C. Çakmak and R. Hamilton, "od: Composing Spatial Multimedia for the Web"
J. Audio Eng. Soc., vol. 68, no. 10, pp. 747–755, (2020 October.).
DOI: https://doi.org/10.17743/jaes.2020.0017

od: Composing Spatial Multimedia for the Web

CEM ÇAKMAKAND ROB HAMILTON
(cakmao@rpi.edu) (hamilr4@rpi.edu)

Rensselaer Polytechnic Institute, New York, USA

Composers have been exploring multi-channel sound field spatialization since the early days of electronic music. However, reproduction of such works outside of specialized concert spaces and research facilities or even their accurate reproduction within those spaces remain difficult and unpredictable at best. Combining the reach and simplicity of web browsers with ambisonic to binaural rendering, Web Audio-based tools can ensure greater accessibility for existing spatial works as well as acting as a platform upon which new ones can be implemented. At times with such practices the developing technologies become deprecated or obsolete during the period of making the work. This paper describes the technical design and artistic conception of *od*, a spatial multimedia production for binaural listening on the web. The project has led us to develop a workflow without relying on specific tools that can be of use as a framework for documenting existing spatial works or novel browser-based creative applications.

0 INTRODUCTION

Spatial, or multi-channel, approaches in electronic music composition are as old as the practice itself. While many of such works furthered experimentation in sound, light, and multimedia, they also surrounded and sought to immerse their audiences in unique spatiotemporal experiences. Distinct as these performances were, today they are often talked about while only a handful of people have truly experienced them on site; the creation and presentation of spatial electronic music often relies on the support and patronage of large-scale research institutions due to their experimental nature and funding required. Issues of accessibility have maintained such works as niche and their recreations seldom feasible.

The challenges facing spatial music performance come in many different categories, such as aesthetic, structural, social-cultural, and most often technical. Often such work is composed in a space different than that used for the performance, meaning that the influences and characteristics of at least two spaces are superimposed during diffusion [1]. Variances in listener experience are due to factors that include audience seating, architecture, or external disruptions. Since such works often focus on timbre and space rather than pitch or rhythm, they are much less tolerant to disturbances, faulty equipment, and lower qualities of production [2]. Binaural technologies create a listening space that is more accurate, personal, and less prone to disruptions by replacing complex multi-channel systems with headphones for presentation, making spatial experiences more affordable and accessible in binaural listening mode.

This paper builds on the authors' recent artistic research [3] to propose a generalized workflow in order to schematize and classify the components needed to present spatial multimedia in web browsers as well as to better illustrate their relationships within a workflow in order to compensate for specific tools and technologies that become obsolete very quickly. In this manner the methods outlined in this workflow can remain valid for longer rather just describe a singular spatial multimedia project. The workflow can be applied as post-production to existing spatial works in order to document and present them online. Or as the second half of the paper will demonstrate with a recent work, *od*, the workflow can also be employed as a strategy for the production of new spatial multimedia.[1] In order to achieve this, the authors seek to bring a number of online and offline tools together in an artistically meaningful way. The main goals of the research are (1) propose a generalized workflow to achieve spatial multimedia production and post-production for the web, (2) apply the workflow to facilitate the composition and dissemination of a novel spatial music piece, (3) exercise binaural spatial listening on the browser, and (4) enable access to a physical site that is typically difficult to visit.

1 BACKGROUND

Beyond stereophony lies an infinite playground, one where loudspeakers work together not just to form a direc-

[1] https://cem.works/od

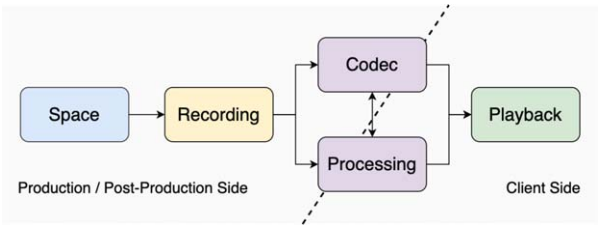

Fig. 1. A simple workflow for producing spatial multimedia. The scheme takes into account a site (Space), its audiovisual capture (Recording), preparation for the web with software and various file formats (Processing/Codec), and finally the presentation on the user side (Playback). In the cases where the audio or visuals are designed virtually, *Space* and *Recording* can be omitted or merged together with *Processing/Codec* as one component. Furthermore, *Processing/Codec* can take place on either production or client side or both, depending on project specifics.

tional image but also to both encompass the listeners and innovate new spatial complexities. Most of the theory behind spherical audio goes back to the 1970s, led by Gerzon's research and practice [4]. But while ambisonic systems aim to recreate a spherical soundscape as accurately as possible, composers also use unconventional speaker arrangements where speakers are treated as instruments that contribute to the music with their characteristics and form a relationship with their surroundings, at times with the inclusion of extra-musical elements. Some of these are GRM's Acousmonium, Xenakis' polytopes, or other custom structures with multi-channel speaker configurations like the EXPO '58 Philips and EXPO '70 Pepsi Pavilions [5–7]. Restaging such works however are often expensive undertakings with dubious fidelity to the original experiences [8].

