A wearable vision-to-audio sensory substitution device for blind assistance and the correlated neural substrates

Yaozu Ye[1], Kaiwei Wang*[1], Weijian Hu[1], Huabing Li[1], Kailun Yang[1], Lei Sun[1], Zuobing Chen[2]

[1]State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, Hangzhou 310027, China

[2]Department of Surgery, First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, Zhejiang Province, China

* wangkaiwei@zju.edu.cn +86-571-8795-3154

Abstract. There are a very few people who have the ability to "see" the surroundings by the echoes, which is called echolocation. The study of the brain mechanism of echolocation can not only help to improve the blind assistance device, but also provides a window into the research of brain's plasticity. In this paper, we developed a wearable system to transform the spatial information captured by camera into a voice description and fed it back to blind users which is inspired by echolocation. After our online virtual scene training, users can easily discriminate object location in the camera's view, motion of the objects, even shape of the objects. Compared with natural echolocation, it's easier to learn and be applied in daily life. In addition, the device achieves high spacial resolution. In this study, two trained blind subjects and two non-trained sighted subjects were tested by using functional Magnetic Resonance Imaging (fMRI). We obtain the fMRI images of the subjects' brain activity when they were listening to the sound of the wearable prototype. Intriguingly, we find that after training with the blind assistance system, the blind' visual area of the brain have been activated when they are dealing with the acoustic feedback from the device.

1. Introduction

It is well known that bats can emit ultrasound waves and use the returned ultrasound echoes to sense the surrounding environments, and even use it to track preys. In fact, bats simply follow some specific rules to understand the echo signal. For example, the bats will regard the path as free space to pass if there is no echo sound in front [1], and the place where there is no sound wave back from the horizontal ground will be considered as the water surface [2]. Actually, some humans have similar capabilities. They can make their own sounds and then use the echo of the sound to sense the surrounding environments. Daniel Kish [3] was reported to sense the surrounding scenes by using the month-click and the echo, and even determine what kind of object that is in front of the path. He was blind from retinal cancer since childhood, but by echolocation he could climb and even ride a bicycle. The ability which can be used to sense the surrounding environments without relying on vision is called echolocation [4].

However, because the information contained in the echo is too complicated, it is extremely difficult to train the blind to master it. The spatial resolution of echo localization is too low and echolocation

cannot be used when they can't make signal noises, for example, in the library. Inevitably, blind individuals will suffer from being given excessive attention by other people, since vocalization (making signal noises) is one of the essential steps of echolocation. Although echolocation can't be widely used due to many restrictions, it has already been demonstrated that humans do have the ability to perceive the environments through sound, and vision-to-audio sensory substitution is achievable.

It is a challenging problem for a wearable device to obtain information such as the current terrains and scenes [5,6], and even more interesting regarding how to effectively and quickly feedback to the users. One straightforward way of human-computer interaction is through semantic expression, that is, using a piece of speech to describe the current scene. However, for a complex scene with multiple objects distributed at different orientations and distances, the efficiency of the interaction of the voice description is insufficient to meet the requirements for real-time access to information.

Some researchers took example of bats' echolocation to design their vision-to-audio Sensory Substitution Devices (SSDs) [7-9]. They utilized ultrasonic sensors to record the ultrasonic signal echoes, then lowered their frequencies into the human auditory range, finally fed it back to users. These types of SSDs, like human echolocation, are difficult to learn.

In addition to the direct use of echolocation principles, there were also SSDs that obtained spacial information through camera and then encoded environmental information into the sound. The vOICe [10], a widely investigated vision-to-audio SSD, used a camera to get a gray-scale map of the current scene, scanned and encoded each column of the image, and then serially fed the sound back to the user through headphones. Although vOICe could convey some information about shapes and edges, it doesn't work when the objects have complex textures since it only conveyed the gray-scale map of the current scene. More importantly, it couldn't directly interpret the depth information which is of more importance for the visually impaired to avoid obstacles, and it is infeasible to deliver real-time assistance as it needs to scan the gray-scale map.

