# Towards Large Scale Evaluation of Novel Sonification Techniques for Non Visual Shape Exploration

Andrea Gerino
Dept. of Computer Science
Università degli Studi di Milano
EveryWare Technologies
andrea.gerino@unimi.it

Lorenzo Picinali
School of Design Engineering
Imperial College London
l.picinali@imperial.ac.uk

Cristian Bernareggi
Dept. of Computer Science
Università degli Studi di Milano
EveryWare Technologies
cristian.bernareggi@unimi.it

Nicolò Alabastro
Dept. of Computer Science
Università degli Studi di Milano
nicolo.alabastro
@studenti.unimi.it

Sergio Mascetti
Dept. of Computer Science
Università degli Studi di Milano
EveryWare Technologies
sergio.mascetti@unimi.it

## ABSTRACT

There are several situations in which a person with visual impairment or blindness needs to extract information from an image. Examples include everyday activities, like reading a map, as well as educational activities, like exercises to develop visuospatial skills.

In this contribution we propose a set of 6 sonification techniques to recognize simple shapes on touchscreen devices. The effectiveness of these sonification techniques is evaluated though *Invisible Puzzle*, a mobile application that makes it possible to conduct non-supervised evaluation sessions. *Invisible Puzzle* adopts a gamification approach and is a preliminary step in the development of a complete game that will make it possible to conduct a large scale evaluation with hundreds or thousands of blind users.

With *Invisible Puzzle* we conducted 131 tests with sighted subjects and 18 tests with subjects with blindness. All subjects involved in the process successfully completed the evaluation session, with high engagement, hence showing the effectiveness of the evaluation procedure. Results give interesting insights on the differences among the sonification techniques and, most importantly, show that, after a short training, subjects are able to identify many different shapes.

## Categories and Subject Descriptors

H.5.2 [**INFORMATION INTERFACES AND PRESENTATION**]: User Interfaces—*Evaluation/methodology, Auditory (non-speech) feedback, User-centered design, Interaction styles*; K.4.2 [**COMPUTERS AND SOCIETY**]: Assistive technologies for persons with disabilities


*ASSETS'15,* October 26–28, 2015, Lisbon, Portugal.

ACM 978-1-4503-3400-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2700648.2809848.

## Keywords

Accessibility; Visual impairment or blindness; Sonification; Image recognition; Gamification; Remote evaluation; Touchscreen devices

## 1. INTRODUCTION

Access to images is a challenge for people with visual impairment or blindness and this causes hindrances in developing visuospatial skills as well as in education, for example to study STEM subjects. People that can rely on residual sight generally use magnifiers in the form of a physical lens or an electronic device. A different approach is needed by people with blindness or people that cannot rely on residual sight[1].

To overcome these difficulties, one common approach consists in transferring the image on a physical support that can be haptically perceived. Examples include tactile drawings (either printed with special printers or hand-drawn) and the arithmetic board. These solutions have the main advantage of being of immediate use and conceptually simple but also have several drawbacks. First, in many cases these supports need to be manually created (e.g., a tactile drawing) or they need to be printed with expensive hardware. Second, physical supports are bulky and this is a limitation, for example, when they are used for educational purposes. Third, in order to be clearly readable, these physical supports cannot contain too many details, including text and other symbols.

In order to overcome these limitations, a different approach consists in conveying the image information through the acoustic channel by sonification. Several solutions adopting this approach have been proposed in the scientific literature. Some of them do not involve an interacting system or are based on an interaction mediated by a keyboard. In both cases, a blind person cannot rely on proprioception to support the image understanding process. Vice versa, other solutions are based on touchscreen devices that allow users to explore the image with their fingers, hence taking advantage from proprioception. As we show in this contribution,

[1]In this contribution we refer to "people with blindness" or "blind people" to also indicate people with low vision that cannot rely on residual sight to explore enlarged images.

these solutions, which we call *sonification modes*, can adopt different combinations of exploration paradigms, audio rendering techniques and sound generators. Given the large variety of possible solutions, a challenging task is to compare them to evaluate advantages and drawbacks.

In this contribution we propose a set of 6 sonification modes that combine elements of existing techniques with novel solutions, such as sound spatialisation (employing both interaural level and time differences) and music equalisation filtering. The proposed solutions have been designed to sonify binary images representing shapes and to be effective also with a short training.

This contribution also shows the application of a methodology to quantitatively compare these sonification modes through large scale experiments. The core idea of the methodology is that tests should be totally automated, in the sense that no intervention by the supervisor should be needed. This is achieved through *Invisible Puzzle*, an iPhone application that instructs the user on how to use one sonification mode, challenge him/her to recognize some (hidden) shapes and measures the performances. *Invisible Puzzle* adopts a gamification approach that engages test subjects and allows them to quickly get proficient with the sonification modes.

Our ultimate goal is to publish a complete and publicly accessible game that makes it possible to conduct an evaluation of sonification techniques with hundreds or thousands of subjects with visual impairments. In its current version *Invisible Puzzle* is a fully functional prototype aimed at evaluating the effectiveness of the sonification modes as well as challenges in the design and architecture. *Invisible Puzzle* presents 16 tasks to the user; In each task the user has to recognize a shape through sonification and distinguish it from other three. In order to evaluate the proposed sonification modes, as well as the scalability of the proposed methodology, we conducted fully-automated (no supervisor intervention) tests with 131 sighted people and 18 blind people. Results show that after a short training, in most cases, subjects correctly recognize a shape after only a few seconds of active exploration. Furthermore, thanks to the increasing difficulty of the presented tasks, *Invisible Puzzle* can effectively guide users through the sonification learning process, allowing them to gradually become proficient.

This paper is organized as follows. Section 2 describes the related work. The 6 proposed sonification modes are presented in Section 3, while Section 4 describes the proposed evaluation methodology as well as the user-centered design process of *Invisible Puzzle*. Section 5 presents the experimental results. Finally, Section 6 concludes the paper and identifies future work.

