MODEL-BASED OBSTACLE SONIFICATION FOR THE NAVIGATION OF VISUALLY IMPAIRED PERSONS

Simone Spagnol, Omar I. Johannesson, Arni Kristjansson, Runar Unnthorsson

University of Iceland,
Reykjavík, Iceland
{spagnols,omarjo,ak,runson}@hi.is

Charalampos Saitis, Kyriaki Kalimeri

ISI Foundation,
Turin, Italy
{charalampos.saitis,kyriaki.kalimeri}@isi.it

Michał Bujacz

Lodz University of Technology,
Łódź, Poland
michal.bujacz@p.lodz.pl

Alin Moldoveanu

University Politehnica of Bucharest
Bucharest, Romania
alin.moldoveanu@cs.pub.ro

ABSTRACT

This paper proposes a sonification model for encoding visual 3D information into sounds, inspired by the impact properties of the objects encountered during blind navigation. The proposed model is compared against two sonification models developed for orientation and mobility, chosen based on their common technical requirements. An extensive validation of the proposed model is reported; five legally blind and five normally sighted participants evaluated the proposed model as compared to the two competitive models on a simplified experimental navigation scenario. The evaluation addressed not only the accuracy of the responses in terms of psychophysical measurements but also the cognitive load and emotional stress of the participants by means of biophysiological signals and evaluation questionnaires. Results show that the proposed impact sound model adequately conveys the relevant information to the participants with low cognitive load, following a short training session.

1. INTRODUCTION

In audio-based software applications, such as auditory displays and audio games, sonification is used to represent various actions, objects or situations in order to virtually describe scenes. *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" [1].

Sonification is also used in health care, for instance in motor rehabilitation systems [2], electronic travel aids (ETAs, i.e., devices which aid in independent mobility through obstacle detection or help in orientation and navigation) [3], and other assistive technologies for visually impaired persons (VIPs). Most of these systems are still in their infancy and mostly still at a prototype stage. Furthermore, available commercial products have limited functionalities, small scientific/technological value and high cost [3].

Available ETAs 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 of the scenes, acquired through camera technologies. The auditory outputs of such systems range from simple binary alerts indicating the presence of an obstacle in the range of a sensor, to complex sound patterns carrying almost as much information as a graphical image [4]. Finding the most suitable accuracy/simplicity trade-off in order to provide valuable information about the environment surrounding the user through sound is therefore a pivotal and challenging task.

This study aims to explore a novel scheme for translating 3D representations of a scene or an environment, represented as a list of objects with properties, into auditory feedback. The remainder of the paper is organized as follows. Section 2 introduces the sound model as well as two alternative models inspired by previous literature. Section 3 describes an experiment targeted at comparing the performance of the three models in a navigation task, through both psychophysical and psychophysiological measurements. Section 4 reports the results of the experiment, and Section 5 concludes the paper.

2. MODEL-BASED OBSTACLE SONIFICATION

Different sonification approaches for representing visual scenes to blind users have previously been studied. The most common natural mappings between object and sound properties are related to the spatial position of the object; the most recurring are

- azimuth → stereo panning / Head-Related Transfer Function (HRTF) filtering [5, 6, 7];
- elevation → HRTF filtering [6] / pitch [8];
- distance → amplitude [5, 9] / pitch [5, 9].

This Section provides details on a sonification model designed by the authors with the help of blind volunteers and specialists in training and rehabilitation of VIPs. Mappings within the model were both inspired by the previous literature shown above and original design. Parameter tuning was refined following a preliminary investigation using psychophysical evaluation methods only [10].

2.1. Sonification through impact sounds

The model we propose treats each object in the frontal hemisphere of the user as an independent virtual sound source that continuously emits impact sounds, as if the VIP was hitting it with a white cane. The pitch and timbre of the sound resulting from the impact are considered dependent on the object's width and category. The distance between object and user is coded into loudness: the closer the object, the higher the sound level. Furthermore, each

sound is spatialized in accordance with the direction of the object with respect to the user.

Single impact sounds are generated through a physical model of non-linear impact between two modal objects. This model is part of a number of sound models included in the Sound Design Toolkit (SDT),[1] an open-source (GPLv2) software package suitable for research and education in Sonic Interaction Design [11]. The SDT consists of a library of physics-based sound synthesis algorithms, available as externals and patches for Max and Pure Data.[2] The Pure Data version was used in the development of this model.

The physical model receives as input parameters related to the striking object (modal object 1) and the struck object (modal object 2), as well as the interaction between the two. The most relevant fixed parameters are strike velocity, set to 1.85 m/s, and striker mass, set to 0.6 kg. These were considered as reasonable parameters for a long white cane and the act of striking with it. Parameters of the struck object, i.e., the object that needs to be sonically represented, change with respect to the width and category of the object itself.

In particular, width is directly mapped to the frequency f of the single mode of the struck object. In order to maximize the available frequency range, this was chosen to vary from values as low as 50 Hz (very wide objects such as walls) to 4 kHz (20-cm narrow objects) according to the following mapping,

$$f = \frac{840}{w} \quad [Hz] \tag{1}$$

where w is the actual width of the object in meters.

