Exploring eye-tracking-driven sonification for the visually impaired

Michael Dietz, Maha El Garf, Ionut Damian, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Dietz, Michael, Maha El Garf, Ionut Damian, and Elisabeth André. 2016. "Exploring eye-tracking-driven sonification for the visually impaired." In *Proceedings of the 7th Augmented Human International Conference 2016 on - AH '16, Geneva, Switzerland — February 25 - 27, 2016*, edited by Jean-Marc Seigneur, Jose M. Hernandez-Munoz, and Paul McCullagh, 5. New York, NY: ACM Press. https://doi.org/10.1145/2875194.2875208.

Exploring Eye-Tracking-Driven Sonification for the Visually Impaired

Michael Dietz
Human Centered Multimedia
Augsburg University
Augsburg, Germany
dietz@hcm-lab.de

Maha El Garf
Faculty of Media Engineering
and Technology
German University in Cairo
Cairo, Egypt
maha.elgarf@guc.edu.eg

Ionut Damian
Human Centered Multimedia
Augsburg University
Augsburg, Germany
damian@hcm-lab.de

Elisabeth André
Human Centered Multimedia
Augsburg University
Augsburg, Germany
andre@hcm-lab.de

ABSTRACT

Most existing sonification approaches for the visually impaired restrict the user to the perception of static scenes by performing sequential scans and transformations of visual information to acoustic signals. This takes away the user's freedom to explore the environment and to decide which information is relevant at a given point in time. As a solution, we propose an eye tracking system to allow the user to choose which elements of the field of view should be sonified. More specifically, we enhance the sonification approaches for color, text and facial expressions with eye tracking mechanisms. To find out how visually impaired people might react to such a system we applied a user centered design approach. Finally, we explored the effectiveness of our concept in a user study with seven visually impaired persons. The results show that eye tracking is a very promising input method to control the sonification, but the large variety of visual impairment conditions restricts the applicability of the technology.

CCS Concepts

•**Human-centered computing** → **Auditory feedback;** *User centered design;* •**Hardware** → Signal processing systems; Sound-based input / output;

Keywords

Sonification; Eye Tracking; Visually Impaired; Sound Synthesis; Signal Processing

AH 2016, February 25 - 27, 2016, Geneva, Switzerland

Figure 1: Visually impaired participants during our user study

1. INTRODUCTION

Vision is one of our primary senses to perceive the real world and thus the diagnosis of visual impairment presents a great challenge for the affected people. According to the latest report of the World Health Organization (WHO) there were over 285 million visually impaired people around the world in 2010 [15]. Compared to their previous estimation from 2002, this is an increase of more than 77%. Since the rising number of affected people is not or only partly able to perceive the environment with their eyes, researchers have been inspired for a long time to make the visual world more accessible to them. One of the most commonly adopted notions to deal with this problem is sensory substitution. The idea behind this concept is to transform the stimuli of one sensory modality into another one to compensate for a defect of the initial sensory modality. Probably the most well known application of it is the *Braille* reading method. It aims at replacing the sight sense with the touch sense and gives visually impaired people the ability to read text through tactile feedback. Another example are text to speech converters, which target the replacement of the sight sense with the hearing sense by enabling the ability of visually impaired people to hear text instead of reading it.

In general, converting visual information into sound has been the most popularly used approach among all the methods for sensory substitution. To this end a very promising concept is the automatic generation of semantic descriptions based on image contents, similar to sighted persons explaining blind users what they see. Even though researchers recently achieved some significant progress in this domain [12, 21], there is one considerable drawback to it. The method takes away the direct perceptual experience and the impressions of actively exploring the images from the visually impaired. Besides that, the generated descriptions only give a rough overview of the image contents while details such as the visual appearance of individual objects, their position and color effect are usually not included.

In this work we therefore propose a system which enables blind and visually impaired people to explore and perceive the environment through their remaining senses. More precisely, we transform certain image aspects such as colors, texts and facial expressions from the field of view of the users into acoustic signals, while it is still their task to analyze and interpret them. In order to give the users the ability to decide which information is relevant to them at any point in time, we use the eye movements to control the interactive exploration of the field of view. This enables a perception experience which is similar to that of sighted persons. Finally, we evaluate the feasibility of our concept in a user study with seven blind and visually impaired persons. The study yielded that four out of seven users were able to successfully use eye tracking as an input method for the sonification system. In the remaining three cases, the nature of the medical condition prohibited the detection of the user's pupil on which the eye tracking algorithm relies. These results suggest that eye tracking has a very high potential to be used as an input method for a certain group of visually impaired persons.

