Blind wayfinding with physically-based liquid sounds

Simone Spagnol[a,*], Rebekka Hoffmann[a,b], Marcelo Herrera Martínez[a], Runar Unnthorsson[a]

[a] *Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, Dunhagi 5, 107 Reykjavík, Iceland*
[b] *Faculty of Psychology, University of Iceland, Sæmundargata 10, 101 Reykjavík, Iceland*

Abstract

Translating visual representations of real environments into auditory feedback is one of the key challenges in the design of an electronic travel aid for visually impaired persons. Although the solutions currently available in the literature can lead to effective sensory substitution, high commitment to an extensive training program involving repetitive sonic patterns is typically required, undermining their use in everyday life. The current study explores a novel sensory substitution algorithm that extracts information from raw depth maps and continuously converts it into parameters of a naturally sounding, physically based liquid sound model describing a population of bubbles. The proposed approach is tested in a simplified wayfinding experiment with 14 blindfolded sighted participants and compared against the most popular sensory substitution algorithm available in the literature - the vOICe (Meijer, 1992) - following a short-time training program. The results indicate a superior performance of the proposed sensory substitution algo-

*Corresponding author
Email addresses: spagnols@hi.is (Simone Spagnol), rebekkah@hi.is (Rebekka Hoffmann), marcelo@hi.is (Marcelo Herrera Martínez), runson@hi.is (Runar Unnthorsson)

rithm in terms of navigation accuracy, intuitiveness and pleasantness of the delivered sounds compared to the vOICe algorithm, supporting its usability for the visually impaired community.

Keywords: sensory substitution, sonification, electronic travel aid, physical sound model

1. Introduction

The technique of data sonification is used as an alternative or a complement to data visualization for representing various actions, objects or signals. Sonification can be defined as *"a mapping of numerically represented relations in some domain under study to relations in an acoustic domain for the purposes of interpreting, understanding, or communicating relations in the domain under study"* [40]. Widely accepted sonification techniques include *audification* (i.e., direct playback of data streams as sound waves), *auditory icons* (i.e., discrete environmental sounds), *earcons* (i.e., discrete symbolic sounds), *parameter mapping sonification* between data dimensions and auditory dimensions, and *model-based sonification* (i.e., based on dynamic models of virtual sounding objects) [23, 17].

Sonification is used in very different contexts to represent a great variety of data, ranging from molecular information [19] to geophysical data [16]. Of particular interest are applications in health care, such as in motor rehabilitation systems [1, 39] where task-related auditory information is able to support motor learning and increases attention and engagement levels during rehabilitation tasks. Another widely explored area is that of electronic travel aids [15] and other assistive technologies for visually impaired persons

(VIPs) [14], where sonification techniques are designed to substitute visual information [28]. Unfortunately, the majority of the systems exploiting such techniques are still in their infancy and have limited functionalities, small scientific and/or technological value and high cost [15].

Available electronic travel aids for VIPs range from simple *obstacle detectors* with a single range-finding sensor (e.g. ultrasound, infrared), to *environmental imagers* employing data generated from visual representations acquired through camera technologies. The most common sonification schemes of obstacle detectors, which only receive range information, are either earcons indicating the presence of an obstacle, or an inversely proportional transform mapping one or more range readings to the loudness and/or pitch of synthetic sounds or musical tones [10]. On the other hand, environmental imagers (i.e., devices able to deliver a representation of the layout of an environment) allow for greater flexibility in sonification mappings. The most significant example is provided by the well-known image sonification algorithm used in the vOICe system [30].

The vOICe algorithm can be thought of as an inverse spectrogram transform, i.e., a time-varying sound whose spectrogram approximately matches an input grayscale image. In particular, the algorithm periodically scans the image from left to right, while associating each row to a different sinusoidal oscillator with fixed frequency (in ascending order from lower to upper rows) and using the brightness of each pixel in turn to control the amplitude of the oscillator. The sound output is then spatialized left to right according to the current scanning point. It has been shown that, following extensive periods of training and exploiting the neural plasticity of the human brain, the

vOICe sonification mechanism can lead to effective sensory substitution [31], both in object recognition [53] and spatial learning [35].

Although the original vOICe algorithm was designed to sonify 2D grayscale images, its use in blind wayfinding is supported by the observation that a depth map can be directly converted into a grayscale image where brightness corresponds to depth. The use of depth information for the sonification of 3D scenes through either the original vOICe algorithm or slight variations of it has already been proposed and investigated [11, 52]. Furthermore, improvements to the pleasantness of sounds (such as using musical tones instead of pure sines) as well as to the spatial feeling and real-time conveyance of the sounds (e.g. presenting independently to each headphone channel simultaneous scans from the left and right edge to the central column of the image) were proposed [3].

The main drawback of most existing sensory substitution devices (SSDs), including the vOICe, is that even though in some cases the conveyed auditory information can be successfully interpreted by naïve users, they demand extremely high commitment on the user's side. A lengthy and strenuous training of up to one year is required in order to enable users to perform most tasks, thus undermining the use of SSDs in everyday life [35]. As Fontana *et al.* point out [18], the prolonged use of SSDs *"leads to the strain of the user [...] due to the continuous listening of the same signal at regular time intervals. This sound, even if spatialized, produces an unnatural effect and causes a progressive fatigue."* Therefore, the choice of the type of sound as well as the way it is generated should be regarded as a key issue in the design of any sensory substitution algorithm.

The current study explores a novel model-based sonification algorithm for translating continuous representations of a dynamic real environment, coded into sequences of depth maps, into auditory feedback. The sensory substitution algorithm we propose is meant to be used for real-time blind wayfinding, with minimum latency between data acquisition and sonification, and with available off-the-shelf hardware technologies. It was designed in an attempt to improve the vOICe algorithm from both an ergonomic and a functional point of view, eventually reducing the required training time, and to be efficiently scalable depending on the available computational resources. The algorithm we propose here directly maps low-order statistics from the raw depth map into the parameters of a physically-based liquid sound model. In this model, physical descriptions of sound events are intentionally simplified to emphasize the most perceptually-relevant timbral features, and to reduce computational requirements as well [4]. The model was specially selected and tuned in order to sound both natural (yet significantly discernible from most daily environmental sounds) and aesthetically pleasant.

The remainder of the paper is organized as follows. In Section 2 we describe the generation mechanism of liquid sounds and its use in the design of our *fluid flow* sensory substitution algorithm. In Section 3 we introduce an experiment designed in order to assess the performance and individual preference of the sensory substitution algorithm in a blind wayfinding task. Results are reported in Section 4 and finally discussed in Section 5.

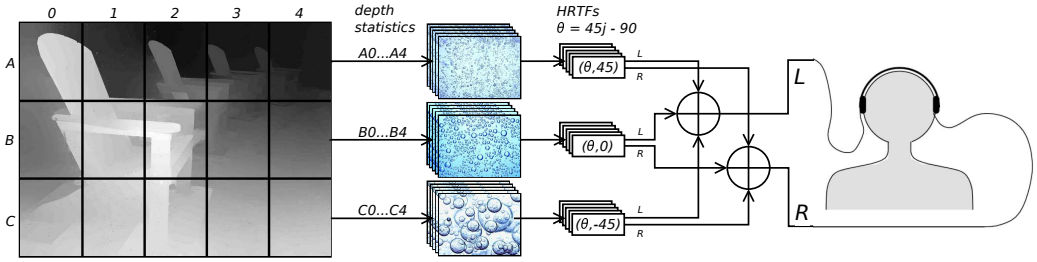


Figure 1: Simplified scheme of the proposed sensory substitution algorithm.

2. Sensory substitution with liquid sounds

The *fluid flow* sensory substitution algorithm that we propose in this paper receives a sequence of depth maps as input. Each depth map is divided into 15 equally sized sectors given by the combination of 3 rows and 5 columns. Every sector corresponds to an independent and uncorrelated instance of a liquid sound generator, and its position within the depth map is spatialized in the frontal hemisphere, allowing for effective source separation. Figure 1 reports a simplified scheme of the proposed algorithm.