Different categories of objects are on the other hand represented by different decay times of the frequency mode. Categorization of objects may follow different rules, e.g. be based on object material (with rubber, wood, glass and steel having increasing decay times [12]) or object type (simple objects, walls, poles or trees, holes or ponds, and so on). Having defined category $C = 1, 2, 3, \ldots$, the mapping to the corresponding acoustic parameter, i.e. decay time t_d, is

$$t_d = 0.02C \quad [s] \tag{2}$$

heuristically set in order to enable association to impacts on different materials [12]. Default parameters were used for the test reported in this paper, with category 1 assigned to wall sounds and category 5 assigned to wall edges (replacing two adjacent wall sounds).

The absolute distance r between the subject and the object is also considered as a parameter. Assuming all obstacles to be sonified lying further than 1 m (closer objects are in the reach of the white cane), thus in the subject's acoustic far field [13], this is directly mapped into the amplitude of the sound by following the classic $1/r$ pressure attenuation law [14]. The overall number n of objects present in the scene influences the repetition rate of the impact sound instead: the period T between two consecutive impacts on the same object is set to

$$T = 0.2(n - 1) \quad [s]. \tag{3}$$

The point associated to the object is either the estimated barycenter in the case of small objects, or the intersection between the closest surface and its normal vector crossing the observer in the case of bigger objects, such as walls.

Last but not least, the direction of the object with respect to the observer taken in angular coordinates (azimuth, elevation) is directly mapped to the corresponding parameters of a generic HRTF filter provided through the `earplug~` Pure Data binaural synthesis external. In particular, 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 [15].[3] This way, the sound is spatialized along the actual direction of the object. It has to be highlighted that spatialization is non-individual; however, models for HRTF individualization [16, 17] or individual HRTFs themselves can be integrated (at an additional measurement cost) if higher spatial accuracy is needed [18].

There were two reasons for choosing impact sounds to convey information about objects. 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. Second, the peculiar pattern of impact sounds, whose rich frequency content and short duration of the attack phase allow for improved sound localization on the horizontal plane [19]. Furthermore, choices about the mappings between object and sound properties were either adopted from previous literature (distance and direction) or based on the nature of the impact model. Actually, the association of higher pitches to smaller objects and different decay times to different categories, e.g. materials, has physical ground [12].

The model was implemented as three Pure Data patches. Both static scenes and simple dynamic scenes with a fixed number of objects are supported. The main Pure Data patch receives as input a text file containing one row per object present in the scene. Each row includes information about the object ID, azimuth angle (between -90 and 90 degrees [15]), elevation angle (between -40 and 90 degrees [15]), distance (above 1 m), width (above 20 cm), object category ($1, 2, 3, \ldots$), and mode (static $= 0$, dynamic $= 1$), separated by spaces.

At the beginning, sources are ordered by increasing azimuth, left to right. In order to avoid simultaneous impacts, the first impact on a given object is played 200 ms after the impact on the object on its immediate left. In the case of a dynamic scene, impacts corresponding to an object stop as soon as the object is behind the listener (i.e., outside the $[-90, 90]$ degree azimuth range).

2.2. Alternative sonification approaches

In the round of testing reported in this paper, the proposed model is compared against two other competing models developed in previous literature in order to solve the same problem. These two alternative approaches are now briefly described.

2.2.1. Depth scanning

The depth scanning model is a sonification method used in Bujacz *et al.* [20]. The main inspiration for the model was the fact that blind persons, especially those blind from birth, have a path-based perception of their environment [21]. The core concept of the method is a virtual scanning plane, i.e., a surface parallel to the observer's frontal plane that moves away from him/her through the scene. As the surface intersects scene elements, sounds originating

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

[2] `https://puredata.info/`

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

from the points of intersection are released. The scanning surface moves for 5 m in 1.5 s, then after a 0.5 s pause it restarts from the observer. This was the default speed chosen by the majority of the blind participants in previous prototype trials [20]. However, in the experiment reported in this paper the scanning surface was slightly sped up to fit 3 cycles into 5-second test samples. Furthermore, reference "tick" sounds are played each time the scanning plane moves 1 m away.

Sounds are designed to naturally correspond to object parameters. Distance, as the most important parameter, is encoded redundantly into the temporal delay inside each cycle as well as into the loudness and pitch of the sound. For instance, if a distant object appears later in a scanning cycle, its sound will be less loud and have lower pitch. The location of an object is encoded via HRTFs and its size through sound duration. The sound coder uses audio files pre-generated with a Microsoft General MIDI calliope synthesizer (no. 83) modulated with 5% noise (14 dB SNR), as previous trials showed that the addition of noise improves spatial localization of a sound [22]. The sounds were stored in collections of 5-s wave files of full tones from the diatonic scale (octaves 2 to 4). Sounds were spatially filtered using the MIT KEMAR generic HRTFs and modulated with a simple ADSR envelope.

2.2.2. Horizontal sweep

The horizontal sweep approach was used in previous sonification studies (e.g. Navbelt [5]), and is sometimes referred to as the "piano scan". It basically translates the distance to pitch in several directions from the observer. The sonification approach is very similar to the one previously described for depth scanning with the main difference being that of the scanning plane; instead of moving away from the frontal plane, it swings left to right around a vertical axis passing through the observer. The scan sweeps from -45 to 45 degrees in 1.5 s. Reference "tick" sounds are played each time the scanning plane moves by 15 degrees.

