Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Noise
Session 1pNSa: Advanced Hearing Protection and Methods of Measurement II

1pNSa6. Sensorial substitution system from vision to audition using transparent digital earplugs

Damien Lescal*, Jean Rouat and Jérémie Voix

***Corresponding author's address: GEGI, Universite de Sherbrooke, Sherbrooke, J1K2R1, QC, Canada, damien.lescal@usherbrooke.ca**

Since the TVSS (Tactile Vision Substitution System) developed by Bach-Y-Rita in 1960's, several sensorial substitution systems have been developed. In general, the so-called "sensorial substitution" system transform stimuli characteristic of one sensory modality (for example, vision) into stimuli of another sensory modality (for example, audition). These systems are developed to help handicapped persons. We developed a sensorial substitution system from vision to audition. An artificial neural network is used to identify the important parts in the image. The Virtual Acoustic Space technic is used to generate localizable sounds. A sound is associated to each important parts of the image. The entire real-time system has been implemented on iOS platforms (iPhone/iPad/iPod Touch{trade mark, serif}). We associated our system with transparent digital earplugs. This way the user is aware of the audio scene happening around him. The system has been tested on non-blind persons and the results are presented.

Published by the Acoustical Society of America through the American Institute of Physics

©2013 Acoustical Society of America [DOI: 10.1121/1.4799670]
Received 23 Jan 2013; published 2 Jun 2013

INTRODUCTION

Sensorial Substitution Systems from Vision to Audition

Substitution systems from vision to audition have a strong potential for assistance to blind people or to a person with low vision, for reeducation in which brain pathways between vision and audition need to be reenforced, for artistic creation where images or patterns of objects in images are mapped to patterns of sounds. Finally, games and education is another application area.

Severals substitution systems from vision to audition have been proposed [1, 2, 3]. So far, the most important systems developed are the vOICe [4], PSVA (Prosthesis for Substitution of Vision by Audition) [5], the device developed by Cronly-Dillon [6] and the Vibe [3]. Other applications [7] are developed to create an auditory representation of the virtual environment, rendering the virtual world entirely through Hearing using a Virtual Acoustical Space representation (VAS) [8].

Although the presented systems are of interest, they all map visual inputs into auditory outputs with no analysis of the image structure. There is a need for a more intelligent mapping in which the sound generation depends on the image structures (for example: disposition of the objects).

From Visual Scenes to Audible Scenes

It is desired to encode has much as possible of the structure of the visual scene into another audible structure which we call auditory scene. None of previously cited substitution systems do so. We therefore propose to map a visual scene into a 3D audible scene after analysis of the image content. An implementation running on handheld devices like iPods/iPads/iPhones™ is also presented. Sounds are synthesized in such way that location of sound sources is perceived in a virtual acoustical space by exploitation of the Intensity Level Differences (ILD), the Interaural Time differences (ITD) and the spectral characteristics of Head Related Transfer functions (HRTF). Figure 1 is an illustration of the proposed system with 3 objects inside the image.

Implemented System

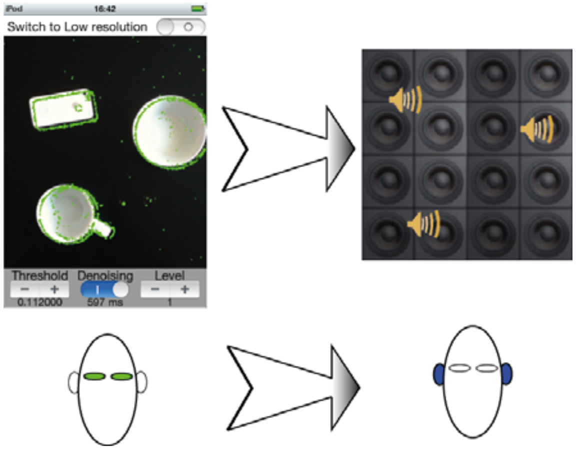

FIGURE 1: A vision to audition substitution system implemented on an iPod™. Regions of interest are automatically selected by a neural network yielding the green areas on the iPod's screen. Then, based on the position of the regions of interest, sounds are created by using the appropriate ILD, ITD and spectral cues obtained from an Head Related Transfer Function (HRTF).

