Journal on Multimodal User Interfaces (2022) 16:107–123
https://doi.org/10.1007/s12193-021-00376-w

ORIGINAL PAPER

SoundSight: a mobile sensory substitution device that sonifies colour, distance, and temperature

Giles Hamilton-Fletcher[1,2] · James Alvarez[2] · Marianna Obrist[3] · Jamie Ward[2]

Received: 29 November 2019 / Accepted: 9 June 2021 / Published online: 2 July 2021

Abstract
Depth, colour, and thermal images contain practical and actionable information for the blind. Conveying this information through alternative modalities such as audition creates new interaction possibilities for users as well as opportunities to study neuroplasticity. The 'SoundSight' App (www.SoundSight.co.uk) is a smartphone platform that allows 3D position, colour, and thermal information to directly control thousands of high-quality sounds in real-time to create completely unique and responsive soundscapes for the user. Users can select the specific sensor input and style of auditory output, which can be based on anything—tones, rainfall, speech, instruments, or even full musical tracks. Appropriate default settings for image-sonification are given by designers, but users still have a fine degree of control over the timing and selection of these sounds. Through utilising smartphone technology with a novel approach to sonification, the SoundSight App provides a cheap, widely accessible, scalable, and flexible sensory tool. In this paper we discuss common problems encountered with assistive sensory tools reaching long-term adoption, how our device seeks to address these problems, its theoretical background, its technical implementation, and finally we showcase both initial user experiences and a range of use case scenarios for scientists, artists, and the blind community.

Keywords Sensory substitution · Sonification · Audition · Depth · Thermal · Colour

1 Introduction and background

1.1 Basic overview of SSDs and issues

'Sensory substitution devices' (or SSDs) can continuously and systematically convert information normally associated with one sense (e.g. vision) into those of another (e.g. hearing, or touch) [1]. SSDs can be useful tools for exploring perception, neuroplasticity, and can operate as an assistive technology for groups like the visually-impaired [2, 3]. For vision-to-sound SSDs, their audiovisual pairings operate at the sensory level, in that fundamental visual features in the image (e.g. location, brightness, angularity) are communicated through specific auditory features (e.g. spatialisation, loudness, pitch-changes) [4]. These provide a consistent audiovisual pairing that allows listeners to mentally reconstruct the original image, upon which they can understand, explore, and interact with the 'visual world' around them. SSDs have the potential to provide an extremely flexible platform given that any visual feature could be mapped to any auditory feature. However, to date, the potential of SSDs has been highly constrained through technical limitations and design choices.

Early SSD research involved linking a moveable TV camera to an array of vibrating pins positioned on the users' back to create live 'tactile images' [5, 6]. The current version of this device, termed the BrainPort, instead uses an eyeglass mounted camera to control patterns of electrical stimulation on the tongue [7]. However, this is expensive for most users and requires extensive training [8], although users can also benefit from customising to their preferences [9]. The spatial resolution of tactile devices is limited by the number of contact points, and the relatively low spatial resolution of the skin. By contrast, auditory devices can conceivably convert each pixel of an image into sound and

✉ Giles Hamilton-Fletcher
giles.hamilton-fletcher@nyulangone.org

[1] Neuroimaging and Visual Science Lab, New York University Langone Health, New York, USA

[2] Sussex Neuroscience and School of Psychology, University of Sussex, Brighton, UK

[3] SCHI Lab, School of Engineering and Informatics, University of Sussex, Brighton, UK

are limited primarily by the ability of the perceiver [10, 11]. Moreover, auditory SSDs can leverage existing hardware in standard smartphones, increasing their accessibility, and potentially providing a pathway into sensory substitution in general. Furthermore, tactile and auditory approaches have been combined to further enhance navigation performance in visually impaired users [12].

Audiovisual SSD research has largely centered around the use of the vOICe [13] which focuses on converting grey-scale images into pure tone frequencies. From the camera stream, the vOICe 'snapshots' a single image and sonifies this over one second, scanning left-to-right through the image by sonifying each column in turn and converting the pixel information into a mix of pitch (denoting verticality), loudness (denoting brightness), and panning (denoting laterality) over-time, and then repeating this process for the next snapshotted image. This process is essentially the reverse of a 'spectrogram' and as such, they can also be used to confirm the preservation of visual information within vOICe soundscapes. The vOICe's approach is intuitive for simple visual images and patterns in highly constrained experimental settings [14], and has been used to explore shifts in behaviour, perception, and neural functioning for both sighted and visually-impaired users [15–18]. However, outside of the lab, there are clear shortcomings for the vOICe: complicated natural images produce unpleasant noise-like sounds; 'snapshotted' images eliminate real-time feedback; and important visual information does not always align with prominent auditory signals – which reduces the device's practicality. This results in a long, effortful, and frustrating learning period, before gains in daily functionality can be delivered.

Beyond the vOICe, this spectrogram-like approach of using pitch-height, panning-laterality, and loudness-brightness has underpinned many other SSD design choices with minor variations, such as presenting the information all-at-once (PSVA [19]; The Vibe [20]), musically (SmartSight [21]), or adding colour information through timbre (EyeMusic [22]). These design choices have been made upfront by their creators with a general lack of opportunity for users to modify the principles of sonification in light of their own experience or interests.

1.2 Lack of adoption

The concept of 'seeing through sound' has received a decent amount of interest in the visually-impaired community considering its lack of public profile. The vOICe has been downloaded regularly by visually-impaired users (~ 60 k on Android by 2016), and after trying a variety of SSDs, potential end users with visual impairments have rated their interest in the technology as high (8.4/10 [23]). Despite this, long-term adoption is rare, with only a handful of visually-impaired SSD experts known to the research community. Research into the reasons behind this lack of adoption can be organised into 3 main factors: hardware design, utility, and the wider situational context.

1.2.1 Hardware design

SSDs have faced a variety of criticisms across a range of devices. In particular, common concerns include: high costs; low availability; difficulty in setting up; being cumbersome; inability to wear easily; and only being 'easy-to-use' in simplified environments (e.g. white objects on black backgrounds) [1, 24–26]. Many of these issues can be avoided by using commonly available technology such as smartphones. These can be low-cost (or no additional cost), small, and familiar to users, and although the hardware is fixed, they allow for some degree of customization. For instance, while some smartphones now feature integrated 3D sensors, more basic models can have their functionality expanded through plug-in distance sensors, thermal cameras, 360° cameras, or micro-cams. Nonetheless, the wearability of smartphone SSDs remains an issue. If attached to clothing (e.g. belt, breast pocket) then the position of the sensor requires whole-body movements for active sensing. Hand-held and head-mounted options enable the user to actively sample from the environment, but have other issues such as not being hands-free (if held) or making the presence of the phone more visible to others—providing concerns around aesthetics or safety [23]. While current solutions might be viable for exploratory at-home use, these problems are likely to remain for public exploration until sensors can be discretely positioned on the body without risking the smartphone.

One potential solution to the wearability problem is the use of smartglasses technology, which provides smartphone-like functionality in a head-mounted set-up. While this does incur additional hardware costs, it has become an increasingly popular way for end users to utilise the vOICe. If the application is run on start-up, smartglasses can effectively provide a closed off sensory substitution system at the touch of a button. However, at present, the cameras on most smartglasses only produce a conventional 2D image. While these can contain cues to the 3D location of objects, the congenitally-blind may not be familiar with these cues [27], the functional resolutions novice users operate at may not be high enough to extract these cues [28], and extracting these in naturalistic environments may take extensive experience [18]. This approach can be contrasted against actual 3D depth sensing which is becoming an increasingly common feature of low-vision tools [29], smartphones, and is highly desired by end users [23]. Depth information allows the additional ability to separate out objects based on 3D location, which has the

dual benefit of prioritising near information and reducing a general sensory overload. While no one ideal solution exists at present, smartphone Apps remain the most readily accessible way for users to access sensory substitution.

1.2.2 Utility

Images can contain a wealth of important information, however if they are to become accessible for the visually-impaired, users need to be able to reconstruct them from their substituted form. There are essentially two information bottlenecks in the process of sensory substitution: one inherent in the design of the device itself (e.g. the choice of which information to sonify), and one inherent in the user (their ability to perceive or otherwise make sense of the information provided to them). In the ideal scenario these two would align: that is, one would only sonify the right type and amount of information to be useful for the user. In practice this can be hard, if not impossible, to achieve not least because sensory substitution is intended to be multi-purpose rather than task-specific.

For instance, the vOICe reduces information presented to the user by disregarding chromatic information and by sonifying a snapshot image, column-by-column over time [13]. In this approach, basic shapes and textures remain intuitive for users [14, 16], but users can have difficulty in counting and tracking multiple simultaneous frequencies in a soundscape [14, 30, 31]. Brown et al. [32] found that for single object shape discrimination, the spatial resolution being sonified by the vOICe could be reduced to as low as 8*8 pixels before novices experienced any drop in functional performance. But it is not always the case that 'less is best.' For example, introducing colour information (via timbre) to the soundscape will increase complexity, but these differences in timbre (e.g. guitar, flute) provide new distinct auditory objects in the soundscape for the user. This helps them segment targets from backgrounds [22], increases their effective spatial resolution [33], and provides key information for both object and scene identity [34, 35]. Knowing which information to retain in the sonification process will depend on the task at hand (e.g. prioritising depth information for navigation, colour for object recognition, or heat for cooking) and as such, SSDs should be flexible in their operation. Relatedly, the amount of information presented to users may need to be scaled according to their level of expertise or the task at hand—for example, progressing from sonifying the presence of a single feature (e.g. detecting light sources or single-point distance) to mapping out whole scenes (e.g. full coloured depth maps).