This model generates sounds from scratch using a simple Moog synthesizer and an ADSR envelope. Pitches are selected from the middle three octaves of the pentatonic scale. A difference with respect to the depth scan approach is that instead of smoothly moving sound sources along the intersection of the sweeping plane and walls, the scene was divided into discrete regions 15 degrees wide (according to the previously set reference ticks), and for each region a sound was produced corresponding to the nearest object.

3. MATERIALS AND METHODS

An experiment was designed where the above described sonification approaches were compared using methods from the fields of behavioural psychology and psychophysiology, namely response time and accuracy, electroencephalography (EEG), and monitoring of electrodermal activity (EDA). The goal was to explore various alternatives in rendering basic 3D visual scenes through sound signals to be delivered to VIPs, through assessing both functionality (psychophysics) and cognitive performance (psychophysiology). This study was accepted by the National Bioethical Committee of Iceland, with reference number VSN-15-107.

3.1. Participants

Five VIPs and five sighted students from the University of Iceland participated in the experiment (6 female; average age = 34 yrs, range = $21 - 52$ yrs) on voluntary basis. One VIP was fully blind, two had vision less than 5%, and two had vision between 5% and 10%. Three of them were congenitally or early blind (first $2-3$ yrs of life) and two had become blind later in life (generally after the age of 3). All participants spoke English fluently and reported having no hearing impairment as well as no general health issues. Two VIPs mentioned having some experience, one reported substantial experience, and two said that they are very experienced with IT technology. All participants gave free and informed consent.

3.2. Psychophysiological approach

Electrodermal activity is a well-known indicator of physiological arousal and stress activation [23]. EDA is more sensitive to emotion related variations in arousal as opposed to physical stressors, which can be better reflected in measurements of cardiovascular activity such as heart rate. Electroencephalography, on the other hand, can provide neurophysiological markers of cognitive and emotional processes induced by stress and indicated by changes in brain rhythmic activity [24]. Taking advantage of their inherent and complementary properties, EEG and EDA signals were collected and analysed concurrently with the more traditional behavioural measures of response time and accuracy.

A measurement of EDA is characterized by two types of behaviour: short-lasting phasic responses (which can be thought of as rapidly changing peaks) and a long-term tonic level (which can be thought of as the underlying slow-changing level in the absence of phasic activity) [23]. Phasic responses are primarily elicited by specific external stimuli, and are typically observed superposed in states of high arousal or short interstimulus interval paradigms such as those employed in cognitive research.

EDA was registered with the Empatica E4 wristband [25], which measures skin conductance through two ventral (inner) wrist electrodes ($f_s = 4$ Hz). Signals were analysed with Ledalab, a Matlab-based toolbox.[4] Ledalab implements a signal decomposition method based on standard deconvolution, which results in one single continuous measure of phasic activity. Time-integration over a specified window after the stimulus onset yields a simple and unbiased (i.e., avoiding biases due to superposing peaks) indicator of phasic EDA, namely integrated skin conductance response (ISCR) [26]. ISCR can be thought of as the cumulative phasic activity within the specified response time period. Our hypothesis was that a pleasant, easy to understand, and less stressful sonification mapping will generally elicit lower phasic activity as indexed by ISCR.

Brain activity is characterized by rhythmic patterns (waves) across distinct frequency bands, the definition of which can vary among studies. Here we analysed EEG in five bands, namely theta (4–7 Hz), alpha-1 (7.5–10 Hz), alpha-2 (10–12.5 Hz), beta (13–30 Hz), and gamma (30–60 Hz). Beta activity is associated with psychological and physical stress, whereas theta and alpha-1 frequencies reflect response inhibition and attentional demands such as phasic alertness [27]. Alpha-2 is related to task performance in terms of speed, relevance, and difficulty [24]. Gamma waves are involved in more complex cognitive functions such as multimodal processing or object representation [28].

EEG was recorded using the Emotiv EPOC+, a wireless headset with 14 passive electrodes (channels) registering over the 10-20 system locations AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and FC4 (sampling rate $f_s = 128$ Hz) [29]. For each

[4] `http://www.ledalab.de`

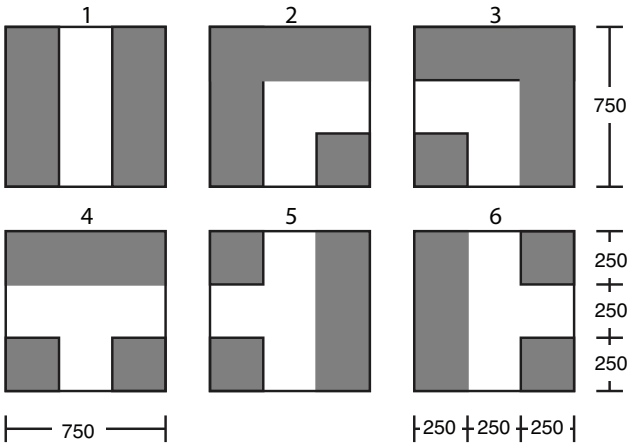

Figure 1: The six modules used in the experiment. The participant always entered each model from the bottom. White areas represent the virtual path and grey areas represent obstacles (delimited by walls). All measures are in cm.

channel we studied the relative spectral power in each of the bands described above over a specified time window following the stimulus onset using the Fourier transform. The total power across all bands (4–60 Hz) was also examined. Analyses were focused on a subset of 6 channels, namely AF3, F7, T7, T8, F8, and FC4, which are often considered suitable enough to monitor brain activity under emotional stress [30]. Our hypothesis was that a less mentally demanding and stressful sonification model will generally involve lower power across the whole EEG spectrum, smaller theta activity and larger alpha-2 power.

