SPOTTY: IMAGING SONIFICATION BASED ON SPOT-MAPPING AND TONAL VOLUME

Grigori Evreinov

Computer-Human Interaction Group
Dept. of Computer and Information Sciences
FIN-33014 University of Tampere, Finland
evreinovg@usa.net

ABSTRACT

A basic question at image sonification is the image segmentation. A cognitive model of visual processing in a greater degree could define possible ways of sound mapping. For instance, the scanpath theory suggests that a top-down internal cognitive model, of what *we see*, drives the sequences of rapid eye movements and fixations or glances that so efficiently travel over scene or picture of interest. The scanpath theory may be applied at sonification of visual image. But it is necessary to solve, what is more important in each stage of the image recognition process: the scan trajectory or the optical characteristics of its extreme positions? That is to say, what is dominant - scanpath or the spot of glance? I hope a solution of these questions will allow to develop new tools for VR applications as well as to continue designing of visualization system for blind people on a basis of blind-eye tracking.

1. INTRODUCTION

There is an opinion that a picture is a static image and therefore there are difficulties in a static parameters' transformation of the image into something that is time-varying as sound, by natural way. Therefore many different sonification methods assume to set only quantitative conformity of visual parameters to the sound ones. It is supposed too that a person may learn any sound code (like as Morse code) and substitute visual mental notions (percepts) by hearing complicated sound patterns. Sometimes investigators argue applicability of their methods are based on such phenomenon as synesthesia. However, till now only simple contour objects (geometric or graphic primitives) evoke visual sensations similar to their prototypes. A similarity degree of the subjective sound image and visual one depends on how the natural mechanisms of visual processing were used at sound mapping.

Except linear function, the quadrangle is frequently used object at sonification. From the viewpoint of sonification the quadrangle, as well as its spatial projections, is enough by entire and convenient visual model. We can consider quadrangle as a piece of the picture, as math function or as a geometric figure. In any case we may observe visual object which presents a piece of the plane restricted from four sides (in two dimensions) and by some way oriented in a space.

A perception of the quadrangle is a very complex process. Notwithstanding what a figure is static object, its perception is dynamic process occupying relatively much time. During this time human eyes are moved at *the most informative parts* of the image [1] and successively fixate relative shifts of qualities or parameters of the spots of glance. Thus a perception includes elements of an active orientation, which are *anticipations* of the future results of action. Not follows to think that the activity of perception can act only during macro intervals of time and is expressed as the motions of eyes, head, hands etc. The shifts of attention in a visual field, well familiar to each, can occupy only insignificant part of one visual fixing, a duration of which is about one third of second. Nevertheless these changes of a position of *consciousness' focus* are necessary, according to Irwin Rock, for invariant perception of a mutual layout of internal details of the object [2]. They base on a preliminary rough observation including localization of the object or its piece in a space, because for a subsequent detailed analysis it is natural to choose comparison points or coordinates or other comparative characteristics of stage's objects, while these primitives or entire images could already exist as mental models. Therefore at sonification follows to take into account possible analogous hearing processes.

From the behavioral viewpoint an internal representation (a model of notion) is formed in a brain during conscious observation and active examination. An active examination is aimed toward the finding and memorizing of functional relationships between the applied actions and the resulting changes in sensory information. An external object becomes *known* and may be recognized when the system is able to subconsciously manipulate the object and to predict the object's reactions to the applied actions. According to this paradigm, the internal object representation contains chains of alternating traces in motor and sensory memories. Each of these chains reflects an alternating sequence of elementary motor actions and sensory signals (proprioceptive and external) which are expected to arrive in response to each action. Perhaps, the brain uses these chains as *behavioral programs* in subconscious *behavioral recognition* when the object is known (or is assumed to be known). The *behavioral paradigm* was formulated and developed in the context of visual perception and recognition in a series of works [1, 3-5]. A recognition process is supposed to consist of an alternating sequence of eye movements (evoked from the motor memory and directed by attention) and of verifications of the expected image fragments (recalled from the sensory memory). In other words, the scanpath theory suggests that a top-down internal cognitive model, of what *we see*, drives the sequences of rapid eye movements and fixations, or glances, that so efficiently travel over scene or picture of interest.

Hence, if visual image is static, an analyzer may simultaneously apply several & different ways of the image processing (Figure 1). Analogously at sonification, not only the subjective sound image arising in the end of hearing might be subject to a specific recognition of object features, but it is necessary (it is extreme important) directly during playback to choose a particular sequence of the pattern representation (or of its pieces) according of a prospective cognitive model that *we assume to see*.