1.2.3 Situational factors and wider context

Further barriers to adoption include the lack of access to training, and the time and effort needed to develop expertise. The optimal balance between time-investment and functional-payoff has been difficult to identify and will likely vary between devices and scenarios [36].

Learning to use SSDs such as the vOICe and EyeMusic involve not only learning and deciphering the visual-to-audio conversion rules, but also learning (or re-learning) the rules that govern visual information more generally (e.g. perspective, occlusion, colour). Self-training is facilitated by online learning guides (www.seeingwithsound.com) or on-board training such as with the EyeMusic. By contrast, the 'voicevision' project (https://voicevision.ru/) provides private personal exposure training, and vOICe experts have been able to compete and win the Neurothlon 2018 games when competing against other synthetic vision-restoration approaches. This approach provides explicit guidance on what aspects of the signal to listen out for, building up their overall understanding piece-by-piece, and motivating users to reach the next milestone [24].

Finally, SSDs compete with other assistive technologies and user strategies. Overlapping function is a common reason for rejection: "I already have my dog for that" and "I can already get around just fine with my cane!" (pg 12–13 [26]). Immediately improving function alongside a positive user experience seems to have driven adoption of applications such as the 'Light Detector' for the fully blind, which converts overall luminance into a single tone. However, new technology may also interfere with previously learned skills. For instance, keeping the head still can be important in orientation and mobility (O&M) training, however active-searching with head-mounted SSDs might interfere with this strategy.

1.3 SoundSight: a mobile SSD

In order to address prior barriers to adoption, the SoundSight App takes a different approach to sensory substitution. In terms of hardware, it leverages the accessibility of smartphones to provide a low cost, versatile, and scalable sensory substitution solution. Here users can sonify their integrated cameras/sensors, as well as expand their functionality with additional plug-in sensor options. In terms of utility, the SoundSight App allows users to choose their preferred sensor / image-type as well as the audio files and presentation method that produce the final soundscape. This flexibility means that practical information can be prioritised (e.g. distance), as well as have this information presented in ways that aid comprehension (e.g. colour-timbre) or have high auditory aesthetics (e.g. music). To address situational factors like the necessity of training for functional gains,

the SoundSight can be scaled down and simplified for novice end users (e.g. single-point) to aid comprehension and immediate functionality, similar to other popular Apps for the fully blind that do not require user training (e.g. Light Detector). From this, users can expand their abilities and scale up over time, by adding complexity (e.g. expanding from single to multiple points simultaneously). This provides steppingstones for users to reach expertise. Furthermore, the SoundSight App can also function as a research tool. Researchers can represent multiple image types through a much wider range of sonification styles that they control.

The SoundSight App was originally developed in response to feedback from 10 visually-impaired end users (6 blind, 4 low-vision; age 30.4 ± 15.6; 4 female) who had tried out three SSDs covering a wide variety of design iterations: The Synaestheatre; the Creole; and the depth-vOICe ([23]; see also youtube.com/watch?v = AAEP2fpmxW8). Participants were able to use each device for ~ 15 min to freely explore an office space and images, before conducting an hour-long semi-structured qualitative interview. This feedback highlighted many of the issues identified earlier in the literature (e.g. hardware, functionality, training), however, they also provided more insights into what makes long-term adoption more likely. In aid of fostering a 'good user experience,' users were concerned that SSDs and assistive technology in general focused too much on functionality at the expense of aesthetics, with users requesting a pleasurable premium-feeling high-quality auditory experience. In terms of aiding function, users prioritised spatial information as the most relevant source of information and found the devices easier to learn when they responded immediately to movement of the sensor. As a result, the SoundSight App was originally designed to enable the rapid changing of hundreds of high-quality sound files in response to movement from distance sensors. With this foundation, the SoundSight expands to support a wide range of wearables, sensors (see Fig. 1), and sonification-styles that make it suitable for a wider range of tasks desired by end users such as navigation, cooking, or access to images. Crucially none of these choices are hard coded into the design, meaning that this flexibility allows for optimal combinations to rise to the top for end users, and for new questions to be explored within the scientific literature. Finally, the name 'SoundSight' was chosen based on simplicity, making the App easier to say, spell and search for than previous SSD solutions listed here (of which the Synaestheatre is the closest ancestor).

2 SoundSight architecture

The architecture of the app is split into three main parts; the sensors, the sonification controller, and the sound engine (see Fig. 2). In the application there are currently four main sensor objects which provide different pixel values: the iPhone's inbuilt camera (single lens [RGB], dual lens/LiDAR [RGB-D]), an external infrared depth sensor (Occipital Structure Sensor [D]), a thermal sensor (FLIR One [RGB-D]), and access to stored images and movies [RGB]. The selected sensors provide values for each pixel in a down-sampled image, which the sonification controller then uses as an input. As an example, the sonification controller can use pixel colour to determine timbre, pixel depth to determine volume, and sound file timing onsets from user settings (which is driven by an internally specified rhythmic pulse called the 'heartbeat'). The volumes and onset times

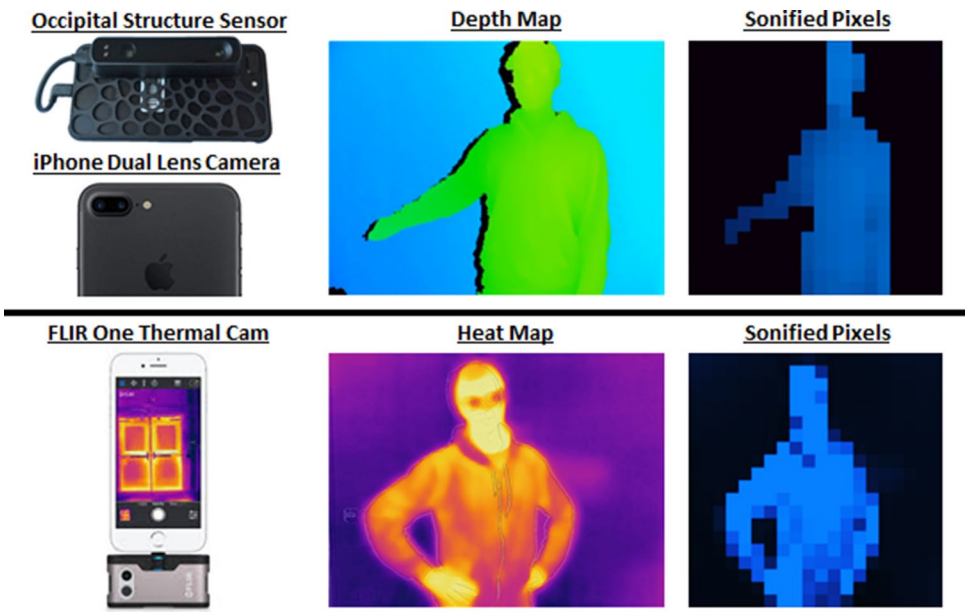

Fig. 1 The SoundSight App is able to take a variety of sensors (left column) and use their images (middle column) to drive the selection and loudness of thousands of high-quality sound files (each represented by a single pixel) in real-time (right column)

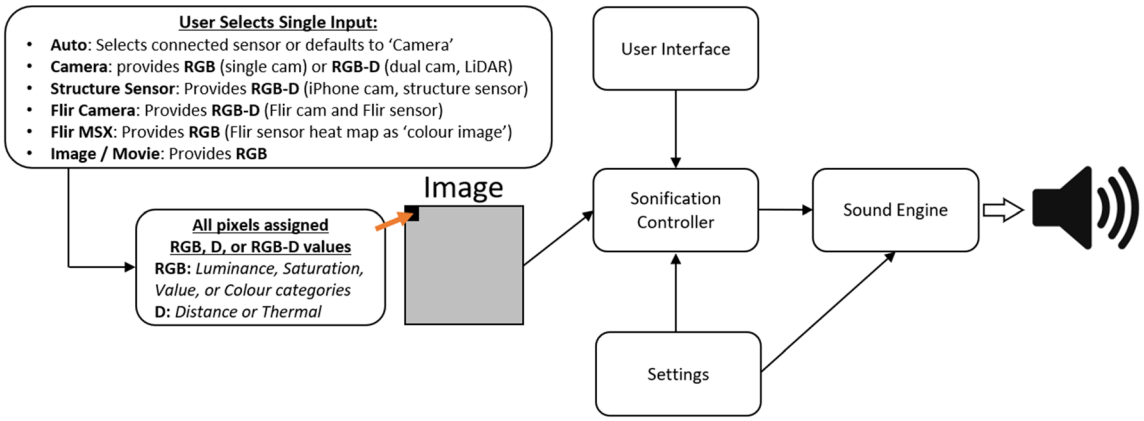

Fig. 2 SoundSight Architecture. The SoundSight consists of sensors, a sonification controller, and a sound engine. The user selects a single camera/sensor input option, which provides values for all pixels in a down sampled image. These values can control the loudness of an array of sound files that have been loaded into the system's memory by the sonification controller. The user interface and settings drive multiple factors of how this sonification occurs and in combination, the sound engine plays all of the sound files to produce the final soundscape

are sent to the sound engine, which controls the playback of the loaded sound samples. Currently the SoundSight app provides support for iOS. Future iterations of the SoundSight on Android will be designed for the specific sensing and processing capabilities of these devices. Before describing how these different parts of the architecture work in more detail, first we describe how the user can choose any sound as well as specifying how they are to be played. For convenience and standardization purposes the application also comes with a variety of inbuilt pre-specified options (Fig. 3).

