Colorophone 2.2 – A Spatial, Real-Time Color-to-Sound Sensory Substitution Device

Dominik Osiński, *Member IEEE*, Marta Łukowska, Weronika Kałwak, Michał Wierzchoń and Dag Roar Hjelme

***Abstract*— Objective: Blindness deprives a person of a significant part of sensory information resulting in limited perceptual abilities and decreased quality of life. Although some aspects of visual information, like the shape of an object, can be conveyed by other senses, there is no easy way to perceive color without using sight. To address this issue, we developed the Colorophone sensory substitution device.**

Method: In a way analogous to pixels in visual displays, we introduced auxels (auditory pixels) that can be used as basic building blocks of a generic auditory display. The developed auxel-based system realizes real-time, spatial, color-to-sound conversion. We created a dedicated software suite that enables the independent introduction of various system features to ensure effective training. Four blind participants assessed the prototype's usability. The evaluation methods included: auditory color recognition, object identification, and virtual sound source localization tasks, as well as two self-descriptive methods: the System Usability Scale and the NASA Task Load Index.

Results: The developed wearable system generates spatially calibrated colorful soundscapes. It enables auditory color recognition and object identification significantly above chance. However, analyzing complex natural scenes remains challenging. Users judged the system's usability from good to best imaginable. The identified usability issues are discussed together with the proposed solutions.

Conclusion: The Colorophone device shows promise for the future development of a useful visual rehabilitation device; however, further work is needed to eliminate existing usability issues.

Significance: The presented work contributes to developing a universal, affordable and user-friendly visual rehabilitation device.

***Index Terms*—Colorophone, Sensory Substitution, Color sonification, Image sonification, Assistive device, Visual impairment, Wearables, Usability, Human-computer interface**

I. Introduction

THE population of blind people exceeds 43 million [1], and it is prognosed that this number will increase to 115 million by 2050 [2]. Visual impairment can negatively influence motor, language, emotional, social and cognitive development [3]. It also reduces the quality of life, independence, mobility and contributes to the decline of physical and mental health [4]. The inability to process visual information creates multiple challenges in everyday functioning, like object identification and description [5]. One of the proposed solutions for the visual rehabilitation of the blind are sensory substitution devices (SSDs) [6]. These devices convert the information from the unavailable sensory modality to another [7], [8]. In the case of visual-to-auditory sensory substitution devices (VASSDs), the acquired image is converted to sound. One of the main challenges in VASSD design is developing the conversion method, which will deliver useful information without distorting the ability to process incoming environmental sounds [9]. It is a daunting task due to the mismatch in sensory data throughput between the visual and auditory sensory channels [10]. However, this challenge might be addressed by developing color sonification devices. In low-resolution images, color information begins to play an important role in scene recognition [11]. Moreover, color is the only perceptual dimension that is not accessible via other senses, and there is no practical way to adjust the environment to provide this information. Several camera-based color sonification systems have already been developed [12], [13], [14], [15], [16]. However, none of the existing systems provide real-time, spatialized color sonification that might enable effective exploration of the visual environment. The detailed presentation, analysis, and design considerations regarding color sonification systems can be found in our previous work

This work was supported by the National Science Centre, Poland, grant OPUS 711 (2016/23/B/HS6/00275) given to Michał Wierzchoń and internal NTNU's Kompetanseløft grant.

The experimental protocol was approved by the Committee for Research Ethics of the Institute of Psychology of Jagiellonian University (decision KE/01/062017) and adhered to the tenets of the Declaration of Helsinki.

D. Osiński is with the Department of Electronic Systems, Norwegian University of Science and Technology, 7491 Trondheim, Norway (e-mail: dominik.osinski@ntnu.no).

M. Łukowska is with Trauma, Health and Eating Lab, Institute of Psychology, University of Silesia in Katowice, Poland, also with Consciousness Lab, Institute of Psychology, Jagiellonian University, Ingardena Street 6, Kraków 30-060, Poland (e-mail: marta.lukowska@us.edu.pl),

W. Kałwak is with Qualitative Research Lab, Institute of Psychology, Jagiellonian University, Ingardena Street 6, Kraków 30-060, Poland (e-mail: weronika.kalwak@uj.edu.pl)

M. Wierzchoń is with Consciousness Lab, Institute of Psychology, Jagiellonian University, Ingardena Street 6, Kraków 30-060, Poland, also with Centre for Brain Research, Jagiellonian University, Kraków, Poland and Jagiellonian Human-Centered Artificial Intelligence Laboratory, Jagiellonian University, Kraków, Poland (e-mail: michal.wierzchon@uj.edu.pl).

D. R. Hjelme is with is with the Department of Electronic Systems, Norwegian University of Science and Technology, 7491 Trondheim, Norway (e-mail: dag.hjelme@ntnu.no).

[17]. Here, we present the new version of the Colorophone system that converts the acquired camera image to sound by generating continuous, real-time stereo soundscapes. The implemented color sonification method is based on the association of individual color components with corresponding sound components. Spatial color sonification is realized by applying the newly proposed design framework that introduces auxels as the basic building blocks of a generic auditory display. The details regarding the development of a spatially calibrated auxel-based system will be described in the following sections. The operation of the system is presented in the videos that can be found in link1 (using headphones is recommended). In addition, we developed a dedicated software suite that enables an effective training process. Finally, we evaluated the system's usability in the context of color recognition, colorful object identification, and virtual sound source localization tasks with four blind users. We assessed usability with the System Usability Scale (SUS) with follow-up questions and NASA-TLX [18], [19]. We identified the existing Colorophone's 2.2 usability issues and proposed some adjustment to increase the acceptance probability of the future versions of the system.

II. The Colorophone

The Colorophone initiative aims to develop an affordable, user-friendly SSD that will enhance the perceptual and cognitive capabilities of the visually impaired [20].

A. Iterative development process

The system has been developed in an iterative way, where the results from the previous development round were used as recommendations for the subsequent round. The idea of color sonification, inspired by the human visual system, was introduced in [21]. The next development iteration, Colorophone 2.0 [17], introduced a dedicated opponent-process-based auditory color space and spatial color sonification. It was evaluated by three independent experts in a usability audit that resulted in recommendations implemented in the Colorophone 2.1 system (see supplementary material for details). The 2.1 version of the system, which provided calibrated virtual acoustic space (VAS), was evaluated by a single blind participant in the pilot study. The pilot study provided insights applied in the current system version.

