The Visual Computer
https://doi.org/10.1007/s00371-024-03707-6

RESEARCH

Minimally disruptive auditory cues: their impact on visual performance in virtual reality

Daniel Jiménez-Navarro[1] · Ana Serrano[2] · Sandra Malpica[2,3]

Accepted: 28 October 2024
© The Author(s) 2024

Abstract
Virtual reality (VR) has the potential to become a revolutionary technology with a significant impact on our daily lives. The immersive experience provided by VR equipment, where the user's body and senses are used to interact with the surrounding content, accompanied by the feeling of presence elicits a realistic behavioral response. In this work, we leverage the full control of audiovisual cues provided by VR to study an audiovisual suppression effect (ASE) where auditory stimuli degrade visual performance. In particular, we explore if barely audible sounds (in the range of the limits of hearing frequencies) generated following a specific spatiotemporal setup can still trigger the ASE while participants are experiencing high cognitive loads. A first study is carried out to find out how sound volume and frequency can impact this suppression effect, while the second study includes higher cognitive load scenarios closer to real applications. Our results show that the ASE is robust to variations in frequency, volume and cognitive load, achieving a reduction of visual perception with the proposed hardly audible sounds. Using such auditory cues means that this effect could be used in real applications, from entertaining to VR techniques like redirected walking.

Keywords Virtual reality · Multimodality · Human perception · Suppression effect

1 Introduction

Virtual reality (VR) empowers users with an active, immersive role. In this context, users can control the virtual environment and interact with their surroundings through their senses. Particularly, the virtual environment exposes users to digital information involving several human senses (usually a combination of vestibular, visual and auditory information), a concept known as multimodality. As expected, there are still differences between reality and VR that are technically difficult to overcome, in terms of both hardware and software or content generation. However, they are similar enough for VR to achieve realistic responses, thus being a powerful tool to study human perception as it offers great control over sensory stimuli, reproducibility and affordability [64].

In general, VR is mainly focused on delivering audiovisual experiences. According to the literature, sight is the sense on which people most heavily rely to perceive extra-personal space while audition is used to perceive environment zones we cannot see (occluded regions, outside our field of view or under low luminance) [20, 51]. It is important to note that not all sensory information collected by our brain is processed in the same way; however, more relevant information is selected while discarding the rest, while keeping a stable and coherent perception of our surroundings [39]. This information selection process goes continuously unnoticed on a daily basis. For example, when performing visual saccades, which are fast eye movements between fixation points, our brain wisely ignores the blurred visual information collected during this fast movement that may cause discomfort. Other situations could be during eye blinking, when visual continuity is not affected by these intervals without visual input, and the steady visual suppression of our nose, located inside our field of view (FoV).

Similarly, inhibitory interactions can also occur when dealing with multimodal information. Crossmodal percep-

✉ Daniel Jiménez-Navarro
djimenez@mpi-inf.mpg.de

Ana Serrano
anase@unizar.es

Sandra Malpica
smalpica@unizar.es

[1] Max Planck Institute for Informatics, Saarbrücken, Germany

[2] University of Zaragoza, I3A, Zaragoza, Spain

[3] Centro Universitario de la Defensa, Zaragoza, Spain

Published online: 10 December 2024

tion is defined as those multimodal effects that involve interactions between two or more different sensory modalities [35]. These interactions can be facilitatory (for example, decreasing reaction time in a search task when two modalities are presented synchronously) or inhibitory, depending on how the cerebral cortex is activated to process the perceived sensory signals.

In this work, we build upon a previous work by Malpica et al. [34] which studies an audiovisual suppression effect (ASE) in immersive environments. This previous work reported how auditory stimuli can degrade visual performance in virtual scenarios when certain conditions are met, namely a spatial incongruence between the auditory and the visual part while keeping the temporal synchronization. This effect was observed to be robust in both detection (the visual target was perceived or not) and recognition (once detected, the specific shape of the visual target was correctly identified) while varying the type and location for both visual and auditory stimuli. Since Malpica et al. employed overt audio sources, our main goal is to research whether this inhibitory effect can still be elicited when the auditory cues are barely audible, resulting in the loss of visual information without the awareness of the user.

For example, visual suppression techniques are used in foveated rendering [41], accounting for eye lower resolution in the periphery to wisely distribute image resolution allowing saving computing time, and redirected walking [47] algorithms, using visual suppression during eye saccades to alter the scene. However, both methods are resource-demanding since foveated rendering needs high-accuracy eye tracking and redirected walking needs to act on the rendered content in real-time. Since most of the content experiences are audiovisual, achieving a similar visual suppression effect through minimally intrusive hearing cues would be the ideal condition to act also on the visual input but without degrading the experience. This method overcomes the limitations previously mentioned and could also be used to improve the existing techniques. The recorded data, including eye-tracking information, are available at https://minimalase.mpi-inf.mpg.de/ to encourage further analysis of this phenomenon.

The main findings of this work can be summarized as follows:

- We present a user study that builds upon the work of Malpica et al. which explores a space of auditory features including volume and frequency under an increased cognitive load environment. We find that the ASE can still be elicited under varying conditions, including changes in volume and frequency.
- We confirm that the ASE can be used in a less disruptive way by using subtle sounds associated with frequency limits, which could potentially increase its direct applicability.

2 Related work

2.1 Multimodality and audiovisual interactions

Multimodality stands as a prominent investigation subject within the VR field, becoming a multidisciplinary topic of interest for neuroscientists, graphics practitioners and content creators alike [35]. Within the context of our research, the integration of multimodal techniques can have a great impact on users' virtual experience and sense of immersion by properly integrating information from different sensory modalities in a realistic way. We refer the reader to the work of Martin et al. [35] for a comprehensive study of how multimodality is being used to improve VR experiences.

Feelings like presence and realism are heavily linked to the sensory information perceived by the users. Previous research refers to sight as the sense on which people depend most daily thus being key in the interaction with our surroundings [8, 56]. Consequently, visual information is also the primary sensory modality in VR [48]. As such, research in VR is often dependent on visual modality, from attention prediction and redirecting techniques to visual feedback methods applied in fields such as medicine or psychology treatments [38, 50]. On the other hand, hearing is the other sense on which we rely for extra-personal space perception because of survival reasons. Information related to occluded areas or regions outside our field of view, where our eyesight does not reach, is collected mainly via hearing [57]. Similarly, useful information regarding surfaces and materials can also be extracted from auditory cues produced by their interaction with other entities, which can potentially influence how they are perceived [60]. In VR, auditory information and feedback have been used in exploration tasks, using only acoustic cues to navigate and reconstruct the environment [32, 44] or in rehabilitation processes [5, 6].

Audiovisual information has proven to be useful in VR to redirect users' attention [10], self-orientation [27], increasing sound localization accuracy [1] and learning tasks [53], among other applications. Interactions between visual and auditory cues, when properly synchronized, tend to produce facilitatory effects that improve performance in tasks such as visual learning [30] or speech processing [3]. If the crossmodal synchronization is temporally or spatially not preserved, the visual information becomes not consistent with the auditory cues or vice versa, and the interaction between modalities can result in the perception of two different events or even trigger inhibitory effects. These inhibitory or suppression effects are the ones we focus on in our work.

2.2 Illusions: inhibitory effects

Illusions and distortion effects are inherent to how we perceive reality, and its origin and action field can be very varied: memory illusions, recalling events that never occurred but are seen as past episodes [42, 43]; time perception illusions, where temporal duration, order and simultaneity can be altered [18, 33]; and sensory illusions, like the well-known rubber hand illusion [29] where the rubber hand is perceived as one's own. In all these cases, the human brain can be easily deceived to misconceive our surroundings due to different reasons which are not fully understood yet in some cases [63]. In this sense, Just Noticeable Differences (JND) [62] define the minimum stimuli variations required to be perceived by our brain/human perception system. Avoiding such changes helps to provide illusory effects to deceive our perception, also regarding audiovisual modalities and their interaction [16, 22].

Sometimes, these sensory illusions translate into suppression effects where some information is lost or wrongly recognized. This information misperception has proven to be useful in some applications. For example, it was demonstrated how golf putting performance can be enhanced when the hole size is misconceived owing to surrounding it with smaller or larger circles during training [12]. Suppression effects can involve a single sense such as vision [14, 21], where eye saccades [37, 49] and eye blinks [67] lead to a loss of visual input for a period of time. Suppression effects can also be found involving more senses like hearing, where sound intensity and simultaneity play a role [45], or haptic, whose intensity can vary depending on the movement of a specific body part while also occurring in a static state [11].