3.3. Setup and procedure

Participants of the virtual navigation task were instructed to use their dominant hand to respond using the arrow keys of a keyboard: left arrow for turning left, up arrow for walking straight and right arrow for turning right. The down arrow was removed from the keyboard during the experiment to help avoid potential confusion. All participants were blindfolded, and sound stimuli were delivered through a pair of in-ear headphones.[5] Information and feedback concerning the progress of the experiment were also presented to the participant through prerecorded audio files.

A virtual walk was designed, comprising 6 different scenes (modules) each representing a different configuration of a free path between walls. The 6 configurations can be seen in Fig. 1: the gray blocks depict walls while the walkable path is in white. Auditory representations of each configuration were created from each of the tested sonification models (Model 1: impact model; Model 2: horizontal sweep; Model 3: depth scanning).

Initially there was a training phase, where in order for the subject to comprehend the rationale of each sound model, a 3D representation of the 6 modules was created using Lego blocks. While touching each block, the participant listened to the corresponding stimulus along with prerecorded explanations of the task. This procedure was repeated two times for each of the sound models.

Subsequently, the physiological sensors were placed. Participants were asked to find a comfortable position and to avoid any unnecessary movement. EDA was recorded from the non-dominant hand (wrist) of the participants to minimize motion artifacts largely due to pressing response buttons [23]. EEG was recorded continuously from the 14 scalp electrodes of the EPOC headset. Signals were transmitted from the headset via a USB wireless receiver to proprietary Emotiv software running on a laptop.

The next three phases were repeated once for each of the three tested sound models, whose order during the test was randomized in order to avoid any bias.

3.3.1. Training

Upon setting up the sensors, a short training session started wherein each module was presented to the participant four times in random order. Responses were recorded but only used to provide feedback to the participant after each response and to calculate their accuracy. If less than 75% of the given responses were correct, the training session was repeated but never more than two times. Right after the training session, the participant was asked to relax completely for 300 s in order to record spontaneous resting state physiological activity.

3.3.2. Testing

During the virtual walk, the participant always entered each module from the bottom (see Fig. 1). The virtual walking speed was chosen to be 1 m/sec. For each module the participant had to make a decision no later than 5 m (5 s) after entering the module: whether to turn (and into which direction) or to continue straight ahead. Participants were instructed to respond as fast and as accurately as possible. Each module was presented 15 times in random order. One full virtual walk lasted approximately 6 to 10 minutes.

If the participant did not respond within the time limit (i.e., after 5 s), or if his decision was incorrect, the virtual walk was stopped and a short sound indicating an error was played. After 0.5 seconds the virtual walk would start again. In the case of registering a correct response, the model stimulus was stopped and the participant instantly moved to the next module where the same procedure was repeated.

Upon completion of testing a model, participants were asked to relax completely for 120 s while their spontaneous physiological activity was being registered.

3.3.3. Questionnaire

As soon as the resting period ended, the participant was asked to evaluate the model on five 5-point Likert scales:

- Q1 - I found the sounds pleasant to the ear.
- Q2 - I could imagine the sounds originating from the environment (as opposed to originating inside my head).
- Q3 - I found it easy to understand what each sound represents.
- Q4 - I found the task stressful (the sounds were too fast to understand).
- Q5 - I think that, with sufficient training, I would understand what each sound represents at even faster rates.

Subjects were also asked to freely comment on the functionality and pleasantness of the sound stimuli. Verbal responses were recorded.

[5] `https://earhero.com`

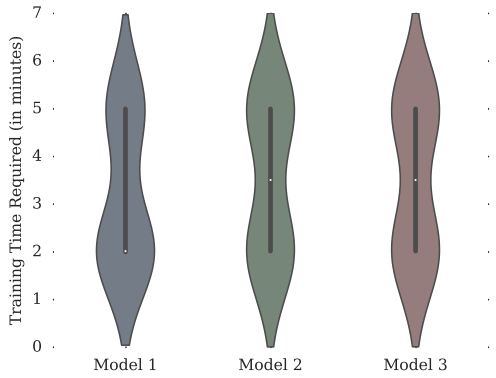

Figure 2: Violin plots of the probability density distribution of learning time required by subjects. The white dot represents the median value.