## 2. RELATED WORK

The problem of non-visual exploration of images has been widely studied in the scientific literature. The solutions proposed can be broadly divided in two groups: tactile or haptic representations and image sonification. We now focus on the latter group with particular attention to the techniques that can be adopted to recognize shapes on touchscreen devices.

Sanz et al. [10] present a survey of sonification systems used to represent visual scenes. The sonification techniques are categorized according to the sonification function (proportional or derivative [8]), the channel (monaural or binaural) and the paradigm adopted to convey visual information (psychoacustic, artificial or mixed).

Another review on image sonification methods and studies has been carried out by Sarkar et al. [11]. Several algorithms have been reviewed and categorised considering two macrocategories: *Transfer Function Generation* (image-to-sound mapping model) and *Rendering Auditory Data* (how to render the auditory data from the auditory representation of the image).

Yeo and Berger [14] created a framework for designing image sonification methods, categorising various aspects of the sonification process, such as time and spatial exploration, image analysis and sonification mapping. The authors point out the difference between two methods for organizing data for auditory display: *Scanning* and *Probing*. In the *Scanning* method, image data is scheduled to be sonified in a predefined order. Differently, in the *Probing* method, the user can inteactively change the portion of the image to be sonified.

Dallas and Erickson [2] propose a technique to convert an image into sound by mapping the vertical position of each pixel to frequency, the horizontal position to time and brightness to loudness. For example, an image can be sonified by scanning it from left to right. At each instant the sound represents the current vertical portion of the image. This technique is adopted in the vOICe project [7] that aims at enabling sight impaired persons to explore frames captured through a camera.

Yoshida [15] investigates a method for exploring images on a touchscreen device through sound. Two sonification modes are designed: local area sonification and distance-to-edge sonification. In the first mode, when the person slides the finger over an edge, a sound representing the line is played. In the second mode, a pulse train signal is used to represent the finger's distance to the closest edge.

Su et al. [12] developed TimbreMap, namely an iPhone application aimed at enabling people with sight impairments to explore maps through sonification. TimbreMap presents two sonification modes: line hinting and area hinting. The former mode produces a sound while the user is following a path segment. If the user's finger wanders away from the path segment, a stereo sound cue is played to guide the finger towards the nearest path segment. The latter mode fills with sound different areas surrounded by paths. This mode was introduced to convey global information about the map.

Taibbi et al. [13] propose *AudioFunctions*, an iPad application to support visually impaired students in exploring function graphs. *AudioFunctions* adopts three sonification techniques to convey information about a function graph. The "non-interactive" technique sonifies the whole function with an approach similar to the one proposed by Meijer. The "mono-dimensional interactive" technique is similar but allows the user to choose the vertical portion of the image to be sonified by dragging a finger along the horizontal bar that represents the $x$ axis. Finally, the "bi-dimensional interactive" technique is similar to Yoshida's "local area sonification".

## 3. SONIFICATION MODES

Several sonification modes have been developed, all based on a parameter mapping approach [5]. The sonified parameter is the luminance of a specific area of the image. Each sonification mode is characterized by three main components:

- *Exploration paradigm* - defines how the image can be explored, and which portion is considered for sonification given the position of the finger on the screen. Two exploration paradigms are defined, both based on a *probing* approach as defined by Yeo and Berger [14].
- *Audio rendering technique* - transforms the selected image portion into higher level information.
- *Sound generator* - finally generates the sound signals.

In this contribution we describe a single sonification mode based on the *bi-dimensional* (2-D) exploration paradigm (Section 3.1). A larger number of sonification modes (5) were developed for the *uni-dimensional* (1-D) exploration paradigm (Section 3.2). An example of each sonification mode is available online[2].

## 3.1 2-D exploration

We designed the 2-D exploration paradigm to provide a benchmark paradigm likely to be intuitive for the user, as it mimics how a person with visual impairment or blindness explores drawings on swell paper. A similar solution is defined as "bi-dimensional interactive" by Taibbi et al. [13] and "local area notification" by Yoshida et al. [15]. 2-D exploration allows the user to touch the screen and to sonify the specific pixel touched in that moment[3]. The rendered sound therefore depends on both the horizontal and the vertical position of the finger on the screen, hence the name *bi-dimensional (2-D) exploration*.

Considering that the parameter to sonify is solely the luminance of a single point, no further data transformation is required for this particular exploration mode. The luminance value of the single pixel is therefore rescaled within the audio rendering stage, sent to the sound generator and associated with the frequency of a sinusoidal wave with fixed amplitude. The frequency range goes from 100 Hz for the lowest luminance value (i.e., black colour), and 1000 Hz for the highest (i.e., white colour).

## 3.2 1-D exploration

The 2-D exploration paradigm requires users to explore the image along two dimensions to perceive all image features. Our intuition, also confirmed by experimental results (see Section 5), is that 2D exploration is time consuming.

1-D exploration paradigm was designed to tackle this issue by allowing the user to explore the image by touching the screen and moving their finger up and down on a single dimension only. The sonified portion of the image does not correspond only to the touched pixel (as for the 2-D exploration), but to the whole horizontal line (*flush line*) at the same height of the touch point. The horizontal (left-right) position of the finger on the screen is not relevant for this particular exploration paradigm.

1-D exploration paradigm is similar to the technique proposed by Dallas [2], as it simultaneously represents all image features on the flush line. However there are two main differences: firstly, our solution is interactive as it adopts a *probing* approach, while the one by Dallas adopts a *scanning* approach. Secondly, while the solution proposed by Dallas sonifies image features along a vertical line, our approach sonifies an horizontal line. This choice was driven by the fact that we use audio spatialization (based on both interaural time and level differences) to convey additional information about the explored images. In fact, our intuition is that it is more natural for the user to associate left-right spatialized audio information to graphical features on an horzontal line.

