VIGIL 17: EMBEDDED MONOCULAR DEPTH ESTIMATION AND GRADIENT SONIFICATION IN SITE-SPECIFIC PRACTICE

Daniel MANZ (d.manz@arts.ac.uk) (0009-0004-6716-9596) and
Daniel HEARN (d.hearn@arts.ac.uk) (0000-0001-8437-4517)

Creative Computing Institute, **University of the Arts London**, UK

ABSTRACT

Vigil 17 is a site-specific installation, tracking depth signals from a camera input mapped to an emergent soundscape. Drawing on technical dialogues within Sensory Substitution of Vision by Audition (SSVA) research, this paper presents a novel application of Monocular Depth Estimation (MDE) algorithms to artistic practice. Through practice-led exploration, a framework for translating spatial distance to a procedural audio system is proposed. Here, depth is explored as material, situating the project within established practices of stochastic processes for procedural composition and site-specific sound-art installation. The implementation advances existing SSVA approaches through gradient-based sonification mapping — capturing contours and boundaries, mapped to wavetable synthesis geometry. The viability of real-time embedded MDE-based sonification is demonstrated, performing on-device depth extraction and audio processing. Through public deployment, Vigil 17 suggests a broader ecology — a co-creation process, unfolding through authored technological mediation, between our device and the outside world.

1. INTRODUCTION

Depth is a fundamental spatial signal, encoding the geometric relationship between an observer and their environment [1]. Its perception is central to navigation, object recognition, and spatial reasoning across biological systems. In echolocating species (bats, dolphins, and certain cetaceans [2]), depth is reconstructed actively from reflected acoustic signals, producing continuous, high-resolution spatial maps that guide movement and behaviour in real time. Humans have evolved advanced depth inference capabilities, integrating binocular disparity, motion parallax and perspective cues [3].

Despite its importance to embodied experience, depth has received comparatively little exploration in formal media-arts practices. Where visual and sonic media have each developed rich traditions of procedural and generative composition, depth (as an agent within these systems) remains relatively untapped. Vigil 17 leverages depth as a medium, modulating the timbre of a generative soundscape. Within this proposition, emergent aesthetic properties of (computational) representations of depth are explored.

This response is cast through the lens of unique sensory worlds — each organism constructs its own perceptual reality from the signals available to it. Through the sonification of depth, bio-inspired fabrication, and situated public deployment, Vigil 17 gestures toward a connection between organic and synthetic sensory fields. This suggests a question for practice-based research: how can these self-worlds be authored through the design of autonomous systems, and through what forms might they be made legible?

It is within this framework the present paper situates Vigil 17, a site-specific installation that translates depth signals (taken from live monocular camera input) into a continuously evolving procedural soundscape. The work engages with the technical lineage of sensory substitution by audition devices, where spatial aspects of an environment are encoded as audio in real-time. Where these precedents have been oriented towards rehabilitation and assistive technology, this paper presents a turn towards procedural sound art, foregrounding aesthetic, situated, and experiential criteria.

In advancing this position, a novel framework combining deep learning-based monocular depth estimation with gradient-based signal analysis is introduced, enabling the extraction of spatial contours and boundaries that are subsequently mapped to wavetable synthesis geometry. This is deployed on a Raspberry Pi 5, demonstrating the viability of real-time modern depth extraction and audio processing in an embedded context — it is hoped, that this can feed back into vision substitution by audition in an assistive technology context.

Through deployment in public space (Blackfriars bridge, London) and real-time depth sensing, Vigil 17 engages materially with its site — its unique spatial ecology, sensor configuration, and the contingent processes that unfold within it. The system responds to the environment itself, enacting a form of co-creation between system and passing signal modulators.

2. RELATED WORK

2.1 Depth Sonification

Depth sonification approaches have been used widely in the design of visual prostheses [4]—artificial systems enabling blind subjects to interact with spatial aspects of an

Copyright: © 2026. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

environment, that are normally visually perceived. Early Sensory Substitution of Vision by Audition (SSVA) devices relied on echolocation, based on the same principles as sonar. An ultrasound source emits a series of clicks, using a receptor and telemetry to determine distance between the source and the distant object. Initial telemetric SSVA devices transcribed distance as pitch, and horizontal position as inter-aural disparity [5, 6].