For the visually impaired, depth information is arguably more important than gray-scale information. Therefore, we encode the current 3D scene information (depth-map) into an audio representation in real time. Specifically, we modulate the depth, elevation and azimuth of every pixel in the current observed image into the loudness, frequency, and phase difference information of the sound (see in 2.1). After training, the blind can form their own ability of "hear the object".

Human echolocation and vision-to-audio SSDs are currently opening up a vibrant area of research in psychology and neuroscience. The first study about the neural mechanisms of echolocation was performed by De Volder et al [8]. Specifically, they investigated the neural networks involved when using an ultrasonic echolocation device. They found that, among the blind subjects, the processing of sound from the ultrasonic echolocation device would activate their brain in Brodmann area (BA) 17/18 (i.e., the early 'visual' cortex). In 2005, their laboratory conducted a study by using a SSD that has been called artificial retina [11]. They demonstrated that some brain areas of the visual cortex are relatively multimodal and may be recruited for depth processing via a sense other than vision. In 2007, by using vOICe, another study [12] reached a conclusion that the lateral-occipital tactile-visual area is driven by the presence of shape information. As for human echolocation, Lore Thaler et al [13] found that it activates the visual region of the brain, and the brain images of these activation regions are modulated by the information carried by the echo.

Although there're already many SSDs mentioned above, we aim to develop a more intuitive, real-time and high-resolution blind assistance device, which enables independent mobility for visually impaired individuals. The purpose of this study is to introduce a new vision-to-audio SSD and investigate the neural substrates when the SSD is used which would indicate some cognitive process.

2. Material and method

2.1. Image to sound mapping principles

Based on our previous work [5,6], we have developed a wearable system to transform the spatial information captured by camera into a voice description and feed it back to the users, which is inspired by echolocation. (See figure 1 for an intuitive understanding of the vision-to-audio mapping principle) There are only three simple rules in our system's general image to sound mapping.

1. Left and right:
   By binaural adjustments of the left-to-right ear's sound delay and the sound intensity, we can make user perceive the azimuth angle of the object. (Azimuth angle range: 0 ~ 180 degrees)
2. Up and down:
   Pitch encodes elevation: the higher the pitch, the higher the position of the visual pattern. (Pitch range: 220 Hz ~ 3520 Hz)
3. Far and near:
   Loudness encodes distance: louder means closer.

First we use Intel RealSense D435 to obtain spatial information of the current scene (depth map) in real time. Afterwards, we down-sample the depth map and obtain a low resolution depth map of 43×43 pixels. Traversing each pixel of the depth map after down-sampling, each pixel is a sound source whose frequency, loudness and left and right channels are modulated according to the above principles of image to sound mapping.

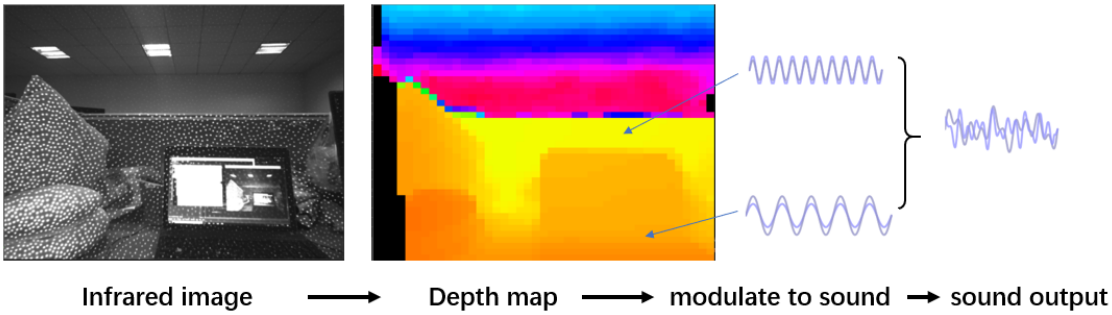

Figure 1. Image to sound mapping principle.

In this study, in order to make it easier for users to carry and wear our equipment, the hardware such as processors must be small and light, which limits the performance of the processor. The program calls the API provided by Intel's RealSense SDK 2.0 to enable and stream from the RealSense D435 camera. FMOD is a professional game sound engine that can easily generate sound effects and produce stereo effects. This program calls the FMOD API to control the output sound. Figure 2 shows the flow chart of the program.