2.1 User-driven sound specification

2.1.1 How visual properties control audio

The core of SoundSight is its ability to link the playback of an array of sounds (provided or user-uploaded) to real-time

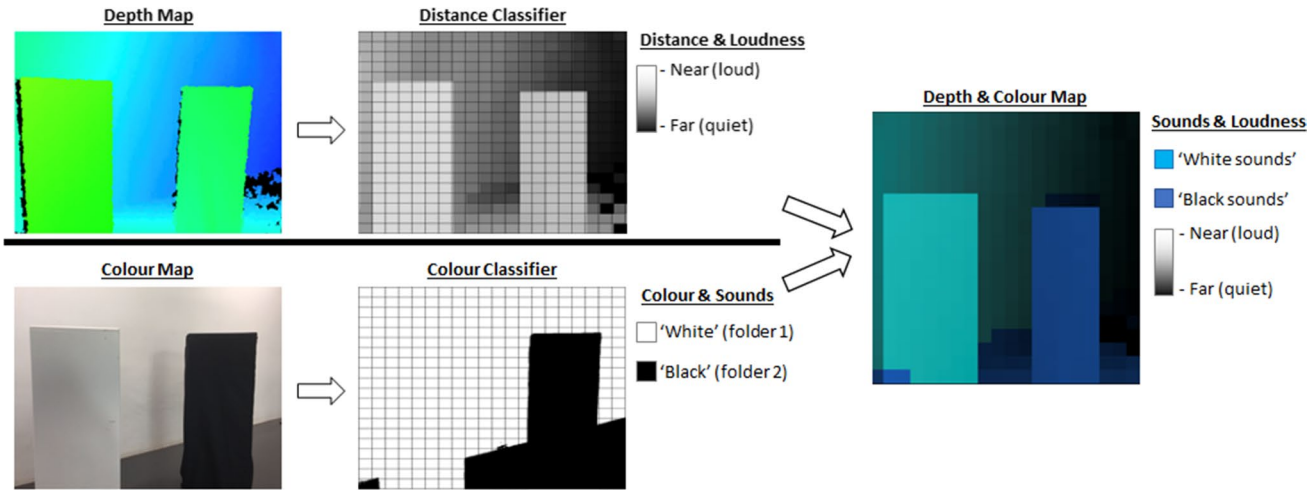

Fig. 3 An example of how sensor images can drive sound file selection. As illustrated in the first column, selected sensors can provide input values such as depth (or heat) and/or colour. Here, the depth image (top row) is provided alongside a colour image (bottom row). A colour classifier on the colour map selects which folders the sound files are selected from for each pixel, while a distance classifier on the depth map drives the loudness of these selected sound files for each pixel in the 20*20 array. Typically, different colour classifications would refer to folders that contain sounds that vary distinctly from one another (e.g. 'white folder' containing high-frequency sounds, 'black folder' containing low-frequency sounds). The colour and depth maps are co-registered together (either by default or external calibrator software) into the final 20*20 depth and colour map (final column), and this illustrates the selected sound file and loudness associated with each pixel location

information from the active sensors. Each mode has a 'synt' file (see Sect. 2.6) which specifies the X*Y resolution that the input image (e.g. depth map) will be down sampled to. Each pixel in this new image would then have its value (e.g. distance) assigned to control the loudness-level of a specific audio file in real-time. This is done via a naming convention, with '0.wav' being the top-left pixel, '1.wav' being the next pixel to the right, and so on until all pixels have been assigned their numbered audio file name.

2.1.2 Creating the audio array

How is the array of X by Y sounds created? The most straightforward example is to start with a single sound file (e.g. a sample of raindrops) that can then be systematically varied in its acoustic properties. This could be achieved by changing its spatial properties such that the first sound file sounds like raindrops coming from the top left, while the last sound file sounds like raindrops coming from the bottom right (and more files are created for everything in between). In natural hearing, the spatial properties of sounds are carried by differences in sound properties between the two ears that reflect relative differences in the timing and loudness of sounds, as well as distortions introduced to the sound by the head and outer ear (referred to as 'head-related transfer function' or HRTF). The SoundSight supports automatically creating spatialisation of audio files evenly along an azimuth angle set by the user using Apple's HRTF processing (see Sect. 2.6).

To create a matrix of sound files for use in the SoundSight, sound files are created externally, loaded into the iPhone via iTunes, and arranged according to a naming convention. As such, users can record or generate (e.g. https://fxive.com/) and then upload a series of original sound files, which are either arranged in the matrix via their file names, or a smaller number of sound files can be automatically spatialised by the SoundSight to fill out the matrix. This requires the same number of sound files as pixel positions from the down sampled input image (X by Y). For each additional colour category specified (N), this uses the same number of sub-folders (N), each of which contains its own matrix (X by Y). This allows the user to specify a sound for each colour across every pixel position (X by Y by N). Furthermore, future iterations will incorporate concurrent tactile feedback (phone vibration) to convey specific sensory features such as the depth value of the central pixel.

The following section explores how users and designers can produce their own auditory experiences with the SoundSight alongside important considerations necessary for the effective communication of visuospatial properties through auditory soundscapes.

2.2 Designing auditory experiences

2.2.1 Auditory properties

The spectral properties of individual sound files are important to consider in terms of how they will come together in the final soundscape. For example, sounds could be dynamic (e.g. a banjo pluck) or sustained (e.g. continuous rainfall). Dynamic sounds have envelopes that can be described in terms of ADSR: attack, decay, sustain, and release. In these sounds, there would typically be both shifts in the spectral energy (distribution of sound frequencies) as well as volume. This allows a precise control over the distribution of sound files over time, as each sound can fade to allow the next sound to be heard. This can be used to convey shape over time to the user. Whereas for sustained sounds, there would only be a volume shift when an object enters or leaves the sensor field (see Fig. 4). This results in motion being easier to track, as any movement is immediately followed by a change in sound. Combinations are also possible, such as with dynamic sounds with a long sustain. This allows the dynamic 'attack/decay' components to 'draw' shape information over time, with the sustain period constantly providing sound to 'catch' and convey motion as and when it occurs. Research is needed to understand the relative advantages of these types of sounds for sonifying different visual features, in terms of intuitive associations, information processing, and aesthetics.

2.2.2 Auditory perception

To ensure the highest possible functional resolutions for end users, audio designers should operate within the users' perceptual discrimination thresholds. In terms of frequency, adults can identify tonal changes of 0.2–0.3% in the range of 250-4000 Hz and get increasingly insensitive above 4 kHz [37]. Similarly, intensity discrimination can be done for level changes as low as 1–2 dB. For spatial hearing, listeners can discriminate a minimum audible angle between two pure tone locations on the azimuth plane to 1–2° in the central field. However, this exponentially decreases towards the periphery to 7–8° when the targets are 75° off-centre. Of note is that spatial discrimination also varies non-linearly with different pure tone frequencies (e.g. reduces for 1500-2000 Hz, or 8000 Hz) [38]. Discrimination of auditory elevation requires the use of broadband sounds that can have their spectral content differentially affected by factors such as listener's pinnae shape [39], with higher frequencies preserved in sounds above the listener, and dampened in those below [40]. Listeners can discriminate elevation to an angle of 3.65° [41], however since this is via natural hearing, reaching this acuity digitally may require the use of personalised rather than generic HRTFs. In addition, there should

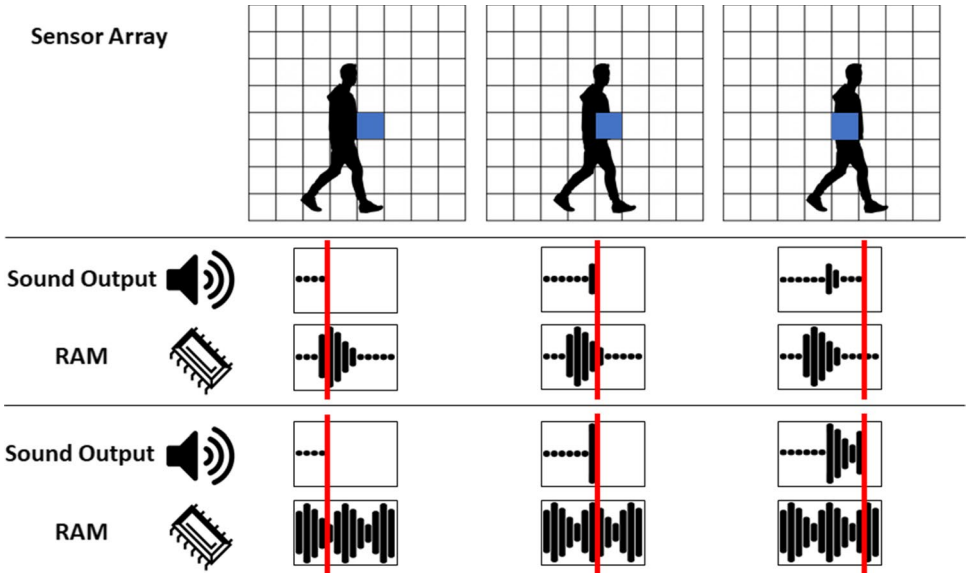

Fig. 4 Illustration of sound output associated with a single pixel. Here the sound output is shown for a single pixel (blue square) over time as a person walks into the pixel's active area. The RAM is repeatedly scanned through over time (red line). Output is shown for two different types of audio loaded into the RAM, either a dynamic sound (middle row), or a sustained sound (bottom row). When the person is not touching the active pixel (1st column) the audio loaded into the RAM is outputted at zero volume (silence), however when they reach the active pixel, the volume of the loaded audio increases (2nd column), and this continues for as long as the pixel is stimulated (3rd column). In the 3rd column, if the audio in the RAM is currently during a silent period (middle row), then this 'silence' is outputted as there is no audio affected by amplification. As the audio loaded into RAM is continuously outputted with only the volume varying in accordance with sensor stimulation, the audio can have its volume increased immediately and at any point in the scan through the audio in the RAM

be a consideration of the end user's perceptual discrimination abilities, as these are altered by a wide range of factors including musical-training, blindness, and age [42–44].

