Bring the Environment to Life: A Sonification Module for People with Visual Impairments to Improve Situation Awareness

Angela Constantinescu[1], Karin Müller[1], Monica Haurilet[2], Vanessa Petrausch[1], Rainer Stiefelhagen[1,2]
[1]Study Center for the Visually Impaired
[2]Computer Vision for Human-Computer Interaction Lab
Karlsruhe Institute of Technology, Karlsruhe, Germany
{angela.constantinescu,karin.e.mueller,haurilet,vanessa.petrausch,rainer.stiefelhagen}@kit.edu

ABSTRACT

Digital navigation tools for helping people with visual impairments have become increasingly popular in recent years. While conventional navigation solutions give routing instructions to the user, systems such as GoogleMaps, BlindSquare, or Soundscape offer additional information about the surroundings and, thereby, improve the orientation of people with visual impairments. However, these systems only provide information about static environments, while dynamic scenes comprising objects such as bikes, dogs, and persons are not considered. In addition, both the routing and the information about the environment are usually conveyed by speech. We address this gap and implement a mobile system that combines object identification with a sonification interface. Our system can be used in three different scenarios of macro and micro navigation: orientation, obstacle avoidance, and exploration of known and unknown routes. Our proposed system leverages popular computer vision methods to localize 18 static and dynamic object classes in real-time. At the heart of our system is a mixed reality sonification interface which is adaptable to the user's needs and is able to transmit the recognized semantic information to the user. The system is designed in a user-centered approach. An exploratory user study conducted by us showed that our object-to-sound mapping with auditory icons is intuitive. On average, users perceived our system as useful and indicated that they want to know more about their environment, apart from wayfinding and points of interest.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility design and evaluation method**; *User studies*; *Auditory feedback*; • **Computing methodologies** → *Computer vision.*

KEYWORDS

Assistive technologies; Sonification; Mixed reality; Evaluations of intelligent user interfaces; Object Localization; Computer Vision

ACM Reference Format:
Angela Constantinescu[1], Karin Müller[1], Monica Haurilet[2], Vanessa Petrausch[1], Rainer Stiefelhagen[1,2] . 2020. Bring the Environment to Life: A Sonification Module for People with Visual Impairments to Improve Situation Awareness . In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3382507.3418874

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418874

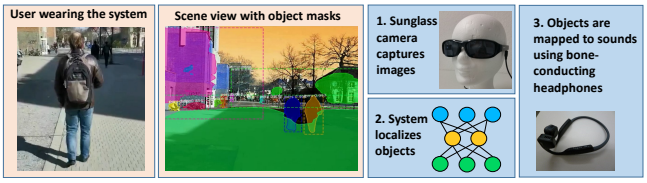

Figure 1: Overview of our three-step process: (1) A camera captures images of the environment; (2) The system localizes objects of the environment using a deep learning neural network; (3) The user is informed about the recognized objects via sonification.

1 INTRODUCTION

In order to improve the safety of a person with visual impairment while traveling, one needs to move beyond stationary information about the visual scene (e.g. buildings), as dynamic entities can become a hazard to the user, *e.g.* knowing in advance that a group of people is coming in one's direction or that a bike is parked on the sidewalk. This work introduces a framework for helping people with visual impairments navigate in unconstrained environments (see Figure 1) by addressing both static and dynamic scenes and tackling the problem of *conveying the recognized semantic information to the user.* In this context, we address three questions: (1) *How to recognize objects in the environment in real-time?* (2) *How to transmit this information to the user?*, and (3) *Is the output intuitive*[1] *and, thus, easy to learn by the target group?*

Regarding the first question, we apply computer vision methods to localize objects and demonstrate their effectiveness by "expanding" the users' perception of the environment.

To answer the second question, there are several ways in which the discovered information can be transmitted to people with visual impairments, the most common ones being speech, vibrotactile feedback, and sonification. Since we want to localize *multiple* objects in a scene at once, our system must inform the user about several objects in a very short period of time. However, speech is often too slow [21] and vibrotactile output is not versatile enough. Additionally, speech might interfere with the voice instructions from a routing system that tells the user which way to go. Moreover, when used

[1]Note that "intuitive" is used here to mean "directly apprehended" or "readily [...] understood" [17].

alone, *i.e.*, not in combination with speech instructions, sonification is language-independent. Finally, sonification does not require expensive additional hardware and can be more intuitive than some vibrotactile approaches. Due to these advantages, we focus in this paper on the sonification modality. Among available sonification approaches, we chose the following for transmitting information about objects in the scene to the user: (1) auditory icons and parameter mapping and (2) spearcons combined with parameter mapping.

For the final question, we evaluated the sonification interface with visually impaired people, and also assessed the intuitiveness and learnability of the sounds. Until now, there are only few studies that evaluate auditory icons for people with visual impairments [5, 9], and even fewer that consider the duration of the sounds [9]. To our knowledge, auditory icons have not yet been evaluated with the target group for intuitiveness and learnability.

Figure 1 illustrates the three-step process of our system applied in an outdoor study: (i) A camera captures images of the environment. (ii) The computer vision system localizes objects in the surrounding with a deep learning method and (iii) informs the user about the recognized objects via sonification. In all stages of the development, at least one user with visual impairment was involved.

This paper makes the following main contributions: (1) combining object identification with a sonification interface which can be used both for macro and micro navigation, (2) creation of a user friendly mixed-reality interface which is adaptable to the user's needs, and (3) insights from an exploratory user study with people with visual impairments.

The paper is organized as follows: In Section 2, we discuss related work regarding assistive systems that apply sonification techniques and computer vision for people with visual impairments. Then, in Section 3, we give an overview of the entire system and elaborate on the computer vision and sonification modules. In Section 4, we present an exploratory user study in which we evaluate our system in an outdoors scenario. We discuss the results of our user study and the design implications as well es the limitations in Sections 5 and 6. In Section 7, we draw the conclusions.

2 RELATED WORK

Sonification is used by many applications to give people with visual impairments access to visual information, *e.g.* to make graphs accessible [61, 63], to explore graphics or virtual maps [6, 25, 52], to indicate rotation instructions in indoor navigation environments [1], to guide the user or to support navigation tasks [5, 13, 33, 43, 59, 60, 65, 67], or to present nearby features, points of interest, and fixed obstacles [65].

Few of these approaches, however, leverage auditory icons. This sonification method was first used in computer systems in the 1980s [23] and later in both mainstream and audio games [26, 31]. Literature on this subject is scarce, especially with the focus on people with visual impairments.

Ferati *et al.* propose audemes, which are based on auditory icons, to facilitate interaction with large collections of spoken educational essays for pupils with visual impairments [20]. However, in contrast to our system, they do not apply them to convey objects in a real-time scenario, and the design of the auditory icons does not consider the duration of the sounds.

Aziz *et al.* [5] explore several sonification methods for auditory routes overviews, including text-to-speech, earcons, and auditory icons. They conclude that auditory icons are appropriate to convey information about points of interest. While they do mention that the auditory icons should be intuitive, they also suggest that a training phase should be provided to help users understand the mapping scheme.