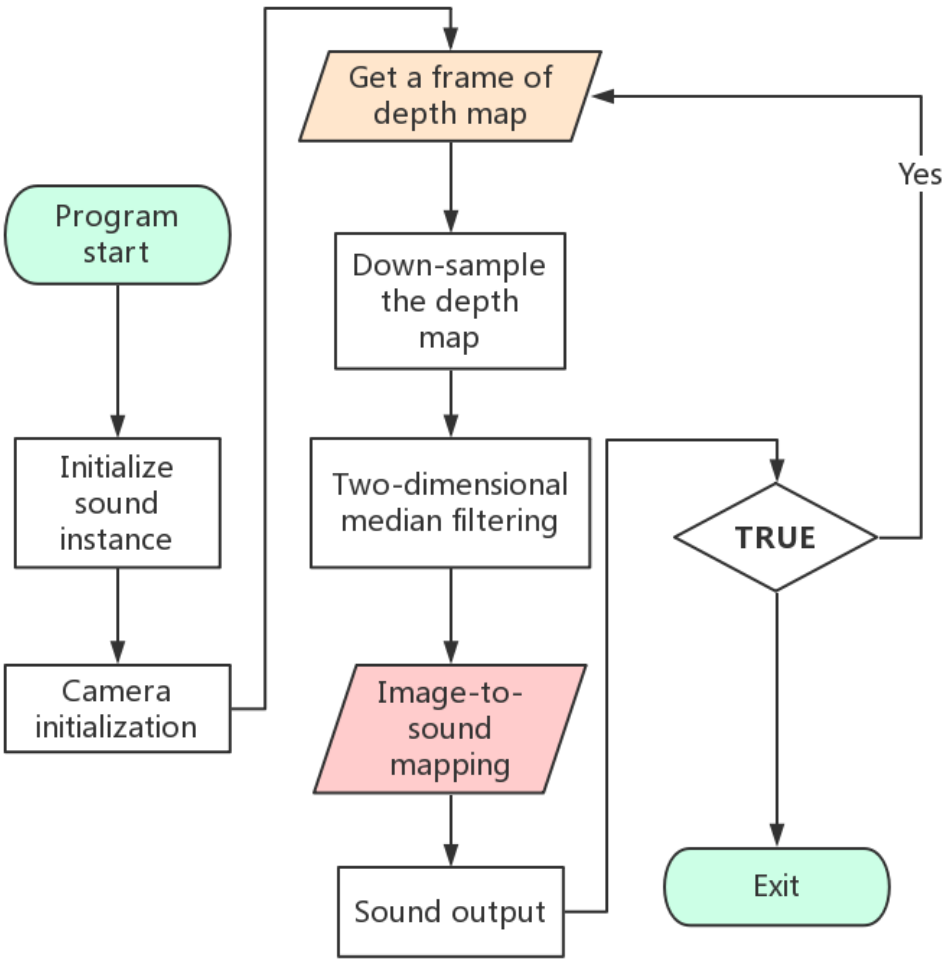

Figure 2. Flow chart of the program.

2.2. Hardware

As figure 3 shows, the blind assistance device consists of a depth camera (Intel RealSense D435), a processor unit and a pair of earphones. The depth camera captures the spacial information and transfers the depth map to the processor and the processor translates the depth map into sound descriptions based on our rules.

- Intel RealSense D435: Includes an infrared speckle emitter, two left and right infrared cameras and a RGB camera. Get the spatial information of the current scene in real time and get the depth map.
    - Depth Field of View (FOV) — (Horizontal × Vertical × Diagonal): 91.2 × 65.5 × 100.6 (+/ - 3°).
    - Depth Stream Output Resolution: 640 × 480.
    - Depth Stream Output Frame Rate: 30 fps.
    - Minimum Depth Distance (Min-Z): 0.2m.
    - Maximum Range: Approx.10 meters; Varies depending on calibration, scene, and lighting condition

- Customized processing unit [14]: A mini-computer with a portable power source.
- A pair of earphones: Transfer the audio information to the user.