B. Colorophone 2.2 system

The prototype, depicted in Fig. 1, consists of Bose Frames Bluetooth glasses, a centrally attached Vsightcam USB mini camera, a processing unit in the form of a Windows Surface tablet, and a Bluetooth controller. The image acquired by the camera is processed in the algorithm described in section E, to generate live, stereo soundscapes. The system enables the sonification of color information from a single area of interest (non-spatial, point mode) or multiple, horizontally organized areas of interest (spatial, panoramic mode). The wireless controller enables volume control, switching between point and panoramic modes, and turning on and off sounds associated with achromatic color components, i.e., black and white.

Fig. 1. The Colorophone 2.2 system: USB mini camera, Bose Frames, Bluetooth controller, and Microsoft Surface Pro as a processing unit.

C. Color sonification method

The applied color sonification method is based on previous work described in detail in [17]. Nonetheless, the method has been slightly modified based on the feedback from [17], and the pilot study. Every color component is still associated with a corresponding auditory color component based on cross-modal correspondences. However, the black color component previously represented by silence is now represented by a low-pass filtered white noise. Adding the black color component allows for a direct auditory representation of black objects. The sound representing the black auditory color component was chosen based on cross-modal correspondences and considering possible masking effects. Other auditory color representations have not been altered; hence white is coded by a rainfall, red by a high pitch sound (1027 Hz), yellow by a middle-high pitch (647 Hz), green by a middle-low pitch (408 Hz), and blue by a low pitch sound (256 Hz). The developed auditory color space can therefore be called an RYGBWBk color space. Color component changes in every type of color transition are shown in Fig. 2-4. The transition depicted in Fig. 2 will look similarly

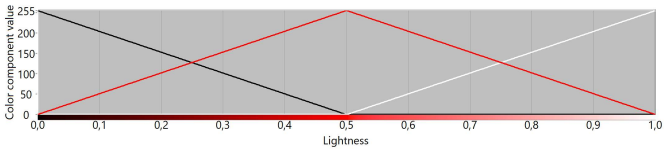

Fig. 2. Visualization of the black to red to white color component transitions in the RYGBWBk color space.

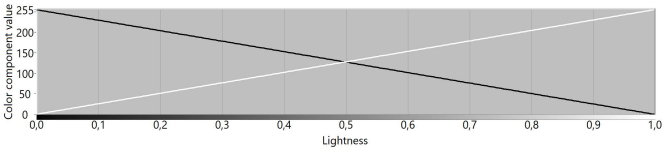

Fig. 3. Visualization of the black to white, achromatic color component transitions in the RYGBWBk color space.

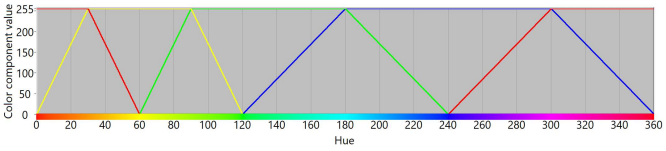

Fig. 4. Visualization of the chromatic color component transitions in the RYGBWBk color space adapted from [17].

for other chromatic color components, i.e., yellow, green, and blue. The system's operations and live spectrograms illustrating soundscape frequency composition can be found in link 2.

D. Spatial color sonification

Here, we propose a design framework that allows for a coherent, general description of spatial color sonification methods. We start with definitions and interrelations between a pixel, a superpixel and an auxel.

A pixel originates from the words 'picture element' and is the fundamental atomic element of an image [22]. Usually, every pixel is described by a set of numbers that define its position and color components of the used color space; thus, a pixel can be noted as a pixel(x, y, R, G, B). Consequently, a superpixel is a rectangular area representing averaged color information from a given number of input pixels. It can be noted as a superpixel$(x_1, y_1, x_2, y_2, \bar{R}, \bar{G}, \bar{B})$, where x_1, y_1 and x_2, y_2 describe the coordinates of the rectangle's corners and $\bar{R}, \bar{G}, \bar{B}$ denote the average color values for these color components. The natural way to define pixel coordinates is a Cartesian system of two dimensions. It is an obvious choice for defining the position of a light source on a two-dimensional screen. However, it is not the natural coordination system for defining the position of sound sources; hence the position of a sound source is often described by the azimuth and elevation parameters inherent to the spherical coordination system. As presented in the previous section, the Colorophone auditory color space consists of RYGBWBk color components. Here, we define an auxel as a virtual sound source characterized by a set of numerical values describing its position and the values of the auditory color components.

Thus, an auxel used in the Colorophone system shall have the following set of parameters: auxel$(\theta, \phi, R, Y, G, B, W, Bk)$, where θ and ϕ represent azimuth and elevation angles. The definition of auxel has been adapted from [23].

1) Superpixel-auxel position calibration

The evaluated version of the Colorophone system was configured with 23 rectangular superpixels. Each superpixel covers approximately 5° of the camera's field of view (FoV). For each superpixel, color and position information are converted to corresponding auxel parameters. The color information is converted from RGB into RYGBWBk color space, and the azimuth θ of every auxel is defined at the center of the corresponding superpixel. The elevation angle ϕ is equal to zero because the processed superpixels are taken from the vertical center of the image. It is necessary to calibrate the system to establish the perceived spatial alignment between the positions of superpixels and corresponding auxels. Defining exact Interaural Time Differences (ITDs) is crucial for the system's ability to emulate the relative position of virtual sound sources. The calibration ensures that every auxel is associated with the exact spatial position of the corresponding superpixel. The supplementary material provides detailed descriptions of the superpixel parameters, calibration procedure, and calibration results.

E. Sonification algorithm

Here, we present the algorithm that is used to generate soundscapes. Before running the main part of the sonification algorithm, it is necessary to define the sizes and positions of superpixels and calculate the ITDs that result from the relative position of every superpixel and associated auxels. As mentioned in the previous section, every auxel is positioned in the center of the corresponding superpixel. The algorithm is presented in Fig. 5.