Tislar *et al.* [58] investigate sonifications of objects through music, earcons, spearcons, and lyricons regarding their learnability, the relatedness of sounds, their attributed meanings as well as their intuitiveness. However, they do not include auditory icons and do not evaluate with visually impaired people. Thus, their results are not transferable, as people with visual impairments and sighted people have different preferences for user interfaces [63], as well as different cognitive loads when using them [37].

Dingler *et al.* [18] propose a work most similar to ours in terms of sonification. They compare auditory icons, earcons, spearcons, and speech used for representing objects in terms of learnability. But this work does not mention any limitation of the *duration of the sounds*, which is relevant to present various objects in a short time. They also trained the sounds first, while we investigated the intuitiveness first (see Section 4.2). Finally, they exclusively evaluate their system with sighted students.

Some approaches also use computer vision to detect objects in indoor or outdoor environments informing the user via speech [3, 8, 30] or vibrotactile feedback [66]. Recent systems combine speech with other scene recognition techniques such as: landmark recognition (*i.e.*, classifying the location captured in the image) [49], image captioning (*i.e.*, generating a textual description of the entire scene) [39], or multi-labeled image classification (*i.e.*, predicting a list of object classes in the scene) [2]. A disadvantage of these approaches is the use of a phone camera, which is very difficult for people with blindness to hold straight and also impractical when already holding a white cane in the dominant hand [49]. In comparison, our system continuously captures the entire scene, conveys both near and far objects to the user, is hands-free and does not require the user to hold the camera straight trying to capture an object or area of interest.

Other related methods try to incorporate a combination of computer vision systems and non-speech sounds to help the orientation of people in indoor environments [47], to identify known people in the environment through face recognition techniques, to help users find close objects using sonification methods [4, 40, 50, 51, 57], to warn cyclists about objects not in their field of vision [54], to give guidance during road crossing [38], or to warn with beeping sounds when other people block the way [34].

The most relevant approaches are the works proposed by Katz *et al.* [33] and Presti *et al.* [45]. Katz *et al.* [33] use computer vision methods to detect objects in outdoor environments, which are then sonified using 3D sound. The main difference to our approach is that their system can only localize a *single* class of objects at once upon user's request. So basically, the user is searching for a certain object. In comparison, we can localize up to 18 different object classes *simultaneously* and sonify them using several different sounds. Presti *et al.* [45] developed an iOS application to detect obstacles. If several obstacles are detected, the system informs the user about only one

Figure 2: Computer vision module: object detection (YOLO [46], left), instance segmentation (Mask-RCNN [15], center), and depth estimation (MonoDepth [27], right) of an image captured by our camera.

of them by transmitting its distance, size and direction via a combination of a base sound and an auditory icon. In comparison to this approach, our system does not only localize different objects in the scene but also *recognizes* them. People with visual impairments can use our system both for (1) orientation or macro navigation [56], by hearing static objects like walls, traffic lights, benches, and for (2) obstacle avoidance or micro navigation, by hearing static and dynamic objects at close range like persons or bicycles.

3 OVERVIEW OF THE SYSTEM

In this section, we describe our hardware encompassing four components: a smartphone, a laptop in a backpack, a sunglasses camera, and Bluetooth bone conduction headphones (the latter was first suggested by [65]). The software of our system includes a computer vision module and a sonification module.

We use a commercial, monocular HD camera running at around 25 fps and seamlessly integrated in a pair of sunglasses to capture the surrounding environment. The video frames are transmitted to the laptop through a short, discreet USB cable that starts behind the user's ear. We run the computer vision module on a laptop consisting of a i7-6700HQ processor, 16 GB of RAM, and a GTX 1070 GPU. A smartphone app [48] is the core of our system and connects the laptop with the headphones, handles the sonification and controls the preferred settings chosen by our users.

Figure 1 summarizes the interaction between the different components: The camera captures objects from the environment and transmits them through Bluetooth to the computer vision module. The detected objects are then transferred to the sonification module, which maps them to sounds and passes them again per Bluetooth on to the user. In one case, the hearing aid of a participant was directly connected to the smartphone instead of the headphones.

3.1 Computer Vision Module

In the following, we describe how we localize objects with computer vision models and, to that end, how we couple fine-grained object localization with depth estimation maps.

3.1.1 Choice of Objects. We analyzed several projects dealing with navigation systems for people with visual impairments as well as interviews with people from the target group to learn which information is important to the users [16, 29, 35, 62]. Based on this analysis, we selected the objects that meet the needs of people with visual impairments. We then compared the obtained list with the objects available in the datasets, resulting in 18 object classes (Table 2 shows an overview of the selected objects). These represent static and dynamic objects that can appear in urban environments.

3.1.2 Data Annotation. Since only a subset of the selected 18 classes are included in COCO [36] (a dataset for instance segmentation), we augmented labels from the COCO-Stuff [11] benchmark containing pixel-wise annotations. In total, we trained on around 80k images annotated with pixelwise labels of our 18 selected classes. The data is biased towards some classes, such as "car" and "person" which are frequently present in the dataset, while images including classes like "stairs" are scarce. Nonetheless, we had at least 800 images for each of our selected classes which we used to train the deep-learning models.

3.1.3 From Semantic Segmentation to Instance-Recognition. We experienced a discrepancy between localization types as COCO-Stuff includes *only* labels used for semantic segmentation. Thus, COCO-Stuff comprises object masks in the form $height \times width$ with natural numbers between 1 and the number of classes. Each value in these matrices associates the corresponding pixel with one of our classes. In comparison, the labels in COCO are in the form of instance segmentation annotations with bounding boxes surrounding each object and associated segmentation masks. More specifically, for each image, we had a set of annotations of the form: bounding box, binary mask (values of 0 and 1) of the size of the bounding box, and the class of the current instance.

Since we had annotations for two different problems, we needed to either change our model to be able to handle both tasks or transform the labels from one task into the other. A possible solution is to directly transform the instance segmentation labels to semantic segmentation ones. This strategy is straight-forward, as one can directly map the masks into the image. However, this deprives the model from the ability to distinguish between different instances, such as counting. Thus, we proposed to transform the semantic segmentation task into the instance segmentation setting, allowing the model to discriminate between instances. To do this, we employed the following protocol: (1) We divided the segmentation labels from COCO-Stuff class-wise obtaining binary maps for each class. (2) If a class was present, we calculated the smallest bounding box enclosing the pixels of the mask. (3) Finally, we combined these annotations into instance segmentation labels comprising the class, the enclosing bounding box, and the mask.

3.1.4 Object Recognition with Mask-RCNN. To localize the objects in a scene, we employed the popular Mask-RCNN [15] model for instance segmentation that implements a detect-and-segment strategy. This consisted of detecting the object, cropping the detected object from the scene using ROI-Align [15] and then generating the binary mask and the class of the detected instance. The network

was trained end-to-end with all the annotations from COCO, as well as the annotations that we mapped from the COCO-Stuff dataset.