2. RELATED WORK

Sonification has been intriguing many researchers in the past decades and one of its most typical tracks, is its use as an alternative to visualization for the visually impaired. Common sonification applications for this user group aim at object recognition such as the vOICe [18]: an application for smart phones which enables users to recognize objects and to locate them in space. This is usually performed through feature extraction like color, shape and texture. Among the approaches to color sonification are the techniques presented in [8] and [2]. In the first, each row is subdivided into 12 segments and color information about each segment is sonified and played back to the user. Similarly, in the latter, each image is processed column by column from left to right emitting a combination of sounds that represents the color information. These approaches that sonify multiple features of the image at once may present a great challenge for the visually impaired. This is because sonification faces the problem of the high number of visual features that can be represented by a visual frame opposed to the inability of a single audio stream to represent an equal amount of characteristics at once. The solution proposed by the authors in [6] and [23] is to use a touch screen. Then, the visually impaired person can explore an image using the tip of his or her fingers and consequently receives audible feedback only about the region underneath his or her fingers. Although we applied an approach based on the method in [6] for color sonification, we believed that the use of touch input in that system restricted the user experience to touchable devices only such as computers, electronic tabs or mobile phones. Since our system is aimed for ubiquitous use, we decided to use eye tracking glasses instead. These are consequently going to sonify the closest area to the point where the eyes of the user are directed in realtime. This way, the person can use the sonification system during normal tasks, such as going shopping or interacting with other persons. A similar approach has been proposed by Twardon et al. [20]. In their work, they use a head mounted eye tracker with a *Microsoft Kinect* attached on top of it, to sonify the distance towards the object at the current gaze point of the user. Although their evaluation yielded some interesting results, it was only done with sighted people and did not investigate, whether the eye tracking device or the applied sonification approach might irritate the users if the system is used for a longer period of time. Besides, their work only focuses on depth sonification, while our system aims at enabling the perception of color, text and facial expression information through acoustic signals.

Several attempts like [3], [4] and [17] have targeted text sonification. The aim of those approaches was to enable the users to have an idea about the intent of a message from the tone that signals the receipt of the message on a mobile phone before actually reading it. Consequently, these applications mainly focused on the intent and mood of the received text message rather than the actual content of it. The intent of the message was analyzed through checks for punctuation characters and emoticons. Text sonification for the visually impaired will have to take a different course though. For a visually impaired person the actual content of the text is of vital importance to give him or her the ability to understand the meaning and implications of it. Therefore, we aimed at transmitting the content of the texts in our sonification approach.

In order to actively engage visually impaired people in the daily communication process, some sonification applications target facial expressions. For example, in [16], the authors have developed a system to detect facial expressions and generate instruments' sounds accordingly. Although the system uses the different facial features for facial expression recognition, the sounds generated always correspond to specific emotions such as happiness, sadness, anger or surprise. However, current state of the art in emotional recognition is not able to provide perfect accuracy, especially in real-time out-of-lab scenarios. Considering this, we investigate and compare the sonfication of low-level facial actions and task the user with interpreting these, as well as higher level emotions extracted from these facial actions automatically using a modern recognizer. Another application introduced in [11] generates an orchestra sound where each instrument represents a feature of the face. The different frequencies of each instrument represent the different actions of the facial features. This means that to recognize a facial expression, the user has to distinguish between four or five sounds with their corresponding frequencies simultaneously. This makes it more difficult for the visually impaired to train on the system and psychologically accept it. Consequently, in our application we only used two important facial features: the mouth and eyebrows in order to simplify the process for the visually impaired users.

It is also worth noting that most of the previously implemented sonification systems did not undergo any user testing [16], or performed a user study on a very small sample of visually impaired users [6]. This calls the viability and the efficiency of the systems for use by the visually impaired into question. In some other applications such as [11], [20] and [23], the system was only tested on normally sighted people. This might also have yielded inaccurate results because normally sighted people may be able to quickly identify an object even when blindfolded based on their previous visual experience. Thus, we performed our case study on seven visually impaired people in order to maximize the accuracy of the results.

3. PARTICIPATORY DESIGN WORKSHOP

In this manuscript we propose a sensory-substitution approach specifically targeted at blind and visually impaired. Considering the special target user group we decided to address, getting user input at an early development stage was a top priority. To this end we made contact with a local association for the blind and visually impaired and conducted a design workshop. One administrative personnel of the association and two visually impaired persons which were also involved in the association took part in the workshop.

The workshop was structured into two sessions. First, we presented our concept using a very basic prototype of the system. The prototype consisted of a simple color sonification demo using a head-mounted camera. The aim of this first session was to give the participants a general impression of the capabilities of sensory substitution systems as well as to gather information regarding the perception of such systems by visually impaired. Furthermore, we discussed possible incompatibilities of medical conditions with eye tracking solutions. The second session consisted of a brainstorming exercise to identify, on one hand, daily activities visually impaired struggle most with and on the other hand, which of those activities could be realistically assisted with sensory-substitution approaches.