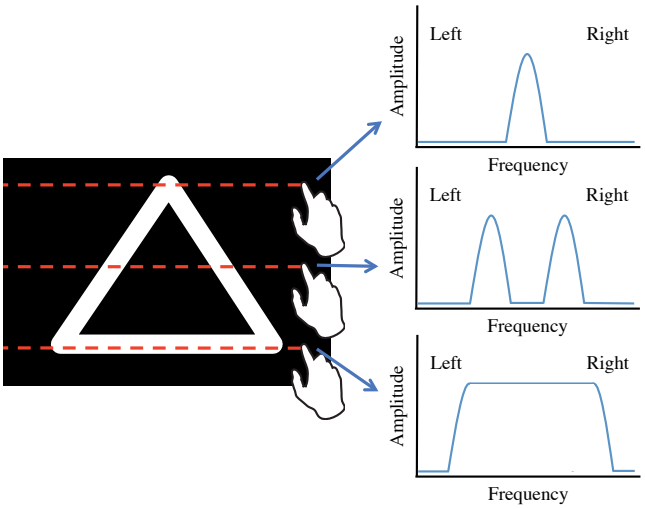


Figure 1: Schematic overview of the 1-D interaction and sonification modes.

In general terms, the luminance of the pixels on the *flush line* is rendered generating a low-frequency sound for pixels located on the left part of the screen, gradually changing to high frequency for pixels located on the right part of the screen. Furthermore, sounds generated from pixels on the left part of the screen are spatialised on the left, gradually changing to the centre and the right for pixels on other parts of the *flush line*. A schematic representation of the 1-D interaction and sonification modes can be found in Figure 1.

It is important to underline that all the sounds corresponding to a single *flush line* are reproduced at the same time, not sequentially moving on the line from left to right. In order to allow for a clear discrimination between concurrent low and high frequency sounds, the spatialisation was implemented using both Interaural Level Differences (ILD) and Interaural Time Differences (ITD). The ILD range was set to a maximum of 20 dB for left-right position, linearly scaled down to 0 dB for the centre position. Similarly, the ITD range was set to a maximum of 1 ms. These values are consistent with spatial hearing literature [9].

The high-low frequency and left-right spatialisation mapping were developed to be as intuitive as possible, taking inspiration from the keyboard of a piano, with the low frequency notes on the left and the high frequency notes on the right. The frequency ranges utilised for the sonification are consistent with the equal loudness curve, therefore with the frequency range for which the human hearing has enhanced loudness sensitivity [9].

Two different audio rendering techniques have been developed using this exploration paradigm, namely *Variable Frequency* (VF) and *Fixed Frequency* (FF).

### 3.2.1 1-D Variable Frequency

[2] `http://webmind.di.unimi.it/assetsip15/`

[3] Clearly, the fingertip touches more than a pixel, however we rely on the mobile OS function that identifies a single pixel that intuitively represents the center of the touch.

The 1-D Variable Frequency audio rendering technique has been designed in order to represent image features on the flush line at the highest possibile resolution (i.e. there is a continous mapping between the x coordinate of each pixel on the flush line and the generated sound's frequency and spatialisation parameters). A luminance threshold is established. Each pixel on the *flush line* with luminance above this threshold value is sonified with a sound generator, whose frequency is associated with the horizontal position of the correspondent pixel. The frequency range is scaled between 100 Hz for the first pixel on the left, and 1618 Hz for the last pixel on the right. Furthermore, low frequency sounds are spatialised on the left, gradually moving towards the right for high frequency sounds, as described previously.

With this particular audio rendering technique, a horizontal white line corresponds to a number of sounds equal to the line length in pixels (possibly a few hundreds). This could potentially create issues in terms of saturation of the output audio channel and, more importantly, it creates a redundancy of information. The ability of the human hearing system to discriminate between several sounds at different frequencies is in fact ultimately limited.

To address this problem, a minimum distance is established between sonified pixels. This is achieved as follows. Starting from the left side of the screen, when a pixel is found with luminance above the threshold, a few following pixels are not sonified regardless of their luminance. The number of these pixels is determined in order to allow for a maximum of 24 concurrent sounds. This value was established considering the maximum sensitivity of the human hearing system in terms of frequency bands detection (the Bark scale [16]).

Two sound generators are usually employed in literature for image sonification: pure tone and noise. We implemented both of them.

- **1-D VF Pure**. The sound generators produce pure sinusoidal sounds.
- **1-D VF Noise**. The sound generators produce narrow-band noise (1/3 octave band width).

### 3.2.2 1-D Fixed Frequency

Due to the particular features of the audio rendering process (i.e., variable frequency and fixed amplitude), both 1-D VF sonification modes allow for smooth frequency changes when exploring an image. However, several sounds with very similar frequencies might be present at the same time and at the same amplitude, creating comb filters and phasing, which are perceived as a marked *vibrato* effect. Our intuition is that this *vibrato* effect could be unpleasant for users.

To tackle this problem we designed the 1-D Fixed Frequency audio rendering technique that generates sounds at 24 predefined frequencies. The *flush line* is divided into 24 equally-sized sectors. The average luminance of each sector is directly sonified modifying the amplitude of 24 sound generators, each reproducing continuously a signal at a fixed frequency, from 100 Hz for the generator correspondent with the sector at the extreme left of the screen, to 1440 Hz for the generator correspondent with the band at the extreme right. The number of sectors is consistent with the number of concurrent sound generators described in Section 3.2.1.

For example, the amplitude of a generator correspondent with a sector in which there are only black pixels is 0, gradually scaled up to 1 (maximum) for a sector in which there are only white pixels. The sound of each generator is spatialised from the left to the right, considering the horizontal position of the associated sector.