3.1.5 Monocular Depth Estimation. Depth estimation methods approximate the distance of the captured scene to the camera. We employed a deep architecture that is able to generate the depth values from a *single* image. Thereby, we leveraged the popular MonoDepth network, an hourglass model comprising an encoder and a decoder. While the encoder decreases the size of the feature maps for an expansion in the receptive field, the decoder increases the shape of the feature map back to the original size of the image. The final prediction is a matrix of real values of the size $height \times width$. Since the network requires ground truth depth values during the learning phase, we evaluated two datasets containing depth maps. The benchmarks that we evaluated are: the popular large-scale KITTI dataset [24] as well as the outdoor Cityscapes benchmark [14]. While in KITTI the depth maps were computed using a Velodyne LiDAR scanner, Cityscapes employs two stereo cameras. As confirmed by our experiments, the model trained on KITTI was able to generate more precise depth maps. Thus, in the following, we report results of the MonoDepth model trained on KITTI.

3.1.6 Combining the Depth Values with the Semantic Information. To generate the final output of the computer vision module, we combined the detected instances with the depth estimation. For each instance, we used the mask to obtain all the depth values of the corresponding object. We approximated the final distance of the camera to the object instance by *averaging* this set of depth values. For obtaining the x-displacement, we used the location of the center point of the instance. Finally, we generated the set of instances with the following associated information: the object class (from the Mask-RCNN model), the distance to the object (extracted from the Mono-Depth network), and the x-displacement (approximated from the location of the instance found by the Mask-RCNN model).

3.1.7 Technical Evaluation. We evaluated the Mask-RCNN instance segmentation model on 278 frames of a video sequence captured by a blind participant with the sunglasses camera. The network achieved a mean Average Precision of 65% at an intersection over union of at least 0.5 and averaged over the 9 object classes that were present in the video sequence (bicycle, bus, car, motorcycle, truck, person, stop sign, bush, and wall).

3.2 Sonification Module of the Interface

In this section, we describe how we chose the sonification type, created the sounds, and how we mapped the objects to auditory icons.

3.2.1 Choice of Sonification Method. In our system, we used sonification for announcing objects and their position with respect to the user. To better understand sonification within a navigation context, we conducted a survey of existing systems and theoretic sonification concepts in advance. The results indicated that for object recognition, auditory icons combined with parameter mapping (parametric auditory icons [53]), are best suited. The only parameters that we mapped to object data were loudness (mapped to object distance) and panorama (mapped to x-axis displacement). This means that we only need to use one auditory icon for one object class and can set the parameters in real time. The second preferred method is to combine spearcons with parameter mapping for displaying information about objects.

3.2.2 Creating the Sounds. We sonified all objects with distinct sounds by creating an auditory icon for each object. Additionally, we used the parameter mapping approach to map distance to volume and displacement on the x-axis to stereo sound (panning). Panning of an audio signal is the procedure of making the sound seem to come from a certain direction, in our case from the object. Mapping distance to volume has the advantage that far away objects, which are less important or less urgent, will be more quiet and will therefore not mask environmental sounds.

To create the auditory icons, we chose sounds that are intuitive and thus can be easily matched with an object, *e.g.* a bicycle was represented by the turning sound of a bicycle's wheel. The sounds for "wall" and "door" are similar since both belong to building.

The sounds have to be extremely short since there are usually several objects in one frame which are played sequentially. So we aimed at creating sounds no longer than 500 ms – the shorter the better, while keeping intact their conceptual mappings. This was a very difficult task, as sounds loose their recognizability very quickly when shortened. As an example: the object "stairs" is represented by the sound of a person going down two steps on a wooden stair. When shortening the sound, the object "stairs" can not be recognized anymore. In the end, the duration of the 18 sounds ranged between 240 ms (dog) and 850 ms (truck) with the exception of traffic light, which exceeded 1 second (1240 ms).

We compared the durations of our auditory icons in milliseconds (ms) with the durations of their corresponding spoken words. The speech was generated using Gespeaker [12], a free GTK+ frontend for espeak [22], with the German-mbrola-5 voice, the default speed of 175 words per minute and a delay of zero. For all our 18 sounds, the average duration is 538 ms for the auditory icons and 660 ms for speech - which is 23% longer. The difference is 122 ms and the standard deviation (SD) 230 ms. When omitting the sounds for traffic light because of its exceptional long duration, the average difference between auditory icons and speech is 164 ms (speech on average 33% longer), $SD = 151$ ms. This shows that the chosen auditory icons are shorter than speech and are thus more suitable for our task.

The fact that our sounds [2] are so compact has another advantage: they do not resemble as much the original, natural sounds, so they are more easily distinguishable from environmental sounds. This is a key aspect when using a system outdoors.

3.2.3 Mapping Localized Objects to Auditory Icons. The objects localized by the computer vision module were converted to sounds according to the mapping described in the previous section. The following procedure guaranteed that only the most recent localized objects were sonified. Once the system was turned on, the first captured frame was processed and all localized object instances were sorted by the distance to the camera (*i.e.* user). Next, the items were filtered based on the user's configuration regarding the object selection and the range at which objects should be localized. Finally, the objects were sonified and the auditory icons were played sequentially starting from the object with the shortest distance to the

[2] The sounds for the 18 object classes can be downloaded at https://www.szs.kit.edu/accesslab/downloads.

user. The closer the item, the louder the sound. The displacement of the object on the x-axis mapped to sound panning, so the user heard the sound as if coming from the object itself on the horizontal plane. Our system discarded all subsequent frames until the auditory icons of the first frame had finished playing. Then, the next most recent captured frame was processed. We processed at most one frame per second, allowing the sonification to play during the remaining time window. Depending on the duration of each auditory icon played (in our case between 240 and 1240 ms), and on the number of objects identified in the current frame, the system could play between 0 and 4 sounds per second.

When the user selected a large number of object classes to be localized, the system often had to sonify more than four objects in one frame. Subsequent frames were discarded until all objects had been sonified and played out. Thus, the user received a detailed overview of the scene, at the expense of a timely feedback, as the system only processed one frame in a few seconds. When the user picked only few object classes, or reduced the distance at which objects should be localized, most of the time, up to four objects are localized and sonified in one frame, so that the system had to process only one frame per second. The user could then leverage the information from the different frames to observe the change in environment by tracking certain objects as he walks.

3.2.4 Adaptability of the Interface. The sonification of each object could be individually turned on and off from the interface and the maximum distance of the objects could be changed at any time. By default, the distances were as follows: for vehicles (cars, motorcycles, buses, trucks, and trains), it was 30 m, for bicycles 15 m, people and dogs 10 m, and static objects 10 m. This means that if only cars were turned on and the distance for vehicles was set to 15 m, then only information about cars within 15 m from the camera was transmitted to the user.

4 USER STUDY AND EVALUATION

We pre-evaluated the sounds and the interface with two sighted accessibility experts before we conducted our user study. We evaluated our interface and the system in an exploratory study with five people with visual impairments. According to Nielsen [42] and Pernice *et al.* [44], five people are sufficient to qualitatively evaluate a system and to draw design implications. Aziz *et al.* also argue that fewer participants is "common practice when working with a niche population" [5]. The general aim of our study was to investigate whether and how an object localization module based on computer vision and combined with a sonification interface can be helpful to people with visual impairments. To that end, in our study, we address the following questions:

(1) How to design the interface to meet the users' needs?
(2) Are auditory icons suitable to convey information about objects in a scene?
(3) Are the mappings of objects to specific auditory icons appropriate to facilitate memorizing them?
(4) Is the computer vision algorithm fast enough to localize multiple object instances in real-time?
(5) Which objects including obstacles, points of interest, and orientation landmarks are most important for people with visual impairments?