Therefore, suppression effects can also take place involving several senses. In those cases, stimuli associated with one modality can alter the information perceived by another, evidencing crossmodal interactions among senses at neural level [35]. Previous works reported how olfactory cues can influence and modulate haptic perception *(olfactory-haptic)* [15], how visual information affects tension levels when experiencing a musical performance *(visual-auditory)* [66], and how visual discrimination can be degraded by tactile stimulation under certain conditions *(visual-haptic)* [25]. Similarly, the ventriloquism effect is a well-known audiovisual illusion where the synchrony between acoustic and visual signals encodes information on sound location [23]. This work addresses a visual suppression effect, where auditory information is used to degrade visual performance.

2.3 Multimodal suppression: audiovisual suppression effect (ASE)

Focusing on multimodal suppression effects, Hidaka et al. reported how auditory [24] stimulation can degrade visual discrimination performance, finding that auditory signals could degrade visual performance in a conventional display setting.

Regarding VR, Malpica et al. [34] later observed how visual performance can be further degraded under specific circumstances when auditory stimuli are included. In their experiment, visual information was presented standalone or accompanied by auditory cues. A degradation of the visual performance was found when both auditory and visual information (the bimodal condition) were presented synchronously under specific conditions in the virtual environment with respect to the visual-only condition. In particular, they introduce a novel approach including auditory cues that are spatially incongruent with the visual stimuli. Moreover, they test visual performance at two different levels: detection (if the visual target was perceived or not) and recognition (once perceived, if the shape of the visual target was correctly identified). The visual cues used as targets were simple shapes (acting as flashes) that the participants had to look for in the virtual environment.

Furthermore, different types of audio sources were presented to the user immersed in the virtual environment, while studying if both location and type of visual and auditory information had any influence. Analyzing the recorded gaze data, they claim that visual performance degradation was not caused by oculomotor phenomena (i.e., saccades toward the sound source) but by neural interactions. In their work, Malpica et al. use exogenous sounds (such as a train horn or a human scream) that are orthogonal to the task and context of their experiment. This limits the usability of their findings since it is not always possible to include these types of sounds depending on the context and needs of a particular application.

Our main goal in this work is to use subtle sounds that are almost inaudible to the users and discover if this suppression effect keeps happening even without users being aware of it. As a result, the virtual experience would not be interrupted by the sounds and the suppression effect could be used more easily in any type of application.

3 Methodology

In this section, we address the procedure followed to perform the user studies as well as specify the hardware employed. Firstly, a pilot experiment is carried out to analyze how volume can affect the ASE. In this first experiment, two volume settings are tested with the limit frequency values obtained through an auditory frequency test in which we adjust the frequencies to be barely audible for each participant. According to these first results, the most promising value for the volume variable is set for study 2, where hearing limit frequencies are

Fig. 1 Complex living room where participants perform the experiments. The participant can move in the area behind the couch, which has a real physical space available of 180 × 480 cms. The scene is obtained from Barking Dog Unity 3D

further sampled by considering also in- and out-of-hearing range values.

3.1 Environment

The virtual environment (VE) where participants are immersed is a complex living room displayed in Fig. 1. Both experiments have been performed in this scenario and developed using the Unity game engine.

Background sounds are also included in the virtual scene to create a more complex environment closer to an application case and to avoid habituation effects during the experiments. Similarly to Malpica et al. [34], we include bird sounds and a football podcast as atmospheric auditory information. Birdsong sound comes from outside the windows, while the radio podcast is played from the speakers next to the TV, on both sides beneath it. Both background sounds are played throughout the whole experiment.

3.2 Hardware

The virtual experience was provided using an HTC VIVE Pro Eye 2 headset. It offers a nominal FoV of 110 visual degrees (3.5 inches OLED screen and 1440 × 1600 pixel resolution per eye) and a frame rate of 90 Hz. It also includes high-impedance headphones that can reproduce the full spectrum of high-resolution audio. This headset supports eye-tracking technology, in this case, powered by Tobii. This eye tracker has a frame rate of 120 Hz and a precision of one visual degree. The HMD is calibrated for each participant by using a six-dot calibration before any experiment.

3.3 Multimodal stimuli conditions

As presented in Malpica et al. [34], we keep generating visual-only, auditory-only and bimodal stimuli throughout the experiment to analyze the suppression effect. Visual stimuli are aimed to detect visual acuity to settle a baseline of normal visual performance. The purpose of the auditory stimuli was originally to avoid habituation effects toward bimodal condition on the participants. In our case, and given that we work with auditory information around the limits of the hearing range, the purpose of the auditory stimuli is also to set a hearing acuity baseline when auditory information is presented standalone. Finally, bimodal condition is used to test the ASE by comparing their results with the other two stimuli modalities previously mentioned.

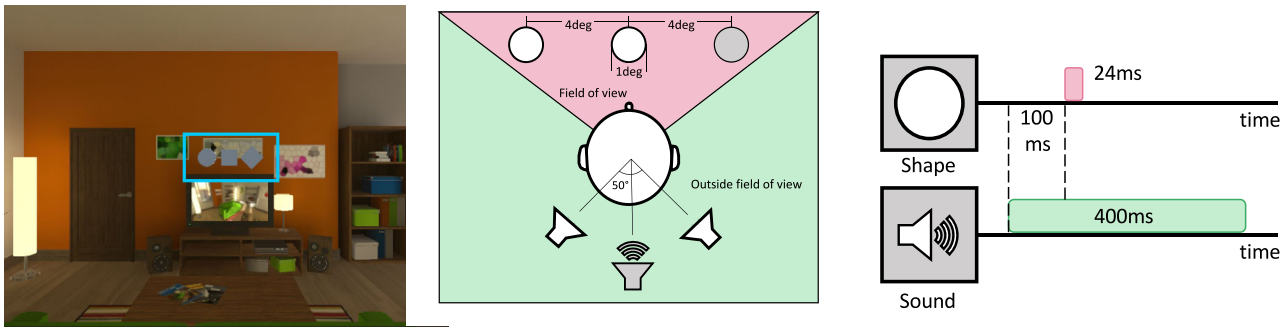
3.3.1 Visual stimulus

Regarding the visual information presented to the user, visual targets are generated in three fixed locations inside the user field of view (FoV). The central location is placed in the middle of the FoV, at its equator line, while the eccentricity of each side location is four visual degrees (see Fig. 2b). The size of all visual targets is one visual degree. Regarding spawning time, all visual targets have a lifetime of 24 ms. Since the statistical analysis carried out by Malpica et al. did not show a significant effect of the target shape or its position on the ASE, we chose to randomize these factors to keep the size of our experiment tractable. Therefore, visual targets can appear in any of these fixed locations during the experiment, while the shape is also randomly selected among three different options: circle, square, or rhombus. These visual targets can be found in Fig. 2a.

3.3.2 Auditory stimulus

Moving into the sound-related part, hearing information is also generated in three fixed locations but in this case outside the user's FoV. The central location is exactly behind the user at 0.2 ms. Each side location is located using a rotation of 50° from the center sound source and always at a distance of 0.2 ms. This can be properly seen in Fig. 2b. These auditory sources play a sound that lasts 400 ms, and after that period, the sound stops.

Contrary to previous work, the audio sources included in our experiments are pure frequency tones located at the limits of the user's hearing range, consequently obtaining barely audible sounds. Details and more information regarding these frequency tones used are explained in further sections, while all participants needed to perform the frequency test included in the Supplementary Material (S1). Considering also Pink noise for comparison reasons with previous work, we handle a total of five auditory cues used in study 1 while seven are considered throughout study 2.



(a) Visual Targets (b) Spatial Scheme (c) Temporal Profile

Fig. 2 **a** Simple shapes available as visual targets, composed of a circle, a square, and a rhombus. All appear enlarged for a correct visualization. In both visual and bimodal stimuli conditions, the visual targets have a lifetime of 24 milliseconds. **b** Spatial configuration for both visual and auditory cues. Visual targets have a size of one visual degree while the eccentricity of the side targets is four visual angles, all of them are presented inside the user FoV. On the other hand, auditory cues are located outside the user FoV, at a constant distance of 0.2 ms from the user and the side sounds displaced 50° from the central one. A bimodal condition example is represented with the right target and the central audio location. **c** Temporal setup followed in bimodal condition. After being the hearing cue played for 100 ms, the visual target spawns and disappears after 24 ms. When the auditory source has been playing for 400 ms, it stops thus ending the bimodal stimuli. **b** and **c** adapted from Malpica et al. [34]

3.3.3 Bimodal stimulus

In the last considered condition, both hearing and visual cues are combined to generate the bimodal stimuli, which involve both vision and hearing modalities. While audio sources and visual targets are spatially incongruent, since they are generated outside and inside the user FoV, respectively, it is important to keep temporal consistency according to the temporal profile displayed in Fig. 2c. As can be noticed, the target appears after the audio source starts to sound and then disappears while the auditory cue keeps being played. The position of the auditory and visual stimuli as well as the shape of the visual target is randomized, while each auditory source is balanced across trials. For more information we refer the reader to the procedure of study 1 and study 2.