Later systems used optical images picked up by a camera converted into sound, with approximations for relative depth. Meijer's seminal vOICe [7] mapped grey tones of a 64×64 pixel natural image into sound through a left-to-right scanning of the visual scene, embedded within off-the-shelf (digital) components. Further pixel-to-frequency SSVA solutions [8,9] were explored in the following years.

The emergence of RGB-D (Depth) kinect-like devices facilitated a new generations of SSVA devices [10–12], based on explicit depth map computation. Stoll et al. present MeloSee [13], a head-mounted device mapping distance to sound intensity, stereo-modulated laterally with pitch representing verticality. Despite a leap in depth mapping accuracy, such implementations were constrained to structured environments with specialised hardware.

The rise of Deep Learning (DL)-based Monocular Depth Estimation (MDE) algorithms [14–16] removed these constraints, enabling real-time depth extraction from a single camera on embedded devices. The lack of ground-truth depth labels in real-world scenes led the emergence of early self-supervised MDE techniques [17, 18]. SC-DepthV3 [19] is a self-supervised MDE algorithm, using an external pretrained MDE to generate single-image depth prior ("pseudo-depth"), resulting in a low-latency model, performant in dynamic scenes. Despite the rise of MDE technologies, there has been limited conversion to SSVA systems in literature.

Leveraging the SC-DepthV3 [19] model, Vigil 17 demonstrates the viability of an embedded MDE-based sonification system. Here, a novel gradient-based mapping from depth-to-audio is developed — capturing contours and boundaries, mapped to wavetable (synthesis) geometry. The work presented is situated outside of the assistive technology paradigm, engaging instead with the material and technical precedents of SSVA devices and their emergent aesthetic potential (both in sound and image).

2.2 Umwelt

Introduced by Baltic-German biologist Jakob von Uexküll in his *Foray into the Worlds of Animals and Humans* [20], the notion of *Umwelt* is defined as the local and subjective environment of an organism. This is a representation of an external, objective outside world - what Kant described as *Dass Ding an Sich* [21].

As Nöth [22] and Emmeche [23] articulate, parallels can be drawn between organic and synthetic Umwelt instances. As a biological organism is equipped with species-specific perceptual organs and receptors, a robot has perceptual modules to process environmental signals. This selective processing of stimuli (following sensor and algorithm selection) can lead to a subjective inner world, contrasting the more encompassing *Umgebung*. Depth input connects Vigil 17 to Uexküll's original work through his Umwelt analysis of a housefly [20] - long nerve structures ("rhabdomes") catching an image at varying depths, an organic counterpart to the MDE module.

Existing artistic practice has explored the phenomenology of other species—*In The Eyes of an Animal* [24] leverages Virtual Reality (VR) to immerse an audience in the sensory fields of four woodland species in Grizedale forest (UK). Vigil 17 seeks to advance practice-led exploration of Umwelt through the design and implementation of *synthetic* self-worlds, with embedded real-time responsiveness to environment.

2.3 Sound Art

Vigil 17 echoes Max Neuhaus' experiments in place [25], actualising "imaginary places" set covertly into public infrastructure. In *Walkthrough* [26], at the entrance to Jay Street Subway station (NYC), Neuhaus deployed 22 loudspeakers networked to a small computer, with pings and clicks responding to weather conditions. Commuters, who encountered the installation daily, were invited to observe aural changes to a familiar environment - for many, this went unnoticed.

More broadly, Neuhaus was interested in how sound can be defined in relation to space (rather than time). Vigil 17 gestures to these place works, embedding installation work directly into public spaces, with both covert design and attentiveness to context. Neuhaus' interplay with site is expanded - commuters become signal modulators, and the mediation process between the site itself and procedural audio approaches are defined more explicitly.

As depth is mapped to a distributed sound scape, the installation draws broadly upon Iannis Xenakis' stochastic approaches to music [27]. In his composition *GENDY13* [28] Xenakis formulates "dynamic stochastic synthesis". Polygonal representations of waveforms, defined through anchor points (x_i, y_i), are recursively generated and perturbed.

For polygonal coordinates at i, defined for two successive waveforms j and $j+1$, stochastic variation is added s.t.

$$\begin{aligned} x_{i,j+1} &= x_{i,j} + f_x(z) \\ y_{i,j+1} &= y_{i,j} + f_y(z) \end{aligned} \tag{1}$$

where $f_{x,y}$(z) are probability density functions, parameterised by latent variable z and sampled per-waverform cycle. Stochastic variation is introduced, simulating and sampling mathematical distributions mapped to polygonal coordinates — Thus, we see a *procedural approach* to wavetable synthesis.