The workshop yielded valuable insights. First, all three stakeholders showed great interest in sensory-substitution solutions. However, concerns have been vocalized regarding the visual appearance of the system. According to our stakeholders, many visually impaired fear the social stigma associated with their condition, a reason for which many also refuse to use white canes or other mobility supporting instruments. While this is indeed a valid concern for technology-enhanced sonification systems, one can speculate that with the rapid advancement of wearable devices, development of inconspicuous solutions are only a matter of time. We also learned that a large part of our target user group may develop pathological nystagmus, or more commonly called “dancing eyes”, which causes the user to loose oculomotor control. Because this condition can strongly impact the accuracy of eye tracking systems, we decided to restrict our target group to blind and visually impaired persons which do not suffer from pathological nystagmus.

The second session gave us some clear examples of daily activities visual impaired persons struggle with. More specifically, all participants pointed out activities such as reading text, identifying objects, avoiding obstacles or navigating unknown streets as most encumbering. One of the visually impaired participants also mentioned that help with reading the emotions of others would greatly ease the burden on social interactions. He explained that the loss of the ability to tell when your loved ones are happy or sad can be especially difficult to overcome for recently diagnosed.

4. SONIFICATION SYSTEM

In order to explore the feasibility of eye tracking as input method for blind and visually impaired people, we implemented a sonification system that uses the eye tracking data to control which part of the user’s field of view should be sonified. Having in mind the outcome of the participatory design workshop (Section 3) and considering technical limitations, we decided to focus on the sonification of color, text and facial expression information. We therefore created a sonification component for each of those aspects.

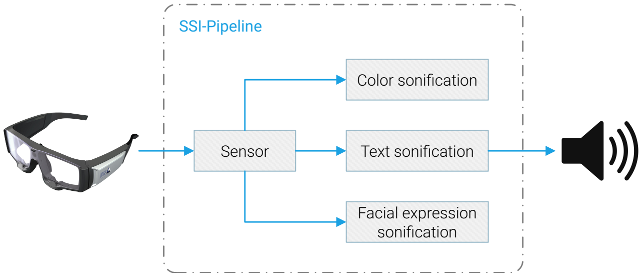

Figure 2: System architecture

The system itself is based on a signal processing pipeline of the Social Signal Interpretation (SSI) Framework [22]. It is used since it provides a high degree of flexibility due to its modular architecture and already contains several open source libraries including OpenCV, ARToolKit and SHORE. Furthermore the framework supports a large variety of sensors like the SMI Eye Tracking Glasses (ETG)[1] which we used in our system. As shown in Figure 2, a sensor component reads the eye tracking data and the video stream of the user’s field of view and passes them to the sonification modules within the pipeline. Thereby the video signal is split into a sequence of frames, which can be processed sequentially by each component. Since every module runs in parallel, the framework also ensures that the data is properly synchronized across all of them. This guarantees that the components always process the same events at the same time. Additionally, due to the independent structure of each module, it is possible to use them in any desired combination. This allows the system to be adapted to the user’s current needs in every situation.

4.1 Color Sonification

The color sonification module is based on the idea that sounds can be mixed similarly to colors. As proposed in [5] and [6] we create an “audible color space” by mapping certain color values of the HSL color space to an appropriate counterpart within the sound space. Through that, the primary colors are represented by their respective sounds while mixed colors can be identified by the mixture of two primary sound components. In combination with eye tracking as input method the user should then be able to explore his environment just by moving his eyes. For example the user can differentiate between red and green apples while buying groceries or he can identify the color of his clothes when doing the laundry. Furthermore with a bit of training it might even be possible to recognize objects through the color differences of their contours as shown in [2], [6] and [8]. However, while those approaches only use static images for

[1] http://www.smivision.com

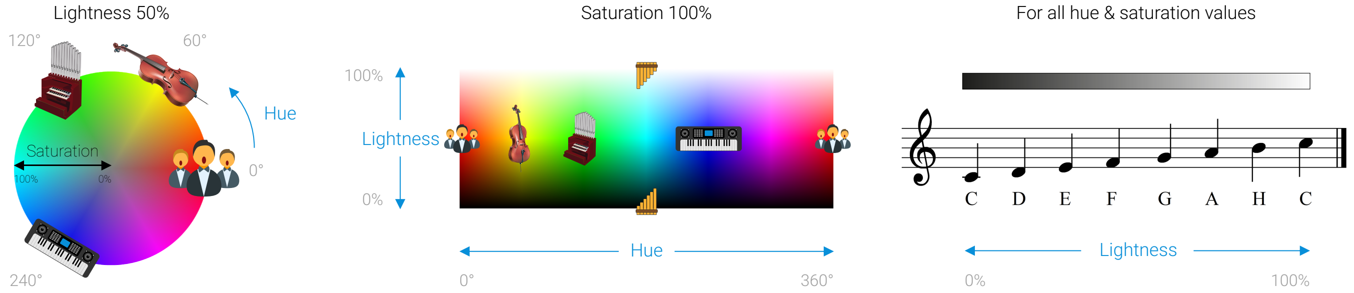

Figure 3: Assignment of colors to MIDI instruments

processing, we want to be able to sonify the video stream of the user's field of view in realtime. Therefore we can not use those concepts since they contain certain image operations which require too much processing time and would cause performance problems if applied to every frame of a video stream. As a result we came up with our own approach based on the work of Banf et al. [6].