In 1-D FF we use the two sound generators already defined for 1-D VF. However, considering the level of annoyance generated by listening for long period of times to pure tones and random-generated noise, we decided to also introduce a third solution which employs a music signal as sound generator. This option is expected to be more enjoyable.

- **1-D FF Pure**. The sound generators produce pure sinusoidal sounds.
- **1-D FF Noise**. The sound generators produce narrow-band noise (1/3 octave band width).
- **1-D FF Music**. Instead of 24 sound generators, one for each sector, a single sound generator consisting of a music track player is used. The sound is split into 24 bands, each with centre frequencies going from 100 Hz to 1440 Hz in 1/3 octave bands. The average luminance of each sector on the *flush line* is associated with the amplitude of the corresponding band (left to right, low to high frequency, left to right spatialisation). As an example, if the luminance of a sector on the right of the *flush line* is high, and for all other sectors is low, then only some high frequency components of the actual music will be audible, spatialised on the right. The track chosen for the playback is an 8-seconds extract from a pop song, continuously looped as soon as the user keeps the finger on the screen. Thanks to the presence of a drum-kit and of several tuned instruments (no voice), the frequency spectrum is rather broad (40 Hz to 20 kHz).

Due to the particular features of the audio rendering process (i.e., fixed frequency and variable amplitude), the 1-D FF Noise and 1-D FF Pure sonification modes generates stepped frequency variations without the *vibrato* effect perceivable in the VF modes. Furthermore, the variable amplitude, associated with the pixel luminance, allows for a higher compatibility with grayscale images if compared with the threshold rendering technique adopted for the VF modes. We leave the evaluation of this sonification mode with grayscale images as a future work.

## 4. EVALUATION METHOD

The overall goal of the experimental evaluation is to understand how effectively each sonification mode allows subjects to perceive images. To achieve this goal, we measure how precisely subjects can recognize images through sonification and how large is the effort involved in this process. We also associate these measures with users characteristics, like age, experience with games, etc.

The challenge is to design a scalable evaluation procedure that makes it possible to involve a large number of subjects with a limited management effort. Our approach is to automate the evaluation procedure in such a way that it can be administered without supervision, possibly remotely. The automated procedure first trains the subject to use the sonification. Then, it instructs him/her to complete some tasks and collects quantitative usage data. Finally, the procedure also administers a questionnaire.

In our preliminary experiments we observed a problem with the automated evaluation: in absence of a supervisor, subjects were less motivated to concentrate on the tasks and to complete the evaluation procedure. This poses an additional challenge: the evaluation procedure should engage the subjects, so that they are encouraged to devote an effort in completing it. This requires the evaluation procedure to be carefully tuned in order to avoid being boring (e.g., tasks should not be too easy) or frustrating (e.g., tasks should not be too difficult).

## 4.1 Invisible Puzzle

In order to conduct the experimental evaluation tackling the aforementioned objectives, we designed the *Invisible Puzzle* application. Its main functionality is to challenge the user to complete tasks, each one consisting in the recognition of a shape. The interaction works as follows: in the *exploration view* the user explores a (hidden) shape through sonification. When the user believes to have recognized the shape, he/she enters a double tap gesture. *Invisible Puzzle* then presents the answer view (see Figure 2) that contains the explored shape together with other three ones, in random order. The user has to identify the explored shape by touching it. For blind users, the four possible choices are described with a text-to-speech synthesizer.

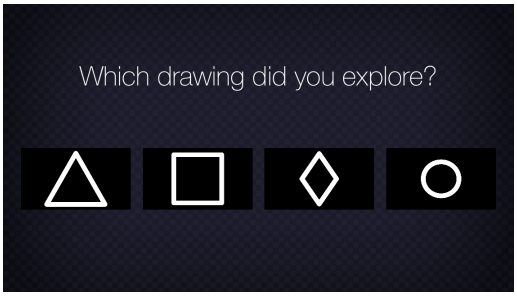


Figure 2: An answer view.

*Invisible Puzzle* performs an instrumented remote evaluation [4] automatically collecting, for each task, a number of metrics that include the answer and the *exploration time* i.e., the time during which the user actually touches the screen in the exploration view.

For a given subject, *Invisible Puzzle* conducts the entire evaluation procedure using a single sonification mode. There are two motivations for this choice. Firstly, from the methodological point of view we want to exclude that the results collected for one task are biased by the fact that the user has previously experienced a different sonification mode. Secondly, from the user point of view, it is simpler to learn how to explore shapes with a single sonification mode.

Regarding subjects' training, the goal is to let the user quickly grasp the basics of the sonification mode without requiring supervisor intervention or a long explanation. At the beginning of the evaluation procedure, *Invisible Puzzle* introduces the user to the challenge (i.e., recognize hidden images) and to the basics of the interaction paradigm by playing a very short video (15 seconds) that also includes a spoken explanation. After the video, a "learning task" is presented in which the shape is visible (a textual description of the shape is also provided for blind users). Users can freely explore the image, hence getting acquainted with the sonification mode.

As we show in Section 5, after these two training steps a large portion of users (up to 50% for some sonification modes) is still unable to recognize a simple shape (i.e., a dot on the center right of the screen). In order to offer further training, *Invisible Puzzle* adopts two solutions. First, upon a wrong answer is given, *Invisible Puzzle* displays the correct answer and asks the user to repeat the task in order to precisely associate the audio feedback with the known shape. Second, following a gamification approach [3], exploration tasks are designed to gradually increase in difficulty and are organized in four groups, each one consisting of four tasks. The first three groups focus on a particular type of shape each (i.e., points, line segments and polygons, respectively). A "learning task" is presented at the beginning of each of the first three groups, but not at the beginning of the fourth. The shapes in the fourth group include elements from previous ones. This is intuitively more challenging as the user does not know what to expect (e.g., a line segment or a polygon). We recorded two videos, for 1-D and 2-D exploration paradigms, showing how users are introduced to *Invisible Puzzle*[2].