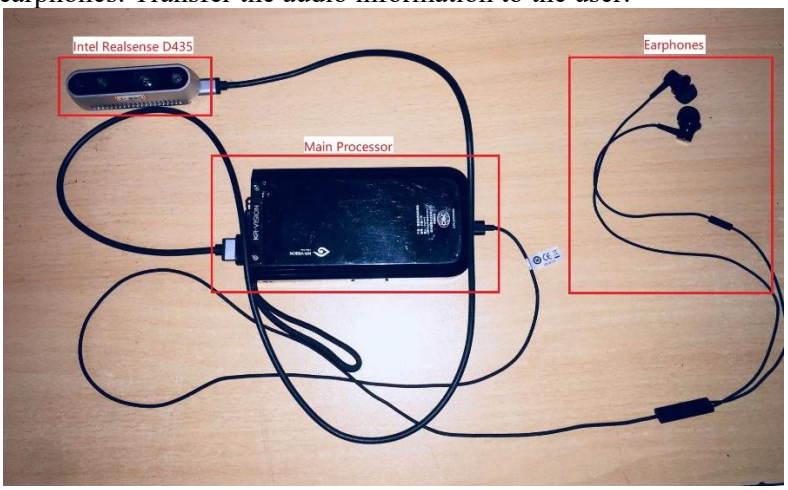

Figure 3. The blind assistance system.

2.3. Simulation experiment

To verify the feasibility of the program of this study, we have first recruited 20 blind volunteers to participate in our "Sound Recognition Training Camp" and make them familiar with the image-to-sound mapping principles.

The basic component of a spatial graphic is two-dimensional shapes, since learning from the sound of an isolated two-dimensional shaped figure is an ideal choice. However, the actual scene always has interference from other objects, and it is difficult to achieve an ideal situation with only one two-dimensional shaped object. In order to solve this problem, this study adopted a computer programming simulation method for the basic training of the blind, and used the simulated depth map (as shown in figure 4) instead of the depth map of the real scenes to train the volunteers to learn the sound coding scheme. Since the depth map at this time is simulated by a computer, it is convenient to obtain sufficient audio samples of graphics of different shapes and different depth positions.

In order to make the training process more attractive for the blind volunteers, so that they were more willing to participate in, we divided the training into two parts: basic training and problem-solving training.

In the basic training, we provide the sound of points in different positions, motion states, sounds of lines, and examples of sounds of triangles, circles, squares, and rectangles of different positions, sizes, and rotation states.

Users can learn the basic principles of our sound mapping by listening to these sounds, and learn to recognize the sounds of various basic graphics.

After the volunteers went through the basic training session, we provided test questions to test them. The content of the test questions was also the judgment of different positions, different sizes, different rotation angles, and different motion states of the basic shapes. The test questions were also divided into general mode and difficult mode. In the general mode, only one graphic would be presented on the depth map. In the difficult mode, the depth map would contain two graphics, and blind volunteers were required to determine the position, shape, motion state, size and other information of the two graphics in the depth map.

In order to facilitate the training and testing of volunteers, we have made the above basic training and test into the form of web links, allowing volunteers to receive remote online training. Blind volunteers completed the training and testing in the link with the help of mobile phone screen reader software.

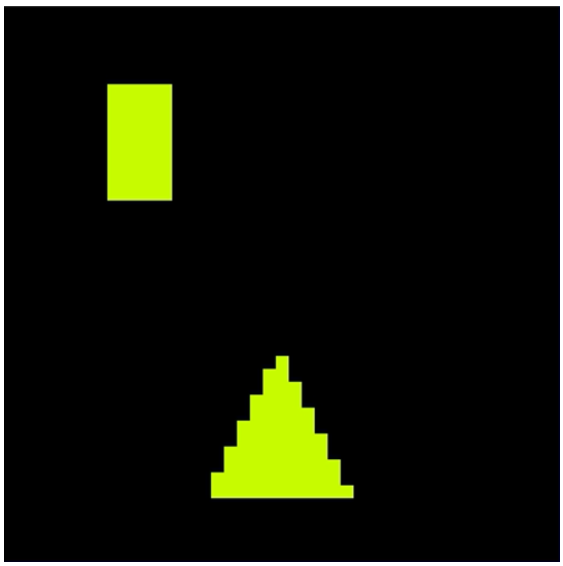

Figure 4. The simulated depth map.