4 Study 1: volume and frequency

The goal of study 1 was to test how sound intensity could affect the ASE as well as to observe participants' behavior during the virtual experience. Therefore, this study 1 was aimed at fixing sound variables and checking the correct operation toward study 2. In this first experiment, participants were told to explore the virtual room while trying to detect the stimuli generated throughout the experience. Once any kind of stimuli was detected, the participants had to notify the experimenter what was perceived: a visual shape (visual stimuli), any kind of sound (auditory stimuli), or both at once (bimodal stimuli).

4.1 Participants

Study 1 was performed by a total of 20 participants (ages 21–26, eight defined as women and none as non-binary). All of the participants had normal or corrected to normal vision. On the other hand, none of them reported having any hearing problems. Nineteen participants reported having heard of VR before, while 15 participants reported having used VR in the past. All participants were naive to the purpose of the experiment. The study procedure was approved by the corresponding ethics committee.

4.2 Procedure

The experiment consisted of 45 different trials, 15 trials per stimuli condition (visual-only, auditory-only, and bimodal). The participants were told to explore the area behind the couch by walking within the available region and looking around. The order of the trials was randomized to account for order effects. Whenever the participants perceived any stimuli, they pressed the trigger button on the controller to pause the experiment and then notify the experimenter about what was perceived. Afterward, the next trial was generated after some random time to avoid learning and prediction effects. The experiment ended when all trials were presented to the user.

Two factors were tested in study 1: volume (or sound intensity) and frequency (or sound type). Two different volume values were used in hearing cues to explore how sound intensity could influence the ASE. The auditory cues (pure frequency tones and pink noise) and the background sound

Table 1 Frequency test results obtained in the pilot experiment

ID	Lower limit ($F1$)	Upper limit ($F2$)
01	200	13,500
02	200	13,300
03	150	11,800
04	120	12,650
05	150	13,200
06	100	15,500
07	200	14,600
08	200	10,200
09	200	13,000
10	100	15,700
11	200	13,000
12	200	14,200
13	300	14,800
14	200	12,000
15	200	15,000
16	200	14,300
17	300	13,000
18	200	12,700
19	200	8,300
20	200	11,000

Both the lower and upper bounds of the hearing range per participant can be found. All values are measured in Hertz (Hz)

had different volumes: tested volumes for the auditory stimuli were 55 dB ($V1$, lower) and 75 dB ($V2$, higher), while the background noise (radio podcast and the birdsong) intensity was set to 40–45 dB on average. Regarding frequency, since our goal is to test frequency values at the limits of the hearing range such that the ASE remains minimally invasive, we use two frequencies that account for the lower ($F1$) and upper limits ($F2$) of the hearing range for each participant. Results obtained for the frequency test aiming to obtain $F1$ and $F2$ are shown in Table 1.

Pink noise results at both volumes ($V1$ and $V2$) are combined and used for comparison with previous work. This means study 1 had 5 different auditory stimuli and each one appeared once in each location thus making up for 15 bimodal trials. The auditory-only and visual-only conditions also have 15 trials each: the auditory-only condition has the same balanced 15 sounds as the auditory part of the bimodal conditions, while the variables of the visual-only condition are randomized.

4.3 Results

Data collected throughout the study 1 provide information regarding performance in detecting and recognizing visual targets presented to the user, which could be accompanied or

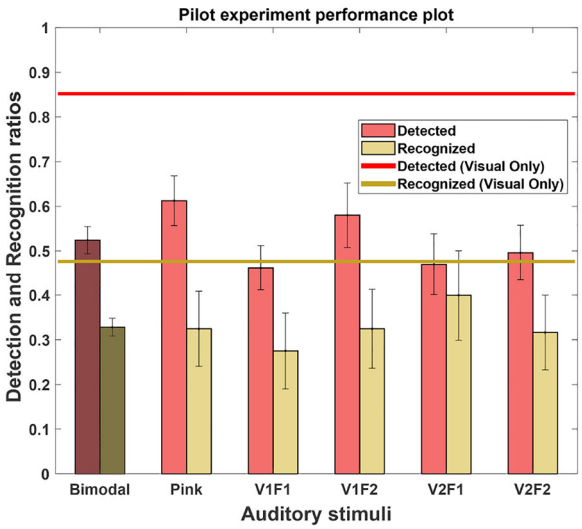


Fig. 3 Study 1 results. Detection (red) and recognition (yellow) ratios for the bimodal condition (separated by sound source types) and visual-only condition (horizontal lines). Error bars on top stand for standard error of the mean (SEM) in all figures unless otherwise specified

not by auditory information. The results of the study 1 for the visual-only and bimodal stimuli are shown in Fig. 3. We can see that there is a decrease in visual performance when sound is present, following the trend observed by Malpica et al. The performance values achieved when no sound was included (visual-only condition) had a detection ratio of 0.852 ± 0.0233 (standard error of the mean) and a recognition ratio of 0.476 ± 0.0274. Meanwhile, bimodal condition had a mean detection ratio of 0.5241 ± 0.0304 and a recognition ratio of 0.3279 ± 0.0201. Details of detection and recognition ratios of bimodal conditions sorted by sound type can be found in Fig. 3. Considering the bimodal condition with different volumes and frequencies, slightly higher differences are found to be associated with the lower volume ($V1$), suggesting the effect has more influence with lower-intensity sounds.

Results for study 1, related to sound intensity values, suggested that the setup in which the suppression effect was slightly more pronounced happened when using the lower intensity value $V1$ (55 dB). We hypothesize that loud enough sounds can act as a warning signal so the participant's awareness can be increased for this alarmed state. Following this hypothesis, the results of the study 1 and in order to keep the size of study 2 tractable, we decide to only use $V1$ for our stimuli.

The ASE still appeared indeed but the magnitude of the suppression effect was not as big as expected (considering the work of Malpica et al. as a baseline) when comparing visual and bimodal conditions. One reason for this difference could be related to how participants were instructed to behave during the experiment: they reported being more focused on finding stimuli than on exploring the scene. To alleviate this

behavior, we decided to include another task in study 2 in order to increase the cognitive load and to encourage a proper exploration of the virtual environment (see the procedure of study 2 for additional details).

5 Study 2: exploring hearing range frequency limits

Once the most promising sound intensity value is set according to study 1 results, we further explore frequency limits in study 2 explained in this section.

5.1 Participants

Another group of 20 people carried out study 2 (ages 21–28, nine defined as women and none as non-binary). Also, all reported to have normal or corrected to normal vision and not having any hearing condition. All of them reported having heard of VR before while 18 participants reported having used VR in the past. More information about the surveys and questionnaires, both questions and results, presented to the participants can be found in the supplementary material (section S.2). The study procedure was approved by the corresponding ethics committee.

5.2 Procedure

Study 2 consisted of a total of 42 trials evenly distributed through the three tested conditions: 14 visual-only, 14 auditory-only and 14 bimodal. As in study 1, participants are instructed to explore the area behind the couch as much as possible by walking and looking around, while training trials are also presented to the participants so they can get used to the virtual environment and the stimuli before starting. During some free time, they were able to perceive different visual target shapes as well as hearing cues, while exploring the virtual environment. Generally, participants decided to start the experiment before the tenth training mock case. Once the real experiment phase starts, the trials are presented in random order and the procedure is the same as the previous user study: participants have to detect, and recognize when appropriate, the perceived stimuli.

In this experiment, volume is fixed according to study 1 results which is 55 dB ($V1$) for all cases, while the influence of sound frequency on the ASE is further studied. Consequently, more frequency values associated with each hearing limit are tested. For each frequency limit, lower bound ($F1$) and higher bound ($F2$), new values are considered located inside and outside the hearing range aiming to observe the ASE behavior regarding these extreme regions for users' hearing perception. Regarding the lower bound, these in- and out-range values are F1U and F1D, respectively, being the F1D smaller and F1U higher the $F1$ lower limit, both reduced and increased, respectively, by 40 Hz. On the other hand, the same procedure is done regarding the higher bound $F2$. In this case, the in- and out-range values are F2D and F2U, respectively, being the F2D smaller and F2U higher the $F2$ upper limit, both reduced and increased, respectively, by 200 Hz. The difference in these intervals is considering human hearing sensitivity [36]. Results for the frequency tests performed to obtain these frequency limit values are shown in Table 2. More details about these new values associated to each frequency limit can be found in the supplementary material (section S.1). Therefore, there is a total of 7 audio sources available, 6 related to the frequency limits, which are obtained for each participant using the frequency test, and the pink noise for comparison reasons. Each audio source appeared twice while randomizing the location of the audio-visual cues and also the type of the visual target. This makes a total of 14 bimodal trials, while auditory-only and visual-only conditions also have 14 trials each. The auditory-only condition has the same balanced 14 sounds as the auditory part of the bimodal conditions, while the visual-only location and type variables are fully randomized.

