Learning to perceive with a visuo-auditory substitution system: Localization and object recognition with "The vOICe"

Malika Auvray[1, 2], Sylvain Hanneton[3], & J. Kevin O'Regan[4]

1. Department of Experimental Psychology, Oxford University, Oxford, UK

2. COSTECH-Groupe Suppléance Perceptive, Université de Technologie de Compiègne, Compiègne, France

3. Laboratoire Neurophysique et Physiologie du Système Moteur, CNRS UMR 8119 & UFR STAPS, Université Paris 5 René Descartes, Paris, France

4. Laboratoire Psychologie de la Perception, CNRS FRE 2929 & Université Paris 5 René Descartes, Paris, France

RUNNING HEAD: Learning to perceive with a visuo-auditory substitution system

ADDRESS FOR CORRESPONDENCE:

Malika Auvray, Department of Experimental Psychology, University of Oxford,
South Parks Road, Oxford, OX1 3UD, UK

E-mail: malika.auvray@psy.ox.ac.uk

TEL: +44-1865-271380

FAX: +44-1865-310447

RESUBMITTED TO: *PERCEPTION* (Accepted, May 2006)

Text pages: 36

Number of figures: 6

Abstract

We investigated to what extent participants can acquire the mastery of an auditory substitution of vision device ("The vOICe") using dynamic tasks in a 3-dimensional environment. After extensive training, participants took part in four experiments. The first experiment explored locomotion and localization abilities. Participants, blindfolded and equipped with the device had to localize a target by moving the hand-held camera, walk towards the target, and point at it. In a second experiment, we studied the localization ability in a constrained pointing task. A third experiment explored participants' ability to recognize natural objects via their auditory rendering. In a fourth experiment we tested the ability of participants to discriminate among objects belonging to the same category. We analyzed participants' performance both from an objective and a subjective point of view. The results showed that participants, through sensorimotor interactions with the perceptual scene while using the hand-held camera, were able to make use of the auditory stimulation in order to obtain the information necessary for locomotor guidance, localization and pointing, as well as for object recognition. Furthermore, analysis from a subjective perspective allowed insights into participants' qualitative experience and into the strategies they used in order to master the device, and thus to pass from a kind of deductive reasoning to a form of immediate apprehension of what is being perceived.

Key words: Sensory substitution, perception, sensory modality, prosthesis, space

Introduction

Sensory substitution systems allow information coming from an artificial receptor to be processed by a different sensory organ from the one normally used to transduce this information (Bach-y-Rita et al 1969). Two main categories of systems designed to compensate for the loss of vision exist: Visual-to-tactile substitution devices that convert visual pictures into “tactile” pictures and visual-to-auditory substitution devices that convert visual images into sounds.

Many visual-to-tactile substitution devices have been developed. In most systems, optical images picked up by a video camera are translated into electrical or vibratory stimulation applied to the skin of a part of the body (abdomen, back, fingertip, forehead, tongue, etc.). Many studies have shown the feasibility of such devices. Research has shown the ability of participants to perform simple form recognition (Kaczmarek and Haase 2003; Sampaio et al 2001), reading (Bliss et al 1970; Craig 1976, 1980, 1981, 1983; Loomis 1974, 1980, 1981), and localization (Janson 1983; Lemaire 1999). Some studies have also shown that users of visual-to-tactile substitution devices can make perceptual judgments using perspective, parallax, looming, and zooming, as well as depth estimates (Bach-y-Rita et al 1969; Epstein 1985). However, visual-to-tactile conversion systems are faced with certain technological limitations such as the need to use a highly-sensitive skin surface and problems such as skin irritation or pain (Bach-y-Rita 1972). Furthermore, the substantive energy consumption of tactile stimulators limits the autonomy of portable versions of these devices (Lenay et al 2003).

The other important field of visual substitution research to have been investigated concerns the auditory substitution of vision. Audition presents many advantages: The human auditory system is able to deal with complex and rapidly changing sound patterns such as speech, even in a noisy environment (Hirsh 1988). The auditory system has fine frequency-discrimination and intensity-discrimination thresholds. Furthermore visual-to-auditory substitution systems depend on only a simple interface, namely headphones (Capelle et al 1998) and the generation of auditory stimuli requires little energy. Digital sound processing is a very common technology and is included in many mobile smart systems (mobile phones, pocket computers, etc.). However, experimental studies involving the auditory substitution of vision are less numerous than those involving the tactile substitution of vision. The two main auditory display-based substitution systems involve echolocation devices and devices based on image-to-sound conversion.

Echolocation devices are based on the same principles as sonar. An ultrasound source/receptor emits a stream of clicks and/or FM signals. Receptors use telemetry in order to determine the distance between the source and the distant object. This method consists of calculating the time taken by an ultrasonic signal to reach an object and return by reflection to the generator. Signals are then converted into an audible frequency range and transmitted to the user’s ears via headphones, giving them an indication concerning the distance and direction of a distant object. For instance, distance can be coded by pitch and the horizontal position by inter-aural disparity (e.g., UltraSonic Torch, Sonic Glasses: Kay 1964, 1980, 1985; Sonic Pathfinder: Heyes 1984). These systems can be helpful for locomotion and guiding movements of blind persons (Heyes 1984; Kay 1964) and can also give information about the spatial layout of three-dimensional scenes (Hughes 2001).

In systems that do not use a telemetry approach, optical images picked up by a camera are converted into sound and transmitted to users via headphones. Three main systems have been studied: The vOICe developed by Meijer (1992), the PSVA (Prosthesis for Substitution of Vision by Audition) developed by Capelle et al (1998), and the device developed by Cronly-Dillon (1999). These three systems convert the vertical position of a luminous object in the video into different audio frequencies, with high pitched sounds corresponding to upper locations and low pitched sounds corresponding to lower locations in the video image. This scheme has been shown by Melara and O'Brian (1987) to correspond to a natural cross-modal correspondence. Participants were presented with a sequence of visual stimuli whose elevation (upper vs. lower) varied and at the same time with a low or high frequency tone. Participants responded faster to the location of the object when the stimuli were congruent (a high-pitched sound with an object in an upper location) than when they were incongruent (a low-pitched sound with an object in an upper location) (see also Evans and Treisman 2005; Pratt 1930). The device developed by Cronly-Dillon differs from the other devices in having a more limited range of frequencies, corresponding to musical notes. All three systems convert the luminosity of the object into sound amplitude.