4.1 Participants

Five male adults with ages ranging from 21 – 50 years, and an average age of 30.2 years (SD 11.7) participated in the study. One of them (P1) was severely visually impaired and could see objects at 10 – 20 m ahead, depending on the light conditions. Two (P2, P4) were congenitally blind and could not see at all, and two (P3, P5) were legally blind according to German law, one of whom (P3) could see objects at very close range, according to his own statements. All of them travelled most of the time alone and on foot and four also used public transportation besides walking. Three out of five (P3-P5) used digital navigation aids, one (P2) very rarely and one (P1) not at all. All participants had received mobility training before. Four of them had a white cane and used it during the evaluation. One (P3) had neither a white cane nor a guide dog. An overview of the participants is given in Table 1.

4.2 Methodology

We started our study by explaining the main idea of the system and the procedure to the participants. The subjects then signed the statement of consent and filled in a questionnaire on demographics. The study was divided into four phases: sound evaluation, outdoor evaluation, assessment, and comparison of two sonification methods. The duration of the entire study was around 2 – 3 hours per participant.

(1) Sound evaluation phase. The participants were first accustomed to the auditory icons, while we also tested the intuitiveness and learnability of the sounds at the same time. In a first step, we read aloud the list of 18 objects. Then, the sounds for all objects were played in a random order, and after each sound, the participants were asked to guess the object it represented (testing the intuitiveness). If the participant was not able to guess it correctly, the test leader told them the object name as well as the conceptual mapping (how the sound was created and how it relates to the object). Any sound could be repeated upon user's request. In a second step, the sounds were played again in a random order to test if the participants could remember them correctly.

(2) Outdoor evaluation phase. After the evaluation of the sounds, we asked our participants to walk on a familiar urban route and say aloud everything they were thinking [41] to retrieve as many comments as possible. With this evaluation, we wanted to learn how people with visual impairments use audible information about the outdoor environment. The test leader accompanied the participants, ensuring their safety and taking notes. The session was recorded with an action camera carried by the participants and attached to the backpack strap.

(3) Assessment phase. After the walk, the participants were asked to fill in a questionnaire that was partially based on the "I like, I wish, what if" method [32]. We asked them: (1) what they liked, (2) what they did not like, (3) suggestions that may not have a link to the prototype; and additionally, (4) how useful they found each of the 18 object classes (on a Likert scale rating from 1 to 5) and in what situation, (5) to name three other relevant objects not included in the test, (6) at what distance and (7) at which angle should objects be identified, (8) if they liked the camera integrated in the glasses, and (9) if and in what situations they would use the system.

(4) Comparison of sonification methods phase. Although auditory icons were our main interest, we also wanted to see how

Table 1: Demographic data of participants.

P	Age	Gend.	VI	Onset of VI	See objects?
1	30	m	low vision	since 9 ys	Yes, light dep.
2	22	m	blind	since birth	No
3	28	m	blind	since 26	Yes, in front
4	50	m	blind	since birth	No
5	21	m	blind	after childhood	No

they compare to spearcons in terms of intuitiveness and users' subjective assessment (like/dislike). At the end of the study, we tested the intuitiveness of spearcons in the same way as for auditory icons. We played the spearcons in random order, and asked the users to guess which object it represents. Users were also asked which approach they prefer, auditory icons or spearcons.

4.3 Results

We evaluated our study by analyzing (1) the times and success rates from the sound evaluation phase and (2) the questionnaire together with the comments given by the participants during their walk. From the questionnaire, we wanted to find out what objects in general were most relevant to people with visual impairments during navigation. Moreover, we wanted to learn how the users intend to use the interface in general regarding the number of objects and the maximum distance at which objects are announced.

4.3.1 Intuitiveness and Learnability of the Object-to-Sound Mapping. We evaluated the intuitiveness of the auditory icons by analyzing the success rates of the participants in guessing which sound belongs to which of the objects. In a second round, we checked the learnability rate by assessing how easily the participants remembered the object classes associated with the sounds, as proposed by Hermann *et al.* [28].

The values of the rating were as follows:

1 - if the user did not guess or did not remember the meaning of the sound
2 - if the user guessed or remembered the sound after thinking for longer than 3 s
3 - if the user guessed or remembered the sound within less than 3 s

Table 2 shows the average intuitiveness rates for all five participants (note that we are missing 3 values for intuitiveness and 2 values for learnability, from 180 values in total). **The most intuitive sounds were "bicycle", "door", "bush", "dog", and "motorcycle"**. The least intuitive sounds were for "truck", "stop sign" and "bench". "Truck", however, had the highest learnability rates, possibly also because the sound was very stringent and different from most others. Once the conceptual mapping was known, the sound was indeed easier associated with a truck. For "stop sign" and "bench" it was difficult to find a good conceptual mapping. The sound for "stop sign" was the "beep" tone that some car navigation systems make to indicate a traffic sign. The sound that we used for "bench" is very similar to the one for "chair" (we selected them this way as the two concepts are similar), but this caused them to be frequently confused. We logged the time of four participants until they chose the correct auditory icon. It took them on average 8.5 minutes

Table 2: Intuitiveness and learnability rates of auditory icons in descending order through: Average (AV) and Standard deviation (SD). The ranking of the importance from 5=useful (dark blue) to 1=useless (light blue): AV and SD.

	Intuitiveness		Learnability		Ranking	
Objects	AV	SD	AV	SD	AV	SD
Bicycle	2.8	0.45	2.4	0.89	4.6	0.89
Door	2.6	0.55	2.8	0.45	4.0	1.22
Bush	2.4	0.89	2.4	0.8	2.4	1.14
Dog	2.4	0.89	2.8	0.45	2.8	0.84
Motorcy.	2.4	0.89	2.6	0.55	4.4	0.89
Bus	2	0.71	2.2	0.84	3.2	1.30
Train	2	0	2.84	0.45	3.6	0.89
Fence	2	1	2.4	0.89	4.0	0.71
Traffic Light	1.8	0.84	2.6	0.55	4.8	0.45
Train Tracks	1.6	0.89	2.6	0.55	3.8	1.10
Car	1.4	0.89	2.6	0.55	3.8	1.30
Person	1.4	0.98	2.75	0.50	3.2	1.10
Stairs	1.4	0.89	2.6	0.55	4.8	0.45
Chair	1.4	0.89	2.8	0.45	3.4	1.67
Wall	1.2	0.45	2.6	0.55	3.2	1.30
Truck	1	0	3	0	4.2	1.10
Stop Sign	1	0	2.75	0.50	1.6	1.34
Bench	1	0	2.2	0.84	4.2	0.8

(SD 2.15 minutes) to guess correctly the auditory icons for the 17 classes (we excluded "stop sign", which was simply introduced by the test leader). Explanations and comments from the test leader as well as the time for playing the sounds are also included in the guessing time. **The average learnability time for all** 18 **objects for these** 4 **participants was** 3.2 **minutes.** Given the times and the rates for both intuitiveness and learnability, we infer that the sounds used were easy to learn, and some of them even intuitive. It will have to be further analyzed whether the poor intuitiveness rates for some sounds are due to the conceptual mappings or the sound design.