Table 2 Frequency test results obtained in study 2

ID	$F1$	F1D	F1U	$F2$	F2D	F2U
21	40	20	80	13,900	13,700	14,100
22	60	20	100	15,300	15,100	15,500
23	60	20	100	17,900	17,700	18,100
24	60	20	100	14,000	13,800	14,200
25	100	60	120	14,800	14,600	15,000
26	80	40	100	12,800	12,600	13,000
27	60	20	100	15,700	15,500	15,900
28	60	20	100	17,200	17,000	17,400
29	60	20	100	15,200	15,000	15,400
30	100	60	140	16,600	16,400	16,800
31	60	20	100	13,000	12,800	13,200
32	60	20	100	15,400	15,200	15,600
33	60	20	100	15,000	14,800	15,200
34	60	20	100	15,900	15,700	16,100
35	160	120	200	15,400	15,200	15,600
36	80	40	120	12,900	12,700	13,100
37	60	20	100	17,600	17,400	17,800
38	40	20	80	16,500	16,300	16,700
39	40	20	80	15,500	15,300	15,700
40	60	20	100	15,300	15,100	15,500

Both the lower and upper bounds of the hearing range per participant can be found, as well as the respective close values inside and outside the hearing range. All values are measured in Hertz (Hz)

Regarding the higher-cognitive-load task included, participants are told to mainly look for scene changes while detecting spawning stimuli.

These changes are attached to some scene objects that keep disappearing and appearing again every 30 s. The scene objects selected to disappear are a couple of drawings, two chairs, a small lamp, a flowerpot and a bigger lamp that can be seen in Fig. 1. These objects were selected to be easy enough to find when disappearing without being extremely obvious, as it would have been if using bigger objects like the TV or the couch. When noticing a change in the scene, participants verbally reported so to the experimenter. We do not report any performance metric associated with this task since it was intended to ensure scene exploration only.

6 Results and analysis

In this section, we present and analyze the results obtained in study 2 (Sect. 5). For this purpose, we perform an aligned rank transform [68] as a previous step before running an ANOVA analysis between three modes: visual-only, bimodal with perceived auditory part (bimodal listened), and bimodal where the auditory part was not perceived (bimodal not listened). This mode, the auditory source type (seven levels: Pink, F1D, $F1$, F1U, F2D, $F2$ and F2U), auditory location (three levels; left, center or right), visual location of the targets (three levels: left, center or right), and visual shape (three levels: circle, square or rhombus) were the fixed effects, while the user ID was modeled as a random effect. Two analyses were carried out: one with detection as the response variable and another with one with recognition as the response variable.

First, we report the detection ratios of the auditory stimuli considering both bimodal and auditory-only conditions, which are displayed in Fig. 4.

The mean perception rate observed for all audio sources used is 0.3678 ± 0.0229, meaning that the auditory stimuli were not perceived in more than half of the trials, which suggests that the auditory stimuli were difficult to perceive as expected. Without taking pink noise into account, the other audio sources are usually perceived in less than half of their appearances, and the F1D audio source (below the lower limit of hearing frequency) is practically inaudible. As can also be expected, those values inside the hearing range (F1U and F2D) have higher ratios than the limits themselves ($F1$ and $F2$), which are also higher than the values outside the hearing range (F1D and F2U).

Results associated with the detection task can be found in Fig. 5. We differentiate between bimodal trials in two modes, where the audio was listened and those where the audio was not listened. These two bimodal modes are defined since we aim to find differences between them, considering that sometimes the barely audible audio sources are not listened

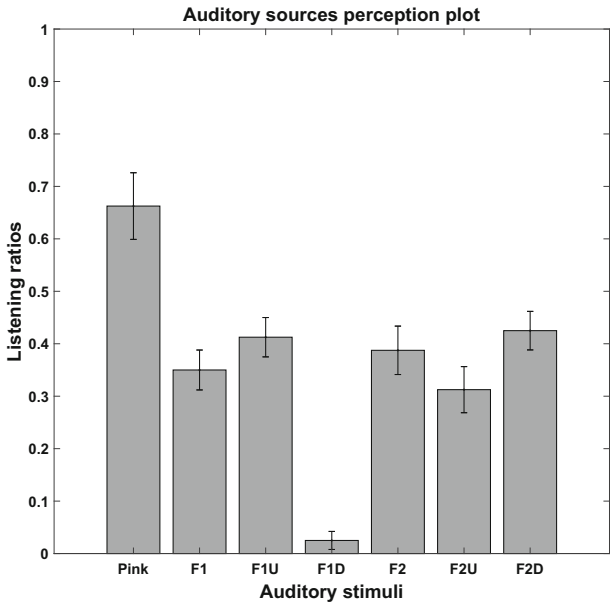


Fig. 4 Audio sources perception plot. Horizontal axis: Each type of audio source that was used throughout the experiment. Vertical axis: Means of audio listening ratios, considering both auditory-only and bimodal conditions

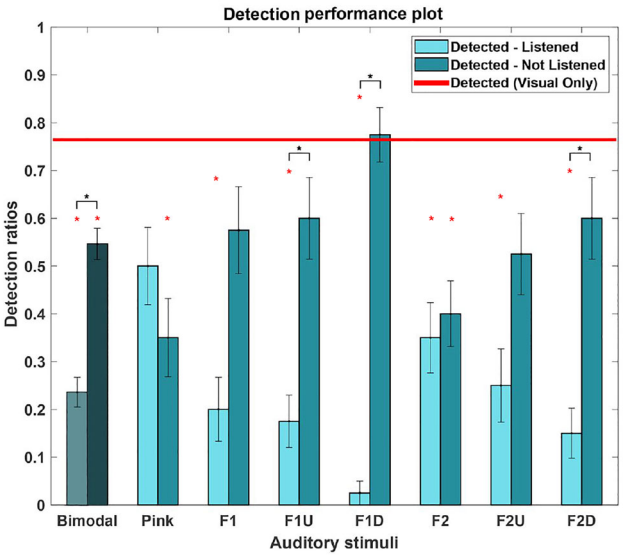


Fig. 5 Detection performance plot. Horizontal axis from left to right: mean detection for all bimodal conditions and detection ratios separated by auditory source type of the bimodal stimuli. The bimodal conditions where the auditory stimuli were listened (light turquoise) or not (deep turquoise) are differentiated. For comparison, the visual-only detection rate is also shown (red horizontal line). Red stars mark significant differences between each bimodal mode (listened and not listened) and the visual-only mode, while black stars mark significant differences between both bimodal modes

Table 3 Stimuli modes contrast regarding detection task (post-hoc analysis of the mode factor for the aligned rank transform ANOVA)

Modes contrast	*t* ratio	*p* value	Sig.
Visual only—bimodal (listened)	7.277	< 0.0001	****
Visual only—bimodal (not listened)	3.386	< 0.001	***
Bimodal (listened–not listened)	−7.781	< 0.0001	****

We can see that the three levels are significantly different from each other. Significance levels are indicated with stars following *** and **** for $p < 0.001$ and $p < 0.0001$ respectively

to. Additionally, even when these subtle auditory cues are listened to, we want to observe if the visual performance is still affected which may be the most promising scenario by achieving visual suppression without being invasive. The detection rate of the visual-only condition is 0.764, while the mean of bimodal condition with listened sound is 0.2357 ± 0.0311. The mean detection rate of bimodal condition when the auditory information was not listened is 0.546 ± 0.0328. We can see a significant effect of the MODE factor ($F(2,266) = 12.2468$, $p < 0.0001$) with significant differences between all three levels (Table 3). The effect size, calculated as partial eta squared (η_p^2), was 0.084 thus indicating a medium effect. We can see a decrease in detection performance in the presence of sound, which confirms the existence of the ASE. Additionally, this difference on detection between the visual-only and bimodal conditions is similar to the findings reported by Malpica et al., which suggests that the inclusion of the additional task worked as intended. The visual detection rate was higher in general for the bimodal condition when the auditory part was not listened for all tested audio sources except for pink noise. Regardless of if the auditory part of the bimodal stimuli was listened or not, the ASE can still be triggered, which suggests minimally disruptive sounds can be used to alter visual perception in immersive environments. We also find a significant interaction between MODE and the auditory source ($F(6,266) = 8.9685$, $p < 0.0001$, $\eta_p^2 = 0.1682$, large effect). Table 4 shows the results of the post-hoc analysis for each auditory source.