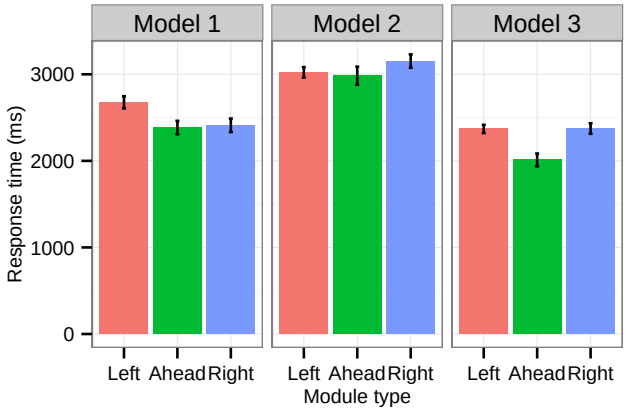

Figure 3: Average RTs by model and response type. Error bars represent the within subjects standard error.

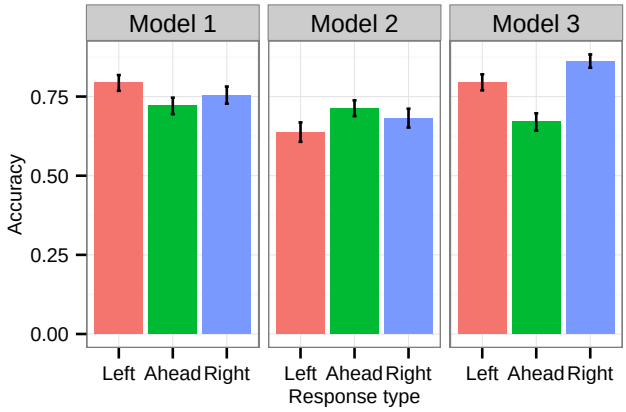

Figure 4: Average accuracy by model and response type. Error bars represent the within subjects standard error.

Once all models were tested, the participant was asked to freely respond to the following question: "Which of the three sonification approaches did you prefer? Can you explain why?". Verbal responses were again recorded. Finally, the EPOC and E4 sensors were detached from the participant. During the experimental session, the experimenter sat outside of the room and monitored the stimulus presentation and the recorded physiological data. The experiment lasted approximately 90 minutes in total.

4. RESULTS

4.1. Model learnability

As described in the previous Section, the experiment involved a short training session. Using response times from the training sessions, we looked at the time required from all participants to "understand" each model (i.e., how the respective sounds mapped to the different modules of the virtual walk task).

Figure 2 depicts the probability density of the learning interval required for each model. From this representation, Model 1 outperforms the other two, since it is the only one that required a median learning interval equal to two minutes. Note that since the training section was not repeated more than two times, the difference between Model 1 and Models 2 and 3 is relevant, even if small.

4.2. Response time and accuracy

The average response time and accuracy were computed for each model and compared using repeated measures ANOVA. Notice from Fig. 1 that modules 1 to 3 had only one correct response out of three (left, ahead, right), while modules 4 to 6 had two correct responses out of three. This means that random responses would lead to an average accuracy of 0.33 and 0.67 in the long run, respectively. Therefore, the average random accuracy in the experiment is 0.5. A response is considered correct only when given within the maximum response time of 5 s.

The average response time (RT) was 2441 ms (SD = 991 ms) for the visually impaired group and 2780 ms (SD = 1059 ms) for the sighted group. The difference between these groups was not significant [$t(7.57) = 1.11$, $p = 0.303$]. The average response time divided by model and response type (left, ahead, right) is reported in Fig. 3. A two-way repeated measures ANOVA with model and response type as factors revealed a significant main effect of response type [$F(2, 18) = 6.08$, $p = 0.010$] but not of model [$F(2, 18) = 2.65$, $p = 0.098$] and significant interaction between models and response type [$F(4, 36) = 2.96$, $p = 0.033$].

Average accuracy was 0.73 (SD = 0.44) and is significantly different from random [paired-t(9) = 6.44, $p < 0.001$]. The average accuracy was 0.73 (SD = 0.45) for the visually impaired group and 0.74 (SD = 0.44) for the sighted group. This very small difference was not significant [$t(4) = 0.51$, $p = 0.638$]. The average accuracy divided by model and response type is reported in Fig. 4. A two-way repeated measures ANOVA with model and response type as factors revealed a significant main effect of response type [$F(2, 18) = 6.71$, $p = 0.007$] but not of model [$F(2, 18) = 2.24$, $p = 0.135$], and the interaction was not significant [$F(4, 36) = 0.87$, $p = 0.491$]. Accuracy was therefore comparable across models.

4.3. Phasic electrodermal response

Prior to analysis, skin conductance data obtained from the E4 EDA sensor were filtered with a first-order Butterworth low-pass filter

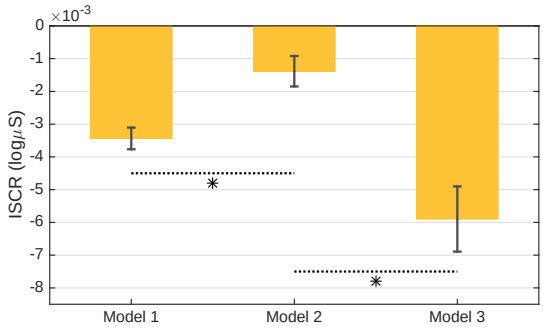

Figure 5: Average cumulative phasic activity for each of the tested models. Error bars represent the standard error of the mean. Star signs (*) indicate statistically significant differences.

using a cutoff frequency of 1 Hz to remove steep peaks stemming from artifacts such as pressure exerted on the electrodes [23]. The filtered time series were subsequently analyzed with Ledalab using the continuous decomposition method (see Section 3.2). As sound stimuli were presented as soon as the response to the previous module was registered (i.e., variable interstimulus intervals), a variable response window was considered, starting at 1 s after stimulus onset and ending at 4 s after stimulus offset. Integrals of the continuous phasic driver within a specified response time period were normalized by means of dividing by the duration of the respective window. To reduce inter-individual variation prior to averaging, means were subtracted from the trial-by-trial ISCR values. The resulted data were further transformed using the formula $y = log(1 + x)$ to improve distributional characteristics.

