Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-025-03065-y

Localization abilities with a visual-to-auditory substitution device are modulated by the spatial arrangement of the scene

Camille Bordeau[1,2] · Florian Scalvini[3,4] · Cyrille Migniot[3] · Julien Dubois[3] · Maxime Ambard[1]

Accepted: 16 March 2025
© The Psychonomic Society, Inc. 2025

Abstract

Visual-to-auditory substitution devices convert visual images into soundscapes. They are intended for use by blind people in everyday situations with various obstacles that need to be localized simultaneously, as well as irrelevant objects that must be ignored. It is therefore important to establish the extent to which substitution devices make it possible to localize obstacles in complex scenes. In this study, we used a substitution device that combines spatial acoustic cues and pitch modulation to convey spatial information. Nineteen blindfolded sighted participants had to point at a virtual target that was displayed alone or among distractors to evaluate their ability to perform a localization task in minimalist and complex virtual scenes. The spatial configuration of the scene was manipulated by varying the number of distractors and their spatial arrangement relative to the target. While elevation localization abilities were not impaired by the presence of distractors, the ability to localize the azimuth of the target was modulated when a large number of distractors were displayed at the same elevation as the target. The elevation localization performance tends to confirm that pitch modulation is effective to convey elevation information with the device in various spatial configurations. Conversely, the impairment to azimuth localization seems to result from segregation difficulties that arise when the spatial configuration of the objects does not allow pitch segregation. This must be considered in the design of substitution devices in order to help blind people correctly evaluate the risks posed by different situations.

Keywords Sensory substitution · Cocktail party · Sonification · Image-to-sound conversion · Localization · Visual impairment · Auditory scene analysis · Feature segregation

Introduction

Sensory substitution devices (SSDs) are intended to convey information about the surrounding environment through an alternative sensory modality. Visual-to-auditory SSDs convert visual information into soundscapes by mapping visual features into auditory cues. Despite promising results on the feasibility of visual-to-auditory SSDs, these devices are still not widely used by blind people in their everyday lives.

It has frequently been asserted that the auditory information provided by a visual-to-auditory SSD can result in an auditory overload when environmental sounds are heard simultaneously or when the visual scene perceived with the SSD is complex (Elli et al., 2014; Maidenbaum et al., 2014). In everyday life, the many different objects present in the surrounding environment may make it difficult to interpret the auditory information provided by the SSD. For this reason, and as stated in Hamilton-Fletcher and Chan (2021), it is important to evaluate the ability to perceptually segregate

✉ Camille Bordeau
bordeau.camille@gmail.com

Florian Scalvini
Florian.Scalvini@imt-atlantique.fr

Cyrille Migniot
Cyrille.Migniot@u-bourgogne.fr

Julien Dubois
Julien.Dubois@u-bourgogne.fr

Maxime Ambard
maxime.ambard@u-bourgogne.fr

1 University of Burgundy, CNRS, LEAD Umr5022, 21000, Dijon, France

2 Aix Marseille University, CNRS, CRPN, Marseille, France

3 Imvia UR 7535–University of Burgundy, Dijon, France

4 IMT Atlantique, LaTIM U1101 INSERM, Brest, France

Published online: 25 April 2025

the soundscape in order to improve the way SSDs make it possible to localize obstacles. However, most studies on SSD have used simple scene configurations. Studies involving localization tasks have usually presented scenes containing only a single object (Ambard et al., 2015; Auvray et al., 2007; Bordeau et al., 2023; Brown et al., 2011; Commère et al., 2020; Hanneton et al., 2010; Levy-Tzedek et al., 2012; Mhaish et al., 2016; Pourghaemi et al., 2018; Proulx et al., 2008), although such scenarios rarely occur outside a laboratory context. In real-life contexts, the relevant SSD information has to be separated both from natural sounds (e.g., people talking, car horns) and from SSD information that might not be particularly relevant for pedestrian mobility.

The work of Buchs et al. (2019) investigated the effect of irrelevant background sounds on the ability to perform a task with the EyeMusic SSD (Abboud et al., 2014). Encouragingly, these authors showed that blind participants are able to efficiently use SSD soundscapes to identify the color and shape of visual stimuli and ignore irrelevant environmental sounds. Therefore, in a similar way to what has been reported regarding the cocktail party problem (Bronkhorst, 2000; Cherry, 1953), participants in the Buchs et al. (2019) study were able to focus their attention on the SSD soundscapes played through bone-conduction headphones while ignoring the irrelevant background noise. However, in this study, the irrelevant auditory information was emitted by a real sound source, and the authors therefore did not directly evaluate the ability to segregate relevant information within complex synthesized SSD soundscapes.

Some studies have investigated the ability to use an SSD when distinct objects are present in the scene and supposed to be perceived through the SSD soundscapes. For instance, Richardson et al. (2019) showed that participants could distinguish between two objects perceived with the SSD Synaestheatre when these were located at distinct distances or distinct elevations. Participants in their study were, to a certain extent, able to discriminate the distance or elevation of the two objects provided that they were far enough apart. Ambard et al. (2015) observed difficulties in segregating two objects perceived with an SSD when they were simultaneously displayed at the same elevation. In their work, Brown et al. (2015) used the vOICe SSD (Meijer, 1992) to investigate the ability to segregate two distinct lines that were sonified with the SSD, in the light of the consonance (frequency component) of the resulting SSD soundscape. They found that the perceptual segregation of the two horizontal lines into distinct objects was impaired when the SSD soundscape contained consonant harmonic relations. However, in the above-mentioned SSD studies, the two displayed objects (or visual features) were relevant for the respective tasks, while in a real context of SSD use, many irrelevant objects may be present in the environment and have to be ignored in order to make it possible to process the relevant information.

The ability to localize a real sound source emitted among an irrelevant acoustic background has been assessed in previous studies using sound maskers (Brungart et al., 2005, 2014; Lorenzi et al., 1999). These "cocktail party" configurations made use of sequential localization tasks, during which a localization target in the form of a broadband noise (Brungart et al., 2005) or broadband environmental sound source (Brungart et al., 2014) was "added" to a mix of up to five irrelevant sounds in Brungart et al. (2014) and 13 irrelevant sounds in Brungart et al. (2005). These studies showed that localization performance decreased as the number of concurrent sound sources increased, reaching a severe level of impairment when more than five sound sources were played simultaneously. These studies using real sound sources give us an insight into the extent to which it is possible to separate relevant sound sources in an ecological context in which the participants can use natural auditory cues.

The ability to distinguish specific sound sources against a background has also been investigated using simulated sound sources (Best et al., 2004; Feierabend et al., 2019; Kawashima & Sato, 2015). For instance, using sound sources spatialized with individualized head-related transfer functions (HRTFs), Best et al. (2004) investigated the ability to spatially segregate two simultaneously emitted broadband sound sources that were separated either in azimuth or in elevation. They showed that the ability to segregate the two sound sources depended on their spatial alignment (azimuth or elevation). When the sound sources were aligned along the same elevation but located at different azimuths around the median axis, the azimuth separation required for participants to perceive two distinct sound sources was smaller than when the two sound sources were located more laterally. In contrast, when the sound sources were aligned along the same azimuth but located at different elevations, segregation abilities were lower when the sound sources were located closer to the median axis than when they were located laterally. In another study with simulated environmental sounds presented either alone or with four other sound sources, Feierabend et al. (2019) showed a decrease in localization performance in both sighted blindfolded and blind participants in the cocktail party configuration.

These studies provide clear evidence that localizing a simulated sound source among other sounds that are emitted simultaneously may result in localization impairments. Furthermore, the spatial arrangement of the sound sources influences this effect (Kwak & Han, 2020). For instance, spatially separating the sound sources reduces the localization impairments (Kawashima & Sato, 2015), an effect known as the spatial release from masking, and segregation abilities are influenced by the dimension (azimuth or elevation) along which the sound sources are aligned (Best et al., 2004).

It is thus well established that, in the context of auditory scene analysis, the ability to separate (and localize) a real or simulated sound source against an irrelevant background is limited and depends both on the spatial arrangement of the auditory scene and on the number of sound sources present. However, this has never been directly investigated in the context of SSDs that are intended for use in complex situations with multiple simultaneous obstacles. Also, SSD soundscapes are quite different from sounds presented in auditory experiments because they are often composed of a succession of short sounds, spatialized or not, that are combined depending on the visual information of the image captured by the camera. The presented combined sounds are therefore changing frame after frame, therefore less steady than in an auditory localization task. The present study therefore evaluates the ability to use an SSD to localize an object in a complex setting containing multiple irrelevant objects that are also transmitted as part of the SSD soundscape, while taking account of both the number of objects and their spatial disposition.

We used a pointing localization task to assess participants' localization abilities after first briefly familiarizing them with the SSD encoding scheme. The SSD encoding scheme uses spatial binaural acoustic cues for the azimuth dimension and combines spatial spectral acoustic cues and pitch modulation for the elevation dimension. Since the number of simultaneous sound sources and their spatial disposition are known to influence localization performance, the number of simultaneous distractors in the scene was manipulated (zero, two, and four), as was their spatial organization relative to the target (horizontally aligned, vertically aligned or nonaligned). We predicted that localization performance would decrease with increasing scene complexity (i.e., as the number of distractors increased), because it was also observed in studies with real sound sources (e.g., Brungart et al., 2014). We expected scene complexity to have a greater effect on azimuth localization abilities than elevation localization abilities, since the pitch modulation used for the elevation dimension should convey relevant spectral information. We also hypothesized that localization abilities should be less impaired when the spatial disposition of the objects in the scene results in object-specific spectral signatures in the SSD soundscape (i.e., when the objects are located at distinct elevations).

Method

Participants

Nineteen participants took part in the study (age: $M = 23.7$ years, $SD = 3.3$, 14 men, 17 right-handed). The sample size was based on a previous work (Bordeau et al., 2023) using the same task and measurements where 19 participants per group were used. None of them reported any hearing impairments, psychiatric illnesses, or neurological disorders in their medical histories. The experimental protocol was approved by the local ethical committee Comité d'Ethique pour la Recherche de Université Bourgogne Franche-Comté (CERUBFC- 2021–12–21–050) and followed the ethical guidelines of the Declaration of Helsinki. All participants provided written informed consent before participating in the experiment and they did not receive any monetary compensation.

Material and apparatus

Virtual environment

The experiment was conducted in a minimalist UNITY3D virtual environment composed of a virtual camera and virtual objects (the target and the distractors). Four HTC VIVE base stations were used to track the participant's head and a pointing tool, to which HTC VIVE Trackers 2.0 were attached. The virtual environment could not be visually explored since the participants were blindfolded and did not wear a virtual reality headset.

Visual-to-auditory SSD

The visual-to-auditory SSD acts in real time and converts a video stream into soundscapes containing the 3D spatial information, as explained in the following sections.

Video acquisition and processing The video recording was captured with a virtual camera with a field of view of 90 × 74° (Horizontal × Vertical) and a frame rate of 60 Hz. The raw video consisted of a depth map encoded into grayscale images ranging from white (0.01 m) to black (5.01 m) gray levels in 0.2-m steps. It resulted in 26 layers of gray level: ranging from [0.01 m, 0.2 m] = layer #25 to [4.9 m, 5.1 m] = layer #0). Therefore, visual information further than 5-m away from the camera are not conveyed.

The processed video frame contained pixels for which the absolute difference in gray levels between consecutive frames (frame differencing) was larger than a threshold of 10. The processed grayscale image was then scaled to 160 × 120 pixels (Horizontal × Vertical) and contained only new visual information—that is, the "active" graphical pixels selected during video processing.

Visual-to-auditory conversion The extracted visual features (i.e., the active pixels) contained in the processed video frame were then converted into a soundscape in accordance with an encoding scheme that mapped 3D spatial information (azimuth, elevation, and distance) to acoustic cues. A

schematic overview of the visual-to-auditory encoding scheme can be seen in Fig. 1. The soundscape consisted of summed "auditory pixels" constituting the audio frame, and consecutive audio frames were combined in real time to form a continuous audio stream. The auditory pixels were 35-ms stereophonic spatialized Gaussian-modulated monotones that were associated with both the location (azimuth and elevation) and the gray level (directly linked to distance) of a given graphical pixel position in the processed image. Elevation location was mapped to the pitch of the monotone with lower pitches corresponded to lower elevation locations. A linear mel scale ranging from 344 mel (for the bottom row) to 1,286 mel (for the top row) was used, which corresponds to frequencies ranging from 250 to 1492 Hz. The monotone was spatialized in azimuth and elevation using HRTFs from the CIPIC database (Algazi et al., 2001). To convey the distance dimension, the sound intensity of the auditory pixels was modulated in such a way that the sound intensity increased with increasing gray level (i.e., decreasing distance). The mean amplitude of each auditory pixel was modulated following the function $f(d) = \frac{1}{1+d^2}$, where d is the distance in meter associated with the gray level of the pixel. Soundscapes were delivered in real time with a SONY MDR-7506 headphone.