We note that all limit frequency audio sources ($F1$, F1U, F1D, $F2$, F2U, and F2D) have lower detection ratios than the visual-only condition when auditory cues are listened. This indicates these sounds when attended to, distract from or impair visual detection tasks, showing the appearance of the ASE. Also, for sounds not listened to ("Not Listened" mode), Pink and F2 show a significantly lower detection rate than the visual-only condition, suggesting that even without direct attention, these sounds can still impact visual detection negatively.

Moving into the visual recognition task, recognition ratios obtained in study 2 can be found in Fig. 6. Recognition trials are calculated over detected stimuli such that a recognition rate of 1 means that every detected trial was also recognized. In this case, the recognition rate of the visual-only condition is 0.514. We see a decrease in performance for the bimodal (listened) condition to 0.1556, while the mean recognition rate when the auditory information was not listened is 0.306. Details for each type of audio source can be seen in Fig. 6. As happened with the detection task, we report how visual recognition performance is decreased for the bimodal condition, where auditory information was listened, with respect to the visual-only condition. These results are in line with previous works, suggesting the appearance of the suppression effect also in recognition tasks. We do not find a significant effect of the MODE factor ($F(2,266) = 1.3718$, $p = 0.2554$, $\eta_p^2 = 0.0102$, small effect), but we do find a significant interaction effect between MODE and the auditory source ($F(6,266) = 4.8937$, $p < 0.0001$, $\eta_p^2 = 0.0994$, medium effect). Results for the ANOVA analysis between modes using the aligned rank

Table 4 Post-hoc analysis of the interaction between the mode factor and the sound source for the detection task

Sound	Modes	*t*	*p* value	Sig.
	Listened–visual	3.1847	0.1053	
Pink	Not listened–visual	4.7148	0.0003	***
	Listened–not listened	1.5300	1	
	Listened–visual	−6.2189	< 0.0001	****
*F*1	Not listened–visual	−2.4456	0.7855	
	Listened–not listened	−3.7732	0.1529	
	Listened–visual	−6.5084	< 0.0001	****
F1U	Not listened–visual	−2.4715	0.7461	
	Listened–not listened	−4.0368	0.0056	**
	Listened–visual	−8.0643	< 0.0001	****
F1D	Not listened–visual	−0.2024	1	
	Listened–not listened	−7.8619	< 0.0001	****
	Listened–visual	−4.688	0.0004	***
*F*2	Not listened–visual	−3.923	0.0086	**
	Listened–not listened	−0.7650	1	
	Listened–visual	−5.7434	< 0.0001	****
F2U	Not listened–visual	−3.1847	0.1053	
	Listened–not listened	−2.5586	0.6305	
	Listened–visual	−6.746	< 0.0001	****
F2D	Not listened–visual	−2.2079	1	
	Listened–not listened	−4.5382	0.0007	***

Holm-Bonferroni correction method was used, while significance levels are indicated with stars following **, ***, **** for $p < 0.01$, $p < 0.001$, and $p < 0.0001$, respectively

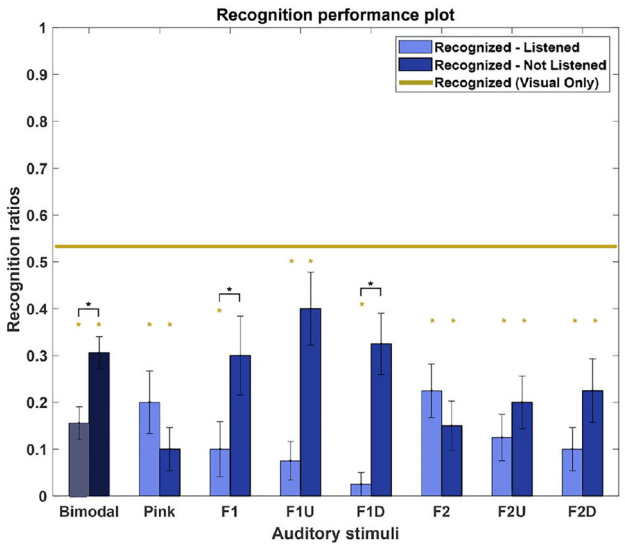


Fig. 6 Recognition performance plot. Horizontal axis: Auditory sources sorting the bimodal stimuli generated in the experiment and the mean of all of them (Bimodal). The bimodal conditions where the auditory stimuli were listened (light blue) or not (deep blue) are differentiated, while the visual information was always recognized. Vertical axis: Visual information recognition ratios for bimodal and visual-only conditions by participant. The yellow horizontal line accounts for the recognition rate of the visual-only condition. Yellow stars mark significant differences between each bimodal mode (listened and not listened) and the visual-only mode, while black stars mark significant differences between both bimodal modes

Table 6 Post-hoc analysis of the interaction between the mode factor and the sound source for the recognition task

Sound	Modes	t	p value	Sig.
	Listened–visual	4.7703	0.0003	***
Pink	Not listened–visual	5.9528	< 0.0001	****
	Listened–not listened	1.1825	1	
	Listened–visual	−6.3009	< 0.0001	****
$F1$	Not listened–visual	−2.4051	1	
	Listened–not listened	−3.8958	0.0114	*
	Listened–visual	−6.1627	< 0.0001	****
F1U	Not listened–visual	−3.7975	0.0164	*
	Listened–not listened	−2.3651	1	
	Listened–visual	−6.9971	< 0.0001	****
F1D	Not listened–visual	−3.0298	0.2231	
	Listened–not listened	−3.9673	0.0087	**
	Listened–visual	−4.2123	0.0032	**
F2	Not listened–visual	−5.2566	< 0.0001	****
	Listened–not listened	1.0443	1	
	Listened–visual	−5.6047	< 0.0001	****
F2U	Not listened–visual	−4.5604	0.0007	***
	Listened–not listened	−1.0443	1	
	Listened–visual	−5.9528	< 0.0001	****
F2D	Not listened–visual	−4.4222	0.0013	**
	Listened–not listened	−1.5306	1	

Holm–Bonferroni correction method was used while significance levels are indicated with stars following *, **, ***, **** for $p < 0.05$, $p < 0.01$, $p < 0.001$ and $p < 0.0001$ respectively

transform are shown in Table 5, while results of the post-hoc analyses for each auditory source can be found in Table 6.

The post-hoc results (Table 6) show that all limit frequency audio sources ($F1$, F1U, F1D, F2D, $F2$, and F2U) have significantly lower recognition ratios compared to the visual-only condition when they are listened to, as happened with detection task. This indicates a notable distraction or interference effect of these sounds on visual recognition tasks when attention is given to them. Additionally, Pink, F1U, $F2$, F2U, and F2D sounds also show significantly lower recognition ratios compared to the visual-only condition even when they are not listened to, suggesting that the presence of these sounds, even without focused attention, can still impact visual recognition negatively.

Results for the remaining factors obtained from this statistical analysis can be found in Table 7. As can be observed, there is no significant influence between different sound types toward the effect on detection or recognition tasks in the bimodal condition, which suggests the tested locations and sources are a valid set where the ASE works. There is a significant effect of shape in the recognition task for the visual-only condition, which is in line with the work of Malpica et al. We can also see how the location and type of the visual targets do not influence the effect either, which is also in line with the results reported by Malpica et al. Similarly, auditory location has been observed not to influence the ASE in any case. Finally, we find a significant effect of the auditory source in the auditory-only condition, which suggests that some sounds are easier to perceive than others in the absence of a visual target.

Table 5 Stimuli modes contrast regarding recognition task (post-hoc analysis of the mode factor for the aligned rank transform ANOVA)

Modes contrast	t ratio	p value	Sig.
Visual only—bimodal (listened)	7.337	< 0.0001	****
Visual only—bimodal (not listened)	5.397	< 0.0001	****
Bimodal (listened–not listened)	−3.880	< 0.001	***

Similarly to the detection task, the three levels are significantly different from each other. Significance levels are indicated with stars following *** and **** for $p < 0.001$ and $p < 0.0001$ respectively

Table 7 Statistical analysis for the visual location, visual shape, auditory location and auditory source factors for the bimodal, visual-only and auditory-only conditions

Condition	Visual task	Fixed effect	F	p value
Bimodal	Detect	Vis. location	1.2925	0.2773
		Vis. shape	2.1880	0.1154
		Aud. location	0.4765	0.6218
		Aud. source	0.3951	0.8813
	Recognize	Vis. location	0.3483	0.7063
		Vis. shape	0.3086	0.7348
		Aud. location	0.1291	0.8789
		Aud. source	0.1645	0.9857
Visual	Detect	Vis. location	0.3875	0.6791
		Vis. shape	0.4932	0.6112
	Recognize	Vis. location	0.4053	0.6671
		Vis. shape**	6.5059	0.0017
Auditory	–	Aud. location	0.0636	0.9383
		Aud. source****	12.2752	< 0.0001

Significance levels are indicated with stars following **, **** for $p < 0.01$ and $p < 0.0001$, respectively

7 Discussion

In previous sections, we have affirmed how the ASE can still take place when subtle sounds are used, *even when those sounds are not consciously listened by the participants*. As we can see in Figs. 5 and 6, including subtle auditory cues that follow a specific spatiotemporal setup (Fig. 2b, c) degrades visual performance in both detection and recognition tasks. These results follow the same trends and meet the conclusions reported by Malpica et al. [34].