With regard to the encoding of the horizontal position of objects, The vOICe and the device developed by Cronly-Dillon use left-to-right scanning in time in order to encode the horizontal position, whereas the PSVA uses a frequency mapping in addition to the one used for the vertical position. More specifically, the frequency associated to each pixel increases from left to right and from bottom to top. Furthermore, in order to increase the similarity with the encoding used by the visual system, the receptor field of the PSVA has a higher resolution in the center of the picture (the fovea has four times the resolution of the periphery). An additional feature of the device developed by Cronly-Dillon is that it has a system for feature extraction designed to enable users to deconstruct a complex optical image into a set of simpler representations.

Studies that have been conducted with auditory devices have shown the possibility of object localization (Renier et al 2005-a) and form recognition (Arno et al

1999, 2001 with the PSVA; Cronly-Dillon et al 1999, 2000; Pollok et al 2005 with The vOICe). Interestingly, recent studies have also demonstrated the possibility of recreating visual illusions with the PSVA (Renier et al 2005-b).

To date, however, no behavioral study has combined investigation of the objective performance with a study of the qualitative experience that participants obtain after extensive training with a visuo-auditory conversion system. The aim of the present study was therefore to investigate participants' performance while actively using the visuo-auditory substitution system The vOICe through dynamic tasks in a 3-dimensional environment. We chose this device because of its ease of implementation and ready availability. We chose two kinds of tasks involving object localization and recognition. We focused on these two tasks because they are ecologically relevant for everyday use and because they involve two different channels of visual processing: The "where" and the "what" system (Ungerleider and Mishkin 1992), also named the "pragmatic" and "semantic" dorsal and ventral cortical pathways (Jeannerod 1994). We were also interested in questioning the qualitative experience of participants who learnt to use this device. Does the subjective experience felt while using this device more resemble vision or audition? Does this subjective feeling depend on the task? Which are the strategies used by participants in order to master the device? Can they reach a form of immediate apprehension of what is perceived?

General methods

Participants

6 sighted participants (2 females and 4 males) took part in these experiments. Their ages ranged from 23 to 32 years (mean: 27 ± S.D. of 3.3 years). All of the participants reported normal auditory perception and none of them was familiar with

The vOICe system. The participants were instructed to choose freely the hand in which they held the webcam. They all chose the right hand. The participants received neither money nor course credit for their participation. The experiments took approximately 15 hours to complete and were performed in accordance with the ethical standards of the 1991 Declaration of Helsinki.

Apparatus and materials

We used the visual-to-auditory conversion system The vOICe developed by Peter Meijer (see Meijer 1992 for details). The vOICe is an experimental system for the conversion of images into sound patterns. Pictures provided by a portable webcam are converted into a 16-greyscale picture of 176*64 pixels and cyclically scanned from left to right, once per second. They are then translated into a complex sound that varies as a function of the position and brightness of the pixels in the column of the image that is under the scan at each moment. Each pixel in the column under the scan corresponds to a different frequency: The higher the position of the pixel, the higher the frequency, the lower the position of the pixel the lower the frequency. The set of frequencies can in principle be chosen arbitrarily but two well-defined benchmark sets are proposed, namely a linear and an exponential distribution. For the experiments conducted here, we chose an exponential distribution going from 500 Hz to 5000 Hz. These values are within the spectral sensitivity of the human auditory system. Brightness of pixels in the image is coded in terms of the amplitude of the sinusoid that is emitted for that pixel: Silence means black and a loud sound means white. Anything in-between is a shade of grey.

For practical reasons, in the study reported here, it appeared easier to present dark objects on a white background than to present white or luminous objects on a dark background. This design offered better control of the overall illumination of the experimental room and a better contrast between objects and background. However, as the auditory translation of a black object results in a pattern of silence that is hard to recognize for participants, we used the contrast reversing function of the conversion program (negative video). Thus, with the use of this reversing function, a straight dark line on a white background running from the bottom left to the top right sounds like a tone steadily increasing in pitch. Two horizontally displaced dark dots sound like two short beeps, and so on.

In the study reported here, we chose to use a hand-held camera. It would also have been possible to use a head-mounted camera. We hypothesized that, for a first experience with a sensory substitution device, holding the camera in the hand may allow the participants better control of the position of the camera and more specifically on the orientation of the camera. However, it would be interesting in future research to compare the use of both hand-held and head-mounted camera, for example, to address the question of whether the use of a hand-held camera allows for the more precise extraction of the sensorimotor invariants while exploring the visual scene?

Procedure

The experiments were conducted in a uniformly illuminated white room. The participants were fitted with occluding goggles and held a miniature Philips ToUcam Pro webcam in their right hand in order to explore their environment. The webcam provided a 32 degree view of the scene. It was connected to a Sony PCG-FX401 personal computer. The processing system The vOICe hosted in the PC translated the images into an auditory output. The resulting signal was provided to the participants through a pair of Sennheiser H280 pro headphones. After a training session, participants underwent two experiments involving pointing tasks and two experiments involving object recognition. After the experiments, participants were given a questionnaire in order to try and understand their qualitative apprehension of the use of the device through these different tasks.

Experiments and results

1- Training session

The participants were given a short verbal explanation about the functioning of the device. The experimenter explained the relation between the time scanning and the horizontal location, the relation between sound frequency and vertical location, and the relation between brightness and loudness. Next, the participants were trained with the device. They were taught how to localize a target and how to keep contact with the target while moving. A black target (a small cylinder of 5 cm long with a 4 cm diameter) was placed on the table in front of them. Participants were asked to keep contact with the auditory signal provided by The vOICe and to move slightly up, down, left, and right. They were then asked to increase the amplitude of their movements. Subsequently, participants were asked to stand up and to move their fingers, hand, and shoulder and to move around, while paying attention to the sounds. Then, we moved the black target and participants were asked to follow this target. The training session lasted for approximately three hours.