2.1. Generation of liquid sounds

The building block of the fluid flow algorithm is the *liquid sound generator*. In the physical world, liquid sounds are mostly caused by gas bubbles trapped inside the liquid rather than by the liquid mass itself. For this reason, sound is generated through a stochastic process modeling the temporal evolution of a population of bubbles, a synthesis approach previously referred to as *physically informed sonic modeling by granular synthesis* [57]. The liquid sound generation algorithm considers individual bubbles to be atomic units (or *grains*, according to the granular synthesis terminology [37]), synthesized

using the well-known physically based Minnaert model [32]. Spherical bubbles effectively act as exponentially decaying sinusoidal oscillators: the compressible gas region of the bubble, surrounded by an incompressible liquid mass, gradually dissipates the energy involved in its creation by a periodic pulsation, as it would happen in a spring-mass system.

Every single bubble k, whose impulse response is

$$i_k(t) = a_k \sin(2\pi f_k^0 t) e^{\zeta_k t} \tag{1}$$

is fully defined by means of its radius r_k and depth factor D_k, that uniquely determine the individual damping factor ζ_k, resonant frequency f_k^0, and amplitude a_k as follows:

$$\zeta_k = \frac{0.13}{r_k} + 0.0072 r_k^{-\frac{3}{2}} \quad f_k^0 = \frac{3}{r_k} \quad a_k = D_k r_k^{\frac{3}{2}} \tag{2}$$

Here the depth factor D_k models the lumped effect of the depth of a bubble, and the effect of different excitation strengths of the bubbles. Bubbles that are submerged more will be attenuated more. Factor D_k is a dimensionless number between 0 and 1, where 1 corresponds to a bubble created at the surface and 0 to a fully submerged bubble.

The creation of bubbles is then modeled as a Bernoulli process occurring at audio rate with success probability $p = 1/\Lambda$, where Λ is the average bubble rate (bubbles per second). The radius of each successfully produced bubble k is set to

$$r_k = x_k^{\gamma_r} (r_{MAX} - r_{MIN}) + r_{MIN} \tag{3}$$

where $x_k \in [0, 1]$ is a number drawn from a uniform distribution function, r_{MIN} and r_{MAX} are the minimum and maximum bubble radius values, and

γ_r is the radius gamma factor, which allows to increase the ratio of bigger bubbles relative to smaller bubbles ($0 < \gamma_r < 1$) or *vice versa* ($\gamma_r > 1$). Similarly, the depth factor D_k is set to

$$D_k = y_k^{\gamma_D}(D_{MAX} - D_{MIN}) + D_{MIN} \tag{4}$$

where $y_k \in [0, 1]$ is a number drawn from a uniform distribution function, D_{MIN} and D_{MAX} are the minimum and maximum depth factor values, and γ_D is the depth gamma factor, which allows to increase the ratio of bubbles close to the surface relative to deeper bubbles ($0 < \gamma_D < 1$) or *vice versa* ($\gamma_D > 1$).

Bubble sounds often exhibit a characteristic rise in pitch, especially when approaching the surface. The phenomenon is mostly caused by the pressure reduction as the liquid mass above the bubble becomes thinner and thinner. The effect is modeled in the synthesis algorithm by a global rise factor parameter ξ. Since bubbles with a rising pitch are created close to the surface, it seems reasonable to assume they are generally louder than average. This effect is modeled by a rise cutoff parameter K_ξ. When it is set to a value $0 < K_\xi < 1$, only bubbles with a depth factor $D_k > K_\xi$ have a nonzero rise factor ξ. According to the physically based bubble sound model described in [57], a rising bubble is modeled by making its frequency time-dependent according to

$$f_k(t) = f_k^0(1 + \sigma_k t) \tag{5}$$

where σ_k is the slope of the frequency rise related to the vertical velocity of the bubble, modeled as

$$\sigma_k = \xi\zeta_k. \tag{6}$$

An implementation of the liquid sound generator described above (*fluid flow* module) is included in the Sound Design Toolkit (SDT),[1] an open-source (GPLv2) library of physically based sound synthesis algorithms for Max and Pure Data [4]. In this implementation the stochastic process drives an oscillator bank, whose number of voices can be set as a parameter. The size of the oscillator bank defines the polyphony of the algorithm, i.e. the maximum number of bubbles that can be active at the same time. If the maximum number is exceeded, a voice stealing mechanism takes place and the new bubble is assigned to the oscillator that currently has the minimum instantaneous amplitude envelope, resetting all its parameters, base frequency included. Phase alignment allows to avoid audible artifacts during the generation of a new bubble [46].

The liquid sound generator is a slightly improved version of the *bubble simulator* proposed by van den Doel [57]. The main improvement with respect to the van den Doel simulator lies in the use of a single Bernoulli process for a population of bubbles with different radii (i.e., with different base frequencies) rather than 50 Bernoulli processes each set to a fixed base frequency. This strategy allows to represent bubbles of arbitrary size, improving the versatility of the algorithm especially with small oscillator banks.

2.2. Model-based sonification

A global d_{MAX} parameter is defined in order to consider only those points in the depth map whose depth is no greater than this defined parameter. Then, for each sector, two descriptive depth metrics are calculated: *map*

[1] `http://soundobject.org/SDT/`

density and *average depth*. Design choices for mappings between depth map properties and liquid sound features are the following:

- map density → average bubble rate;
- average depth → maximum bubble depth factor.

Map density ρ is defined as the number of pixels with depth value no greater than d_{MAX} divided by the total number of pixels in that sector. It is mapped to the *average bubble rate* Λ according to

$$\Lambda = 500\rho^2 \tag{7}$$

so that the denser the sector, the more the generated bubbles. The upper limit of 500 bubbles/second was heuristically set following informal investigations on the pleasantness and intelligibility of the associated liquid sound.

Average depth $\bar{d}$ is defined as the mean depth value (in meters) of all pixels with depth no greater than d_{MAX} in that sector. It is mapped to the *maximum bubble depth factor* D_{MAX} as

$$D_{MAX} = \left(\frac{d_{MAX} - \bar{d}}{d_{MAX}}\right)^2. \tag{8}$$

In this way, closer obstacles are transformed in a larger amount of bubbles close to the surface of the water, thus increasing their average loudness and sharpness. As an analogy, it might help to think of the scene as a big aquarium seen from above, with the water surface just in front of the observer and all objects producing bubbles.

In order to provide a spatial dimension of the depth map, the sound produced by each liquid sound generator is binaurally spatialized by mapping

the corresponding depth map sector (R_i,C_j) to the azimuth and elevation parameters (θ,ϕ) of a generic HRTF filter as follows:

$$\theta = 45j - 90 \tag{9}$$

$$\phi = 45 - 45i \tag{10}$$

where θ and ϕ are expressed in degrees with respect to the observer according to a vertical polar coordinate system, $i = 0, 1, 2$ is the row number (top to bottom), and $j = 0, \ldots, 4$ is the column number (left to right). However, since elevation cues greatly differ from subject to subject [48] and lead to high variance in vertical localization performance with generic HRTFs [33], elevation information is redundantly coded into another liquid sound feature. In particular, sectors belonging to different rows of the depth map are assigned different bubble radius intervals $[r_{MIN}, r_{MAX}]$ as follows:

$$\begin{aligned} R_0 &: r_{MIN} = 0.2mm, r_{MAX} = 1mm; \\ R_1 &: r_{MIN} = 1mm, r_{MAX} = 5mm; \\ R_2 &: r_{MIN} = 5mm, r_{MAX} = 20mm. \end{aligned} \tag{11}$$

Thanks to the inversely proportional relation between bubble radius and resonant frequency (see Eq. 2), the above heuristically defined intervals allow for different characteristic liquid sounds to be produced depending on elevation, i.e., ranging from light, fizzy sounds for higher elevations (row R_0) to low, gurgling sounds for lower elevations (row R_2).