Vigil 17 echoes this formulation, transplanting stochastic processes $f(z)$ into the aleatory of the *real world* — depth contours mapped to wavetable anchor points (x_i, y_i). High-dimensional polygons are generated in real-time (enabled by technical advances). In doing so, Vigil 17 seeks to draw connections between Xenakis' stochastic approaches to soundscape, and Neuhaus' studies of site as a *medium* itself.

3. METHODOLOGY

3.1 Depth Sonification Approaches

This phase began with exploring depth map acquisition, auditioning common MDE models, their performance on the Raspberry Pi 5 and aesthetic properties of the depth matrices. Notably, Depth Anything V2 [29] and ZoeDepth [15] provide state of the art inference. Unfortunately, when deployed on the Raspberry Pi 5 they are limited to $\sim$ 1 to 2 fps. This is too slow for continuous wavetable modulation, and smooth visual output (presented in a media-arts context). In contrast, SC-DepthV3 [19] is designed for edge deployment with an optimised build for the Hailo AI accelerator hat, achieving up to 30 fps. Imperfect depth predictions (and resultant distortion of image) were instead embraced as aesthetic artifacts. Some dummy footage, taken from the site, was fed through our depth processing pipeline and used to iterate on depth sonification approaches, with the following schema.

- Mapping depth values directly to pitch (closer objects $\rightarrow$ higher frequencies). A scan across the depth image (left $\rightarrow$ right) is implemented, playing tones sequentially. Subtle stereo modulation is used to indicate lateral position. Gestures to existing SSVA approaches [7–9, 13], with the mobility-relevant property of "hearing shapes".
- Texture-based sonification, generating micro-sounds using granular synthesis. A depth frame is divided into square patches, where each generates cluster of grains. The average depth of each region can modulate both amplitude and density of grain emission. Amplitude modulation approaches cite SSVA precedents [11, 13].
- Spatial audio mapping, in which the depth map is converted into a 3D auditory scene. The (x, y, z)-axes mapped to respective ambisonic coordinates (where z is depth). Each depth pixel becomes a unique sound source in space. This felt like the most "real-world" mapping schema with precedents in SSVA research [10–13].
- Edge-based sonification, using gradient analysis to extract depth boundaries. Here, column-wise local depth maxima are calculated, defining a contour function that falls into the steepest ridges (traced from left $\rightarrow$ right). This trace is mapped to a wavetable buffer, modulating the timbre of the soundscape.

Conventional SSVA mappings are largely defined by behavioural heuristics — pitch encoding distance, amplitude signalling proximity, stereo position indicating laterality. Despite the salience of depth boundaries as indicators of spatial structure and motion, gradient-based approaches have remained underexplored in the literature. Therefore, edge analysis is adopted as the primary sonification approach, tracing contours in depth space and mapping them to wavetable geometry. It was found this mapping approach leads to emergent string/flute-like characteristics, which flutter as the depth map moves. The wavetable output is spectrally shaped to foreground melodic frequencies, and passed through a reverberation algorithm, establishing a sustained ambient bed whilst preserving the evolving timbral character of the depth contour signal. For specific technical implementation please see section 4.1.

As a secondary layer, samples collected from the adjacent Blackfriars train station are swept across the stereo field, tracking regions of high depth variation. This affords a more intuitive and perceptually legible mapping between depth signal and sonic output. Through the use of field recordings, the work is further grounded in site.

3.2 System Prototype

Following selection and embedding of the MDE and sonification approaches, initial prototyping was conducted in a controlled environment to evaluate system response to live camera input. This was explored across varying lighting conditions and movement speed, informing iterative refinement of the sonification schema, including the weighting of the depth contour function's contribution to the final wavetable. A sample-based layer was introduced to reinforce the perceptual relationship between depth signal and sonic output, as at some points this became difficult to gauge. Hardware configuration was finalised during this stage, including the physical organisation of the modules, informing the subsequent fabrication process.

Ahead of site-deployment some basic performance benchmarking was performed, to ensure the feasibility of the installation. Frame rate moved between 25 and 30 fps, and battery life (from a full charge) was around 1.5 hours — sufficient for this stage of deployment. For specific hardware and system architecture, please see section 4.2.