2.2.3 Communicating distance

Currently distance can be conveyed through timing differences and/or a linear/exponential ramping of amplitude of the audio files (see Sect. 2.3.2). Furthermore, there are options to create a sense of space through optional 'reverb' post processing effects altering wet/dry properties (0–100% reverb) using 13 representations of space (e.g. small room, Cathedral). However, while these options can enhance distinctiveness and aesthetic quality for the user, currently their use is not informed by external space. One such way this could be introduced into future iterations of the App is through recording external reverberation levels by the device's mic, or 3D mapping the environment to inform appropriate reverberation qualities. In particular, manipulating the initial time delay gap and ratio of direct-to-reflected sound cues can help to create a convincing estimate of environment size and feeling of perceptual presence [45–47]. Studies have also shown that these reverberant cues can influence visual processing of distance [48]. Furthermore, realism could be increased by using natural reductions in sound levels over distance (the 'inverse square law'), higher frequencies being more muted at further distances, and closer sounds producing higher level differences between the ears. However, psychophysical studies show that listeners tend to overestimate the distance of sounds in peripersonal space, and underestimate them in extrapersonal space, and so re-weighing these to suit listener bias' may result in better user accuracy [46, 49].

2.2.4 Naturalistic approaches: inspiration from blind individuals

In the 2016 film 'Notes on Blindness' Prof. John Hull describes how, in the 'real' world, the sound of rain makes silent objects audible ("…it was raining, I stood for a few minutes, lost in the beauty of it… the rain brings out the contours of what's around you, in that it introduces a continuous blanket of differentiated and specialised sound… if only there could be something equivalent to rain falling inside, then the whole of the room would take on shape and

dimension… instead of being isolated… you are presented with a world…"). Here differences in auditory feedback from rain falling on different objects in the environment like grass or cars allow the listener to localise these in the environment. Through the SoundSight app we can create a comparable auditory augmented reality using rainfall sounds, with different visual characteristics altering the audio, such that green objects may play the muffled sound of rain falling on grass, while red objects may play a more reverberant sound, similar to rain falling on the metallic surface of a car. But there is no strong commitment to raindrops. In fact, our pilot users wanted a wide variety of auditory experiences, ranging from distinctive 'clicks/snaps' that are reminiscent of the tapping of the cane, to otherworldly synth sounds, to out-and-out music. While there is a huge amount of potential for fun and creative sonifications, other practical factors should also be considered, such as avoiding audio which is difficult to understand, or that could mask or be confused for natural sounds.

2.2.5 Abstract approaches: tones, rainfall and music

As people are relatively poor at judging the vertical position of sounds using only HRTF, another possibility is to represent verticality in an abstract manner such as by using differences in pitch, similar to those used by the vOICe. Here the SoundSight can replicate this, using puretones to convey shape information (youtube.com/watch?v=_O3s9IWrRgA). Pitch-height mappings can also be conveyed using a completely different timbre, for instance, by vertically arranging rainfall sounds that have been constrained to higher or lower frequency ranges (youtube.com/watch?v=5L9mUE1YcAc). Finally, we have recently incorporated multi-track audio from a single musical piece, where different constituent tracks fade into and out of a coherent musical piece, driven by the position of objects in 3D space. For example, vertical space can be represented through vocals, guitar, synth, bass and drums, ordered from top to bottom (youtube.com/watch?v=RNUcLCq7ytM). We strongly encourage the reader to visit the above videos to gain a better understanding of the capabilities of the SoundSight App. This approach leaves open the option to convey complex visual information through varying components of full musical tracks from any genre—funk, rap, rock etc., making it suitable for a wide variety of musical preferences by different end users. In addition, by constraining the total information (e.g. just sonifying the Y axis), users can map out entire scenes by moving the sensor back and forth to map out the X axis. This method eliminates the need for using spatialised sound or headphones, as all the essential auditory information can be produced by inbuilt phone speakers. Ultimately, users can choose the type of audio and presentation style most appropriate to them, selecting the level of aesthetics, distinctiveness, and complexity that provides the best user experience or functionality during exploration.

2.2.6 *Abstract approaches: m*etaphors, perceptual spaces and associations

The quality of sound can be used to convey specific features or aid in the perceptual segmentation of the image. For instance, cold or hot objects detected by the thermal cam could play sounds that are natural metaphors for heat, like howling winds, a crackling fireplace, or cold/hot water being poured. This could aid users in their environmental interactions as this conveys expectations as to the tactile feel of an object prior to physical interaction [23]. If audio designers wish to increase the distinctiveness of specific visual attributes, sounds can be chosen for each of these properties that are perceptually dissimilar from one another [50]. Designers can even go further to preserve the multidimensional structure of a visual space in sound, for instance, the perceptual distance between specific colours can be communicated through a similar perceptual distance in their auditory representations [34]. However, perceptual similarity is not just evaluated in terms of acoustical features (e.g. bass or drum sharing lower frequencies), but can also be done through causal inference (e.g. bass or violin both occurring from a plucked string), or semantic identity (e.g. musical vs natural sounds) [51]. Furthermore, sounds can be chosen that reflect intuitive associations to visuospatial features in fully blind individuals [52, 53].

Here we explored potential sonification styles in terms of their auditory dynamics, perceptual discriminability, and illustrated a variety of overall 'themes' and their potential impact. Having considered the design of the final soundscape, the next section explores the technical specifics of how these are produced from the sonification controller, sensors, and sound engine working together (see Fig. 2).

2.3 The sonification controller

The sonification controller receives information (colour, depth, heat) from the active sensor and this specifies the volume of each sound file in the array. The user can also specify the timing onset assigned to each sound file according to position or colour category. This volume and timing information is passed to the sound engine, and updated in real-time (~ 30 ms), which gives the device the ability to sonify rapid changes in an image such as the movement of objects. This is possible because all sound files in the array are 'played' simultaneously (i.e. active within RAM) but, typically, most will be silent. Whilst it may seem counter-intuitive to play sound files with zero volume (as opposed to not playing them at all), the advantage of this approach is that it avoids the lag associated with constantly loading

and reloading sound files for playback. This is key for controlling thousands of high-quality sound files in real-time, while keeping changes in the overall soundscape smoother. Overall, this approach allows our sonification controller to obtain a higher quality and responsivity in auditory outputs compared to previous approaches in sensory substitution.

2.3.1 Heartbeat and sound onset

All of the audio files are played once within each soundscape, and the length of time this is played before starting over is determined by the 'heartbeat' timing parameter: so named because it can produce a pulsing rhythmic quality with certain sounds. This timing of starting the soundscape over, is not constrained by the length of the constituent sound files. As already noted, each audio files' timing onset within the soundscape is set by the sonification controller. This timing can be set by horizontal, vertical, and depth position, as well as by colour category. For example, horizontal offsets would start playing the sound files in the leftmost column first, adding in the sound files from each successive column to the right over time, similar to the left-to-right scanning of the vOICe SSD. These onsets can act to separate sounds over time according to position or colour, which can also act as cues to perceptual grouping (e.g. objects with different colours could pulse out-of-sync with one another). When all timing onsets are set at zero, all sounds begin simultaneously, which is suitable for musical tracks, where timing is important for aesthetics. When a timing onset is set to '1', then the sound onsets are spread evenly across the entire heartbeat interval. For dynamic sounds, it is good to allow sufficient time for the envelope to be played, otherwise there may be silent periods in which sensed objects cannot be heard (see Fig. 4).

2.3.2 Depth, simple colours and volume

With the heartbeat (and associated timing offsets) controlling the onset of the sounds at a regular pace, the volume of these sounds are updated in real time at the speed of the sensor updates. For depth sensors, the user can set two parameters which map the depth to the volume of the sound: the closest depth, and the range. The closest depth marks the point at which objects closer than this value are sonified at maximum volume, so if this is set at 500 mm, any objects closer than half a meter are sounded at maximum volume. The depth range determines the drop off rate, so if the depth range is 3000 mm (with closest depth as 500 mm), then objects further than 3500 mm are silent, and objects halfway (e.g. 2000 mm) are played at half volume. This volume fall-off can be linear or exponential (see Fig. 5). This allows the user to adjust the sonification to provide more detail for close objects or far away objects as required. While the maximum and minimum volume are usually 100% and 0% respectively, they can be altered to other values. Occipital recommends using the Structure Sensor mk1 up to 3.5 m indoors, and although distance values are reported up to 9 m, the image becomes increasingly degraded and unreliable. The SoundSight can use any furthest depth value (e.g. 9 m), which may be more suitable for depth sensors which have further operating ranges than the mk1 version. When connected to the Flir thermal sensor then the same principle applies such that hot objects are loud and colder objects are silent—or alternatively the heatmap can be read as a simple or complex colour image.