Generally one of the challenges of color sonification is the fact that color values in images often change rapidly from one pixel to another even though the overall color of the material or texture is roughly the same. The reason for that is mostly due to image noise caused by the camera which leads to faulty pixels with differing color and brightness values. Since the sonification of those pixels would lead to wrong impressions and could confuse the users we need to remove them first. However, doing that for the whole image region would cause performance issues and is not necessary in our case. Instead, we first extract an area of 100 × 100 pixels around the current gaze point (x, y) from the video frame and apply a bilateral filter to it. Thus the noise gets removed while the edges within the image are still preserved [19]. After that the image section is converted to the HSL color space in order to extract the smoothed values for hue $h(x, y)$, saturation $s(x, y)$ and lightness $l(x, y)$. Depending on their combination an appropriate sound is played.

As shown in Figure 3 we currently use the following assignment of MIDI instruments to colors similar to [5]: *flute* (black, white, gray), *choir* (red), *organ* (green), *synthesizer* (blue) and *cello* (yellow). However, this assignment is just a suggestion, which can be adjusted according to the preferences of the user. For example, we tried to use birds' twittering for green and wind noises for blue to simplify the mental mapping of the colors. Although this was a bit more intuitive, we found that those sounds are likely to irritate the user if the system is used for a longer period of time, which is the reason why we went back to MIDI instruments. In conjunction with that, the sonification of secondary colors is achieved by playing the instruments assigned to both involved primary colors simultaneously with a certain sound level. Thereby the amount of each instrument is controlled through a specific volume shape ($v(h, s, l)$ with $v \in [0, ..., 1]$) which maps each combination of hue h, saturation s and lightness l to a value between 0 and 1 [6]. For instance the volume shape $v_{choir}(h, s, l)$ returns 1 for $h = 0°$, $s = 100\%$ and $l = 50\%$ while the volume shapes for all the other instruments return 0 and therefore only the choir is played with maximum loudness. In addition to the volume we also adjust the pitch according to the current lightness. For that, each lightness value l between 0 and 1 is mapped to one of the eight tones of a musical scale from C4 (261,6 Hz) to C5 (523,2 Hz). The resulting tone is then played by all instruments even if they can not be heard due to the value of their volume shape. This enables a more precise sonification of colors since it not only allows the user to distinguish between different colors, but also between dark and bright color variations.

4.2 Text Sonification

One of the most frequently mentioned problems during our interviews with blind and visually impaired people was the loss of the ability to read texts. While there are already some approaches like [3], [4] and [17] which convert text information to audio signals, most of them only focus on the intention of the sentences rather than the actual content. For the sonification of name plates, street signs or shop signs though, this is not very helpful as in those cases only the meaning of the text is relevant to the user. Therefore we came up with the following approach:

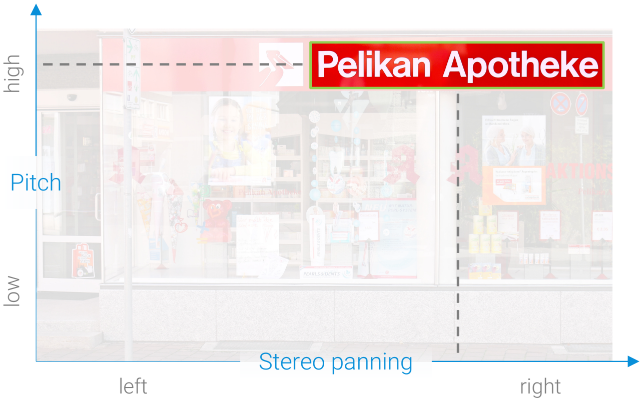

Figure 4: Conversion of text position to sound

In the first step, we analyze if there are any visible texts present in the field of view of the user. Since this is a computationally intensive operation, it can not be done for every frame of the video stream. Instead, we only execute this process as soon as the previous run has finished which results in an average execution rate of once every 2-3 seconds. For the analysis of the image data we use the Stroke Width Transform (SWT) algorithm [10] due to the fact that it has a very high precision while only requiring a relatively short processing time compared to other text detection algorithms. It is also language independent and works with many different fonts and sizes. After every execution the algorithm returns a set of rectangles designating the areas in which texts have been detected. The rectangles are then sorted by the y-value of their center points in descending order so that the sonification of multiple texts is always done from top to bottom. In this regard the x and y-coordinates of each detected text area are used to generate a sound through which the user can locate its position [7]. As shown in Figure 4 the x-coordinate thereby determines the stereo panning of the sound while the y-coordinate is mapped to a two octave musical scale from C3 (130,8 Hz) to C5 (523,2 Hz). Initially we only used one octave but after a few tests we found that a two octave musical scale allows for a more precise localization. Using more than two octaves however did not lead to any further measurable benefits. The sound itself is generated through a vibraphone MIDI instrument. It is chosen since it resembles a pleasant notification tone and distinctively differs from the sounds of the other modules.