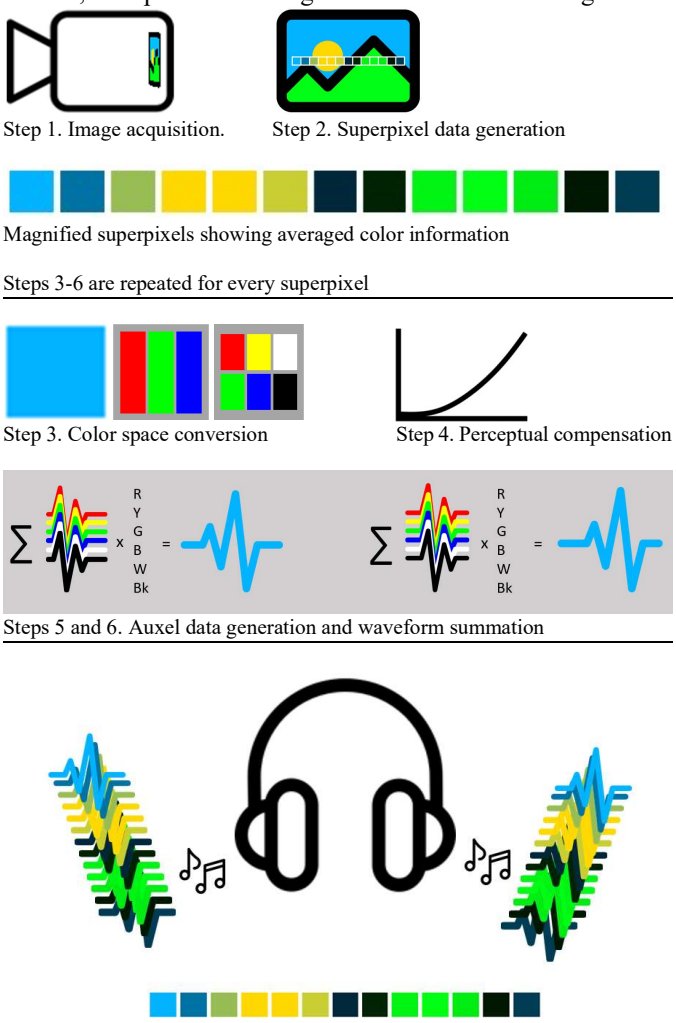

Fig. 5. The seven-step Colorophone sonification algorithm.

1) Image acquisition

The first step is acquiring a colorful image from the USB mini camera.

2) Superpixel data generation

The second step involves averaging color information for every superpixel based on pre-defined sizes and positions. As a result of this step, we receive an array of average RGB values for each superpixel.

Steps 3-5 are performed for every superpixel

3) Color space conversion

In the current step, the color information coded in RGB color space is converted to the RYGBWBk color space based on the color space conversion method described in the section "Color sonification method" and in [17].

4) Perceptual compensation

During this step, the numerical values representing colors in the RYGBWBk color space are transformed to compensate for non-linearities in human loudness perception [24]. This step produces values of normalized auditory color components.

5) Auxel data generation

Based on the information of the position of every superpixel and corresponding ITDs, pairs of waveforms associated with every auxel are phase-shifted to emulate the position of the created virtual sound source in space. The digital waveform amplitude values are multiplied by the corresponding normalized RYGBWBk values. The perceived loudness of every waveform reflects the perceived intensity of every color component.

6) Waveform summation

Individual waveforms representing every auditory color component of every auxel are summed independently for the left and right sound channels.

7) Sound generation

The last step is generating stereo sound from the digital waveforms constructed in the previous step and playing it via stereo headphones.

F. Hardware

The system is designed in a modular way that enables the easy exchange of used modules in future realizations. The Colorophone 2.2, presented in Fig. 1, consists of the following modules:

1) Camera

The utilized camera uses the OV2735 image sensor and a distortion lens that provides a 140° FoV. The camera is attached to the center of the Bose Frames audio glasses by a semi-rigid joint that enables individual adjustment of the default elevation angle for every user.

2) Bose frames

Bose Frames Bluetooth audio glasses [25] are used as wireless headphones and a mounting frame for the camera.

3) Microsoft Surface

The processing unit is Microsoft Surface Pro 7+, equipped with an i7, 2.8 GHz processor, 16GB of RAM, and 256GB disc space.

4) Controller

The utilized controller is an off-the-shelf product, typically used on a car steering wheel or on a bike handlebar for remote control of music-playing devices.

G. Software architecture

The used programming environment was LabVIEW 2020 with Vision Development Module and Vision Acquisition Software. The chosen environment enables easy hardware integration and rapid system development. Here, we describe the software architecture for the main Colorophone 2.2 program (see Fig. 6 for the functional block diagram). The current version of the system is based on the previous realizations described in [17]. Although the developed programs differ in implementation details, the steps described below can provide an overview of the developed processing pipeline and its interdependencies.

1) Initialization

In the initialization step, the image acquisition parameters and auxel definitions are used to calculate ITDs for every auxel. All hardware connections are also established in this step.

2) Image acquisition

In the image acquisition loop, the RGB image from the camera is acquired, and the horizontal stripe with pixels is extracted.

3) Data processing

In the data processing loop, the extracted stripe is converted to superpixels. Then the color space conversion and perceptual compensation are performed and only the data representing active auxels is sent further through the processing pipeline.

4) Audio generation

In the audio generation loop, previously calculated ITDs are applied to generate delays between waveforms, and auditory color component values are multiplied with waveforms representing corresponding color components. The waveforms for the left and the right auditory channels are respectively added, and the resulting stereo sound is generated on the default audio output device.

5) Graphical User Interface (GUI) update

The update of the GUI is realized in the independent loop so that it will not interfere with the processing part of the system.