2- Experiment 1 - Locomotion and pointing task

In a first experiment, we investigated the participants' ability to localize and to point at a target via The vOICe. Participants, blindfolded and equipped with The vOICe, stood in a white room (see Figure 1). We placed an 11*11 cm square black target on the

70*135 cm white table in front of them. Participants stood one meter from the table. We asked them to localize the target with hand movements, to walk toward the table, and to point at the target. The target could be at any one of nine possible locations differing by 3 vertical positions (20, 40, and 60 cm from the table border) and by 3 horizontal positions (-40, 0, and 40 cm around the table center). Participants completed 18 trials, two for each location. We measured the time taken to move and to point correctly at the target (i.e. to touch the target). Participants underwent two one-hour sessions.

(FIG. 1 ABOUT HERE)

Results

Participants always succeeded in pointing at the target. An ANOVA performed on the time taken to point correctly at the target as a function of its Vertical position (close, middle, and far) and Horizontal position (left, middle, and right) showed a significant effect of the Vertical position of the target [F(2,99)=9.52, p<.001]. A Duncan post-hoc test revealed a significant difference between the vertical position far and the other two positions: close and middle (both ps<.005), but did not show any significant difference between the middle and close positions. Participants took less time to point at the target when it was close to the table border (73.9 ± S.D. of 35.3 s) and in the middle (89.6 ± 36.9 s), than when it was far from the border of the table (135.9 ± 53.2 s). The analysis did not reveal any effect of the Horizontal position of the target [F(2,99)<1, ns], nor any interaction between Vertical and Horizontal position of the target [F(4,99)<1, ns].

An ANOVA performed on the time taken to point correctly at the target as a function of the Presentation order of the target showed no effect of this factor [F(17,90)<1, ns]. However, an ANOVA performed on the time taken to point correctly at the target as function of the Presentation order for each vertical position (6 levels) showed a significant effect of this factor [F(5,102)=3.1, p<.05]. Thus, participants' performance improved significantly over time for each vertical position of the target. As the training progressed, the time taken for correct pointing decreased from about two minutes (105.6 ± 22.3 s) at the beginning to about one minute (65.4 ± 28.1 s) at the end of the experimental session.

3- Experiment 2 - Second localization task

In a second experiment, we studied participants' ability to localize and to point at a target using a constrained pointing task. Participants were seated in front of the table. They held the webcam with their right hand. Their right elbow rested stationary on the table in one position. We placed a small black 4 cm diameter ball at different positions on the table and asked participants to point at it with their left hand. The distance between their elbow and the target could vary from 40 to 80 cm (from 0 to 70 cm vertically and from 0 to 40 cm horizontally). Participants completed 26 trials. We measured the distance between the location pointed by the participants and the centre of the target. Participants underwent two one-hour sessions.

Results

An ANOVA performed on the pointing error as a function of the distance between the target and the elbow showed a significant main effect of this factor [F(21,133)= 1.7, p<0.05]. The further the target was from the elbow, the greater was the pointing error (see Figure 2). The mean error for pointing was 7.8 cm (± S.D. of 5.1 cm). An ANOVA performed on the pointing error as a function of the Presentation order did not show any significant effect of this factor [F(25,129)<1, ns]. Thus participants' performance did not improve over time.

(FIG. 2 ABOUT HERE)

4- Experiment 3 - Recognition task

In a third experiment, we studied if participants were able to recognize objects via their auditory rendering. Participants, blindfolded and fitted with The vOICe, were seated in front of a white table. During a training session, participants were presented with ten common objects: A plant, a book, a pan, a shoe, a statuette, a remote control, a spoon, a bottle, a little table, and a hand bag (see Figure 3). In order to favor the extraction of an auditory signature for each object and in order to avoid the additional difficulty of dealing with interposition of objects, each object was presented separately for 5 minutes. Participants were asked to actively explore the object with the webcam and to recognize it via its auditory rendering. The participants were encouraged to move the webcam actively and to choose different points of view upon the object in order to detect object features by making use of their self-produced movements. Active use of the camera during the recognition tasks was introduced in order to make sure that participants went through a form of perceptuo-motor learning with the device rather than doing the task by simply learning static crossmodal associations between the visual objects and their auditory conversion. At the end of the exploration of each object, the participants were asked to explore the object with their hand. They were told that the same objects would be used during the experiment.

During the evaluation sessions, each object was presented five times in a random order (i.e., participants completed 50 trials in total). The participants were asked to recognize each presented object. We measured the time taken to auditorily recognize the object correctly and the number of objects enumerated before the participant gave the correct response. Responses were collected for 2 minutes after we placed the object on the table. Participants remained blindfolded for the duration of the experiment. Participants were given feedback regarding the correctness of their responses: If they were unable to recognize the object via its auditory rendering, they were asked to explore it with their hand (but they were scored as having failed to recognize the object). Participants underwent a 1 one-hour session for the training and 3 one-hour sessions for the evaluation.

(FIG. 3 ABOUT HERE)

Results

Trials in which participants failed to make a response before the trial was terminated (less than 12% of trials overall) were not included in the data analyses. Participants took a mean of 42.4 ± S.D. of 27.9 s to recognize an object and enumerated 1.6 ± 0.9 objects before giving the correct answer. An ANOVA performed on the recognition time as a function of the Presentation order for each object (5 levels) revealed that performance improved significantly over time [$\underline{F}(4,25)=3.6$, $\underline{p}<.05$]. As the training progressed, the processing time decreased from about 57.6 ± 26.1 s at the beginning to about 34.7 ± 19.3 s at the end of the training.

We also note that the recognition time differed significantly among participants [$\underline{F}(5,54)=4.6$, $\underline{p}<.001$]. For instance, the mean recognition time of one of the participants, who was a musician, across all objects was 27.6 ± 7.6 s whereas the mean of the 5 other participants across all objects was 46.6 ± 21.4 s.

An ANOVA performed on the recognition time as a function of the Type of objects revealed that participants' performance was significantly affected by the Type of objects [$\underline{F}(9,50)=2.15$, $\underline{p}<0.05$]. For instance, the book was recognized after a mean duration of 60.9 ± 19.5 s whereas the plant was recognized after a mean duration of 19.9 ± 13.2 s (see Figure 4).