Virtual objects

Virtual target

The virtual target to be localized was a self-rotating 3D propeller shape consisting of four intersecting bars of a length of 5 cm. The virtual target was rotating around their center because movement in the image is necessary to activate auditory pixels (otherwise all auditory pixels would have been silenced the frame after it was displayed). In all cases, the virtual target was placed on a virtual sphere with a radius of 1 m centered on the position of the virtual camera (Fig. 2)

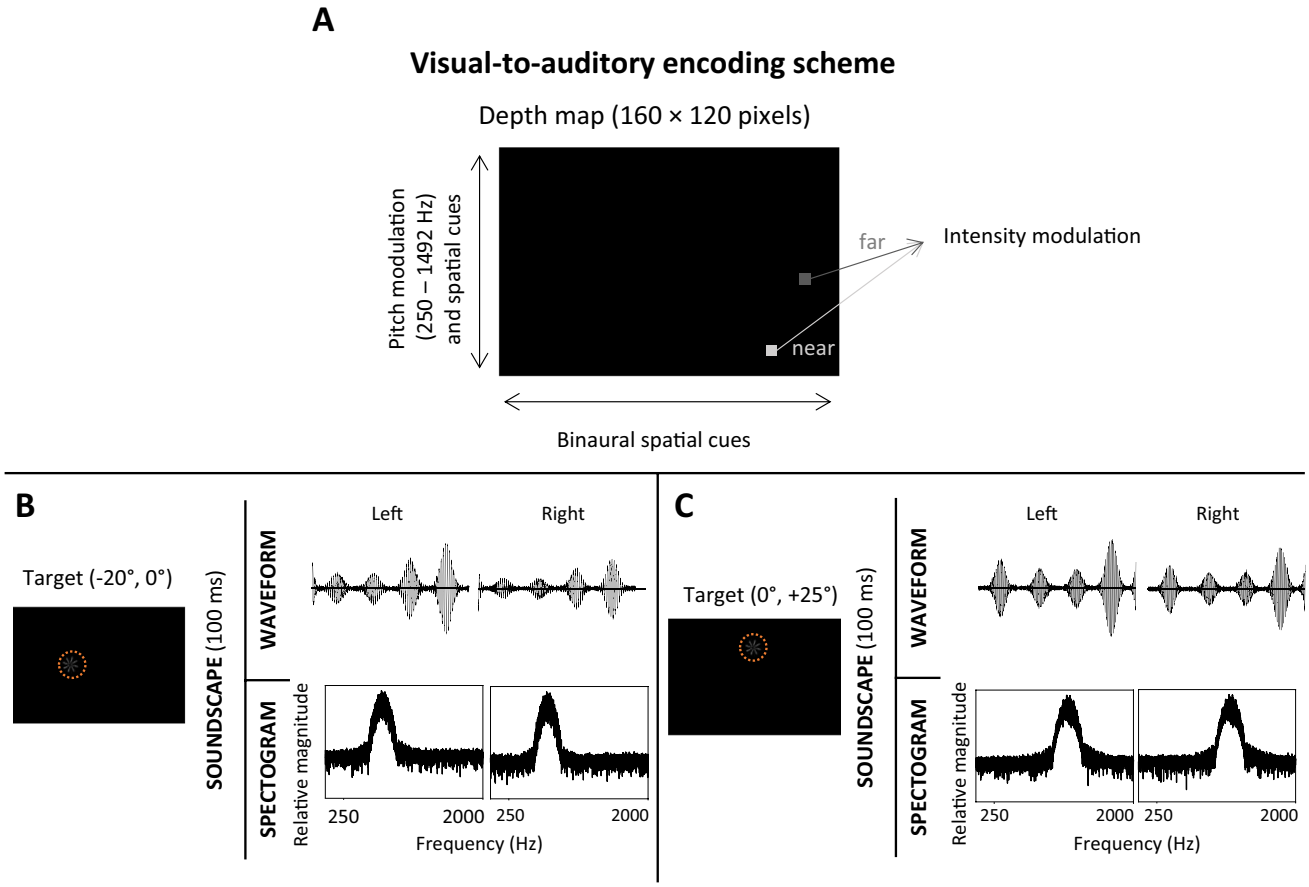

Fig. 1 Visual-to-auditory encoding scheme (**A**). Azimuth is conveyed by binaural cues, elevation by pitch modulation and spatial cues, and distance by intensity modulation. Two examples of 100-ms soundscapes (bottom) corresponding to a target (circled in orange) localized on the left side (**B**) and upper part of the processed depth-map image (**C**). The spectrum shows the difference in frequency components between the two locations (higher frequencies for the upper position). The waveform of the left location shows the binaural spatial cues. (Color figure online)

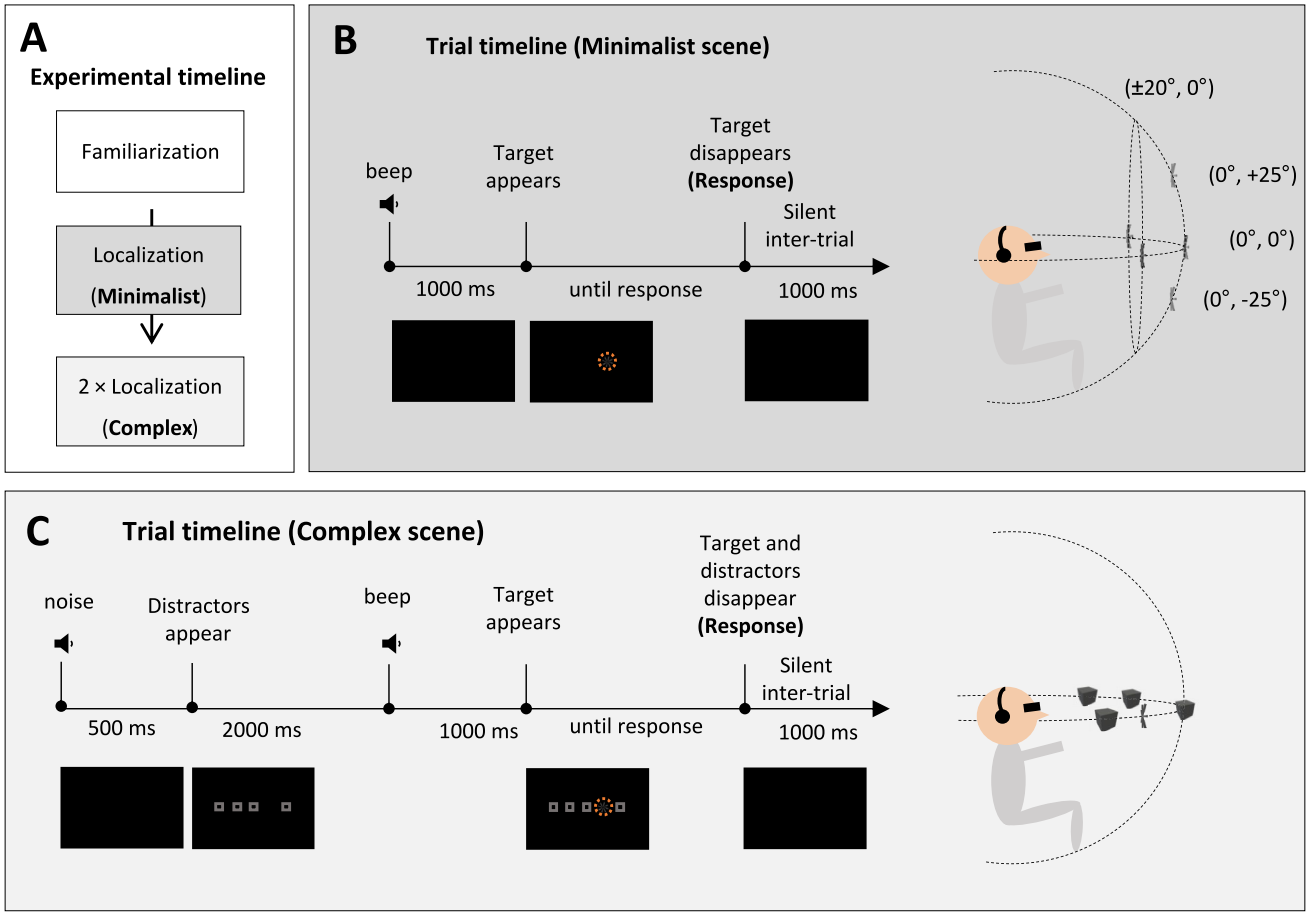

Fig. 2 Experimental timeline (**A**) and trial timeline in the minimalist scene condition (**B**) and complex scene condition (**C**). **A** Immediately after the brief, verbal explanations of the SSD encoding scheme and the short audio-motor familiarization, participants practiced a block of the localization task in the minimalist scene (without distractors), followed by two blocks in the complex scene (with distractors). **B** In the minimalist scene trials, the target (the propeller circled in orange) appeared after a short auditory beep and disappeared when the participants pressed the pointing tool to record the response location. The five possible target locations are depicted in the right figure (opaque and transparent gray propeller). **C** In the complex scene trials, the appearance of the target (propeller circled in orange) was preceded by that of the distractors (gray cubes), which remained displayed until the participant pressed the pointing tool. The experimental view is shown schematically on the right of the figure, where the target (opaque propeller) is located at the right location (+ 20°, 0°) along with four horizontally aligned distractors (gray cubes). (Color figure online)

at one of the five following positions: centered in azimuth at a high position (azimuth = 0°, elevation = + 25°), middle position (0°, 0°), or bottom position (0°, − 25°), or laterally on the right side (+ 20°, 0°) or left side (− 20°, 0°).

Virtual distractors

The virtual distractors were self-rotating cubes measuring 8 × 8 × 8 cm displayed on a sphere of 2 m in radius centered on the virtual camera. The cubes were self-rotating for the same reason than the target (to avoid silencing the auditory pixels). The number of distractors (number: two or four) and their spatial disposition relative to the target (disposition: nonaligned, vertically aligned, or horizontally aligned) were manipulated in a within-subject design. In the vertically aligned condition, the distractors and the target were vertically aligned on the image at an elevation of between − 25° and + 25°, spaced at intervals of 25° in the case of two distractors and 12.5° in the case of four distractors. In the horizontally aligned condition, they were horizontally aligned with the target at intervals of 20° in azimuth along the azimuth range from − 20° to + 20° with two distractors and − 40° to + 40° with four distractors. In the vertically aligned and horizontally aligned conditions, no distractor was displayed at the location of the target, with the result that the target was always separated from the nearest distractor at least by 12.5° in elevation and 20° in azimuth. In the nonaligned condition, the distractors were never aligned with the target since their coordinates were (− 30°, + 27°) and (+ 30°, − 27°) with two distractors, while two additional

positions (− 30°, − 27°) and (+ 30°, + 27°) were used with four distractors. The three spatial dispositions of the distractors relative to the target are depicted in Fig. 3.

Experimental procedure

The experiment consisted of a single 45-min session. After participants had given their informed consent and filled out a demographic questionnaire, the experimenter briefly explained the main principles of the SSD. Participants were then actively familiarized with the SSD and were tested on three blocks of the localization task, each composed of 48 trials. The timeline of a session is given in Fig. 2 and explained in more detail in the following sections.

Verbal explanations of the main principles of the SSD

The experimenter explained the conversion principles of the SSD verbally to each participant at the beginning of the experiment. Participants were informed that they would be blindfolded and then have to localize a virtual target based on sounds that depended on the lateral and vertical position of the target in front of them. The encoding scheme for azimuth (spatialization) and elevation (pitch modulation) was briefly explained to them. Since the distance was not manipulated during the experiment, no information was given to the participants about this dimension.

Audio-motor familiarization with the SSD

Immediately after the verbal explanations, participants were blindfolded and performed a 90-s active familiarization phase. They were instructed to hold the pointing tool in an outstretched arm and point it ahead of them. The virtual target was always placed on a virtual sphere of 1 m in radius centered on the position of the virtual camera, at the intersection between the ray coming from the pointing tool and the sphere. Participants were instructed to pay attention to the sounds they heard depending on the location of the virtual target. They were free to place the target wherever they wanted, but were encouraged to place it at various elevations and azimuth positions and to pay attention to the spatial limit of the space within which the target could be heard (i.e., the field of sonification). Since the position of the virtual camera was updated on the basis of the participant's head tracker position only at the beginning of the familiarization phase, participants were instructed to keep their heads still. The participant's head tracker position was recorded during the familiarization session to check that this instruction was adhered to.

Localization task

After the familiarization phase, participants were tested on 144 trials of the localization task, divided on three blocks of 48 trials. The order of presentation of the scene condition was fixed: For all participants, the first block was always the localization task without distractors (minimalist scene), while the second and third blocks were always with distractors (complex scene; see the following two sections for details on the conditions). In the three blocks, the participants were blindfolded and the task was to localize the virtual target by pointing to it with the pointing tool, relying on soundscapes provided by the SSD. They were asked to point to the target as quickly and accurately as possible, by focusing on the accuracy. Before the first and the second block, participants practiced three practice trials to familiarize themselves with the localization task and to check whether they understood the instructions. Data from these practice trials were not recorded.

During the trials, the response time was not limited. The position of the virtual camera was updated with the participant's head tracker position at the beginning of each trial. Participants were blindfolded during all blocks and were instructed to keep their heads still. The head tracker position was recorded during the localization task to check that they followed this instruction. Participants were given breaks after the familiarization and between the three blocks. Breaks lasted approximately 2 min and they had the possibility to remove the blindfold (all participants did it). In total, participants were tested on 144 trials of the localization task, divided in three blocks of 48 trials (one block in the minimalist scene, following by two blocks in the complex scene).

Minimalist scene (without distractor) In the minimalist scene (first block), a target was displayed without a distractor in each of the 48 trials (Fig. 2, top panel). Each of the 48 trials began with a 500-ms auditory signal (a 400-Hz beep) indicating that the virtual target was going to be displayed in 500 ms. The virtual target was then displayed at one of the five possible locations until the participant pressed the trigger to log the perceived position. The order of trials within the block was randomized so the location of the target varied trial by trial.