Other parameters that define the liquid sound generator are kept constant. These include the radius gamma factor ($\gamma_r = 1$), the minimum bubble depth ($D_{MIN} = 0$), the depth gamma factor ($\gamma_D = 1$), the rise factor

($\xi = 0.5$), and the rise cutoff ($K_\xi = 0.5$). Both gamma factors are set to 1 in order to preserve the uniform distribution of radius and depth values. On the other hand, the choices for the rise factor and rise cutoff allow for an additional auditory depth cue. By combining Eq. 8 and Eq. 4 it can be shown indeed that the average depth value at which pitch-rising bubbles start being produced ($D_k > K_\xi$) roughly corresponds to $\bar{d} \approx 0.3 d_{MAX}$. This translates at auditory level into a peculiar boiling water sound for close objects, and the closer the object (i.e., the lower the average depth value), the higher the number of pitch-rising bubbles and therefore the clearer the boiling effect.

A preliminary version of the fluid flow algorithm was previously presented by the authors in [46]. With respect to the previous version, the main improvements of the algorithm described here lie in the representation of elevation information with different bubble radius values, in using bubble depth as a proper physical depth indicator rather than plain amplitude control, and in the use of the rising pitch cue for close objects rather than elevated objects. These design changes were suggested from both test results and informal comments following preliminary experimental trials with offline video sequences [46], that highlighted above all the difficulty of interpreting elevation cues.

At the same time, the new mappings provide more meaningful correspondences between physical and auditory cues. As a matter of fact, beside the intuitive relationship between physical depth and bubble depth, crossmodal correspondences between pitch (resonant frequency in the bubble model) and elevation are well known in the literature [25] and frequently used in sensory substitution systems (including the vOICe). Furthermore, the boiling ef-

fect that gets more and more prominent while approaching an object can be interpreted as an effective natural warning sound [55].

3. Evaluation

The main goal of the experiment presented here is to assess the performance and individual preference of the *fluid flow* sensory substitution algorithm in a blind wayfinding task. More in detail, the point-by-point objectives are

1. to validate the effectiveness of the proposed sounds of giving reliable and distinguishable information in a simplified wayfinding task with a reasonably sized pool of naïve blindfolded participants;
2. to collect individual judgments about the naturalness, pleasantness and usability of the sounds that are conveyed;
3. to compare the above results and ratings against those collected using the reference sensory substitution scheme provided through the original vOICe algorithm [30].

Our working hypotheses are that: (1) after a short training session, the fluid flow algorithm is able to help participants avoid obstacles in the large majority of the presented cases; (2) performance and completion time are at least comparable to the vOICe algorithm; (3) the individual judgments on the liquid sounds reflect a positive opinion on all the investigated aspects and, in particular, a more positive rating compared to the sounds produced by the vOICe algorithm.

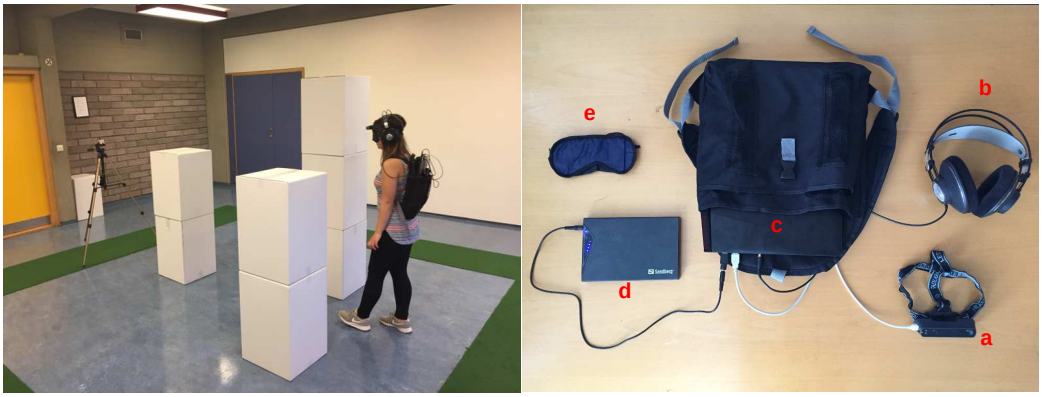
3.1. Sample

Fourteen participants (7F, 7M) participated on a voluntary basis. Ages ranged from 22 to 46 (M = 30.5, SD = 7.2). All participants spoke fluent English and none of them reported either visual or hearing impairments. All participants gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the National Bioethical Committee of Iceland (reference number VSN-15-107).

3.2. Experimental setup

The experiment took place in an empty classroom sized $8m$ (length) $\times$ $6.7m$ (width) $\times 3.5m$ (height) inside a building of the University of Iceland. Four pieces of green carpet, sized $4m \times 0.5m$ each, were placed in the middle of the classroom floor in order to delimit a square $3.5m \times 3.5m$ testing area (see Figure 2a). During the whole experiment, to control for confounding effects, windows were kept closed and artificial light was turned on. The absence of any kind of activity in the neighboring classrooms due to summer break guaranteed a quiet environment throughout the testing sessions. The ventilation system of the classroom produced the only significant, yet constant, environmental sound.

During the tests, white cardboard boxes were placed in predefined locations of the testing area. The size of a single cardboard box was $0.4m$ (length) $\times$ $0.4m$ (width) $\times 0.6m$ (height). The number of boxes inside the testing area during each experimental trial ranged from 5 to 8; when less than 8, the unused boxes were placed along one wall as shown in Figure 2a. Furthermore, a tripod holding a small Bluetooth box speaker (at approximately $1.2m$ height)



(a) (b)

Figure 2: Experimental setup. (**a**) Subject during the experiment. (**b**) Close up of the equipment.

was placed along the end-side of the testing area. The only other significant objects present in the room were a desk and two chairs for the experimenters, all positioned behind the starting point of the participants.

Participants wore the following equipment, pictured in Figure 2b: (a) an elastic headband (originally holding a searchlight) with a Structure Sensor camera[2], a high-performance structured light 3D sensor, tightened to the frontal plastic hold; (b) a pair of open over-ear headphones (AKG K612 Pro) allowing environmental sound to enter the ear; (c) a small backpack carrying a Lenovo Ideapad Y700 laptop running the software to which the camera, headphones and (d) an external battery were connected; (e) a blindfold. In order to ensure regular functioning, the laptop was constantly monitored by an experimenter through a second laptop placed on the desk behind the

[2] `https://structure.io/`

testing area, connected via VPN.

Depth maps with a resolution of 640×480 pixels were acquired from the Structure Sensor at a rate of 10 frames per second with the support of an open-source Matlab Wrapper for OpenNI 2.2,[3] processed in Matlab, and sonified through the Pure Data software implementing the fluid flow and vOICe algorithms. Depth maps spanned the entire field of view of the Structure Sensor, i.e., 58° horizontal, 45° vertical, and a $0.4m$ to $3m$ depth range. Visual information falling beyond these ranges was therefore not sonified.

3.3. Stimuli

The sound stimulus conveyed to participants during the experiment was a continuous sonification of the depth data acquired through the Structure Sensor, either through the fluid flow algorithm, referred to as FF and described in Section 2, or the vOICe algorithm, referred to as VC and described in the following paragraph. Each algorithm was implemented as a Pure Data patch that constantly receives the depth map statistics data through the OSC (Open Sound Control) protocol. In order to avoid audible artifacts, the incoming depth map statistics values were smoothed with a 100-ms ramp function. In the experiment, the d_{MAX} parameter was set to $3m$ and the number of voices of each liquid sound generator to 32. For the sake of consistency, the level of the sound card was kept constant throughout the experiment for all participants.

[3] `http://uk.mathworks.com/matlabcentral/fileexchange/42127-matlab-wrapper-for-openni-2-2`

The vOICe sensory substitution algorithm was implemented following the specifications from Meijer [30]. The algorithm scans each depth snapshot (resized to 64×64 pixels) from left to right, while associating height (i.e. the vertical coordinate of the pixel) with pitch and depth with loudness. More specifically, every row is associated to an amplitude-controlled oscillator whose fixed frequency exponentially ranges from 500 Hz (bottom row) to 5 kHz (top row), while amplitude is inversely proportionally related to the depth value, ranging from 0 for pixels of unknown depth value or where depth is greater than or equal to d_{MAX}, to 1 for pixels of zero depth. The auditory output of the implemented algorithm was compared against the original vOICe software for Windows on a small benchmark set of 10 depth maps from the NYU-Depth Dataset V2[4] [44], and it was found to never exceed 1 dB of spectral distortion in the $0.5 - 5$ kHz range.