To compare our results with the ones reported in Malpica et al. [34], we need to refer to the Pink sound only for a fair analogy. Firstly, regarding detection task, Malpica et al. reported 82.07% ± (4.81%) in the visual-only condition and 20.02% ± (4.86%) for bimodal condition using pink noise. We report a detection rate of 76.4% in the visual-only condition and 39.05% for the bimodal condition with pink noise. Detection rates for visual-only stimuli are similar enough to claim that participants were able to detect visual information when presented standalone, replicating in this sense previous work conditions. We believe that the difference in pink noise bimodal trials could be due to differences in our setup (such as the volume) and random effects such as the participants' familiarity with VR and videogames.

Moving into visual recognition tasks rates, Malpica et al. reported a 59.93% ± (6.76%) in the visual-only condition and a 7.93% ± (4.12%) for the bimodal condition, again using pink noise. In our experiment, we report a recognition rate of 53.3% in the visual-only condition and 23.08% for the bimodal condition with pink noise. Consequently, the same trends are found in both conditions, although we observe a smaller effect which could be caused by differences in audio, participants, etc.

Addressing the barely audible audio sources related to the hearing limits frequencies in the bimodal condition, the ratio for bimodal (listened mode) in the detection task is 23.5% while a 15.5% is achieved regarding the recognition task. These results support that the ASE can still be triggered by subtle sounds, in particular with pure frequencies close to the hearing limits of each participant. But there is still more: the ASE is triggered even when the participants *do not listen to the auditory cue*. This means that we could trigger the suppression effect without the users' awareness. This is a more subtle and versatile effect that can be used in the majority of scenarios in real VR applications. This aspect could potentially allow us to perform slight scene changes while the user has no hints of it. Figure 4 also supports this idea, showing the lower detection rates for these frequency-related audio sources.

As a consequence of previous results, we can allege that the ASE is robust enough to happen in both visual detection and recognition tasks, even with barely audible auditory cues. In our study, participants performed a high-demanding cognitive task, like looking for scene changes, thus bringing the suppression effect closer to a possible real application in a more complex scenario, such as training sessions or other demanding tasks. According to the additional statistical analysis carried out, we found that the ASE has no relation with the tested visual target types or spatial locations, either the spatial location of auditory sources. Previous works on spatial location suggest that there is a crossmodal dependence in exogenous orienting where audition positively influences vision, but not vice versa [55]. Following Spence et al.'s discussion about seemingly incongruent effects with previous research, our suppression effect could be related to the response-priming paradigm [31], where an inconsis-

tent prime can elicit the wrong response, particularly at long stimulus onset asynchronies (SOAs) of up to 100ms, which is consistent with our temporal setup. (All auditory sources appear 100ms before the visual target onset.) These effects are independent of visual awareness of the prime [52] and, in our case, we would have an auditory prime which triggers a decrease in visual performance, whether the auditory cue is perceived or not. Regarding the location of the visual and auditory parts of the bimodal stimuli, previous works have shown a bias toward the center when localizing visual stimuli, while auditory localization is biased toward the periphery [40]. In their work, Odegaard et al. proposed that the bias in one modality dominated the other with the visual modality being stronger for audiovisual events in a location task. However, even when all our visual targets are located toward the center of the field of view, participants fail to correctly detect them in the presence of sound. Different cognitive tasks (we do ask participants to detect or recognize stimuli, not to locate them) may modulate this modal dominance. Audiovisual effects are indeed complex and depend on users perceiving two different auditory and visual events as a single audiovisual event due to their temporal co-occurrence. If participants are perceiving a single audiovisual object their attention can be spread between the auditory and visual modalities, potentially hindering performance in any of them. Busse et al. show that visual attention can modulate the processing of irrelevant, spatially discrepant auditory stimuli, enhancing auditory processing [9]. This diversion of resources may in turn decrease visual performance. Additionally, Chiou et al. report that high and low tones may induce attention shifts to upper or lower locations depending on the pitch, which would deviate attention from the central plane where the visual targets appear [13]. In our case, we are using both higher and lower pitches, which may influence the areas of the environment that participants focus on during our studies. In summary, a co-occurrent auditory event with a SOA of 100ms, spatial incongruency and irrelevant pitch may deviate visual attention from the center of the field of view, causing a significant decrease on visual performance. These crossmodal spatial attention links were also observed by Driver et al. [17], reporting how spatially attending at one modality induces attention shifts in other modalities.

From the statistical analyses performed, visual location has been found to not affect the ASE, in line with previous results. Research has been done on how visual acuity [7], orientation [65], attention [61], and perception [26] relate to visual eccentricity. Visual capabilities tend to degrade toward the periphery, thus originating rapid ballistic eye movements known as visual saccades as a result. All our visual targets are presented in the central plane, aligned with the direction of the participant in the virtual environment or at 4°eccentricity. We believe we follow a conservative approach since visual targets should be more difficult to detect in the peripheral region. No differences have been found between the three different simple shapes employed, as also reported by Malpica et al. In this sense, visual saliency and acuity are affected by visual contrast, luminance, target size and color [19, 28, 46, 54]. Our visual targets have a distinct appearance from the overall virtual environment so that they are easier to perceive, and the relative size is similar regardless of the particular shape of the visual target. However, it is important to note that we have only studied a subset of the possible locations and shapes for the visual target and that any possible extrapolation should be validated. Moving into auditory cues, no differences have been found in sound locations, similar to Malpica et al. The reason behind this may lie in the fact that incongruent sounds tend to be ignored in visually driven tasks if they are perceived as distractors or irrelevant to the task [58]. Spatial incongruency between audiovisual cues favors selective attention toward visual cues instead [59], similar to the cocktail party effect [4]. It is certainly difficult to assess whether the auditory part of the bimodal stimulus is deviating visual attention, if perceiving an audiovisual event is modulating auditory processing or if something entirely different is responsible for the ASE effect. In any case, it is clear that the audiovisual stimuli that we present in our experiments cause a crossmodal effect that results in the degradation of visual performance.

8 Limitations and future work

As in any VR study, the hardware used can somewhat influence results. For example, the resolution of a display and the field of view that a headset covers varies between models. Additionally, some visual artifacts are also present like lens-induced distortions or the screen-door effect [2]. While newer headsets may further minimize such artifacts, the size of visual targets is large enough to be clearly visible in the experiment procedure while the use of the same headset across all conditions ensures consistency. Therefore, we do not expect these factors to compromise the robustness of our findings.

Regarding audiovisual stimuli used in our studies, visual targets are limited to simple shapes while hearing cues are pure frequencies or noise. None of them present any semantic information related to the scene. In this direction, new targets such as more complex shapes, dynamic objects, or even videos could be used to analyze the ASE. Using scene objects as visual targets could be an option as suggested by some participants but those cases would be even harder to perceive. Adding color could introduce a new variable to consider, since higher contrast colors may be more difficult to miss thus reducing the ASE effect. Moving objects or temporal-changing images could be used to extend the problem to 360° panoramas that could be affected differently. Regarding the auditory aspect, further sound types could be added to the

experiment as auditory sources. Hearing cues related to the scene could be used to trigger the ASE, taking advantage of more meaningful sounds that were heard but unnoticed by the participants. Ideally, the ASE could also take place with specific sounds of the scene such as the doorbell. Therefore, the most straightforward direction would be exploring if more complex or dynamic scenes could be employed to perform the experiment. Having not only the radio podcast but with some audiovisual events taking place in the scene that may attract participants' attention could even improve the potential of the ASE.

Next steps could also consider participants' moods or phobias to create a more personalized and engaging experience, aiming to find out its influence on the suppression effect if any. Following this idea, since the ASE does not seem to have oculomotor causes (see the work of Malpica et al.), brain signals could be studied more properly with adequate hardware. New senses could also be included to look at higher-sensory-level suppression effects. It would also be a good practice to test the ASE in a final VR application, like for example fusing the suppression effect with an existing redirected walking technique.