Philosophers have speculated what *we see in our mind's eye*, but till now the scanpath theory is a little evidence in supporting of this conjecture. Eye movements are an essential part of vision. An illusion of clarity exists, that *we see* the

entire visual field with equal high resolution, but this cannot be true because of the fovea only has such resolution within 1/2 - 2 degrees. Eye movements must bring nearer the fovea to each part of a scene or picture or page to be processed with ample resolution. Enough rough, you could imagine that it is necessary to convey an image with the help of a single optical detector (photodiode, for instance), while there is an opportunity to manipulate by its position and to focus an optical system. In such case an image description will include dynamic arrays of brightness and signals of management by mechanical positioning system of the detector and optics.

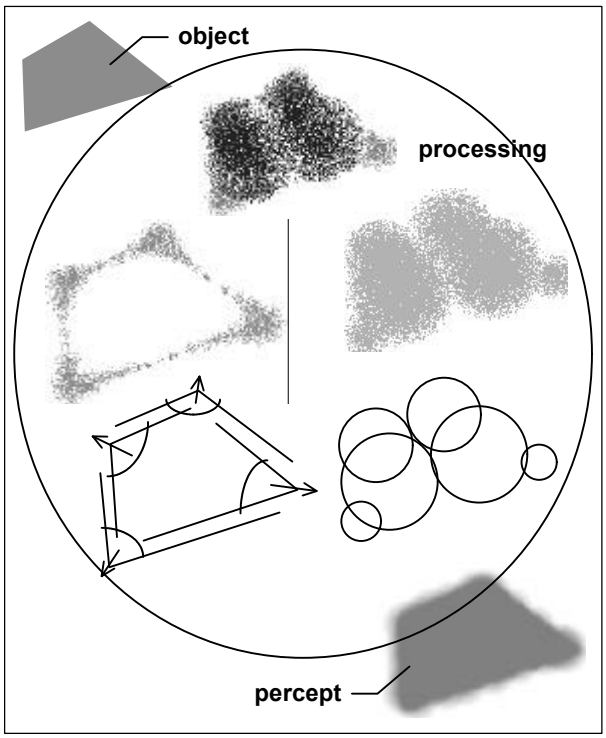

Figure 1. *Image segmentation, detection & evaluation of graphic primitives at observation of visual object.*

Let's return to the quadrangle. Some possible strategies at an observation of visual object is shown in Figures 1, 2. As you can see, depending on gradient of image's density, of local maximums, of borders of heterogeneity a focus of attention may be displaced to processing of the contour or internal surface of object. While, essential deviations in parameters of the contour may initially lead to a deform perception of the homogeneous surface, after systematic eye tracking (so called sweep motions [6]) a revision & improvement of the mental model of observed object occurs. To a conclusion that an observed object is a quadrangle we come after several saccades. However to make a decision that an observed object is not a trapeze we must precisely analyze its angles, sides and their relations. Thus, the graphic prototype contains all necessary information about the elements of the spatial composition and their interrelations.

Hence, a sonification process should at least not reduce a quantity of this information about the prototype. In fact, the image is already fragmented. So the task is that: How to make a sonification most naturally? How to use a suitable modulation of sound parameters for the mapping of the shape and spatial position of these fragments?

2. SPOTTY PROJECT

2.1. Background

Many papers are devoted to a sonification of contour images, graphs and diagrams. Not follows to think that this field was carefully studied, in detail described and was placed into the school textbooks. But much less researchers tried to sonify a arbitrary shape or a planar object, in spite of the fact that such goal et desire exist *ab ovo*.

