Auditory depth map representations with a sensory substitution scheme based on synthetic fluid sounds

Simone Spagnol
University of Iceland
Reykjavík, Iceland
Email: spagnols@hi.is

Stefano Baldan
Iuav University of Venice
Venice, Italy
Email: stefanobaldan@iuav.it

Runar Unnthorsson
University of Iceland
Reykjavík, Iceland
Email: runson@hi.is

***Abstract*—A novel sensory substitution algorithm based on the sonification of depth maps into physically based fluid flow sounds is described. Spatial properties are extracted from depth maps and mapped into parameters of an empirical phenomenological model of bubble statistics, which manages the generation of the corresponding synthetic fluid flow sound. Following minimal training, the proposed approach was tested in a preliminary experiment with 20 normally sighted participants and compared against the well-known vOICe sensory substitution algorithm. Although the accuracy in recognizing visual sequences based on the corresponding sonification is comparable between the two systems, an overwhelming support for the fluid sounds compared to the vOICe output in terms of pleasantness was recorded. Collected data further suggests that ample margins of performance improvement are achievable following thorough training procedures.**

I. 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 may 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" [1]. Sonification is also used in health care, such as in motor rehabilitation systems [2], [3], electronic travel aids [4], [5], and other assistive technologies for visually impaired persons (VIPs). The majority of these systems are still in their infancy and mostly at a prototype stage. Furthermore, available commercial products have limited functionalities, small scientific and/or technological value and high cost [4].

Available electronic travel aids for VIPs provide various information that ranges from simple obstacle detection with a single range-finding sensor, to more advanced feedback employing data generated from visual representations acquired through camera technologies. The auditory outputs of such systems range from simple binary alerts indicating the presence of an obstacle to complex sound patterns carrying almost as much information as an image [5]. An example of the latter output is provided by the most well-known image sonification algorithm, used in the vOICe system [6]. The vOICe sonification mechanism can be thought of as an inverse spectrogram transform, i.e., a time-varying sound whose spectrogram approximately matches an input grayscale image. It has been shown that, following extensive periods of training and exploiting the neural plasticity of the human brain, the vOICe sonification system can lead to effective sensory substitution [7], both in object recognition [8] and spatial learning [9]. The main drawback is that, even though in some cases the input from the device can be successfully interpreted by naïve users, lengthy and frustrating training with the vOICe system as well as other sensory substitution devices is typically required to perform most tasks [9].

978-1-5090-3649-3/17/$31.00 ©2017 IEEE

The present study explores a novel scheme 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. It was designed in an attempt to (1) improve the vOICe sonification system from both an ergonomic and a functional point of view, eventually reducing the required training time, and (2) enhance the usability and reliability of the conveyed information with respect to previous sensory substitution schemes that rely on segmented object information [10]–[13]. As a matter of fact, while these assume an image processing step that requires an extra real-time computational load and introduces noise, the scheme we propose here directly maps low-order statistics from the raw depth map into the parameters of a fluid flow sound model. This model was specially selected and tuned in order to sound both natural (yet significantly discernible from most daily environmental sounds) and pleasant.

The remainder of the paper is organized as follows. Section II introduces the fluid flow sound model, basic block of the sensory substitution algorithm described in Section III. Section IV outlines a preliminary experiment targeted at comparing performance and individual preference of the proposed sensory substitution scheme with respect to the vOICe algorithm in a scene recognition task. Section V reports the results of the experiment, and Section VI concludes the paper.

II. The fluid flow sound model

The fluid flow model used for this study is a slightly improved version of the bubble simulator proposed by Doel [14]. 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 by 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*.

The model considers individual bubbles to be atomic units (or *grains*, according to the granular synthesis terminology), synthesized using the well-known physically based Minnaert model [15]. 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

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

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

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

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 stochastic process drives an oscillator bank, whose number of voices can be set as a parameter. The size of the oscillator bank defines the true polyphony of the algorithm, i.e. the maximum number of bubbles that can be active at the same time. If 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 clicks during the generation of a new bubble.

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 uniformly distributed random number, 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 uniformly distributed random number, D_{MIN} and D_{MAX} are the minimum and maximum bubble depth 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 [14], the rising bubble is then 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 $\sigma_k = \xi d_k$.