3.3 Site Experiments

Three sites across Blackfriar's bridge (London, UK) were short-listed for installation site, balancing optimum foot traffic, field of vision and attractive structural/material properties of host. These included a roadside barrier and a mossy crevice in a damaged section of the bridge, displayed (location) as yellow crosses in Fig. 1.

Further exploration with installation sketches and dummy footage (mapped to sonification) led Vigil 17 as primary site — an electrical box with potential for magnet-based deployment. Vigil 17 was then planted, tuning the audio output on-site with the depth signal processing pipeline and tweaks to the SuperCollider patch.

3.4 Fabrication Procedure

The enclosure was designed using generative CAD tooling, producing an organic, skeletal structure printed in translucent filament. Magnetic mounts, integrated into the structure, ensure non-invasive attachment to the host site. The form functions as a load-bearing exoskeleton, housing the Raspberry Pi stack as a main body with an arm extruding the camera away from the body, gesturing toward the division of cognitive and perceptual apparatus found in biological systems. Due to the size of the Raspberry Pi stack,

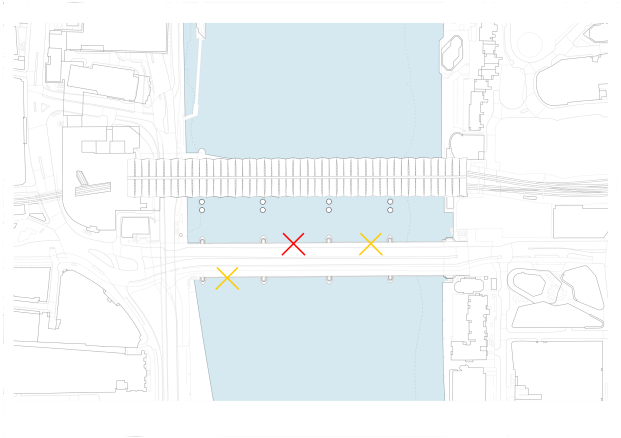


Figure 1: Bird's-eye view of three proposed sites along Blackfriar's bridge, with "Vigil 17" site shown in red cross [30].

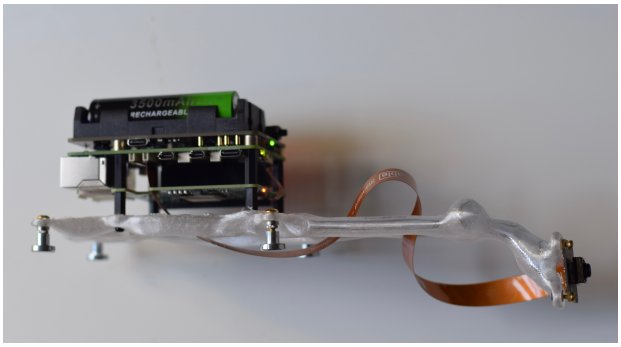


Figure 2: Assembled fabrication setup, integrating magnetic stands and skeletal 3D printed CAD design.

audio is broadcast via WebSockets rather than a DAC hat and speaker, avoiding additional weight and bulk.

4. IMPLEMENTATION

4.1 Depth Signal Processing

To translate the depth signal into meaningful wavetable changes (sonically), a *contour* function f is posited, tracing local gradient maxima across the depth map using gradient analysis. A pre- and post-processing pipeline are implemented, stabilising the output signal and extracting salient variation in the wavetable mapping.

The SC-DepthV3 [19] MDE algorithm is used to extract a depth matrix $D_t \in \mathbb{R}^{m\times n}$ at time t, from the site environment. Foreground extraction is applied, isolating the subjects (moving signal modulators) from the background B_t (ground, sky, bridge). This is adaptive, applying an exponential moving average per-pixel,

$$B_t(x,y) = (1-\alpha)\, B_{t-1}(x,y) + \alpha\, D_t(x,y), \quad (2)$$

where $\alpha = 0.05$ controls the accumulation rate. The foreground residual $F_t = |D_t - B_t|$ is then median-filtered to suppress isolated noise.

A depth gradient map is then extracted using Sobel operators, followed by spatial smoothing using a Gaussian kernel (5×5). This gives the smoothed depth gradient matrix $G_t \in \mathbb{R}^{m\times n}$, defined at time t, A *contour* function $f_t(x)$ that follows local gradient maxima can be posited.