The increasing difficulty of the tasks implies that it is possible to gradually introduce new elements. For example, the first two tasks in the point group present a single point, while the last two present two points each and, in particular, the last task presents two points on the same *flush line*. Thanks to this approach it is possible to avoid that the user faces tasks that are too challenging and hence frustrating. At the same time, by finely tuning the increasing difficulty of the tasks, it is possible to presents tasks that are not too easy hence engaging the user. This is in line with the "Flow theory" from Csikszentmihalyi [1].

At the end of the evaluation procedure, *Invisible Puzzle* automatically administers a questionnaire to the user. The questionnaire is made of a first section of questions about the user (age, experience with music and video games), a second section with a 5-point Likert scale composed by 9 items and, finally, an open text section for comments. The Likert items investigate the users' experience with the application (e.g., whether they enjoyed playing). Table 1 reports the list of Likert items that we take into account in Section 5.

| |
|---|
| **Q1**. I enjoyed playing with *Invisible Puzzle* |
| **Q2**. I would like to play with *Invisible Puzzle* again |
| **Q3**. Playing with *Invisible Puzzle* required me a lot of concentration |
| **Q4**. I have found the sounds comforting/pleasant |

Table 1: Some of the Likert items in the questionnaire.

## 4.2 The design of Invisible Puzzle

We designed *Invisible Puzzle* with a user centered approach through a series of iterations. The whole process benefitted from feedback from many subjects, including one of the designers who is blind and other three blind people. From the interaction design point of view, the most challenging part is to introduce the user to two basic activities:

to explore the screen and to terminate the exploration (with a double tap).

In its first version, *Invisible Puzzle* presented a textual explanation of these activities. However, we observed that most users did not understand the principle of the image exploration.

To tackle this problem in the second version we added, after the textual description, the "learning task". However we observed that many sighted users, in particular with the 2-D sonification mode, tended to tap on the visible dots rather than to slide the finger on the screen; consequently in the following task, with an hidden shape, users did not know how to interact.

To address this problem, we substituted the initial textual explanation with a video that includes a speech-based description. This description is more detailed than the purely textual one and includes the explanation of the challenge (i.e., to recognize hidden shapes). With this solution, most users were able to use the 2-D sonification mode, but still some users had problems with the sonification modes based on the 1-D exploration paradigm. The reason is that users did not understand that the image was sonified in its full width independently of the horizontal coordinate of the point being touched.

To address this problem, we changed *Invisible Puzzle* so that, when using the 1-D exploration paradigm, it forces the user to slide the finger over a small column on the right edge of the screen (all touches outside this column have no effect) and it visually represents the *flush line*. However we observed one additional problem: some users did not understand how to terminate a task and hence remained stuck on the first one. For this reason in its current version, if the user stops touching the screen for a given time, *Invisible Puzzle* shows a text (and reads it for blind users) reminding how to complete the task.

Thanks to this design process we developed the introductory part of *Invisible Puzzle* that allowed all users involved in the experiments to complete the tasks without the need of an external explanation.

# 5. EVALUATION RESULTS

## 5.1 Experimental setting

In order to get statistically significant results, *Invisible Puzzle* was tested by a considerable amount of subjects. Considering the difficulties in recruiting individuals with visual impairment, we decided to carry out the test also on individuals without visual disabilities. Results from the two groups can help obtaining statistically significant data to guide future development and experimental studies for both visually impaired and sighted subjects. Previous studies have followed a similar approach [6]. Table 2 and Table 3 report the number of sighted and blind subjects involved in the tests.

All tests were conducted on iPhone 5 and iPhone 5$S$ devices (that have the same screen size) and during the tests subjects wore Apple *EarPod* headphones. Tests with sighted users were conducted both in Italy and in the UK, while tests with blind users were conducted in Italy only[4]. Tests were conducted in various environments with limited ambient noise, including students' library, laboratories and also the cafeteria.

Tests were not supervised, in the sense that subjects were asked to conduct a test and no other information was provided, excluded the expected duration of the test (i.e., approximately 15 minutes).

## 5.2 Results with sighted subjects

Table 2 reports mean values and standard deviations ($\sigma$) for all the measured parameters and values. We first consider the percentage of correct answers, which are displayed in the boxplot in Figure 3. This value represents the percentage of tasks in which subjects gave the correct answer in the first attempt. Its dual value is what we consider the "error rate" i.e., the number of failed tasks (this value can be easily derived from the percentage of correct answers). We can observe that the mean performance with the 2-D sonification mode is slightly better than with the 1-D. Considering that the data sets are normally distributed, a one-way ANOVA analysis was conducted. The results show that there are no statistically significant differences between the six sonification modes $[F(5, 125) = 1.005, p = 0.418]$.

Interestingly, there are statistically significant differences $[F(1, 129) = 12.502, p = 0.001]$ between subjects that do not play musical instruments (67% correct, $\sigma = 13.38$) and those who do (75% correct, $\sigma = 12.79$). On the other hand, no statistically significant difference can be found between individuals who play computer games (one or more hours per day and one hour per week) and individuals who don't (less than one hour per week) $[F(3, 127) = 1.117, p = 0.345]$.

The performances with the various sonification modes exhibit clearer differences in terms of exploration time, as shown in Figure 4. Indeed 2-D and 1-D FF Music require a longer exploration time with respect to the other four sonification modes. One-way ANOVA analysis reveals that the differences between the six sonification modes are statistically significant $[F(5, 2090) = 88.887, p = 0.000]$. As expected, post-hoc Tukey analysis highlights that the significant differences are between 2-D and 1-VF Pure ($p = 0.000$), 1-VF Noise ($p = 0.000$), 1-FF Pure ($p = 0.000$) and 1-FF Noise ($p = 0.000$); and between 1-D FF Music and 1-VF Pure ($p = 0.000$), 1-VF Noise ($p = 0.000$), 1-FF Pure ($p = 0.000$) and 1-FF Noise ($p = 0.000$). It is important to underline that in this case the inferential analysis was carried out considering the time taken to complete every single trial, and not the average trial time for each user.