2.4. Real-world training based on sound coding scheme

Since the previous spatial graphics simulation experiment was an online training, the actual training time of the volunteers could not be guaranteed. We recruited two blind subjects to participate in our offline training. One of the blind subject is Congenital Blind (CBS) but have light sensation. Another blind subject was blind at the age of 8 (Late Blind Subject, LBS). LBS is totally blind without light sensation. Additionally, the participant CBS was 25 years old and the LBS was 24 years old when we conducted the study.

Training Arrangement: After 5 days of daily two-hour simulation training (online test questions), we had ensured that volunteers had fully learned and mastered the principles and rules of image–to-sound mapping, and had become familiar with the sound of basic two-dimension graphics descriptions. After that, the two blind subjects were required to use the blind assistance device in the daily life, which help them to complete the two-dimension to three-dimension transition training. This phase lasts for 12 days and the subjects are required to train one hour a day.

2.5. Resolution test

To evaluate vision-to-audio mapping scheme of this study in spacial resolution respect, we conducted a resolution test experiment.

The resolution test is mainly divided into three parts of horizontal, height and depth. The experiment was carried out on a flat table with no other objects, ensuring that the experimenter was not disturbed by the sound of other objects on the table. Because the blind assistance device was designed to make the sound discriminable when depth varies within 1.5 meter, the sound feedback outside 1.5 meters would be at a low volume level. Therefore, the resolution test was designed to be conducted at 35cm.

Experimental props: Rectangular cardboard with a length of 14 cm and a width of 10.5 cm.

Depth resolution test: We place two props cardboards (two boards close together) side by side at the space of 35 cm away from the camera (to ensure that the distance between the camera and the two boards is the same). In the testing process, the left cardboard would be slowly moved back and forth to change the distance to the camera while keeping the distance from the right cardboard to the camera, the subjects needed to identified the distance from the cardboard to the camera by listening to the sound feedback from the SSD. The minimum depth difference between the two cardboards when the subject can sense the depth change, is the depth resolution of the device at 35 cm.

Horizontal resolution test: We place a piece of prop board vertically at 35 cm from the camera, and slowly increase the horizontal width of the cardboard. The subjects needed to discern the change in the

horizontal width of the cardboard by listening to the sound feedback from the SSD. The width increased when the subject was exactly able to sense the horizontal width change is the horizontal resolution of the equipment at 35 cm.

Height resolution test: Two props cardboard (40 cm apart) are placed at 35 cm from the camera. We slowly increased the height of the cardboard on the left, while keeping the height of the cardboard on the right. The experimental subjects were required to discern the height change of the cardboard by listening to the sound feedback from the experimental equipment. The minimum height difference between the two cardboards when the subject can exactly sense the height change is the height resolution of the device at 35 cm.

2.6. Functional magnetic resonance imaging

Functional magnetic resonance imaging (fMRI) is a technique performed under high magnetic field. It magnetizes the hydrogen atoms in the human body and resonates with electromagnetic waves to generate signals, which are then converted into images by computer processing to detect activity in different areas of the brain. Moreover, the fMRI procedure is noninvasive and nonradiative.

We recruited six subjects and they were divide to three groups. Written informed consent was obtained before the study.

- Control group: Two Un-Trained Sighted people (UTS1 and UTS2). They would accept the whole brain fMRI imaging when given sound stimulation. The UTS1 and UTS2 didn't even know about the rule of image-to-sound mapping.
- Experimental group 1: Two trained blind people (The CBS and the LBS, see in 2.4). They would accept whole brain fMRI imaging when given sound stimulation.
- Experimental group 2: Two Trained Sighted people (TS1 and TS2). They would accept whole brain fMRI imaging when given sound stimulation. (TS2 was not sufficiently trained.)

There were 5 pre-recorded sound stimuli, each of which lasted for 10 seconds.