The trace falls into the *steepest* ridges in depth space, moving horizontally. A column-wise maximum operator g_t is defined s.t.

$$g_t(x) = \max_{y=1,\dots,m} G_t(x,y) \quad (3)$$

The contour $f_t(x)$ carries this value forward when the column maximum exceeds a site-calibrated threshold τ, and holds its previous value otherwise s.t.

$$f_t(x) = \begin{cases} g_t(x), & \text{if } \max_y G_t(x,y) \geq \tau, \\ f_t(x-1), & \text{otherwise.} \end{cases} \quad (4)$$

Thresholding isolates salient changes in the depth map. Here, both inter-frame (temporal) and inter-column (spatial) smoothing are applied. Spatial smoothing is applied to f_t using a uniform kernel of width $k = 15$, yielding $\tilde{f}_t$. Temporal smoothing is implemented by an inter-frame exponential moving average,

$$\hat{f}_t = (1-\beta)\, \hat{f}_{t-1} + \beta\, \tilde{f}_t, \quad (5)$$

where $\beta = 0.3$. The stabilised contour $\hat{f}_t$ is then superposed with a sinusoidal base,

$$s(x) = \frac{m}{2} + A\sin\left(\frac{2\pi\nu x}{n}\right), \quad (6)$$

where A is the amplitude in pixels and ν the number of cycles across the frame width. A Tukey window $w(x;\alpha_w)$ tapers the mixing weight towards frame edges, giving the displaced contour

$$d_t(x) = \big(1 - w(x)\big)\, s(x) \;+\; w(x)\, \hat{f}_t(x). \quad (7)$$

Finally, d_t is normalised to $[-1,1]$ and transmitted via OSC as a wavetable frame to the synthesis engine. A still frame is shown in Fig. 3. Full implementation is available at the project repository. [1]

4.2 System Architecture

The Vigil 17 system is embedded on a Raspberry Pi 5, with a 2.4 GHz quad-core 64-bit Arm Cortex-A76 CPU and 8 GB of RAM. A fan was installed, preventing thermal throttling experienced at the higher frame rates. The 13 TOPS variant of the HAILO AI HAT+ NPU [2] is attached, accelerating inference modules and opening up the possibility of a *local* implementation, handling the pipeline in real-time. The X1200 module (UPS expansion board), using two 18650 cells, was selected as power source.

The Raspberry Pi Camera Module 3 captures the environment image signal. After on-device MDE extraction, the depth signal is fed into the processing and sonification pipeline. A full-system pipeline is shown in Fig. 4.

[1] https://github.com/danhearn/vigil-17-app

[2] https://hailo.ai/products/ai-accelerators/hailo-8l-ai-accelerator-for-ai-light-applications



Figure 3: Visual overview of contour signal extraction pipeline, beginning with RGB values with MDE module to extract depth map, with contour signal overlaid in red.

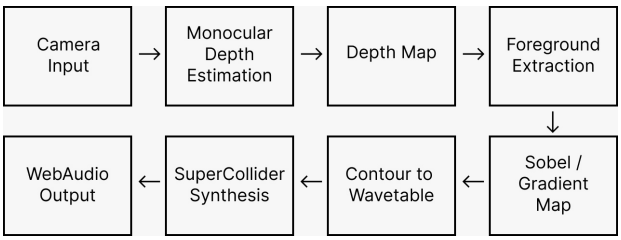


Figure 4: Full system pipeline, embedded and running in real-time within Vigil 17.

4.3 Technologies

A bespoke Vigil 17 app was developed. A Python back end computes the SC-DepthV3 [19] MDE algorithm, implementing pre- and post-procesing and extracting the contour function signal. The output is streamed as a real-time wavetable signal to SuperCollider via OSC, where it is inscribed to a buffer.

The raw depth matrix is streamed to a Next.js front end via WebSocket [3], where each frame is rendered to an HTML5 canvas element as a full-screen, greyscale image, providing a real-time visualisation of the depth signal. The SuperCollider audio output is simultaneously captured from the device audio interface and streamed to the same front end as a continuous binary audio stream over a second WebSocket connection. The browser reconstructs and plays back the audio in real time using the Web Audio API, allowing the sonification and its corresponding depth visualisation to be experienced together in the browser. The application is deployed to Fly.io, ensuring a persistent, publicly accessible connection between the on-device Python producer and the browser client.