The generic HRTF filter that we used is provided through the *earplug~* Pure Data binaural synthesis external. The filter renders the angular position of the sound source relative to the subject by convolving the incoming signal with left and right HRTFs from the MIT KEMAR database[5] [20]. For the sake of consistency, the same HRTF filters were used for both FF and VC.

3.4. Experimental procedure

The experiment was divided in two sessions, each corresponding to a single sensory substitution algorithm (FF or VC). The two sessions were conducted on different days and the order of the sensory substitution algorithms

[4] `http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html`

[5] `http://sound.media.mit.edu/resources/KEMAR.html`

was randomized and balanced. A single experimental session was composed of three parts presented in the following order: a self-training part, a guided training part, and an experimental test. The purpose of the training was to allow for sufficient interaction with the system and to gain experience with the sonification algorithm prior to the experimental test, where the actual performance data was collected. The duration of the self- and guided training was approximately 10 and 65 minutes, respectively, while the average duration of the experimental test was approximately 40 minutes.

3.4.1. Self-training

Basic information about the sensory substitution algorithm was first provided to participants through a short written description (7 lines) on an experimental sheet, transcribed in the Appendix. Then, participants wore the pair of headphones and freely interacted via keyboard with a simplified demo of the system representing a single virtual object in the field of view of the camera. Participants controlled the azimuth, elevation, distance, and size of the object (see key assignment below), and directly listened to the corresponding sonification:

- numpads $1-9$: change the direction of the object on a 3×3 grid: 3 azimuths (left, center, right) and 3 elevations (up, middle, down);
- arrow keys up/down: increase/decrease the distance of the object between $0.5m$ and $3m$, in $0.5m$ steps;
- keys $+/-$: increase/decrease the size of the object (in terms of % of the occupied area in that sector) from 0% to 100%, in 10% steps.

The self-training was designed to introduce participants to the sensory substitution algorithm and the underlying mappings.

3.4.2. Guided training

Participants were equipped with the system (backpack/PC, camera headband, blindfold, headphones) and then guided through five consecutive training steps as follows.

Step A (3 minutes). Participants listened interactively to the sonification of an empty testing area while being allowed to freely explore the empty room (only being stopped when going too close to an obstacle, e.g. the desk or a wall). Additionally to the floor, at this stage, it was important for participants to listen to and recognize the sonification of walls, ceiling and other fixed objects in the room.

Step B (7 minutes). One object (made of two or three boxes on top of each other in turn) was placed in the middle of the testing area and participants were asked to interact with it. Participants were encouraged (guided if necessary) to systematically explore the sonification output in relation to changing their own position, e.g. to (1) go towards/away from the object while facing it, therefore experiencing distance changes, while getting verbal feedback on the current distance; (2) circle the object and stand aside of it while trying to locate it with only head movements; (3) stand $2m$ away, face the object and tilt the head up/down in order to experience elevation changes. At this stage it was important to let participants realize through training that objects closer than $0.4m$ or further than $3m$ were not represented; therefore, participants were invited to explore and experience at what distance the sonification of the object stopped.

Step C (15 minutes). Participants trained scenes with *a single* object (made of two or three boxes on top of each other) positioned in randomly chosen locations of the testing area within the represented distance range. Pink noise was played on the headphones in order to mask the sound of boxes being moved when preparing the next scene. The participants' task was to first point at the object after head movement only, tell its approximate distance (in meters) and size (2 or 3 boxes), and then to go towards it and touch it. From this step onwards, after successful completion of each scene, participants were invited to temporarily remove the blindfold in order to check the scene they just accomplished.

Step D (20 minutes). Participants trained scenes with *two* objects (each made of two or three boxes on top of each other) positioned in randomly chosen locations of the testing area within the represented distance range, provided that they were positioned no less than $0.8m$ apart from each other in order to be able to comfortably pass between them. The participants' first task was to point at each object in turn after head movement only and tell again their approximate distance and size. After successful completion of the first task, participants were asked to walk between and past the two objects trying not to touch or collide with them.

Step E (20 minutes). Participants trained a number of scenes with *two or three* objects (randomized), aiming to find their way towards the small speaker placed at a randomly chosen point on the opposite side of the testing area and playing easy-listening pop music [54] at a comfortable level. The obstacles (again 2 or 3 boxes on top of each other) were placed randomly within the testing area, provided that they were positioned no less than

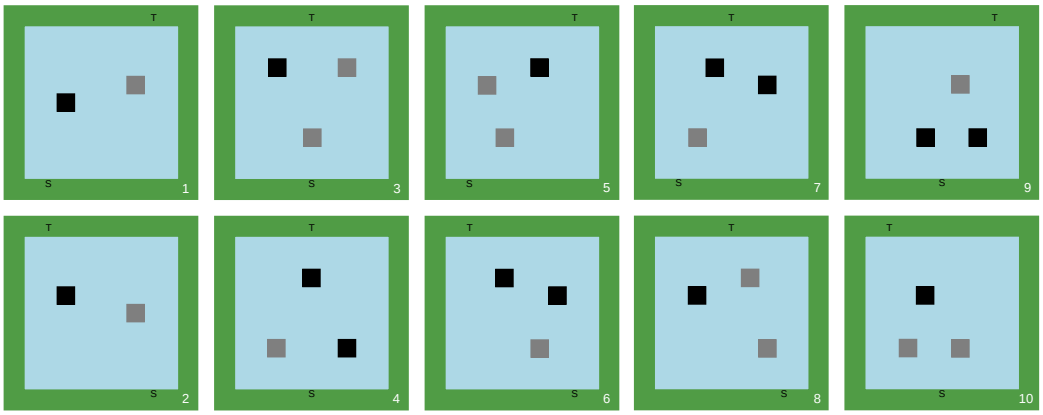


Figure 3: The 10 testing scenes. The 2-box obstacles are depicted as gray squares, and the 3-box obstacles as black squares. The starting and target (end) points are marked with S and T, respectively.

$0.8m$ apart from each other (to all sides). Participants were asked to walk as carefully as possible trying not to touch or collide with the obstacles, to stay inside the testing area all the time, and to scan the environment before moving forward. At this stage it was important to tell participants that the tripod would be represented through sound as well, that they should walk towards the target without detour (especially when starting on the edges of the testing area), and that if close to the target, they should try to touch the target promptly.

In order to reduce fatigue, a mandatory 10-minute break was introduced between Step D and Step E. Participants were invited to take off the system and relax.

3.4.3. Experimental test

Right after the training, the blindfolded participants tested 10 wayfinding scenes with two or three objects always positioned within the path towards

the target, with a task similar to training step E. However, this time the obstacles (2 or 3 boxes on top of each other each) were not placed randomly within the testing area but in predefined locations, as well as the starting and target (end) points, as shown in Figure 3. The order of the 10 scenes was randomized for each participant and each session. Participants were reminded to walk as carefully as possible, to scan the environment before moving forward, and to walk towards the target without detour. Participants were informed that their goal was to reach the target speaker trying to avoid any collision with obstacles and without leaving the testing area, and that all errors would be counted. For each experimental testing, collected data included:

- number of collisions with obstacles, while differentiating between minor collisions (i.e., not moving boxes from their position, for instance brushing on them) and major collisions (i.e., boxes moved);
- number of times the participant left the testing area by treading, even partially, on the carpet (except when in the target's vicinity);
- completion time (in seconds, taken with a timer), defined as the time between the moment when the sonification was turned on and the moment when the participant touched the speaker or tripod.

After completion of all experimental testing scenes, participants were asked to reply to a questionnaire about the corresponding sensory substitution algorithm by ticking one item in each of three 7-point Likert scales (1 = strongly disagree, 7 = strongly agree):

1. I feel I could directly understand the meaning of the sounds without training;
2. I feel that the sounds are pleasant;
3. I would feel comfortable hearing these sounds on a daily basis.