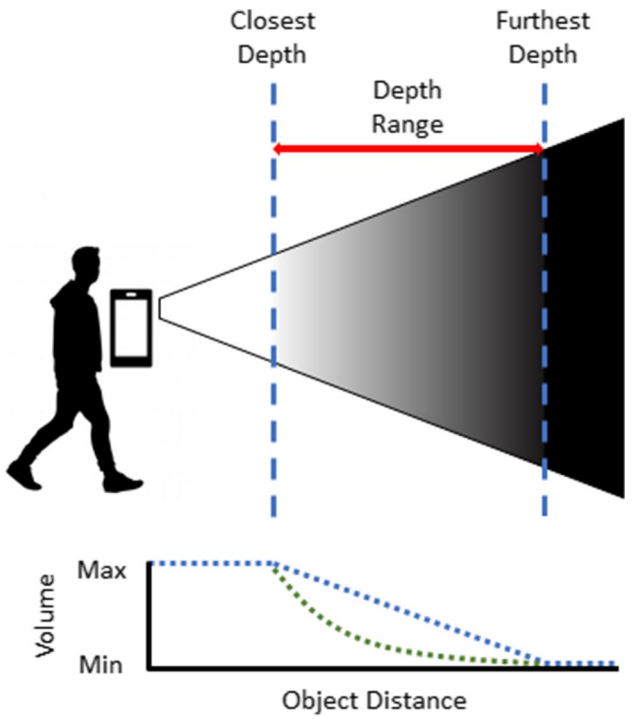

Fig. 5 Illustration of how distance parameters produce sounds of different volumes. Objects closer than the 'closest depth' value are sonified at maximum volume (e.g. 100%), objects further than the closest depth value get progressively quieter in either a linear or exponential fashion (see the blue & green dotted lines respectively), objects more distant than the furthest depth value are at minimum volume (e.g. 0%). The values for the closest depth, depth range, maximum and minimum volumes are modifiable

For camera images where each X,Y point only has a single colour value associated with it, this value can be used to directly control the loudness of a single array of sound files in either a linear or exponential manner. This single value can be luminance (e.g. greyscale images) or utilise saturation/value from HSV colour space. This allows either bright, colourful, or both resulting in louder sounds. The use of a greyscale image, with a horizontal timing offset and an array of dynamic sounds varying in frequency (Y-Axis) and

spatialisation (X-Axis) results in a vOICe-like replica, albeit updating in real-time rather than from static snapshots.

2.3.3 Complex colours and sound selection

For a sensor array of X by Y RGB integers (where X and Y are pixel positions in 2D space, and RGB is the pixel's colour content), this will drive an array of X by Y by N sounds (where N represents the number of distinct colour categories set from colour space). The number of colour categories will be the same as the number of folders, with each folder containing its own X by Y array of sounds. For instance, pixels classified as 'yellow' may be assigned to a folder containing an array of violin sounds, while 'blue' classifications are assigned to a folder of piano sounds, etc. Of course, the assignment of colour to sound need not be based on timbre but could be based on any auditory feature. Prior research has shown that there are more intuitive ways of assigning colours to sounds for the purposes of sensory substitution [35, 54].

2.3.3.1 Discrete colour mode In this mode, each pixel is classified to a specific colour category and plays the sound from the same array position in the relevant 'colour' folder. The colour category that each pixel is assigned for each frame is determined by the ranges that the pixel's HSV colour value falls within. While colour values are natively provided in RGB space, this is transferred to HSV space (hue, saturation, value) due to its low computational cost, which helps to quickly categorise and segment colour space. Here colour category is determined by the following process: if the pixel's saturation is above the saturation threshold set by the user then a saturated colour category is chosen based on which colour category value the hue value is closest to (e.g. red, orange, yellow, green, blue, purple); if the saturation is below a set threshold then an unsaturated colour is chosen based on user set values for black, grey, or white. Given that each 'pixel' has been down sampled from a higher resolution image, each colour is determined by the value of the central pixel in that down sampled area (rather than an average of that region). This is done so that a 'pixel' that is down sampled from a region that contains red and yellow will be sonified as either red or yellow, and not be blended into the colour orange. The number of saturated hues and desaturated shades is customisable but is typically set to correspond to basic colour categories. The model cannot specify pinks/brown as these are dependent on both lightness and hue, pink being light red/purple, and brown being dark orange/yellow. This was a compromise to keep the complexity of the colour model simple to aid in responsivity. Once the colour for a pixel is chosen, it's assigned sound has the potential for its volume to be increased (subject to the dimension driving loudness—e.g. distance), with all other potential colour-sounds for that pixel being completely silenced. Discrete colour mode is suitable even with high spatial resolutions as there is always a fixed number of audio files being played. Furthermore, the added auditory contrast between different coloured areas can actually increase the user's functional resolution [33].

2.3.3.2 Blended colour mode While the discrete colour mode simplifies the colour information into categories—meaning that all shades of red play the same 'red' sound file (and are thus indistinguishable to the user), the blended colour mode allows a smooth transition between different hues through mixing the two nearest colour categories (and their auditory representations) together. This allows the user to determine how close each pixel is to a specific focal colour as well as the perceptual relationship between colours (e.g. that orange is perceptually close to yellow). To accomplish this, instead of playing one sound per pixel, two can play, with their respective volumes tuned to how close the pixel's colour is to the nearest focal colour (e.g. yellow, green). This operates via a linear interpolation between the two closest sounds. Say for instance a pixel is reddish-orange, in this case both the red and orange colour-sounds play, with the volume of each determined by how close the pixel's colour is to each exemplar colour's inputted hue value. If the colour is perfectly red, then only the red sound file plays. This blended approach enables the congenitally blind to gain perceptual experience in understanding how colour is organised for those with prior visual experience [55]. However, the additional sound files being played increases the complexity of the soundscape and chance of cacophonous conflicts, as such, the more complex colour representations like blended mode may be easier for users to understand with lower spatial resolutions (e.g. single-point).

2.4 Sensors

There are a variety of potential input options including sensors both internal and external to the phone. When a new setup (or 'patch') is loaded, external sensors are searched for, and if none are found, internal sensors/cameras are defaulted to. Here the detected sensor is loaded and passed to the sonification controller, this results in the chosen sensor becoming a dependency of the sonification object. However, since the sensor class typically provides a stream of updates, this sensor triggers the flow of states throughout the sonification controller by means of a callback. Typically, the sensor will receive new data, store it raw in memory, then signal to the sonification controller to update. On receiving this signal, the sonification controller calls methods on the sensor to receive data in a uniform format. For depth data, this is an array of X by Y floating point values of mm distance, while for colour data, this is an array of X by Y

RGB colour integers. The user can pick which sensor is to be used, and audio alerts are given when the sensor is not providing updates.

Each sensor provides data (depth, thermal, colour) in a given native resolution. The user provides a new resolution corresponding to the setup of sound files, e.g., if there are 15×15 sound files to map, then the native resolution is resampled to this new resolution. Additionally, the user can adjust the 'window' of the resampling on the image itself, both horizontally and vertically. This is the equivalent of zooming, say if the window is set at 50% horizontally and 50% vertically, then the 15×15 resampling takes place across a central area spanning 25% of the total space in the original image. Finally, the user can also specify that only a subset of audio files remain active. So for a 15×15 image, one could eliminate peripheral rows and/or columns to only sonify the central row (15×1) like the See ColOr [56], or only sonify the central column (1×15) like the EyeSynth (eyesynth.com/?lang = en), or even just the central pixel (1×1) like the enactive torch or EyeCane [57, 58]. As such, the SoundSight has the ability to replicate multiple styles of sensory substitution in a single app albeit with more flexibility in sound selection (see Fig. 6).

Despite the multitude of options, users do not need a detailed knowledge of the working of the app. Most of these settings will be soft-locked for each audio mode in order to provide the best user experience, with options for precisely tuning specific sensor and audio parameters using gestural controls such as swipes and pinches. This tuning is done with optional audio or visual feedback. The parameters that can be controlled via gestures can be selected in advance by the designer, or manually set by the user. In the following section a variety of supported sensors are described.

2.4.1 iPhone inbuilt camera

The SoundSight app supports multiple types of inbuilt cameras on modern iPhones. For iPhones with only single lens cameras, 2D colour images are streamed, allowing the use of simple colour images (e.g. greyscale images) or complex colour images (e.g. colour categorized images). The parameter controlling the loudness of each audio file can be set to colour values such as luminance or saturation, or 'none' (which plays all audio at a set loudness). If the iPhone features a dual lens camera, then the stream contains co-registered colour and depth information in synchronized packets. This provides a depth parameter which can be used to control audio file loudness. This system is also built to support modern integrated sensors that provide synchronized depth and colour packets (i.e. FaceID, LiDAR).

2.4.2 Structure sensor

For precise depth maps, an external Structure Sensor (Occipital ltd) can be plugged into the iPhone. Here arrays of distance information are provided from the infrared Structure Sensor and 2D colour images are provided from the iPhone's inbuilt camera. These arrays are synchronised into a final RGB-D array via Occipital's SDK. An app provided by Structure calibrates the relative positions of the Structure

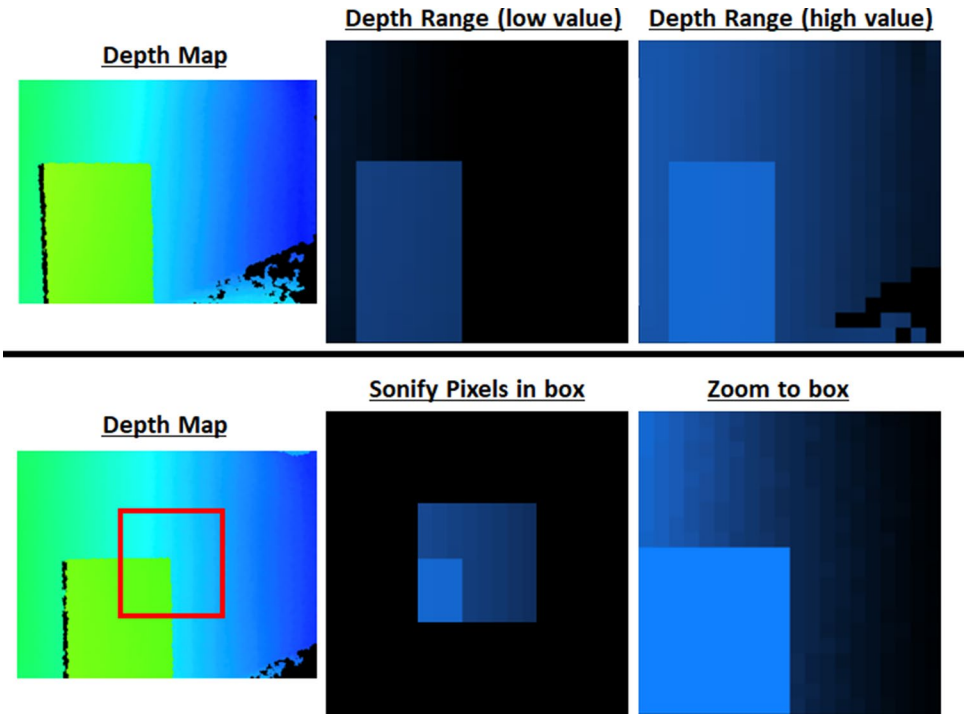

Fig. 6 Illustration of how a variety of options affect the sonification of a depth map image (left column). In the top row, we show how changes to 'depth range' values can eliminate the sonification of further objects (top middle image), or expand the range to sonify objects a further distance away (top right image). Users also have the option to focus on specific areas of an image by either eliminating peripheral rows and columns of 'pixels' (lower middle image), or by expanding the total image to fit the red bounding box in the lower left image, effectively creating a 'zoom' function (lower right image). In these examples the closest depth value remains constant

Sensor and camera to ensure that depth and colour data are accurately co-registered in the final arrays that are sent.