Sound stimuli 1: Silence, as a contrast to other sound stimuli.

Sound stimuli 2 and 3: Acoustic feedback of the simulated spatial pattern.

Sound stimuli 4 and 5: Acoustic feedback of the SSD in real scene.

The fMRI scanned a total of 25 cycles, and the sound stimulation of the cycle varied with the number of cycles. Figure 5 shows the flow chart of the fMRI procedure.

For the control group (sighted people), since they did not know the vision-to-audio mapping principles, we required them to pay attention only to the tone changes and the stereo effects of the sound stimuli. For the experimental group, we required the participants to pay attention to the picture represented by the sound stimulation. During the experiment, the room lights were turned off and the subjects were asked to close their eyes. (The experimental routine and related parameters follow the fMRI experiment in Ref 13.)

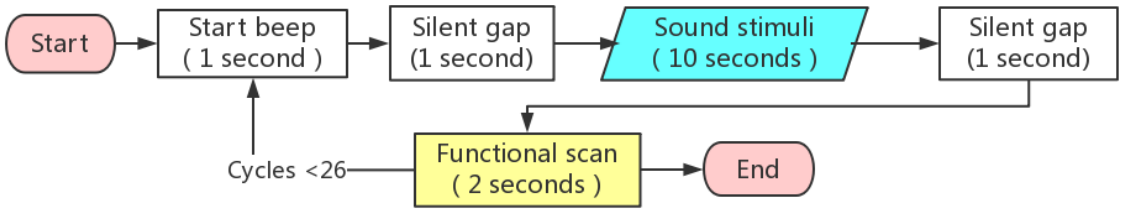

Figure 5. The flow chart of fMRI experiment.

3. Result and discussion

3.1. Simulation experiment

After online training and test, we collected the volunteers' feedback. The feedback collected shows that volunteers can easily discriminate object location in the camera's view, motion of the objects. Even they could discriminate different shapes (triangle\rectangle\circle), the angle that the shapes rotated, although it took time to figure it out. The result has well verified the feasibility of the vision-to-audio mapping principles.

3.2. Resolution test

Table 1. Result of the resolution test at a depth of 35 cm

Subject	Test number	Resolution (cm)					
		Height	Mean (height)	Horizontal	Mean (horizontal)	Depth	Mean (depth)
LBS	1	2.8	2.7	2.1	2.1	0.7	0.8
	2	3.0		2.3		1.0	
	3	2.3		1.9		0.8	
CBS	1	3.6	3.5	2.1	2.0	0.9	1.0
	2	3.3		2.0		1.1	
	3	3.5		1.8		0.9	

As table 1 shows, the SSD achieved high resolution when the object is in the vicinity (at 35 cm away). The resolution obtained in this study was greatly improved compared to the depth resolution of 40 cm at 170 cm in human natural echolocation [15].