The main differences between the described model and the Doel simulator [14] lie in the evolution and parametrization of the stochastic process driving the bubble population. In the original implementation, the sinusoidal oscillator bank is composed of 50 voices, each one set to a fixed base frequency and driven by a dedicated Bernoulli process. This choice allows to represent only as many different bubble radii as the number of oscillators in the bank. On the contrary, our approach uses a single Bernoulli process for the whole bubble population. This strategy allows to represent bubbles of arbitrary size, improving the versatility and sound quality of the algorithm especially with small oscillator banks. This is a very desirable property in an algorithm which needs to run in real-time, possibly on smartphones, handheld devices and embedded systems with low computing power.

An implementation of the presented model is included in the Sound Design Toolkit (SDT),[1] an open-source (GPLv2) software package designed for research and education in Sonic Interaction Design [16]. The SDT is a library of physics-based sound synthesis algorithms, available as externals and patches for Max and Pure Data. Sound generators are organized according to a hierarchical taxonomy of everyday sounds, based on psychoacoustic experiments on sound production and perception. Audio is generated procedurally rather than by sample manipulation, namely by mathematical descriptions of sound-producing mechanical interactions, which are functional to the creation and design of virtual sonic interactions. The SDT has been successfully employed in the sonification of several interactive installations, and is extensively used in the imitation-driven sonic sketching tools developed within the SkAT-VG project[2] [17]–[19]. The Pure Data version of the SDT fluid flow model was used in the development of the sensory substitution scheme described in the next Section.

III. The sensory substitution algorithm

In the proposed sensory substitution algorithm, a scheme of which is reported in Figure 1, the depth map is divided into 10

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

[2] http://skatvg.iuav.it/

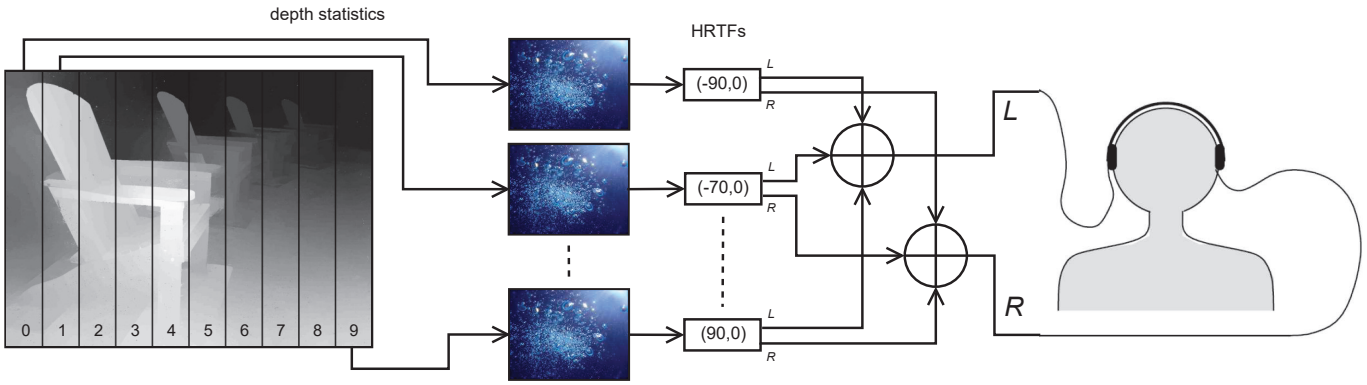


Fig. 1. The proposed sensory substitution scheme. Mapping of sectors 2 to 8 is not shown for the sake of space.

equally wide vertical sectors and each sector corresponds to an independent instance of the fluid flow generator. Descriptive properties of depth information in each sector are mapped into liquid sound features, while the direction of the vertical sector is mapped into spatial sound features. The presence of 10 independent instances of the fluid flow generator guarantees uncorrelation between every pair of outputs, allowing for effective source separation.

More specifically, sounds associated to each vertical sector $N = 0, 1, \ldots, 9$, left to right, are binaurally spatialized by mapping sector N to the azimuth parameter θ of a generic HRTF filter as $\theta = 20N - 90°$, expressed in degrees with respect to the observer according to a vertical polar coordinate system. Even though elevation information is extracted (as detailed in the following), it is not mapped to any spatial parameter, because it is known that in the case of generic HRTF rendering elevation information is not consistent among subjects [20]. Therefore, the elevation parameter of the HRTF filter is set to 0°. The generic HRTF filter 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[3] [21].

A global `max_depth` parameter is defined in order to consider only those points in the depth map whose depth is no greater than it. Then, for each sector, three descriptive depth metrics are calculated: map density, average distance, and average elevation. Map density ρ is defined as the number of pixels where depth is no greater than `max_depth` divided by the total number of pixels in that sector. It is mapped to the average bubble rate Λ according to $\Lambda = 1000\rho^4$: the denser the sector, the more the generated bubbles. The upper limit of 1000 bubbles/second was heuristically set following informal investigations on the pleasantness and intelligibility of the associated stream sound.