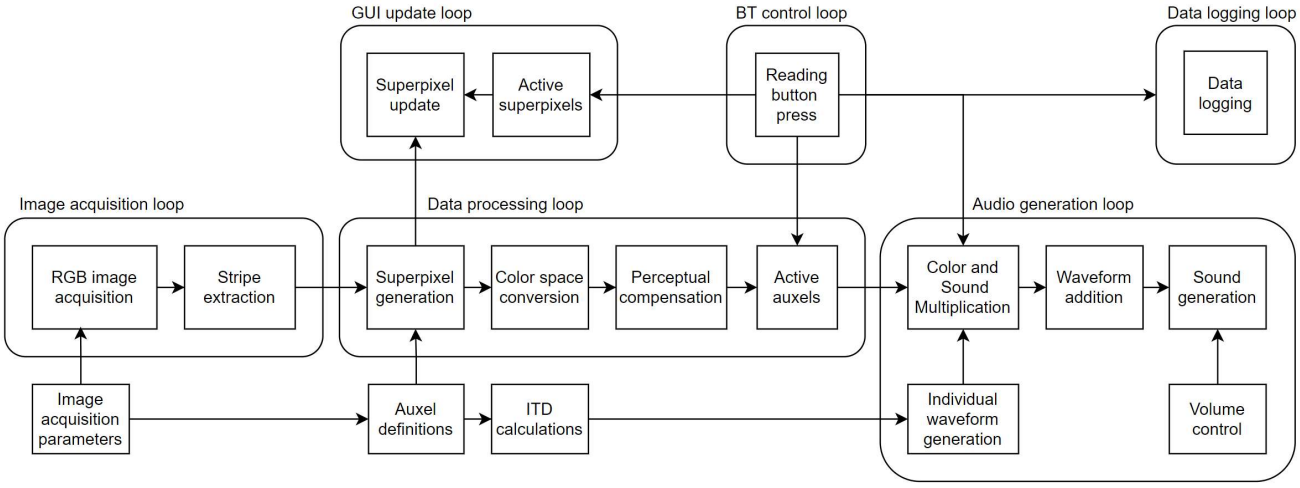

Fig. 6. Functional block diagram of the main Colorophone 2.2 program

6) Bluetooth control

The Bluetooth control loop assures continuous contact with the controller and enables communication with multiple loops that are influenced by option switching.

7) Data logging

The data logging loop enables parameter logging for every experiment (see supplementary material for details).

H. Developed programs

We developed four dedicated programs to enable effective training and evaluation of the system. The supplementary material provides a detailed description of the programs and their functions for the subsequent versions. Here we provide an overview of the functions of the developed programs in version 2.2.

1) VAS generator

The program generates virtual acoustic space (VAS) based on spatialized auditory color representations. It enables individual control of auditory color components for every auxel. VAS generator provides the possibility to present ideal virtual stimuli, which is advisable during the learning process. Although the VAS program generates spatialized sound, the stimuli do not change when the participants move their heads.

2) Image hearer

The program realizes the sonification of the color information indicated by touching the image displayed on the tablet screen. It allows the presentation of continuous color transitions at a trajectory and pace controlled by the user. Image hearer can present any color transition that results from the image used for sonification and applied mouse pointer trajectory.

3) Colorophone main

Colorophone is the main program that realizes spatial color sonification of the image acquired by the camera. It operates in two modes: researcher mode and user mode. Researcher mode enables GUI-based individual auxel control (activation or muting chosen auxels), turning on and off achromatic auditory color components and spatial mode by using screen controls. User mode allows remote toggling of the achromatic auditory color components and spatial mode by using the Bluetooth controller. The program can continuously log the selected operation modes, including point/panoramic modes and active achromatic color components. The graphical user interface of the Colorophone 2.2 program in the user mode is presented in Fig. 7.

4) Colorophone assistant

Colorophone assistant is a simplified version of the main Colorophone application, with a GUI, adapted for Polish-speaking users that does not contain the researcher mode. It was developed for non-professional training assistants (see our paper about SSD training [26] for details). The assistant version of the system continuously logs information about the chosen operation modes.

III. KNOWN SYSTEM LIMITATIONS

The Colorophone 2.2 system has some inherent limitations that result from hardware choices and the interactions between the implemented color sonification method and the human perceptual system. Here, we enumerate the known limitations of the device.

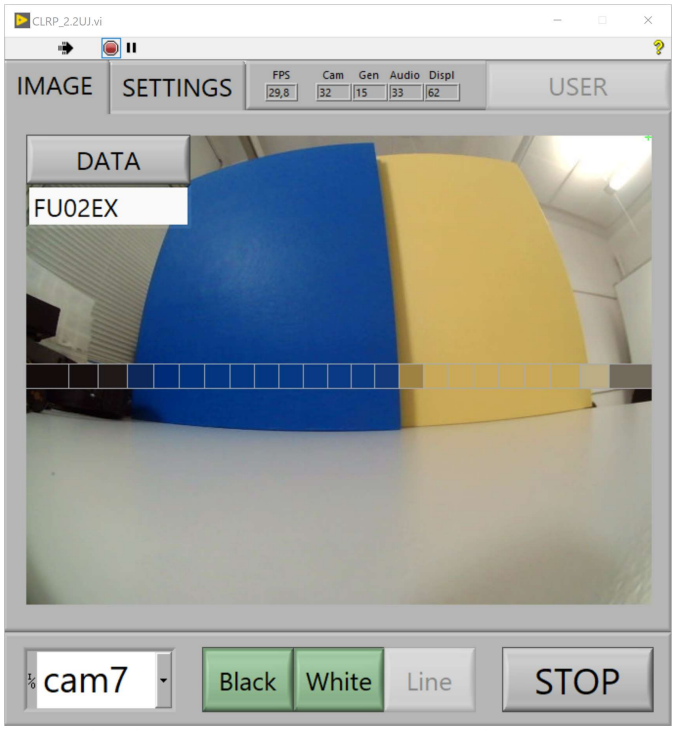

Fig. 7. Colorophone 2.2 main program, GUI in user mode.

A. Temporal resolution and processing delays

The camera hardware determines the temporal resolution of the system. The current frame rate is 30 frames per second (FPS). However, in response to poor light conditions, the camera can automatically adjust the frame rate. In that case, the sounds are still continuously generated, but the system output's reaction to changing input signals becomes delayed.

B. Auditory color interference

Neighboring auxels representing individual auditory color components can be interpreted as a single auxel representing complex color (e.g., two neighboring auxels representing red and yellow can be perceived as one auxel representing orange).

C. Auditory stream overlap

Two objects of the same color simultaneously present in the processed image stripe will affect the objects' perceived position and color composition. In the worst-case scenario, two objects of the same color symmetrically placed on the sides of the sonified scene can be perceived as one centrally placed object. However, this issue can be easily addressed by head movements.

D. Color constancy

In the current version, the system does not use any method that ensures color constancy; therefore, auditory representations of the same object are affected by the background light conditions.

E. Color reproduction

The chosen camera has good low-light performance; however, green-blue colors are not reproduced adequately. In addition, the color naming in Polish (all testers were Polish) and

English, in contrast to, for example, Russian [27] does not use two distinct words for blue and light blue, which also might contribute to the difficulties with color identification and categorization of these particular colors.