3.3. Functional magnetic resonance imaging

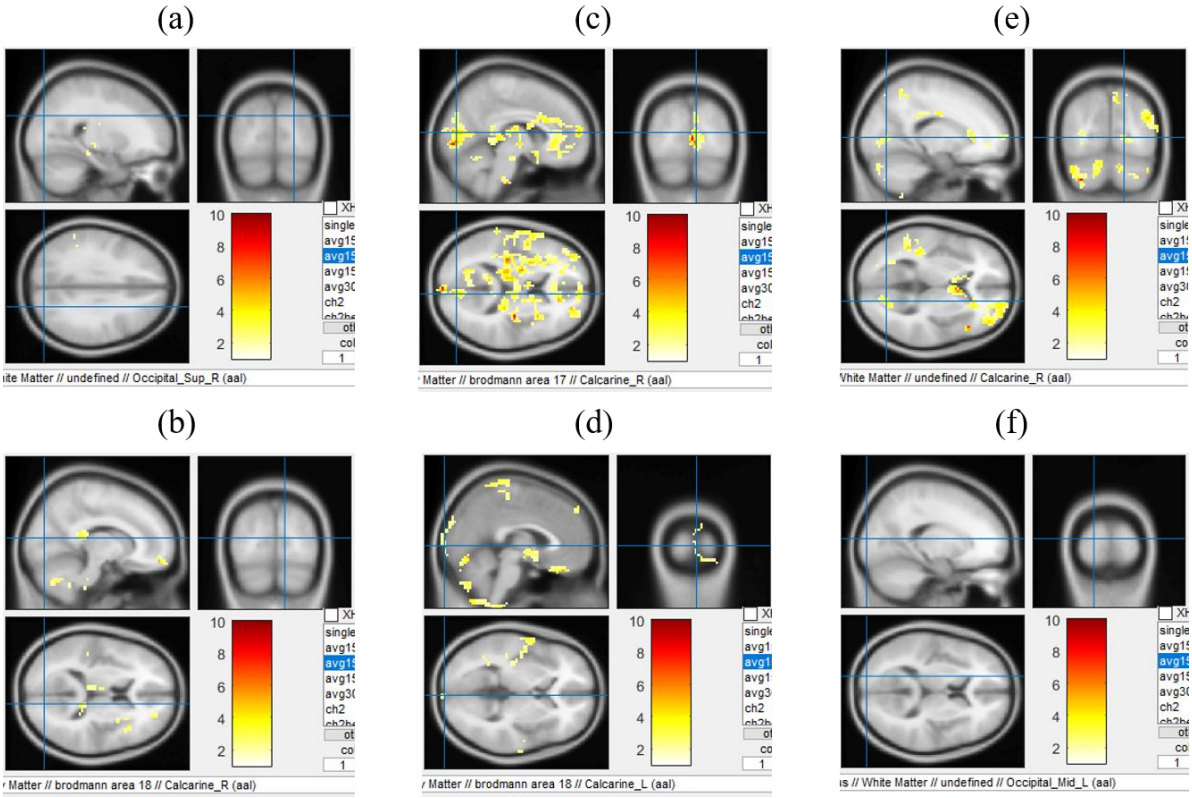

Figure 6. The brain activation when subjects were dealing with sound 3. (a) UTS1 (b) UTS2 (c) LBS (d) CBS (e) TS1 (f) TS2. The highlight area is the active area.

All image was performed at Laboratory of Institute of Biomedical Engineering, Zhejiang university, on a 3-tesla, whole-body fMRI system. In this way, fMRI data were analysed by using SPM12 and xjView (Matlab toolbox) on Matlab2017b.

Generally, there were activations in occipital cortex of CBS and LBS in stimuli 2 to 5 when compared to stimuli 1(silence). When we compared activation with stimuli 3 to stimuli 1 (silence), as figure 6 shows, there were recuitment in occipital cortex in both after-trained blind subjects. In contrast, we didn't observe any activation in occipital cortex in non-trained sighted individual. However, as for after-trained sighted subjects, we found activations in occipital cortex in TS1 but not in TS2. It was seemingly due to the insufficient training of TS2. The UST1 had no activations in all sound stimuli compared to silence, but UTS2's occipital cortex showed evident activation in stimuli 2, 3 and 5, compared to silence. More details are shown in Table 2.

Table 2. Result of neuroimaging experiment.

Group	Subjects	Visual cortex activation compared to stimuli 1 (silence)			
		Sound stimuli 2	Sound stimuli 3	Sound stimuli 4	Sound stimuli 5
Experimental group 1	LBS	√	√	√	√
	CBS	√	√	√	√
Experimental group 2	TS1	√	√	√	√
	TS2	√	×	×	×
Control group	UTS1	×	×	×	×
	UTS2	√	×	√	√

4. Discussion and conclusion

In this study, we have developed a new kind of blind assisting device, which maps the three dimensions of the spatial information (depth map) to the three dimensions of the sound in real time. Users would get high-resolution and real-time environmental information (depth maps) from the SSD. Specifically, users can use the device to identify simple shapes, the position of objects, and the direction of motion in real time. From the results of the fMRI experiment we found that after the training, the visual cortex of the blind subjects has been recruited when processing the sound feedback from the device. This study not only presented a new prototype of blind assisting device, but also provided guidance for the design of blind accessories. On the one hand, the fMRI results helped us better understand the correlated neural substrates of the blind when using SSD. On the other hand, it could be used to evaluate the effect of SSD. This is a promising step in the direction of developing a SSD which enable independent mobility for visually impaired individuals.