Considering again exploration time, statistically significant differences $[F(1, 2094) = 3.850, p = 0.05]$ are found between subjects that do not play musical instruments (mean exploration time 15.61s per task, $\sigma = 11.42$) and those who do (mean exploration time 14.61s per task, $\sigma = 11.93$). There are also statistically significant differences $[F(3, 2092) = 13.147, p = 0.000]$ between the computer games playtime groups. A post-hoc Tukey analysis reveals that the significant differences are between the individuals who play computer games more than 1 hour per day and all the other groups, therefore 1 hour per day ($p = 0.001$), 1 hour per week ($p = 0.004$), and less than 1 hour per week ($p = 0.000$).

Considering data collected through the questionnaire, other interesting aspects emerge (see Table 2 again). Firstly, higher concentration (mean 4.43, SD 0.67) is required for the 1-D FF Music sonification mode, if compared with sonification modes adopting the pure sound generators (mean 3.71, $\sigma =$

[4] *Invisible Puzzle* is localized in Italian and English.

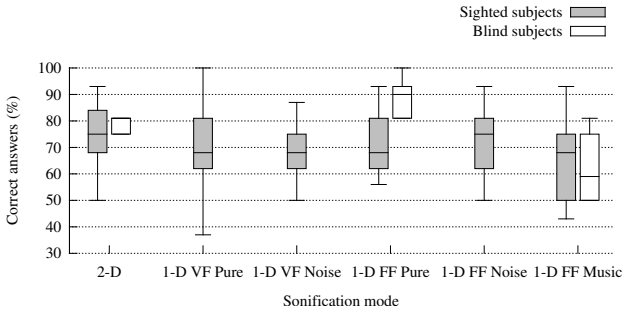


Figure 3: Correct answers for each sonification mode.

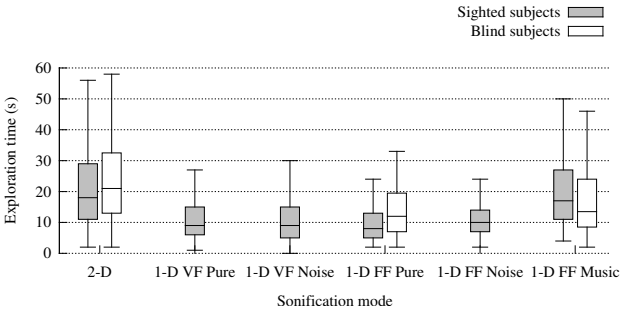


Figure 4: Mean exploration times for each sonification.

| Sonification | Participants | Correct answers | | Exploration time (s) | | Q1 (enjoy) | | Q2 (play again) | | Q3 (concentration) | | Q4 (pleasantness) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | % | $\sigma$ | m | $\sigma$ | m | $\sigma$ | m | $\sigma$ | m | $\sigma$ | m | $\sigma$ |
| 2-D | 24 | 75.0 | 13.0 | 22.1 | 14.1 | 3.4 | 0.7 | 3.1 | 1.0 | 3.8 | 0.9 | 2.2 | 1.0 |
| 1-D VF Pure | 22 | 69.6 | 15.7 | 11.6 | 7.3 | 3.9 | 0.7 | 3.5 | 0.9 | 3.8 | 1.0 | 2.8 | 1.1 |
| 1-D VF Noise | 21 | 67.8 | 11.9 | 11.6 | 7.9 | 3.9 | 0.8 | 3.4 | 1.1 | 3.9 | 0.9 | 3.1 | 0.9 |
| 1-D FF Pure | 22 | 71.3 | 10.8 | 10.6 | 8.3 | 4.4 | 0.6 | 4.1 | 0.5 | 3.4 | 0.9 | 3.7 | 0.9 |
| 1-D FF Noise | 21 | 73.5 | 12.9 | 12.5 | 9.0 | 3.8 | 1.1 | 3.9 | 1.1 | 4.4 | 0.8 | 3.2 | 1.0 |
| 1-D FF Music | 21 | 68.1 | 16.7 | 21.4 | 14.1 | 4.0 | 0.7 | 3.3 | 1.1 | 4.4 | 0.6 | 4.3 | 0.8 |
| Total | 131 | 70.9 | 13.6 | 15.1 | 11.6 | 3.9 | 0.8 | 3.6 | 1.0 | 3.9 | 0.9 | 3.2 | 1.1 |

Table 2: Results with sighted subjects ("m" = mean).

0.97) and the noise sound generators (mean 4.19, $\sigma = 0.89$). One-way ANOVA analysis reveals that these differences are statistically significant $[F(2, 128) = 6.776, p = 0.002]$, and post-hoc Tukey analysis highlights significant differences between pure tone and noise ($p = 0.005$) and pure tone and music ($p = 0.02$).

The least enjoyable sonification mode is 2-D, while the most enjoyable is 1-D FF Pure. One-way ANOVA analysis reveals that differences are statistically significant $[F(5, 125) = 3.241, p = 0.009]$, and post-hoc Tukey analysis highlights significant differences between 2-D and 1-D FF Pure ($p = 0.002$).

Similar results hold considering how subjects declare to be interested in playing again with *Invisible Puzzle*. Indeed, subjects are less interested in playing with 2-D, while 1-D FF Pure is the sonification mode that raised the highest interest.

Finally, as expected, music is the most pleasant sound generator (mean 4.38, SD .805), followed by noise (mean $3, 17$, $\sigma = 0.986$) and pure (mean 2.94, $\sigma = 1, 20$). One-way ANOVA analysis reveals that differences are statistically significant $[F(2, 128) = 14.298, p = 0.000]$, and post-hoc Tukey analysis highlights significant differences between music and noise ($p = 0.000$) and between music and pure ($p = 0.000$).