For spearcons, the average intuitiveness rates for all users and all objects (for P1 we only have the intuitiveness rates for 5 objects) was 2.5 out of a maximum of 3.0 (SD 0.5). This is much higher than for auditory icons, which only had an average intuitiveness rate of 1.7 (SD 0.5). This could be caused by the fact that the words are already hidden in the compressed sound. All users said, however, that they prefer auditory icons. There could be a bias here, as the spearcons were evaluated at the end of the entire study, when the users were already accustomed to the auditory icons. The comment of P5 also suggests this: "*I think the others [auditory icons] are more feasible, because it's easy, [...] it's intuitive. You have to learn it first, but when the brain made the connection, it's just as if I were hearing a car coming from behind. Maybe this also works with words, but it will have to be evaluated*". Thus, spearcons are in any case intuitive and, thus, suitable for sonifying objects, but another study is necessary to investigate whether they are better than auditory icons. However, we suspect that auditory icons are easier to distinguish from language in the navigation context than spearcons.

Table 2 shows the assessment of the participants when asked about the importance of objects during navigation. The **most useful objects selected by our subjects were stairs, traffic lights,**

bicycles, and motorcycles, all crucial during critical situations. When looking at the object "walls", it did not rank very high, but for one participant (P5) it was the most important object class, as it informed him about the existence of buildings, which in turn helped him to orientate. When asked in what situation the users found objects useful, the answers differed. P2 for instance only wanted to know about stationary bicycles, while P3 only about moving ones. P4 could hear walls anyways and, thus, did not need to know about their location at all, P3 only needs walls when it is dark outside, and P5 found them the most useful object class. This shows that the choice of objects is very personal and diverse, and depends on the users' navigation attitudes [64]. Thus, it is essential that a system offers the user the possibility to *choose* from a larger set of object classes. When asked to name three other objects that they find relevant and were not part of the test set, the participants mentioned: pole (P2, P3, P5), garbage can (P4, P5), tree (P3, P4), water: brook, river, lake (P2), puddle (P1), fountain (P3), mailbox (P1), bus stop (P3), open trunk (P4) and truck's loading dock (P4).

4.3.2 Assessment of the System in General. The participants appreciated the idea of the system in general. One of them said "*I like that you get a lot information from the environment, what happens left and right.*", and another one added "*It tells me things that I don't perceive otherwise.*"

The audio interface based on auditory icons and parameter mapping was rated very positively. Two participants (P3, P4) said that they were positively surprised by the interface, as they expected speech to be used. Others also liked the panning (P1, P3, P4) and distance to loudness mapping (P3). Two other participants (P1, P2) praised the interface in general and said they were astonished that **one can learn the sounds so fast**. P5 remarked "*Sounds don't disturb much; one gets used to them quickly*". He also commented that "*it is similar to sight: one turns the head and perceives what's there - direct feedback - quite cool*". Thus, we conclude that the sonification chosen enabled the users to profit from the interface very fast. The comments strongly support the results of our survey that auditory icons combined with parameter mapping are suitable for object localization.

All of the participants named at least two objects that they found very useful. They appreciated that one can choose from so many object classes, and **configurability of the system was assessed paramount**. Each participant had his own set of preferred objects, according to the ranks and also comments. We also found that the **preferred distance for the object classes is different for each person** and may depend on the degree of impairment: P2 and P4, who are blind from birth, would mainly use the system short range (up to 10 m for most things), while P3 and P5, who have some minimal residual sight, found larger distances best. Regarding the angle at which objects should be identified, 4 out of 5 participants said that at least short range, the angle should be wider than the one of the white cane, if possible full range. P2 commented that for searching objects such as a park bench, the angle should be wider, while for obstacle avoidance the same angle as the span of the white cane is sufficient, but at a greater distance. There are several possible reasons for this divergence in opinions, including: the individual's orientation and mobility proficiency, environment familiarity, changes in environment, or navigation aids [7].

Users appreciate the inconspicuousness of the camera integrated into sunglasses. However, they suggested that the lenses should be exchangeable with transparent ones when needed, so as not to disturb their residual perception of light, as it is helpful for navigation (P1, P2). Moreover, they also proposed that a camera should be wireless, and one of the users mentioned that the temple stem interferes with the frame of the bone conduction headphones. One participant (P1) wished he was able to change between a chest camera and one mounted on glasses. Two participants (P3, P5) proposed a headband similar to headlamp or only a light frame under the eyes instead of the full glasses (P5).

The participants were interested in further using the system to orient themselves (walls, traffic lights), to find things (bench, car/taxi, person), and to avoid obstacles (bicycle, car, person). According to P1, the interface can be improved such that dangerous dynamic objects are more prominent. The comments of the participants in general point out that orientation strategies and needs differ very much between people and, thus, supporting systems require a high adaptability to the users' needs which corresponds with the findings of Williams *et al.* [64].

5 DISCUSSION AND DESIGN IMPLICATIONS

We investigated through an exploratory study how information about the environment can be conveyed to people with visual impairments when walking outdoors. In this paper, we demonstrate the feasibility of using a camera-based system to localize multiple objects and to pass this information on to the user via sonification. We observe several issues regarding the camera, the creation of the sounds, the conceptual mappings of objects to sounds, dangerous objects and the interaction with the user interface.

Camera. The advantage of a head-mounted camera is that the camera follows the movement of the head, widening the angle of the surrounding. As a participant expressed it, "*It is similar to sight: one turns the head and perceives what's there*". Moreover, a camera worn on the body is better for people with visual impairments than one held in the hand.

Creation of Auditory Icons. We showed that the sounds can be learned very easily and that on average, they are shorter than speech. Even more, when considering only the 5 most intuitive and easy to learn auditory icons (door, bicycle, dog, motorcycle, and bush), their average duration is 372 ms, while the corresponding speech lasts on average 627 ms - 69% longer (SD=70 ms). Or considering the sound for dog: the auditory icon was the shortest of the 18 sounds, namely 240 ms long, and had an intuitiveness of 2.4 and learnability of 2.8 (out of 3.0). This shows that one can find auditory icons that are very short and still intuitive, thus easy to remember. One participant expressed his preference for auditory icons as follows "*I generally prefer sounds. Somehow you perceive them directly while you have to interpret speech (spearcons)*".

The main challenge is to find sounds that preserve the meaning but are short enough, such that sufficient information about the environment is passed on the move. It is unclear how many sounds a user can discern and interpret in a short period of time and on a continuous basis. This needs further investigation.

It is possible that advanced users who listen to speech at rates of up to 500 words per minute [19] will profit more from using spearcons instead of auditory icons. However, since not everyone

uses the same speech rate, well-designed auditory icons offer the possibility to address all users.