Average distance $\bar{d}$ is defined as the mean depth value (in meters) of all pixels with depth no greater than `max_depth` in that sector. It is mapped to the maximum bubble depth D_{MAX} as $D_{MAX} = \min(1/\bar{d} - 0.2, 1)$. 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. 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.

Average elevation $\bar{\varphi}$ is defined as the normalized (0 to 1 starting from below) row index of the sector above which 50% of the reciprocals of depth values no greater than `max_depth` sum up. This parameter basically defines the average elevation where obstacles are concentrated in that sector. It is mapped $1:1$ to the rise factor parameter ξ of the fluid flow generator. It is therefore an elevation indicator in the sense that obstacles concentrated in the upper half of the depth map sector will produce more pitch-rising bubbles than obstacles concentrated in the lower half. As limit cases, when there is no obstacle closer than `max_depth` in the upper half of the sector none of the produced bubbles will rise, while when the same happens in the lower half of the sector then all of the bubbles will do.

Other parameters that define the fluid flow generator are kept constant. These include the minimum and maximum bubble radius (set to $r_{MIN} = 0.15$ mm and $r_{MAX} = 10$ mm), the radius gamma factor ($\gamma_r = 3$), the minimum bubble depth ($D_{MIN} = 0$), the depth gamma factor ($\gamma_D = 1$), and the rise cutoff ($K_\xi = 0.5$). The reasons for fixing these parameters are mainly of perceptual nature (as for r_{MIN} and r_{MAX}, in order for the streaming sounds to be pleasant and "liquid") or due to the need for a reference value, i.e. setting the K_ξ value to correspond to the horizontal bisecting line of the depth map. Beyond perceptual pleasantness, the reason for choosing a greater concentration of small bubbles is due to efficiency.

The algorithm is 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 clicks, the incoming depth map statistics values are smoothed with 100-ms ramps. The number of voices needs to be sufficiently high in order to accommodate for pleasant streaming sounds with a high average bubble rate. However, efficiency issues arise as the number of voices grows. An acceptable trade-off was found by setting the number of voices of each generator to 16.

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

IV. Comparison with the vOICe algorithm

A preliminary experiment was designed in order to have an initial assessment of the performance and individual preference of the proposed sensory substitution algorithm in conveying environmental spatial information through sound. More specifically, the objectives of the designed experiment are to

1) assess the capability of the fluid flow sounds to give reliable and distinguishable information about visual scenes following minimal user training;
2) investigate the degree of difficulty in understanding the sensory substitution scheme with naïve sighted users;
3) collect individual judgments about the pleasantness and usefulness of the sounds that are conveyed;
4) compare all of the above results and ratings against those collected using the reference sensory substitution scheme provided through the vOICe algorithm.

It has to be stressed that although the original vOICe system was designed to sonify simple 2D grayscale images, a depth map can be easily converted into a grayscale image where brightness corresponds to depth. Furthermore, 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 [22]–[24].

An implementation of the vOICe sensory substitution algorithm was carried out in Pure Data following the specifications from Meijer [6]. 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 `max_depth`, to 1 for pixels of zero depth. The auditory output of the implemented system was compared against the original vOICe software for Windows on a small benchmark depth map set, and it was found to never exceed 1 dB of spectral distortion in the $0.5 - 5$ kHz range. For the sake of consistency with the fluid flow scheme, the output was binaurally spatialized from $90°$ left to $90°$ right on the horizontal plane with the `earplug~` Pure Data external.

The experiment consisted in identifying through sound a number of video sequences taken from a third-party depth map dataset, the NYU-Depth Dataset V2[4] [25], sonified with either the vOICe algorithm or the fluid flow algorithm. The dataset is comprised of video sequences from a variety of indoor scenes (rooms, kitchens, stores, offices, and so on) as recorded by both the RGB and depth cameras from a Microsoft Kinect. Here we used the *raw* dataset, containing the raw image and accelerometer dumps from the Kinect. According to the dataset documentation, the RGB and depth camera sampling rate lies between 20 and 30 FPS (variable over time). Therefore, the timestamps for each of the RGB, depth and accelerometer were synchronized to produce continuous video sequences. Furthermore, the raw depth images were projected onto the RGB coordinate space in order to align the images. Following a thorough check of the database, some video sequences were discarded because of insufficient duration (less than 10 s) or major skips. A total of 470 sequences were finally kept. The `max_depth` parameter was set to 5 meters in this experiment. The experimental procedure was implemented in MATLAB.