(FIG 4 ABOUT HERE)

5- Experiment 4 - Discrimination task

In a fourth experiment, we investigated if participants, once they were able to recognize different objects, could discriminate among different versions of the same object. We used the ten objects of Experiment 3 and some variants of these objects that varied in size or in form, but that belonged to the same category (e.g., a different kind of shoe). The ten objects used previously were a plant, a book, a pan, a shoe, a statuette, a remote control, a spoon, a bottle, a little table, and a hand bag. The nine new objects consisted of two shoes, two pans, a plant, a little bottle, a handbag, a folder, and a fork (see Figure 5). We used the same experimental protocol as in Experiment 3. During a training session, the participants were presented with each of the nine new objects for five minutes. The participants were asked to actively explore the object with the webcam and to recognize it via its auditory rendering. For objects that had variants, we introduced simple labels such as plant 1 and plant 2. During the evaluation sessions, participants were presented one at a time, and in a random order, with objects from the set of 19 objects (the 10 objects of experiment 3 and the 9 additional objects). Participants were asked first to recognize the category of the object (e.g., plant) and then to discriminate which object was used among this category (e.g., plant 2). Participants were given feedback regarding the correctness of their responses: If they failed to recognize the object via its auditory rendering or to discriminate among objects belonging to the same category, they were allowed to explore it with their hand (but were scored as having failed to recognize the object). Participants underwent a 1 one-hour session for the training and 3 one-hour sessions for the evaluation. They completed 50 trials in total. We measured the time taken to recognize the objects, the number of objects enumerated before the participants gave the correct response, and the correctness of the discrimination response among objects belonging to the same category.

(FIG. 5 ABOUT HERE)

Results

Trials in which participants failed to make a response before the trial was terminated (less than 13% of trials overall) were not included in the data analyses. Participants took a mean of 38.7 ± S.D. of 28.4 s to recognize an object and enumerated 1.7 ± 0.9 objects before giving the correct answer. With regard to the discrimination abilities, results showed that when there were three objects belonging to the same category, the percentage of correct responses was 56.4 ± 11.8 %, while when there were two possible objects in the category, the percentage of correct responses was 74.2 ± 9.9 % (see Figure 6). In order to determine if these results were above the chance level (50% and 33%, respectively) we compared the measured percentage to the corresponding chance levels with a t-test. This test showed that these percentages of correct answers are (for $\alpha=0.05$) above the chance level (respectively $t = 7.7$ for $n = 159$ and $t = 3.94$ for $n = 104$, the threshold value for t is 1.96). We also note that participants took a little less time when they responded correctly (38.7 ± 27.5 s) than when they responded incorrectly (42.0 ± 31.9 s). They also took less time for objects that did not have a variant (32.2 ± 24.0 s) than for objects that have variants (39.1 ± 28.5 s).

(FIG. 6 ABOUT HERE)

6- Results concerning the qualitative experience of participants

We gave the participants questionnaires in order to understand their qualitative experience while using the device (see Appendix 1). We asked them to which sensory modality their experience resembled more for the two localization tasks and for the recognition and discrimination tasks. Results showed that localization tasks were more likely apprehended as either a visual sense or a new sense. Recognition and discrimination tasks were perceived as resembling more to audition (see Table 1). Two participants mentioned a resemblance with the tactile modality. We raise here the possibility that this result could be due to the tactile contact with the object allowed when participants failed to recognize the objects auditorily and to the use of a hand-held camera. The hypothesis raised by sensorimotor theories of perception is that the use of a head-mounted camera shares more sensorimotor resemblance with vision than the use of a hand-held camera (e.g., Lenay 2002; Noë 2005; O'Regan and Noë 2001). Thus, it could be the case that the use of a hand-held camera might have diminished an association between perception via the device and the visual sensory modality as compared with the situations in which a head-mounted camera had been used instead. However, we stress that, in our experiments, the participants held the camera during the 4 experiments but associated their qualitative experience differently as a function of the task. It thus seems that there were other components, such as top down influences, involved in the phenomenological sensation associated with perception via the device.

(TABLE 1 ABOUT HERE)

7- Qualitative results concerning motor behavior and perceptuo-motor strategies of participants

Since perceptual abilities originate through coupling between actions on the webcam and auditory sensations that result from them, we carefully observed the motor behavior of participants during the experiments. With regard to their mastery of the device, the participants had no problem in understanding the functioning of the device:

They easily understood the different rules linked to the spatial location of the objects. During the training session, all of the participants attested that making contact with objects by the use of the device was almost immediate. According to them, the signal was easily interpretable, moving the webcam was intuitive, and as a consequence it was easy to manipulate the signal. Furthermore, participants had an instantaneous ability to detect the specific signal that indicates the presence of an “object” over the “scene noise”. They exhibited no problem in finding the target.

With regard to the motor behavior of participants, we observed that they seemed to have had some problems during their displacement. These problems were not very noticeable for small movements when the web cam was locked on the target. But the participants encountered greater difficulties when they initiated large movements: Participants often lost the target and had problems getting back to exactly the same place or to contact the target again. This was the most critical problem with this device. It could be due to limitations of proprioceptive information or memory involved in the task. It could be that the memory of either participants’ actions or their body posture was not sufficiently accurate. When participants lost contact with the object, they were often unable to come back to their preceding position and instead completely restarted their exploration from the starting position, which was used as a kind of reference position. At first, to overcome this, most participants realized that they could not take into account how their whole body moved. As a consequence, they tried to move only one body part and to keep other parts still. They often immobilized their elbow and wrist and moved only their fingers, or they could immobilize their fingers and shoulder and move only their wrist or their forearm. We observed that people used different strategies. Little by little, participants managed to integrate more movements of their body, but even at the end of the experiments they did not succeed in moving entirely with ease. These difficulties may explain the problems of gaining precise access to depth information, as revealed by the two pointing tasks.

The delay involved in scanning the image also had an effect upon participants' motor behavior. As the visual scene was sampled only once a second, the sensory feedback was not immediate. To adapt to this constraint participants waited till the end of each scan before initiating a movement in order to adopt another point of view with respect to the scene. Thus they often used jerky movements in order to be in synchrony with the sensory return. Participants had to practice a lot in order to move with ease. At the end of the 15 hours of experiments, they were able to move more easily.

