Audio Engineering Society

Convention Paper 9801

Presented at the 142nd Convention
2017 May 20–23 Berlin, Germany

Usability and Effectiveness of Auditory Sensory Substitution Models for the Visually Impaired

Adam Csapo[1], Simone Spagnol[2], Marcelo Herrera Martinez[2], Michal Bujacz[3], Maciej Janeczek[3], Gabriel Ivanica[4], Gyorgy Wersenyi[1], Alin Moldoveanu[4], and Runar Unnthorsson[2]

[1] *Széchenyi István University, Győr, Hungary*

[2] *University of Iceland, Reykjavik, Iceland*

[3] *Lodz University of Technology, Lodz, Poland*

[4] *Politechnica University of Bucharest, Bucharest, Romania*

Correspondence should be addressed to Adam Csapo (csapo.adam@sze.hu)

ABSTRACT

This paper focuses on auditory sensory substitution for providing visually impaired users with suitable information in both static scene recognition and dynamic obstacle avoidance. We introduce three different sonification models together with three temporal presentation schemes, i.e. ways of temporally organizing the sonic events in order to provide suitable information. Following an overview of the motivation and challenges behind each of the solutions, we describe their implementation and an evaluation of their relative strengths and weaknesses based on a set of experiments conducted in a virtual environment.

1 Introduction

The goal of the Sound of Vision H2020 project is to develop an all-purpose wearable solution to visual perception through auditory and haptic sensory substitution for the visually impaired [14]. The solution is expected to be non-invasive and self-sufficient, i.e. without any external infrastructural requirements. It is also expected to be suitable for a wide range of use-case scenarios, including static scene recognition and dynamic obstacle avoidance. The project includes teams for developing auditory and haptic models, developing / carrying out testing protocols in virtual and real-world environments, and implementing a wearable hardware solution.

This paper provides a summary of the most recent audio solutions developed within the project, as well as a short evaluation of their relative strengths and weaknesses. For the purposes of separating the influence of different factors on usability and effectiveness, we distinguish between audio models (i.e. ways of producing sonic events which arise as an image-to-sound transformation to provide a representation of reality) and temporal presentation schemes (i.e. ways of temporally organizing the sonic events). Three different audio models were used in current investigations: 1. a metallic bar impact physical sound model with varying characteristics of pitch, duration and amount of oscillation; 2. a bursting bubble sound model with varying characteristics of starting and ending pitch as well as sweeping velocity between the two; and 3. a granular synthesis model with varying grain and grain texture characteristics including grain pitch, length and density. Several different temporal presentation schemes were used in conjunction with the three audio models (though not in every combination for reasons discussed in the paper) including : 1. an "*objects as loudspeakers*" scheme in which each object was regarded as an individual sound-producing source at each point in time; 2. a "*left-to-right scanning*" scheme in which sonic events from different horizontal regions occurred in a time-division multiplexed, non-overlapping fashion; and 3. an "*expanding sphere*" scheme in which the presentation of sonic events was organized according to a periodical depth scan.

In previous work, preliminary studies were carried out using earlier variants of these models in a virtual training environment. Findings suggested that different sonification schemes were suitable for different kinds of tasks, while subjective user questionnaires showed that the models were perceived as rather similar in terms of their "cyclical nature" and "lack of continuity". For this reason, the enhanced models presented in this paper were designed with partially separate use-cases and larger perceptual variety in mind. In particular, the metallic bar impact and granular models were revised, while the bursting bubble sound model was added to the set of candidate solutions. Similarly, possibilities opened through the continuous nature of the "objects as loudspeakers" temporal presentation scheme were explored to a fuller extent.

The paper is structured as follows. Section 2 introduces the three audio models and describes their implementation. Section 3 describes the three temporal presentation schemes and the hypotheses formulated with respect to their effectiveness. Section 4 describes the results of preliminary tests. Finally, Section 5 presents conclusions drawn from this work and discusses plans for future work.