IV. Evaluation & Usability

In order to evaluate the Colorophone 2.2, we conducted several tasks during a structured and supervised training [26] with 4 visually impaired users (two congenitally, one early, and one late blind; 2 female; age = 39.75 ± 7.32). The training lasted 10 days and ca. 22 hours in total. All participants gave written informed consent to study participation. The experimental protocol was approved by the Committee for Research Ethics of the Institute of Psychology of Jagiellonian University (decision KE/01/062017) and adhered to the tenets of the Declaration of Helsinki. Importantly, the training was preceded by a five-day-lasting introductory training with the former version of the Colorophone (i.e., v. 1.0) underwent by all four users [28] as well as a pilot study with one, highly insightful, congenitally blind user (User 1; for details – see [26]). To keep it consistent with the users' numbering in [26], the user numbering in the current paper starts with User 2.

We aimed to assess users' ability to recognize colors using the Colorophone (by the auditory color recognition task applied on day 2 of the training); ability to identify simple objects (by the object identification tasks applied on day 2 of the training), and ability to localize auditory color representations in space (by the virtual sound source localization tasks applied on day 3 of the training). Additionally, we aimed to assess users' subjective experience when using the device (by the System Usability Scale applied on day 9 of the training) and task load (by the NASA Task Load Index applied after every training day). Moreover, we analyzed users' responses to the follow-up questions after each SUS item to identify existing usability issues.

A. Auditory color recognition and object identification tasks

We administered two auditory color recognition (CR; i.e., pre-recorded auditory color representations – *sounds task*; the natural unicolor objects – *socks task*) and one object identification (OI; i.e., the natural multicolor objects – *cans task*) tasks.

In the *Sounds task*, we presented basic auditory color components RYGBWBk (6 sounds * 4 reps = 24 trials; shuffled). In each trial, users listened to a sound for as long as they wanted and then were asked to recognize the color indicating one of the six possible colors associated with the sound (6-alternatives-forced-choice task, 6AFC). Group mean CR accuracy equaled 95.94 ± 3.41%. All participants responded significantly above the 29% chance level, see Fig. 8; (i.e., corrected chance level based on the binomial distribution, chance [95%] – accuracy for the threshold of 5% of guessing - see supplementary materials for the details)

In the *Socks task*, users have to indicate the color of unicolor socks (4 colors [RYGB] * 3 reps = 12 trials; shuffled) using the non-spatial mode. In each trial, they were allowed to inspect the socks without time limits until they selected the answer (4-alternatives-forced-choice task, 4AFC). Group mean CR accuracy equaled 85.42 ± 12.5%. All participants responded above the 42% chance level, see Fig. 8.

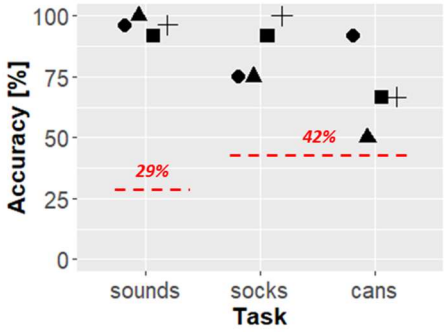

Fig. 8. Accuracy in the color recognition and object identification tasks with respect to users (coded with shapes). Dashed lines mark the chance level. Users' markers: User 2 – circle, User 3 – triangle, User 4 – square, User 5 – cross.

In the *Cans task*, users have to indicate a type of multicolor soda cans (4 types [Pepsi, Mirinda, ChaiKola, Tymbark] * 3 reps = 12 trials; shuffled) using the non-spatial mode. The task started with a familiarization part, during which users were allowed to freely explore and memorize the cans' appearance without time limits. Then, in the test part, in each trial, they were allowed to explore the cans without time limits until they selected the answer (4AFC). Group mean CR accuracy equaled 68.75 ± 17.18%. All participants responded above the 42% chance level, see Fig. 8.

B. Virtual sound source localization tasks

Users underwent three virtual sound source localization tasks using the virtual acoustic space (VAS) generator to learn how to discriminate sound localization within the acoustic space of the spatial mode: one auxel location detection (1 sound location - *1SL task*), two auxels (right-left) discrimination (2 sounds location - *2SL task*), and multiple auxels detection (multiple sounds location - *MSL task*). The locations of the 23 auxels were noted from -11 for the leftmost auxel, 0 for the central auxel, and 11 for the rightmost auxel.

In the *1SL task*, one sound (R/Y/G/B) was played at a time and users were asked to determine its location (3-alternatives-forced-choice task, 3AFC: right/left/center). The task was divided into three blocks with respect to a possible location of the sound: peripheral (auxels: -11:-9 & 9:11), medial (auxels: -8:-5 & 5:8), central (auxels: -4:4). Each user underwent 18 trials (3 block * 6 reps; randomized within blocks). Group mean SL accuracy equaled 79.16 ± 12.32%. All participants responded above the 56% chance level, see Fig. 9. The performance of the participants was comparable in peripheral, medial and central blocks (see supplementary material).

In the *2SL task*, two sounds (R/Y/G/B) were played simultaneously and users were asked to determine the location of a target sound (2AFC: right/left). The task was divided into three blocks with respect to a possible location of the two sounds: symmetric (one sound on the left, second on the right in the same distance from the central auxel, e.g. -7 and 7), asymmetric (one sound of the left, second on the right with a

different distance from the central auxel, e.g. -7 and 4), close (the sounds distance between 0 to 2 auxels). Each user underwent 18 trials (3 block * 6 reps; randomized within blocks). Group mean SL accuracy equaled 75.96 ± 17.31%. Two participants responded above, one below, and one at the 72% chance level, see Fig. 9. Note that participant 3 exhibits an especially low level of performance in a symmetric and close block (see supplementary material).

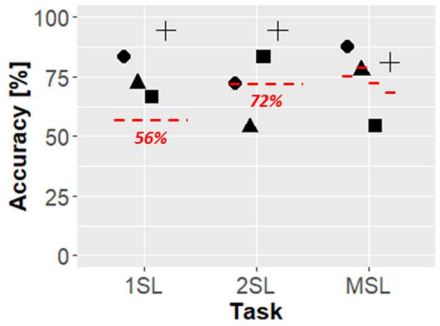

Fig 9. Accuracy in the virtual sound source localization tasks with respect to users (coded with shapes). Dashed lines mark the chance level. Users' markers: User 2 – circle, User 3 – triangle, User 4 – square, User 5 – cross.

