DEPTH - MELODY SUBSTITUTION

Vincent Fristot[1], *Jérémy Boucheteil*[1], *Lionel Granjon*[1], *Denis Pellerin*[1], *David Alleysson*[2]

[1]Gipsa-Lab, Grenoble Université
11, rue des mathématiques, Campus Universitaire, Grenoble, France
phone: + (33) 4 76 57 43 50, fax: + (33) 4 76 57 47 90, email: firstname.lastname@gipsa-lab.grenoble-inp.fr
web: http://www.gipsa-lab.inpg.fr
[2]Laboratoire de Psychologie et NeuroCognition, Grenoble Université
1251, Av. centrale, Campus Universitaire, Grenoble, France
phone: + (33)4 76 82 56 74, fax: + (33) 4 76 82 78 34, email: firstname.lastname@upmf-grenoble.fr
web: http://webu2.upmf-grenoble.fr/LPNC/

ABSTRACT

We present a system for visuo-auditory substitution that takes as input a depth map, and produces as output a corresponding melody. A real-time apparatus based on the Kinect®, a computer and a headphone, is developed within our team. Earlier experiment with the system shows that, when equipped, a sighted person temporary deprived from vision can therefore move in an unknown indoor environment without any training.

1. INTRODUCTION

Sensory substitution can offer a therapeutic for improving daytime for blind people (Bach-y Rita [4]) as well as an experimental research paradigm to study sensorimotor models [13] or neural plasticity [2, 15]. Among all the propositions for sensory substitution, replacing vision by audition is by far the simplest and most efficient system. Nevertheless, visuo-auditory systems need two main steps to achieve vision replacement by audition. They first encode visual information coming from a video camera, and then using a sound generator, they produce sounds that can be interpreted by a person as an image.

Many prototypes have already been developed in the field of visuo-auditory substitution. The earlier devices like The vOICe, Meijer 1992 [14], PSVA [7] converted only a single column of the input image at a time; a sequential play of the columns. Cronly-Dillon [8] introduced musical notes where frequencies of a sound increased from bottom to top for each column. Gonzales [11] proposed Head-Related Transfer Functions (HRTF) technique to allow spatial discrimination of sounds. More recently, Bologna *et al.* developed a new The Vibe system [12], called See ColOr [5], where different colors of the scene are coded with different musical instruments, whereas the sound duration corresponds to the depth of the sound source.

In this study, we developed original methods to encode the visual environment and generate sounds. We used a depth map provided by the Kinect® camera as an input to the visuo-auditory system. We re-sampled this map by several receptive fields imitating the neurons in human retina [10]. This parallel encoding of neuron's receptive field ensures compression of the visual information. Finally, the response of neurons associated to each receptive field were transformed into sound. For user comfort and discrimination, we associated each neuron with a sound frequency in the musical scale. This generated a melody from the image.

The proposed system provides valuable information transmission from camera image to sound since participants require no training period to use it. The paper first presents a description of the apparatus. Then, we discuss some properties of the system within the context of sensorimotor theories and brain plasticity.

2. DESCRIPTION OF THE SYSTEM

Our sensory substitution system is designed for real-time navigation. The device converts the depth's video stream into an auditory stream delivered to the user through headphones.

The system shown in Figure 1 is composed of a Kinect®video camera, which outputs a depth map that is used as input to the retinal encoder. The encoder delivers a compressed signal to the sound generator, which produces a stereo melody.

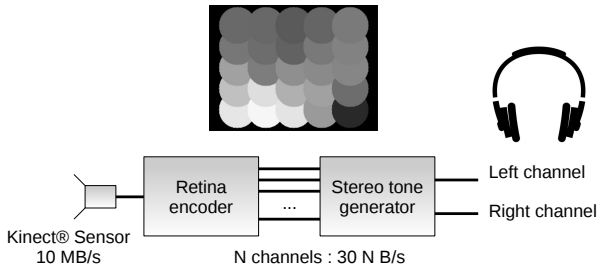


Figure 1: Diagram of the system

As stated in detail in the following sections, each step of encoding is designed to improve user comfort, discriminability and time lag.

2.1 The Kinect® sensor

The system uses Kinect sensor (Figure 2) to provide a depth image of the scene. The sensor is designed for Microsoft Xbox360 video-game console. It creates two types of images: one in usual RGB format, and second a depth map encoded as grayscale. The depth video-stream outputs in VGA resolution (640 by 480 pixels), coded on 11-bits at a frame rate of 30 Hz. Kinect system projects an infrared pattern on the scene, and measures the deformation of the pattern to guess its depth. For this reason, the system did not work properly in outdoor conditions.