2 Audio Models

The general goal of the audio models presented in this section is to provide real-time feedback to users on the geometric properties of objects segmented from a video stream that is generated based on input from a head-mounted camera. All models presuppose that properties such as the quantity of visible objects, as well as their height, width, elevation, distance and azimuth (i.e. direction in the horizontal plane) are available. In the following, we describe the three models and their implementation.

2.1 Bar impact model

2.1.1 General overview of the model

One of the first approaches we considered in the project was to simulate the actual striking of obstacles with a white cane [18]. The model treated each object in the frontal hemisphere of the user as an independent virtual sound source that continuously emits impact sounds. The pitch and timbre of the sound resulting from the impact were considered dependent on the object's width and category, while the distance between object and user was coded into loudness: the closer the object, the higher the sound level. Furthermore, each sound was spatialized in accordance with the direction of the object with respect to the user. Experimental results suggested that the adopted sonification approach might lead to improved results if adequate modifications were performed, either to the mapping schemes or to the chosen sound stimuli.

The reason for choosing impact sounds to convey information about objects was twofold. First, the ecological validity of physics-based sounds, whose nature allows a direct association to the virtual act of detecting the object by striking it with a cane, was considered as an advantageous property. Second, the peculiar pattern of impact sounds, whose rich frequency content and short attack phase give rise to dynamic perceptual qualities, could be expected to result in improved sound localization on the horizontal plane [4]. Furthermore, design choices concerning the mappings between object and sound properties were made based on physical ground [1].

The model was improved by considering an alternative implementation of a struck metal bar and by varying some of the above mappings in order to provide an exclusive set of relations between object parameters and physical parameters.

2.1.2 Implementation of the model

We implemented a simplified physical model of a struck metal bar using the *barmodel* Csound opcode, developed by Bilbao and Fitch [3][1]. The model uses

[1] http://www.csounds.com/manual/html /barmodel.html

a differential equation simulating wave propagation in a metal bar, allowing to control various properties of the bar (e.g. dimensions and support clamps), its material (e.g. stiffness, wave propagation speeds) and impacting hammer (size and velocity). In the model, each generic object in the scene is associated with its estimated barycenter, whose direction with respect to the cameras is spatialized. The sounds are filtered using the KEMAR HRTFs built into CSound, although personalized HRTFs can be used as well [6]. Optionally, the output can be directed to custom multi-speaker headphones that allow panning not only between left and right, but also up and down [5].

2.1.3 Parameter mappings used

The bar model allows for control of the following parameters of a metal bar struck by a hammer:

- bar properties: stiffness, loss of high-frequencies, 30dB decay time
- boundary conditions and output scanning speed
- strike properties – position on the bar, width and strike velocity.

In our parameterization, both ends of the bars are clamped in order to guarantee a degree of consistency between the timbres of different instances. The estimated width of the object (ranging from 0.4m to 5m) is linearly mapped to the stiffness of the bar, such that wider objects correspond to smaller stiffness values ranging from 100 to 400. Since stiffness has a direct relation with pitch, wider objects correspond to lower pitches. Additionally, width is linearly translated to duration, from 0.3s to 1s for smallest to largest objects. Absolute distance between the listener and the object (from 0 up to 5m) is inversely mapped onto the strike velocity of the exciter (from 2000 to 15000), so that closer objects produce significantly louder sounds. The position and width of strike are kept constant at 50 (middle position) and 0.5 (half the width of the bar). Object elevation is mapped to the timbre of the sound: face-level obstacles are coded with a higher scanning speed, thus producing modulated "ringing" sounds, while grounded obstacles sound more "dry". This mapping was derived experimentally and is a power function of the angle between the camera vector and the elevation of an obstacle, capped at -10 and +10 degrees. The scanning speed is equal to $1.5^{(\text{angle}+2)}$, which ensures that sound modulation for angles above the center of the line of sight is significantly increased.