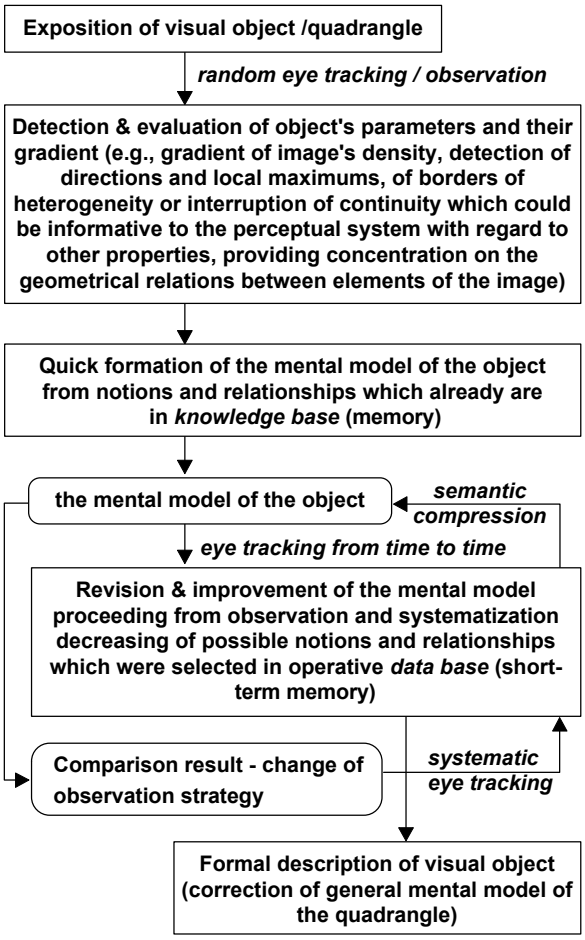

Figure 2. *A formation of mental model at observation of visual object*

A. Hollander [7] studied an opportunity to sonify cylindrical /spherical shells and cubic volume with the help of the virtual sound source. He formed virtual sound source which moved at random within a mathematically defined volume through the Convolvotron(TM). Unfortunately visual/physical model as well as hardware which have been used have not allowed to receive any satisfactory results. He wrote: "To get some sense of the problem, I made visual representations of positions of sound source. I noticed that, even in the visual modality, it was difficult to recognize the shape of a volume with fewer points than a three-thousand. At fifty positions per second, a full minute would be required to visit this number of points. The graphical representation displayed all of the points simultaneously, what would not be the case for the auditory stimuli. The listener would need to rely on memory, or some degree of *persistence of hearing* to build up an image. Since volumes seemed to be a problem, as a potential fix, some of the sound shapes presented were shells instead of volumes. Visual inspection of shells indicated that only three seconds worth of points (150) might do." A positive meaning of this idea is that a sonification, as well as image recognition, would be carried out on a basis of the texture analysis. A lack of idea is that the author assumes to present a homogeneous texture through a random dotted scanning of the object. He does not take into account a stimulus dependent

behaviour of visual analyzer, i.e. a dependence of concentration and switching of attention as a function of spatial density of the image texture, gradient of this parameter in a neighbourhood of edgings of object, change of strategy of observation at detection of known attributes and so on. At absence of positive results, Hollander has turned a research into more real domain - a sonification of contour images.

P. Meijer [8] proposed a sonification method of arbitrary pictures registered by a video camera. His approach is based on agreement that the amplitude of a sinusoidal oscillator is proportional to the pixel gray level, the frequency is dependent on a vertical pixel position, the image is scanned through one column at a time, and the associated sinusoidal oscillator outputs are presented as a sum (chord) followed by a click before the presentation of a new sequence of columns. A significant aspect that was used in this display is image contrast. Meijer's display maps contrast onto the intensity of each frequency component, producing a "16-level gray-scale" output. The image is built up through columns' sequence, sampling from left to right. Thus, the horizontal dimension is simply mapped onto the time interval since the last click. While in a later version Meijer added horizontal localization cues, image scanning is carried out in two dimensions independently from any of known visual models of the eye tracking.

But how to combine a sonification of texture and direction dynamics of glance? What is more important at recognition of image piece: a scan trajectory or optical characteristics of extreme positions (of local maximums, of borders of heterogeneity etc.)? That is to say, what is dominant in sequence of eye actions - scanpath or the spot of glance? What sound parameters may be applied at a sonification, taking into account a possible strategy dynamics of visual perception?

Meantime, as Hollander had mentioned, there is a considerable body of research devoted to non-directional but spatial *size* of sounds such as a tonal volume and that could be considered as a unidimensional manifestation of auditory shape. In particular, von Bekesy [9] had performed a geometrical assessment of this attribute. He asked his listeners to give a numerical estimate (in centimeters) of the *width* of the test tones. As expected, he observed that the subjective width increased with loudness and decreased with frequency. However, von Bekesy also observed that the perceived size of the auditory image was of an actual size of the loudspeaker that produced the sound. Perrott *et al.* [10] have attempted also a two dimensional analysis of volume. Their results reaffirmed the findings of von Bekesy on the horizontal dimension, but they were not able to extend the definition of tonal volume to include the vertical dimension. Judgments of tonal volume on the vertical dimension were not significantly different from judgments of loudness.

Now by using a technique of the formation and spatial positioning of virtual sound source (via interference maximum of four loudspeakers) there is an opportunity to change a tonal volume independently of the actual size of the loudspeakers. Besides that, during sonification of the contour a virtual sound source may be in each spatial position a very limited time (much less than 50 ms), therefore a formation of the envelope through volume modulation could have a sense within entire sound pattern of the contour or of plane piece.