Once the user moves his eye gaze into one of the text areas the text-to-speech conversion is triggered. More precisely, the corresponding section first gets extracted from the image and is then processed using a binary threshold function to simplify the text recognition [14]. After that we extract the actual text with *Tesseract*[2], an open source OCR-Framework maintained by Google. Subsequently, the output is passed to the *Microsoft Speech API*[3] which reads the text to the user. Until that is done the sonification of new texts is disabled in order to prevent any potential disturbances and distractions. The biggest advantage of this approach is, that the user has full control over the text sonification at any given time. This is especially helpful in situations where multiple texts are visible at once, since the sequential sonification of all detected texts would overwhelm the user. Furthermore, the person might not be interested in all the text information in his or her field of view. However, with our method the user can always decide, when which text should be read to him.

4.3 Facial Expression Sonification

The nonverbal part is regarded as a very important and informative component of a conversation. According to [13], while talking about likes and dislikes, the content of the speech contains only 7% of the information while the facial expressions represent 55%. From that emerges the utmost significance of facial expression sonification for the visually impaired. The input to our facial expression sonification system, a video stream along with a 2D point representing the location of the eye gaze of the user are provided via the SMI glasses and fed into the "Intraface" face tracking system [9] which is integrated into SSI. Intraface detects the faces in an input scene using the cascade haar classifier algorithm and then starts tracking the largest face in the scene over time. In our system, the original existing behavior of intraface is altered, so instead of tracking the largest face, the face closest to the user's eye gaze is tracked. The facial expression sonification process is subdivided into two tasks as detailed below.

For the recognition of facial expressions, we implemented several simple geometrical threshold based detectors. These are able to compute different facial actions (e.g. mouth corners down, mouth open, eye brows raised) and map them to facial expressions (laugh, smile, surprise, neutral and sadness/frown). Figure 5 provides some examples of recognized facial expressions. To transform the facial expressions into sounds, our first attempt was to use the *irrKlang*[4] sound library to play back short prerecorded expressive sound segments that correspond to the different detected expressions. For example, a laugh is represented by a happy cheerful sound while a frown is represented by a sad tune. However, initial informal tests revealed that repeated playback of the expressive sounds caused discomfort. To this end, we replaced the prerecorded sounds with the playback of various MIDI instruments. More specifically, the following mapping of facial expressions to sounds is used: *laugh* (drums), *smile* (french horn), *surprise* (bells), *neutral* (piano) and *sadness/frown* (guitar). While not so inherently recognizable, this approach has the benefit of being more pleasant to listen to over longer periods of time, such as while performing daily chores.

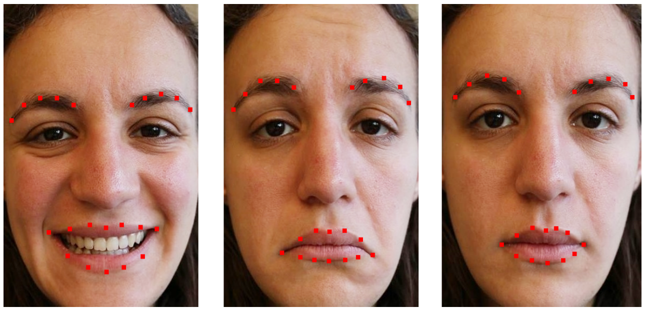

Figure 5: Examples of different facial expressions detected by the system: laugh (left), sadness (center), neutral (right)

5. USER STUDY

In order to get an accurate impression of the system's usefulness, we conducted a user study with blind and visually impaired persons with intact oculomotor control. Since this is a very special user group, which can not be reached without direct contact, we cooperated with the Bavarian Association for Blind and Visually Impaired (BBSB).

ID	Age	Gender	Visual impairment	Method
P1	68	male	Cataract	center
P2	49	female	Cataract (early stage)	gaze
P3	43	female	Optic atrophy	gaze
P4	73	male	Congenital blindness	center
P5	68	male	Optic nerve damage	center
P6	87	female	Macular degeneration	gaze
P7	70	male	Retinal degeneration	gaze

Table 1: List of participants

With their help we were able to find seven users which met all requirements and agreed to participate in our study as shown in Table 1. Even though the average age of the subjects was above 65 years, they were very open-minded towards and interested in new technologies.

[2] https://github.com/tesseract-ocr/tesseract

[3] https://msdn.microsoft.com/en-us/library/ee125663.aspx

[4] http://ambiera.com/irrklang

5.1 Experiments

Within the user study each module of our sonification system was evaluated separately to prevent any mutual influences and to enable reliable conclusions from the individual results. As a consequence each experiment was adjusted to the designated use case scenario of the corresponding module. However, in order to ease the arrival of the participants, the study was conducted at the premises of the BBSB and therefore no complex set-ups could be utilized. Instead, only portable objects and tools were used. After each experiment the participants were asked four questions regarding the difficulty and usefulness of the system and the pleasantness of the sounds, which they could answer on a five point Likert Scale. At the end of each trial, we also asked the users whether they would prefer to always run the modules in parallel or rather to activate them on demand. During all experiments, the users wore the SMI Eye Tracking Glasses which were connected to a notebook powered by an Intel Core i7 processor. For sound generation, we used a pair of common stereo speakers.