2.1.4 Markers and special sounds.

The bar impact model also considers a special division for obstacle categories – i.e. walls vs. generic obstacles. Walls are coded with a lower stiffness range (below 100), making them resonate for longer periods of time than smaller obstacles. Additionally, the duration of oscillations is dependent on the relative orientation of the wall to the observer – the more perpendicular oreintation, the shorter the resulting wall sound.

Further categories of special objects are implemented, to which auditory icons [11, 12, 7] with a duration of 0.5s are assigned. These include specially recognized scene elements such as stairs, doors, holes / discontinuities in the ground and text.

When performing sonification using the expanding sphere or left-right scanning presentation schemes (described later), special marker sounds are played, which are short wooden percussive "ticks" every 1m (expanding sphere) or 30 degrees (left-right). These help the listener to a large extent in estimating the position of obstacles.

2.2 Bubble sound model

2.2.1 General overview of the model

The bubble sound model is a natural sounding model that conveys information about direction, distance, and size, i.e. the fundamental properties of one or more generic objects. The model was designed in an attempt to improve the impact sound model, especially in terms of pleasantness of the synthesized sounds. A key additional property of these sounds is that, while they sound natural, they are also significantly different from normal sounds from the environment.

The atomic sound unit in this model is the *bubble*, defined as a thin sphere of liquid enclosing air. The acoustic mechanism responsible for bubble sounds is volume pulsation, which was first correctly identified by Minnaert [15]. Bubbles are typically formed when the water surface causes air to be trapped in the water, usually accompanied by an energy injection into the bubble at creation time. After formation, the bubble emits a sinusoidal sound that decays as energy is dissipated. The impulse response $i(t)$ of a radially oscillating bubble is expressed as

$$i(t) = a\sin(2\pi f t)e^{-dt} \quad (1)$$

where f is the resonance frequency, d is the damping factor, a is the amplitude, and t is time. If the bubble survives long enough, this is all that happens. If the bubble is formed close to the water-air interface and is rising, the pitch of the bubble rises, giving the familiar "*blooink*" sound that is audible sometimes when a stone is thrown in water, and the appropriate cavity is formed. The rising bubble is modelled by making the frequency time dependent according to

$$f(t) = f_0(1 + Dft) \quad (2)$$

where f_0 is the Minnaert frequency and Df is the slope of the frequency rise, related to the vertical velocity of the bubble. However, a perceptually more relevant parameter is the audible rise in pitch, which also depends on the damping factor d of the bubble sound. By modeling the slope of the frequency rise as $\sigma = \xi d$, the effect of damping is taken into account and ξ roughly parameterizes the audible rise [20].

2.2.2 Implementation of the model

In the bubble sound model, each object in the scene corresponds to a virtual bubble. Single bubble sounds are generated through the above-described physical model, an implementation of which is included in the Sound Design Toolkit (SDT)[2], which is an open-source (GPLv2) software package suitable for research and education in Sonic Interaction Design [8]. The SDT consists of a library of physics-based sound synthesis algorithms, available as externals and patches for Max and Pure Data [3]. The latter version was used in the development of this sound model.

The two parameters used in the implementation of the physical model were the radius r (which controls the starting pitch and the duration of the bubble sound) and the rise factor ξ (which controls the frequency excursion of the "*blooink*" sound). It can be verified that these two parameters uniquely define the damping factor d, the resonant frequency f_0, and the slope of the frequency rise Df [20]. Therefore, the only remaining parameter to fully determine the impulse response of the radially oscillating bubble is amplitude a, which was made proportional to ξ.

2.2.3 Parameter mappings used

Coming to the bubble model design, the size of the object is directly mapped to the bubble radius. Here size is intended as a single parameter merging the object's width w and height h, mapped onto the bubble radius parameter r as

$$r = 10.6\sqrt{\left(\frac{w-0.2}{1.8}\right)^2 + \left(\frac{h-0.2}{1.8}\right)^2}\ [m] \quad (3)$$

with $w, h > 0.2\,m$. In accordance with everyday physics [20], the larger the bubble, the lower is the pitch and the longer the bubble sound. Azimuth and elevation of the object, expressed in degrees with respect to the observer according to a vertical polar coordinate system, are directly mapped to the same parameters of a generic HRTF filter 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 [10][4]. It has to be stressed that the accuracy of virtual sound localization depends on multiple factors independent of the sound model itself, including the type of headphones used, the headphone equalization filter, and the choice of the HRTF set [2]. In particular, spatialization is non-individual; however, models for

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

[3] https://puredata.info/

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

HRTF individualization such as structural models [17,13] or individual HRTFs themselves can be integrated (at an additional measurement cost) if higher spatial accuracy is needed.