A volume could be perceived as a natural remoteness attribute of sound source or of spatial position of separate points of sound trajectory [11]. On the other hand, a remoteness is only a derivative of the perceived size of the object. Studies show tonal volume is increased with the intensity and duration of sound, and is inversely proportional to its pitch [10]. The hypotheses as to the nature of tonal volume are various. Some describe it as a cognitive phenomenon stemming from associations formed between types of sounds and their sources. Other theories are physiological and suggest that tonal volume is correlated with the area of the basilar membrane affected by a sound [12]. Unfortunately, none of the theories of the phenomenon nature have been thoroughly explored. It's clear that on short time scales, loudness is dependent on duration. For the first few hundred milliseconds the sound energy is integrated to produce the subjective impression of loudness [6].

To study tonal volume of virtual sound source and an opportunity of using this phenomenon a special program was designed. Figure 4 shows a snapshot of the program '*Spotty*'.

2.2. Sonification design

Initially, it was assumed that depending on a given accuracy a surface of any image may be segmented into a final amount of spots. Formally, we could say that the spot is a homogeneous piece of the surface (regarding a texture), the form of which is close to the spot of glance. Within a visual designing field each spot may have spatial coordinates, optical characteristics and sound attributes coupled to visual parameters. It was assumed too that a playback of spots must occur in the same sequence that could be registered through known methods of eye tracking. While for a sonification of random sequence of virtual sound sources two oscillators are ample, their parameters should be changed depending on the spatial coordinates and optical characteristics of the spots. A frequency of embedded oscillators of the special sound card which was used [13] may be set within a range 170-5000 Hz and its deviations along of X-Y coordinates occur so that:

$$F(X,Y) = F_0 \times (1 - ((P - Y) / P) \times DevF(Y) / 100) + (X \times DevF(X) / 100) \quad (1)$$

where, DevF(Z), DevF(X) is the frequency deviation percentage along Y, X axes correspondingly; F_0 is the frequency as a percentage of a maximal value (256), a parameter value in Hz you can see in Figure 3; P is a maximal value of the parameter or a number of points in the image plane (256).

%	Frequency, Hz	%	Frequency, Hz
02 ÷10	172 ÷ 188	57 ÷ 65	381 ÷ 481
10 ÷17	188 ÷ 203	65 ÷ 73	481 ÷ 648
17 ÷27	203 ÷ 231	73 ÷ 83	648 ÷ 982
27 ÷37	231 ÷ 267	83 ÷ 91	982 ÷ 1982
37 ÷47	267 ÷ 314	91 ÷ 96	1982 ÷ 4982
47 ÷57	314 ÷ 381		

Figure 3. *Definition of* F_0 *value*, in Hz.

Deviation direction of frequency is chosen proceeding from perception intermodality of the hearing space [13], assuming that zero is the left top point of the visual designing field (or of a virtual acoustic plane).

To create a virtual sound source having loudness B in a random point of the virtual acoustic plane, the amplitude of four loudspeakers (V_A,..V_D), located in front of the listener in the corners of acoustic plane, is redistributed depending on values of X-Y coordinates according to equations:

$$V_B = (P - X)(Y / P) \qquad V_C = X \times (Y / P)$$
$$V_A = (P - X)((P - Y) / P) \qquad V_D = X \times ((P - Y) / P) \quad (2)$$

while $B \sim \Sigma (V_A, ..V_D)$.