5.1.1 Experiment 1: Color Sonification

In this experiment, we examined the ability to recognize objects by the means of their color and shape with the help of the color sonification module. As preparation, we first presented the mapping of colors and corresponding sounds to the participants. Once they had memorized them, we started a training with the examples shown in Figure 6 (left). Thereby the user was asked to move his gaze from left to right and to repeat this process in a vertical direction in order to explore the image from top to bottom. With those examples, we wanted to make the users aware of the color and sound differences between the colored shapes and the black backgrounds. Moreover, we taught the participants that they can use the duration of each sound to identify the shape of the object. In case of the square, assuming constant eye gaze speed, the sound for green always has the same duration when scanning the image while in case of the triangle the sound for yellow is played shortly in the top and longer in the bottom region. Once the participants were familiar with this concept we started the experiment. For that, an apple was placed in the user's line of sight as shown in Figure 6 (right). The participants were told that the object was either a red apple, a banana or an orange. Now the task of the participants was to identify which object was in front of them only by using the color sonification module.

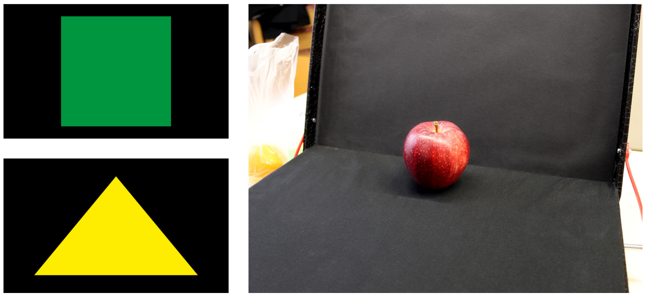

Figure 6: Training examples (left) and experiment set-up (right)

5.1.2 Experiment 2: Text Sonification

The second experiment was used to evaluate the text sonification module. For that, we trained the participants with two examples to clarify the transformation of text positions to acoustic signals. In the first example, the text area appeared in the top right corner of the user's field of view which resulted in a sound with high pitch coming from the right speaker. As soon as the participant looked in the direction of the text it was automatically read to him. In the second example, the text was shown in the bottom left corner. Once the user was familiar with locating the text positions, we began with the actual experiment. Thereby the user had to locate the text and move his or her eye gaze into the corresponding area similar to the examples. For each participant we measured whether the text position was recognized correctly and how much time has been required for the experiment.

5.1.3 Experiment 3: Facial Expression Sonification

The aim of this third experiment was to measure the ability of the system to make the users perceive the emotional state of the speakers more accurately, so that they can use the system during daily social interactions. We started the experiment with a 3-5 minutes training session aimed at familiarizing the participants with the mapping of facial expressions to sounds. During this session, the participant was instructed to look at an experimenter who simulated different facial expressions triggering the corresponding audio cues of the sonification module. We then started the experiment by showing each participant two videos, each roughly two minutes in length and consisting of a monologue of a young woman, in randomized order. Each video showed a different monologue regarding a past experience. The contents of the monologues were neutral in both cases. This was tested during a pre-study in which 12 normally sighted persons rated the content of the monologues to be neutral or ambiguous. However, one monologue was presented more positively while the other one more negatively. The videos were displayed on a 17" LCD monitor located roughly 1.5m in front of the participants. Each user watched one video with the sonification system on and one video with the sonification system off. The user then had to rate each video in terms of the emotional state of the person in the video on a scale from 1 (extreme negative) to 5 (extreme positive). In the case where the system was on, the user was permitted to use both the voice of the speaker and the sounds provided by the sonification system to identify the emotional state of the speaker. Whereas, the user depended only on the voice of the speaker in case where the system was off. Then, the results of the ratings of the videos were compared with the emotional state acted by the speaker during the video.

5.2 Results

Each of the seven participants performed all three experiments successively in one session with an average length of about 40 minutes per user. Generally, all of them completed the tasks without any major problems. However, in three cases we noticed that the eye tracker could not detect the gaze position correctly. In order to still obtain results from those users regarding the system itself, we adjusted the algorithms to use the center point of their field of view instead of the actual eye gaze. This way the participants could control the sonification by moving their head in a certain direction. Table 1 shows in which cases this has been done. With the adjustments in place, six users were able to correctly identify the object in the first experiment only by using the color sonification. In the second experiment it was even possible for all participants to detect the text position. In the third experiment, the emotional state of the person from the positive video was rated as more positive when the system was on ($M = 3.5$, $SD = 0.5$) than when the system was off ($M = 2.66$, $SD = 1.24$). Similarly, the emotional state in the negative video was rated as more negative when the system was activated ($M = 1.66$, $SD = 0.47$) than the case when it was deactivated ($M = 2.0$, $SD = 0.7$). Due to the small sample size no results have been found statistically significant.