Complex scene (with distractors) In the complex scene (second and third blocks), the target was displayed among two or four distractors that were either nonaligned with the target, horizontally aligned with the target, or vertically aligned with the target (Fig. 2). Each of the 96 trials (divided into two blocks) began with a 500-ms white noise audio signal indicating that the distractors were going to appear. After 2 s, during which the distractors were displayed alone, a

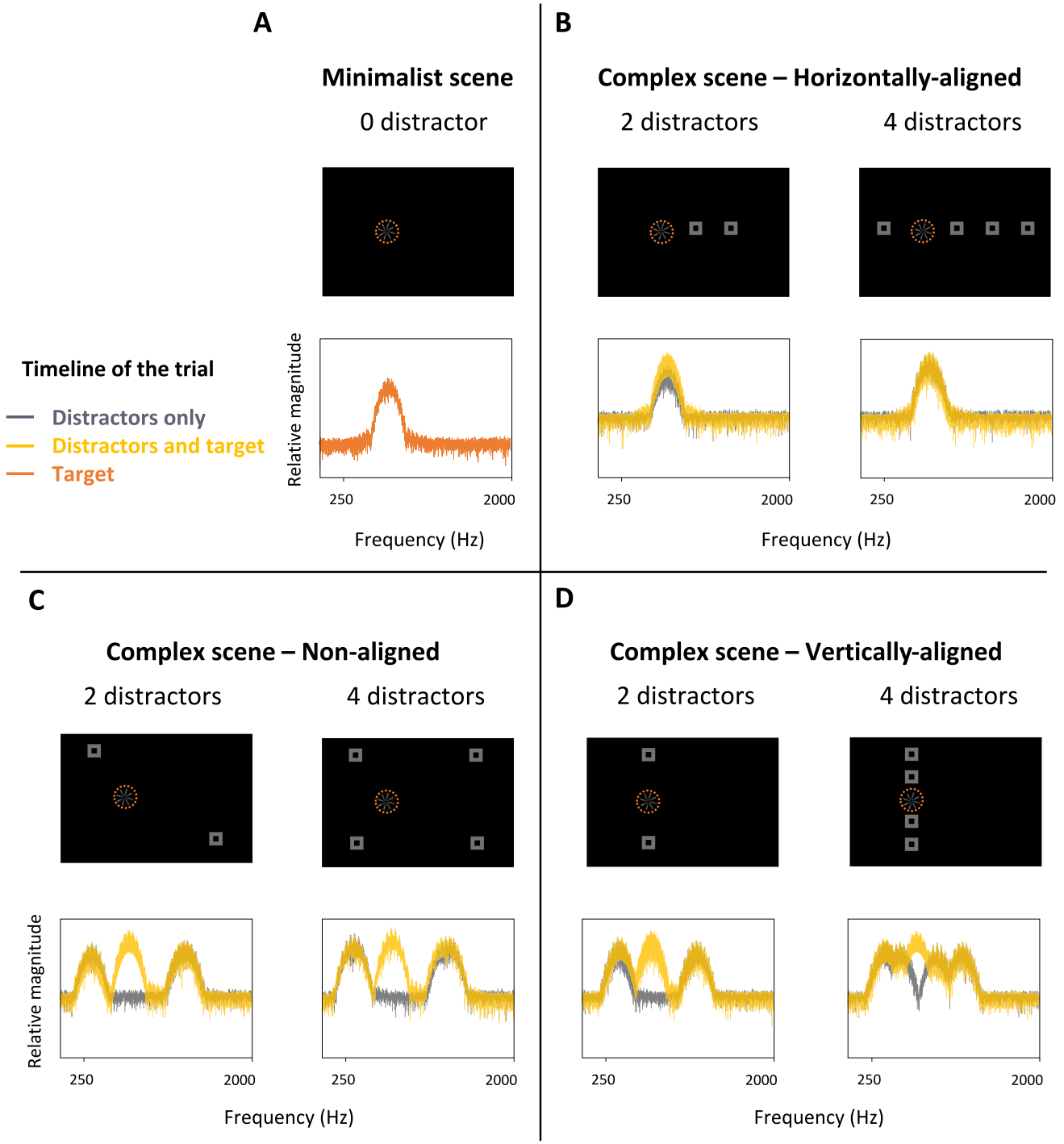

Fig. 3 Processed image captured by the virtual camera and frequency spectrum during a trial as a function of the scene—minimalist (**A**) or complex (**B**), **C** and **D**, the number of distractors (zero, two or four), and their spatial disposition (horizontally aligned, nonaligned, or vertically aligned). The target (circled in orange) was located on the left (− 20°, 0°). The frequency spectrum of the left ear channel corresponding to the SSD soundscapes associated with the processed image is depicted below as a function of the timeline of the trial. In the minimalist scene (without distractor), the frequency spectrum corresponds to the soundscape when the target was displayed (target, orange). In the complex scene (with distractors), the frequency spectrum is provided separately for the phase during which only the distractors were displayed (distractors only, gray) and for the phase during which the target was displayed among the distractors (distractors and target, yellow). (Color figure online)

400-Hz beep was played for 500 ms, and the virtual target was then immediately displayed among the distractors at one of the five possible locations. The distractors and the target disappeared at the same time when the participants logged their response using the pointing tool. The order of trials was randomized within each block so the location of the target, the disposition, and the number of distractors was random.

Data analysis

Version 3.6.1 of the R studio software (Team, 2020) was used for all statistical analyses. In total, 144 response positions (three blocks of 48 trials) per participant were recorded. Localization performance was assessed separately for the azimuth and the elevation dimensions. For each dimension, localization performance was assessed with regression-based and error-based metrics and analyzed with linear mixed models (LMMs) using the *lmerTest* package in R (Kuznetsova et al., 2017). Accuracy was measured with regression-based metrics (gain and bias), while precision was measured with error-based metrics (variable error). A second error-based metric (unsigned error) was used to measure both accuracy and precision. The effects were estimated using analyses of variance (ANOVAs), and the R package *emmeans* (Version 1.7.4; Lenth, 2022) was used for post hoc analyses with Tukey correction.

Regression-based metric: gain and bias

The gain and bias for both dimensions (azimuth and elevation) were estimated based on the predictions of the LMMs (response position as a function of target position). The gain was estimated based on the predicted slope, while the bias was estimated by means of the intercepts. Optimum performance would correspond to a gain value of 1.0 and a bias of 0.0°. In contrast, a gain value of 0.0 could reflect a random pattern of responses. In the azimuth dimension, a negative bias would suggest a leftward bias, while in the elevation dimension, a negative bias would suggest an underestimation bias. Gain and bias were therefore used to measure the accuracy of the localization.

A first LMM on all response positions was fitted with the scene (minimalist scene or complex scene) and target position (− 20°, 0°, and + 20° for the model on azimuth responses, and − 25°, 0°, and + 25° for the model on elevation responses) as fixed factors to investigate the effect of the presence of the distractors on the response pattern. This was done for the azimuth and elevation dimensions separately. Participants were considered as a random factor. A second LMM was fitted only on the response positions in the complex scene (i.e., with distractors) in order to investigate the effects of the number of distractors and their spatial disposition relative to the target, with number (two or four distractors), disposition (nonaligned, horizontally aligned, or vertically aligned) and target position (same modalities as the first model) as fixed factors, and participants as a random factor.

Error-based metrics: unsigned error and variable error

Unsigned errors on elevation (or azimuth) were computed as the absolute value of the difference between the target and the response elevation (or azimuth). The unsigned error metric gives insight about both localization accuracy and precision. The effect of the presence of distractors (scene: minimalist or complex) on the unsigned errors and the effects of the number of distractors (number) and their spatial disposition relative to the target (disposition) were investigated separately by means of two LMMs. The target position was not included as a fixed factor. The estimated marginal means of the unsigned errors provided by the LMMs were used for post hoc pairwise comparisons. Localization precision was investigated with an additional metric: variable error, which was investigated separately for the azimuth and elevation dimensions. For each participant separately, variable error was computed as the mean of the standard deviation of the unsigned error in each experimental condition (minimalist or complex scene; two or four distractors; vertically/horizontally/nonaligned). The effect of the presence of distractors (scene: minimalist or complex) on the variable error and the effects of the number of distractors (number) and their spatial disposition relative to the target (disposition) were investigated separately by means of two LMMs. Target position (elevation or azimuth) was not considered as fixed factor.

Response time

Response time was computed as the duration required for the participant to log its response since the apparition of the target. In the Complex conditions, the period during which the distractors were displayed without the target was not taken into account in the response time computation. The effect of the presence of distractors (scene: minimalist or complex) on response time and the effects of the number of distractors (number) and their spatial disposition relative to the target (disposition) were investigated separately by means of two LMMs. Target position (elevation or azimuth) was not considered as fixed factor.

Results

Head tracker checks

Since participants were instructed to keep their heads as still as possible, the position of the head tracker was recorded every 200 ms during both the familiarization phase and the localization tasks to check that they respected the instructions. During familiarization, the maximum distance of the head from its mean position during the entire familiarization phase was 3.86 ± 1.34 cm ($M \pm SD$) while, during the localization tasks, the maximum distance of the head from its mean position for each trial was on average 1.4 ± 1.2 cm ($M \pm SD$). In both the localization tasks and the familiarization phase, participants thus mostly obeyed the instruction to keep their heads still.

Effect of the presence of distractors on localization performance

The effect of the presence of distractors on localization abilities, without considering the number of distractors or their spatial disposition, was assessed by comparing the azimuth and elevation localization performance without distractors (minimalist scene) and with distractors (complex scene). Regression-based metrics (gain and bias) and error-based metrics (unsigned errors) were analyzed.

Response time with and without distractors

Response time was in average 4.15 s (95% CI [3.56, 4.74]) in the minimalist scene, and 3.93 s (95% CI [3.38, 4.48]) in the complex scene. Response time was not modulated by the presence of distractors, as shown by a nonsignificant effect of scene on response time, $F(1, 18) = 2.60$, $p = 0.124$, $\eta_p^2 = 0.13$.

Azimuth localization performance with and without distractors

For the error-based metrics, the estimated marginal means of the azimuth unsigned error in the minimalist and in the complex scenes are depicted in Fig. 4A. The ANOVA did not reveal any significant effect of scene on the azimuth unsigned error, $F(1, 18) = 0.55$, $p = 0.468$, $\eta_p^2 = 0.03$, with no significant difference being observed between the Minimalist scene (14.5°, 95% CI [11.8, 17.3]) and the complex scene (13.8°, 95% CI [11.7, 15.9]). This result suggests that azimuth localization accuracy was not modulated by the presence of the distractors.

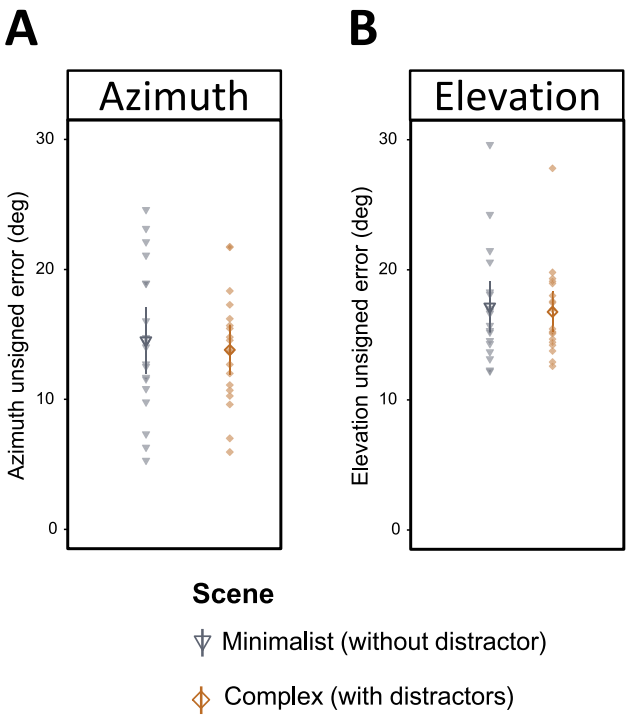

Fig. 4 Estimated marginal mean of the unsigned error on azimuth (**A**) and elevation (**B**) dimensions in the minimalist scene (empty gray triangle) and complex scene (empty brown diamond), all target positions combined. Error bars show the 95% confidence interval of the estimated marginal means. The average unsigned error on azimuth (**A**) and elevation (**B**) for each participant is depicted in the minimalist scene (filled gray triangle) and complex scene (filled brown diamond), all target positions combined. (Color figure online)

For the regression-based metric, the azimuth response positions as a function of the target azimuth in the minimalist scene and in the complex scene are depicted in Fig. 5. The ANOVA revealed a significant interaction effect Target Azimuth × Scene on the azimuth response position, $F(1, 34.5) = 4.85$, $p = 0.034$, $\eta_p^2 = 0.12$, suggesting an effect of scene (minimalist or complex) on the response pattern for the azimuth dimension. Post hoc analyses were conducted to specify the effect of the presence of distractors on the azimuth gain and bias. The azimuth gain was estimated with the slope of the model, and the azimuth bias was estimated with the intercept of the model.

