***I. Kliuiev**, Postgraduate student*
***A. Morozov**, PhD in Tech. Sci., As. Prof.*
Zhytomyr Polytechnic State University

PARTICLE-BASED AUDIO REPRESENTATION OF VISUAL SCENES FOR SENSORY SUBSTITUTION

Sensory substitution is a field of research in which information that is normally received through one sensory channel is delivered through another. For people with severe visual impairments, visual-to-auditory systems are of particular interest because they transform properties of a scene into structured sound. One of the best-known systems of this kind is The vOICe, proposed by P. Meijer in 1992 [1]. Its core idea is to preserve as much raw visual information as possible in the auditory signal without first converting the scene into interpreted objects. In the classical version, the image is scanned from left to right; vertical position is mapped to pitch, brightness to loudness, and horizontal position to time. As a result, the user receives a stream of low-level features from which spatial and structural regularities must be learned and interpreted by the brain itself [1, 2].

Later developments showed that color can also be incorporated into this type of representation. The EyeMusic system follows a similar scanning logic, but uses musical notes instead of simple tones and encodes colors through different musical instruments. Studies with EyeMusic showed that adding a color channel can improve perception: the best achieved visual acuity increased from 20/800 to 20/400 [3, 4]. Related approaches include See ColOr, where color and depth are translated into spatially distributed instrument-like sounds, and Colorophone 2.0, which generates live stereo soundscapes based on color information [5, 6]. These systems suggest that color in sensory substitution should be treated as an informative dimension of the visual signal rather than as a secondary add-on [3–6].

At the same time, the universality of the audio-raster approach also reveals its limitations. The central difficulty lies in feature encoding together with the need to reduce auditory overload. In The vOICe and similar systems, the user receives a dense spectro-temporal flow that still has to be integrated into boundaries, objects, and spatial relations. This increases the demands placed on training and increases the load on short-term memory. In EyeMusic experiments the authors deliberately used a 40×24 resolution in order to avoid overloading the auditory system with too many simultaneous notes [4]. In addition, studies in auditory scene analysis show that for sonified visual structures it is not enough to preserve information alone; the signal must also remain structured in a way that allows the auditory system to group it into meaningful perceptual objects [7].

A different logic is embodied in the Sound of Vision project, launched in 2015 within Horizon 2020 [8]. Unlike The vOICe, it is focused primarily on navigation, collision avoidance and spatial perception and therefore relies on prior scene interpretation and the extraction of relevant environmental elements. In publications describing the system, it is presented as a wearable sensory substitution device that continuously scans the surroundings, detects relevant features, and delivers them through audio and haptic channels [8, 9]. To obtain depth information indoors, in low

light, and in darkness, the system uses a structured-light 3D sensor, while outdoors under normal or bright lighting it uses a stereo vision system. The architecture includes iterative and continuous full-scene rendering modes [9].

For the present work, the most relevant aspect of Sound of Vision is the way it sonifies depth maps. In the paper by Spagnol et al., the algorithm operates on raw depth maps, divides the scene into sectors, and computes two parameters for each sector: map density and average depth [10]. Sector density controls the average rate of bubble generation: the more relevant points appear in a sector, the more intense the bubble stream becomes. Average depth controls the maximum depth at which bubbles are generated in the liquid model: closer objects correspond to bubbles created nearer the surface, which makes them louder and more likely to exhibit a noticeable rising pitch. Vertical position is additionally encoded through bubble size: upper sectors are represented by smaller bubbles with a higher resonant frequency and a lighter, hiss-like character, whereas lower sectors are represented by larger bubbles with a lower, more gurgling timbre [10].

The present work is aimed at a more general representation of a visual scene, closer in its task formulation to The vOICe and EyeMusic. The goal is to transform a broad set of visual characteristics, including contours, surfaces, regions, textural differences, color zones, etc. At the current stage of the research, convolutional layers are being considered as a preprocessing mechanism for simplifying the scene before sonification. This choice is motivated by the fact that early biological vision relies on local receptive fields and multi-scale spatial filtering, while convolutional networks were historically inspired by related ideas and became a practical computational mechanism for extracting local patterns and contours [11–13].

After this preprocessing stage, the study considers a transition to particle-based audio representation. In this framework, the scene is represented through a limited set of short sound events with controllable parameters such as duration, spectrum, density, loudness, spatial position, and motion. Unlike Sound of Vision, where bubbles are primarily tied to the depth matrix, the current work considers several classes of particles for different types of visual structures: separate particle families may be used for edges, surfaces, textured regions, moving objects, or dynamic changes in the scene. This idea is consistent with findings from auditory scene analysis, in which perceptual grouping is essential to the formation of coherent auditory objects [7].

Another idea explored in the current work is subtractive colorification. In EyeMusic, See ColOr, and Colorophone, color is mainly introduced through separate instruments, timbres, or predefined sound classes [3–6]. In the present work, another option is considered: a base signal or a base particle class is modified by spectral masks, that is, by attenuating or removing specific frequency regions. The purpose of this approach is to preserve a more coherent sonic space of the scene and to avoid excessive timbral fragmentation. At this stage, this remains a working hypothesis. However, it is grounded both in the already demonstrated usefulness of color in sensory substitution systems and in the broader goal of reducing unnecessary complexity in the resulting soundscape [4, 6, 7].