## 5.3 Results with blind subjects

In the evaluation with blind subjects we compared only three sonification modes: 2-D, 1-D FF Pure and 1-D FF Music. This is motivated by the fact that, in order to obtain statistically relevant results with the relatively small number of subjects with visual impairments, we preferred to conduct 6 tests for each of the 3 sonification modes listed above, rather than uniformly distributing the tests among the 6 sonification modes. Table 3 reports the results.

Concerning the percentage of correct answers (see boxplot in Figure 3), 1-D FF Pure yields better results than 2-D and 1-D FF Music. Since the dataset is little and not normally distributed, an independent samples Kruskal-Wallis test was conducted, showing statistically significant differences between the sonification modes ($p = 0.045$).

1-D FF Pure yields better results also considering exploration time. In this case the population is normally distributed, therefore one-way ANOVA analysis was used. The differences are statistically significant $[F(2, 285) = 13.583, p = 0.000]$, and Tukey post-hoc shows significant differences between 2-D and 1-D FF Pure ($p = 0.000$) and between 2-D and 1-D FF Music ($p = 0.001$).

Considering the results of the questionnaire, it emerges that 2-D requires more concentration than the other two sonification modes. Also, unexpectedly, 1-D FF Music is the least enjoyable with a large difference with respect to 1-D FF Pure. As a consequence, subjects would be much less interested in playing again using 1-D FF Music rather than 1-D FF Pure. Finally, subjects found the sound of 1-D FF Pure as pleasant as the sound of 1-D FF Music. This was also unexpected.

## 5.4 High-level considerations about experimental results

This contribution focuses on two main aspects: the effectiveness of the sonification modes and the scalability of the evaluation system. For what concerns the sonification modes, we can conclude that, despite the very short training, users can successfully recognize most of the shapes after a few seconds of exploration. Consider for example the third task in the "polygon" group (see Figure 5). Using 1-D exploration paradigm, it is only possible to identify the correct answer by distinguishing two different sounds, one for the left edge (with constant frequency) and the other one for

| Sonification | Participants | Correct answers | | Exploration time (s) | | Q1 (enjoy) | | Q2 (play again) | | Q3 (concentration) | | Q4 (pleasantness) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | % | $\sigma$ | m | $\sigma$ | m | $\sigma$ | m | $\sigma$ | m | $\sigma$ | m | $\sigma$ |
| 2-D | 6 | 77.0 | 7.5 | 25.8 | 18.5 | 3.6 | 1.2 | 3.8 | 0.9 | 4.6 | 0.5 | 3.6 | 1.0 |
| 1-D FF Pure | 6 | 82.2 | 22.8 | 15.2 | 11.5 | 4.5 | 0.8 | 4.3 | 1.0 | 4.3 | 0.5 | 4.5 | 0.5 |
| 1-D FF Music | 6 | 62.5 | 13.11 | 18.0 | 12.9 | 3.5 | 0.8 | 3.3 | 0.8 | 4.0 | 0.8 | 4.5 | 0.8 |
| Total | 18 | 73.9 | 17.1 | 19.7 | 15.2 | 3.8 | 1.0 | 3.8 | 0.9 | 4.3 | 0.6 | 4.2 | 0.8 |

Table 3: Results with blind subjects ("m" = mean).

the other two edges. We expected this to be a challenging task, after the short training allowed to the users, but we were proven wrong, as 86% of the subjects gave the correct answer with the 1-D exploration modes.

(a) Correct (b) Wrong (c) Wrong (d) Wrong

Figure 5: Answers of the third task in the "polygon" group.

In many cases, subjects have not been able to identify the exact shape, but still recognized its main characteristics. Consider the last task, which resulted to be, as expected, the toughest one (see Figure 6). Intuitively, with all proposed sonifications modes it is hard to distinguish the large circle from the pentagon and from the octagon, as they all have the same size. Vice versa, the small circle should be easily ruled out, due to its size. Indeed, only 3% of the subjects gave the small circle as the answer. This is characteristic of the fact that, with the various sonification modes, subjects can easily perceive the dimension of the shape being explored.

It is even more remarkable that, as reported by some subjects, using the FF audio rendering technique, they could perceive discrete levels in the audio frequencies (which is actually a characteristic of FF audio rendering) and hence were induced to give pentagon or octagon as the answer. Indeed, 77% of subjects gave either pentagon or octagon as the answer. While, on one hand, this result highlights a limit of the FF audio rendering technique, on the other hand we were surprised by how quickly subjects got so proficient with the sonification mode to perceive such a fine audio detail.

(a) Correct (b) Wrong (c) Wrong (d) Wrong

Figure 6: Answers of the last task.

Comparing the performance of sighted and visually impaired subjects (see Tables 2 and 3), we can observe that very similar results are obtained by the two classes of users for the 2-D exploration mode in terms of both percentage of correct answers and exploration time. Differently, with 1-D FF Pure visually impaired subjects achieve a higher percentage of correct answers, at the cost of a higher exploration time. On the contrary, with 1-D FF Music, subjects with visual impairments have a lower percentage of correct answers and a lower exploration time. Considering the answers to the Likert Items, there are similar results for blind and sighted subjects. One difference is that blind subjects apparently find that *Invisible Puzzle* is less enjoyable with 1-D FF Music (mean value 3.5) with respect to sighted subjects (mean 4.0).