However, does the activation in visual cortex mean that when using the SSD, the blind subjects has an imagery sense of the current scene in his brain, that is, "seeing" the world through the sound? To answer this questions, more research efforts are needed.

Based on the fMRI results, we can conclude that the blind person has been trained by the vision-to-audio sensory substitution device have visual cortex participation in the phase of processing the sound information fed back by the device. Nonetheless, since there is no fMRI data before the blind subjects trained, the conclusion is not very convincing. On the other hand, due to the time and resource limits, only six subjects were recruited in the fMRI experiment. However, in fMRI experiments, individual differences could influence the results. More subjects are needed to eliminate the effects of individual differences. Therefore, our next step is to conduct a more complete fMRI experiment. Specifically, more blind and sighted subjects will be recruited. The division of sound stimuli will be more refine. Moreover, the fMRI data of brain activation before subjects trained will be collected to make our conclusions more reliable and convincing.

5. References

[1] Greif, S., Zsebők, S., Schmieder, D. and Siemers, B.M., 2017. Acoustic mirrors as sensory traps for bats. Science, 357(6355), pp.1045-1047.
[2] Greif, S. and Siemers, B.M., 2010. Innate recognition of water bodies in echolocating bats. Nature communications, 1, p.107.
[3] Kish, D., 2009. Human echolocation: How to “see” like a bat. New Scientist, 202(2703), pp.31-33.
[4] Griffin, D.R., 1944. Echolocation by blind men, bats and radar. Science, 100(2609), pp.589-590.
[5] Yang, K., Wang, K., Hu, W. and Bai, J., 2016. Expanding the detection of traversable area with RealSense for the visually impaired. Sensors, 16(11), p.1954.
[6] Yang, K., Wang, K., Bergasa, L.M., Romera, E., Hu, W., Sun, D., Sun, J., Cheng, R., Chen, T. and López, E., 2018. Unifying Terrain Awareness for the Visually Impaired through Real-Time Semantic Segmentation. Sensors, 18(5), p.1506.
[7] Ifukube, T., Sasaki, T. and Peng, C., 1991. A blind mobility aid modeled after echolocation of bats. IEEE Transactions on biomedical engineering, 38(5), pp.461-465.
[8] De Volder, A.G., Catalan-Ahumada, M., Robert, A., Bol, A., Labar, D., Coppens, A., Michel, C. and Veraart, C., 1999. Changes in occipital cortex activity in early blind humans using a sensory substitution device. Brain research, 826(1), pp.128-134.
[9] Sohl-Dickstein, J., Teng, S., Gaub, B.M., Rodgers, C.C., Li, C., DeWeese, M.R. and Harper, N.S., 2015. A device for human ultrasonic echolocation. IEEE Transactions on Biomedical Engineering, 62(6), pp.1526-1534.
[10] Meijer, P.B., 1992. An experimental system for auditory image representations. IEEE transactions on biomedical engineering, 39(2), pp.112-121.
[11] Renier, L., Collignon, O., Poirier, C., Tranduy, D., Vanlierde, A., Bol, A., Veraart, C. and De Volder, A.G., 2005. Cross-modal activation of visual cortex during depth perception using auditory substitution of vision. Neuroimage, 26(2), pp.573-580.
[12] Amedi, A., Stern, W.M., Camprodon, J.A., Bermpohl, F., Merabet, L., Rotman, S., Hemond, C., Meijer, P. and Pascual-Leone, A., 2007. Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex. Nature neuroscience, 10(6), p.687.
[13] Thaler, L., Arnott, S.R. and Goodale, M.A., 2011. Neural correlates of natural human echolocation in early and late blind echolocation experts. PLoS one, 6(5), p.e20162.
[14] Krvision: http://www.krvision.cn/
[15] Schörnich, S., Nagy, A. and Wiegrebe, L., 2012. Discovering your inner bat: echo–acoustic target ranging in humans. Journal of the Association for Research in Otolaryngology, 13(5), pp.673-682.