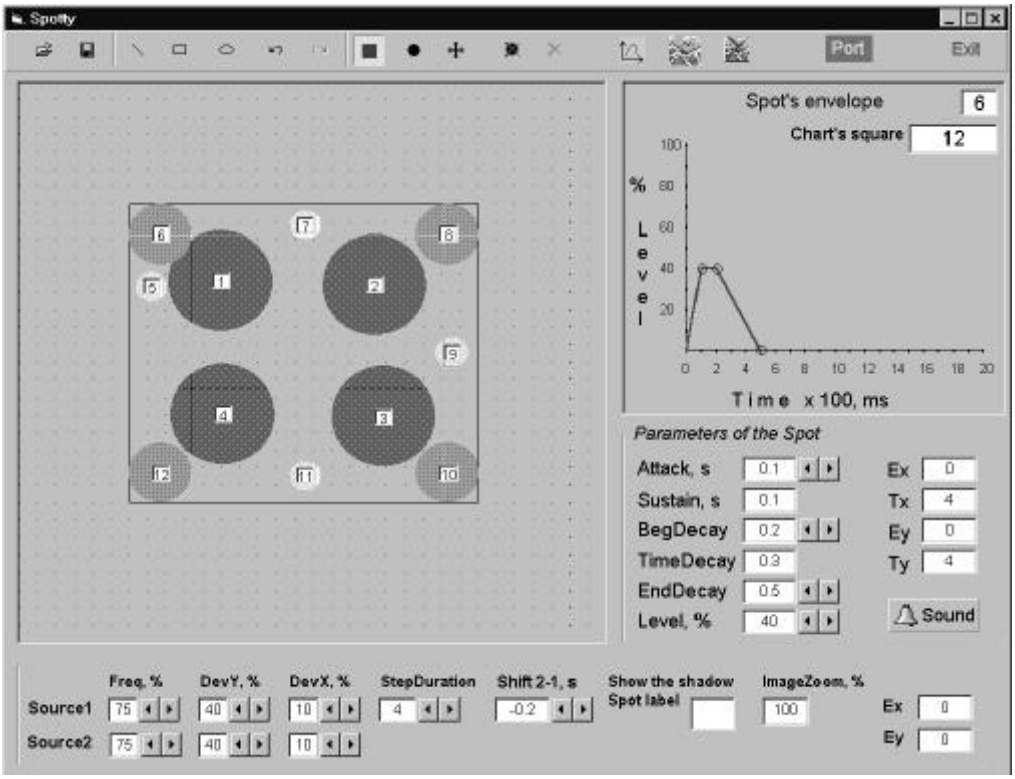

Figure 4. *Snapshot of the program 'Spotty'.*

The program allows to set within a visual designing field any image, to change its scale ('ImageZoom, %') and to substitute the image through sequence of spots (till now by hand). To facilitate the task there are additional options which allow to create standard geometrical forms (line, broken line, rectangle, ellipse/circle), while loudness ('Level, %') was connected with chart's square of the envelope and a gray-scale level of the spot (previously as an exponential function).

To control by tonal volume of the virtual source, sound envelope was formed for each spot. Sound envelope consists of conventional intervals of time: *Attack, Sustain, Decay*. While a step of changing of all parameters is defined by value of 'StepDuration'. I.e. if value of 'Attack' was set 100 ms, 'Level' - 40 % (of 256) and a value of 'StepDuration' was 4, hence on this interval an increment of loudness was 1.6 % . As in previous researches [11, 13] pure tones are not used. A form of sound signal was close to trapeze, taking into account a band-width of playback canal. Therefore additional distortions were not listened and interpolation within envelope was not applied.

However, sequential playback of sound spots belonging to a surface having homogeneous texture evokes essential infringement of homogeneity perception. During visual scanning (switching of attention between two glance spots) a perception persistence is provided even on retinal level, due to definite time of restoration of visual pigments. To provide some degree of *persistence of hearing*, spots' sound envelopes must have a cross. A value of 'Shift 2-1' option is intended for realization of such function. But since we used only two oscillators, a value of 'Shift 2-1' option may not be more than a half of the least duration of sound envelope if an amount of spots more than two. Such limitation evokes inconvenience when sonification duration and loudness of spots varies in a wide range.

Later one more function was added to parameters of the control by envelope of sound spot. Often a form of surface having homogeneous texture is extended in one direction, and spot of glance is not a circle in a strict sense too. During sonification to form not only round spots, such options as eccentricity ('Ex', 'Ey') and deviation period of spot's coordinates (Tx, Ty) were entered. In this program version an eccentricity is dependent on the period Tx and a total amount of sound points within envelope (N) as a sinusoidal function:

$$Ex = Ex_0 \times \mathrm{Sin}(Tx \times \pi \times k / N)$$
$$Ey = Ey_0 \times \mathrm{Sin}(Ty \times \pi \times k / N) \qquad (3)$$

where Ex_0, Ey_0 - are initial values of the eccentricity and k takes values from 0 to N. Accordingly, in equations (1) and (2) all values of coordinates will receive shift onto eccentricity.

2.3. Quadrangle sonification

Initially we have supposed to process the image according to some strategy, for instance along of the texture's gradient. However, real eye tracking is essentially dependent on a content of image and an observation goal [1]. Even an observation of contour images has individual particularities. Therefore there is at least two approaches. According to the first, an image segmentation into a spots' sequence may be implemented after a registration of eye tracking and a scanpaths' analysis during representation of real image were produced. Another approach consist of that to map visual image through spots according to definite strategy, to sonify them and then to implement a registration of eye tracking during representation of virtual sound object, by reaching gradually of a recognizing accuracy. Spotty program provides both methods. In particular, after designing of the virtual sound object via clicking within the form all objects are hide and 'Sound' option (a playback) of the sound object is carried out after pressing space-key, so that through a definite time of delay at forming the strobe into definite 'Port' to start 4250R+ eye tracker of Applied Science Laboratories.