Another interesting point relates to the intuitive or deductive nature of the different tasks. At the beginning, to keep in contact with the object, participants reported that they used a very deductive approach: They consciously sought the highest and lowest pitch provided by the device. Later they went beyond this deductive phase and attained a more intuitive apprehension of the object. They said "we reason less", "we trust the sounds". However, each time they encountered a problem with their movements they lost this intuitive apprehension.

8- Results concerning the subjective perception of recognition tasks

The comments of the participants during the experiments and the results of the questionnaires given at the end of the experiments suggested that participants used a mixture of two strategies during the recognition tasks. Either participants tried to recognize an "auditory signature of the object", a "very characteristic sound" or they tried to deduce the characteristics of the objects from an analysis of the sound pattern. They made four rough categories: Flat and elongated objects, thin and tall objects, objects with a precise form but made of two or more distinctive sounds (as the shoes or the pans), and confused objects (as the plants). They rapidly associated each object to the correct category, but they then had difficulties in refining their recognition. The use of these two strategies was associated with a large difference in the recognition time as a function of the different objects: Some objects produced very specific sounds and were easy to recognize, as in the case of the plant for instance.

There was also a great deal of difference among participants. In particular, one of them was a musician and said that she noticed an “auditory signature” for almost all objects. Correspondingly, her recognition time was extremely fast. Furthermore, she exhibited very fine recognition abilities that allowed her to discriminate between very similar objects, such as two slightly different kinds of shoes for example.

General Discussion

In the study reported here, we investigated the objective and qualitative performance of six blindfolded participants using the visual-to-auditory conversion system The vOICe. In these studies, participants were encouraged to actively move the webcam and to choose several points of view upon objects in order to localize them and to detect pattern features thanks to self-produced movements. During the first two experiments, participants underwent pointing tasks. Their performance showed that they were able to use the auditory conversion of the visual stimuli for locomotor guidance, localization of objects, and for pointing. Participants at first had difficulties in gaining precise access to depth information but their performance improved over trials (in Experiment 1). Renier et al (2005-a) reported similar improvements with the PSVA. Interestingly, these authors reported that before training early blind participants had difficulties in localizing objects in depth whereas blindfolded sighted participants seemed to apply from the beginning the rules of visual depth when they used the PSVA. However, after training, blind participants reached equivalent performance as that of sighted participants. These results show the possibility for early blind individuals to use a visual-to-auditory conversion system in order to localize objects.

Performance during the two recognition tasks showed that participants were able to recognize 3-dimensionnal objects and were able to discriminate among objects belonging to the same category. With training, perception through the device became more and more “automatic” and overall processing time decreased. Similarly, Pollok et al (2005) reported that participants’ performance in the recognition of bi-dimensional natural objects with the Voice improved with practice, even if the practice occured with 3-dimensional objects. So far, no previous behavioral studies had investigated performance in recognition of 3-dimensionnal object using a visual-to-tactile conversion system. However, studies with the PSVA showed that early blind individuals are able to recognize simple bi-dimensional visual patterns thanks to their auditory translation, with better performance than blindfolded sighted participants (Arno et al 2001). These results prefigure the possibility of blind persons to using sensory substitution devices to recognize natural objects. However, we underline that in our study the recognition task was simplified as compared to a natural environment: The objects were presented one at a time on a uniform background. It would thus be interesting in future research to study the extent to which participants can recognize objects in a complex scene and more specifically if, thanks to self produced movements, they are able to deal with interposition of objects.

In the experiment reported here, we were also interested in the qualitative experience of participants who learnt to use a visual-to-auditory conversion system. We first mention that in spite of the long duration of the experiments, all of the participants treated the different tasks like a game and reported being surprised and amused by their ability to acquire the mastery of a sensory substitution device. Similar verbal reports were found during the learning of the TVSS (e.g., Second et al 2005). The second interesting result to emerge from the qualitative experience of participants is that as their learning progressed, participants' perception with the device passed from a form of deductive reasoning to a more intuitive and immediate apprehension of what was perceived. Thus, with practice, the device became progressively more transparent and gave more direct access to the environment. The third result to emerge from this study was that the qualitative experience of participants depended on the task. Some tasks appeared to be more auditory, others more visual, and sometimes participants thought that their apprehension more resembled touch or smell. Some participants emphasized that they simply had the feeling of mastering a new tool. They used this device as a tool, and easily acquired mastery of it because, as one participant said: "We are used to extending our bodies through machines, exactly as when we learn how to drive a car or how to use a computer". However, it could be the case that the use of a hand-held camera favored this association. It would be interesting in future research to compare the use of both hand-held and head-mounted cameras on the mastery of a sensory substitution device and on the phenomenological sensation associated with this perception. Does the use of a head-mounted camera favor the association between perception via the device and the visual sensory modality? We also mention the role that mental imagery could have played in the qualitative experience of participants. Indeed, it has been found that, when tactually exploring objects, blindfolded sighted participants used their memories of the corresponding visual object and thus generated visual images (Heller 2000). A similar intervention of visual imagery in sighted participants could affect the perception reached with a sensory-substitution device. It thus would be interesting to compare the qualitative experience obtained by blindfolded sighted, early blind and late blind participants who are extensively trained with a sensory substitution device.

In summary, in the study reported here, participants were able to use The vOICe in order to obtain the information necessary for locomotor guidance, localization, and recognition of objects in a 3-dimensional environment. Thus, despite the clear difference in stimulated sensory modality, participants were able to learn perceptive abilities of a visual nature even if the sensory inputs were tactile or auditory. These results confirm the functional plasticity of our sensory systems and contribute to raise hopes for the development of non-invasive apparatus to help sensorially-impaired people.

Acknowledgement

We wish to thank the volunteers for their participation to the study. We are also grateful to Charles Spence and Peter Meijer for their helpful comments on this manuscript.