MSL task was divided into three blocks, each differing in the number of sounds played simultaneously within the VAS: block 1: 2 or 3 sounds, block 2: 3 or 4 sounds, and block 3: 4 or 5 sounds. In each trial, users chose one of the two possible options (2AFC), responding how many sounds were played. Then, they were asked to name the colors associated with the sounds starting from left to right. The number of underwent trials differed between users due to the training time restrictions and individual differences in the multiple source localization ability, but there were no more than 6 trials per block. Group mean SL accuracy equaled 75.26 ± 14.39%. Two participants responded above, one below, and one at the individually set (due to the between-user differences in the number of underwent trials) chance levels (see Fig. 9). Also, in the case of MSL task we observed large differences in performance between participants (see supplementary material).

C. Usability

We administered two surveys to measure the Colorophone 2.2 usability and ergonomics. At the end of each training day in the laboratory, users answered the NASA Task Load Index (NASA-TLX) [29]. At the end of the whole training, users filled out a Polish version of the System Usability Scale [30], [31], [32] four times, judging each of the system elements separately, i.e.: glasses, camera, tablet with the controller, and application. Both tools were adjusted to the visually impaired users' needs—i.e., statements and possible answers were read aloud and multiple times if needed. Additionally, we simplified the NASA-TLX evaluation procedure—i.e., we skipped the first Weight part and restricted the response scale in the second Rating part to range from 0 to 10 points to make it more accessible for blind users. Moreover, for the sake of time restrictions, we administered the NASA-TLX scale once per training day instead of after every task.

1) SUS

We calculated SUS scores, grades, and adjective ratings [33] for each user for each system's parts separately, see Fig. 10. Then, we computed group scores (averaged scores of all users) for each element and general user's scores (averaged scores of all system's parts per user).

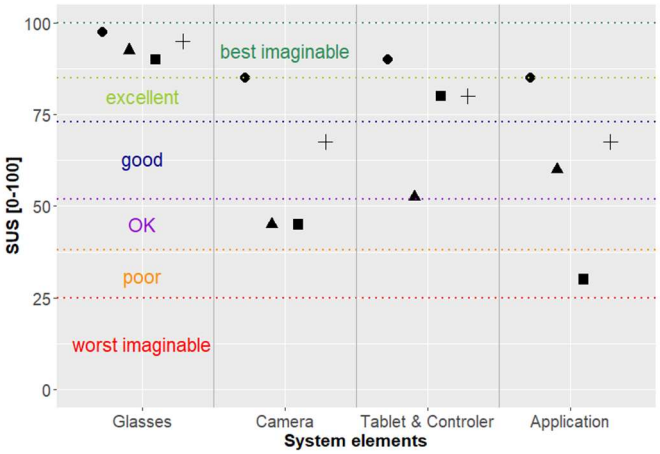

Fig. 10. System Usability Scale results with respect to the system's parts and users (coded by shapes). Dotted lines delineate the boundaries of SUS adjective rating scale [31]. Users' markers: User 2 – circle, User 3 – triangle, User 4 – square, User 5 – cross.

On average, users rated the glasses as the best imaginable (93.75 ± 3.23; grade A; see Fig. 10 and supplementary materials), the tablet with the controller as excellent (75.63 ± 16.12; grade C), and application (60.63 ± 22.95; grade D) and camera (60.63 ± 19.41; grade D) as good.

Nevertheless, we observed large individual differences between users in their ratings, with one user judging the whole system as the best imaginable (User 2: 89.38 ± 5.91; grade B), one as excellent (User 5: 77.5 ± 13.07; grade C), and two others as good (User 3: 62.5 ± 20.92; grade D & User 4: 61.25 ± 28.4; grade D).

2) NASA-TLX

We averaged NASA-TLX ratings for each user in each item separately, (see Fig. 11) and group ratings for each item. Additionally, we ran two-way repeated measures AVOVA to test the effects of item and day on ratings. We found a significant main effect of item ($F_{(5,60)} = 7.35$, $p < .001$). Neither main effect of day nor effect of item and day interaction were significant ($ps > .5$). The post-hoc pairwise comparisons (for details, see supplementary material) revealed that Mental (6 ± 2.25) was, on average, significantly greater than both Physical (3.31 ± 1.57) and Temporal (2.89 ± 1.68) Demand. Effort (7.03 ± 2.27) was greater than Mental, Physical, and Temporal Demand, and also than Frustration (3.62 ± 1.79). Moreover, Frustration was rated lower than Mental Demand and Performance (6.77 ± 1.61).

D. Identified usability issues

The list of identified usability issues was created based on the follow-up questions after each SUS item and is presented in the supplementary material.

The list includes comments of the users on each part of the Colorophone system, i.e., glasses, camera, controller & tablet and the application.

The glasses lack the physical volume control buttons, which makes the process of volume adjusting difficult.

The camera has a cumbersome cable. It is also hard to adjust the tilt angle. Finally, color constancy is questionable.

The controller has too small buttons that are hard to distinguish. It also enters the sleep mode too quickly.

The tablet, used as a processing unit is perceived as cumbersome.

Finally, the application has too slow temporal resolution. Users also commented on the low auditory image resolution (see supplementary material for details). Note that users proposed some suggestions for device development.

V. Discussion

A fundamental challenge every VASSD must answer is coding the relevant visual information into understandable auditory signals. In the case of the Colorophone, the chosen visual perceptual dimension is color. Thus, the basic function of the system is auditory color representation. However, since it is the first system that provides real-time, spatial color sonification that has only been used for several hours, the limits of its usability are yet to be determined. Nonetheless, the performed evaluation of the system allows for the identification of some of the strengths and weaknesses of the current prototype, providing recommendations on the further development of the device.