4.4 Sonification System

Rather than direct mapping to wavetable buffers, the depth contour function delineated in section 4.1 is used to displace a sine oscillator. This is implemented through additive synthesis, resulting in the superposition of the base sine wave and the depth contour signal. This is weighted with a mixing coefficient weighting respective amplitudes. A snapshot of this additive process (for a single frame) is shown in Fig. 5.

This signal is sent (via OSC) from the Python depth processing module to the SuperCollider [4] (SC) programming language for audio synthesis. Here, incoming arrays are shaped into wavetables using envelope interpolation, converted to SC internal wavetable format, and written into a pre-allocated ring of 32 consecutively-addressed buffers.

Playback uses the SC `VOsc` object, a wavetable oscillator that performs continuous band-limited interpolation between adjacent buffers in a consecutive block. Eight detuned instances, spread across the stereo field, read from the buffer ring at positions modulated by independent noise generators, producing a slowly fluttering spectral texture. To prevent write-read collisions at high update rates, SC reports the mean read position back to the language layer at 20 Hz — incoming wavetables are written only to slots outside a safety margin of four buffer positions from the current read head, ensuring no buffer is overwritten while actively being read.

The wavetable output was further shaped through EQ gain across the C3 to C6 frequency range, with reverb applied to extend the spectral decay — setting a melodic character and establishing a sustained ambient bed beneath the evolving timbral texture. This is implemented as four chained `BPeakEQ` filters at C3, C4, C5, and C6, followed by `FreeVerb2`.

Samples (collected from Blackfriars train station) are triggered by global depth gradient threshold, with local gradient-maxima modulating stereo position. Here, a passing signal modulator (a pedestrian or vehicle traversing the field of view) crosses the threshold, with their lateral position within the frame mapped to the stereo distribution of the triggered release.

5. RESULTS

5.1 Deployment

The system was deployed on 13 March 2026 around 14:00, for approximately 1 hour and 15 minutes. This occurred at the Vigil 17 site, attached to the electrical box as is shown in Fig. 6. After this duration the battery died, giving a slightly lower life span than in our prototyping phase. Frame rate stabilised around 27 fps, aligning with initial experiments. The lighting was slightly brighter than initial testing conditions, but this did not interfere with the depth estimation signal quality.

[3] https://developer.mozilla.org/en-US/docs/Web/API/WebSockets_API

[4] https://supercollider.github.io/

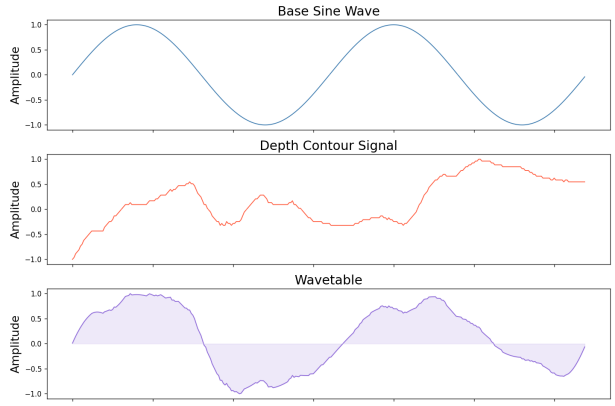


Figure 5: Superposition of base sine wave and depth contour signal by additive synthesis, resulting in final wavetable.



Figure 6: Vigil 17 system installed on electrical box, mapping depth signals to soundscape timbre in real-time.

Passer-by engagement was limited, with occasional interest dissipated upon brief explanation of the project. The installation operated without disruption to the surrounding environment, remaining largely anonymous within the flow of pedestrian activity. Through a phone hotspot, the WebSocket solution was able to stream depth visualisation and audio, allowing observation in real-time.

5.2 Output

A system output sample, comprising depth map visualisation alongside the emergent soundscape, is presented at the following link: `https://www.youtube.com/watch?v=Ik0UL6j7C2Y`. A representative three-minute excerpt has been selected and slowed (maintaining mapping relationships) for clarity of observation.

6. DISCUSSION

The system demonstrates the viability of embedded, real-time depth-to-audio systems, leveraging SC-DepthV3 [19] on a Raspberry Pi 5 with Hailo NPU acceleration. Through advancing this, a practical pathway for modern MDE-based SSVA devices is suggested. The gradient-based contour extraction approach represents a departure from conventional SSVA mappings, which typically operate on raw depth values modulated to pitch, amplitude or spatial position.