Lastly, a more diverse and bigger sample of participants could be tested to see how the ASE is affected by different demographics, including age and familiarity with VR. In this sense, the G*power tool has been used to calculate the required sample size for our experiment. Depending on the effect size f, 7 participants ($f = 0.14$, big effect) to 18 participants ($f = 0.06$, mid-sized effect) would be needed to see significant differences. Therefore, small effects that could be only seen with a larger pool of participants or with more trials per participant may be missing.

9 Conclusion

Throughout this work, we research on the scope of the audiovisual suppression effect (ASE) reported and further studied in VR by previous works, exploring in our case if subtle sounds can trigger the inhibitory effect. By carrying out a frequency test, the frequency values associated with the limits of the hearing range for each participant were obtained. Using proper intervals such that auditory sources are located near the frequency hearing limits of each participant, we observe how the ASE can still degrade visual perception under such circumstances. Compared to visual-only condition, visual target detection and recognition in bimodal condition where the hearing source is almost imperceptible are found to be significantly decreased. Consequently, the potential of this audiovisual suppression effect is shown to be useful in realistic VR scenarios. In such environments, unperceivable sounds would impact visual perception without user awareness, thus potentially improving a wide range of applications and techniques. For example, foveated rendering and redirected walking methods applied to 3D scenarios could benefit from our findings. In such cases, visual input would be modified as desired by taking advantage of this ASE without interfering with the experience and the side effects produced by that intrusion, affecting immersion and realism.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00371-024-03707-6.

Acknowledgements This work has been supported by grant PID2022-141539NB-I00, funded by MICIU/AEI/10.13039/501100011033 and ERDF (EU), the Aragon Institute for Engineering Research (I3A) through the Impulso program, and the Government of Aragon's Departamento de Ciencia, Universidad y Sociedad del Conocimiento through the Reference Research Group "Graphics and Imaging Lab" (ref T34_23R).

Author Contributions D.J-N. carried out the research experiments, wrote the main manuscript and prepared all figures. S.M. and A.S. supervised the research. All authors reviewed the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability Additional details about the frequency test and a demo video clip are included in the supplementary files. The generated eye-tracking data is available at https://minimalase.mpi-inf.mpg.de/.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Ahrens, A., Lund, K.D., Marschall, M., Dau, T.: Sound source localization with varying amount of visual information in virtual reality. PLoS ONE **14**, e0214603 (2019)
2. Angelov, V., Petkov, E., Shipkovenski, G., Kalushkov, T.: Modern virtual reality headsets. In 2020 International Congress on Human–Computer Interaction, Optimization and Robotic Applications (HORA). pp. 1–5. https://doi.org/10.1109/HORA49412.2020.9152604(2020)

3. Arnal, L.H., Morillon, B., Kell, C.A., Giraud, A.-L.: Dual neural routing of visual facilitation in speech processing. J. Neurosci. **29**(43), 13445–13453 (2009)
4. Arons, B.: A review of the cocktail party effect. J. Amer. Voice I/O Soc. **12**(7), 35–50 (1992)
5. Baram, Y., Lenger, R.: Gait improvement in patients with cerebral palsy by visual and auditory feedback. Neuromodulation: Technol. Neural Interface **15**(1), 48–52 (2012)
6. Baram, Y., Miller, A.: Auditory feedback control for improvement of gait in patients with Multiple Sclerosis. J. Neurol. Sci. **254**(1), 90–94 (2007)
7. Battista, J., Kalloniatis, M., Metha, A.: Visual function: the problem with eccentricity. Clin. Exp. Optom. **88**(5), 313–321 (2005)
8. Burns, E., Razzaque, S., Panter, A.T., Whitton, M.C., McCallus, M.R., Brooks, F.P.: The hand is slower than the eye: a quantitative exploration of visual dominance over proprioception. In IEEE Proceedings. VR 2005. Virtual Reality 2005, pp. 3–10 (2005)
9. Busse, L., Roberts, K.C., Crist, R.E., Weissman, D.H., Woldorff, M.G.: The spread of attention across modalities and space in a multisensory object. Proc. Natl. Acad. Sci. **102**(51), 18751–18756 (2005)
10. Chao, F.-Y., Ozcinar, C., Wang, C., Zerman, E., Zhang, L., Hamidouche, W., Deforges, O., Smolic, A.: Audio-visual perception of omnidirectional video for virtual reality applications. In 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 1–6 (2020)
11. Chapman, E., Tremblay, F.: Tactile Suppression. Vol. 10 (2015)
12. Chauvel, G., Maquestiaux, F.: Visual illusions can facilitate sport skill learning. Psychon. Bull. Rev. **22**, 717–721 (2014)
13. Chiou, R., Rich, A.N.: Cross-modality correspondence between pitch and spatial location modulates attentional orienting. Perception **41**(3), 339–353 (2012)
14. Coren, S., Girgus, J.S.: Seeing is Deceiving: The Psychology of Visual Illusions. Routledge (2020)
15. Demattè, M.L., Sanabria, D., Sugarman, R., Spence, C.: Cross-modal interactions between olfaction and touch. Chem. Senses **31**(4), 291–300 (2006)
16. Desantis, A., Mamassian, P., Lisi, M., Waszak, F.: The prediction of visual stimuli influences auditory loudness discrimination. Exp. Brain Res. **232**, 3317–3324 (2014)
17. Driver, J., Spence, C.: Attention and the crossmodal construction of space. Trends Cogn. Sci. **2**(7), 254–262 (1998)
18. Eagleman, D.: Human time perception and its illusions. Curr. Opin. Neurobiol. **18**, 131–6 (2008)
19. Engmann, S., Hart, B.M.T., Sieren, T., Onat, S., König, P., Einhäuser, W.: Saliency on a natural scene background: effects of color and luminance contrast add linearly. Attent. Percept. Psychophys. **71**(6), 1337–1352 (2009)
20. Enoch, J., McDonald, L., Jones, L., Jones, P.R., Crabb, D.P.: Evaluating whether sight is the most valued sense. JAMA Ophthalmol. **137**(11), 1317–1320 (2019)
21. Gregory, R.L.: Visual illusions. Sci. Am. **219**(5), 66–79 (1968)
22. Harrar, V., Harris, L.R.: The effect of exposure to asynchronous audio, visual, and tactile stimulus combinations on the perception of simultaneity. Exp. Brain Res. **186**, 517–524 (2008)
23. Hershey, J., Movellan, J.: Audio vision: using audio-visual synchrony to locate sounds. In Solla, S., Leen, T., Müller, K. (eds.) Advances in Neural Information Processing Systems, Vol. 12. MIT Press (1999)
24. Hidaka, S., Ide, M.: Sound can suppress visual perception. Sci. Rep. **5**(1), 10483 (2015)
25. Ide, M., Hidaka, S.: Tactile stimulation can suppress visual perception. Sci. Rep. **3**, 3453 (2013)
26. Ikeda, H., Blake, R., Watanabe, K.: Eccentric perception of biological motion is unscalably poor. Vis. Res. **45**(15), 1935–1943 (2005)
27. Jeka, J., Oie, K., Kiemel, T.: Multisensory information for human postural control: integrating touch and vision. Exp. Brain Res. **134**, 107–25 (2000)
28. Johnson, C.H.R.I.S.A., Casson, E.J., et al.: Effects of luminance, contrast, and blur on visual acuity. Optom. Vis. Sci. **72**, 864–869 (1995)
29. Kammers, M.P.M., de Vignemont, F., Verhagen, L., Dijkerman, H.C.: The rubber hand illusion in action. Neuropsychologia **47**(1), 204–211 (2009)
30. Kim, R.S., Seitz, A.R., Shams, L.: Benefits of stimulus congruency for multisensory facilitation of visual learning. PLoS ONE **3**(1), 1–5 (2008)
31. Klotz, W., Neumann, O.: Motor activation without conscious discrimination in metacontrast masking. J. Exp. Psychol. Hum. Percept. Perform. **25**(4), 976 (1999)
32. Lecuyer, A., Mobuchon, P., Megard, C., Perret, J., Andriot, C., Colinot, J.-P.: HOMERE: a multimodal system for visually impaired people to explore virtual environments. In IEEE Virtual Reality, Proceedings. pp. 251–258 (2003)
33. Malpica, S., Masia, B., Herman, L., Wetzstein, G., Eagleman, D., Gutiérrez, D., Bylinskii, Z., Sun, Q.: Larger visual changes compress time: the inverted effect of asemantic visual features on interval time perception. PLoS ONE **17**, e0265591 (2022)
34. Malpica, S., Serrano, A., Gutierrez, D., Masia. B.: Auditory Stimuli Degrade Visual Performance in Virtual Reality. Scientific Reports (Nature Publishing Group) 10 (2020)
35. Martin, D., Malpica, S., Gutierrez, D., Masia, B., Serrano, A.: Multimodality in VR: a survey. ACM Comput. Surv. **54**(10s), 36 (2022)
36. Masterton, B., Heffner, H., Ravizza, R.: The evolution of human hearing. J. Acoust. Soc. Am. **45**, 966–85 (1969)
37. Matin, E.: Saccadic suppression: a review and an analysis. Psychol. Bull. **81**, 899–917 (1975)
38. Mestre, D., Ewald, M., Maiano, C.: Virtual reality and exercise: behavioral and psychological effects of visual feedback. Stud. Health Technol. Inform. **167**, 122–7 (2011)
39. Micah, M.M., Wallace, M.T.: The neural bases of multisensory processes (2011)
40. Odegaard, B., Wozny, D.R., Shams, L.: Biases in visual, auditory, and audiovisual perception of space. PLoS Comput. Biol. **11**, e1004649 (2015)
41. Patney, A., Salvi, M., Kim, J., Kaplanyan, A., Wyman, C., Benty, N., Luebke, D., Lefohn, A.: Towards foveated rendering for gaze-tracked virtual reality. ACM Trans. Graph. (TOG) **35**, 1–12 (2016)
42. Payne, D.G., Elie, C.J., Blackwell, J.M., Neuschatz, J.S.: Memory illusions: recalling, recognizing, and recollecting events that never occurred. J. Mem. Lang. **35**, 261–285 (1996)
43. Payne, D.G., Elie, C.J., Blackwell, J.M., Neuschatz, J.S.: Memory illusions: recalling, recognizing, and recollecting events that never occurred. J. Mem. Lang. **35**, 261–285 (1996)
44. Picinali, L., Jaco, A.A., Denis, M., Katz, B.: Exploration of architectural spaces by blind people using auditory virtual reality for the construction of spatial knowledge (vol 72, pg 393, 2014). Int. J. Hum Comput Stud. **72**, 875 (2014)
45. Plack, C.J., Viemeister, N.F.: Suppression and the dynamic range of hearing. J. Acoust. Soc. Am. **93**, 976–982 (1993)
46. Proulx, M.J.: Size matters: large objects capture attention in visual search. PLoS ONE **5**(12), e15293 (2010)
47. Razzaque, S., Kohn, Z., Whitton, M.C.: Redirected walking. Vol. 3. Citeseer (2005)
48. Riecke, B., Cunningham, D., Bülthoff, H.: Spatial updating in virtual reality: the sufficiency of visual information. Psychol. Res. **71**, 298–313 (2007)
49. John Ross, M., Morrone, C., Goldberg, M.E., Burr, D.C.: Changes in visual perception at the time of saccades. Trends Neurosci. **24**, 113–121 (2001)