Since it is hard to correctly convey elevation information in the case of generic HRTF rendering [2], the elevation parameter φ is redundantly mapped onto the rise factor parameter ξ of the bubble as follows:

$$\xi = \frac{\varphi + 40}{80}, \qquad (4)$$

so that a minimum elevation of $\varphi = -40°$ corresponds to a rise factor of zero (bubble fully submerged), and the symmetrical elevation $\varphi = 40°$ corresponds to a rise factor of one, which gives an audible frequency rise of an octave (bubble just below the surface). This is consistent with the fact that frequency rises are increasingly observed for bubbles closer and closer to the water surface [20].

Finally, similarly to other sound models investigated in the Sound of Vision project [5,18], in this model distance information is exclusively conveyed through the "*expanding sphere*" cyclic scan paradigm described in Section 2.3. One difference with respect to other sound models is that the bubble sound model does not use multiple marker sounds for different reference distances, but a single 10-ms 100-Hz pulse signaling the start of a cycle. This design choice was made in order to reduce the number of sonic events within each cycle, thus the complexity of the model.

2.3 Granular synthesis based model

2.3.1 General overview of the model

The goal of this model was to generate sounds that are perceptually varied but still suitable for giving users a general overview of the visual scene. Originally it was expected that this model would yield results that are less crisp in their interpretation, but at the same time interesting enough to be of use in explorative tasks such as user navigation.

Granular synthesis is a well-known sound synthesis technique based on the random selection and temporal overlaying of miniature sound samples taken from a pre-defined source signal [9, 21, 16]. Such sound samples, referred to as grains, generally have a length on the order a few tens of milliseconds, and therefore do not in themselves constitute musical sounds. However, when the sampling process is replicated at different temporal offsets and frequencies, and the resulting grains are overlapped with each other, it is possible to obtain perceptually varied and dynamically rich sonic textures.

2.3.2 Implementation of the model

The granular synthesis model was implemented using the built-in *grain* opcode in Csound. The opcode accepts a number of parameters, including:

- *amplitude*
- *sampling frequency*
- *stream density*
- *maximum variation* of the first two parameters (amplitude and sampling frequency, which are controlled, but random)
- *grain duration*
- *input audio sample id* (stored in an array within memory, referred to as a wave table)
- *window function id* (a wave table used as a multiplicative envelope for individual grains)
- *random sampling* – a boolean parameter that determines whether or not the point at which grain sampling occurs is selected at random within the input audio sample.

In the implementation of the model (irrespective of the temporal presentation scheme used), the visual scene is divided into a 2 (or 3)-dimensional grid of cells, each of which are sonified individually and the output of which are temporally ordered based on the specific presentation scheme. Various input sources were considered as input audio samples to the granular synthesis model, including piano music, recordings of percussion instruments, solo vocal performances and orchestral excerpts. A specific marker sound was also designed using the same opcode, but with specific settings, to produce a pleasant "rumbling" effect for use at the end of scanning phases (specifically in the left-to-right scanning scheme, which will be described later). In order to be better able to control the pitch of the resulting granular stream (at least in terms of a few perceptual categories, such as 'high', 'medium' and 'low' register), bandpass filtering was also experimented with. Later this idea was discarded due to the limited capabilities of the filters to alter the perceptual qualities of the resulting stream. However, random sampling was turned off instead, as well as the potential variation in sampling frequency and stream density maximally reduced, with the goal of reducing the randomness of resulting streams as much as possible.