3.5. Statistical analysis

After an exploratory data analysis on all categories of navigation errors, a more advanced analysis was performed. Due to the dependent, nested structure of the data, and to factor in covariates, linear mixed models with fixed and random effects [36] were fit in R version 3.4.1 (R Development Core Team 2017). The within-subjects design of the current study allowed to statistically control for the differences across participants in every analysis by taking individual variance as random effect into account, which might otherwise distort the results. Additionally, training effects might influence the outcome, meaning that participants accomplished more scenes without navigation errors when they went through the training and testing procedure for the second time compared to the first time, independent of the sensory substitution algorithm. By randomizing the sequence of the two algorithms, any systematical influence due to training effects was experimentally controlled for. Yet, the training effect might lead to substantial additional variance in the data, which is why it was statistically controlled for by being factored in as random effect into all analyses.

3.5.1. Analysis of performance data

In order to compare the performance between the two sensory substitution algorithms, the probability of passing a scene (meaning the participants

did neither collide with any obstacle nor leave the testing area) for each of the two algorithms was calculated, set as outcome variable and fit in a Generalized Linear Mixed Model (GLMM). Due to the categorical nature of the outcome variable, a mixed-effects binomial logistic regression model was performed [22, 24] by executing the the `glmer()` function as part of the `lme4` package in R [5]. For parameter estimation in the GLMM, in order to approximate true likelihood, the Laplace approximation method with an adaptive algorithm using one integration point was performed [7].

A model selection process was the first step of the performance analysis, in which the improvement of model fits for three different models was compared. Firstly, *Model 0* (a baseline model not containing any fixed predictor but only the random effects of individuals and training) was compared to *Model 1* (with algorithm added as one fixed predictor) in order to determine if taking in algorithm as predictor into the model significantly improves the variance explained by the model. If so, algorithm would have a significant effect on the probability of passing a scene. Secondly, *Model 1* was compared to *Model 2* (with time that was necessary for scene completion added as second fixed predictor, besides algorithm) in order to determine if adding time as predictor significantly improves the variance explained. If so, time would have a significant effect on the probability of passing a scene. A Chi-square distributed Likelihood Ratio Test was performed to determine if the difference between models was significant and therefore select the best model. Finally, the model with the best fit was reported with regression coefficients, effect direction, confidence intervals and the predictors significance was ascertained with the Wald statistics [58].

3.5.2. Analysis of time data

In the performance analysis described above, the time that participants needed to complete a scene was only indirectly taken into account as possible predictor for passing as scene. However, we were mainly interested in answering the question if the choice of sensory substitution algorithm results in significantly different times (while statistically controlling for training and individual effects). To address this, a subset of data only including passed scenes was created and analyzed with time as continuous outcome variable. This approach was chosen since the occurrence of navigation errors hint at the possibility that scenes were not represented understandably and participants were not able to interpret the obstacle location, which questions the sense of interpreting failed scenes.

A Linear Mixed Model with algorithm as fixed effect and individual differences and training as random effects was fit using Restricted Maximum Likelihood (REML) [36]. We performed the `lmer()` function as part of the `lme4` package to fit the LMM in R [5], as well as the `lmerTest` package[6] to test if the predictor of the proposed model was significant. The package provides F-test statistics by calculating the degrees of freedom with the Satterthwatie approximation method [41].

3.5.3. Analysis of questionnaire data

We finally investigated for differences in individual questionnaire scores between the two algorithms by running three separate Wilcoxon signed-rank tests, one per questionnaire item (intuitiveness, pleasantness and usability,

[6] `https://CRAN.R-project.org/package=lmerTest`

respectively). The choice of the Wilcoxon signed-rank test was due to the within-participants design and to the non-normal distribution of the questionnaire data. Before applying each test, we verified the assumption that the distribution of the differences between the two related groups was symmetrical in shape by checking that its skew value was between -2 and 2 [27].

4. Results

The complete individual results from the experiment are reported in Table 1. In the table, variables C_{MIN} (number of minor collisions), C_{MAJ} (number of major collisions), N_{OUT} (number of times the participant left the testing area), and T_{TOT} (completion time) are aggregated for the 10 scenes. It can be noticed that a lower average number in all types of navigation errors was registered for FF compared to VC.

4.1. Performance

First, we compared the performance between the two algorithms, FF and VC. To assess whether a scene was reliably and understandably represented by the algorithm, the number of passed scenes was counted. The results show that when using FF, 107 (out of 140) scenes were successfully completed by participants (therefore fulfilling our hypothesis no.1), compared to 77 (out of 140) when the same participants used VC.

In order to assess whether the higher proportion of passed scenes with FF was statistically significant (on alpha level of .05), the influence of the algorithm on the probability of passing a scene was determined as described in Section 3.5.1 following a model selection process. The results for Model 1 and Model 2 are reported in Table 2 with regression coefficients, standard

errors, confidence intervals and Wald statistics per predictor. Whereas all models included individual variance and training as random effects, the basic model (Model 0) did not contain any fixed predictors, which is why it is not presented in the table, but served as baseline model for comparison to Model 1.

According to the Likelihood Ratio Test (LRT), including the predictor of sensory substitution algorithm (Model 1) significantly improved the model fit compared to an empty model without predictors (Model 0), $\chi^2(1, N = 280) = 20.15$, $p < .001$. This result indicates that the choice of algorithm, FF or VC, has a significant effect on the outcome variable of performance, meaning that the probability that participants performed a scene without errors was significantly higher when they followed FF compared to VC.

In Model 2, time was included as additional fixed predictor to test if it had a significant influence on the performance. We expected that a short completion time, even though at first glance seemingly positive, might indicate that participants rushed through the scenes since they were lacking understanding of the scene resulting in collisions. However, including time as predictor (additionally to algorithm) does not significantly improve the model according to the LRT, $\chi^2(1, N = 280) = 2.50$, $p = .105$, meaning that the completion time is not a predictor for more passed scenes.

To summarize, Model 1, only including algorithm as fixed effect while factoring individual variance and training as random effects, explains most of variance in the data. Adding time as predictor does not improve the model fit. The Wald statistics for each fixed predictor of Model 1, reported in Table 2, confirm the significant effect of the sensory substitution algorithm (improving

our expectations as stated in hypothesis no.2) and the non-significant effect of time on the probability of passing a scene.

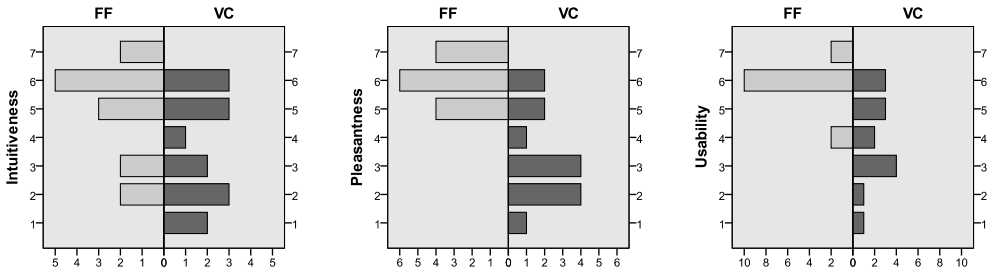
4.2. Time

As shown above, including time as fixed effect to predict if a scene was passed does not significantly improve the model fit, thereby suggesting that if participants completed a scene either quickly or slowly is not related to the fact that the scene was mastered without errors or not.

The aim of the detailed time analysis was to investigate if the different sensory substitution algorithms lead to significantly different completion times. Thus, for the analysis, a subset of data only including passed scenes was created and fit in a LMM with time as continuous outcome variable, algorithm as fixed and individuals and training as random effects, as described in Section 3.5.2. The resulting parameter is the regression coefficient for the fixed predictor of algorithm (on time as outcome variable), $B = 4.98$ $[-1.91, 11.87]$ with SE $= 3.52$, indicating that the choice of algorithm does not influence the time needed for completing the scenes ($F(170, 184) = 2.01$, $p = .159$). In conclusion, using FF does not cause participants to either complete a scene faster or slower, compared to VC.