Another important element is the attention mechanism. In the current concept of the system, the user should be able to control the focus of scene exploration through head movements or mouse input, shifting the region of increased detail and the field of view. This makes it possible to combine a coarse global representation of the scene

with local detailed inspection and better reflects the active nature of perception. In parallel, a separate navigation mode is being considered. In this mode, the scene would first be interpreted by computer vision algorithms, with explicit extraction of obstacles, walls, people, suspended objects, level changes, and other relevant elements, and only then transformed into a more semantic sound representation following a logic closer to Sound of Vision and SoundSight [9, 14].

As for spatial rendering, the current stage of the work focuses on an ITD/ILD-based approach for direct scene exploration. This approach is simpler, less computationally demanding, and naturally suited to horizontal localization. HRTF remains an important reference point for future studies, but its application becomes more difficult when many simultaneous sound sources must be rendered. Recent reviews emphasize that HRTFs depend strongly on the individual geometry of the head, pinnae, and torso, while high-quality spatial reproduction requires either extensive measurement data or reliable personalization. In addition, generic HRTFs often lead to reduced accuracy in vertical localization [10, 15].

Thus, the current research direction combines several lines of prior work: raw-data representation as exemplified by The vOICe [1, 2]; the use of color demonstrated in EyeMusic and related systems [3–6]; navigation modes based on scene interpretation and micro-sound models developed in Sound of Vision [8–10]; and more flexible, user-controlled soundscape approaches such as SoundSight [14]. On this basis, the proposed approach includes convolution-based preprocessing, an attention mechanism, multiple classes of sound particles, subtractive colorification, ITD/ILD-based spatial rendering for the direct mode, and a separate interpreted mode for navigation [16].

REFERENCES

1. Meijer P. B. L. An experimental system for auditory image representations. IEEE Transactions on Biomedical Engineering. 1992. Vol. 39, No. 2. P. 112–121.
2. Haigh A., Brown D. J., Meijer P., Proulx M. J. How well do you see what you hear? The acuity of visual-to-auditory sensory substitution. Frontiers in Psychology. 2013. Vol. 4. Article 330.
3. Abboud S., Hanassy S., Levy-Tzedek S., Maidenbaum S., Amedi A. EyeMusic: Introducing a visual colorful experience for the blind using auditory sensory substitution. Restorative Neurology and Neuroscience. 2014. Vol. 32, No. 2. P. 247–257.
4. Levy-Tzedek S., Riemer D., Amedi A. Color improves visual acuity via sound. Frontiers in Neuroscience. 2014. Vol. 8. Article 358.
5. Bologna G., Deville B., Vinckenbosch M., Pun T. See ColOr: An extended sensory substitution device for the visually impaired. Journal of Assistive Technologies. 2014. Vol. 8, No. 3. P. 135–147.
6. Osiński D., Łukowska M., Kałwak W., Wierzchoń M., Hjelme D. R. Colorophone 2.0: A wearable color sonification device generating live stereo soundscapes—design, implementation, and usability audit. Sensors. 2021. Vol. 21, No. 21. Article 7351.
7. Brown D. J., Simpson A. J. R., Proulx M. J. Auditory scene analysis and sonified visual images. Does consonance negatively impact on object formation when using complex sonified stimuli? Frontiers in Psychology. 2015. Vol. 6. Article 1522.
8. Jóhannesson Ó. I., Balan O., Unnthorsson R., Moldoveanu A., Kristjánsson Á. The Sound of Vision Project: On the Feasibility of an Audio-Haptic Representation of the Environment, for the Visually Impaired. Brain Sciences. 2016. Vol. 6, No. 3. Article 20.
9. Zvorișteanu O., Caraiman S., Lupu R.-G., Botezatu N. A., Burlacu A. Sensory substitution for the visually impaired: A study on the usability of the Sound of Vision system in outdoor environments. Electronics. 2021. Vol. 10, No. 14. Article 1619.
10. Spagnol S., Hoffmann R., Herrera Martínez M., Unnthorsson R. Blind wayfinding with physically-based liquid sounds. International Journal of Human-Computer Studies. 2018. Vol. 115. P. 9–19.

11. Balasubramanian V., Sterling P. Receptive fields and functional architecture in the retina. The Journal of Physiology. 2009. Vol. 587, No. 12. P. 2753–2767.
12. Lindeberg T. A computational theory of visual receptive fields. Biological Cybernetics. 2013. Vol. 107, No. 6. P. 589–635.
13. Lindsay G. W. Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. Journal of Cognitive Neuroscience. 2021. Vol. 33, No. 10. P. 2017–2031.
14. Hamilton-Fletcher G., Alvarez J., Obrist M., Ward J. SoundSight: a mobile sensory substitution device that sonifies colour, distance, and temperature. Journal on Multimodal User Interfaces. 2022. Vol. 16. P. 107–123.
15. Bruschi V., Grossi L., Dourou N. A. et al. A Review on Head-Related Transfer Function Generation for Spatial Audio. Applied Sciences. 2024. Vol. 14, No. 23. Article 11242.
16. SonoVizio Official project website. (2026). SonoVizio — See the world through sound: https://sonovizio.github.io