2.3.3 Parameter mappings used

Mapping from the visual to auditory domains was determined such that the model represents objects that are closer (as opposed to farther away) by a relatively greater number of longer grains – resulting in louder and smoother textures for closer objects. Objects that are taller are generally associated with a higher pitch (i.e. the original sound source is sampled at a higher frequency when producing the grains) than shorter ones. Objects that have a larger volume correspond to a relatively higher grain density, allowing for the interpretation to arise that objects covering more pixels within a cell of fixed length are "more dense". Finally, the direction of objects is represented through generic KEMAR HRTFs using the *hrtfmove* opcode.

In terms of concrete values, the model uses:

- linear mapping from distances between 0 - 5m to amplitude between 1 and 0.
- linear mapping from distances between 0 – 5m to grain duration between 0.5 and 0.02, such that shorter distances are represented using longer grains.
- non-linear (piecewise linear and exponential) mapping from ratio of data points (belonging to the object) within the individual cell and the size of the cell in the visual scene (between 0 and 100%) to grain density. Density values between 2 and 800 are used, with linear inflection points at 30% (density = 10) and 60% (density = 60), and an exponential rise from 60 to 800 grains per second in the top 40%.
- a linear mapping from elevation from ground (on a scale of 0 to 10) to scanning frequency of sampled grains, ranging from 0.1 to 2.5 times a "base playing rate" of the input audio sample, which corresponds to playing the sample in 1 second
- generic KEMAR HRTFs to control the perceived azimuth of the object from -90 to 90 degrees.

Based on the above, it was expected that closer and taller objects would produce louder streams and smoother (as opposed to percussive) streams at higher frequencies, but that objects with a relatively large volume would also produce streams with higher grain densities, hence louder sounds. Initially, we experimented with the randomization of grain amplitudes and frequencies, but eventually it was decided that it would be more effective if the variation among similar point clouds could be reduced. As a result, fixed values of 0 and 0.001 were chosen for variation in amplitude and scanning frequency, respectively.

Finally, the parameters for the marker sounds were selected such that the input sound source was a full sine wave, the sampling frequency was 70Hz, the stream density was 800 grains per second, variation in amplitude and sampling frequency was set to low values (0.01 and 0.5, respectively) and grain duration was set at 1.2 seconds (the point here was to set the grains to have a much longer duration than the duration for which the marker sounds would be played, resulting in a relatively constant "rumbling" sound with an attack but no decay phase).

3 Temporal Presentation Schemes

3.1 Objects as loudspeakers

This presentation scheme renders all the segmented objects received from the 3D module simultaneously and continuously. Each segmented object in the frontal hemisphere of the user becomes an independent virtual sound source that continuously emits a specific sound. The strength of the method lies in the continuity and simultaneity of the encoding. The weakness is that when used with discrete and sparse sound models, sounds might turn out to be unorganized and cluttered [19]. This is the reason why presentation scheme was used primarily in conjunction with the dense granular synthesis model, as described in Section 2.3.

3.2 Left to right scanning

The motivation behind left-to-right scanning was to clearly convey the direction of scene elements. This temporal presentation scheme can either be applied to a list of segmented objects – by playing their sounds in order from left to right – or directly to a depth-map – by sonifying the distance to the nearest obstacle in a number of directions (e.g. every 5-15°). As left-to-right scanning renders object sounds in order from left to right, it can be expected to be a perceptually slower method than the expanding sphere (see next subsection), but possibly clearer to understand. Our working hypothesis was that it would be primarily useful for scene perception and less well-suited to mobility purposes.

The depth-map based version of left-to-right scanning was used in conjunction with granular synthesis. In this case, the depth map was divided into rows and columns (leading to 'cells' on the image characterized by a height and width dimension). As a result of this layout and temporally distinct columns, the model provided a clear representation of direction and number of obstacles.