References

Arno P, Capelle C, Wanet Defalque M C, Catalan Ahumada M, Veraart C, 1999 "Auditory coding of visual patterns for the blind" *Perception* **28** 1013 - 1029

Arno P, Vanlierde A, Streel E, Wanet Defalque M C, Sanabria Bohorquez S, Veraart C, 2001 "Auditory substitution of vision: Pattern recognition by the blind" *Applied Cognitive Psychology* **15** 509 - 519

Bach-y-Rita P, 1972 *Brain mechanisms in sensory substitution* (New York: Academic Press)

Bach-y-Rita P, Collins C C, Saunders F A, White B, Scadden L, 1969 "Vision substitution by tactile image projection" *Nature* **221** 963 - 964

Bliss J C, Katcher M H, Rogers C H, Shepard R P, 1970 "Optical-to-tactile image conversion for the blind" *IEEE Transaction Man-Machine Systems* **11** 58 - 65

Capelle C, Trullemans C, Arno P, Veraart C, 1998 "A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution" *IEEE Transaction on Biomedical Engineering* **45** 1279 - 1293

Craig J C, 1976 "Vibrotactile letter recognition: The effects of a masking stimulus" *Perception & Psychophysics* **20** 317 - 326

Craig J C, 1980 "Modes of vibrotactile pattern generation" *Journal of Experimental Psychology: Human Perception and Performance* **6** 151 - 166

Craig J C, 1981 "Tactile letter recognition: Pattern duration and modes of pattern generation" *Perception & Psychophysics* **30** 540 - 546

Craig J C, 1983 "Some factors affecting tactile pattern recognition" *International Journal of Neuroscience* **19** 47 - 58

Cronly-Dillon J, Persaud K, Blore F, 2000 "Blind subjects construct conscious mental images of visual scenes encoded in musical form" *Proceedings of Royal Society of London* **267** 2231 - 2238

Cronly-Dillon J, Persaud K, Gregory R P F, 1999 "The perception of visual images encoded in musical form: a study in cross modality information transfer" *Proceedings of Royal Society of London* **266** 2427 - 2433

Epstein W, 1985 "Amodal information and transmodal perception" *Electronic Spatial Sensing for the Blind* 421-430

Evans K K, Treisman A, 2005 "Crossmodal binding of audio-visual correspondent features" (abstract) *Journal of Vision* **5** 874

Heller M A, 2000 *Touch, Representation and Blindness (Debates in Psychology)* (Oxford: Oxford University Press)

Heyes A D, 1984 "Sonic Pathfinder: A programmable guidance aid for the blind" *Electronics and Wireless World* **90** 26 - 29

Hirsh I J, 1988 "Auditory perception and speech", in *Handbook of Experimental Psychology* Eds R C Atkinson, R J Hernstein, G Lindzey and R D Luce (New York: John Wiley) pp 377 - 408

Hughes B, 2001 "Active artificial echolocation and the nonvisual perception of aperture passability" *Human Movement Science* **20** 371 - 400

Jansson G, 1983 "Tactile guidance of movement" *International Journal of Neuroscience* **19** 37 - 46

Jeannerod M, 1994 "The representing brain: Neural correlates of motor intention and imagery" *Behavioral and Brain Sciences* **17** 187 - 245

Kaczmarek K A, Haase S J, 2003 "Pattern identification and perceived stimulus quality as a function of stimulation current on a fingertip-scanned electrotactile display" *IEEE Transaction on Neural System Rehabilitation Engineering* **11** 9 - 16

Kay L, 1964 "An ultrasonic sensing probe as a mobility aid for the Blind" *Ultrasonics* **2** 53

Kay L, 1980 "Air sonars with acoustical display of spatial information", in *Animal Sonar Systems* Eds R G Busnel and J F Fish (New York: Plenum Press) pp 769 - 816

Kay L, 1985 "Sensory aids to spatial perception for blind persons: Their design and evaluation", in *Electronic spatial sensing for the blind* Eds D Warren and E Strelow (Dordrecht: Martinus Nijhoff) pp 125 - 139

Lemaire L, 1999 *Approche comportementale de la question de Molyneux (A behavioral approach to Molyneux' question)*, PhD thesis, Université Louis Pasteur, Strasbourg, France

Lenay C, 2002 *Ignorance et Suppléance : la question de l'espace (Ignorance and augmentation: The question of space)* Unpublished Thesis, U.T.C, Compiègne, France

Lenay C, Gapenne O, Hanneton S, Marque C, Genouëlle C, 2003 "Sensory substitution: Limits and perspectives", in *Touching for Knowing* Eds Y Hatwell, A Streri and E Gentaz (Amsterdam: John Benjamins) pp 275 - 292

Loomis J M, 1974 "Tactile letter recognition under different modes of stimulus presentation" *Perception & Psychophysics* **16** 401 - 408

Loomis J M, 1980 "Interaction of display mode and character size in vibrotactile letter recognition" *Bulletin of the Psychonomic Society* **16** 385 - 387

Loomis J M, 1981 "Tactile pattern perception" *Perception* **10** 5 - 27

Meijer P B L, 1992 "An experimental system for auditory image representations" *IEEE Transactions on Biomedical Engineering* **39** 112 - 121

Melara R D, O'Brien T P, 1987 "Interaction between synesthetically corresponding dimensions" *Journal of Experimental Psychology: General* **116** 323 - 336

Noë A, 2005 *Action in perception* (Cambridge, MA: MIT Press)

O'Regan J K, Noë A, 2001 "A sensorimotor account of vision and visual consciousness" *Behavioral and Brain Sciences* **24** 939 - 973

Pollok B, Schnitzler I, Mierdorf T, Stoerig P, Schnitzler A, 2005 "Image-to-sound conversion: Experience-induced plasticity in auditory cortex of blindfolded adults" *Experimental Brain Research* **167** 287 - 291

Pratt C C, 1930 "The spatial character of high and low tones" *Journal of Experimental Psychology* **13** 278 - 285

Renier L, Collignon O, Poirier C, Tranduy D, Vanlierde A, Bol A, Veraart C, De Volder A G, 2005-a "Cross-modal activation of visual cortex during depth perception using auditory substitution of vision" *NeuroImage* **26** 573 - 580

Renier L, Laloyaux C, Collignon O, Tranduy D, Vanlierde A, Bruyer R, De Volder A G, 2005-b "The Ponzo illusion using auditory substitution of vision in sighted and early blind subjects" *Perception* **34** 857 - 867

Sampaio E, Maris S, Bach-y-Rita P, 2001 "Brain plasticity: 'Visual' acuity of blind persons via the tongue" *Brain Research* **908** 204-207.