With regard to the gain, the azimuth gain was significantly higher than the optimal gain 1.0, all $t(18) > 6.62$, all $p < 0.0001$, in both conditions. This reveals a tendency to overestimate the lateral position of the lateral targets. However, the analysis showed that the azimuth gain in the complex scene (1.65, 95% CI [1.46, 1.84]) was significantly lower than the azimuth gain in the minimalist scene (1.81, 95% CI [1.55, 2.06]), $t(18) = 2.2$, $p = 0.0409$, indicating that

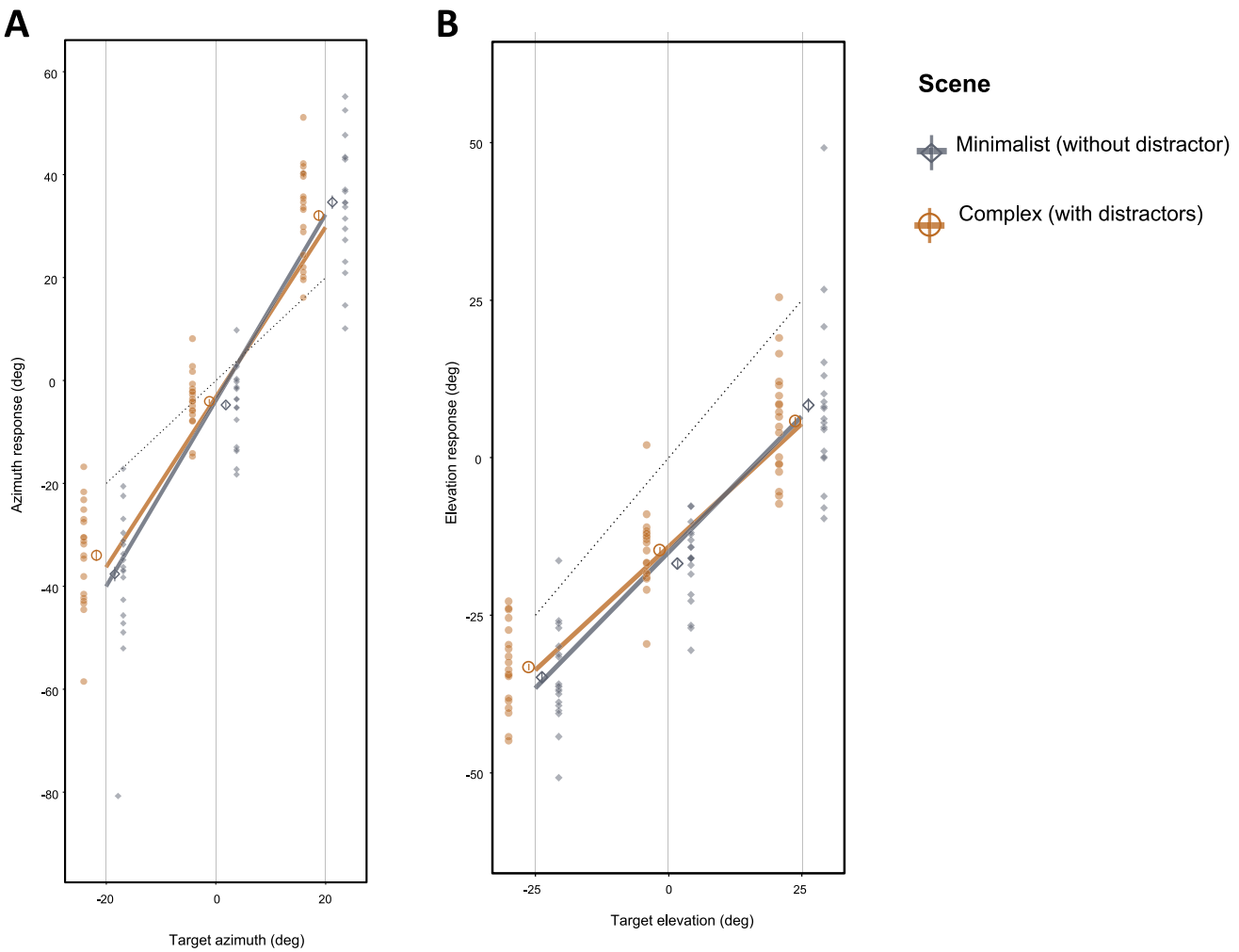

Fig. 5 Mean response position as a function of the target position for the azimuth (**A**) and elevation (**B**) dimensions. Black dashed lines indicate optimal performance with gain = 1.0 and bias = 0°. **A** Azimuth response position as a function of the target azimuth in the minimalist scene (empty gray triangle) and complex scene (empty brown diamond). Error bars show the standard error of the azimuth response position. Solid lines represent the azimuth gains (estimated based on the slopes provided by the LMM) in the minimalist scene (gray) and complex scene (brown). **B** Elevation response position as a function of the target elevation in the minimalist scene (empty gray triangle) and complex scene (empty brown diamond). Error bars show the standard error of the elevation response position. Solid lines represent the elevation gains (estimated based on the slopes provided by the LMM) in the minimalist scene (gray) and complex scene (brown). The average response position for the azimuth (**A**) and elevation (**B**) dimensions is depicted for each participant in the minimalist scene (filled gray triangle) and complex scene (filled brown diamond), all target positions combined. (Color figure online)

the lateral overestimation pattern was lower when the target was displayed among distractors.

With regard to bias, the analysis did not reveal any significant difference between the minimalist scene (− 3.31°, 95% CI [− 5.54, 1.08]) and the complex scene (− 3.97°, 95% CI [− 7.15, − 0.79]). However, the azimuth bias was significantly lower than the optimal bias of 0° in both conditions, all $t(18) > 2.45$, all $p < 0.0248$, suggesting a leftward tendency irrespective of whether or not distractors were present.

With regard to the precision of the localization in the azimuth dimension, analysis did not show a significant effect of Scene on variable error, $F(1, 18) = 1.845$, $p = 0.19$, $\eta_p^2 = 0.09$, with no significant difference being observed between the minimalist scene (9.67°, 95% CI [8.74, 11.7]) and the complex scene (10.22°, 95% CI [8.74, 11.7]). This result suggests that azimuth localization precision was not modulated by the presence of the distractors.

To sum up, these results show that participants were able to perceive the azimuth location of the target regardless of the presence of distractors. They also show a lateral overestimation pattern and a slight tendency to judge the azimuth as being at a more leftwards location.

Elevation localization performance with and without distractors

For the error-based metrics, the estimated marginal means of the elevation unsigned error in the minimalist scene and in the complex scene are depicted in Fig. 4B. The ANOVA did not show a significant effect of scene on the elevation unsigned error, $F(1, 18) = 0.348$, $p = 0.563$, $\eta_p^2 = 0.02$, with no difference being observed between the minimalist scene (17.2°, (95% CI [15.1, 19.3]) and the complex scene (16.8°, 95% CI [15.1, 18.4]). This result suggests that the elevation localization accuracy was not modulated by the presence of the distractors.

For the regression-based metric, the elevation response positions in the minimalist scene (without distractor) and in the complex scene (with distractors) are depicted in Fig. 5. The Target Elevation × Scene interaction effect was not significant, $F(1, 17.998) = 1.85$, $p = 0.19$, $\eta_p^2 = 0.09$, suggesting that the response patterns for the elevation dimension were comparable whether the target was displayed alone or among distractors. To facilitate descriptions, the elevation gains and biases were estimated based on the predictions of the LMM in the minimalist and complex scenes. The elevation gain in the minimalist scene (0.86, 95% CI [0.68, 1.04]), $t(18) = 1.62$, $p = 0.12$, was not significantly different from the optimal elevation gain of 1.0, while the elevation gain was significantly lower than this optimal value in the complex scene (0.78, 95% CI [0.65, 0.91]), $t(18) = 3.63$, $p = 0.0019$. This result suggests an elevation compression pattern when the target had to be localized among distractors, although the response pattern in the elevation dimension did not seem to vary dramatically, since the elevation gain when the target was displayed among distractors was not significantly lower than when it was displayed alone.

The elevation bias was significantly lower than the optimal bias of 0° both in the minimalist scene (− 15.1°, 95% CI [− 17.5, − 12.6]) and in the complex scene (− 14.2°, 95% CI [− 16.5, − 11.9]), all $t(18) > 12.7$, all $p < 0.0001$. This indicates an underestimation of the elevation of the target.

With regard to the precision of the localization in the elevation dimension, variable error was not modulated by the presence of distractors, as shown by a nonsignificant effect of scene, $F(1, 18) = 0.002$, $p = 0.961$, $\eta_p^2 = 0.7$. The precision was 10.5° (95% CI [9.08, 11.9]) in the minimalist scene and 10.5° (95% CI [9.05, 11.9]) in the complex scene. Overall, and without considering the number of distractors and their spatial disposition, participants successfully localized the target with the SSD in the azimuth and elevation dimensions, even when it was displayed in a complex scene containing distractors. The azimuth localization performance was characterized by an overestimation pattern and a slight leftward bias, while the elevation was underestimated, with a slight elevation compression pattern being observed when distractors were displayed.

Effect of the number of distractors and their spatial disposition on localization performance

To investigate the effect of the number of distractors (number: two or four distractors) and their spatial disposition relative to the target location (disposition: nonaligned, horizontally aligned, or vertically aligned), we used another LMM that only took account of the trials in which the target was displayed among distractors (complex scene only). Regression-based metrics (gain and bias computed based on the response positions) and error-based metrics (unsigned errors) were analyzed.

Response time in the complex scene

The ANOVA showed a significant main effect of number, $F(1, 39.39) = 9.224$, $p = 0.004$, $\eta_p^2 = 0.19$, and congruence, $F(1, 36.59) = 5.616$, $p = 0.007$, $\eta_p^2 = 0.23$, on response time. Post hoc analysis showed that response time was significantly longer when the target was horizontally aligned with the distractors (4.11 s, 95% CI [3.52, 4.70]) than when it was vertically aligned (3.85 s, 95% CI [3.32, 4.39]) or nonaligned (3.82 s, 95% CI [3.27, 4.37]), all $t(18) > 2.96$, all $p < 0.0218$. Response time was also longer in trials where four distractors were displayed (4.03 s, 95% CI [3.47, 4.59]) in comparison with trials where only two distractors were displayed (3.83 s, 95% CI [3.47, 4.59]), $t(18) = 3.04$, $p = 0.007$.

Azimuth localization performance in the complex scene

For the error-based metrics, the estimated marginal means of the azimuth unsigned error in the six experimental conditions are depicted in Fig. 6. The ANOVA showed that the Number × Disposition interaction effect was significant, $F(2, 33.83) = 4.0$, $p = 0.0275$, $\eta_p^2 = 0.19$. However, post hoc analyses with Tukey correction did not show any significant differences, all $t(18) < 1.72$, all $p > 0.102$, between the six experimental conditions. The azimuth localization accuracy did not therefore seem to be modulated by the number of distractors or by their spatial disposition relative to the target.

For the regression-based metric, the azimuth response positions in the complex scene are depicted in Fig. 7. The azimuth gains and bias were estimated based on the predictions of the LMM (the slopes and intercepts, respectively) and are summarized in Table 1 for the 6 experimental conditions (with 95% confidence interval).

The azimuth gains in the six experimental conditions were significantly different from the optimal gain of 1.0, all $t(18) > 6.37$, all $p < 0.0001$. However, two distinct response

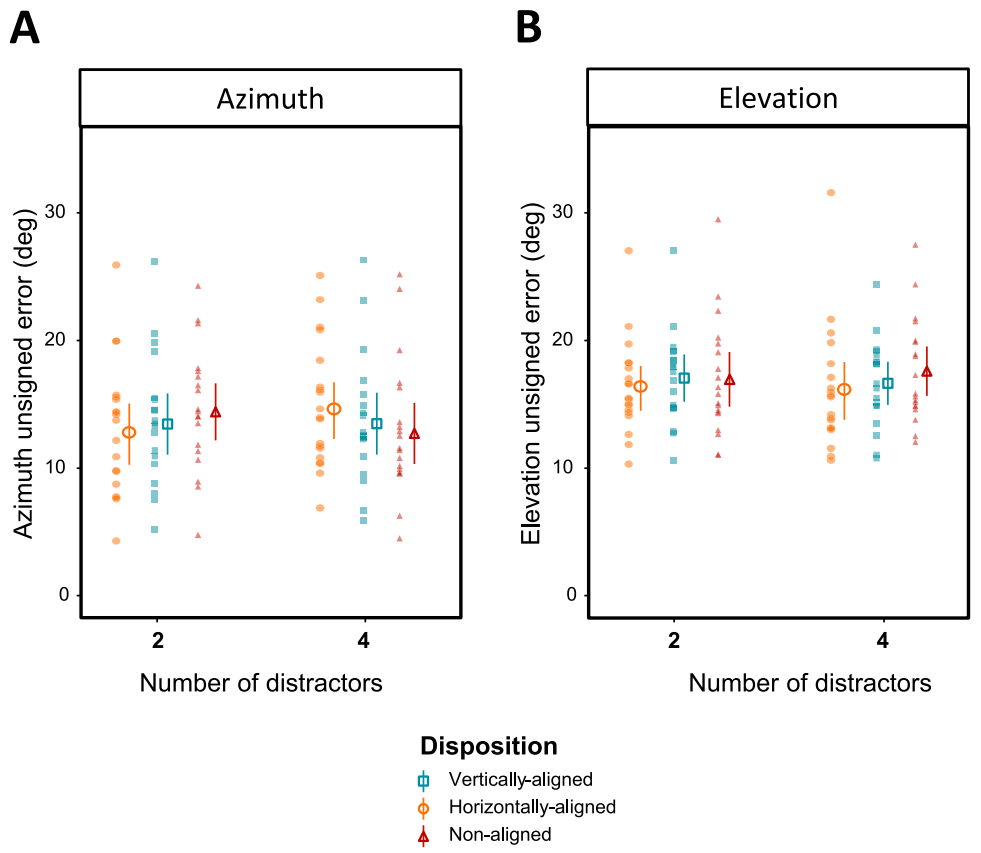

Fig. 6 Estimated marginal mean of the unsigned error on the azimuth (**A**) and elevation (**B**) dimensions in the complex scene. Error bars show the 95% confidence interval of the estimated marginal means. **A** Estimated marginal means of the azimuth unsigned error as a function of the number of distractors (two or four) when the distractors were vertically aligned (empty blue square), horizontally aligned (empty orange circle), and nonaligned with the target (empty red triangle). **B** Estimated marginal means of the elevation unsigned error as a function of the number of distractors (two or four) when the distractors were vertically aligned (empty blue square), horizontally aligned (empty orange circle), and nonaligned with the target (empty red triangle). The average unsigned error on azimuth (**A**) and elevation (**B**) are depicted for each participant as a function of the number of distractors (two or four) when the distractors were vertically aligned (filled blue square), horizontally aligned (filled orange triangle), and nonaligned with the target (filled red triangle). (Color figure online)

patterns were measured depending on the number of distractors and their spatial disposition. Except in the experimental condition where four horizontally aligned distractors were displayed (horizontally aligned condition), the azimuth gains in the other five experimental conditions revealed a tendency to overestimate the lateral position of lateral targets (azimuth gains between 1.61 and 1.95). In contrast, when four distractors were horizontally aligned with the target (horizontally aligned condition), a reverse pattern of azimuth compression was observed. As shown in Fig. 7A (right figure), the azimuth gain decreased dramatically (azimuth gain of 0.79, lower than the optimal gain of 1.0), indicating a lateral underestimation pattern. This decrease did not seem to reflect a random pattern of responses, since the azimuth gain in this condition was still significantly higher than 0.0, $t(18) = 6.424$, $p < 0.0001$. Therefore, participants tended to underestimate the lateral position only when four distractors were horizontally aligned with the target.