3.3 Expanding sphere

The motivation behind the expanding sphere temporal presentation scheme was to clearly convey distance to the nearest obstacles, while still allowing for the incorporation of both generic object sounds (currently synthesized using the physical bar model and the bubble model) and special object sounds (using short auditory icons). In this presentation scheme, object sounds are played in order of proximity. The core concept of the scheme is a virtual scanning sphere that originates at the subject's head and expands throughout the scene. The sphere is preferred to a scanning plane paradigm in order to preserve radial distances between objects and observer. As the surface of the sphere intersects scene elements (generic and special objects as points, walls as surfaces), sounds originating from the places of intersection are released. The scanning sphere radius expands from 0 to 5 m in 1.5 s, and then after a 500 ms pause, it restarts from zero. The minimum (forced) time delay between two consecutive impacts on objects is set to 100 ms.

In the case of audio implementations with impact sounds and bubble sounds, a list of segmented objects is received and objects that intersect the sphere are used to generate the sounds. The strength of this methodology lies in its clear ability to convey distance and number of obstacles, while providing information on the closest obstacles first.

4 TEST SETUP AND PRELIMINARY RESULTS

4.1 Experimental environment

All sonification models presented in this paper were implemented in the Csound scripting language or in Pure Data (in the case of the bubble sound model), and integrated into the Sound of Vision processing Runtime. The Runtime is the core application component responsible for managing the entire processing pipeline, starting from the processing of the visual stream (including 3D processing, object classification) to the coordination of sensory substitution signals (for both audio and haptic output). The 3D processing pipeline stage is responsible for the decomposition of stream information into an abstract 3D scene composed of generic entities and raw depth-map information. A generic entity is defined to be a close representation of a real-world object with position in space (relative to the user), width and height (in meters), as well as a general category classification. After this step, a rudimentary classification is performed to identify special objects (e.g. walls, staircases, discontinuities in the ground), and the output is forwarded to the sensory substitution stage where it is processed by the selected audio/haptic representation. Two different input stream sources are accepted: real-world and virtual scenes. The real-world stream is obtained from depth sensors, a stereo RGB camera and an inertial measurement unit while virtual streams are emulated through the usage of virtual cameras inside the Virtual Training Environment (VTE) application, which is a standalone application that runs in a separate process. Stream information from the VTE is forwarded to the runtime through a TCP based local networking connection.

A prototype headgear with the required cameras, inertial measurement unit (IMU) and audio output was built to provide the input and output requirements described. Thus, in virtual scene testing, the headgear was used by testers to control the virtual camera orientation through head-movements captured by the IMU sensor while audio output feedback was received through a set of speakers attached onto the headgear frame.

Tests were carried out on three different virtual scene categories: scenes with a single randomly positioned generic object (referred to as the ‘random scenes’); scenes with 3-5 randomly positioned generic objects (referred to as the ‘complex scenes’); and finally, scenes with any random number of obstacles together with a target object (referred to as the ‘box scenes’). In the random and complex scenes, the goal was to identify the quantity (1 to 5), spatial location (5 possible distances, 5 possible directions, and 2 possible elevations) and width of objects (3 possible sizes) as quickly and precisely as possible. In the latter ‘box scenes’, the goal instead was to navigate to the target object while evading obstacles along the way. Figure 1 shows screenshots of the user interface for the Runtime, VTE and a sample boxes scene.

Not all sonification models were used with all temporal presentation schemes and all testing scenes. In particular, the granular synthesis model was not tested in the “recognition” type scenes (only in the “boxes” scene, which was designed to test navigation capabilities). Further, the granular synthesis model was not used together with the expanding sphere scheme, as it was the only model that was based on point clouds rather than segmented objects, and so determining the timing of sound events would have further complicated the model.

4.2 Results

Initial tests were carried out with seven participants (four visually impaired, and three normally sighted). Initial results (summarized in Tables I-III) show that different sonification models are fit for different purposes. Within the bar impact model, the expanding sphere paradigm is more suitable for correctly detecting the distance to obstacles, while left/right scanning is more suitable for direction. Elevation information is also clearly conveyed in the bar impact model. The bubble model exhibits clear pros and cons: while it is relatively poor in conveying distance and quantity information, it shows the best results in the perception of direction and width. On the other hand, the “objects as loudspeakers” paradigm is better suited to obstacle avoidance tasks. It also seems to be the case that models operating at less conceptual levels (such as the granular synthesis model, which derives its parameters from a depth / density map of object points) are much better suited to the latter, navigation-type tasks.