In response to such a significant history of research and performance that can not be experienced again, we propose a simple workflow scheme to eliminate specifics and focus on various tasks brought together to create an accessible spatial experience. Illustrated in Fig. 1, the workflow begins with a space, as this is the origin of spatiality, and continues with stages of recording, processing, media conversion, and presentation. We investigate these stages in Sec. 2, and we demonstrate how this workflow can be applied to a specific spatial multimedia project for the web in SECS. 2.6 and 3. While the workflow does not stress specific tools—as these change regularly—it focuses on the use of ambisonics and Web Audio to deliver spatial multimedia online. Ambisonic recordings can capture any kind of sound field with directional information from a fixed location and have the advantage of reproducibility over varying loudspeaker configurations. For example, Kermit-Canfield proposes a configurable diffusion tool for building virtual acousmonia and encoding the output using ambisonics for changing speaker arrangements [9]. Gurevich et al. demonstrate networked ambisonic streaming and diffusion between distributed concert spaces [10]. But this superimposition of composed and listening spaces may lead to undesired results, especially of a finished work. Ambisonic to binaural is a useful conversion to document and present spatial works with a far more accessible experience on the web browser. Lastly, personalized head-related transfer functions (HRTFs) for improved binaural experiences are being implemented in Web Audio [11].

With the ongoing development of web audio technologies and tools, audience exposure to new spatial artworks can be made more accessible and less ephemeral. Sec. 2 of this paper describes the components of the proposed workflow and the tools we employed to compose novel spatial multimedia designed to be experienced and shared online. The workflow is itself not novel and may already be implicit in such projects but on an entry level to the practice, the manners in which such techniques and technologies combine is not readily apparent. And on the professional level of production such standard workflows remain a challenge as components change and deprecate at a rapid rate. Thus we believe that there is a need for generalized and methodological guidelines and suitable workflows for composing spatial multimedia and delivering through web interfaces.

2 WORKFLOW COMPONENTS AND TOOL SELECTION

This section will first detail the workflow with available tools and technologies and subsequently describe their use and selection for a specific project in line with the workflow. A comprehensive survey of each of the following subsections would extend far beyond the scope of this paper. Instead, we will describe their purpose in respect to the general workflow, focusing on web-based presentation. Following this we will detail what technologies we have employed during the design and creation of *od* according to the workflow diagram.

2.1 Spaces

The space is the original source of spatial information and could be a concert hall, outdoor site, virtual environment, or hybrid construction. Spaces designed specifically for audiovisual performances have benefits in terms of technical infrastructure, whereas public spaces are interesting for artistic presentation in that they range from natural sites to bustling cities, each with their own dynamic and unique visual and acoustic environments. Virtual spaces can be built from the ground up, facilitated through existing ones or augmented from site recordings. Educational spaces with technological infrastructure such as the planetarium are alternatives as well; the common architecture and infrastructure found across planetaria implies that when an experience is designed for one, it can work for others as well. Should a shared language form around it and some standards be established, the planetarium space may provide an accessible substitute for virtual reality for larger audiences [12].

2.2 Recording

To capture a spatial event in a physical space, specialized recording devices are necessary for both audio and video. Recording audiovisuals in 3D has come a long way and commercial products that serve such needs are becoming

increasingly available. Nevertheless, recording configurations differ for visual or audio recordings, so we will distinguish between the microphones and cameras.

2.2.1 Microphones

Gerzon's leading theoretical work and practical contribution with the soundfield microphone forms a foundation for ambisonic recording [13]. Today an individual has multiple options for 3D audio recording with a range of features and prices. Other 3D recording methods such as stereo-pairing can also be employed depending on the project [14]. Research comparing different 3D microphones in terms of timbral quality and localization accuracy is available [15]. While high-order ambisonic (HOA) recording is still comparably expensive to standard microphones, affordable first-order ambisonic (FOA) microphones are becoming available, even as smartphone accessories.[2] With ambisonic playback support over major Virtual Reality (VR) platforms, FOA is currently the most enabled ambisonic method for web-based presentation. On the other hand, FOA recordings lack sufficient clarity compared to higher orders. When recording ambisonics, raw audio is stored in A-format where each channel corresponds to one of the microphones. For playback, an FOA recording must be converted to B-format. HOA recordings on the other hand are commonly ordered and normalized for playback using the Ambisonic Channel Number (ACN) ordering. As an alternative specific sound sources can be recorded with non-3D microphones and the scene can be reconstructed later by joining them with their position data, but such object-based audio methods get less feasible with increasingly complex sound fields.