4.3. Questionnaires

The histograms in Figure 4 report the scores given to each of the 3 questionnaire items. The support in favour of the FF algorithm compared to the VC algorithm was almost unanimous and reflected in all scores, in line with our hypothesis no.3. Intuitiveness FF scores were significantly higher



(a) (b) (c)

Figure 4: Histograms of questionnaire scores. (**a**) Intuitiveness. (**b**) Pleasantness. (**c**) Usability.

($Z = -2.81$, $p = .005$) than VC scores (medians: FF = 5.5, VC = 3.5). Similarly, usability FF scores were significantly higher ($Z = -2.96$, $p = .003$) than VC scores (medians: FF = 6, VC = 4). More interestingly, an overwhelming difference was found in the pleasantness scores (medians: FF = 6, VC = 3), according to which participants highly significantly preferred FF to VC ($Z = -3.2$, $p = .001$). All the participants judged FF sounds pleasant, while 9 participants out of 14 negatively judged the pleasantness of VC sounds. Only one participant gave an equal rating to the two types of sounds, while all other participants gave a higher score to FF sounds.

5. Discussion

The *fluid flow* sensory substitution algorithm proved to be a usable and informative sensory substitution scheme for recognizing the location of obstacles in a simplified blind wayfinding task. This conclusion is supported by the experimental results on a pool of blindfolded sighted participants, who managed to complete the task in 76% of the proposed scenes. It has to be

remarked that the majority of the scenes (see Figure 3) required the participants to travel through spaces as narrow as 80cm without even brushing against an obstacle. If we apply a minimum tolerance on the committed navigation errors and allow for one minor collision per scene, which in the majority of cases meant that participants recognized the obstacle but did not keep enough distance while walking past it, the percentage of completed scenes grows to 86%.

Remarkably, our experimental results indicate a statistically significant superior performance of the fluid flow algorithm compared to the vOICe algorithm in terms of obstacle avoidance and navigation accuracy. This finding is supported by qualitative evaluations from the participants collected at the end of each session. For instance, a subset of participants remarked that they preferred to scan the environment themselves by rotating their heads rather than let the algorithm scan at a fixed rate. This remark supports the use of real-time representation of the environment as provided by the fluid flow scheme rather than the vOICe, whose inherently scanning nature combined with head motion results in an unnatural "scan within a scan" not easy to manage for some participants, at least following a short training session. Another subset of participants reported, following a collision with an obstacle, to have "lost" the obstacle vOICe representation while moving; this issue can also be related to the lack of a real-time feedback for effectively tracking obstacles not only during head movement but also during body movement. Due to the high cognitive load on the working memory imposed by the double-scanning with the vOICe algorithm, two participants reported headache after 2 hours of training, which did not occur with the real-time presentation used

by the fluid flow.

On the other hand, one participant deemed the vOICe algorithm to be more convincing in delivering the spatial layout of the obstacles due to the clear left-to-right scanning mechanism. The participant reported that he found the liquid sound representation of obstacles more difficult to separate when there were two or more obstacles in the field of view of the camera, and that he needed head and body movement to resolve the scene layout. This remark may hint at the necessity of a more consistent training with the fluid flow algorithm in static conditions.

As reported in the previous section, the time required to complete the scenes was not significantly different between the two algorithms. Two participants scored exceptionally good performances, completing most scenes without errors and in less than 30 seconds each, independently of the sensory substitution algorithm. This results indicates a ceiling effect for certain participants, meaning that the scenes were too easy for them to accomplish and therefore they were not able to differentiate between the two algorithms. The ceiling effects covers potential differences between the algorithms; however, this issue only applied for two out of 14 participants. Some participants were on average both faster and more accurate with the fluid flow algorithm than with the vOICe, while other participants considerably slowed down when using the fluid flow sounds. When asked about the latter behaviour, one participant (at the end of her second session) stated that she had a much better understanding of the scene with the fluid flow sounds and felt like she had more control about her performance than with the vOICe algorithm, and therefore devoted more attention to complete the scene without errors.

This conduct is consistent with the fact that prior to the experimental test participants were clearly informed that their task was to minimize navigation errors and not race against time.

The proposed algorithm directly receives as input reliable low-level information conveyed through an off-the-shelf depth sensor, contrary to other sensory substitution schemes previously explored by the authors [9, 50, 49, 13] that used obstacle information segmented through computationally heavy image processing techniques. This is a very desirable property in a system that needs to be scalable in order to run on smartphones or embedded systems with low processing power. The scalability of the proposed approach is further supported by the possibility of reducing the resolution of the depth map without considerable loss of information, as well as changing the size of the oscillator bank for each liquid sound generator at the price of sound quality [4]. This would allow for graceful degradation of our rendering approach depending on the available computational resources. Future work will investigate the quality of experience of the sounds produced by the sensory substitution algorithm even in cases of limited computing power.

One limitation of the current study lies in the use of a sensor with limited field of view and range information, that disoriented some participants in that the obstacle sonification stopped when getting close enough to it, and required considerable head rotation (both yaw and pitch) for a full scan of the scene. Furthermore, although not directly investigated in this study, the choice of the spatialization technique has an undeniable impact on the spatial perception of sounds, and therefore on the degree of immersion [34] and overall quality of experience. The most effective solution would be the use of

individual HRTFs measured on the listener with the addition of head tracking and artificial reverberation [6, 56]. However, obtaining acoustically measured individual HRTF data is only possible with tailored equipment and invasive recording procedures [12]. On the other hand, even though one participant to our study commented that he could "clearly visualize columns of bubbles" where the obstacles were, using non-individual HRTFs is only effective for a limited number of individuals. Different alternative approaches towards HRTF-based spatial rendering were proposed throughout the last decades, ranging from HRTF selection [43, 21] to structural HRTF models [8, 47] and numerical HRTF simulations [26, 59]. Such approaches are expected to progressively bridge the gap between accessibility and accuracy of individual spatial audio [51]. Still, in cases of limited computing power, HRTF rendering can be substituted by constant-power panning [29] to represent horizontal direction at least.

Validation with sighted users implies that these results should only be generalized to the visually impaired population with caution. Blind users are generally more adapted to rely on their sense of hearing for orientation and solving daily mobility challenges compared to sighted, e.g. by using echolocation techniques [42]. This might result in even lower training time required for VIPs to successfully apply the fluid flow algorithm. Furthermore, dynamic postural stability is affected by the visual system, which is why the postural stability of sighted individuals with eyes closed has been shown to be superior to that of blind people [2]. This might result in more collisions when VIPs perform the same task compared to sighted people, even when the obstacle is correctly located in the first place. Hence, to control for these

possible differences between sighted and blind, similar evaluations of the fluid flow algorithm are currently being carried out within the *Sound of Vision*[7] project, ranging from virtual to complex real world environments [13], required for assessing the usability of the system outside the laboratory.

In the final questionnaire, participants reported a clear preference for the fluid flow sounds compared to the vOICe sounds, in terms of intuitiveness, pleasantness, and usability. This result is of great relevance for the integration of the fluid flow sounds in a sensory substitution system for VIPs. Our belief, supported by several participant comments in addition to the questionnaire scores, is that a natural, intuitive, and aesthetically pleasant sonic representation requires little time and effort to be learned while at the same time allowing for longer and less fatiguing practice sessions [45]. In a seminal paper from 2003, yet still as current today as ever, Rocchesso *et al.* [38] assert that *"an aesthetic mismatch exists between the rich, complex, and informative soundscapes in which mammals have evolved and the poor and annoying sounds of contemporary life in today's information society"*, recognizing *"the need for sounds that can convey information about the environment yet be expressive and aesthetically interesting."* In our view, the use of physically based, natural-sounding liquid sounds perfectly matches this need within the field of sensory substitution.

[7] `https://soundofvision.net/`

Acknowledgments

The authors would like to thank Stefano Baldan for his support with the SDT and all the participants involved in this study. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 643636.

