: 10.3406/intel.2007.1266