Conceptual mappings of objects to sounds. The conceptual mappings of objects to sounds is a topic that has not yet been explored in detail. For instance, for bicycles, we chose the sound of a spinning wheel, where a flexible object is held in the spokes. The users of our evaluation found this sound very intuitive (average rate of 2.8 out of 3.0) and easy to use (average rate of 2.4 out of 3.0). In the literature, however, a bicycle bell is commonly used [10]. This shows that it is necessary to develop a better understanding of the conceptual mappings of objects to sounds which are most intuitive for people with visual impairments.

Dangerous objects. Currently all objects are treated the same. If they occur in the scene, they are sonified and the user will be informed. However, the user must be able to recognize when object positions change significantly and can pose a danger, for instance, those that move towards the user. One possibility would be to track these objects and change the property of the sound or combine it with a certain warning sound.

Interaction with the User Interface. Our study showed that the needs and capabilities of people with visual impairments are very diverse. The users exploited the options to switch on and off certain objects while walking (which was accomplished by the test leader during the experiment). For instance, one of the participants dropped the object class “car” because there were too many parked cars along the way, thus this information was not useful. This shows that an assistive system must contain a large pool of object classes from which the users can choose easily and therefore adapt to their needs. The user should either have the option of making changes on the move without stopping, or the system should automatically make changes depending on the environment, task, and user characteristics.

6 LIMITATIONS

The promising results of our initial work indicate many possibilites for future investigation. The most obvious one is the hardware: while the system is mobile, carrying a laptop in a backpack is unacceptable, thus, we are currently working on minimizing the hardware. Some users also complained about concomitant use of sunglasses and bone conducting headphones. One solution would be to use glasses that have integrated earphones.

The number of participants in our study could also be regarded as a limitation. A final system should be tested with more users, and with more diverse demographics. However, for a prototype, we consider that five users provide us with enough feedback to drive the development in a positive direction and inform future research.

A comparison between auditory icons and spearcons requires a subsequent test, where both sonification approaches are tested similarly and interchangeably with users - whether with the same users but in interchangeable order or with different users. The problem with our test was that spearcons were only tested at the end, when the users were already accustomed with the auditory icons, so a comparison is difficult to make based only on our data.

The think-aloud method that we used during the user study has several disadvantages: first of all, people may not say what they are thinking [55]; secondly, describing one’s thoughts may be difficult and "alter the process" [55]; finally, it is difficult for a person with visual impairments to talk while walking, since talking interferes with the user’s awareness of the environment. While we cannot influence the first two factors, for the third one, we made sure that the users walk on routes that are well known to them. We also observed that they often stopped in order to make a longer comment. Since we did not record the times walked, this fact did not influence our study.

The computer vision algorithm needs to be further improved, as some objects, such as doors, were poorly identified throughout the test. Others, such as stairs, were wrongly identified, for instance as benches (presumably due to the small number of samples seen during training for these categories). Additionally, since the preferences of the users for object classes were very diverse, including a larger pool of classes to choose from would ensure that more people would use the system.

7 CONCLUSIONS

In this paper, we present a mobile system to support people with visual impairments in both micro and macro navigation in an outdoor environment. Our framework combining computer vision with sonification shows great promise in revealing otherwise inaccessible environmental information to people with visual impairments. Our exploratory study demonstrated that parametric auditory icons are best suited for conveying information about objects in the scene to the target users. We also showed that parameterized auditory icons can be learned very fast, despite their shortness. A major finding is that it is crucial for the system to be adaptable to the user’s current needs regarding the object classes and the distance at which objects are sonified. We also found that the objects that are relevant for the pedestrians’ safety, such as stairs, traffic lights, bikes, and motorcycles were assessed very important. Moreover, auditory icons were considered better than spearcons by all five participants, in the given test setting. In general, we show that sonification and especially auditory icons are a suitable means for enabling people with visual impairments to better understand their environment. Finally, the results of our study will be helpful in developing mixed reality mobility solutions for outdoor use.

The current work will be further developed in different aspects. Considering the sonification method, we intend to further improve and investigate the conceptual mappings of auditory icons with a larger group of participants. In a next step, our goal is to make our computer vision module more portable, *i.e.* to implement the object localization on a different platform. Furthermore, we plan to introduce additional relevant object classes. We also plan to test our system for objects localization and sonification together with a text-to-speech navigation module in a more extensive study with a larger number of participants with visual impairments.

8 ACKNOWLEDGEMENTS

The project is partially funded by the German Federal Ministry of Labour and Social Affairs (BMAS), grant number 01KM151112 and by the German Federal Ministry of Education and Research (BMBF), grant number 16SV7609.

REFERENCES

[1] Dragan Ahmetovic, Federico Avanzini, Cristian Bernareggi, Gabriele Galimberti, Luca Ludovico, Sergio Mascetti, and Giorgio Presti. 2019. Sonification of Rotation Instructions to Support Navigation of People with Visual Impairment. 1–10. https://doi.org/10.1109/PERCOM.2019.8767407

[2] Aipoly Inc. 2020. Aipoly Vision: Sight for Blind Visually Impaired. Retrieved August 8, 2020 from https://apps.apple.com/us/app/aipoly-vision-sight-for-blind-visually-impaired/id1069166437

[3] Aipoly Inc. 2020. Orcam MyEye 2. Retrieved August 8, 2020 from https://www.orcam.com/en/myeye2/

[4] Malika Auvray, Sylvain Hanneton, and J Kevin O'Regan. 2007. Learning to Perceive with a Visuo - Auditory Substitution System: Localisation and Object Recognition with 'The Voice'. *Perception* 36, 3 (2007), 416–430.

[5] Nida Aziz, Tony Stockman, and Rebecca Stewart. 2019. An Investigation Into Customisable Automatically Generated Auditory Route Overviews for Pre-navigation. https://doi.org/10.21785/icad2019.029

[6] Michael Banf and Volker Blanz. 2012. A Modular Computer Vision Sonification Model for the Visually Impaired. In *The 18th International Conference on Auditory Display (ICAD2012)*.

[7] Nikola Banovic, Rachel L. Franz, Khai N. Truong, Jennifer Mankoff, and Anind K. Dey. 2013. Uncovering Information Needs for Independent Spatial Learning for Users Who Are Visually Impaired. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2513383.2513445

[8] Fereshteh S. Bashiri, Eric LaRose, Jonathan C. Badger, Roshan M. D'Souza, Zeyun Yu, and Peggy Peissig. 2018. Object Detection to Assist Visually Impaired People: A Deep Neural Network Adventure. In *Advances in Visual Computing*, George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Matt Turek, Srikumar Ramalingam, Kai Xu, Stephen Lin, Bilal Alsallakh, Jing Yang, Eduardo Cuervo, and Jonathan Ventura (Eds.). Springer International Publishing, Cham, 500–510.

[9] Brandon Biggs, James M. Coughlan, and Peter Coppin. 2019. Design and evaluation of an audio game-inspired auditory map interface. In *Proceedings of the 25th International Conference on Auditory Display (ICAD2019)*.