Images are analyzed with a one layer neural network that performs contrast and texture enhancement analysis [9, 10]. The neural network has smaller outputs and activities in contrasted and highly textured areas of the image. This activity is used to locate regions of interest inside images. The position of these regions is then encoded into azimuth and elevation of specific sounds to create a Virtual Acoustical Space (VAS) representation. The choice of the sound characteristics is left to the user of the system. Also, for versatility, the sensitivity of the image analysis is modifiable and any headsets can be used.

THE VISUAL SCENE ANALYSIS

Real-Time Visual Scene Analysis Constraints

Visual Scene Analysis is far from being trivial. Current state of the art methods use complex image and video analysis (see for examples [11]). An accurate visual scene analysis system requires a bidirectional architecture in which bottom-up (analysis) and top-down (attention process, memory) paths are required. Although such real-time systems can be implemented on a conventional computer [12], they still are too complex to run in real time on handheld devices. For this reason a feedforward bottom-up implementation is presented (for the image analysis module) and no explicit object recognition is done in this work. Regions of interests are only detected, located and characterized. Still, the implementation has been designed so that prior knowledge of the environment could be used to modulate the neural activity in regions of interest of the images depending on prior knowledge. This has not been implemented in the reported experiments and results. Even so, results show a strong potential of the approach.

Search of Regions of Interest

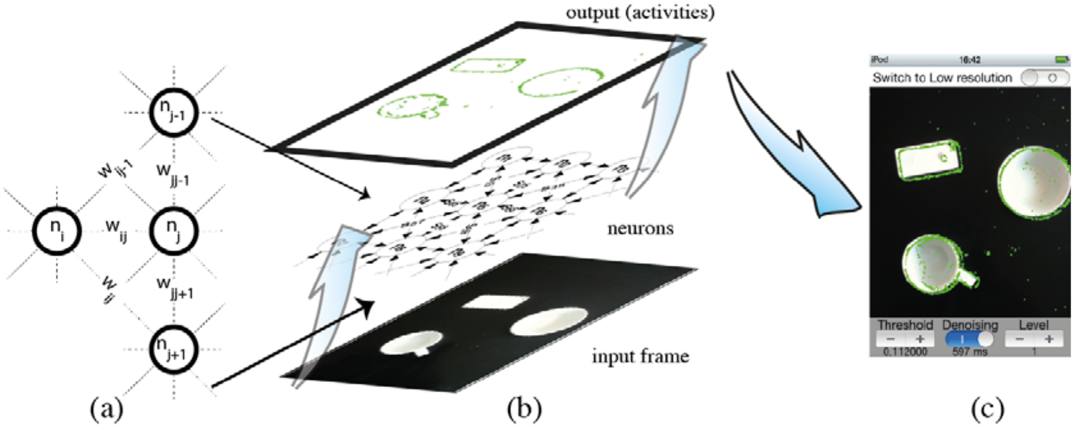

FIGURE 2: Video frames are presented as input to the neural network (b). The neural network (a) is configured in such way that contours and contrasts emerge first in the neuron's output. The activities of the neural network are illustrated as the last layer (b). The same output is also plotted and superimposed on the screen of the iPod (c). The small shift between the input image and the output result is due to the movement of the iPod. In fact, the system does not need to be static to analyze snapshots of the frontal camera of the iPod.

Neural Network Description

Each pixel of the input image is associated to a neuron in the network and synapses connect each neuron to its 8 neighboring neurons. The model has been inspired from our knowledge in neurophysio- logy. In some way, it is equivalent to a recurrent spiking neural map for feature extraction – but much faster. The neurophysiological spiking neural network would comprise *fast local excitatory* and *slower global inhibitory* synapses. Neurons firing first (those who reach first a threshold $THRES$) would contribute immediately to each neuron it is connected. Contributions are important between two neurons that are strongly connected. Inhibition would then take place and reduce transmenbrane electrical potential of all neurons – except for those that are in synchrony with the firing neuron. For a fast implementation and quick execution time, the spiking neural network has not been implemented, but approximated with neurons which outputs are continuous time variables [10] that are thresholded after each iteration. Thresholding of the neural network has been preserved in the implemented model through a shared identical threshold. A neuron has a state variable s_i and a pixel value p_i. A synapse is characterized by a weight w_{ij} which is function of the difference in the pixel value of the two neurons n_i and n_j it connects. The algorithm iteratively updates the state of the neurons depending on their current state, the weight of their synapses and the state of their neighbors [9]. The goal of this algorithm is to differentiate homogeneous regions of the image from the ones with dense contours or high textures. In some way, the algorithm performs a spatial integration of local gradients with no need of explicit edge detections. In principle, this approach is more robust to contrast changes and in practice it is much faster than the true spiking neural network. First iteration enhances contours of objects and highly textured regions. Subsequent iterations propagate and attract contours towards homogeneous areas of the image.