Figure 5 depicts the average ISCR for each model computed across all participants and all stimuli. It can be immediately observed that phasic electrodermal activity was substantially higher in Model 2 and lower in Model 3. To test whether these differences were statistically significant, a repeated measures ANOVA with models as factor was run. This analysis revealed a significant difference in physiological arousal between the three sonification alternatives [$F(2) = 10.6$, $p = 0.02$, using the Greenhouse-Geisser correction for sphericity]. Pairwise comparison of the means using Bonferroni post hoc tests showed that Model 2 is significantly different from Models 1 and 3 [$p = 0.04$ and $p = 0.05$, respectively], whereas the observed difference between Models 1 and 3 is marginally not significant [$p = 0.06$].

4.4. EEG power spectra

The Emotiv EPOC+ system involves a number of internal signal conditioning steps. Analogue signals are first high-pass filtered with a 0.16 Hz cut-off, pre-amplified, low-pass filtered with a 83 Hz cut-off, and sampled at 2048 Hz. Digital signals are then notch-filtered at 50/60 Hz and down-sampled to 128 Hz prior to transmission. Prior to analysis, the EEG data obtained from the headset was baseline-normalized by subtracting for each participant and for each channel the mean of the resting state registrations.

As described in the previous section, sound stimuli were presented as soon as the response to the previous module was registered. Considering the shortest interstimulus interval (3.4 s), EEG epochs lasting 3 s after stimulus onset were extracted for each model-module condition, resulting in a total of 2700 epochs per

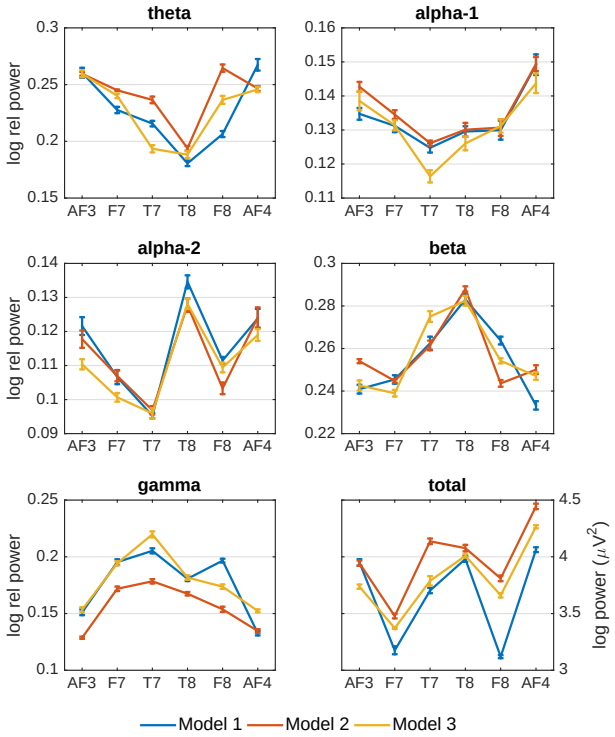

Figure 6: Mean relative and total band power for each of the tested models in 6 frontal electrode positions. Error bars represent the standard error of the mean.

EEG channel. A Hann window was applied to each epoch to minimize spectral leakage, and then the Fourier transform of the windowed data was used to calculate power spectral density estimates in the theta (4–7 Hz), alpha-1 (7.5–10 Hz), alpha-2 (10–12.5 Hz), beta (13–30 Hz), and gamma (30–60 Hz) bands as well as across the total 4–60 Hz range. Individual band power estimates were normalized by means of dividing by the across-band power. Before averaging, a logarithmic transformation $y = log(1 + x)$ of single-trial values was applied to improve their distributional characteristics.

Figure 6 shows the average band power in the AF3, F7, T7, T8, F8, and FC4 channels for each model, calculated across all participants and modules. A first look at the different plots suggests that Model 1 resulted in better cognitive performance during the experimental task than Models 2 and 3. Gamma activity, related to information representation and processing, was particularly low for Model 2, which had the largest total power. To test whether model differences were statistically significant, a repeated measures ANOVA with model as the between-subjects factor and electrode location as the within-subjects factor was run for each frequency range. Where appropriate, p-values were corrected by means of the Greenhouse-Geisser method. Bonferroni post-hoc tests were used for pairwise comparison of means.