Considering the scalability of the evaluation system, one challenge was to quickly introduce the user to the exploration paradigm. Thanks to the application of the user-centric approach described in Section 4.2, all subjects successfully completed the evaluation procedure without the need of any external explanation. This does not mean that the introductory explanation (i.e., the video) provides an exhaustive explanation on its own. Rather, it is the combination of the explanation with the following user experience that actually guided subjects in getting proficient with the sonification modes. To support this conclusion, consider that, with the 1-D sonification modes, only 66% of the subjects provided a correct answer in the first task. This mean that, before starting the first task, 1/3 of subjects have not really grasped how the sonification mode worked. However more than 90% of subjects that gave a wrong answer in the first task, after repeating it, gave a correct answer in the second one.

Another challenge involved in the evaluation procedure is to engage the subjects, so that they complete the procedure without distractions. Results show that *Invisible Puzzle* achieve this objective and, in particular, there are some sonification modes that help involving the subjects. We expected 1-D FF Music to be more involving. Instead, it resulted that the music being played requires higher attention and overall the sonification mode that results more enjoyable by both sighted and blind subjects is 1-D FF Pure.

## 6. CONCLUSIONS

This paper presents 6 sonification modes that allow people with blindness to explore shapes in images. The large number of tests, conducted with both sighted and blind subjects, give evidence that the overall technique is effective and that there are some sonification modes that allow a faster and more enjoyable detection. In particular, 1-D FF Pure has the best trade-off among percentage of correct answer, exploration time and general user satisfaction. Thanks to *Invisible Puzzle*, the evaluation process did not require the presence of a supervisor, and this made it possible to collect 149 tests in less than 2 weeks, with a limited managing effort.

As a future work, we intend to further develop *Invisible Puzzle* and to publish it as a video game in on-line stores (e.g., AppleStore). This would make it available to a much broader audience. In particular we expect that, by advertising *Invisible Puzzle* in communities of people with visual

impairments, it will be possible to remotely collect a large amount of evaluation data from sighted and blind people using devices with different screen sizes (e.g., iPad).

*Invisible Puzzle* can be extended along a number of directions. First, it is possible to improve the technique to evaluate how clearly subjects identify the hidden shape, by asking them to draw the shape on the device. Then the drawn shape can be automatically compared to the hidden one. We believe that this solution, possibly coupled with the already adopted multiple choice question, could give more insights on the actual user's understanding of the image. Second, it is interesting to evaluate to which extent more complicated images can be perceived. This also include the use of grayscale images and, by designing new sonification modes, color images. While in theory the three solutions based on 1-D FF can be already used on real-world grayscale images, their effectiveness for this kind of application needs to be carefully evaluated. Third, with a longer training it could be possible to make several sonification modes available to the same user. This allows the user to choose the preferred one and, possibly, to switch among them while exploring a single image in order to capture different aspects of it. Fourth, it is possible to introduce additional features to make the exploration more interactive, like the possibility to zoom in and out in the image or the separation of the image into layers, each one explorable separately from the others.

## 7. ACKNOWLEDGMENTS

Authors would like to thank Ginevra Are Cappiello and Alice Brovelli from the organization "SIT - Social Innovation Teams" for their help in the design of *Invisible Puzzle* and all the volunteers who tested our application providing precious data and valuable feedback.

## 8. REFERENCES

[1] M. Csikszentmihalyi. *Flow*. Harper Perennial Modern Classics. HarperCollins, 2009.

[2] S. A. Dallas Jr and A. J. Erickson. Sound pattern generator, Mar. 29 1983. US Patent 4,378,569.

[3] S. Deterding, M. Sicart, L. Nacke, K. O'Hara, and D. Dixon. Gamification. Using game-design elements in non-gaming contexts. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2011.

[4] H. R. Hartson, J. C. Castillo, J. Kelso, and W. C. Neale. Remote evaluation: the network as an extension of the usability laboratory. In *Proc. of SIGCHI*. ACM, 1996.

[5] T. Hermann and H. Ritter. Listen to your data: Model-based sonification for data analysis. *Advances in intelligent computing and multimedia systems*, 1999.

[6] B. F. Katz and L. Picinali. *Spatial audio applied to research with the blind*. INTECH Open Access Publisher, 2011.

[7] P. B. Meijer. An experimental system for auditory image representations. *Biomedical Engineering, IEEE Transactions on*, 39(2), 1992.

[8] E. Milios, B. Kapralos, A. Kopinska, and S. Stergiopoulos. Sonification of range information for 3-d space perception. *IEEE Neural Systems and Rehabilitation Engineering*, 2003.

[9] B. C. Moore. *An introduction to the psychology of hearing*. Brill, 2012.

[10] P. R. Sanz, B. R. Mezcua, J. M. S. Pena, and B. N. Walker. Scenes and images into sounds: a taxonomy of image sonification methods for mobility applications. *Journal of the Audio Engineering Society*, 62(3), 2014.

[11] R. Sarkar, S. Bakshi, and P. K. Sa. Review on image sonification: a non-visual scene representation. In *Recent Advances in Information Technology (RAIT), 2012 1st International Conference on*, pages 86–90. IEEE, 2012.

[12] J. Su, A. Rosenzweig, A. Goel, E. de Lara, and K. N. Truong. Timbremap: enabling the visually-impaired to use maps on touch-enabled devices. In *Proc. of the 12th Int. Conf. on Human computer interaction with mobile devices and services*. ACM, 2010.

[13] M. Taibbi, C. Bernareggi, A. Gerino, D. Ahmetovic, and S. Mascetti. Audiofunctions: Eyes-free exploration of mathematical functions on tablets. In *Computers Helping People with Special Needs*. Springer, 2014.

[14] W. S. Yeo and J. Berger. A framework for designing image sonification methods. In *Proceedings of International Conference on Auditory Display*, 2005.

[15] T. Yoshida, K. M. Kitani, H. Koike, S. Belongie, and K. Schlei. Edgesonic: image feature sonification for the visually impaired. In *Proc. of the 2nd Augmented Human Int. Conf.* ACM, 2011.

[16] E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The Journal of the Acoustical Society of America*, 2(33), 1961.