Segond H, Weiss D, 2005 "Human spatial navigation via a visuo-tactile sensory substitution system" *Perception* **34** 1231 - 1249

Ungerleider L G, Mishkin M, 1992 "Two cortical visual systems" in *Analysis of visual behavior* Eds D J Ingle, M A Goodale and R J W Mansfield (Cambridge, Mass.: MIT Press)

Table 1

	Localization tasks	**Recognition and discrimination tasks**
S 1	Between visual and another sense (smell)	Tactile
S 2	Visual	Between visual and tactile
S 3	Another sense (sonar)	Auditory
S 4	Another sense (sonar)	Auditory
S 5	Visual	Auditory
S 6	Visual	Visual

Results extracted from questionnaires concerning the qualitative experience of participants.

Figure Captions

Figure 1. Participant performing the pointing task in Experiment 1.

Figure 2. Error surface (cm) as a function of the horizontal and vertical distance (cm) to the participants' elbow. The surface was obtained by a spline type model.

Figure 3. Examples of objects used in the recognition task.

Figure 4. Mean time over participants taken to recognize the different objects. The error bars represent the standard deviation to the mean.

Figure 5. Examples of objects used in the generalization task.

Figure 6. Mean over participants of the percentage of correct response as a function of objects. Numbers in brackets represents the number of possible answers for this object. The error bars represent the standard deviation to the mean.

FIGURE 1

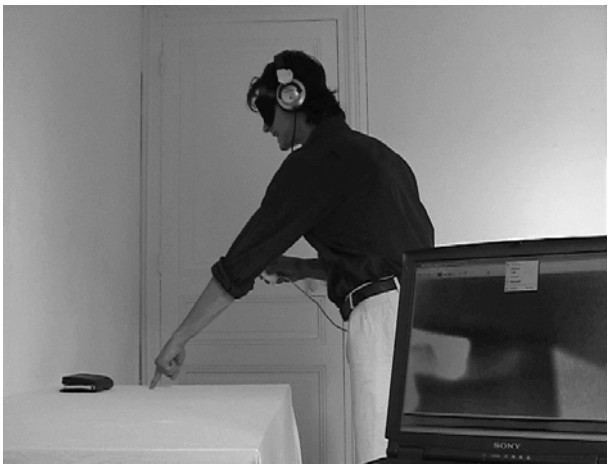


FIGURE 2

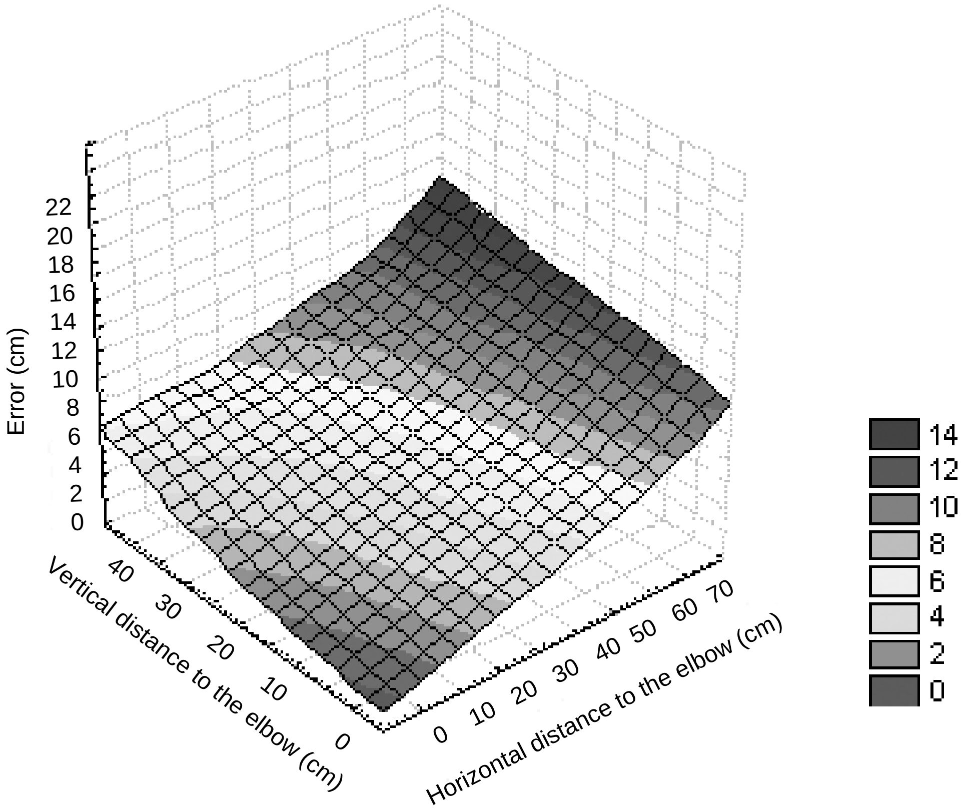
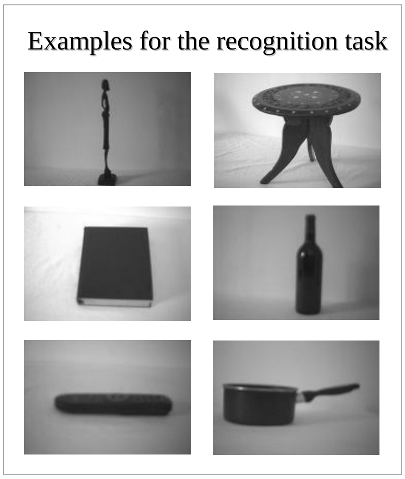

Error (cm)
22
20
18
16
14
12
10
8
6
4
2
0
Vertical distance to the elbow (cm)
40
30
20
10
0
Horizontal distance to the elbow (cm)
0
10
20
30
40
50
60
70
14
12
10
8
6
4
2
0