There was a significant effect of model on total power [$F(10) = 61.14, p \ll .001$] and on relative power in each band [theta, alpha-2, beta, gamma: $F(10) \geq 6.39, p \ll .001$; alpha-1: $F(10) = 3.31, p = .0013$]. Total power for Model 1 was significantly

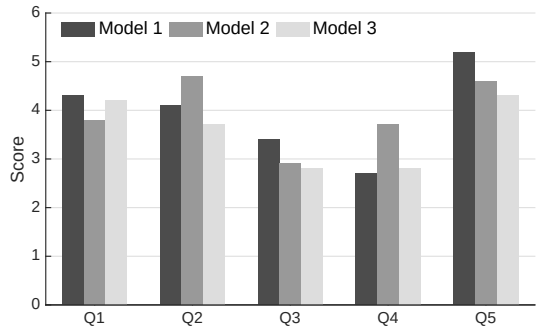

Figure 7: Average scores across the five usability scales for each of the tested models.

lower than for Models 2 and 3, and for Model 2 it was significantly higher than for Model 3 [$p \ll .001$]. This suggests that participants were more cognitively engaged when responding to Model 2 stimuli. However, gamma activity was significantly lower for Model 2 than for the other two models [$p \ll .001$] and no significant difference between Models 1 and 3 was revealed [$p = 1$]. This would imply that the larger total power for Model 2 reflected its reduced ability to convey the relevant information [28]. Model 2 further resulted in higher theta [significant effect, $p < .001$], alpha-1 [not significant, $p \geq .21$], and beta [not significant, $p \geq .62$] power than the other two models, suggesting higher response inhibition, attentional demands, and stress. Model 1 had the smallest theta [significantly so than Model 2 but not Model 3] and largest alpha-2 power [no significant differences between the three models, $p \geq .1$], suggesting better cognitive performance [24].

Analyses were repeated for all 14 channels of the Emotiv headset. The same effects and differences were observed, thus confirming that the selected 6 frontal electrode locations were suitable and sufficient to assess stress-related cognitive performance [30].

4.5. User experience questionnaire

The majority of the participants (8 out of 10) preferred Model 1 over the other two models referring to it as the easiest to use. The sounds in Model 1 were mainly described as pleasant and none of the subjects seemed to have strong negative opinions about their unpleasantness. Model 2 appeared to be the least favored one as participants reported having trouble understanding what the sounds were intended to convey. A few subjects thought the sounds were too similar to each other and had difficulty telling them apart, with some even describing them as confusing. Despite this, the majority of subjects found the model's sounds to be pleasant. A few participants had issues understanding some of the sounds in Model 3 while others described it as functional and easy to use. The sounds in this model were reported as the most annoying and/or irritating although some found them pleasant.

Complementary to the analysis of the verbal comments, the scores of each model in each of the 5 Likert-scales previously described in Section 3.3.3 were computed. Figure 7 shows that Model 1 was perceived as slightly more pleasant (Q1), easier to understand (Q3), less stressful (Q4), and easier to learn (Q5). Model 2, on the other hand, was characterized as the most natural sounding (Q2).

5. DISCUSSION AND CONCLUSIONS

Psychophysical results show overall that the proposed impact sound model (Model 1) adequately conveys the relevant information to the participants, who are able to use this information to guide their virtual walk. Although no significant differences in RTs and accuracy were found in the comparison against two other models, Model 2 was found to perform slightly worse, and Model 3 slightly better than Model 1. These results are consistent with the event-related analysis of phasic EDA: Model 2 appears to elicit the highest stress-related physiological arousal compared to Models 1 and 3, whereas Model 3 is shown to be marginally less stressful than Model 1.

This last finding, however, appears to disagree with the perceptions emerging from the verbal comments of the participants, where Model 1 came out as the most preferred and was ranked slightly higher than Model 3 in terms of functionality and pleasantness. Further analysis is necessary to examine the origins of this discrepancy. Furthermore, Model 1 had the best cognitive performance compared to the other two models. Finally, Model 1 was the easiest to learn among the three models.

The above results suggest that the adopted sonification approach may lead to improved results if adequate modifications, either to the mapping schemes or to the chosen sound stimuli, are performed. Ongoing work in this direction involves attempting to improve the model by combining impact sounds with the depth scan paradigm. Future work related to the model presented in this paper will therefore explore variations of the basic sound components used for encoding (impact sounds) and the scanning paradigm, as well as combining discrete encodings with continuous encodings for various object categories (e.g. walls, stairs, doors). We further plan to test the functionality and cognitive performance of the model in indoor and outdoor navigation scenarios using similar biosignal monitoring and analysis methods [31].

6. ACKNOWLEDGMENTS

The authors wish to thank the participants for their collaboration as well as the administration and O&M instructors at the National Institute for the Blind, Visually Impaired, and Deafblind in Iceland for their valuable input and generous assistance. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 643636.[6]

7. REFERENCES

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

[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., chapter 12, pp. 273–283. MIT Press, Cambridge, MA, USA, March 2013.

[3] D. Dakopoulos and N. G. Bourbakis, "Wearable obstacle avoidance electronic travel aids for blind: A survey," *IEEE*

[6] http://www.soundofvision.net/

Trans. Syst. Man Cybern., vol. 40, no. 1, pp. 25–35, January 2010.