Appendix A. Experimental sheet descriptions

FF. The system converts the video stream into a liquid streaming sound produced through superposition of bubble sounds. Bubbles simultaneously come from the visible objects direction in space. The bigger the volume occupied by an object in the visible space, the richer the texture of the corresponding streaming sound (i.e., more bubbles produced). The higher the position of the object in the visible space, the fizzier the bubbles sound. The closer an object within the represented distance range, the louder the liquid streaming sound. If the object gets closer than $1m$, bubbles begin to present a characteristic boiling sound.

VC. The system converts the video stream into a sound made of the superposition of simple tones. The acquired image is scanned in a left to right scanning order, at a rate of one scan per second. Hearing some sound on your left or right thus means having a corresponding object pattern on the left or right side, respectively. During every scan, the higher the pitch, the higher the position of objects in that direction in the visible space. Loudness means distance: the louder the sound, the closer the objects in that direction in the visible space. The bigger the volume occupied by an object in the

visible space, the richer (i.e., more simultaneous tones) and the longer the corresponding sound.

REFERENCES

[1] Avanzini, F., Spagnol, S., Rodá, A., De Götzen, A., March 2013. Designing interactive sound for motor rehabilitation tasks. In: Franinovic, K., Serafin, S. (Eds.), Sonic Interaction Design. MIT Press, Cambridge, MA, USA, Ch. 12, pp. 273–283.

[2] Aydoğ, E., Aydoğ, S. T., Cakci, A., Doral, M. N., May 2006. Dynamic postural stability in blind athletes using the biodex stability system. Int. J. Sports Med. 27 (5), 415–418.

[3] Balakrishnan, G., Sainarayanan, G., Nagarajan, R., Yaacob, S., 2008. A stereo image processing system for visually impaired. Int. J. Signal Process. 2 (3), 136–145.

[4] Baldan, S., Delle Monache, S., Rocchesso, D., 2017. The Sound Design Toolkit. Software XIn press.

[5] Bates, D., Mächler, M., Bolker, B., Walker, S., October 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1), 48 pp.

[6] Begault, D. R., Wenzel, E. M., Anderson, M. R., October 2001. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. J. Audio Eng. Soc. 49 (10), 904–916.

[7] Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., White, J. S. S., March 2009. Generalized linear mixed models: A practical guide for ecology and evolution. Trends Ecol. Evol. 24 (3), 127–135.

[8] Brown, C. P., Duda, R. O., September 1998. A structural model for binaural sound synthesis. IEEE Trans. Speech Audio Process. 6 (5), 476–488.

[9] Bujacz, M., Kropidlowski, K., Ivanica, G., Moldoveanu, A., Saitis, C., Csapó, A., Wersényi, G., Spagnol, S., Jóhannesson, O. I., Unnthórsson, R., Rotnicki, M., Witek, P., July 2016. Sound of Vision - Spatial audio output and sonification approaches. In: Miesenberger, K., Bühler, C., Penaz, P. (Eds.), Computers Helping People with Special Needs - 15th International Conference (ICCHP 2016). Vol. 9759 of Lecture Notes in Computer Science. Springer Int. Publishing, Linz, Austria, pp. 202–209.

[10] Bujacz, M., Strumiłło, P., October 2016. Sonification: Review of auditory display solutions in electronic travel aids for the blind. Arch. Acoust. 41 (3), 401–414.

[11] Capp, M., Picton, P., June 2000. The optophone: an electronic blind aid. Eng. Sci. Educ. J. 9 (3), 137–143.

[12] Cheng, C. I., Wakefield, G. H., April 2001. Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space. J. Audio Eng. Soc. 49 (4), 231–249.

[13] Csapó, A., Spagnol, S., Herrera Martínez, M., Bujacz, M., Janeczek, M., Ivanica, G., Wersényi, G., Moldoveanu, A., Unnthórsson, R., May 2017. Usability and effectiveness of auditory sensory substitution models for the visually impaired. In: Proc. 142nd Conv. Audio Eng. Soc. No. 9801. Berlin, Germany, p. 10 pp.

[14] Csapó, A., Wersényi, G., Nagy, H., Stockman, T., December 2015. A survey of assistive technologies and applications for blind users on mobile platforms: A review and foundation for research. J. Multimod. User Interf. 9 (4), 275–286.

[15] Dakopoulos, D., Bourbakis, N. G., January 2010. Wearable obstacle avoidance electronic travel aids for blind: A survey. IEEE Trans. Syst. Man Cybern. 40 (1), 25–35.

[16] Dell'Aversana, P., Gabbriellini, G., Amendola, A., January 2017. Sonification of geophysical data through time-frequency analysis: Theory and applications. Geophys. Prospect. 65 (1), 146–157.

[17] Dubus, G., Bresin, R., December 2013. A systematic review of mapping strategies for the sonification of physical quantities. PLoS One 8 (12), 28 pp.

[18] Fontana, F., Fusiello, A., Gobbi, M., Murino, V., Rocchesso, D., Sartor, L., Panuccio, A., 2002. A cross-modal electronic travel aid device. In: Human Computer Interaction with Mobile Devices. Vol. 2411 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 393–397.

[19] Garcia-Ruiz, M. A., Gutierrez-Pulido, J. R., July 2006. An overview of auditory display to assist comprehension of molecular information. Interact. Comput. 18 (4), 853–868.

[20] Gardner, W. G., Martin, K. D., June 1995. HRTF measurements of a KEMAR. J. Acoust. Soc. Am. 97 (6), 3907–3908.

[21] Geronazzo, M., Spagnol, S., Avanzini, F., August 2017. Do we need individual head-related transfer functions for vertical localization? The case study of a spectral notch distance metric. Submitted for publication.

[22] Hartzel, J., Agresti, A., Caffo, B., July 2001. Multinomial logit random effects models. Stat. Model. 1 (2), 81–102.

[23] Hermann, T., Hunt, A., Neuhoff, J. G. (Eds.), November 2011. The Sonification Handbook, 1st Edition. Logos Publishing House, Berlin, Germany.

[24] Hosmer, D. W., Lemeshow, S., Sturdivant, R. X., August 2013. Logistic regression models for multinomial and ordinal outcomes. In: Applied Logistic Regression, 3rd Edition. John Wiley & Sons, Inc., New York, NY, USA, pp. 269–311.

[25] Jamal, Y., Lacey, S., Nygaard, L., Sathian, K., 2017. Interactions between auditory elevation, auditory pitch and visual elevation during multisensory perception. Multisens. Res. 30 (3-5), 287–306.

[26] Katz, B. F. G., November 2001. Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. J. Acoust. Soc. Am. 110 (5), 2440–2448.

[27] Kim, H.-Y., February 2013. Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis. Restor. Dent. Endod. 38 (1), 52–54.

[28] Kristjánsson, A., Moldoveanu, A., Jóhannesson, O. I., Balan, O., Spagnol, S., Valgeirsdóttir, V. V., Unnthórsson, R., October 2016. Designing sensory-substitution devices: Principles, pitfalls and potential. Restor. Neurol. Neurosci. 34 (5), 769–787.

[29] Lee, S.-L., Han, K.-Y., Lee, S.-R., Sung, K.-M., August 2004. Reduction of sound localization error for surround sound system using enhanced constant power panning law. IEEE Trans. Consum. Electr. 50 (3), 941–944.

[30] Meijer, P. B. L., February 1992. An experimental system for auditory image representations. IEEE Trans. Biomed. Eng. 39 (2), 112–121.

[31] Merabet, L., Battelli, L., Obretenova, S., Maguire, S., Meijer, P. B. L., Pascual-Leone, A., January 2009. Functional recruitment of visual cortex for sound encoded object identification in the blind. Neuroreport 20 (2), 132–138.

[32] Minnaert, M., 1933. On musical air-bubbles and the sounds of running water. Phil. Mag. 16, 235–248.

[33] Møller, H., Sørensen, M. F., Jensen, C. B., Hammershøi, D., June 1996. Binaural technique: Do we need individual recordings? J. Audio Eng. Soc. 44 (6), 451–469.