The Colorophone algorithm is based on color substitution, so an obvious first choice is to use the device to recognize the colors of objects. *Auditory color recognition and object identification tasks* results clearly show that test participants were able to successfully use the device to differentiate both artificial and natural objects based on simple colors (*Socks task*), but also more complex color patterns (*Cans task*). However, the Colorophone aims at representing not only colors, but also spatial localization and visual features of objects. As shown with the *virtual sound source localization task,* the former can be achieved with single sources. Results with more virtual sound sources are more ambiguous, however show that most of the testers were able to detect the location of multiple sources in space. This result is promising as this functionality is much less frequently successfully supported by SSDs proposed to date. It is also worth considering why some participants were not able to make use of the Colorophone successfully to perform the tasks. This may be the result of too short training, technical system limitations, or perceptual system limitations. The first limitation results from the timeline of the tests. Note that the a*uditory color recognition and object identification tasks* have been introduced on the second day of the tests, whereas the v*irtual sound source localization task* on the third day. Given the panoramic mode of the device has been introduced on the third day of the protocol, it may be that prolonged training can result in higher accuracy of the tests. This might be true, assuming that the other two types of limitations can be avoided. It is worth to note that the system's limitations (see section III) are twofold. Avoiding limitations from the first group (temporal delays, auditory color interference or auditory stream overlap) would support the more successful acquisition of the sensorimotor contingencies, whereas solving problems related to color consistency and reproduction would support color recognition itself. The order in which those limitations should be addressed taps into the question of the major function the device should serve. Our studies [26] indicated that over the training, participants started to prefer the panoramic mode. Importantly, they preferred the panoramic mode being faced with real-life tasks such as product category recognition (milk, can, bar, bread) or navigation in the natural environment (door counting at a corridor) – see [26] for details. Taking those results together, it seems the auditory representation of the spatial features of the visual scene should be supported by the device in the first place. Finally, it is worth remembering the limitations of the perceptual system itself. Recognition of the object's position as well as precise recognition of its features may be challenging due to the limitations of the human auditory system. Those functions could be supported by developing machine learning algorithms that recognize objects and provide automated feedback. However, such solutions would require a fundamental redesign of the system and would make it much more complex, that would affect usability judgments.

Another important challenge each SSD faces is the acceptance of the device by its target users. To test the usability of the Colorophone 2.2., we applied the System Usability Scale (SUS) with follow-up questions and NASA-TLX. We also identified the existing Colorophone's 2.2 usability issues and proposed some adjustments to increase the acceptance rate for future versions of the system. SUS results have been collected for each element of the system separately. The elements with the highest ratings were Glasses, Tablet and Controller. This is not surprising, as those were parts of the system making use of the commercialized products that unavoidably passed multiple usability tests before. The evaluation of the camera and application is more ambiguous. Nevertheless, the two out of four participants rated both components at least as good. The same ambiguity could be observed at the level of the averaged score, but all participants rated the system at least as good, whereas two of them assessed the system as excellent and best imaginable respectively. NASA-TLX also shows great interindividual variability between the participants. Note however, that due to limited availability of the end users group the number of participants is low. Future studies focused on the usability should include more users so that the numerical effects would be more conclusive. Nevertheless, it is worth noticing the system received the highest average scores in Effort, Performance and Mental Demand, whereas Physical and Temporal Demand as well as Frustration were rated relatively low. This means that using the device requires significant effort from participants, so longer training seems advisable. This is additionally supported by the lack of significant day effect, which suggests that participants did not reach the level of expertise that would lead to less effortful interaction with the Colorophone. Low frustration value suggests longer training is possible and hopefully would lead to device incorporation. However, this would require studies with super users, which is not possible at this level of system development. Future development should focus on the improvement of the system elements. Table V of the supplementary material lists a list of usability issues that should be addressed in the next version of the system. Many of the suggestions refer to the technical details that can be relatively easy to address (e.g., volume control on glasses, adjustable camera angle, size of the buttons, sleep mode, or tablet size). However, it is worth noticing that on top of that, in line with what has been noted above, some of the usability issues relate to two distinctive functionalities that

may be supported with the Colorophone device. Again, some of the comments refer to the veridicality of the color reproduction supporting the color recognition function, whereas others tap into functions supporting sensorimotor interaction (such as processing delay, but also adjustable camera angle, which is necessary to build such interaction successfully). Again, it puts into question which visual function users would want to support with the device. As mentioned, one can try to respond to this question based on the preference of the panoramic mode reported in [26]. However, an equally acceptable solution is to allow participants to configure the device so participants can individually select their preferred function and set a function of the device so it could be successfully supported.

Individual differences in the usability scores as well as those observed in the recognition and localization task, seem to clearly show that it is difficult to develop a device that would be equally accepted by all the testers. The solution might be to personalize the device so it would fit individual needs and already existing compensations of users, as well as their preferences as for which part of the available visual information is to be substituted by the device. It is however worth noting that individual customization of the device would influence the end price of the final product, resulting in lower availability. Thus, finding the balance between universality and customization of the devices providing optimal acceptance rate is yet to be determined.

VI. Conclusions

The introduced auxel-based framework provides an effective tool for the development of auditory displays. The system produces real-time, colorful stereo soundscapes that are interpretable by the users enabling color, color-based object recognition and object identification. The usability of the current version of the system was assessed from good to best imaginable, indicating high individual differences in users' evaluation of the system. Usability issues were identified and discussed. The second part of the evaluation, the NASA Task Load Index, indicates that although the use of the system is mentally demanding, the frustration rate remains low. Moreover, the self-assessed performance is high. However, the individual variability indicates the complex nature of factors influencing SSD acceptance that include the design of the device, quality and quantity of the performed training, as well as individual differences and needs of the users. Despite the abovementioned complexity, we conclude that the developed system extends the perceptual abilities of visually impaired persons and, after eliminating the existing usability issues, might become a useful visual rehabilitation device.

Acknowledgment

Special thanks to Simon Hviid del Pin for his valuable help with data analysis.

References

[1] R. Bourne *et al.*, "Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study," *The Lancet Global Health*, vol. 9, no. 2, pp. e130–e143, Feb. 2021, doi: 10.1016/S2214-109X(20)30425-3.

[2] R. R. A. Bourne *et al.*, "Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis," *The Lancet Global Health*, vol. 5, no. 9, pp. e888–e897, Sep. 2017, doi: 10.1016/S2214-109X(17)30293-0.

[3] "Vision impairment and blindness." https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment (accessed Feb. 21, 2022).

[4] E. National Academies of Sciences *et al.*, *The Impact of Vision Loss*. National Academies Press (US), 2016. Accessed: Feb. 21, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK402367/

[5] E. Brady, M. R. Morris, Y. Zhong, S. White, and J. P. Bigham, "Visual challenges in the everyday lives of blind people," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, Paris, France, 2013, p. 2117. doi: 10.1145/2470654.2481291.