50. Sato, K., Fukumori, S., Matsusaki, T., Maruo, T., Ishikawa, S., Nishie, H., Takata, K., Mizuhara, H., Mizobuchi, S., Nakatsuka, H., Matsumi, M., Gofuku, A., Yokoyama, M., Morita, K.: Non-immersive virtual reality mirror visual feedback therapy and its application for the treatment of complex regional pain syndrome: an open-label pilot study. Pain Med. **11**(4), 622–629 (2010)
51. Schifferstein, H.N.J.: The perceived importance of sensory modalities in product usage: a study of self-reports. Acta Physiol. (Oxf) **121**(1), 41–64 (2006)
52. Schmidt, T., Seydell, A.: Visual attention amplifies response priming of pointing movements to color targets. Percept. Psychophys. **70**(3), 443–455 (2008)
53. Seitz, A., Kim, R., Shams, L.: Sound facilitates visual learning. Curr. Biol.: CB **16**, 1422–7 (2006)
54. Shlaer, S.: The relation between visual acuity and illumination. J. Gen. Physiol. **21**(2), 165–188 (1937)
55. Spence, C., Driver, J.: Audiovisual links in exogenous covert spatial orienting. Percept. Psychophys. **59**(1), 1–22 (1997)
56. Spence, C., Lee, J., Van der Stoep, N.: Responding to sounds from unseen locations: crossmodal attentional orienting in response to sounds presented from the rear. Eur. J. Neurosci. **51**, 1137–1150 (2017)
57. Spence, C., Lee, J., Van der Stoep, N.: Responding to sounds from unseen locations: crossmodal attentional orienting in response to sounds presented from the rear. Eur. J. Neurosci. **51**(5), 1137–1150 (2020)
58. Spence, C., Lee, J., Van der Stoep, N.: Responding to sounds from unseen locations: crossmodal attentional orienting in response to sounds presented from the rear. Eur. J. Neurosci. **51**(5), 1137–1150 (2020)
59. Spence, C., Ranson, J., Driver, J.: Cross-modal selective attention: on the difficulty of ignoring sounds at the locus of visual attention. Percept. Psychophys. **62**, 410–424 (2000)
60. Spence, C., Zampini, M.: Auditory contributions to multisensory product perception. Acta Acust. Acust. **92**, 1009–1025 (2006)
61. Staugaard, C.F., Petersen, A., Vangkilde, S.: Eccentricity effects in vision and attention. Neuropsychologia **92**, 69–78 (2016)
62. Stern, M.K., Johnson, J.H.: Just noticeable difference. Corsini Encycl. Psychol. **2010**, 1–2 (2010)
63. Striedter, G.F.: Brain evolution. The human nervous system, pp. 3–21 (2004)
64. van Veen, H.A., Distler, H.K., Braun, S.J., Bülthoff, H.H.: Navigating through a virtual city: using virtual reality technology to study human action and perception. Fut. Gener. Comput. Syst. **14**(3), 231–242 (1998)
65. Vandenbussche, E., Vogels, R., Orban, G.A.: Human orientation discrimination: changes with eccentricity in normal and amblyopic vision. Investig. Ophthalmol. Vis. Sci. **27**(2), 237–245 (1986)
66. Vines, B.W., Krumhansl, C.L., Wanderley, M.M., Levitin, D.J.: Cross-modal interactions in the perception of musical performance. Cognition **101**(1), 80–113 (2006)
67. Volkmann, F.C., Riggs, L.A., Moore, R.K.: Eyeblinks and visual suppression. Science **207**(4433), 900–902 (1980)
68. Wobbrock, J.O., Findlater, L., Gergle, D., Higgins, J.J.: The aligned rank transform for nonparametric factorial analyses using only anova procedures. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 143–146 (2011)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Daniel Jiménez-Navarro is PhD student at Max Planck Institute for Informatics, Saarbrücken (Germany). He is a researcher of the Computer Graphics Department (D4), where he works under the supervision of Prof. Dr. Karol Myszkowski and Prof. Dr. Hans-Peter Seidel. Daniel obtained his bachelor degree at University of Zaragoza, majoring in Electronic and Automatic Engineering in 2020. Later, he obtained his master degree at the same university, majoring in Robotics, Graphics and Computer Vision in 2022. He spent some months as research intern at Graphics and Imaging Lab, working with Dr. Ana Serrano and Dr. Sandra Malpica. His research is focused on studying human perception and performance in virtual environments, particularly audiovisual interaction and visual response behavior.

Ana Serrano is an Associate Professor at Universidad de Zaragoza (Spain). Previously, she was a Postdoctoral Research Fellow at the Max- Planck- Institute for Informatics. She was the recipient of an Adobe Research Fellowship honorable mention in 2017, and a NVIDIA Graduate Fellowship in 2018. Her doctoral thesis was recognized with one of the Eurographics 2020 Ph.D. awards. She was recognized with the Eurographics Young Researcher Award in 2023 and with the IEEE VGTC Significant New Researcher Award in 2024. Her research spans several areas of visual computing with a focus on perceptually- motivated solutions. She has served in several technical papers program committees, including ACM SIGGRAPH, Eurographics, or IEEE VR, and she has been program chair of conferences such as the ACM Symposium on Applied Perception and CEIG.

Sandra Malpica is an Assistant Professor at the Centro Universitario de la Defensa, Zaragoza (Spain). She is a researcher of the Graphics and Imaging Lab, where she carried out her PhD under the supervision of Prof. Belen Masia and Prof. Diego Gutierrez until 2023. Sandra obtained her bachelor degree at Universidad de Zaragoza, majoring in Computer Engineering with a Computer Science mention in 2017. Later, she obtained her master degree at the same university, majoring in Biomedical Engineering in 2018. She has worked as a research intern with Adobe Research and Facebook Reality labs, and spent some months as a postdoc researcher working with Prof. Nuria Pelechano and Prof. Oscar Argudo. Her research interests revolve about the particularities of human perception and virtual reality, with a particular focus in audiovisual and multimodal perception.