2.4.3 FLIR one sensor

The FLIR SDK provides synchronized temperature and 2D colour data from its own built-in camera. When the FLIR One is plugged in, this RGB camera is used as the default. The FLIR One SDK provides temperature in Kelvin, but the SoundSight app fixes the active temperature window to lie between 0 and 50 degrees Celsius (with temperatures above and below that defaulting to specified minimum and maximum values). When sonifying only temperature, we default to hotter objects being louder to increase the saliency of a variety of potential risks, hot objects, people, or service animals. Temperature data is treated as equivalent to depth data in the app, such that 50C is equivalent to 0 mm, and 0C to be 3000 mm, thus we applied a formula to convert Kelvin into depth: mm = −60* (K—273.15) + 3000. By treating heat information as equivalent to depth information within the application, the minimum, maximum, closest depth (heat), and (heat) depth range values can be specified and controlled by the user. The FLIR One SDK also allows the outputting of a blended MSX RGBA image format, this is a hybrid of a coloured heat map (red indicating heat and transitioning to blue for cold) with additional edge detection from the accompanying RGB camera. The SoundSight app is able to read this as a coloured image, so that categories for various colours/sounds can be used to indicate different heat levels.

2.4.4 Image/movie sensor

This 'sensor' option provides the ability for the user to load an image or movie from the iPhone, and to have its colour data sonified. An internal 20 fps timer starts the sonification callback, and the 'sensor' provides colour data during active playback (with automatic repeat). This data is otherwise treated identically to live RGB camera images, and can be sonified according to simple colour values (luminance, saturation, or value) or assignment to complex colour categories.

2.5 Sound engine

The sound engine loads all sounds from the currently selected audio patch into memory to avoid any latency. The number of sound files loaded is determined by the size of the array and the number of colours. For instance, with a 7 by 11 array, with 7 colours, the number of files is $7 \times 11 \times 7 = 539$. Sound files are temporally arranged in memory according to their allotted time offset (see above 'heartbeat and sound onset'), and the audio rendering procedure then mixes all sounds simultaneously (usually with the majority being silent). The sounds do not loop, and so require constant triggering, which happens at the rate of the heartbeat in the sonification controller. Volumes are updated in real-time, via a separate pathway. To avoid artefacts due to noise, the rate of volume change is limited internally by a ramp with a set speed.

The amount of audio files that can be loaded is limited only by the amount of RAM accessible to the sound engine—i.e. the summation of the size of all audio files cannot exceed this RAM limit, otherwise the patch will autoclose. In general, this increases with more and longer sound files. A designer-specified limit of 1 GB is used, although this is primarily to have a consistent experience across a range of devices, some of which may have lower RAM sizes than others. This value effectively provides the upper 'resolution' limit of the SoundSight app. During use this upper resolution can be dynamically reduced, through silencing peripheral rows and columns to focus on central regions (see Fig. 6). Designers can work within the RAM limitations to prioritise different types of resolution—for instance, higher spatial resolutions can be achieved through the use of brief sound files with many X and Y points in an array, while higher colour resolutions can be achieved by prioritising RAM for additional folders for each colour category. It is worth remembering that the primary bottleneck in information processing tends to be with the perceiver rather than the SSD [30–32], and hence focusing on sound qualities that enhance user performance are likely to have larger performance gains than straight increases in spatial resolution.

2.6 Settings and user control

All of the parameters that are adjustable in the settings screen, the parameters for sound files, and the settings for colours are specified in a human-readable configuration file, called a '.synt' file. A synt file is a JSON formatted dictionary with a pre-determined structure, with keys specifying multiple adjustable parameters for the SoundSight. Here is an example of the contents of a simple synt file:

```
{
"rows": 20,
"cols": 20,
"name": "Puretones",
"hrtf": 1,
"hrtf_angle": 60,
}
```

This synt file specifies that the input image will be subdivided into 20 rows and 20 columns, this creates 400-pixel locations which would eventually require 400 audio

files (named 0–399.wav). The name parameter determines the directory in which the sound files will be found and/or created to. A 'hrtf' value of 1 specifies that these 400 audio files will be created from a smaller subset of stereo sounds that are located in a 'hrtf' subfolder (e.g. Puretones/hrtf/). Here the user places audio files equal to the number of rows (i.e. 20, named 1–20.wav), and that the algorithm will create new audio files that are spatialised across all column positions. The 'hrtf_angle' parameter specifies that an angle of 60 degrees (centred on the forward position) should separate the leftmost and rightmost sounds. This HRTF-preprocessing assumes an equal angle between each column, so a 60° angle divided by 20 column positions results in each column being 3° spatially separated in azimuth. These new HRTF-spatialised audio files end up being outputted to the main folder using the SoundSight's naming convention, so that the sound engine can immediately use them in operation. It is also possible to fill out the entire array from just a single audio file. Here a file named 0.wav in the /hrtf/ folder can be pitch-shifted into a variety of new audio files (e.g. 1.wav, 2.wav), in the /hrtf/ folder, which can then have HRTF processing applied to fill out the entire array.

While all parameter settings can be set in the synt file (e.g. heartbeat interval, timing onsets), users can also manually adjust them during use and save these values to the synt file. These can be used by researchers to create a desired setup, share the same mode across multiple devices, or simply restrict users to a particular set up. A full documentation of the synt file can be found as supplementary material.

Visually-impaired users and creators have the ability to set up their own modes in a variety of ways. If a user wishes to create a mode from scratch, the synt file is editable as a text-file, the user can upload the relevant synt file and folder containing the constituent audio files via iTunes, or they can modify and save changes to the synt file during use. Users and creators have the ability to edit both their modes and those of others, so that they can be modified to meet their specific preferences. In the future, it may be possible to further streamline this through an automated stage-by-stage guide to assist users through the creation process.

2.7 Gestural controls

Almost all of the stated options and parameters can be assigned to standard gestural controls (e.g. swiping, pinching, tapping), where users can precisely tune values during device operation. There is also the option of visual or verbal feedback on the newly assigned parameter values. This is useful, both for blind users understanding how changing various values changes sonification, but also for experimenters who can map multiple features simultaneously to quickly evaluate options. Novice users are likely best limited to a carefully-selected subset of key options (e.g. swiping to control depth range) in order to keep the best possible practical user experience, or alternatively, having gestural controls disabled in order to avoid accidental changes in sonification during device use.

In summary, the SoundSight is a highly flexible mobile app-based sensory substitution device that can sonify, without perceptible lag, data from different input sources (RGB, depth and thermal cameras). In the final two sections, we explore feedback from visually-impaired users for the SoundSight and discuss potential usage scenarios in the future across multiple fields.

3 Preliminary end user testing

To provide an example of the flexibility of this approach, we conducted end user testing with both fully blind ($N=3$) and visually-impaired ($N=4$) participants recruited via advertising at a local blind charity (age $=32.6 \pm 16.9$; 2 female). Testing was ethically approved by the University of Sussex and conducted in accordance with the Declaration of Helsinki. Here we evaluated their perceptions of a variety of sonification styles during freeform exploration of an experimental room featuring walls, tables, and chairs. Participants were provided with an iPod and Structure Sensor and had 10 min with each sonification mode to explore the room. Each mode provided the same spatial information from the structure sensor (Field of view: 58° horizontal × 45° vertical; Range: 3 m; Resolution: 7×7) and was sonified using either pure-tones, banjo recordings, or rainfall sounds. All sonification styles used HRTF spatialization on the horizontal axis, logarithmic increases in frequency for the vertical axis, and linear increases in loudness for increasing proximity on the depth axis. The pure-tones and banjo sounds were presented with a slight left-to-right timing shift over 1 s, similar to sonifications by the vOICe. After trying all sonification styles, they rated each mode out of 10 on five usability metrics – whether the sounds were clear, engaging, relaxing, distracting, or conveyed tangible external objects. A two-way repeated measures ANOVA revealed that there was a significant interaction effect between the different sonification-modes and their usability-metric rating, $F(8,40)=2.34$, $p=0.036$, $\eta_p{}^2=0.318$. Here we found that pure-tones had the highest overall rating for clarity (6.42), while banjos were the most engaging (6.57) and the best at conveying external objects (6.85), and finally, rainfall was the most relaxing (7.21) and least distracting (4.07) for users.