[10] João Paulo Cabral and Gerard Bastiaan Remijn. 2019. Auditory icons: Design and physical characteristics. *Applied Ergonomics* 78 (2019), 224 – 239. https://doi.org/10.1016/j.apergo.2019.02.008

[11] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. Coco-stuff: Thing and stuff classes in context. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[12] Fabio Castelli. 2020. Gespeaker. Retrieved August 8, 2020 from http://www.muflone.com/gespeaker/english

[13] Angela Constantinescu and Tanja Schultz. 2011. Redundancy Versus Complexity in Auditory Displays for Object Localization - A Pilot Study. In *International Conference on Auditory Display, ICAD 2011*. 222–235. http://hdl.handle.net/1853/51742

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *IEEE conference on computer vision and pattern recognition*.

[15] Jifeng Dai, Kaiming He, and Jian Sun. 2016. Instance-aware semantic segmentation via multi-task network cascades. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[16] Alireza Darvishy, Hans-Peter Hutter, and Alexander Seifert. 2016. *Altersgerechte mobile Applikationen. Grundlagen und Empfehlungen*. Zürcher Hochschule für Angewandte Wissenschaften. https://www.zhaw.ch/storage/engineering/institute-zentren/init/HII_-_Human_Information_Interaction/IAL_-_ICT-Accessibility_Lab/AppBrochure_20160822_GzD2.pdf.

[17] Dictionary by Merriam-Webster. America's most-trusted online dictionary. 2020. intuitive. Definition of intuitive. Retrieved August 8, 2020 from https://www.merriam-webster.com/dictionary/intuitive

[18] Tilman Dingler, Jeffrey Lindsay, and Bruce N. Walker. 2008. Learnability of Sound Cues for Environmental Features: Auditory Icons, Earcons, Spearcons, and Speech. In *Proceedings of the 14th International Conference on Auditory Display (ICAD)*.

[19] Alistair D.N. Edwards. 1991. *Speech Synthesis: Technology for Disabled People*. Paul Chapman Educational Publishing.

[20] Mexhid Ferati, Steve Mannheimer, and Davide Bolchini. 2011. Usability Evaluation of Acoustic Interfaces for the Blind. In *Proceedings of the 29th ACM International Conference on Design of Communication (SIGDOC '11)*. Association for Computing Machinery, New York, NY, USA, 9–16. https://doi.org/10.1145/2038476.2038479

[21] Emerson Foulke. 2006. Listening Comprehension as a Function of Word Rate. *Journal of Communication* 18, 3 (02 2006), 198–206.

[22] Free Software Foundation, Inc. 2020. eSpeak text to speech. Retrieved August 8, 2020 from http://espeak.sourceforge.net

[23] William W. Gaver. 1986. Auditory Icons: Using Sound in Computer Interfaces. *Human–Computer Interaction* 2, 2 (1986), 167–177.

[24] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research* (2013).

[25] Michele Geronazzo, Alberto Bedin, Luca Brayda, Claudio Campus, and Federico Avanzini. 2016. Interactive spatial sonification for non-visual exploration of virtual maps. *International Journal of Human-Computer Studies* 85 (2016), 4–15. https://doi.org/10.1016/j.ijhcs.2015.08.004 Data Sonification and Sound Design in Interactive Systems.

[26] GMA Games. 2020. ShadesOfDoom. Retrieved August 8, 2020 from https://www.audiogames.net/db.php?id=shadesofdoom

[27] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. 2017. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[28] Thomas Hermann, Andy Hunt, and John G. Neuhoff (Eds.). 2011. *The Sonification Handbook*. Logos Publishing House, Berlin, Germany. 1–586 pages. http://sonification.de/handbook

[29] HFC Human-Factors-Consult GmbH 2019. *OIWOB - Orientieren, Informieren, Warnen. Orientierungshilfe für Blinde*. HFC Human-Factors-Consult GmbH. http://www.oiwob.de/start.

[30] Andreas Hub, Joachim Diepstraten, and Thomas Ertl. 2003. Design and Development of an Indoor Navigation and Object Identification System for the Blind. Association for Computing Machinery, New York, USA. https://doi.org/10.1145/1028630.1028657

[31] Impression Games. 2020. CaesarIII. Retrieved August 8, 2020 from https://de.wikipedia.org/wiki/Caesar_(Spieleserie)#Caesar_III

[32] Interaction Design Foundation. 2020. I-Like-I-Wish-What-If. Retrieved August 8, 2020 from https://public-media.interaction-design.org/pdf/I-Like-I-Wish-What-If.pdf

[33] Brian Katz, Slim Kammoun, Gaëtan Parseihian, Olivier Gutierrez, Adrien Brilhault, Malika Auvray, Philippe Truillet, Michel Denis, Simon Thorpe, and Christophe Jouffrais. 2012. NAVIG: Augmented reality guidance system for the visually impaired: Combining object localization, GNSS, and spatial audio. *Virtual Reality* 16 (01 2012), 17.

[34] Seita Kayukawa, Keita Higuchi, João Guerreiro, Shigeo Morishima, Yoichi Sato, Kris Kitani, and Chieko Asakawa. 2019. BBeep: A Sonic Collision Avoidance System for Blind Travellers and Nearby Pedestrians. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3290605.3300282

[35] Kreis Soest. 2020. m4guide - mobile multi-modal mobility guide. Retrieved August 8, 2020 from https://www.kreis-soest.de/guide4blind/ueberuns/folge/m4guide/m4guide.php

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer.

[37] M. Martinez, A. Constantinescu, B. Schauerte, D. Koester, and R. Stiefelhagen. 2014. Cognitive Evaluation of Haptic and Audio Feedback in Short Range Navigation Tasks. In *Proc. 14th Int. Conf. Computers Helping People with Special Needs (ICCHP)*. Springer, Paris, France.

[38] Sergio Mascetti, Lorenzo Picinali, Andrea Gerino, Dragan Ahmetovic, and Cristian Bernareggi. 2016. Sonification of guidance data during road crossing for people with visual impairments or blindness. *International Journal of Human-Computer Studies* 85 (2016), 16–26. https://doi.org/10.1016/j.ijhcs.2015.08.003 Data Sonification and Sound Design in Interactive Systems.

[39] Microsoft. 2020. Seeing AI. Turning the visual world into an audible experience. Retrieved August 8, 2020 from https://www.microsoft.com/en-us/ai/seeing-ai

[40] R. Nagarajan, S. Yaacob, and G. Sainarayanan. 2003. Role of object identification in sonification system for visually impaired. In *TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region*, Vol. 2. 735–739 Vol.2. https://doi.org/10.1109/TENCON.2003.1273276

[41] Jakob Nielsen. 1994. *Usability engineering*. Elsevier.

[42] Nielsen Norman Group 2019. *How Many Test Users in a Usability Study?* Nielsen Norman Group. https://www.nngroup.com/articles/how-many-test-users/.

[43] G. Parseihian, C. Gondre, M. Aramaki, S. Ystad, and R. Kronland-Martinet. 2016. Comparison and Evaluation of Sonification Strategies for Guidance Tasks. *IEEE Transactions on Multimedia* 18, 4 (2016), 674–686.