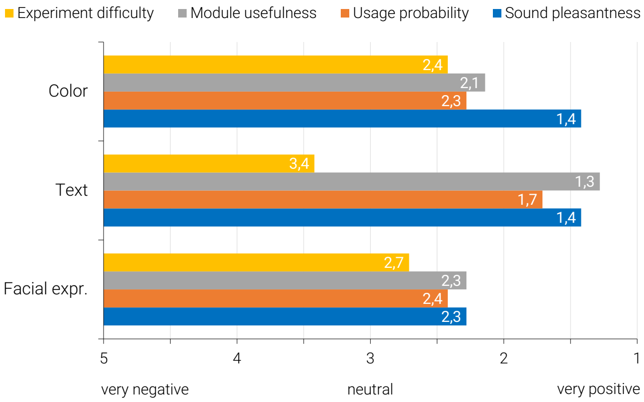

Figure 7: Questionnaire results

Figure 7 shows the results from the questionnaires where each number represents the average value across all participants. Generally, most of the results are in the positive area. The only exception from that is the difficulty of the text sonification experiment. Though in return, the module usefulness, usage probability and sound pleasantness was rated more positive in this case than for the other two modules. A further result which is not shown in the diagram was regarding the question whether the modules should always run in parallel or if they should be activatable when needed. With 85, 7%, the majority of all participants voted for an activation on demand.

6. DISCUSSION

The evaluation study yielded overall positive results. Both color and text sonification modules proved to be useful and usable by our target group. The results from the third experiment suggest that the facial expression sonification may improve the user's ability to recognize emotions. Furthermore, the lower standard deviations for the condition with the system turned on can be interpreted as a reduction in the confusion of the users. From the questionnaire data we can extract that the likeliness of such a system being used by members of our target group is reasonably high. We were also pleased by the positive ratings of the sound pleasantness which suggests that the system could be used over a longer period of time. Nevertheless, some users stated that they would have needed a longer training session to be more proficient in recognizing the different audio cues. P4 was more critical stating that while the color sonification was useful for recognizing simple objects (the apple), it might prove to be difficult to use on more complex objects. Some participants pointed out that the usefulness of the modules might be influenced by the activity they wish to perform: "I could imagine using it [the color sonification] in certain situations" (P3).

We also observed some technical limitations of the system. For the color sonification module, we found that under certain conditions, the camera we used would falsify the colors. More specifically, during our user study a brick wall outside of the window significantly shifted the colors of the objects into the red spectrum. Here, a different camera might resolve the issue. The text sonification module currently also suffers from a relatively slow update rate, allowing text fields to be recognized only once every 2-3 seconds. This problem could be addressed by utilizing more efficient hardware. In our study setup we also used a pair of stationary speakers, however wireless bone conductance headphones could be used as well. This would make the system mobile while not limiting the user's ability to hear and react to outside events. Furthermore, it would be interesting to evaluate the system in a more real-world setting over a longer period of time, as opposed to the controlled environment in our current study. Thereby it could be investigated how the system handles more complex visual scenes.

Although the eye tracking technique proved to be successful, it did not work correctly for three of the seven cases, where the center point of the image had to be used instead of the eye gaze to identify the information to be sonified. There were however, no prominent differences in the results between the participants which were able to use the eye tracking method and those for which the center of the field of view was used instead. The reason for this were the medical conditions which caused the visual impairment and how these presented themselves in the participants. One user was diagnosed with cataract, a condition which causes the lens of their eye to turn cloudy and misty rather than being clear and transparent. This interfered with the eye tracking glasses' ability to track the position of the pupil, as it is designed to detect dark colored pupils. Here, adapting the tracking algorithm might make it more robust towards pupil color. In another case, the congenital eye blindness caused one participant to have difficulty controlling his eye movements since he has never purposefully used the eye muscles. The third participant suffered from optic nerve damage which only permitted him to partially open his eyes. This did not give the eye tracker a clear view of the pupil. While all these issues proved to be unsurmountable by the eye tracking glasses we used, this might have been caused by the fact that these technologies are designed for and tested on normally sighted persons. Specifically accounting for such variations in the human visual system during the design and development of eye tracking solutions might make them more robust towards the blind and visually impaired.

Furthermore, after our workshop, we intentionally constrained the target group by elimination of users with pathological nystagmus: a medical eye condition where the eye movements are repetitive, random and uncontrollable. An interesting area we are considering for our future work is the deployment of the system in a bio-feedback therapy scenario [1]. This may be able to help users train on controlling their eye movements by giving them auditory feedback about their eye locations at specific moments.