Figure 2: Kinect®sensor

To implement the image-processing program in Window 7 environment, we used libfreenect library (*openkinect.org* project) as a low-level API to read Kinect's depth image.

2.2 Image simplification: Retinal encoder

As the entire visual information of the image (all the pixels) cannot be converted simultaneously into sounds without making a cacophony, compression of visual information is needed. Hanneton [12] and Durette [10] introduced receptive fields to under-sample the input images using a bio-inspired approach, allowing sounds to be generated in parallel in the so-called The Vibe system, in contrast to, the former The vOICe system [14].

In Figure 3 an example of this approach is shown for a gray-scaled intensity map. Extracted pixels (white crosses) are grouped in receptive fields (RFs) and associated to a neuron, as shown by the dotted line circle. Previously, Durette's work use 50 to 100 RFs per image, whereas we used 64 RFs uniformly distributed over the image field.

Figure 3: Example of retinal encoding applied to a gray-scaled image according to B. Durette. Crosses are pixels p used for the calculation. Filled circles are centers of the receptive fields.

The way receptive fields are placed is to keep only lower luminance frequencies. In his thesis, Durette [9] refined the Auvray's model for the calculation of RF activation. Let d denotes the mean distance between receptive fields, and the spectrum of the sampled image has to be cut off to f_c according to the Nyquist Shanon criterion.

$$f_c = \frac{1}{2d}$$

A Gaussian low-pass filter, centered at the RFs, could be used to weight the contribution of pixels contained within. To reduce the heavy computation needs for the large Gaussian convolution mask, we propose to replace this filtering by averaging fewer samples. The samples are selected according to a 2D normal distribution (with same standard deviation for each RF) at the centers of the RFs. The sample positions are randomly chosen, different for each RF, and they are computed once initially. This non-regular sampling of the pixels reduces the aliasing effect. Durette [9] shows that a set of $p = 10$ samples, and a standard deviation $\sigma = 20$ for each RF is sufficient for good estimation of the mean. Our implementation is based on these parameters.

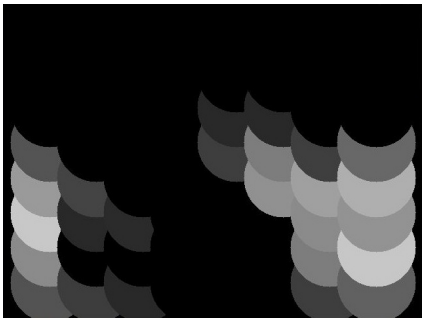
The activity Act_i of the neuron i (a set of p pixels with luminance $l_{i,k}$ for the receptive field RF_i), is normalized into the interval [0,1] using the following function:

$$Act_i = \frac{1}{255p} \sum_{k=1}^{p} l_{i,k}$$

Figure 4 summarizes an example of retinal encoding with 64 RFs, as it is implemented in our system.

(a) Input gray image

(b) The depth map



(c) Intensity of the RFs

Figure 4: Retinal encoding with 64 RFs

In summary, the retinal encoder extracts 64 activities of the RFs placed on a regular grid in the image field. Each receptive field response is the average of the activity of 10 pixels chosen randomly in its neighborhood but fixed in a configuration file. The depth map is thresholded in four levels corresponding approximately to a distance between 50 cm to 2 m, in order to increase the contribution of nearest objects related to the background objects. Restricting the number of levels in the depth map also removes the sound generation, when there are no objects nearby, thus improves the comfort for the participant.

2.3 From visual to auditory: tone generator

Each neuron's activity is transformed into a particular sound source in real-time. Sine wave frequencies are associated to the neuron according to its vertical position (lower frequencies from the bottom of the image to higher frequencies at the top). The horizontal position defines left, and right gain (for amplification or attenuation) applied to the sound source. This intensity modulation allows for padding coding of the position of the RF. The sounds for all neurons are played in parallel. A stereo tone generator generates the auditory scene by adding the contributions of all RFs as a synthesizer with linear contributions from oscillators.

From the activity Act_i of the neuron with its receptive field RF_i, we defined two attenuation parameters of the desired sound, evolving exponentially with intensity to match the sensation of the human audition system:

- A sound source intensity A_i, which rises with the activity Act_i, in the range from -60 dB to 0 dB:

$$A_i = 10^{-3(1-Act_i)}$$

- The Interaural Level Difference (ILD) which depends on the horizontal position a_i of the source (a_i = -1 at the left and 1 at the right of image), $P_{i,l}$ and $P_{i,r}$, for left and right channel gain within range from -12 dB to +12 dB:

$$\begin{cases} P_{i,l} = 10^{\frac{-3a_i}{10}} \\ P_{i,r} = 10^{\frac{3a_i}{10}} \end{cases}$$