[44] Kara Pernice and Jakob Nielsen. [n. d.]. *How to Conduct Usability Studies for Accessibility. Methodology Guidelines for Testing Websites and Intranets With Users Who Use Assistive Technology*. Nielsen Norman Group.

[45] Giorgio Presti, Dragan Ahmetovic, Mattia Ducci, Cristian Bernareggi, Luca Ludovico, Adriano Baratè, Federico Avanzini, and Sergio Mascetti. 2019. WatchOut: Obstacle Sonification for People with Visual Impairment or Blindness. Association for Computing Machinery, New York, USA. https://doi.org/10.1145/3308561.3353779

[46] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[47] F. Ribeiro, D. Florêncio, P. A. Chou, and Z. Zhang. 2012. Auditory augmented reality: Object sonification for the visually impaired. In *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*. 319–324. https:

//doi.org/10.1109/MMSP.2012.6343462

[48] Sebastian Ritterbusch, Patrick Frank, and Vanessa Petrausch. 2018. Modular Human-Computer Interaction Platform on Mobile Devices for Distributed Development. In *International Conference on Computers Helping People with Special Needs*. Springer, 75–78.

[49] Manaswi Saha, Alexander J. Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. 2019. Closing the Gap: Designing for the Last-Few-Meters Wayfinding Problem for People with Visual Impairments. Association for Computing Machinery, New York, USA. https://doi.org/10.1145/3308561.3353776

[50] G. Sainarayanan, R. Nagarajan, and Sazali Yaacob. 2007. Fuzzy Image Processing Scheme for Autonomous Navigation of Human Blind. 7, 1 (2007). https://doi.org/10.1016/j.asoc.2005.06.005

[51] Boris Schauerte, Manel Martinez, Angela Constantinescu, and Rainer Stiefelhagen. 2012. An Assistive Vision System for the Blind That Helps Find Lost Things. In *Computers Helping People with Special Needs*, Klaus Miesenberger, Arthur Karshmer, Petr Penaz, and Wolfgang Zagler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 566–572.

[52] Boris Schauerte, Torsten Wörtwein, and Rainer Stiefelhagen. 2015. A Web-based Platform for Interactive Image Sonificationn. In *Accessible Interaction for Visually Impaired People (AI4VIP)*. Stuttgart, Germany.

[53] Joram Schito. 2012. *Effizienzanalyse der akustischen Wahrnehmung einer Parameter Mapping Sonification eines digitalen Höhenmodells durch interaktive Datenexploration: Masterarbeit über die Verwendung von Sonifikation in der GIScience*. Master's thesis. https://doi.org/10.13140/RG.2.1.4316.2326

[54] Eldon Schoop, James Smith, and Bjoern Hartmann. 2018. HindSight: Enhancing Spatial Awareness by Sonifying Detected Objects in Real-Time 360-Degree Video. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 18)*. Association for Computing Machinery, New York, NY, USA, Article Paper 143, 12 pages. https://doi.org/10.1145/3173574.3173717

[55] B. Shneiderman, C. Plaisant, M. Cohen, S. Jacobs, N. Elmqvist, and N. Diakopoulos. 2018. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson Education.

[56] Thomas Strothotte, Steffi Fritz, Rainer Michel, Andreas Raab, Helen Petrie, Valerie Johnson, Lars Reichert, and Axel Schalt. 1996. Development of Dialogue Systems for a Mobility Aid for Blind People: Initial Design and Usability Testing. In *Proceedings of the Second Annual ACM Conference on Assistive Technologies (Assets '96)*. Association for Computing Machinery, New York, NY, USA, 139–144. https://doi.org/10.1145/228347.228369

[57] Titus J. J. Tang and Wai Ho Li. 2014. An Assistive EyeWear Prototype That Interactively Converts 3D Object Locations into Spatial Audio. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers (ISWC '14)*. Association for Computing Machinery, New York, NY, USA, 119–126. https://doi.org/10.1145/2634317.2634318

[58] Kay Tislar, Zackery Duford, Madeline Peabody Brittany Nelson, and Myounghoon Jeon. 2018. Examining the Learnability of Auditory Displays: Music, Earcons, Spearcons, and Lyricon. In *Proceedings of the 24th International Conference on Auditory Display (ICAD2018)*. 197–202. https://doi.org/10.21785/icad2018.029

[59] Christoph Urbanietz, Gerald Enzner, Alexander Orth, Patrick Kwiatkowski, and Nils Pohl. 2019. A Radar-based Navigation Assistance Device With Binaural Sound Interface for Vision-impaired People. In *Proceedings of the 25th International Conference on Auditory Display (ICAD 2019)*. Department of Computer and Information Sciences, Northumbria University. https://doi.org/10.21785/icad2019.023

[60] Antonio Vasilijevic, Kristian Jambrosic, and Zoran Vukic. 2018. Teleoperated path following and trajectory tracking of unmanned vehicles using spatial auditory guidance system. *Applied Acoustics* 129 (2018), 72 – 85. https://doi.org/10.1016/j.apacoust.2017.07.001

[61] Karen Vines, Chris Hughes, Laura Alexander, Carol Calvert, Chetz Colwell, Hilary Holmes, Claire Kotecki, Kaela Parks, and Victoria Pearson. 2019. Sonification of numerical data for education. *Open Learning: The Journal of Open, Distance and e-Learning* 34, 1 (2019), 19–39. https://doi.org/10.1080/02680513.2018.1553707 arXiv:https://doi.org/10.1080/02680513.2018.1553707

[62] Mark Vollrath, Kathrin Leske, Berhard Friedrich, and Steffen Axer. 2015. Innerstädtische Mobilitätsunterstützung für Blinde und Sehbehinderte: InMoBS: Schlussbericht: Teilvorhaben Technische Universität Braunschweig. Retrieved August 8, 2020 from https://doi.org/10.2314/GBV:856433810

[63] Bruce Walker and Lisa Mauney. 2010. Universal Design of Auditory Graphs: A Comparison of Sonification Mappings for Visually Impaired and Sighted Listeners. *ACM Transactions on Accessible Computing (TACCESS)* 2 (03 2010), 12. https://doi.org/10.1145/1714458.1714459

[64] Michele A. Williams, Amy Hurst, and Shaun K. Kane. 2013. "Pray before You Step out": Describing Personal and Situational Blind Navigation Behaviors. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2513383.2513449

[65] J. Wilson, B. N. Walker, J. Lindsay, C. Cambias, and F. Dellaert. 2007. SWAN: System for Wearable Audio Navigation. In *2007 11th IEEE International Symposium on Wearable Computers*. 91–98. https://doi.org/10.1109/ISWC.2007.4373786

[66] Limin Zeng, Markus Simros, and Gerhard Weber. 2017. Camera-Based Mobile Electronic Travel Aids Support for Cognitive Mapping of Unknown Spaces. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3098279.3098563

[67] Tim Ziemer and Holger Schultheis. 2018. Psychoacoustic auditory display for navigation: an auditory assistance system for spatial orientation tasks. *Journal on Multimodal User Interfaces* 13 (11 2018), 205–218.