7. CONCLUSION

The aim of this work was to explore the feasibility of eye tracking as an input method to control the sonification for blind and visually impaired people. In order to identify the most useful applications for that, we conducted a design workshop with members of the target user group. Based on their feedback we implemented an eye-tracking-driven sonification system capable of converting colors, texts and facial expressions from the field of view of the user into acoustic signals. Through that, the users are able to decide, which elements should be sonified at any point in time just by moving their eyes. To evaluate the effectiveness of our approach, we conducted a user study with seven blind and visually impaired persons. Generally, all modules of the sonification system worked as intended and were perceived rather positive by the participants. Although we limited our target user group to persons who can move their eyes and do not suffer from pathological nystagmus, three participants were still not able to use their eyes to control the system. The reasons for that can be mostly attributed to the nature of their visual impairment. For example, the cataract of one participant was so far progressed that the pupil was almost white and thus could not be detected by the eye tracker. In another case the visual impairment caused by an accident only allowed for restricted eye movements. Therefore, future work should focus on the identification of the conditions which enable the usage of such a system. For persons who can actually utilize our approach, eye tracking appears to be a very promising input method to control the sonification.

8. ACKNOWLEDGMENTS

The work was partially funded by the German Federal Ministry of Education and Research (BMBF) under project "GLASSISTANT".

9. REFERENCES

[1] R. Abadi, D. Carden, and J. Simpson. A new treatment for congenital nystagmus. *British Journal of Ophthalmology*, 64(1):2–6, 1980.

[2] S. Abboud, S. Hanassy, S. Levy-Tzedek, S. Maidenbaum, and A. Amedi. Eyemusic: Introducing a visual colorful experience for the blind using auditory sensory substitution. *Restorative neurology and neuroscience*, 2014.

[3] F. Alt, B. Pfleging, and A. Schmidt. Sonify - a platform for the sonification of text messages. In S. Boll, S. Maaß, and R. Malaka, editors, *Mensch & Computer 2013: Interaktive Vielfalt*, pages 149–158, München, 2013. Oldenbourg Verlag.

[4] F. Alt, A. S. Shirazi, S. Legien, A. Schmidt, and J. Mennenöh. Creating meaningful melodies from text messages. In K. Beilharz, B. Bongers, A. Johnston, and S. Ferguson, editors, *Proc. NIME*, pages 63–68, 2010.

[5] M. Banf and V. Blanz. A modular computer vision sonification model for the visually impaired. In *Proc. ICAD*, Atlanta, USA, 2012.

[6] M. Banf and V. Blanz. Sonification of images for the visually impaired using a multilevel approach. In *Proc. Augmented Human*, pages 162–169, New York, NY, USA, 2013. ACM.

[7] M. Brock and P. O. Kristensson. Supporting blind navigation using depth sensing and sonification. In *Proc. UbiComp*, pages 255–258, 2013.

[8] S. Covaco, J. Henriques, M. Mengucci, N. Correia, and F. Medeirous. Color sonification for the visually impaired. In *Procedia Technology*, pages 1048–1057, Amsterdam, Netherlands, 2013. Elsevier.

[9] F. De La Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn. Intraface. In *Proc. Automatic Face and Gesture Recognition*, 2015.

[10] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *Proc. CVPR*, pages 2963 – 2970, 2010.

[11] I. Guizatdinova and Z. Guo. Sonification of facial expressions. In *Proc. new interaction techniques*, 2003.

[12] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014.

[13] A. Mehrabian. Communication without words. *Psychology Today*, pages 53–56, 1968.

[14] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 09:62–66, 1979.

[15] D. Pascolini and S. P. Mariotti. Global estimates of visual impairment: 2010. *British Journal of Ophthalmology*, 2011.

[16] V. Patil, M. Q. Akhtar, A. Parab, and A. Fernandes. Sonification of facial expression using dense optical flow on segmented facial plane. In *Proc. ICCCE*, India, 2012. Coimbatore institute of technology.

[17] B. Pfleging, F. Alt, and A. Schmidt. Meaningful melodies: Personal sonification of text messages for mobile devices. In *Proc. MobileHCI*, pages 189–192. ACM, 2012.

[18] E. Striem-Amit, M. Guendelman, and A. Amedi. Visual acuity of the congenitally blind using visual-to-auditory sensory substitution. *PLoS ONE*, 7, 03 2012.

[19] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proc. ICCV*, pages 839–846, 1998.

[20] L. Twardon, H. Koesling, A. Finke, and H. Ritter. Gaze-contingent audio-visual substitution for the blind and visually impaired. In *Proc. PervasiveHealth*, pages 129–136, ICST, Brussels, Belgium, Belgium, 2013. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[21] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.

[22] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André. The social signal interpretation (ssi) framework: multimodal signal processing and recognition in real-time. In *Proc. MM*, pages 831–834, New York, NY, USA, 2013. ACM.

[23] T. Yoshida, K. M. Kitani, H. Koike, S. Belonge, and K. Schlei. Edgesonic: Image feature sonification for the visually impaired. In *Proc. Augmented Human*, New York, NY, USA, 2011. ACM.