This heterogeneity in the results was confirmed by an ANOVA that revealed a significant Target Azimuth × Number × Disposition interaction effect, $F(2, 69.004) = 13.88$, $p < 0.0001$, $\eta_p^2 = 0.29$. Post hoc analyses were therefore conducted to further specify the interaction effect between disposition (nonaligned, horizontally aligned, or vertically aligned) and number (two distractors or four distractors) on the azimuth gains and bias.

To this end, we first investigated whether increasing the number of distractors modulated the azimuth gain as a function of the spatial disposition of the distractors relative to the target (i.e., effect of the number of distractors on the azimuth gain for the three disposition conditions, separately). The analysis confirmed that the decrease in the azimuth gain

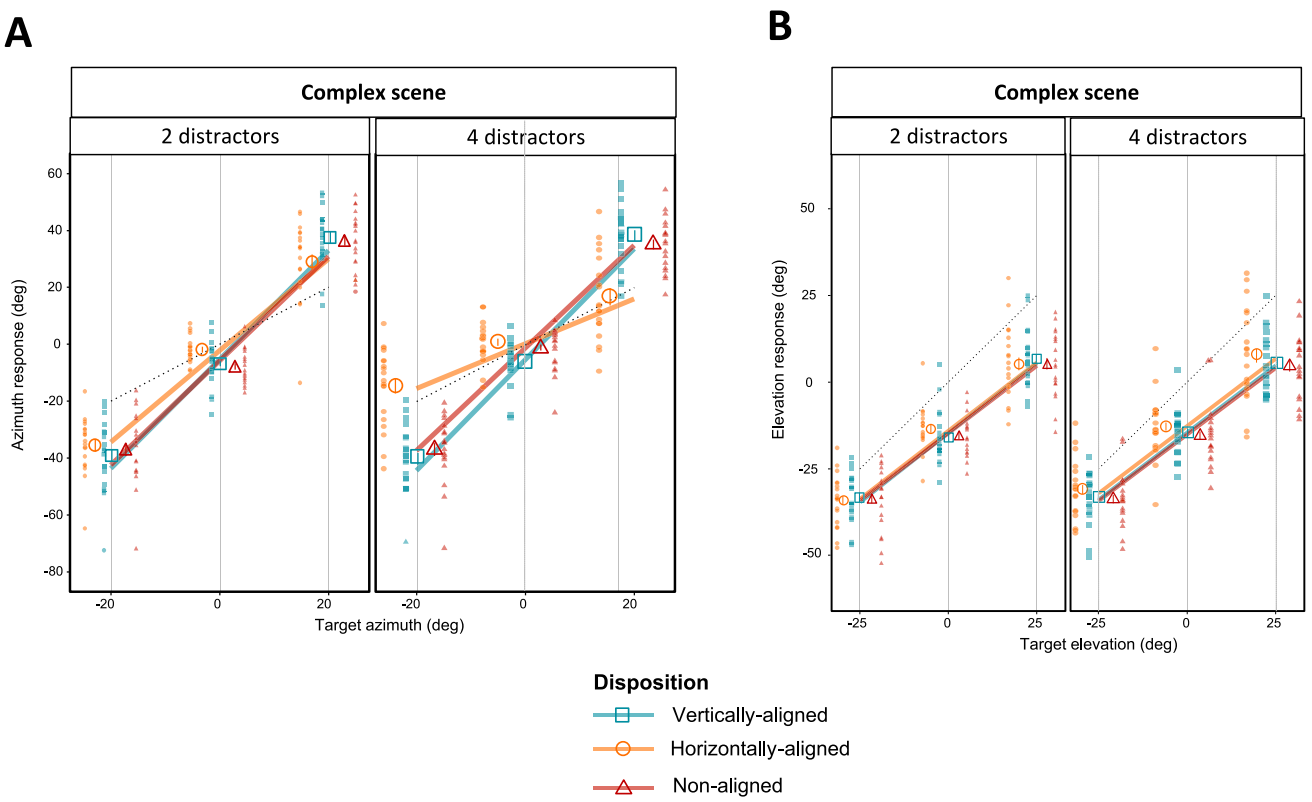

Fig. 7 Mean response position as a function of the target position for the azimuth (**A**) and elevation (**B**) dimensions in the complex scene. Black dashed lines indicate optimal performance with gain = 1.0 and bias = 0°. **A** Azimuth response position as a function of the target azimuth, and the number of distractors (two distractors in the left figure; four distractors in the right figure). Symbols represent the mean azimuth response positions when the distractors were vertically aligned (empty blue square), horizontally aligned (empty orange circle), and nonaligned with the target (empty red triangle). Solid lines represent the azimuth gains (estimated based on the slopes provided by the LMM) when the distractors were vertically aligned (blue), horizontally aligned (orange), and nonaligned with the target (red). **B** Elevation response position as a function of the target elevation, and the number of distractors (two distractors: left figure; four distractors: right figure). Symbols represent the mean elevation response positions when the distractors were vertically aligned (empty blue square), horizontally aligned (empty orange circle), and nonaligned with the target (empty red triangle). Solid lines represent the elevation gains (estimated based on the slopes provided by the LMM) when the distractors were vertically aligned (blue), horizontally aligned (orange), and nonaligned with the target (red). The average response position for the azimuth (**A**) and elevation (**B**) are depicted for each participant as a function of the number of distractors (two or four) when the distractors were vertically aligned (filled blue square), horizontally aligned (filled orange triangle), and nonaligned with the target (filled red triangle). (Color figure online)

Table 1 Azimuth gain and bias for each disposition condition (nonaligned, vertically aligned, horizontally aligned) and number condition (two distractors, four distractors)

		Non-aligned	Vertically aligned	Horizontally aligned
2 distractors	Azimuth gain	1.84 [1.59, 2.08] (*)	1.92 [1.66, 2.17] (*)	1.61 [1.38, 1.84] (*)
	Azimuth bias	− 5.90° [− 8.68, − 3.12] (*)	− 5.37° [− 8.59, − 2.14] (*)	− 2.26° [− 4.76, 0.23]
4 distractors	Azimuth gain	1.80 [1.55, 2.06] (*)	1.95 [1.71, 2.19] (*)	0.79 [0.53, 1.05] (*)
	Azimuth bias	− 1.24° [− 4.50, 2.01]	− 5.36° [− 8.70, − 2.02] (*)	0.26° [− 2.93, 3.45]

95% confidence interval is given in brackets. The (*) symbol indicates a significant difference between the azimuth gain and the optimal gain 1.0, or between the azimuth bias and the optimal bias 0.0°

observed when the number of distractors increased from two to four was only present when the distractors were horizontally aligned with the target (from 1.61 with two distractors to 0.79 with four distractors), $t(18) = 6.29$, $p < 0.0001$.

Secondly, we investigated whether the spatial disposition of the distractors had an influence on the azimuth gain when there were either two or four distractors (i.e., effect of the disposition of the distractors on the azimuth gain depending on the number of distractors). This analysis confirmed the specificity of the configuration with four horizontally aligned distractors. With four distractors, the azimuth gain in the horizontally aligned disposition (0.79) was significantly lower than with the two other spatial dispositions (nonaligned: 1.80, vertically aligned: 1.95), all $t(18) > 7.58$, $p < 0.0001$. In contrast, with two distractors, the azimuth gain in the horizontally aligned condition was only marginally lower than when the distractors were vertically aligned with the target (1.92), $t(18) = 2.529$, $p = 0.0523$, while no difference in azimuth gain was found in the nonaligned condition (1.84), $t(18) = 1.89$, $p = 0.171$.

For the azimuth bias (Table 1), post hoc analyses showed that a leftward bias was observed only when the distractors were vertically aligned with the target (vertically aligned with two distractors: $-5.37°$, vertically aligned with four distractors: $-5.36°$), and when two distractors were nonaligned with the target (nonaligned with two distractors: $-5.9°$), all $t(18) > 3.14$, all $p < 0.0056$. There was no significant shift to the right or left in the other experimental conditions.

With regard to the precision (Table 2), analysis on the variable error revealed a significant interaction effect of Congruence × Number, $F(2, 72.001) = 4.536$, $p = 0.014$, $\eta_p^2 = 0.11$. Post hoc analysis showed that the only significant difference in variable error was when two distractors were displayed, where variable error was significantly higher (lower precision) when the distractors were not aligned with the target (11.02, 95% CI [9.25, 12.915]) than when they were vertically aligned with it (9.33, 95% CI [7.67, 11.0]), $t(46.6) = 2.616$, $p = 0.0315$.

To sum up the results for the azimuth, the accuracy measured in terms of the unsigned error suggests that when the target was displayed among distractors, the accuracy with which the participants localized the azimuth of the target did not vary much irrespective of the spatial disposition of the distractors and their number. However, the azimuth localization pattern measured based on the azimuth gains revealed a lateral underestimation of the azimuth when four distractors were horizontally aligned with the target. In contrast, an overestimation pattern similar to that observed without distractors was measured in all the other configurations. A slight tendency to judge the azimuth at a more leftward location was systematically observed when the distractors were vertically aligned with the target, and also when two distractors were displayed but not aligned with the target.

Elevation localization performance in the complex scene

For the error-based metrics in the elevation dimension, the estimated marginal means of the elevation unsigned error in the six experimental conditions are depicted in Fig. 6. No significant effect of number (two distractors or four distractors) or disposition (nonaligned, horizontally aligned, or vertically aligned) and no Number × Disposition interaction effect were observed in the analysis(all $p > 0.226$) This suggests that the number of distractors and their spatial disposition relative to the target did not modulate the accuracy of target elevation localization.

For the regression-based metric, the elevation response positions in the complex scene are depicted in Fig. 7B, as a function of the number of distractors (number) and their spatial disposition relative to the target (disposition). The corresponding values for the elevation gains and the bias (with the 95% confidence interval) are summarized in Table 3 for the six experimental conditions. The gain and bias of the elevation in all six experimental conditions were compared with the optimal gain of 1.0 and the optimal bias of 0.0°, respectively. A compression pattern was measured for the elevation gains in the six experimental conditions since they were all significantly lower than the optimal gain of 1.0, all $t(18) > 2.38$, all $p < 0.0286$. For the elevation bias, an underestimation was also observed in the six experimental conditions, with the elevation bias being significantly lower than 0°, all $t(18) > 7.868$, all $p < 0.0001$.

Table 2 Precision in the azimuth dimension for each disposition condition (nonaligned, vertically aligned, horizontally aligned) and number condition (two distractors, four distractors)

		Nonaligned	Vertically aligned	Horizontally aligned
2 distractors	Azimuth precision	11.02° [9.25, 12.8]	9.33° [7.67, 11.0]	9.69° [8.02, 11.4]
4 distractors	Azimuth precision	9.36° [7.92, 10.8]	9.49° [8.18, 10.8]	10.65° [9.26, 12.0]

95% confidence interval is given in brackets

Table 3 Elevation gain and bias for each disposition condition (nonaligned, vertically aligned, horizontally aligned) and number condition (two distractors, four distractors)

		Nonaligned	Vertically aligned	Horizontally aligned
2 distractors	Elevation gain	0.78 [0.62, 0.94] (*)	0.80 [0.67, 0.94] (*)	0.79 [0.63, 0.94] (*)
	Elevation bias	− 14.7° [− 17.1, − 12.32] (*)	− 14.6° [− 17.3, − 11.85] (*)	− 13.9° [− 16.5, 11.29] (*)
4 distractors	Elevation gain	0.77 [0.63, 0.90] (*)	0.77 [0.64, 0.91] (*)	0.78 [0.58, 0.97] (*)
	Elevation bias	− 15.0° [− 18.2, − 11.73] (*)	− 14.5° [− 16.9, − 12.03] (*)	− 12.5° [− 15.9, − 9.18] (*)

95% confidence interval is given in brackets. The (*) symbol indicates a significant difference between the elevation gain and the optimal gain of 1.0, or between the elevation bias and the optimal bias of 0.0°

Table 4 Precision in the elevation dimension for each disposition condition (nonaligned, vertically aligned, horizontally aligned) and number condition (two distractors, four distractors)

		Non-aligned	Vertically aligned	Horizontally aligned
2 distractors	Elevation precision	10.46° [8.87, 12.0]	9.77° [8.18, 11.4]	11.20° [9.61, 12.8]
4 distractors	Elevation precision	10.36° [9.12, 11.6]	10.15° [8.91, 11.4]	10.08° [8.84, 11.3]

95% confidence interval is given in brackets

The number of distractors and their spatial disposition relative to the target did not influence the response pattern in the elevation dimension since the ANOVA on the elevation response position did not reveal any significant effect except the main effect of target elevation, $F(1, 17.96) = 166.94$, $p < 0.0001$, $\eta_p^2 = 0.9$.