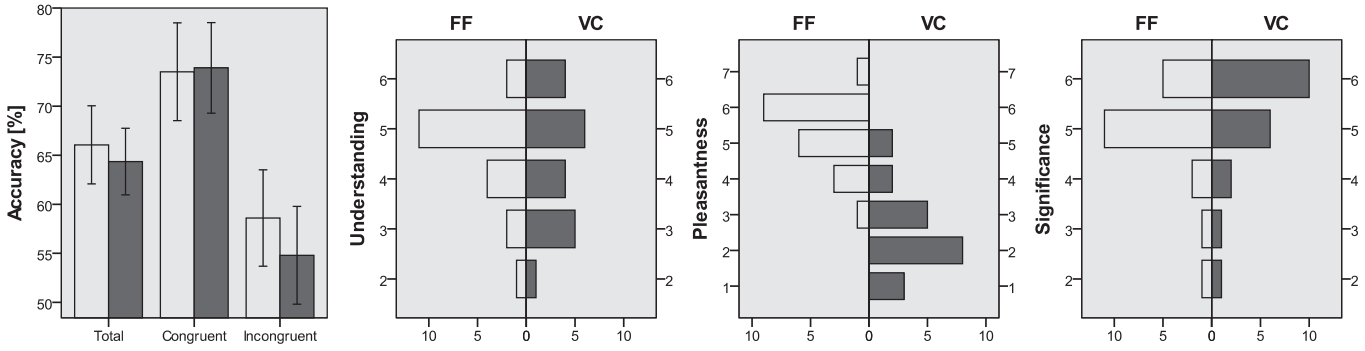
[4]http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html

Fig. 2. Participant during the experiment.

Twenty participants (10F, 10M) participated on a voluntary basis. Ages ranged from 18 to 65 ($\mathrm{M} = 30.65$, $\mathrm{SD} = 11.75$). All of them self-reported normal or corrected vision and normal hearing. Every participant signed and dated an informed consent form, which included short descriptions (7 lines) of each of the two tested sensory substitution mechanisms, referred to as "Bubble" (fluid flow) and "Scan" (vOICe). The experiment was run on a Dell XPS 13 laptop inside a silent office environment, and sound was conveyed through a pair of Sony MDR-1R headphones (see Figure 2).

The experiment was split into two experimental sessions, each testing one single sensory substitution algorithm (fluid flow and vOICe). The order of presentation of the two systems was randomized and balanced across participants. Within a single experimental session, participants watched 100 randomly drawn RGB subsequences of 250 frames (about 10 seconds each) from the database while at the same time listening to an auditory representation, that could be either congruent (direct sonification of the corresponding depth subsequence) or incongruent (sonification of a subsequence from a different randomly chosen sequence). At the end of each trial, participants judged whether the auditory representation was congruent with the video sequence (by pressing the *Y* key) or not (by pressing the *N* key).

Two short training sessions introduced the experimental session. In the first training session 10 sample sequences with congruent auditory information were shown to the participant.



(a) Accuracy scores. (b) Understanding scores. (c) Pleasantness scores. (d) Significance scores.

Fig. 3. Descriptive statistics divided by sensory substitution algorithm: FF (light gray) = fluid flow, VC (dark gray) = vOICe.

The second training session followed the same procedure as the experimental session, except that the participant was given yes/no information about the correctness of his/her feedback right after each trial (10 in total). At the end of the experimental session, participants replied to a brief questionnaire about the corresponding sensory substitution scheme by ticking one item in each of three 7-point Likert scales (1 = strongly disagree, 7 = strongly agree):

1) I think that the meaning of the sounds is easy to understand;
2) I feel that the sounds are pleasant;
3) I think that the sounds provide significant information about the visual sequence.

The total duration of the experiment was $60 - 70$ minutes.

V. Results

Figure 3 reports descriptive statistics of the experiment divided by sensory substitution algorithm (FF = fluid flow, VC = vOICe). The barplot in Figure 3a reports the mean and 95% confidence interval of the individual scores (percentage of correct responses), both for all trials and divided by congruent and incongruent trials. The histograms in Figures 3b–3d report the scores given to each of the 3 questionnaire items (understanding, pleasantness and significance, respectively).