A comfortable use of the system is observed when frequencies coupled to neuron are well-tuned, similarly as are notes in the musical scale (Figure 5). Finally, using the pulsation $\omega_i = 2\pi f_i$ associated to the RF_i receptive field, we compute the contribution from all RFs, $S_l(t)$ and $S_r(t)$, for the output left and right channels:

$$\begin{cases} S_l(t) = \sum_{i=1}^{N\_RF} A_i P_{i,l} \sin(\omega_i t + \phi_i) \\ S_r(t) = \sum_{i=1}^{N\_RF} A_i P_{i,r} \sin(\omega_i t + \phi_i) \end{cases}$$

N_RF is the total number of RFs, ϕ_i are randomly chosen into the interval [0,2π] for allowing a good discrimination of the sound source.

In summary, the left and right sounds computed by the tone generator constitute activities of the receptive fields located in the image. We also tuned the melody or chords by addition of sinusoids for salient objects.

We also pay attention about good reactivity of the device for sound generation. Then Asio driver (*www.asio4all.com*) for Windows Driver Model sound cards allows asynchronous operation (a callback function is called to fill the sound buffer). We measured 2 ms latency in our configuration that depends on the optimization of the sound buffer size.

note	frequency (Hz)
C_4	264,0
D	297,0
E	330,0
F	352,0
G	396,0
A	440,0
B	495,0
C_5	528,0

Figure 5: Frequencies of the 'just intonation' scale

2.4 System performances

In a sensorimotor approach, compute time and latency are key features for improving a system. The filling of sound buffers adds a latency time lat, with $length_{buffer}$ the number of samples of the sound buffer and f_{sampl} the sampling frequency:

$$lat = \frac{length_{buffer}}{f_{sampl}}$$

We have to limit the buffer size to reduce this latency, but more frequent calls to the sound callback function increase overhead, and compromise computation efficiency.

We assessed the latency of whole processing using a webcam for input image (30 fps). A blinking LED was shot by the webcam, the LED also triggered an oscilloscope to process the latency, which is the difference of time from the LED to the arrival of the beep. With a mid-range laptop, we measured this to be about 100 ms, which is consistent with the goal of a real-time system.

3. PRE-TESTS

We equipped several sighted people with the system, and covered their eyes with a mask. Some of them were unfamiliar to the building. All of them started walking after a few explorations with head movements, saw the door aperture, and went through it to move into the corridor. Several of them roamed along the corridor for more than thirty minutes.

Malika Auvray [3] defined several steps necessary for a subject to appropriate himself with a prosthesis, and explore the possibilities offered by the apparatus. The steps are: contact, distal attribution, control and localization, and creation of a proper distal experiment. Amazingly, the proposed system gives very fast impregnation, with all participants confident with the system, and less hesitant to walk slowly despite being deprived of their vision. The fact of no training period needed means that depth is a pertinent cue in the visual world when transformed into sound. However, to be more accurate on the capabilities of the system for mobility, we need to perform more careful and well-designed experiments.

Figure 6: A blindfolded participant walking in an office corridor.

4. DISCUSSION

The main difficulty in transforming one sense into another is that both behave differently. Visual information is sampled with two two-dimensional array of sensors (the two eyes) which constantly move across scenes. Audition is a mono-dimensional temporal signal that is sampled for two positions using a finely tuned multi-frequency sensor. The common properties shared by these two senses that can be exploited to substitute one sense with another, still remains an open question.

Sensorimotor theories think of a human body as an automaton, which processes sensory input, and produces an output command in consequence [16]. In this hypothesis, brain is the control, working with sensorimotor contingency. For some, all the cognitive construction would be built upon this rule. Therefore, sensory substitution can work very much alike the natural condition, if it provides the brain with same sensorimotor contingency. A number of situations might appear where the brain constructs contingencies by learning.

Another approach studies cerebral plasticity to explain the substitution of a sense by another. It has been shown for long that people with an illness have a different brain than ones healthy. Blind people extend their auditory region in the cortex to zones used by vision [1, 6]. Based on this hypothesis, sensory substitution works because the brain adapts to the substituted sense using brain region dedicated to the substituting sense.

Actually, these two theories rely upon the fact that if the true sensorimotor contingency is used by the system, the brain will better adapt. However, both imply a training period for the substituting sense to take over the substituted one. Training used by visuo-auditory substitution systems can be hurtful to blind people. It needs effort, and may not remain fixed after all.

Depth map image is simpler than an intensity image because it categorizes all object reflectances into the same range of depth. For locomotion, depth information could be superior to other visual information cues. Often estimated from a binocular, or other depth from focus rules. Here, we show that it can be estimated more easily, somewhat simpler, but it can be accurately transmitted to audition.