With regard to the precision (Table 4), analysis on the variable error in the elevation dimension did not reveal any significant effect of congruence or number, nor interaction effect (all $p > 0.258$). Therefore, the precision in the elevation dimension did not seem to be modulated by the number of distractors or their spatial disposition relative to the target. To sum up the results concerning elevation, when the target was displayed among distractors, the accuracy with which the participants localized the elevation of the target did not much vary, irrespective of the spatial disposition of the distractors and their number. In all the conditions, the perception of the elevation of the target was similarly compressed and underestimated.

Discussion

In this study, we investigated the ability to use a visual-to-auditory SSD to localize an object displayed among other irrelevant objects considered as distractors. After participants had been familiarized briefly with the principles of the SSD, their early-stage abilities were assessed with a blindfolded pointing localization task in a virtual environment. The effect of the presence of distractors on localization abilities was assessed by comparing localization performance with and without distractors, while the effect of the spatial disposition of the scene was investigated by manipulating the number of distractors and their spatial disposition relative to the target. The results of the study are discussed in accordance with previous research conducted with visual-to-auditory SSDs, but also with virtual and real sound sources. Although, as mention in the introduction, the soundscapes of an SSD can differ from auditory localization experiments, where stimuli are generally

more steady, the comparison remains relevant because the encoding scheme of our SSD and the experimental design with self-rotating objects result in the same or very close auditory pixels activation from frame-to-frame.

Localization abilities in a minimalist scene

Sound spatialization is an effective encoding scheme for the azimuth dimension although an overestimation pattern is observed

In the minimalist scene without distractors, localization performance for the azimuth dimension showed a lateral overestimation pattern (azimuth gain of 1.81) and a slight leftward bias (− 3.31°). In the field of SSD research, the task described in the current study is comparable to the one presented in Bordeau et al. (2023), which used the same SSD encoding scheme (called monotonic encoding in their study) and a similar experimental set-up (pointing task and familiarization method). In this previous study, the authors did not observe a leftward bias (− 1.8°, but not significant). Nevertheless, a leftward shift has been measured with other tested encoding schemes. Although the reasons for this slight leftward bias are unclear, Bordeau et al. (2023) suggest that it could be due to a perceptive or proprioceptive bias. However, this latter study found a lateral overestimation pattern in the monotonic encoding condition with an azimuth gain of 1.23, a finding which is consistent with the current research. Although present, this previously observed lateral overestimation pattern therefore seems to be lower than the one reported in the current study (1.81). It could be due to the fact that the azimuth range tested in the current study ([− 20°, + 20°]) was smaller than in the previous one ([− 40°, + 40°]). This pattern of overestimation for lateral sound sources has also been reported in many auditory localization studies with real sound sources (e.g., Bruns et al., 2024; Ocklenburg et al., 2009; Odegaard et al., 2015; Oldfield & Parker, 1984), simulated sound sources using nonindividualized HRTFs (Wenzel et al., 1993) or in virtual environments (Ahrens et al., 2019).

The azimuth unsigned error without distractors of 14.2° found in our study is comparable with the range of 15° to 19° measured after familiarization in Bordeau et al. (2023). Although azimuth error has also been measured with other SSDs, the values are difficult to compare since the tasks were different. For instance, using pointing tasks on a table, Hanneton et al. (2010) measured an angular error of about 5° with the Vibe SSD, while Commère and Rouat (2023) measured azimuth errors of between about 8° and 45° with their SSD, which conveys azimuth location by means of stereo panning. Using a body pointing task (i.e., participants had to face the target) with an SSD, Scalvini et al. (2022) measured an azimuth error of about 7°. This value is only half that observed in the current study, but this could be due to the different pointing methods used. Using simulated sound sources spatialized with the same nonindividualized HRTFs database as in this study, Mendonça et al. (2013) measured an azimuth unsigned error of about 15°, which is similar to the value reported here.

Overall, a strong laterality overestimation pattern was observed when the participants had to localize the target in a minimalist scene without distractors, which is consistent with previous auditory localization experiments conducted with real and simulated sound sources with nonindividualized HRTFs. When using an SSD, the lateral overestimation of an obstacle could result in a collision. However, in a real context of use, SSD users would be able to move their heads and improve their perception by aligning the target with the median axis, as observed in experiments using a body or head-pointing method.

Pitch modulation is an effective acoustic cue for the elevation dimension although an underestimation bias is observed

Elevation localization performance measured based on both gain and bias indicated a good ability to discriminate the three tested elevations. The elevation gain of 0.86 was not significantly different from the optimal gain of 1.0, although an underestimation bias of − 15.1° was measured. In Bordeau et al. (2023), the elevation gain obtained using a similar SSD encoding scheme (i.e., monotonic encoding) was 1.015, which was also comparable to the optimal gain of 1.0. The underestimation bias measured in the current study was also observed in Bordeau et al. (2023) and seems comparable (− 14.15°). As explained in Bordeau et al. (2023), it is possible that this underestimation bias was due to the high position of the head tracker, which was located on the front of the participants' heads and associated with the virtual camera. This resulted in a field of view in which the 0° elevation coordinate corresponded to the axis straight ahead of the head tracker, i.e., higher than ear level.

The unsigned elevation error observed in the current study was about 16°, which is comparable to the range of 17° to 24° after familiarization found in Bordeau et al. (2023). Using the Synaestheatre SSD (Hamilton-Fletcher et al., 2016), which conveys elevation and azimuth using only spatial cues based on nonindividualized HRTFs, Richardson et al. (2019) measured an elevation discrimination score of 14°, which is also similar to the current findings. As a comparison in the auditory localization field, Mendonça et al. (2013) measured an elevation error of about 25° after the participants had been familiarized with sounds spatialized using the same HRTF database as used in this study. However, Mendonça et al. (2013) tested higher elevation locations (elevation locations ranging from 0° to + 90°)

and this probably resulted in larger localization errors since elevation localization abilities are poorer for high elevations (Makous & Middlebrooks, 1990).

Overall, in the minimalist scene without distractors, the brief period of audio-motor familiarization with the SSD encoding scheme was sufficient to enable participants to localize the elevation of the target based on spectral information resulting from the pitch modulation in the elevation encoding scheme. Although elevation was greatly underestimated, this was probably due to the spatial disparity between the egocentric spatial mental representation and the head tracker location. As participants become more familiar with the SSD, such errors could be reduced by recalibrating their auditory spatial perception, as observed in studies using nonindividualized HRTFs (e.g., Stitt et al., 2019).

Localization abilities in a complex scene

To assess localization abilities with the SSD in a complex scene, we assessed localization performance of a target displayed among distractors. The distractors were either aligned with the target (horizontally or vertically) or nonaligned. Overall, localization performance in the elevation dimension was not impaired with the distractors, although the presence of distractors modulated the response pattern in the azimuth dimension. It is important to mention that the order of the blocks was fixed so the minimalist scene was always tested first, followed by two blocks in the complex scene. This might have resulted in an order effect due to, for example, fatigue. However, this experimental choice, also made in the studies of Buchs et al. (2019) and Feierabend et al. (2019), was made to avoid a too difficult task at the beginning of the experiment.

Causes of the azimuth underestimation pattern when four distractors are horizontally aligned with the target

The laterality overestimation pattern observed without distractors was also present when the target was displayed among distractors, with an azimuth gain of 1.65, although it was lower than the azimuth gain of 1.81 without distractors. Actually, the decrease in the azimuth gain observed with distractors is mainly due to one specific configuration in which four distractors were horizontally aligned with the target. In this condition, we observed a reverse pattern of lateral underestimation, with an azimuth gain of 0.79. This decrease in the azimuth gain and the resulting lateral underestimation pattern may reflect a localization impairment in the azimuth dimension since it shows a bias toward the median axis (i.e., 0° azimuth). Since the precision was not significantly modulated in this condition, this underestimation pattern does not seem to reflect random responses but rather a decrease in discrimination ability. The average response time was longer in the complex scene when distractors were horizontally aligned compared with when they were vertically aligned or nonaligned. Response times were also longer when four distractors were displayed rather than two. Previous studies investigating assistive devices for the blind have shown that the time to complete a task is linked to task difficulty (Hicks et al., 2013; Kolarik et al., 2014), which supports the idea that task difficulty was higher in the condition where four distractors were horizontally aligned with the target. The potential reasons for these difficulties are discussed below.

What mainly differentiate the condition where distractors were horizontally aligned with the target from the other conditions is the similar narrowband frequency spectrum associated with each object. Since the elevation encoding scheme used by the SSD involves pitch modulation, the configuration in which the distractors and the target are horizontally aligned results in a similar narrowband frequency spectrum associated with each object (Fig. 3). The frequency composition of a soundscape is known to influence sound source segregation abilities due to the auditory filters emerging from the cochlear tonotopy (Bregman, 1990), also known as critical bands (Glasberg & Moore, 1990; Zwicker, 1961). This phenomenon implies that the segregation of two narrowband sounds is impaired when they share frequency components within a given bandwidth, called the critical band. In other words, a masking effect occurs when the frequency spectrum of the sound masker is too close to the frequency of the target tone and leads to the perceptual fusion of the masker and the target.

Ambard et al. (2015) also observed a bias toward the median axis when two horizontally aligned objects were simultaneously perceived through an SSD soundscape. A masking effect probably also occurred when the target and the distractors were horizontally aligned and thus shared the same spectral composition. In this configuration, the perceptual segregation of the target and the distractors could only be done with the spatial binaural cues resulting from the spatialization with nonindividualized HRTFs. However, in the current work, the localization impairments do not seem to be due entirely to the spectral similarity resulting from the spatial disposition since they were observed only when four horizontally aligned distractors were displayed (but not two horizontally aligned distractors).

If we consider the target as the signal and the distractors as the noise, increasing the number of distractors resulted in a decrease in the signal-to-noise ratio (SNR). In the context of auditory localization, it is well known that azimuth localization abilities are impaired as SNR decreases (Kerber & Seeber, 2012; Lorenzi et al., 1999). For instance, a decrease in SNR has been associated with a decrease in azimuth gains, resulting in a lateral underestimation pattern in Kerber and Seeber (2012), as well as an increase in azimuth angular error in Lorenzi et al. (1999). Therefore, the

underestimation pattern observed in our study is more likely to be caused by a pitch-masking effect occurring when the SNR within a narrow frequency range falls below a critical threshold. In this case, it is plausible that participants had difficulties to segregate the target from the distractors due to perceptual fusion.

Not only the spectral composition of the sounds to be segregated influences segregation abilities, but their angular separation also plays a crucial role. When two sound sources are played simultaneously at lateral locations, minimal audible angle (MAA) between 10° for low tones (Perrott, 1984) and 20° for spatialized broadband sounds (Best et al., 2004) have been measured. These values are equal to or lower than the 20° angular separation used in our horizontally aligned condition. The maximum number of sound sources that can be separately perceived has been found to be about three for real tones ranging from 313 to 5051 Hz (Zhong & Yost, 2017), and between four and 5 five spatialized environmental sounds using nonindividualized HRTFs (Eramudugolla et al., 2005; Kawashima & Sato, 2015). In the abovementioned studies, the angular separation of the sound sources in azimuth was 30°, which is greater than the 20° separation used in our horizontally aligned condition. When multiple sounds are played from different locations, spatial separation prevents masking effects, a phenomenon known as spatial release from masking (Kawashima & Sato, 2015). Auditory experiments investigating the cocktail-party problem have shown that spatial release from masking based on binaural cues remains robust with broadband sounds, even when maskers are spatially distributed around the acoustic signal in both hemifields, as shown in Hawley et al. (2004).

If we turn to the leftward bias, we found that it was not systematic when the scene was complex but only when the distractors were vertically aligned with the target, or when they were not aligned but only two distractors were present. The presence of a systematic leftward bias when the distractors were vertically aligned with the target is consistent with the leftward bias observed in the minimalist scene (without distractors), since the azimuth location of the distractors and the target was the same in this complex scene disposition. However, it is unclear whether this bias has a perceptual or proprioceptive origin.

Taken together, the results concerning azimuth localization abilities in the complex scene show that spatial binaural cues can be used efficiently even when no target-specific spectral signature is provided, although this capacity seems to decrease rapidly when many distractors are located at the same elevation (horizontally aligned), resulting in a pitch-masking effect when the SNR is too low. If such masking occurred, then the masking effect was nevertheless incomplete since participants could still report that the target was located at a lateral location, even if they underestimated the eccentricity. However, when we measured the accuracy in localizing the azimuth of the target based on the azimuth unsigned error, there was no effect of the spatial disposition of the complex scene. If such a masking effect occurs, it should make it difficult to detect the appearance of the target among the distractors without the 440-Hz beep signal preceding the target display. Assessing the reaction time to detect the target appearance among distractors could help us confirm this explanation (for a related study, see Eramudugolla et al., 2005).

Elevation localization performance is not impaired in the complex scene

Localization performance for the elevation dimension was not greatly impaired by the presence of distractors, whatever their number or spatial disposition. The accuracy in the elevation dimension and the downward bias were not significantly different between the complex and minimalist scenes (15.9° and 15.8° for the unsigned error and − 14° and − 15° for the downward bias, respectively). However, the presence of distractors resulted in a compressive bias, as suggested by the fact that the elevation gain in the complex scene was significantly lower than the optimal gain of 1.0 (elevation gain of 0.78). This bias was not observed when the target was displayed alone in the minimalist scene, with an elevation gain of 0.86, which was not significantly lower than 1.0. However, this nonsignificant result has to be considered with caution because of the small sample size (19 participants), since it may have influenced the statistical power of the analysis.