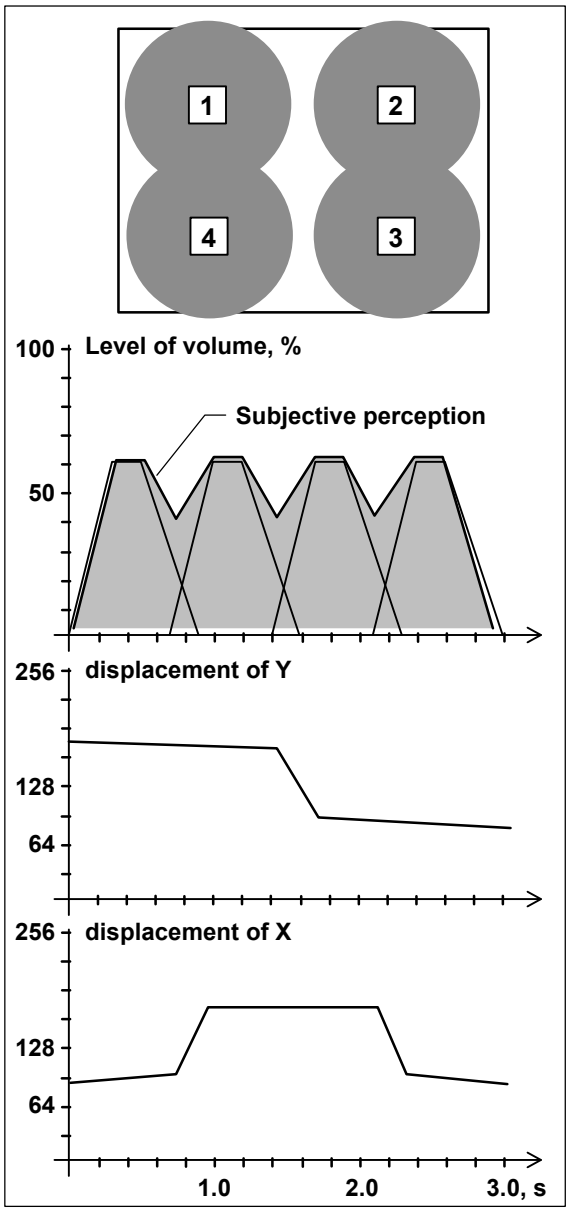

Figure 5. *Surface's sonification of the quadrangle.*

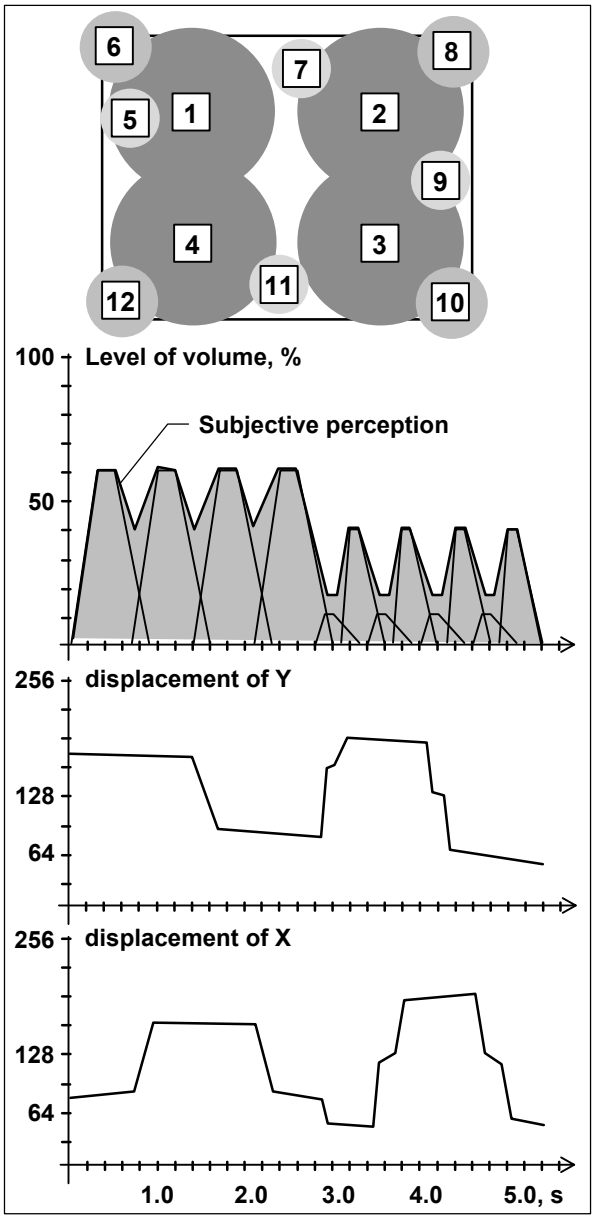

Figure 6. *Emphasizing of the corners.*

The system of eye's tracking registration does not contact the user and allows the subject to move freely. Recorded data may include time, X and Y eye position coordinates and pupil diameter. External data events/marks can be recorded along with eye tracker data. Eye position coordinates correlate to specific areas on the monitor screen being viewed. Head position and orientation can also be recorded. That allow to fulfill the objective investigation: a comparison of visual and sound mapping of the same images.