The assumption of data normality was verified for all accuracy scores through the Shapiro-Wilk test. Therefore, we ran paired t-tests across participants on total accuracy scores, which revealed no significant differences between FF and VC ($t = 0.51$, $p = 0.62$). However, we found a significantly higher accuracy in identifying congruent rather than incongruent audiovisual sequences, both with FF ($t = 3.44$, $p = 0.006$) and with VC ($t = 3.67$, $p = 0.004$). According to a one-sample t-test, accuracy scores were strongly significantly higher than the 50% chance level in both algorithms (FF: $t = 8.44$, $p < 0.001$; VC: $t = 8.83$, $p < 0.001$), but when divided into congruent and incongruent sequences, the accuracy of incongruent sequences only showed a mildly significant difference from chance level for VC ($t = 8.83$, $p = 0.06$). Significance scores for accuracy barely change when selecting for analysis only those subjects who had scores above 60% in both sessions, which is considered a safe margin from chance performance.

For what concerns the questionnaires, we investigated for differences in individual scores between the two algorithms by running three separate Wilcoxon signed-rank tests, one per questionnaire item. Participants were mostly agreeing with the understanding and significance items for both algorithms. Understanding scores for FF were slightly higher (medians: FF = 5, VC = 4.5), but not significantly ($Z = -0.63$, $p = 0.53$). The opposite occurred with significance scores, that were slightly higher for VC (medians: FF = 5, VC = 5.5), but again not significantly ($Z = -0.65$, $p = 0.51$). However, an overwhelming difference was found in the pleasantness scores (medians: FF = 5.5, VC = 2), according to which participants significantly and clearly preferred FF to VC ($Z = -3.85$, $p < 0.001$). Only 1 participant out of 20 judged the FF sounds mildly unpleasant, while 16 participants out of 20 negatively judged the pleasantness of VC sounds.

Further analyses were conducted between groups in order to check whether scores differed by sex or by order of presentation of the two algorithms. No differences were found between males and females in any score. On the other hand, a mild impact of the order of presentation on the accuracy scores was seen: while VC had roughly the same average accuracy when presented as first or second (means: 64.4% and 64.3%, respectively), FF had mildly significantly higher accuracy ($t = 1.81$, $p = 0.09$ according to an independent samples t-test) when presented as second rather than first (means: 62.8% first, 69.3% second). This result suggests that the proposed scheme could lead to even higher scores following a more thorough learning of the task under consideration.

VI. Conclusion

Our preliminary evaluation of the proposed sensory substitution scheme proved that sighted participants were able to recognize coherent audio-visual information with the same accuracy as with the reference sensory substitution algorithm provided through the vOICe algorithm. It has to be pointed out that the relatively low accuracy results may likely be associated to the minimal amount of training and to the nature of the

experimental task, that required participants to discriminate between congruent and incongruent information without having an absolute reference to separate the two categories. Some participants scored performances close to chance level that might also be justified with an insufficient understanding of the sonification rationale. However, other participants could reach accuracy scores as high as 80% with the fluid flow algorithm even with little training and no previous knowledge of the aims and methods of the experiment. Furthermore, the overwhelming support of the fluid flow scheme in terms of pleasantness of the sounds conveyed promotes its usability.

The fluid flow sonification strategy applied to blind navigation is currently being tested in Iceland within the Sound of Vision[5] project, both in virtual and real world environments and with a pair of hear-through headphones not blocking environmental sounds [26]. Reports from the first training sessions with VIPs are encouraging and providing insightful feedback towards possible improvements of the sonification scheme. As an example, the considered sectors from the depth map can be extended to cover a 2D grid in order to provide more accurate elevation information. A hybrid sonification scheme combining both raw depth map information and a basic object segmentation will also be investigated in order to allow discriminating between generic obstacles and walls.

Acknowledgment

The authors would like to thank Marcelo Herrera Martínez and Rebekka Hoffmann for pilot testing and fruitful discussions, 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 (Sound of Vision) and from the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission under FET-Open grant No. 618067 (SkAT-VG).

References

[1] C. Scaletti, "Sound synthesis algorithms for auditory data representations," in *Auditory Display: Sonification, Audification, and Auditory Interfaces*, G. Kramer, Ed. Reading, MA, USA: Adison-Wesley, 1994, vol. 1, pp. 223–251.

[2] F. Avanzini, S. Spagnol, A. Rodá, and A. De Götzen, "Designing interactive sound for motor rehabilitation tasks," in *Sonic Interaction Design*, K. Franinovic and S. Serafin, Eds. Cambridge, MA, USA: MIT Press, March 2013, ch. 12, pp. 273–283.