In further explorations with the pure-tone style of image-sonification (Resolution: 7×7), we asked users

to indicate the location of a free-floating square object (14 cm^2) in 3D space across 10 trials, this was done by asking users to indicate the X and Z location of objects by positioning their white cane tip under the object on the floor, and measuring the Y axis through hand positioning. We found that across the last 5 trials, fully blind users had an average Euclidian distance error of 5.9 cm to the centre of the object, while visually-impaired (and blindfolded) had an average error of 8.07 cm. After using the SoundSight, users completed the AttrakDiff questionnaire, where users rated their overall impressions of the device on a 7-point Likert scale between opposing adjective-pairs (e.g. complicated-simple, isolating-connective, cautious-bold). This helps identify how users view this particular image-sonification style (3D space into pure-tones) in terms of pragmatic and hedonic values on a -3 to $+3$ scale, here we find that novice users are largely neutral in pragmatic terms (0.102), however, the App is seen as attractive in terms of its overall quality (1.16), and in terms of its hedonic properties, it is seen as interesting (0.9) and that users personally-identify with using the App (1.43). Finally, after having experience with the SoundSight, participants were asked to rate their interest in the topic of sensory substitution as an assistive technology (out of 5), here fully blind individuals indicated a higher level of interest (4.31) than the visually-impaired (3.5).

We note several observations during use, particularly by our fully blind participants. When evaluating the height of an object, participants adopted a wide range of strategies including: Moving their hand through the image until it matched the object's pitch; identify the 'middle' pitch and then slowly lower the sensor until the object hit the same 'middle' pitch; or use the 'edge' of the sensor's field of view, to determine when the object entered the image, and use this to guide hand movements. Some of these approaches appear to arise from difficulties with sonifying a whole field of view as well as having the height of objects determined from their position in the image, rather than their physical height relative to the floor. These observations indicate that novice users may benefit from either only sonifying central regions (e.g. 'single-point') or having these regions be conveyed in a qualitatively different manner so as not to require pitch-comparisons. Furthermore, having pitch reflect the height in the sensors' field-of-view rather than the object's physical height created some initial difficulties. Audio communicating an object's physical height could be produced in future iterations by utilising information either from gyroscopic tilt, or whole room scanning.

Users also gave their initial impressions for exploring the room with the 3D mode (tones, banjo, rainfall) and exploring images and clothing with the blended colour mode (using sound-colour combinations from [23, 35]. For the 3D mode, fully blind participants DH and DB wanted to continue testing the App in daily life and were excited about future directions, while JM liked the idea but felt the 3D mode was overwhelming, complicated, and did not see practical applications for herself. The colour mode was positively received by JM ("Love it, I want a copy of it now."), DH liked immediately and enjoyed the variety of sounds, while DB wanted more shades of colour to avoid confusion ("red in pink is confusing"). Participants with residual sight wanted the 3D mode to prioritise sonifying steps, with the colour mode not enhancing their functionality beyond existing apps. Blind users envisioned the 3D mode as being suitable for indoor navigation, following others, mapping new / unexpected routes, checking seat availability, while the colour mode was seen as suitable for exploring images on social media as well as identifying and cleaning clothes. Beta-testing to further refine features is due to open up in early 2021. Additional details on the study, individual results, and further impressions can be found in our supplemental materials.

4 Use case scenarios

The vignettes below are based on actual scientific projects, feedback from potential blind users, and discussions with other researchers.

4.1 Sensory substitution and augmentation

The overall aim is for the SoundSight to function as a wide-ranging tool to assist with locating objects in the environment, navigation, and interacting with the sighted world—as described elsewhere more generally for SSDs. Below we consider some more specific scenarios.

One possible scenario for blind users is to access information in images ranging from photographs to scientific graphs and illustrations. These could be taken by the phone's camera or downloaded from the internet or social media. Using the SoundSight, images can either be scanned through in a similar way to the live-viewing mode, but the user also has the option to scan through static images manually using a finger, such that only the colours directly underneath their finger are sonified. Since these sounds are also spatialized, it is possible to paint the image over time in their mind's eye.

A scenario for sensory augmentation, is the conversion of temperature (from a plug-in thermal camera) into sound in order to detect humans or guide dogs. A user can control what temperatures are turned into sound and apply a suitable threshold so that only warm objects (such as other people and service animals) are turned into sound. Using a plug-in thermal camera for their iPhone and the SoundSight App, this form of sensory augmentation allows people to be located at a distance and affords interesting learning

opportunities about how objects at a distance appear smaller in the visual world.

A final use case scenario concerns visual field defects such as tunnel vision. In this scenario, the SoundSight could extend the users' field-of-view by sonifying peripheral objects. The existence of some central vision would enable the user to learn the visual-sound conversion rules and extend them into regions of space that cannot be seen.

4.2 Interactive art

Imagine a scenario in which an artist wants to convey sounds that are associated with different visual elements of a painting, such that viewers can point to different parts of the painting and experience a dynamic and interactive audio feedback. This can be achieved by placing sound files of different qualities in the appropriate part of the sound array that corresponds to different elements of the painting (rather than the more standard SoundSight approach of creating an array of spatialised sounds based on a HRTF). A depth-sensor mounted near the painting detects approaching objects and triggers the appropriate real-time sonification.

A variation of this approach would be to create real-time sonifications of the human body itself. For instance, by having 3D space control which sounds are playing at a given moment, the artist can use their phone to capture the body in 3D space, so that any given pose will produce a different collection of sounds. Or a camera mounted to a mirror could sonify people in front of it according to the colour of the clothes they are wearing. The resultant interactive composition would reflect a combination of their clothing, movements, and body shape. Augmenting this with a thermal camera would be an alternative way of segmenting people from their surrounding objects and background.

4.3 Research on human perception

The SoundSight provides a platform that enables researchers to tackle a number of problems in scientific research more effectively as illustrated by the examples below.

Researchers who are interested in spatial hearing could use the SoundSight to set up their experiment. Instead of setting up an entire array of speakers to be controlled with custom-built software, the researcher can use one speaker to pre-record the sounds from a variety of locations. After uploading them into the SoundSight, it is possible to control which of thousands of potential 'speakers' are played at a given moment using (abstract) visual images as the 'trigger' for these sounds. This simplifies the process beyond integrating binaural encoders into experiments as no programming experience is required and the 'triggering' images provide easy to understand visual feedback of when any file is played.

The rapid updating of sounds following movement of the sensor, or objects in front of it, make the SoundSight suitable as a research tool for studying sensory-motor interactions. Researchers could upload various auditory representations of 3D space to evaluate how specific changes in audition influences user perception while still allowing subjects to freely explore environmental stimuli using head movements. For instance, users could upload a matrix of spatial recordings of frontal space as well as other matrices with changes to reverberation or cues like timing, level, or spectral changes. Now subjects wearing a head-mounted iPhone can explore these different matrices of sound files with audio files triggered by visual stimuli. This method aligns the iPhone and audio files with the subject's head position, allowing them to actively explore the sound signal with head movements to help them extract cues regarding the stimulus' position.

A variation of this, would be to sonify the subject's own arm/hands as it enters the SoundSight's field-of-view. In this way, one can explore the relationship between exteroceptive signals of limb position (spatialized audio via the SoundSight) and internal signals from the joints (proprioception). As the subject's arm passes through the 3D space captured by the iPhone, the sounds are played back to the subject in real-time. Various types of auditory feedback can be implemented (e.g. lagged feedback, spatial misalignments), allowing the researcher to evaluate how variations in auditory feedback influence the subject's perception of their own arm position.

5 Conclusions

After interviews with visually impaired end users about SSD technologies, several features were specified in making their adoption more likely. Here we showcase the outcome of this process with the SoundSight. The SoundSight takes a unique approach in sensory substitution because it delivers on the aspects requested by end users (e.g. responsivity to movement, conveying 3D space, smartphone-based) but then leaves the exact style of sonification open-ended to the wider community (of both end users and researchers). Any type of recorded audio can be replayed in this dynamic fashion. Similarly, the complexity faced by the end user can also be scaled, for instance, by only sonifying certain colours, temperatures, distances, or even just one pixel at a time, this provides users with the steppingstones from simple to advanced sensory substitution. Our preliminary SoundSight testing with visually-impaired users covered a wider range of factors than is typical for SSD research. The results, feedback, and observations revealed that the device was viewed positively in terms of aesthetic appeal, personal

identity, with fully-blind users enjoying the exploration of sonified colours, accurately localising free-floating objects in 3D space, and expressing enthusiasm regarding sensory substitution. However, users were overall neutral regarding the App's perceived practicality, which might be partially explained by difficulties using the camera's field-of-view and deciphering complex soundscapes to isolate the desired information. Moving forward, further development and testing (both inside the lab, and in beta-testing) will explore how the initially perceived practicality can be improved while maintaining the other positive characteristics of the App. To address this, future development is focusing on using computer-vision techniques to further simplify visual content and its resulting sonification, as well as label visual objects to aid the users' immediate comprehension of the environment and assist the learning process. In terms of being a scientific tool, the SoundSight currently allows robust testing of the effectiveness of different forms of sonification, since the same information can be represented in a multitude of different ways [35]. The SoundSight allows SSD researchers to push beyond identifying problems for end users for specific SSD designs (which are normally hard-coded by the designers) and implement alternative sonifications that address the problems inherent in prior SSD designs.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12193-021-00376-w.

Acknowledgements This work was developed from prior projects funded by the RM Phillips Foundation. In addition, we would like to thank the visually impaired participants who took part in the preliminary end user testing.