Figure 5 shows an example of the quadrangle sonification through tonal volume of virtual sound source and temporal diagrams of forming the subjective image. Sound spots may be placed within graphic object or closer to corners. A sequence of playback is marked by numbers.

Figure 6 shows an example of the quadrangle sonification with more detailed sound description. Emphasizing of the corners is implemented by using the spots 6, 8, 10, 12. The spots 5, 7, 9, 11 are necessary to intensify an impression of angles' direction through attention switching between triads 5-6-7, 7-8-9, 9-10-11, 11-12-5.

Figures 7-8 show a spatial modulation of the sound spots at using of E, T factors (3). At Figure 7 there is an attempt to combine a tonal volume and spatial deviation of sound source to present a perimeter of graphic object. Figure 8 shows an opportunity of sonification of a surface and corners of the quadrangle through spatial modulation and tonal volume of sound source. A perception of a surface formed by interference of two sinusoids (deviations of X & Y) is enough diffuse.

Author is conscious that a proposed type of intermodal transformation is not a natural mapping, but this is an attempt to use possible mechanisms of perception in sonification and vice versa to use sonification as a tool for cognitive research of perception issues.

3. REFERENCES

[1] Yarbus, A.L., *Eye movements and vision*. New York, Plenum Press, 1967.

[2] Rock, I., *An Introduction in Perception.* "Pedagogika", V. 1. Moscow, 1980.

[3] Stark, L.W., *Top-Down Vision in Humans and Robots*. BISC Seminar, University of California at Berkeley, 1997.

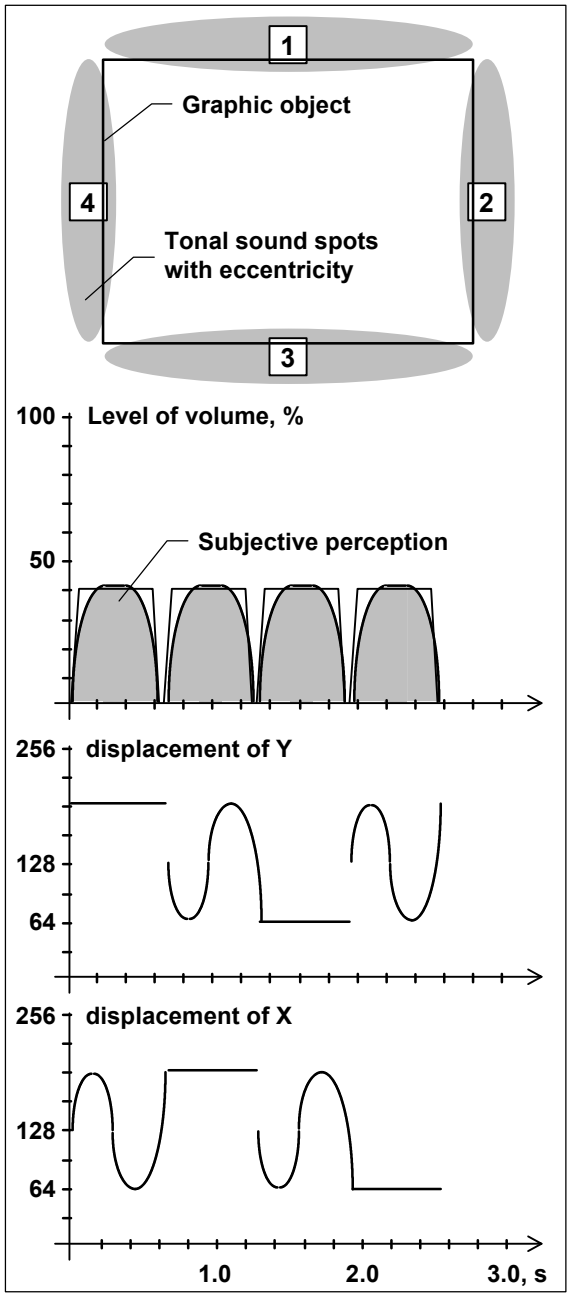

Figure 7. *Perimeter's sonification of the quadrangle via spatial modulation of sound spots.*