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

[4] D. Dakopoulos and N. G. Bourbakis, "Wearable obstacle avoidance electronic travel aids for blind: A survey," *IEEE Trans. Syst. Man Cybern.*, vol. 40, no. 1, pp. 25–35, January 2010.

[5] M. Bujacz and P. Strumiłło, "Sonification: Review of auditory display solutions in electronic travel aids for the blind," *Arch. Acoust.*, vol. 41, no. 3, pp. 401–414, October 2016.

[6] P. B. L. Meijer, "An experimental system for auditory image representations," *IEEE Trans. Biomed. Eng.*, vol. 39, no. 2, pp. 112–121, February 1992.

[7] A. Kristjánsson, A. Moldoveanu, O. I. Jóhannesson, O. Balan, S. Spagnol, V. V. Valgeirsdóttir, and R. Unnthórsson, "Designing sensory-substitution devices: Principles, pitfalls and potential," *Restor. Neurol. Neurosci.*, vol. 34, no. 5, pp. 769–787, October 2016.

[8] E. Striem-Amit, M. Guendelman, and A. Amedi, "Visual acuity of the congenitally blind using visual-to-auditory sensory substitution," *PLoS One*, vol. 7, no. 3, March 2012.

[9] A. Pasqualotto and T. Esenkaya, "Sensory substitution: the spatial updating of auditory scenes mimics the spatial updating of visual scenes," *Front. Behav. Neurosci.*, vol. 10, no. 79, April 2016.

[10] M. Bujacz, K. Kropidlowski, G. Ivanica, A. Moldoveanu, C. Saitis, A. Csapó, G. Wersényi, S. Spagnol, O. I. Jóhannesson, R. Unnthórsson, M. Rotnicki, and P. Witek, "Sound of Vision - Spatial audio output and sonification approaches," in *Computers Helping People with Special Needs - 15th International Conference (ICCHP 2016)*, ser. Lecture Notes in Computer Science, K. Miesenberger, C. Bühler, and P. Penaz, Eds. Linz, Austria: Springer Int. Publishing, July 2016, vol. 9759, no. II, pp. 202–209.

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

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

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

[14] K. van den Doel, "Physically-based models for liquid sounds," *ACM Trans. Applied Perception*, vol. 2, no. 4, pp. 534–546, October 2005.

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

[16] S. Baldan, S. Delle Monache, D. Rocchesso, and H. Lachambre, "Sketching sonic interactions by imitation-driven sound synthesis," in *Proc. 13th Int. Conf. Sound and Music Computing (SMC 2016)*, Hamburg, Germany, September 2016.

[17] D. Rocchesso, G. Lemaitre, S. Ternström, P. Susini, and P. Boussard, "Sketching sound with voice and gesture," *ACM Interactions*, vol. 22, no. 1, pp. 38–41, January/February 2015.

[18] D. Rocchesso, D. A. Mauro, and S. Delle Monache, "miMic: The microphone as a pencil," in *Proc. 10th ACM Int. Conf. on Tangible, Embedded, and Embodied Interaction (TEI'16)*, Eindhoven, Netherlands, February 2016, pp. 357–364.

[19] O. Houix, S. Delle Monache, H. Lachambre, F. Bevilacqua, D. Rocchesso, and G. Lemaitre, "Innovative tools for sound sketching combining vocalizations and gestures," in *Proc. Audio Mostly 2016*, Norrköping, Sweden, October 2016, pp. 12–19.

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

[21] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.*, vol. 97, no. 6, pp. 3907–3908, June 1995.

[22] M. Capp and P. Picton, "The optophone: an electronic blind aid," *Eng. Sci. Educ. J.*, vol. 9, no. 3, pp. 137–143, June 2000.

[23] G. Balakrishnan, G. Sainarayanan, R. Nagarajan, and S. Yaacob, "A stereo image processing system for visually impaired," *Int. J. Signal Process.*, vol. 2, no. 3, pp. 136–145, 2008.

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

[25] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. 12th Eur. Conf. on Computer Vision (ECCV'12)*, Florence, Italy, October 2012, pp. 746–760.

[26] S. Spagnol, G. Wersényi, M. Bujacz, O. Balan, M. Herrera Martínez, A. Moldoveanu, and R. Unnthórsson, "Current use and future perspectives of spatial audio technologies in electronic travel aids," *J. Audio Eng Soc. (submitted for publication)*, June 2017.

[5] https://soundofvision.net/