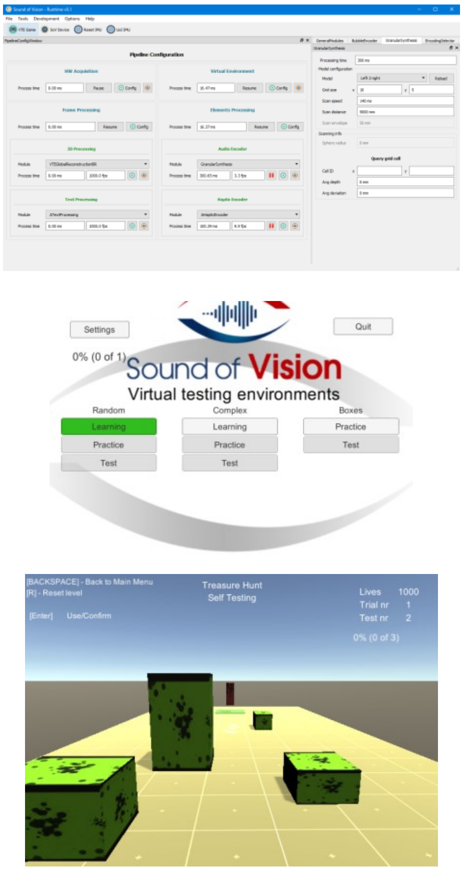

Figure 1. User interface of Runtime, Virtual Training Environment and a sample “boxes scene” with the navigation target shown at the back in red.

Model / Task	Width	Distance	Direction	Elevation
Bar impact (ES)	64.29/ 24.40	70.00/ 15.28	52.86/ 32.00	72.86/ 22.15
Bar impact (LR)	59.52/ 12.70	40.48/ 21.94	78.57/ 22.19	90.95/ 16.20
Bubble (ES)	74.29/ 22.25	25.71/ 22.25	80.00/ 23.09	68.57/ 19.52

Table I: Mean/standard deviation of success rates in random scene tests (abbreviations ES - expanding sphere, and LR - left to right scanning)

Model / Task	Quantity	Closest object	Widest object
Bar impact (ES)	78.39 / 9.11	70.83/ 24.12	49.41/ 15.37
Bar impact (LR)	82.92 / 16.99	69.17/ 13.84	54.17/ 27.00
Bubble (ES)	64.58 / 25.52	64.17/ 5.40	47.50/ 30.62

Table II: Mean/standard deviation of success rates in complex scene tests (abbreviations: ES - expanding sphere, and LR - left to right scanning)

Model / Task	No. collisions	Total time (s)
Bar impact (ES)	4	452
Bar impact (LR)	4	475
Granular (OaL)	2.7	409
Granular (LR)	3.1	315

Table III: Means of key indicators in box scene tests (abbreviations: OaL - objects as loudspeakers, LR - left to right scanning and ES - expanding sphere)

5 CONCLUSIONS AND FUTURE WORK

Tests are being continuously carried out in order to assess both the effectiveness and perceived effectiveness / comfort associated with each model and to achieve a tight feedback loop towards design teams. Also important to note is that our goal is to select models that are stable in their perception and effectiveness, even if those models are sub-optimal in a significant percentage of cases.

Preliminary results show that both the bar impact and bubble models are promising solutions (based on tests within the virtual training environment), and are soon expected to be tested in real-world scenarios where the ability to correctly assess the distance, direction and spatial extension of individual objects is important.