The SSD used in this study conveys the elevation dimension in part through spatial cues provided by HRTFs, but mostly through pitch modulation in the frequency range [250, 1492 Hz]. The measured ability to segregate sound sources which do not share the same spectral composition is in line with the findings of Zhong and Yost (2017) in the field of auditory scene analysis, where the segregation of speech sounds and tones has been found to be based not only on spatial processing but also on other acoustic features such as pitch. In this latter study, multiple speech sounds or tones could be detected even when they were played from the same loudspeaker (i.e., the same location). However, it was rare for more than three sounds to be identified correctly when simultaneously played, even if none of the tones constituted a harmonic series and the employed frequencies were separated by at least 106 Hz. In the current study, although it is difficult to assess whether participants were able to perceive five distinct objects in the most complex configuration (with four distractors), they were clearly able to segregate the elevation location of the target.

The frequency range chosen for the SSD used for this study ([250 Hz, 1492 Hz]) covers a range of approximately 30 successive equivalent rectangular bands (ERBs) as

defined in Glasberg and Moore (1990). Theoretically, this would limit the elevation segregation abilities to a maximum of 30 distinct elevation locations although, in practice, the ability to segregate sound sources tends to be restricted to three to five simultaneous sound sources (Brungart et al., 2005; Zhong & Yost, 2017). However, the ability to segregate two (or more) objects located at distinct elevations depends on their proximity to each other (i.e., depends on the angular separation between elevations). In the current study, no more than five elevation locations were ever occupied simultaneously, namely in the situation when four distractors were vertically aligned with the target. In this spatial disposition, each object was separated by 12.5° in elevation. Given the size of the distractors and the target in the video frame with a vertical resolution of 120 pixels, we can estimate that each of the five objects was separated by at least two independent ERBs as defined in Glasberg and Moore (1990) and was therefore characterized by a specific spectral signature.

Overall, pitch modulation seems to be an efficient acoustic cue for use in SSD encoding schemes since it can be interpreted quickly and thus help localize the elevation of an object in a complex scene where it has to be segregated from other irrelevant objects, a common situation when moving around on foot. Due to the auditory filtering of the auditory system, the ability to segregate the auditory scene based on spectral cues depends on the frequency range and resolution. The current SSD, which uses a frequency range of 250–1492 Hz and a frequency resolution of 120 frequencies distributed across approximately 30 separate ERBs, seems sufficiently reliable to make it possible to distinguish one object among four others.

Implications for SSD design for the blind

Although the virtual environment used in the current study is still very different from a rich real environment such as a crowded street, our results point to segregation difficulties of objects in the visual scene on the basis of the SSD soundscape in some configurations. In a real context of SSD use, this situation would arise when multiple objects of similar heights are located at distinct azimuth locations and at a similar distance from the camera, such as a line of small posts in a street. Concretely, the results of our study suggest that in this case, the blind user would have difficulties to distinguish the distinct small posts. In a situation where a blind SSD user needs to reach a friend (the target) in a dense area composed of multiple objects such as small posts, a trash, a bench (the distractors), the blind individual could have difficulties to distinguish the friend arriving among this visual background.

Many SSDs use the modulation of the sound frequency for the elevation/vertical dimension (Capelle et al., 1998; Cronly-Dillon et al., 1999; Gonzalez-Mora et al., 2006; Mhaish et al., 2016; Neugebauer et al., 2020) and spatial acoustic cues for the azimuth/horizontal dimension (Bizon-Angov et al., 2021; Commère et al., 2020; Paré et al., 2021; Ribeiro et al., 2012; Richardson et al., 2019; Spagnol et al., 2017; Ton et al., 2018), with many SSDs combining both cues (Abboud et al., 2014; Ambard et al., 2015; Bordeau et al., 2023; Hamilton-Fletcher et al., 2016; Hamilton-Fletcher et al., 2022; Hanneton et al., 2010; Meijer, 1992; Stoll et al., 2015). Therefore, our results have implications for other SSDs, where similar localization patterns would be observed.

In our study, the degree of spatial separation between the target and the distractors was defined in order that distractors and the target largely occupy the field of view of the camera when the number of distractors is of four (the maximum). Therefore, different degrees of spatial separation were used in the azimuth (20° and 40°) and elevation dimensions (12.5° or 25°). With our SSD, the acoustic cues used for the azimuth and elevation dimensions were different (spatial acoustic cues only versus spatial acoustic cues and pitch modulation), but it could be important to test if the angular separation in both azimuth and elevation has an effect on performance.

The results raises the question of the size of the amount of information that is transmitted through the SSD soundscape. Reducing the flow of auditory information seems to be of value in order to prevent cognitive overload when other relevant information present in the scene has to be perceived and processed. In the current study, the horizontal field of view of the virtual camera was 90° and the objects (target and distractors) were located between −40° and +40° in azimuth. If the visual scene is too complex to permit accurate segregation by users then there is little point in conveying this large amount of information to them. Instead, with a narrower horizontal field of view, users could turn their heads to the side to make a lateral scan of the visual scene and thus temporally break down the complex scene into a succession of simpler ones. For instance, the SSD proposed by Neugebauer et al. (2020) transmits only the central column of the image via the soundscape (corresponding to a resolution of about 5.6°) to limit the amount of simultaneously transmitted auditory information. However, such a narrow field of view has the drawback of limiting the detection of lateral obstacles or mobile objects approaching from the side (i.e., common events in everyday use). In a real context of SSD use, we think that parameters should be modulated by the user itself such as the distance limit that is transmitted through the SSD soundscape, the visual field of view of the camera, and the frequency range and resolution used in the encoding scheme. In the context of really dense area, one would focus on closer visual information by conveying only the closer visual information.

In our study, we did not vary the distance between the distractors and the target. With our SSD, the distance of an object from the user (i.e., from the camera) is conveyed through intensity modulation, with higher intensity associated to closer visual information. Therefore, we could expect a stronger masking effect from the distractors if they were located closer to the camera than the target. For instance, in a protocol where the distance between the distractors and the target vary, we could predict lower localization abilities when the distractors are located closer to both the target and the participant.

It is interesting to note that in the current study, we used a brief familiarization period (90 s), whereas longer training sessions have been used in other studies—for example, 3 h in the studies of Auvray et al. (2007) and Pesnot Lerousseau et al. (2021). Training or experience can result in improving performance in the context of SSD (e.g., Proulx et al., 2008) or auditory localization experiments with non-individualized HRTFs (e.g., Mendonça et al., 2013, which used the same CIPIC database for the HRTFs). Therefore, better performance could be expected after training and experience with our SSD, such as higher precision and accuracy, as well as a decrease in the underestimation of the eccentricity of the lateral targets when 4 distractors are horizontally aligned with the target. Before testing a longer training, we could implement a familiarization session of the same duration, in which participants move a target in front of them while other distractors are displayed simultaneously.

The current study was conducted with sighted blindfolded participants and comparable research with blind participants remains to be performed. The localization task with distractors used in the current study was comparable with a cocktail party configuration, as described in Feierabend et al. (2019). These latter authors showed localization impairment in a cocktail party configuration, although the localization performance of blind people was comparable with that of sighted (but blindfolded) participants. Replicating the current study with blind participants might result in comparable or higher performance since studies have observed that blind participants outperform normal participants in auditory localization (Doucet et al., 2005; Voss et al., 2015) and pitch discrimination tasks (Gougoux et al., 2004). Although Doucet et al. (2005) and Voss et al. (2015) showed strong interindividual variability in azimuth localization abilities, they also suggested that the better localization abilities of the blind resulted from their more effective use of spatial spectral cues for the azimuth dimension. However, while showing that blind participants were more effective in the use of spectral cues for the azimuth dimension, Voss et al. also found, by contrast, that they had impaired elevation localization abilities, suggesting that they use spectral cues differently rather than possessing superior localization abilities. Since our SSD mainly conveys elevation through pitch modulation, elevation localization performance can be expected to be primarily dependent on pitch perception. While it has been found that blindfolded sighted participants find the use of pitch modulation for the elevation dimension in SSD to be intuitive (Bordeau et al., 2023; Stiles & Shimojo, 2015), it has been suggested that the cross-modal correspondence between pitch height and spatial height is weaker in the blind population (Deroy et al., 2016). Assessing the abilities to use SSDs in environment with increasing complexity is essential in order to increase our understandability of the difficulties occurring in real context of use of an SSD, and resolve these issues. For this reason, the protocol presented in our study has been designed to be easily replicated with blind participants.

Conclusion

The current study investigated the early-stage ability to localize a target in a minimalist and a complex scene using a visual-to-auditory SSD that uses spatial acoustic cues and pitch modulation to convert the image captured by the camera into a soundscape. After a brief period of familiarization with the principles of the SSD, blindfolded participants were able to perceive the location of a target in a minimalist scene composed only of the target. In the complex scene composed of a target and distractors, participants still succeeded in determining the location of the target, a task that was more difficult when no target-specific spectral signature was available.

This work suggests that the ability to segregate a complex visual scene on the basis of an SSD soundscape depends on the availability of a specific spectral signature when pitch modulation is used as an acoustic cue in the SSD encoding scheme. The study highlights the need to consider both the auditory filtering frequency of the auditory system and the resolution of the SSD's field of sonification in order to facilitate segregation abilities and limit perceptual overload, both of which are necessary in the context of SSDs for pedestrian locomotion assistance.

Acknowledgements This research was funded by the Conseil Régional de Bourgogne Franche-Comté (2020_0335), France and the Fond Européen de Développement Régional (FEDER) (BG0027904). The authors thank the Conseil Régional de Bourgogne Franche-Comté, France and the Fond Européen de Développement Régional (FEDER) for their financial support, and the Université de Bourgogne and the Centre National de la Recherche Scientifique (CNRS) for providing administrative and infrastructural support.

Authors' contributions C.B. and M.A. contributed to the conception and design of the experiment and interpreted the data. C.B. ran the study, performed data analysis and wrote the manuscript in close

collaboration with M.A. F.S., C.M., and J.D. provided important feedback. All authors have read and approved the manuscript and contributed substantially to it.

Funding This research was funded by the Conseil Régional de Bourgogne Franche-Comté (2020_0335), France and the Fond Européen de Développement Régional (FEDER) (BG0027904).

Data availability The experiment was not preregistered. The data that support the findings of this study are available on request from the corresponding author (C.B.).

Code availability The R code generated for the statistical analysis that support the findings of this study are available on request from the corresponding author (C.B.).

Declarations

Conflicts of interest/Competing interests The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethics approval The study involving human participants was reviewed and approved by Comité d'Ethique pour la Recherche de Université Bourgogne Franche-Comté (CERUBFC- 2021–12 - 21–050) and was performed in accordance with the ethical standards as laid down in the Declaration of Helsinki.

Consent to participate The participants provided their written informed consent to participate in this study.

Consent to publication The participants provided their written informed consent so the data collected during their participation is used for scientific publication.

References

Abboud, S., Hanassy, S., Levy-Tzedek, S., Maidenbaum, S., & Amedi, A. (2014). EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution. *Restorative Neurology and Neuroscience, 32*(2), 247–257. https://doi.org/10.3233/RNN-130338

Ahrens, A., Lund, K. D., Marschall, M., & Dau, T. (2019). Sound source localization with varying amount of visual information in virtual reality. *PLOS ONE, 14*(3), e0214603. https://doi.org/10.1371/journal.pone.0214603

Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). The CIPIC HRTF database. *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 99–102. https://doi.org/10.1109/ASPAA.2001.969552

Ambard, M., Benezeth, Y., & P, P. (2015). Mobile video-to-audio transducer and motion detection for sensory substitution. *Frontiers in ICT, 2*. https://doi.org/10.3389/fict.2015.00020

Auvray, M., Hanneton, S., & O'Regan, J. K. (2007). Learning to perceive with a visuo-auditory substitution system: Localisation and object recognition with 'The Voice.' *Perception, 36*(3), 416–430. https://doi.org/10.1068/p5631

Best, V., van Schaik, A., & Carlile, S. (2004). Separation of concurrent broadband sound sources by human listeners. *The Journal of the Acoustical Society of America, 115*(1), 324–336. https://doi.org/10.1121/1.1632484

Bizoń-Angov, P., Osiński, D., Wierzchoń, M., & Konieczny, J. (2021). Visual echolocation concept for the colorophone sensory substitution device using virtual reality. *Sensors, 21*(1), 237. https://doi.org/10.3390/s21010237

Bordeau, C., Scalvini, F., Migniot, C., Dubois, J., & Ambard, M. (2023). Cross-modal correspondence enhances elevation localization in visual-to-auditory sensory substitution. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1079998

Bregman, A. S. (1990). *Auditory scene analysis* [eBook]. MIT Press eBooks.https://doi.org/10.7551/mitpress/1486.001.0001

Bronkhorst, A. W. (2000). The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics, 77*(5), 1465–1487. https://doi.org/10.3758/s13414-015-0882-9

Brown, D., Macpherson, T., & Ward, J. (2011). Seeing with sound? Exploring different characteristics of a visual-to-auditory sensory substitution device. *Perception, 40*(9), 1120–1135. https://doi.org/10.1068/p6952

Brown, D., Simpson, A. J. R., & Proulx, M. J. (2015). Auditory scene analysis and sonified visual images. Does consonance negatively impact on object formation when using complex sonified stimuli? *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.01522

Brungart, D. S., Cohen, J., Cord, M., Zion, D., & Kalluri, S. (2014). Assessment of auditory spatial awareness in complex listening environments. *The Journal of the Acoustical Society of America, 136*(4), 1808–1820. https://doi.org/10.1121/1.4893932

Brungart, D., Simpson, B., & Kordik, A. (2005). Localization in the presence of multiple simultaneous sounds. *Acta Acustica United with Acustica, 91*, 471–479.