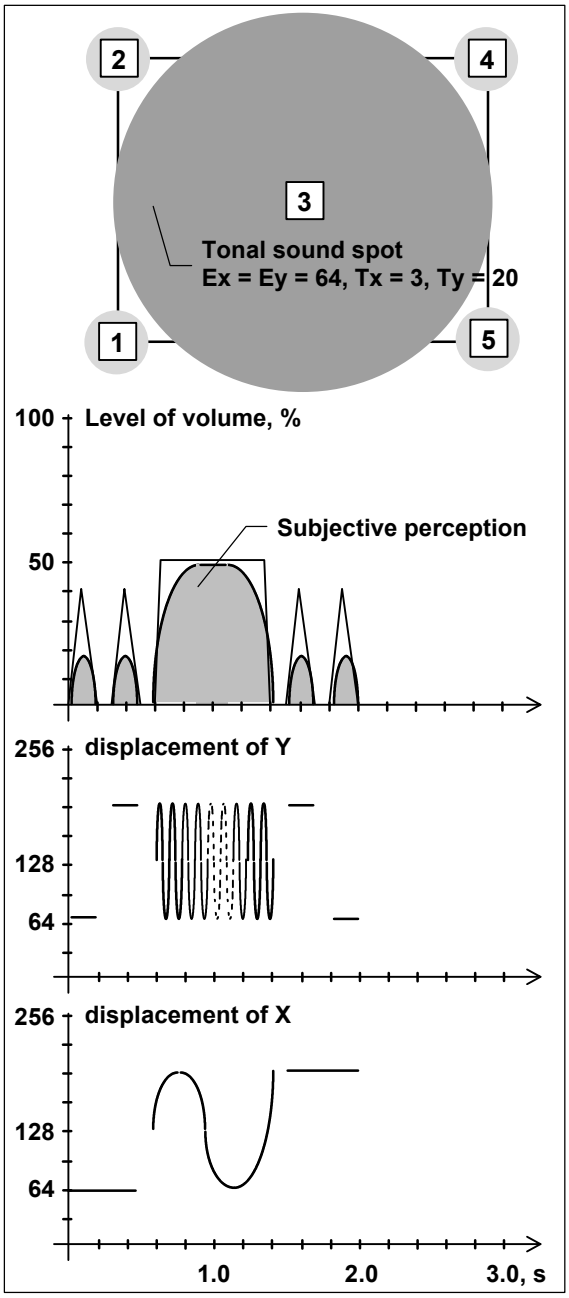

Figure 8. *Sonification of surface & corners of the quadrangle through spatial modulation and tonal volume of sound source.*

[4] Deubel, H., & Schneider, W.X., "Saccade target selection and object recognition: Evidence for a common attentional mechanism," *Vision Research*, 36, pp. 1827-1837, 1996.

[5] Driver, J., & Baylis, G. C., "Movement and visual attention: The spotlight metaphor breaks down," *J. of Experimental Psychology: Human Perception & Performance*, 15, pp. 448-456, 1989.

[6] Aaltonen, A., Hyrskykari, A., Raiha, K.-J. "101 Spots, or How Do Users Read Menus?" *Human Factors in Computing Systems*, *Proc. of CHI'98*, April 1998, Los Angeles, pp. 132-139.

[7] Hollander, A. J., "An Exploration of Virtual Auditory Shape Perception." Ph.D. thesis, Human Interface Technology Lab, University of Washington, USA, 1994, URL: www.hitl.washington.edu/publications/hollander/

[8] Meijer, P. "An experimental system for auditory image representation," *IEEE Transactions on Biomedical Engineering*, vol.39, no.2, pp.112-121, Feb. 1992.

[9] von Bekesy, G., Wever, E. G.(Ed), *Experiments in Hearing*, New York, NY: McGraw Hill, 1960.

[10] Perrott, D., Musicant, A., & Schwethelm, B., "The expanding image effect: The concept of tonal volume revisited," *J. of Auditory Research*, 20, pp. 43-55, 1980.

[11] Edwards, A.D.N., Evreinov G. E., Agranovski A.V., "Isomorphic Sonification of Spatial Relations," in: *Human-Computer Interaction: Ergonomics and User Interfaces, V. 1 of the Proc. of HCI International '99*, Munich, Germany, August 22-26, 1999, pp. 526-530.

[12] Perrott, D., & Buell, T., "Judgements of sound volume: Effects of signal duration, level, and interaural characteristics on the perceived extensity of broadband noise," *J. Acoust. Soc. Am*., vol. 72, no. 5, pp. 1413-1417, 1982.

[13] Agranovski A., Evreinov G., "Creating Virtual Sound Objects," *J. Acoustical Physics* ISSN 1063-7710, vol. 46, no. 1, pp. 8-14, 2000.