However, a tension emerges between the system's media-arts framing and its assistive potential. Imperfect depth predictions from the MDE module were embraced as aesthetic artefacts within a media-arts context, leading to both visual and sonic distortion. In an SSVA context, this would constitute signal degradation, potentially compromising spatial legibility for a visually impaired user. Future work might resolve this tension by isolating embedded DL-based MDE approaches, evaluating within a dedicated assistive technology context, where depth accuracy and perceptual clarity are the primary constraints.

Treating depth as compositional substrate, modulating the timbral evolution with gradient-based depth analysis, yielded genuinely interesting textural variation, with emergent string-like qualities arising from the continuous displacement of the wavetable geometry. However, the perceptual legibility of the depth-to-sound mapping is blurry. This points to a fundamental distinction between Xenakis' algorithmic stochastic processes, where probability distributions are explicitly bounded and parametrised, and the aleatory of real-world depth signals, where variation is unbounded and resistant to clean delineation in distribution space. This suggests further exploration of alternative mapping schemas: texture-based granular approaches and ambisonic spatial mapping, explored in early prototyping, may afford more semantically salient correspondences between depth structure and auditory space.

For the authors, the convergence of visual and sonic output suggests a system operating beyond a deterministic depth pipeline — one that enacts a form of synthetic *Umwelt*, selectively mediating environmental perception through the constraints of sensor, algorithm, and synthesis architecture. This seems resonant with with Uexküll's notion of subjective perceptual worlds nested within a shared space. At the level of fabrication, skeletal CAD geometry and organic material choices reinforce this framing. It could be interesting to explore bi-directionality in more depth — the degree to which the soundscape might actively shape movement through space, rather than merely responding to it. Such an extension would deepen the entanglement between synthetic and organic perceptual fields, and raise broader questions around how the deliberate authorship of synthetic self-worlds shape public encounter with automata.

7. CONCLUSION

Vigil 17 presents a framework for translating monocular depth signals into an evolving procedural soundscape, demonstrating the viability of embedded, real-time MDE-based audio mapping on consumer hardware. A novel gradient-based contour extraction approach is introduced, departing from conventional SSVA mappings and foregrounding timbral variation as a carrier of spatial information. Deployed publicly on Blackfriars Bridge, the system engages materially with its site — pedestrians and vehicles functioning as signal modulators within a broader ecology.

In doing so, Vigil 17 advances practice-led exploration of synthetic self-worlds, situating depth as a compositional substrate within these systems. It is hoped that the technical framework developed here, demonstrating embedded MDE-based sonification outside of controlled laboratory conditions, may inform future development of assistive SSVA devices, alongside continued inquiry into depth as a medium for procedural sound-art practice.

Acknowledgments

We would like to thank Saina Akhond for her guidance, helping us to shape our thoughts and research process, from somewhat scattered starting points. For open access to his lab throughout the fabrication process, we thank our good friend Ruiyang Wang.

8. REFERENCES

[1] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt and Co., Inc., 1982.

[2] D. R. Griffin, *Listening in the Dark: The Acoustic Orientation of Bats and Men*. Oxford, England: Yale University Press, 1958.

[3] I. P. Howard and B. J. Rogers, *Binocular Vision and Stereopsis*. Oxford University Press, 02 1996. [Online]. Available: https://doi.org/10.1093/acprof:oso/9780195084764.001.0001

[4] C. Veraart and M.-C. Wanet, *Sensory Substitution of Vision by Audition*. Dordrecht: Springer Netherlands, 1985, pp. 217–238. [Online]. Available: https://doi.org/10.1007/978-94-017-1400-6_15

[5] L. Kay, "An ultrasonic sensing probe as a mobility aid for the blind," *Ultrasonics*, vol. 2, no. 2, pp. 53–59, 1964. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0041624X64903828

[6] L.Kay, *Air Sonars with Acoustical Display of Spatial Information*, R.-G. Busnel and J. F. Fish, Eds. Boston, MA: Springer US, 1980. [Online]. Available: https://doi.org/10.1007/978-1-4684-7254-7_35

[7] P. Meijer, "An experimental system for auditory image representations," *Biomedical Engineering, IEEE Transactions on*, vol. 39, pp. 112 – 121, 03 1992.