[6] M. Ptito, M. Bleau, I. Djerourou, S. Paré, F. C. Schneider, and D.-R. Chebat, "Brain-Machine Interfaces to Assist the Blind," *Front. Hum. Neurosci.*, vol. 15, 2021, doi: 10.3389/fnhum.2021.638887.

[7] T. D. Wright and J. Ward, "Sensory Substitution Devices as Advanced Sensory Tools," in *Sensory Substitution and Augmentation*, Oxford: British Academy, 2018. doi: 10.5871/bacad/9780197266441.003.0012.

[8] J. Ward and T. Wright, "Sensory substitution as an artificially acquired synaesthesia," *Neuroscience & Biobehavioral Reviews*, vol. 41, pp. 26–35, Apr. 2014, doi: 10.1016/j.neubiorev.2012.07.007.

[9] Á. Kristjánsson *et al.*, "Designing sensory-substitution devices: Principles, pitfalls and potential," *RNN*, vol. 34, no. 5, pp. 769–787, Sep. 2016, doi: 10.3233/RNN-160647.

[10] K. Kokjer, "The Information Capacity of the Human Fingertip," *IEEE Trans. Syst., Man, Cybern.*, vol. 17, no. 1, pp. 100–102, Jan. 1987, doi: 10.1109/TSMC.1987.289337.

[11] A. Torralba, "How many pixels make an image?," *Vis Neurosci*, vol. 26, no. 1, pp. 123–131, Jan. 2009, doi: 10.1017/S0952523808080930.

[12] 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*, vol. 32, no. 2, pp. 247–257, 2014, doi: 10.3233/RNN-130338.

[13] S. Cavaco, J. T. Henriques, M. Mengucci, N. Correia, and F. Medeiros, "Color Sonification for the Visually Impaired," *Procedia Technology*, vol. 9, pp. 1048–1057, 2013, doi: 10.1016/j.protcy.2013.12.117.

[14] "Eyeborg," *Wikipedia*. Jan. 14, 2021. Accessed: Mar. 30, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Eyeborg&oldid=1000269982

[15] Z. Capalbo and B. Glenney, "Hearing Color: Radical Pluralistic Realism and SSDs," in *Proceedings of AP-CAP*, Tokyo, Japan, Oct. 2009, pp. 135–140.
[16] G. Bologna, B. Deville, and T. Pun, "On the use of the auditory pathway to represent image scenes in real-time," *Neurocomputing*, vol. 72, no. 4–6, pp. 839–849, Jan. 2009, doi: 10.1016/j.neucom.2008.06.020.
[17] D. Osiński, M. Łukowska, D. R. Hjelme, and M. Wierzchoń, "Colorophone 2.0: A Wearable Color Sonification Device Generating Live Stereo-Soundscapes—Design, Implementation, and Usability Audit," *Sensors*, vol. 21, no. 21, p. 7351, Nov. 2021, doi: 10.3390/s21217351.
[18] S. Maidenbaum, S. Abboud, and A. Amedi, "Sensory substitution: Closing the gap between basic research and widespread practical visual rehabilitation," *Neuroscience & Biobehavioral Reviews*, vol. 41, pp. 3–15, Apr. 2014, doi: 10.1016/j.neubiorev.2013.11.007.
[19] G. V. Elli, S. Benetti, and O. Collignon, "Is There a Future for Sensory Substitution Outside Academic Laboratories?," *Multisens Res*, vol. 27, no. 5–6, pp. 271–291, 2014, doi: 10.1163/22134808-00002460.
[20] "Colorophone," Mar. 29, 2021. https://www.colorophone.com (accessed Mar. 29, 2021).
[21] D. Osinski and D. R. Hjelme, "A Sensory Substitution Device Inspired by the Human Visual System," in *2018 11th International Conference on Human System Interaction (HSI)*, Gdansk, Jul. 2018, pp. 186–192. doi: 10.1109/HSI.2018.8431078.
[22] J. F. Blinn, "What is a pixel?," *IEEE Computer Graphics and Applications*, vol. 25, no. 5, pp. 82–87, Sep. 2005, doi: 10.1109/MCG.2005.119.
[23] R. G. Farrell and J. R. Kozloski, "Sound source selection for aural interest," US9693009B2, Jun. 27, 2017 Accessed: Jun. 22, 2021. [Online]. Available: https://patents.google.com/patent/US9693009/en
[24] S. S. Stevens, "On the psychophysical law.," *Psychological review*, vol. 64, no. 3, p. 153, 1957.
[25] "Bluetooth Audio Sunglasses | Bose," Mar. 09, 2021. https://www.bose.com/en_us/products/frames.html (accessed Mar. 09, 2021).
[26] M. Lukowska, W. Kałwak, D. Osinski, J. Janik, and M. Wierzchon, "How to teach a blind person to hear colours? Multi-method training for a colour-to-sound sensory substitution device – design and evaluation," *International Journal of Human-Computer Studies*, p. 102925, Sep. 2022, doi: 10.1016/j.ijhcs.2022.102925.
[27] J. Winawer, N. Witthoft, M. C. Frank, L. Wu, A. R. Wade, and L. Boroditsky, "Russian blues reveal effects of language on color discrimination," *Proceedings of the national academy of sciences*, vol. 104, no. 19, pp. 7780–7785, 2007.
[28] P. Gwiaździński, M. Reuter, J. Dubis, D. Osiński, and M. Wierzchoń, "Research protocol. Usability of Sensory Substitution systems: test and comparison of BrainPort and Colorophone devices." Rochester, NY, Jul. 18, 2022. Accessed: Jul. 26, 2022. [Online]. Available: https://papers.ssrn.com/abstract=4165732
[29] N. NASA, "Task Load Index (TLX) v. 1.0 Manual," *NASA, NASA-Ames Research Center Moffett Field*, 1986.
[30] P. W. Jordan, B. Thomas, I. L. McClelland, and B. Weerdmeester, *Usability Evaluation In Industry*. CRC Press, 1996.
[31] A. Borkowska and K. Jach, "Pre-testing of polish translation of System Usability Scale (SUS)," in *Information Systems Architecture and Technology: Proceedings of 37th International Conference on Information Systems Architecture and Technology–ISAT 2016–Part I*, 2017, pp. 143–153.
[32] J. Brook, "SUS: a 'quick and dirty' usability scale," *Usability evaluation in industry*, 1996.
[33] A. Bangor, P. Kortum, and J. A. Miller, "Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale," vol. 4, no. 3, p. 10, 2009.