THE AUDIBLE SCENE SYNTHESIS

Versatility of Use

Various configurations of the system are possible. Depending on the application, conventional or specialized headsets that allow for external sound hearing can be used. A hardware interface has been designed for the use of Sonomax headsets to be used by blind or low vision people. For more conventional applications, standard headsets can be used. A prototype to illustrate the use for game and musical training has been realized and will be available for evaluation on the iTunes™ store. Twelve different notes are played depending on positions of regions of interest that are in the image.

Virtual Auditory Scene Generation

A *Virtual Acoustic Space* [8] is created by convolving the generated sounds with an Head Related Transfert Function (HRTF) filter before playing through headsets. For a given sound source, the 3D illusion is obtained by convolving the left and right channels with the HRTF filters corresponding to the azimuth and elevation of the virtual sound source. The listener has then the feeling that the perceived sound has been generated externally to the head from a specific direction and elevation. The HRTF filters are chosen amongst the set of possible filters that are defined according to the ILD, the ITD and the spectral weighting cues for the direction and elevation associated to the position in the image of the region of interest that is associated, by the system, to the virtual sound source.

Three prototypes have been realized:

1. A first evaluation research platform used the HRTFs measured by the CIPIC Interface Laboratory at UCD [13, 14].
2. A second platform for educational applications has been created with a new set of HRTFs. Measures were done in a 95 m^3 semi-anechoic room that has been transformed into an anechoic room. The G.R.A.S. 45CB head [15] has been used.

3. A third platform has been realized for applications that require the user to still perceive external sounds while using the vision to audition substitution system (for example, blind or low vision users). An interface has been designed and built for the use of Sonomax Auditory Research Platform (ARP ™) dynamic headsets [16]. These headsets are used as active hearing protectors in noisy industrial workplaces: they are instantly custom fitted to the user's ear canal with the help of the inflatable earpiece show in Fig. 3a. They comprise an external microphone and an internal loudspeaker for each ear (Fig. 3). Depending on the sound level and on the type of sound (speech, no-speech, etc.) the external microphone can drive the internal loudspeakers of the ARP headset (Fig. 4). Effects of the headset on the HRTFs has been evaluated by measuring the changes in ITD, ILD and spectral cues with the microphones from the artificial head. It is then possible to correct the HRTFs so that, the user can still localize the sound sources even when wearing the ARP earpiece.

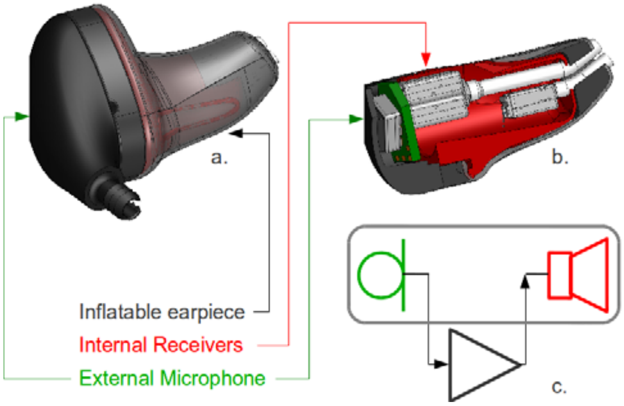

FIGURE 3: Overview of the digital custom earpiece (a), its electro-acoustical components (b), and equivalent schematic with external amplifier (c).