[8] J. Cronly–Dillon, K. C. Persaud, and R. Blore, "Blind subjects construct conscious mental images of visual scenes encoded in musical form," *Proceedings of the Royal Society B: Biological Sciences*, vol. 267, no. 1458, pp. 2231–2238, 11 2000. [Online]. Available: https://doi.org/10.1098/rspb.2000.1273

[9] P. Arno, C. Capelle, M.-C. Wanet-Defalque, M. Catalan-Ahumada, and C. Veraart, "Auditory coding of visual patterns for the blind," *Perception*, vol. 28, no. 8, pp. 1013–1029, 1999, pMID: 10664751. [Online]. Available: https://doi.org/10.1068/p281013

[10] F. Ribeiro, D. Florêncio, P. A. Chou, and Z. Zhang, "Auditory augmented reality: Object sonification for the visually impaired," in *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, 2012, pp. 319–324.

[11] M. Brock and P. O. Kristensson, "Supporting blind navigation using depth sensing and sonification," in *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, ser. UbiComp '13 Adjunct. New York, NY, USA: Association for Computing Machinery, 2013, p. 255–258. [Online]. Available: https://doi.org/10.1145/2494091.2494173

[12] D. Massiceti, S. L. Hicks, and J. J. van Rheede, "Stereosonic vision: Exploring visual-to-auditory sensory substitution mappings in an immersive virtual reality navigation paradigm," *PLOS ONE*, vol. 13, no. 7, pp. 1–32, 07 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0199389

[13] C. Stoll, R. Palluel-Germain, V. Fristot, D. Pellerin, D. Alleysson, and C. Graff, "Navigating from a depth image converted into sound," *Applied Bionics and Biomechanics*, vol. 2015, pp. 1–9, 02 2015.

[14] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 2366–2374.

[15] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," 2023. [Online]. Available: https://arxiv.org/abs/2302.12288

[16] R. Ranftl, K. Lasinger, D. Hafner, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, pp. 1–1, 08 2020.

[17] R. Garg, V. K. B.G., G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 740–756.

[18] T. Zhou, M. A. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6619, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:11977588

[19] L. Sun, J.-W. Bian, H. Zhan, W. Yin, I. Reid, and C. Shen, "Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46,

no. 1, p. 497–508, Jan. 2024. [Online]. Available: https://doi.org/10.1109/TPAMI.2023.3322549

[20] J. von Uexküll, *Streifzüge durch die Umwelten von Tieren und Menschen: Ein Bilderbuch unsichtbarer Welten*, ser. Verständliche Wissenschaft. Berlin: Verlag von Julius Springer, 1934, vol. 21.

[21] I. Kant, *Critique of Pure Reason*, J. M. D. Meiklejohn, Ed. Cambridge: Cambridge University Press, 1929.

[22] W. Nöth, "Semiosis and the umwelt of a robot," *Semiotica*, vol. 2001, pp. 695–699, 01 2001.

[23] C. Emmeche, "Does a robot have an umwelt? - reflections on the qualitative biosemiotics of jakob von uexküll," *Semiotica*, vol. 134, 11 2001.

[24] Marshmallow Laser Feast. (2015) In the eyes of the animal. [Online]. Available: https://marshmallowlaserfeast.com/project/in-the-eyes-of-the-animal/

[25] J.-C. Ammann, "Plädoyer für eine neue kunst im öffentlichen raum," *Parkett*, no. 2, 1984.

[26] M. Neuhaus. (1973) Walkthrough. Sound installation, Jay Street Subway Station, New York, extant 1973–1977. [Online]. Available: https://www.max-neuhaus.estate/en/sound-works/vectors/place/walkthrough-jay-street-subway-station-new-york-1973

[27] I. Xenakis and S. Kanach, *Formalized Music : Thought and Mathematics in Composition*, revised edition / additional material compiled and edited by sharon kanach. ed., ser. Harmonologia series ; no.6. Stuyvesant, N.Y: Pendragon, 1992.

[28] M.-H. Serra, "Stochastic composition and stochastic timbre: Gendy3 by iannis xenakis," *Perspectives of New Music*, vol. 31, no. 1, pp. 236–257, 1993. [Online]. Available: http://www.jstor.org/stable/833052

[29] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 21 875–21 911. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/26cfdcd8fe6fd75cc53e92963a656c58-Paper-Conference.pdf

[30] A. M. K. Coonar, "Site plan for undergraduate final project with sonography concept," 2025, unpublished architectural drawing, BA Architecture, London South Bank University.