5. CONCLUSION

We present a system for visuo-auditory substitution using correspondence between depth and sound. A compression of visual depth information is realized at several points in the image. All these points are transformed into a note where the height is associated with the vertical position of the point, and gain is associated with left-right panning. The points responding accordingly to the activity of their receptive fields are played in parallel. The realized system then allows a sighted person to move in an unknown environment without any training when deprived of his vision.

The factors for the system effectively replacing vision by audition in locomotion are still unclear. It could be because of the encoding of depth by the compression system, the melodious sound produced, or the conjunction of both. Moreover, there is a minimum delay between image and sound, which certainly favors the sensorimotor contingency.

There are no real experiments showing that visuo-auditive substitution is of actual benefit for the blind. The main drawback of these systems is that natural hearing is extensively used by the blind, and such system could hinder their ability to locate an object using their sounds. We conclude that the system could mainly be used as a novel paradigm for sensorimotor experiments. In particular, we would like to know whether the system's use improves performance at this baseline, or if there are other visual auditory modalities that accord as well as depth and melody.

5.1 Acknowledgments

We thank Barthélémy Durette for his thesis work leveraged in this paper. This research was supported by the "Pôle Grenoble Cognition" and the "Structure fédérative de recherche (SFR) santé société". We also thank Anis Rahman for proofreading.

REFERENCES

[1] A. Amedi, L. Merabet, F. Bermpohl, and A. Pascual-Leone. The occipital cortex in the blind. *Current Directions in Psychological Science*, 14(6):306, 2005.

[2] A. Amedi, W. Stern, J. Camprodon, F. Bermpohl, L. Merabet, S. Rotman, C. Hemond, P. Meijer, and A. Pascual-Leone. Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex. *Nature neuroscience*, 10(6):687–689, 2007.

[3] M. Auvray, S. Hanneton, and J. K. O'Regan. Learning to perceive with a visuo-auditory substitution system: localisation and object recognition with 'the vOICe'. *Perception*, 36(3):416–30, 2007.

[4] P. Bach-y Rita, C. Collins, F. Saunders, B. White, and L. Scadden. Vision substitution by tactile image projection. 1969.

[5] G. Bologna, B. Deville, T. Pun, and M. Vinckenbosch. Transforming 3d coloured pixels into musical instru-

ment notes for vision substitution applications. *J. Image Video Process.*, 2007:8–8, August 2007.

[6] H. Burton. Visual cortex activity in early and late blind people. *The Journal of neuroscience*, 23(10):4005, 2003.

[7] C. Capelle, C. Trullemans, P. Arno, and C. Veraart. A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution. *IEEE Transactions on Biomedical Engineering*, 45(10):1279 –1293, oct. 1998.

[8] J. Cronly-Dillon, K. Persaud, and R. P. Gregory. The perception of visual images encoded in musical form: a study in cross-modality information transfer. *Proceedings: Biological Sciences*, 166:2427–2433, Dec 1999.

[9] B. Durette. *Traitement du signal pour les prothèses visuelles: approche biométrique et sensori-motrice*. Phd thesis, Université Joseph-Fourier - Grenoble I, July 2009.

[10] B. Durette, N. Louveton, D. Alleysson, and J. Hérault. Visuo-auditory sensory substitution for mobility assistance: testing TheVIBE. In *Workshop on Computer Vision Applications for the Visually Impaired*, Marseille, France, 2008. James Coughlan and Roberto Manduchi.

[11] J. González-Mora, A. Rodríguez-Hernández, L. Rodríguez-Ramos, L. Díaz-Saco, and N. Sosa. Development of a new space perception system for blind people, based on the creation of a virtual acoustic space. In J. Mira and J. Sanchez-Andres, editors, *Engineering Applications of Bio-Inspired Artificial Neural Networks*, volume 1607 of *Lecture Notes in Computer Science*, pages 321–330. Springer Berlin / Heidelberg, 1999. 10.1007/BFb0100499.

[12] S. Hanneton, M. Auvray, and B. Durette. The Vibe: a versatile vision-to-audition sensory substitution device. *Applied Bionics and Biomechanics*, 7(4):269–276, Sept. 2010.

[13] K. Kaczmarek. The tongue display unit (tdu) for electrotactile spatiotemporal pattern presentation. *Scientia Iranica*, 2011.

[14] P. Meijer. An experimental system for auditory image representations. *IEEE Trans. Bio. Eng.*, 39(2):112–121, 1992.

[15] L. Merabet, L. Battelli, S. Obretenova, S. Maguire, P. Meijer, and A. Pascual-Leone. Functional recruitment of visual cortex for sound encoded object identification in the blind. *Neuroreport*, 20(2):132, 2009.

[16] J. O'Regan and A. Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24(5):939–972, 2001.