Bruns, P., Thun, C., & Röder, B. (2024). Quantifying accuracy and precision from continuous response data in studies of spatial perception and crossmodal recalibration. *Behavior Research Methods, 56*(4), 3814–3830. https://doi.org/10.3758/s13428-024-02416-1

Buchs, G., Heimler, B., & Amedi, A. (2019). The effect of irrelevant environmental noise on the performance of visual-to-auditory sensory substitution devices used by blind adults. *Multisensory Research, 32*(2), 87–109. https://doi.org/10.1163/22134808-20181327

Capelle, C., Trullemans, C., Arno, P., & Veraart, C. (1998). A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution. *IEEE Transactions on Biomedical Engineering, 45*(10), 1279–1293. https://doi.org/10.1109/10.720206

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*(5), 975–979. https://doi.org/10.1121/1.1907229

Commère, L., & Rouat, J. (2023). Evaluation of short range depth sonifications for visual-to-auditory sensory substitution. *AriXiv Preprints*. http://arxiv.org/abs/2304.05462

Commère, L., Wood, S. U. N., and Rouat, J. (2020). *Evaluation of a vision-to-audition substitution system that provides 2D WHERE information and fast user learning* [Tech. Rep.]. https://doi.org/10.48550/arXiv.2010.09041

Cronly-Dillon, J., Persaud, K., & Gregory, R. P. F. (1999). The perception of visual images encoded in musical form: A study in cross-modality information transfer. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 266*(1436), 2427–2433. https://doi.org/10.1098/rspb.1999.0942

Deroy, O., Fasiello, I., Hayward, V., & Auvray, M. (2016). Differentiated audio-tactile correspondences in sighted and blind individuals. *Journal of Experimental Psychology: Human Perception and Performance, 42*(8), 1204–1214. https://doi.org/10.1037/xhp0000152

Doucet, M.-E., Guillemot, J.-P., Lassonde, M., Gagné, J.-P., Leclerc, C., & Lepore, F. (2005). Blind subjects process auditory

spectral cues more efficiently than sighted individuals. *Experimental Brain Research, 160*(2), 194–202. https://doi.org/10.1007/s00221-004-2000-4

Elli, G. V., Benetti, S., & Collignon, O. (2014). Is there a future for sensory substitution outside academic laboratories? *Multisensory Research, 27*(5/6), 271–291. https://doi.org/10.1163/22134808-00002460

Eramudugolla, R., Irvine, D. R. F., McAnally, K. I., Martin, R. L., & Mattingley, J. B. (2005). Directed attention eliminates 'change deafness' in complex auditory scenes. *Current Biology, 15*(12), 1108–1113. https://doi.org/10.1016/j.cub.2005.05.051

Feierabend, M., Karnath, H.-O., & Lewald, J. (2019). Auditory space perception in the blind: Horizontal sound localization in acoustically simple and complex situations. *Perception, 48*(11), 1039–1057. https://doi.org/10.1177/0301006619872062

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research, 47*(1–2), 103–138. https://doi.org/10.1016/0378-5955(90)90170-T

Gonzalez-Mora, J. L., Rodriguez-Hernandez, A., Burunat, E., Martin, F., & Castellano, M. A. (2006). Seeing the world by hearing: Virtual Acoustic Space (VAS) a new space perception system for blind people. In *2006 2nd international conference on information & communication technologies* (pp. 837–842). https://doi.org/10.1109/ICTTA.2006.1684482

Gougoux, F., Lepore, F., Lassonde, M., Voss, P., Zatorre, R. J., & Belin, P. (2004). Pitch discrimination in the early blind. *Nature, 430*(6997), 309–309. https://doi.org/10.1038/430309a

Hamilton-Fletcher, G., & Chan, K. C. (2021). Auditory scene analysis principles improve image reconstruction abilities of novice vision-to-audio sensory substitution users. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC),* 5868–5871. https://doi.org/10.1109/EMBC46164.2021.9630296

Hamilton-Fletcher, G., Mengucci, M., & Medeiros, F. (2016). Synaestheatre: Sonification of coloured objects in space. *Proceedings of the 2016 International Conference on Live Interfaces* (pp. 252–256).

Hamilton-Fletcher, G., Alvarez, J., Obrist, M., & Ward, J. (2022). SoundSight: A mobile sensory substitution device that sonifies colour, distance, and temperature. *Journal on Multimodal User Interfaces, 16*(1), 107–123. https://doi.org/10.1007/s12193-021-00376-w

Hanneton, S., Auvray, M., & Durette, B. (2010). The Vibe : A versatile vision-to-audition sensory substitution device. *Applied Bionics and Biomechanics, 7*(4), 269–276. https://doi.org/10.1080/11762322.2010.512734

Hawley, M. L., Litovsky, R. Y., & Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America, 115*(2), 833–843. https://doi.org/10.1121/1.1639908

Hicks, S. L., Wilson, I., Muhammed, L., Worsfold, J., Downes, S. M., & Kennard, C. (2013). A depth-based head-mounted visual display to aid navigation in partially sighted individuals. *PLOS ONE, 8*(7), e67695. https://doi.org/10.1371/journal.pone.0067695

Howard, P. (1966). Human spatial orientation. *The Journal of the Royal Aeronautical Society, 70*(670), 960–961. https://doi.org/10.1017/S0368393100082778

Kawashima, T., & Sato, T. (2015). Perceptual limits in a simulated "Cocktail party." *Attention, Perception, & Psychophysics, 77*(6), 2108–2120. https://doi.org/10.3758/s13414-015-0910-9

Kerber, S., & Seeber, B. U. (2012). Sound localization in noise by normal-hearing listeners and cochlear implant users. *Ear & Hearing, 33*(4), 445–457. https://doi.org/10.1097/AUD.0b013e318257607b

Kolarik, A. J., Timmis, M. A., Cirstea, S., & Pardhan, S. (2014). Sensory substitution information informs locomotor adjustments when walking through apertures. *Experimental Brain Research, 232*, 975–984. https://doi.org/10.1007/s00221-013-3809-5

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13). https://doi.org/10.18637/jss.v082.i13

Kwak, C., & Han, W. (2020). Towards size of scene in auditory scene analysis : A systematic review. *Journal of Audiology and Otology, 24*(1), 1–9. https://doi.org/10.7874/jao.2019.00248

Lenth, R. V. (2022). *emmeans: Estimated marginal means, aka least-squares means*. R package version 1.7.4–1.

Levy-Tzedek, S., Hanassy, S., Abboud, S., Maidenbaum, S., & Amedi, A. (2012). Fast, accurate reaching movements with a visual-to-auditory sensory substitution device. *Restorative Neurology and Neuroscience, 30*(4), 313–323. https://doi.org/10.3233/RNN-2012-110219

Lorenzi, C., Gatehouse, S., & Lever, C. (1999). Sound localization in noise in normal-hearing listeners. *The Journal of the Acoustical Society of America, 105*(3), 1810–1820. https://doi.org/10.1121/1.426719

Maidenbaum, S., Abboud, S., & Amedi, A. (2014). Sensory substitution: Closing the gap between basic research and widespread practical visual rehabilitation. *Neuroscience & Biobehavioral Reviews, 41*, 3–15. https://doi.org/10.1016/j.neubiorev.2013.11.007

Meijer, P. B. L. (1992). An experimental system for auditory image representations. *IEEE Transactions on Biomedical Engineering, 39*(2), 112–121. https://doi.org/10.1109/10.121642

Makous, J. C., & Middlebrooks, J. C. (1990). Two-dimensional sound localization by human listeners. *The Journal of the Acoustical Society of America, 87*, 2188–2200. https://doi.org/10.1121/1.399186

Mendonça, C., Campos, G., Dias, P., & Santos, J. A. (2013). Learning auditory space: Generalization and long-term effects. *PLOS ONE, 8*(10), e77900. https://doi.org/10.1371/journal.pone.0077900

Mhaish, A., Gholamalizadeh, T., Ince, G., & Duff, D. J. (2016). Assessment of a visual to spatial-audio sensory substitution system. *2016 24th Signal Processing and Communication Application Conference (SIU)* (pp. 245–248). https://doi.org/10.1109/SIU.2016.7495723

Neugebauer, A., Rifai, K., Getzlaff, M., & Wahl, S. (2020). Navigation aid for blind persons by visual-to-auditory sensory substitution: A pilot study. *PLOS ONE, 15*(8), e0237344. https://doi.org/10.1371/journal.pone.0237344

Ocklenburg, S., Hirnstein, M., Hausmann, M., & Lewald, J. (2009). Auditory space perception in left- and right-handers. *Brain and Cognition, 72*(2), 210–217. https://doi.org/10.1016/j.bandc.2009.08.013

Odegaard, B., Wozny, D. R., & Shams, L. (2015). Biases in visual, auditory, and audiovisual perception of space. *PLOS Computational Biology, 11*(12), e1004649. https://doi.org/10.1371/journal.pcbi.1004649

Oldfield, S. R., & Parker, S. P. A. (1984). Acuity of sound localisation: A topography of auditory space I. Normal Hearing Conditions. *Perception, 13*(5), 581–600. https://doi.org/10.1068/p130581

Paré, S., Bleau, M., Djerourou, I., Malotaux, V., Kupers, R., & Ptito, M. (2021). Spatial navigation with horizontally spatialized sounds in early and late blind individuals. *PLOS ONE, 16*(2), e0247448. https://doi.org/10.1371/journal.pone.0247448

Perrott, D. R. (1984). Concurrent minimum audible angle: A re-examination of the concept of auditory spatial acuity. *The Journal of the Acoustical Society of America, 75*(4), 1201–1206. https://doi.org/10.1121/1.390771

Pesnot Lerousseau, J., Arnold, G., & Auvray, M. (2021). Training-induced plasticity enables visualizing sounds with a visual-to-auditory conversion device. *Scientific Reports, 11*(1), 14762. https://doi.org/10.1038/s41598-021-94133-4

Pourghaemi, H., Gholamalizadeh, T., Mhaish, A., Duff, D. J., & Ince, G. (2018). Realtime shape-based sensory substitution for object localization and recognition. *Proceedings of the 11th International Conference on Advances in Computer-Human Interactions.*

Proulx, M. J., Stoerig, P., Ludowig, E., & Knoll, I. (2008). Seeing 'where' through the ears: Effects of learning-by-doing and long-term sensory deprivation on localization based on image-to-sound

substitution. *PLOS ONE, 3*(3), e1840. https://doi.org/10.1371/journal.pone.0001840

Ribeiro, F., Florencio, D., Chou, P. A., & Zhang, Z. (2012). Auditory augmented reality: Object sonification for the visually impaired. In *2012 IEEE 14th international workshop on multimedia signal processing (MMSP)* (pp. 319–324). https://doi.org/10.1109/MMSP.2012.6343462

Richardson, M., Thar, J., Alvarez, J., Borchers, J., Ward, J., & Hamilton-Fletcher, G. (2019). How much spatial information is lost in the sensory substitution process? Comparing visual, tactile, and auditory approaches. *Perception, 48*(11), 1079–1103. https://doi.org/10.1177/0301006619873194

Scalvini, F., Bordeau, C., Ambard, M., Migniot, C., & Dubois, J. (2022). Low-latency human-computer auditory interface based on real-time vision analysis. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 36–40). https://doi.org/10.1109/ICASSP43922.2022.9747094

Spagnol, S., Baldan, S., & Unnthorsson, R. (2017). Auditory depth map representations with a sensory substitution scheme based on synthetic fluid sounds. In *2017 IEEE 19th international workshop on multimedia signal processing (MMSP)* (pp. 1–6). http://ieeexplore.ieee.org/document/8122220/

Stiles, N. R. B., & Shimojo, S. (2015). Auditory sensory substitution is intuitive and automatic with texture stimuli. *Scientific Reports, 5*(1), 15628. https://doi.org/10.1038/srep15628

Stitt, P., Picinali, L., & Katz, B. F. G. (2019). Auditory accommodation to poorly matched non-individual spectral localization cues through active learning. *Scientific Reports, 9*, 1063. https://doi.org/10.1038/s41598-018-37873-0

Stoll, C., Palluel-Germain, R., Fristot, V., Pellerin, D., Alleysson, D., & Graff, C. (2015). Navigating from a depth image converted into sound. *Applied Bionics and Biomechanics, 2015*(1), 1–9. https://doi.org/10.1155/2015/543492

Team, R. C. (2020). *R: A language and environment for statistical computing*. R Core Team.

Ton, C., Omar, A., Szedenko, V., Tran, V. H., Aftab, A., Perla, F., Bernstein, M. J., & Yang, Y. (2018). LIDAR assist spatial sensing for the visually impaired and performance analysis. *IEEE Transactions On Neural Systems And Rehabilitation Engineering, 26*(9), 1727–1734. https://doi.org/10.1109/tnsre.2018.2859800

Voss, P., Tabry, V., & Zatorre, R. J. (2015). Trade-off in the sound localization abilities of early blind individuals between the horizontal and vertical planes. *The Journal of Neuroscience, 35*(15), 6051–6056. https://doi.org/10.1523/JNEUROSCI.4544-14.2015

Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America, 94*(1), 111–123. https://doi.org/10.1121/1.407089

Zhong, X., & Yost, W. A. (2017). How many images are in an auditory scene? *The Journal of the Acoustical Society of America, 141*(4), 2882–2892. https://doi.org/10.1121/1.4981118

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *The Journal of the Acoustical Society of America, 33*(2), 248. https://doi.org/10.1121/1.1908630

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.