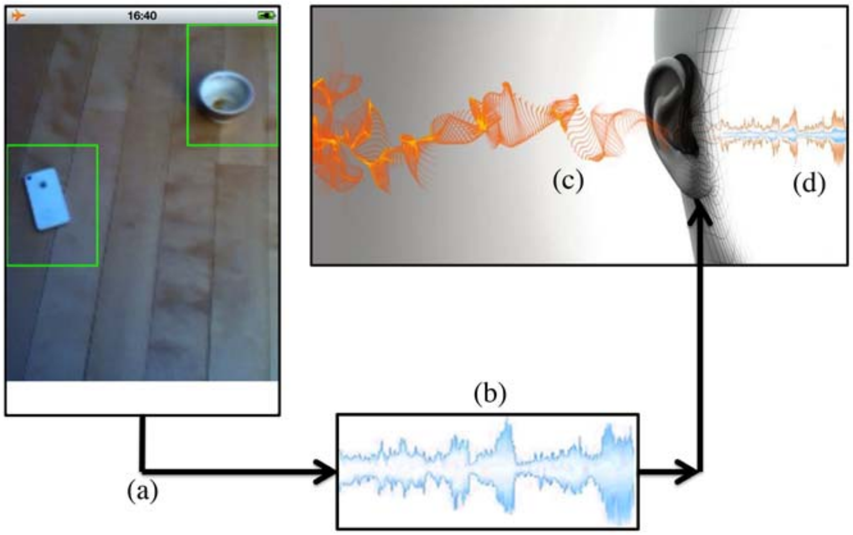

FIGURE 4: Third platform which can mix the sound generated by the iPod/iPad/iPhone (b) and the external sounds (c). Two objects are localized by the iPod system (a) and the two appropriate sounds are synthesized (b). The surrounding sound (c) is captured and mixed with the artificial sound (b) by the digital earplugs. The listener can then perceive simultaneously the surround and the virtual sounds (d).

TESTS TO LOCATE OBJECTS OR SOUNDS WITHOUT VISION

Five subjects have been asked to *i)* move in a cluttered corridor with different objects on the floor *ii)* grasp objects on a table and *iii)* locate sounds. They do not have any visual deficiencies and were not trained. A band was placed on their eyes. The Apple conventional headset (iPod 4th. generation headset) has been used to conduct all experiments.

Finding the way from one room to another room through a cluttered corridor

TABLE 1: Number of collisions with objects. Five subjects and five trials; three objects were placed in the corridor: office's bin, school rucksack and a pair of shoes. Objects were displaced between each experiment.

	Experiment 1	**Exp. 2**	**Exp. 3**	**Exp. 4**	**Exp. 5**
Subject 1	2	1	**0**	1	**0**
Subject 2	1	**0**	**0**	**0**	**0**
Subject 3	2	1	1	**0**	1
Subject 4	3	1	2	1	1
Subject 5	**0**	1	2	1	**0**

When taking into account all trials for each subject, the averaged success rate is 36% with no collision with objects, and 80% for at least one collision. Nevertheless, when looking at results for the last 2 experiments for each subject, one sees that the number of collision is at maximum equals to 1. This suggests that after training performance can increase.

Locate objects on a table

Subjects sat in front of a table and had to locate three objects: a computer mouse, an iPhone and a coffee mug. Each subject had 30 seconds to locate as much objects they can. Objects were displaced between each experiment. Five experiments have been conducted. The last was the most difficult as 2 objects were very close (between 2 and 4 centimeters).

The location success rate is of 88%. Two subjects decided not to be static and moved the iPod around the table and the objects for a better and faster localization.

Sound Localisation

TABLE 2: Sound source localization: YES for a correct localization and NO for an incorrect localization. A is for Azimut (-90<=>left, 0<=>middle and 90<=>right) and E is for elevation (-40<=>bottom et 45<=>top)

	Test 1	**Test 2**	**Test 3**	**Test 4**	**Test 5**
	A=-90E=0	**A=90E=0**	**A=0E=0**	**A=0E=-40**	**A=-90E=45**
Subject 1	**YES**	**YES**	**YES**	NO	NO
Subject 2	**YES**	NO	**YES**	**YES**	NO
Subject 3	**YES**	**YES**	NO	NO	**YES**
Subject 4	NO	**YES**	**YES**	**YES**	**YES**
Subject 5	**YES**	**YES**	NO	NO	NO

To test the new HRTF, subjects were sitting and were asked to locate sounds played with a computer through the HRTF and the conventional headset. The sounds are the same than the ones for the iPod's application – musical notes. Subjects were asked to give a vertical location (top, middle, bottom) and a horizontal location (left, middle or right). Five different locations were used. The first three were presented in the horizontal plan with azimuths of -90, 0 and + 90 degrees (left, front or right) and the two last ones where at a different elevation.