References

1. Kristjansson A, Moldoveanu A, Johannesson OI, Balan O, Spagnol S, Valgeirsdottir VV et al (2016) Designing sensory-substitution devices: principles, pitfalls and potential. Restor Neurol Neurosci 34(5):769–787
2. Nau AC, Murphy MC, Chan KC (2015) Use of sensory substitution devices as a model system for investigating cross-modal neuroplasticity in humans. Neural Regen Res 10(11):1717–1719
3. Proulx MJ, Ptito M, Amedi A (2014) Multisensory integration, sensory substitution and visual rehabilitation. Neurosci Biobehav Rev 41:1–2
4. Wright T, Ward J (2018) Sensory substitution devices as advanced sensory tools. Sensory substitution and augmentation. Oxford University Press, Oxford
5. Bach-y-Rita P, Collins CC, Saunders FA, White B, Scadden L (1969) Vision substitution by tactile image projection. Trans Pac Coast Otoophthalmol Soc Annu Meet 50:83–91
6. Bach-y-Rita P (2004) Tactile sensory substitution studies. Ann N Y Acad Sci 1013:83–91
7. Grant P, Spencer L, Arnoldussen A, Hogle R, Nau A, Szlyk J et al (2016) The functional performance of the BrainPort V100 device in persons who are profoundly blind. J Visual Impair Blind 110:77–88
8. Vincent M, Tang H, Khoo W, Zhu Z, Ro T (2016) Shape discrimination using the tongue: Implications for a visual-to-tactile sensory substitution device. Multisens Res 29(8):773–798
9. Richardson ML, Lloyd-Esenkaya T, Petrini K, Proulx MJ, editors. Reading with the Tongue: Individual Differences Affect the Perception of Ambiguous Stimuli with the BrainPort. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020
10. Jacobson H (1951) Information and the human ear. J Acoust Soc Am 23(4):463–471
11. Kokjer KJ (1987) The information capacity of the human fingertip. IEEE Trans Syst Man Cybern 17(1):100–102
12. Jicol C, Lloyd-Esenkaya T, Proulx MJ, Lange-Smith S, Scheller M, O'Neill E et al (2020) Efficiency of sensory substitution devices alone and in combination with self-motion for spatial navigation in sighted and visually impaired. Front Psychol 11:1443
13. Meijer PB (1992) An experimental system for auditory image representations. IEEE Trans Biomed Eng 39(2):112–121
14. Stiles NR, Shimojo S (2015) Auditory sensory substitution is intuitive and automatic with texture stimuli. Sci Rep 5:15628
15. Amedi A, Stern WM, Camprodon JA, Bermpohl F, Merabet L, Rotman S et al (2007) Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex. Nat Neurosci 10(6):687–689
16. Auvray M, Hanneton S, O'Regan JK (2007) Learning to perceive with a visuo-auditory substitution system: localisation and object recognition with "the vOICe." Perception 36(3):416–430
17. Murphy MC, Nau AC, Fisher C, Kim SG, Schuman JS, Chan KC (2016) Top-down influence on the visual cortex of the blind during sensory substitution. Neuroimage 125:932–940
18. Ward J, Meijer P (2010) Visual experiences in the blind induced by an auditory sensory substitution device. Conscious Cogn 19(1):492–500
19. Capelle C, Trullemans C, Arno P, Veraart C (1998) A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution. IEEE Trans Biomed Eng 45(10):1279–1293
20. Auvray M, Hanneton S, Lenay C, O'Regan K (2005) There is something out there: distal attribution in sensory substitution, twenty years later. J Integr Neurosci 4(4):505–521
21. Cronly-Dillon J, Persaud K, Gregory RP (1999) The perception of visual images encoded in musical form: a study in cross-modality information transfer. Proc Biol Sci 266(1436):2427–2433
22. 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. Restor Neurol Neurosci 32(2):247–257
23. Hamilton-Fletcher G, Obrist M, Watten P, Mengucci M, Ward J. "I Always Wanted to See the Night Sky" Blind User Preferences for Sensory Substitution Devices. Proceedings of the 2016 CHI

Conference on Human Factors in Computing Systems: ACM; 2016. p. 2162–74
24. Maidenbaum S, Abboud S, Amedi A (2014) Sensory substitution: closing the gap between basic research and widespread practical visual rehabilitation. Neurosci Biobehav Rev 41:3–15
25. Elli GV, Benetti S, Collignon O (2014) Is there a future for sensory substitution outside academic laboratories? Multisens Res 27(5–6):271–291
26. Chebat DR, Harrar V, Kupers R, Maidenbaum S, Amedi A, Ptito M. Sensory substitution and the neural correlates of navigation in blindness. Mobility of Visually Impaired People. Springer, Cham.2018. p. 167–200
27. Renier L, De Volder AG (2010) Vision substitution and depth perception: early blind subjects experience visual perspective through their ears. Disabil Rehabil Assist Technol 5(3):175–183
28. Haigh A, Brown DJ, Meijer P, Proulx MJ (2013) How well do you see what you hear? The acuity of visual-to-auditory sensory substitution. Front Psychol 4:330
29. van Rheede JJ, Wilson IR, Qian RI, Downes SM, Kennard C, Hicks SL (2015) Improving mobility performance in low vision with a distance-based representation of the visual scene. Invest Ophthalmol Vis Sci 56(8):4802–4809
30. Brown DJ, Simpson AJ, Proulx MJ (2015) Auditory scene analysis and sonified visual images. Does consonance negatively impact on object formation when using complex sonified stimuli? Front Psychol 6:1522
31. Brown DJ, Proulx MJ (2016) Audio–vision substitution for blind individuals: addressing human information processing capacity limitations. IEEE J Sel Topics Signal Process 10(5):924–931
32. Brown DJ, Simpson AJ, Proulx MJ (2014) Visual objects in the auditory system in sensory substitution: how much information do we need? Multisens Res 27(5–6):337–357
33. Levy-Tzedek S, Riemer D, Amedi A (2014) Color improves "visual" acuity via sound. Front Neurosci 8:358
34. Hamilton-Fletcher G, Ward J (2013) Representing colour through hearing and touch in sensory substitution devices. Multisens Res 26(6):503–532
35. Hamilton-Fletcher G, Wright TD, Ward J (2016) Cross-modal correspondences enhance performance on a colour-to-sound sensory substitution device. Multisens Res 29(4–5):337–363
36. Bertram C, Stafford T (2016) Improving training for sensory augmentation using the science of expertise. Neurosci Biobehav Rev 68:234–244
37. Moore BC (1973) Frequency difference limens for short-duration tones. J Acoust Soc Am 54(3):610–619
38. Mills AW (1972) Auditory Localization. In: Tobias JV (ed) Foundations of modern auditory theory, vol 2. Academic Press, New York, pp 303–348
39. Blauert J (1997) Spatial hearing: the psychophysics of human sound localization. MIT press
40. Parise CV, Knorre K, Ernst MO (2014) Natural auditory scene statistics shapes human spatial hearing. Proc Natl Acad Sci U S A 111(16):6104–6108
41. Perrott DR, Saberi K (1990) Minimum audible angle thresholds for sources varying in both elevation and azimuth. J Acoust Soc Am 87(4):1728–1731
42. Micheyl C, Delhommeau K, Perrot X, Oxenham AJ (2006) Influence of musical and psychoacoustical training on pitch discrimination. Hear Res 219(1–2):36–47
43. Sinnott JM, Aslin RN (1985) Frequency and intensity discrimination in human infants and adults. J Acoust Soc Am 78(6):1986–1992
44. Battal C, Occelli V, Bertonati G, Falagiarda F, Collignon O (2020) General enhancement of spatial hearing in congenitally blind people. Psychol Sci 31(9):1129–1139
45. Paquier M, Côté N, Devillers F, Koehl V (2016) Interaction between auditory and visual perceptions on distance estimations in a virtual environment. Appl Acoust 105:186–199
46. Kolarik AJ, Moore BC, Zahorik P, Cirstea S, Pardhan S (2016) Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. Atten Percept Psychophys 78(2):373–395
47. Khenak N, Vezien J, Bourdot P (2020) Spatial presence, performance, and behavior between real, remote, and virtual immersive environments. IEEE Trans Vis Comput Graph 26(12):3467–3478
48. Etchemendy PE, Abregu E, Calcagno ER, Eguia MC, Vechiatti N, Iasi F et al (2017) Auditory environmental context affects visual distance perception. Sci Rep 7(1):7189
49. Zahorik P, Wightman FL (2001) Loudness constancy with varying sound source distance. Nat Neurosci 4(1):78–83
50. Peeters G, Giordano BL, Susini P, Misdariis N, McAdams S (2011) The Timbre Toolbox: extracting audio descriptors from musical signals. J Acoust Soc Am 130(5):2902–2916
51. Lemaitre G, Houix O, Misdariis N, Susini P (2010) Listener expertise and sound identification influence the categorization of environmental sounds. J Exp Psychol Appl 16(1):16–32
52. Hamilton-Fletcher G, Pisanski K, Reby D, Stefanczyk M, Ward J, Sorokowska A (2018) The role of visual experience in the emergence of cross-modal correspondences. Cognition 175:114–121
53. Hamilton-Fletcher G, Pieniak M, Stefanczyk M, Chan KC, Oleszkiewicz A (2020) Visual experience influences association between pitch and distance, but not pitch and height. J Vis 20(11):1316
54. Hamilton-Fletcher G, Witzel C, Reby D, Ward J (2017) Sound properties associated with equiluminant colours. Multisens Res 30(3–5):337–362
55. Shepard RN, Cooper LA (1992) Representation of colors in the blind, color-blind, and normally sighted. Psychol Sci 3(2):97–104
56. Gomez JD, Bologna G, Pun T (2014) See ColOr: an extended sensory substitution device for the visually impaired. J Assist Technol 8(2):77–94
57. Froese T, McGann M, Bigge W, Spiers A, Seth AK (2012) The enactive torch: a new tool for the science of perception. IEEE Trans Haptics 5(4):365–375
58. Maidenbaum S, Hanassy S, Abboud S, Buchs G, Chebat DR, Levy-Tzedek S et al (2014) The "EyeCane", a new electronic travel aid for the blind: Technology, behavior and swift learning. Restor Neurol Neurosci 32(6):813–824