[4] Á. Csapó, G. Wersényi, H. Nagy, and T. Stockman, "A survey of assistive technologies and applications for blind users on mobile platforms: A review and foundation for research," *J. Multimod. User Interf.*, vol. 9, no. 4, pp. 275–286, December 2015.

[5] S. Shoval, J. Borenstein, and Y. Koren, "Auditory guidance with the Navbelt - A computerized travel aid for the blind," *IEEE Trans. Syst. Man Cybern.*, vol. 28, no. 3, pp. 459–467, August 1998.

[6] J. L. González-Mora, A. Rodríguez-Hernández, L. F. Rodríguez-Ramos, L. Díaz-Saco, and N. Sosa, "Development of a new space perception system for blind people, based on the creation of a virtual acoustic space," in *Engineering Applications of Bio-Inspired Artificial Neural Networks*, vol. 1607 of *Lecture Notes in Computer Science*, pp. 321–330. Springer Berlin Heidelberg, 1999.

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

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

[9] E. Milios, B. Kapralos, A. Kopinska, and S. Stergiopoulos, "Sonification of range information for 3-D space perception," *IEEE Trans. Neural Syst. Rehab. Eng.*, vol. 11, no. 4, pp. 416–421, December 2003.

[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)*. Springer-Verlag, Berlin, Germany, July 2016.

[11] S. Delle Monache, P. Polotti, and D. Rocchesso, "A toolkit for explorations in sonic interaction design," in *Proc. 5th Audio Mostly Conference (AM '10)*, New York, NY, USA, September 2010, number 1, ACM.

[12] F. Avanzini and D. Rocchesso, "Controlling material properties in physical models of sounding objects," in *Proc. Int. Computer Music Conf. (ICMC'01)*, La Habana, Cuba, September 2001.

[13] S. Spagnol, "On distance dependence of pinna spectral patterns in head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 137, no. 1, pp. EL58–EL64, January 2015.

[14] D. H. Ashmead, D. LeRoy, and R. D. Odom, "Perception of the relative distances of nearby sound sources," *Percept. Psychophys.*, vol. 47, no. 4, pp. 326–331, April 1990.

[15] 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.

[16] S. Spagnol, M. Geronazzo, D. Rocchesso, and F. Avanzini, "Synthetic individual binaural audio delivery by pinna image processing," *Int. J. Pervasive Comput. Comm.*, vol. 10, no. 3, pp. 239–254, July 2014.

[17] S. Spagnol and F. Avanzini, "Frequency estimation of the first pinna notch in head-related transfer functions with a linear anthropometric model," in *Proc. 18th Int. Conf. Digital Audio Effects (DAFx-15)*, Trondheim, Norway, December 2015, pp. 231–236.

[18] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?," *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469, June 1996.

[19] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, Cambridge, MA, USA, 2nd edition, October 1996.

[20] M. Bujacz, P. Skulimowski, and P. Strumiłło, "Naviton - A prototype mobility aid for auditory presentation of three-dimensional scenes to the visually impaired," *J. Audio Eng. Soc.*, vol. 60, no. 9, pp. 696–708, September 2012.

[21] M. Brambring, "Language and geographic orientation for the blind," in *Speech, Place, and Action: Studies in Deixis and Related Topics*, R. J. Jarvella and W. Klein, Eds., pp. 203–218. John Wiley & Sons, Chichester, UK, 1982.

[22] F. L. Wightman and D. J. Kistler, "Factors affecting the relative salience of sound localization cues," in *Binaural and Spatial Hearing in Real and Virtual Environments*, pp. 1–24. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1997.

[23] W. Boucsein, *Electrodermal Activity*, Springer, New York, NY, USA, 2nd edition, 2012.

[24] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis," *Brain Res. Rev.*, vol. 29, no. 2–3, pp. 169–195, April 1999.

[25] M. Garbarino, M. Lai, D. Bender, R. W. Picard, and S. Tognetti, "Empatica E3 - A wearable wireless multisensor device for real-time computerized biofeedback and data acquisition," in *EAI 4th Int. Conf. Wirel. Mob. Commun. Healthcare (Mobihealth)*, Athens, Greece, November 2014, pp. 39–42.

[26] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *J. Neurosci. Methods*, vol. 190, no. 1, pp. 80–91, June 2010.

[27] W. J. Ray and H. W. Cole, "EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes," *Science*, vol. 228, no. 4700, pp. 750–752, May 1985.

[28] A. Keil, M. M. Müller, W. J. Ray, T. Gruber, and T. Elbert, "Human gamma band activity and perception of a gestalt," *J. Neurosci.*, vol. 19, no. 16, pp. 7152–7161, August 1999.

[29] N. A. Badcock, P. Mousikou, Y. Mahajan, P. de Lissa, J. Thie, and G. McArthur, "Validation of the Emotiv EPOC EEG gaming system for measuring research quality auditory ERPs," *PeerJ*, vol. 19, 2013.

[30] W. L. Zheng and B. L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, September 2015.

[31] C. Saitis and K. Kalimeri, "Identifying urban mobility challenges for the visually impaired with mobile monitoring of multimodal biosignals," in *Universal Access in Human-Computer Interaction - 10th International Conference*, M. Antona and C. Stephanidis, Eds. Springer-Verlag, Berlin, Germany, July 2016, In press.