FIGURE 3



Examples for the recognition task

FIGURE 4

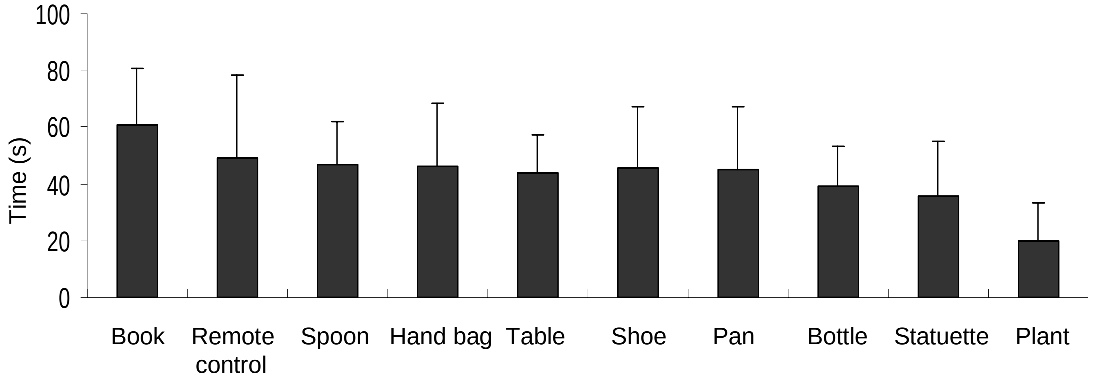
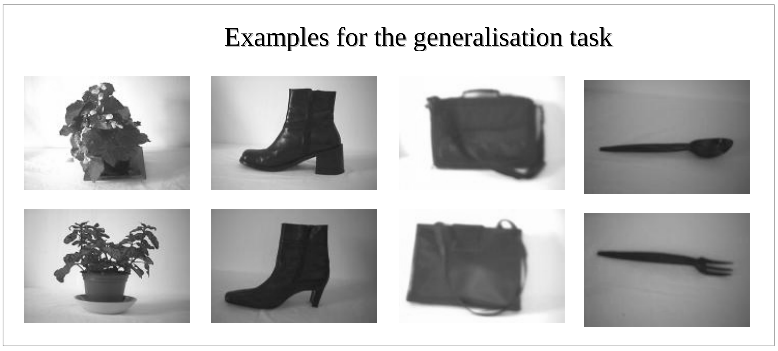
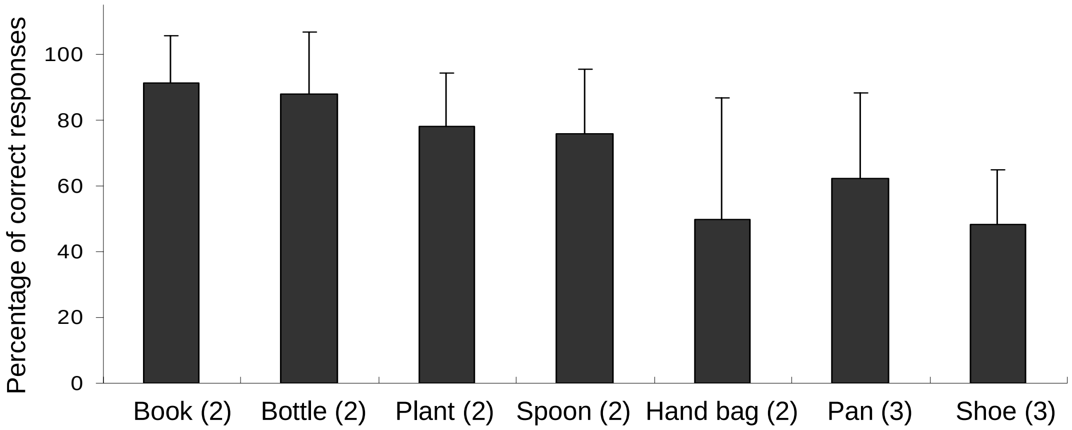

100
80
60
40
20
0
Time (s)
Book
Remote control
Spoon
Hand bag
Table
Shoe
Pan
Bottle
Statuette
Plant

FIGURE 5



Examples for the generalisation task

FIGURE 6


Percentage of correct responses
0
20
40
60
80
100
Book (2)
Bottle (2)
Plant (2)
Spoon (2)
Hand bag (2)
Pan (3)
Shoe (3)

Appendix 1: Questionnaires used for the experiments

I- Learning session

1- At first, did you find it easy to use the device?

2- How much time did it take you to acquire the basics needed to use the device?

3- Did you have problems in making contact with the target and in centering it?

4- Did you have problems in moving in front of the target

- With small movements
- With large movements
- Did you have the impression that your movements were jerky?
- Was it easier to move horizontally or vertically? Why?

5- Did you adopt a particular strategy with your movements?

6- How did you find the object when you lost it?

7- Did you find this task difficult?

8- Would you describe this task as being intuitive, or would you rather say deductive?

9- Were you conscious of the device, the equipment, or did you come to ignore it?

10- To which sensory modality would you compare your experience?

- Can you propose analogies?

II- Experiment 1 & 2

11- Did you have problems in moving in front of the target

- With small movements
- With large movements
- Did you have the impression that your movements were jerky?
- Was it easier to move horizontally or vertically? Why?

12- Did you adopt a particular strategy with your movements?

13- How did you find the object when you lost it?

14- Did you find this task difficult?

15- Would you describe this task as being intuitive, or would you rather say deductive?

16- Were you conscious of the device, the equipment, or did you come to ignore it?

17- To which sensory modality would you compare your experience?

- Can you propose analogies?

II- Experiment 3 & 4

18- Did you have problems in moving in front of the target

- With small movements
- With large movements
- Did you have the impression that your movements were jerky?
- Was it easier to move horizontally or vertically? Why?

19- Did you adopt a particular strategy with your movements?

20- How did you find the object when you lost it?

21- Please could you describe the strategies you used to recognize objects?

22- When we introduced a new series of objects, did this perturb you?

- Did it interfere with recognition of the objects from the first phase of the experiment?
- Did you change your strategy?

23- Is the comparison with similar objects difficult?

- How did you manage to do this?

24- Did you find this task difficult?

25- Would you describe this task as being intuitive, or would you rather say deductive?

26- Were you conscious of the device, the equipment, or did you come to ignore it?

27- To which sensory modality would you compare your experience?

- Can you propose analogies?