[34] Nilsson, N., Nordahl, R., Serafin, S., November 2016. Immersion revisited: A review of existing definitions of immersion and their relation to different theories of presence. Hum. Tech. 12 (2), 108–134.

[35] Pasqualotto, A., Esenkaya, T., April 2016. Sensory substitution: the spatial updating of auditory scenes mimics the spatial updating of visual scenes. Front. Behav. Neurosci. 10 (79).

[36] Pinheiro, J. C., Bates, D. M., 2000. Mixed-Effects Models in S and S-PLUS. Statistics and Computing. Springer-Verlag, New York, NY, USA.

[37] Roads, C., Summer 1988. Introduction to granular synthesis. Comput. Music J. 12 (2), 11–13.

[38] Rocchesso, D., Bresin, R., Fernström, M., April–June 2003. Sounding objects. IEEE Multimedia 10 (2), 42–52.

[39] Rosati, G., Oscari, F., Reinkensmeyer, D. J., Secoli, R., Avanzini, F., Spagnol, S., Masiero, S., June 2011. Improving robotics for neurorehabilitation: Enhancing engagement, performance, and learning with auditory feedback. In: Proc. IEEE 12th Int. Conf. Rehab. Rob. (ICORR 2011). Zurich, Switzerland, pp. 341–346.

[40] Scaletti, C., 1994. Sound synthesis algorithms for auditory data representations. In: Kramer, G. (Ed.), Auditory Display: Sonification, Audification, and Auditory Interfaces. Vol. 1. Adison-Wesley, Reading, MA, USA, pp. 223–251.

[41] Schaalje, G. B., McBride, J. B., Fellingham, G. W., December 2002.

884 Adequacy of approximations to distributions of test statistics in complex
885 mixed linear models. J. Agric. Biol. Environ. Stat. 7 (4).

886 [42] Schenkman, B. N., Nilsson, M. E., April 2010. Human echolocation:
887 Blind and sighted persons' ability to detect sounds recorded in the pres-
888 ence of a reflecting object. Perception 39 (4), 483–501.

889 [43] Seeber, B. U., Fastl, H., July 2003. Subjective selection of non-individual
890 head-related transfer functions. In: Proc. 2003 Int. Conf. Auditory Dis-
891 play (ICAD03). Boston, MA, USA, pp. 259–262.

892 [44] Silberman, N., Hoiem, D., Kohli, P., Fergus, R., October 2012. Indoor
893 segmentation and support inference from RGBD images. In: Proc. 12th
894 Eur. Conf. on Computer Vision (ECCV'12). Florence, Italy, pp. 746–
895 760.

896 [45] Singh, A., Piana, S., Pollarolo, D., Volpe, G., Varni, G., Tajadura-
897 Jimnez, A., CdeC Williams, A., Camurri, A., Bianchi-Berthouze, N.,
898 2016. Go-with-the-Flow: Tracking, analysis and sonification of move-
899 ment and breathing to build confidence in activity despite chronic pain.
900 Hum.-Comput. Interact. 31 (3–4), 335–383.

901 [46] Spagnol, S., Baldan, S., Unnthórsson, R., October 2017. Auditory depth
902 map representations with a sensory substitution scheme based on syn-
903 thetic fluid sounds. In: Proc. IEEE 19th Int. Work. Multi. Signal Pro-
904 cess. (MMSP 2017). Luton, UK.

905 [47] Spagnol, S., Geronazzo, M., Avanzini, F., March 2013. On the relation

between pinna reflection patterns and head-related transfer function features. IEEE Trans. Audio, Speech, Lang. Process. 21 (3), 508–519.

[48] Spagnol, S., Hiipakka, M., Pulkki, V., September 2011. A single-azimuth pinna-related transfer function database. In: Proc. 14th Int. Conf. Digital Audio Effects (DAFx-11). Paris, France, pp. 209–212.

[49] Spagnol, S., Saitis, C., Bujacz, M., Jóhannesson, O. I., Kalimeri, K., Moldoveanu, A., Kristjánsson, A., Unnthórsson, R., September 2016. Model-based obstacle sonification for the navigation of visually impaired persons. In: Proc. 19th Int. Conf. Digital Audio Effects (DAFx-16). Brno, Czech Republic, pp. 309–316.

[50] Spagnol, S., Saitis, C., Kalimeri, K., Jóhannesson, O. I., Unnthórsson, R., October 2016. Sonificazione di ostacoli come ausilio alla deambulazione di non vedenti. In: Proc. XXI Colloquium on Music Informatics (XXI CIM). Cagliari, Italy, pp. 47–54.

[51] Spagnol, S., Wersényi, G., Bujacz, M., Balan, O., Herrera Martínez, M., Moldoveanu, A., Unnthórsson, R., September 2017. Current use and future perspectives of spatial audio technologies in electronic travel aids. Submitted for publication.

[52] Stoll, C., Palluel-Germain, R., Fristot, V., Pellerin, D., Alleysson, D., Graff, C., January 2015. Navigating from a depth image converted into sound. Appl. Bionics Biomech. 2015.

[53] Striem-Amit, E., Guendelman, M., Amedi, A., March 2012. Visual acu-

ity of the congenitally blind using visual-to-auditory sensory substitution. PLoS One 7 (3).

[54] The High Llamas, March 1996. Hawaii [CD]. CD WOOL 2, Alpaca Park.

[55] Ulfvengren, P., 2003. Design of natural warning sounds in human-machine systems. Ph.D. thesis, Royal Institute of Technology, Stockholm, Sweden.

[56] Välimäki, V., Parker, J. D., Savioja, L., Smith, J. O., Abel, J. S., July 2012. Fifty years of artificial reverberation. IEEE Trans. Audio, Speech, Lang. Process. 20 (5), 1421–1448.

[57] van den Doel, K., October 2005. Physically-based models for liquid sounds. ACM Trans. Applied Perception 2 (4), 534–546.

[58] Wald, A., November 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Am. Math. Soc. 54 (3), 426–482.

[59] Ziegelwanger, H., Majdak, P., Kreuzer, W., July 2015. Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization. J. Acoust. Soc. Am. 138 (1), 208–222.

Table 1: Individual experimental results: number of minor/major collisions (C_{MIN}/C_{MAJ}), number of times the participant left the testing area (N_{OUT}) and total completion time (T_{TOT}), divided by participant and sensory substitution algorithm.

	C_{MIN}		C_{MAJ}		N_{OUT}		T_{TOT} [s]	
Participant ID	FF	VC	FF	VC	FF	VC	FF	VC
01	2	1	0	1	0	0	771	702
02	0	2	0	1	0	0	2451	1453
03	0	6	0	1	0	0	1458	1840
04	2	7	1	5	0	0	909	1292
05	6	4	2	7	1	7	717	1648
06	2	2	1	10	1	0	2172	1138
07	8	8	1	6	2	1	1983	1202
08	0	0	1	0	0	1	963	1506
09	0	1	1	4	0	0	1466	1824
10	0	0	1	0	0	0	230	228
11	3	5	5	13	0	0	822	882
12	0	0	0	0	0	0	407	344
13	1	2	1	0	0	0	410	1021
14	4	9	5	9	0	1	1880	1645
Mean	2	3.4	1.4	4.1	0.3	0.7	1188.5	1194.6
SD	2.5	3.1	1.6	4.4	0.6	1.9	713.4	513.4

Table 2: Results of calculating Generalized Linear Mixed Model for Model 1 and Model 2 including one additional predictor, each with individual variance and training as random effects. The model parameter estimates are calculated basing on Laplace approximation with 1 integration point. Shown are regression coefficients with associated standard errors (SE) and confidence intervals (CI), and Wald statistics (z-value and p-value).

	Predictor	**Coeff.**	**SE**	**CI [LL,UL]**	**z-value**	***p*-value**
Model 1	**Algorithm**	-1.30	0.33	[-1.94,-0.66]	-3.96	$p < .001$
Model 2	**Algorithm**	-1.30	0.33	[-1.95,-0.65]	-3.94	$p < .001$
	Time	-0.01	0.01	[-0.02,0.00]	-1.64	$p = .101$