Two possible improvements have been identified for the bubble model. First, similarly to the other sound models, the distance to the object can be redundantly coded into the amplitude of the bubble sound in order to use absolute loudness as a further distance cue to natural sounding events and to convey a stronger sense of urgency (louder sound) for closer objects. Second, given the limited number of parameters of the physical sound model, the rise factor parameter can be used to signal selected "dangerous" objects. In this case, elevation information needs to be conveyed exclusively through spatial sound.

Results also suggest that the current granular synthesis model (combined with left-to-right scanning) is not a suitable solution, due to some reported discomfort and a relatively large degree of variation in objective effectiveness (possibly as a result of the confounding nature of proximity and elevation, which both contribute to perceived loudness in the model). At the same time, further investigation of the model would be useful in conjunction with the objects as loudspeakers presentation scheme (at least in navigation as opposed to recognition tasks).

Acknowledgements

This project received funding from the EU Horizon 2020 research and innovation programme under grant agreement No 643636 "Sound of Vision". Thanks is also due to the SoV testing team for the design of the testing procedure used.

References

[1] F. Avanzini and D. Rocchesso, "Controlling material properties in physical models of

sounding objects," in *Proc. Int. Computer Music Conf.* (ICMC'01), Cuba, 2001.

[2] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "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), pp. 904–916, 2001.

[3] S. Bilbao, Numerical Sound Synthesis: Finite Difference Schemes and Simulation in Musical Acoustics, *Wiley*, 2009.

[4] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, 1996.

[5] M. Bujacz *et al.*. Sound of Vision - Spatial audio output and sonification approaches. In *Computers Helping People with Special Needs (ICCHP)*, pp. 202-209, Springer Int. Publishing, 2016.

[6] A. Dobrucki, P. Plaskota, P. Pruchnicki, M. Pec, M. Bujacz, and P. Strumillo. "Measurement system for personalized head-related transfer functions and its verification by virtual source localization trials with visually impaired and sighted individuals." *J. Audio Eng. Society*, 58(9), pp. 724-738, 2010.

[7] A. Csapo, G. Wersenyi. "Overview of auditory representations in human-machine interfaces", *ACM Computing Surveys*, vol. 46, no. 2, 2014

[8] S. Delle Monache, P. Polotti, and D. Rocchesso, "A toolkit for explorations in sonic interaction design," in *Proc. 5th Audio Mostly Conference (AM '10)*, USA 2010.

[9] D. Gabor, "Acoustical Quanta and the Theory of Hearing", *Nature* 159[1044]: pp. 591-594

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

[11] W. W. Gaver. "Auditory Icons: Using sound in computer interfaces". *Human-Computer Interactions*, 2, 2, 167-177, 1986.

[12] W. W. Gaver, "Auditory Interfaces". In, *Handbook of Human-Comp. Int.*, 4, 67-94. , 1997.

[13] M. Geronazzo, S. Spagnol, and F. Avanzini, "Mixed structural modeling of head-related transfer functions for customized binaural audio delivery," in *Proc. 18th Int. Conf. on Digital Signal Processing*, Greece 2013.

[14] A. Kristjánsson *et al.*, "Designing sensory-substitution devices: Principles, pitfalls and potential," *Restor. Neurol. Neurosci*. 34(5), pp. 769–787, October 2016.

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

[16] C. Roads, "Introduction to Granular Synthesis", *Computer Music Journal*, vol. 12, no. 2, pp. 11-13, 1988

[17] S. Spagnol, M. Geronazzo, and F. Avanzini, "Structural modeling of pinna-related transfer functions," in *Proc. 7th Int. Conf. on Sound and Music Computing (SMC 2010)*, pp. 422–428, Spain, 2010.

[18] S. Spagnol *et al.*, "Model-based obstacle sonification for the navigation of visually impaired persons," in *Proc. 19th Int. Conf. on Digital Audio Effects*, pp. 309–316, Czech Republic, 2016.

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

[20] K. van den Doel, "Physically-based models for liquid sounds," *ACM Trans. on Applied Perception* 2(4), pp. 534–546, 2005.

[21] I. Xenakis, "Formalized Music: Thought and Mathematics in Composition". *Bloomington and London: Indiana University Press*, 1971