With 73% of good location on the horizontal plane but only 40% on the vertical plane, it is observed that location on the vertical plane is in fact more complex and subtle. The same HRTF (the one corresponding to the artificial head) is used and seems to be more adapted to subject 4 for the elevation (which is mostly based on the spectral cues of the HRTF) and to subject 1 for the azimuth (which is mostly based on ILD and ITD).

Conclusion

A polyvent system has been presented. It is simple and efficient. Reported preliminary results show that – even with that simple video analysis performed by a bio-inspired neural network – many applications can be designed. They can be from reeducation to games. There is also a potential for assistance to blind or low vision people. A prototype is planned to be available for evaluation through the iTunes store.

ACKNOWLEDGMENTS

Financial support to Damien Lescal and Jean Rouat is from FQRNT and Université de Sherbrooke through the support of NECOTIS research project on sensorial substitution. HRTFs measures have been done with the ICAR facilities which is a result from a collaborative research effort between ÉTS and the *Institut de recherche Robert-Sauvé en santé et en sécurité du travail* (IRSST). The authors would also like to thank the Sonomax Technologies Inc. and its Industrial Research Chair in In-ear Technologies for providing specific equipment required for the ARP experimental setup. ICAR is composed by a 125 m^3 reverberation room and a 95 m^3 semi-anechoic room. NECOTIS is an interdisciplinary (Signal Processing and Neurosciences) and inter-institutional research group (U. Sherbrooke, U. Montréal, CRC Ottawa).

References

[1] T. Akinbiyi, C. E. Reiley, S. Saha, D. Burschka, C. J. Hasser, D. D. Yuh, and A. M. Okamura, "Dynamic augmented reality for sensory substitution in robot-assisted surgical systems.", Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society. **1**, 567–70 (2006), URL `http://www.ncbi.nlm.nih.gov/pubmed/17945986`.

[2] L. B. 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**, 132–8 (2009), URL `http://www.ncbi.nlm.nih.gov/pubmed/19104453`.

[3] S. Hanneton, M. Auvray, and B. Durette, "The vibe: a versatile vision-to-audition sensory substitution device", Applied Bionics and Biomechanics **7**, 269 –276 (2010), URL `http://www.informaworld.com/10.1080/11762322.2010.512734`.

[4] P. Meijer, "An experimental system for auditory image representations", Biomedical Engineering, IEEE Transactions on **39**, 112 –121 (1992).

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

[6] 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 / The Royal Society **266**, 2427–33 (1999).

[7] M. Torres-Gil, O. Casanova-Gonzalez, and J. Gonzalez-Mora, “Applications of virtual reality for visually impaired people”, WSEAS Transactions on Computers **9**, 184–193 (2010).

[8] J. Schnupp, I. Nelken, and A. King, *Auditory Neuroscience: Making Sense of Sound* (The MIT Press) (2011).

[9] D. Lescal, L.-C. Caron, and J. Rouat, “Neural visual objects enhancement for sensorial substitution from vision to audition”, in *IEEE int. Conf. on Information Science, Signal Processing and their Applications* (2012).

[10] J. Rouat, J. Bergeron, L.-C. Caron, V. de Ladurantaye, and F. Mailhot, “Method, system and aggregation engine for providing structural representations of physical entities”, (2012), PCT patent application WO2012167359.

[11] V. de Ladurantaye, J. Rouat, and J. Vanden-Abeele, “Models of information processing in the visual cortex”, in *Visual Cortex - Current Status and Perspectives, Stéphane Molotchnikoff and Jean Rouat (Ed.)* (2012), http://dx.doi.org/10.5772/50616.

[12] V. de Ladurantaye, J. Rouat, and S. Molotchnikoff, “Object perception using biologically realistic binding by synchrony”, in *Society for Neurosciences annual meeting*, 712.06 (2012), URL `http://www.abstractsonline.com/`.

[13] V. Algazi, R. Duda, D. Thompson, and C. Avendano, “The CIPIC HRTF database”, in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 99 –102 (2001).

[14] URL `http://interface.cipic.ucdavis.edu/`.

[15] G.R.A.S, `http://ansihead.com/` (2012).

[16] Sonomax Technologies Inc., “(QC, Canada)”, `http://critias.etsmtl.ca/arp` (2012).