2.2.2 Cameras

With an expanding range of products, 360 or VR camera arrays are becoming more available and affordable for projects documenting spatial performances. Camera rigs vary significantly in design, in particular with regards to the positioning and orientation of the cameras, their choice of lenses, and even sensors. Most VR cameras record monoscopic images, which is the most widespread method, but for further degrees of freedom others can record stereoscopic images for binocular vision effects. Whichever camera rig is employed for the specific project, the recordings often need to be stitched together to produce the desired spherical video. The quality of the resultant 360 recording depends on both hardware and software capabilities of the product as well as the site conditions.

2.3 Codec

Navigating through various audio, visual, and multimedia environments, projects using spherical video are often faced with changing formats, the need for efficient compression, and the complexities offered by overlaying media. Types and methods vary depending on project specifics, particularly with the tools employed in recording and processing stages. Ambisonic data exchange formats vary widely and no standard method has been established due to the borrowing of the spherical harmonics formulations from other disciplines with different conventions. Ambisonics provides a decoupling of the codec where the generalized multi-channel format is applied in the encoding stage, but the playback configuration can be specific to the decoding stage. Ambisonic encoding and decoding stages have several ad-hoc formats and incompatible specifications: 3D microphones often come with proprietary software for encoding a given format. Open-source libraries for ambisonic data processing are available as well.[3]

2.4 Processing

The processing stage is where all media are imported, edited, composed, and exported for playback. As with recording the technicalities change significantly among auditory and visual methods; thus we will split the processing stage into audio and visual components.

2.4.1 Audio Processing

3D audio is receiving increasing attention partly due to its complimentary nature to VR projects. The challenges of spatial audio differ among projects depending on their needs; the accurate documentation of a real-world soundfield may require no more than a few touches on spectral and dynamic qualities in this stage. On the other hand, artificial environments with virtual sound sources and projections may be constructed without any real-world reference. Spatial audio tools for programming environments provide flexibility in configuration of the soundfield as well as sound design [16, 17]. Objects can be composed and redesigned for changing configurations or creative investigations. Despite digital-audio workstations (DAWs) being overwhelmingly concerned with stereo imaging, plugins that undertake ambisonics within DAWs provide another working environment for spatial audio production [18]. Furthermore, plugins that pipe data streams between DAWs and standalone 3D production software help further integrate DAW usage with spatial audio practices [19, 20].

2.4.2 Visual Processing

Visual intentions can vary from project to project; if the desired outcome is the accurate documentation of a physical space, this stage will mostly be concerned with stitching raw camera footage together and eliminating errors in continuity by calibration, exposure, or color adjustments. The software used to produce such video compositions often come with a given hardware product and are generally specific to the given camera array. Nevertheless, open source and camera-independent stitching solutions are also available.[4] On the other hand, if the intended result does not necessitate an accurate site recording, further production techniques can be employed for creative

[2] http://yun.twirlingvr.com/index.php/home/lite/lite-en.html

[3] https://github.com/udio

[4] https://github.com/stitchEm/stitchEm

outcomes; these can range from subtle stitch manipulations to video effects and animations. Various VR tools can be applied to further increase immersive or narrative qualities of the environment; such tools will be discussed in Sec. 3.2.

2.5 Playback

The playback component is where users access the spatial content using a range of devices. We are concerned with creating or recreating spatial multimedia that is accessible to audiences without undue physical restrictions or state-of-the-art tools. Thus we will consider the internet browser and headphones as the primary playback environment. Browsers are available across many devices with internet connections, such as computers, tablets, mobile devices, head-mounted displays, and smart-televisions. Stereo headphones are already in widespread use and are a favorable solution for distributed spatial playback. The ease of access for binaural renditions using browsers and headphones both allows for faithful recreation and documentation of an existing work as well as an expansion of the audience for spatial multimedia practices.

The Web Audio API provides three audio nodes, or signal processing blocks, that are utilized in ambisonic processing: *Gain Node*, *Convolver Node*, and *Channel Splitter/Merger Nodes*. *Gain Node* provides user-controlled signal multiplication; a combination of these handle the value stream on runtime efficiently. *Convolver Node* is used in the binaural decoding stage for linear convolution with user-specified filters. And *Channel Splitter/Merger Node* combines and separates channels when sending and processing between nodes. HRTFs are necessary for binaural reproduction of the soundfield; non-individualized or dummy-head HRTFs are typically used in projects, but personalized HRTFs pulled from a server database may also be available, though these have challenges of implementation in Web Audio [21–23]. Direct binaural synthesis within Web Audio is also possible and can potentially increase spatial resolution and accuracy. This method may skip the ambisonic processes for certain projects, for example in the case of a navigable space. But direct synthesis can also cause a setback if too many source files have to be loaded on the user end; the total file size to load could exceed beyond even an HOA file and increase CPU load significantly. Streaming sources from a server or providing custom synthesis scripts through *AudioWorklet* nodes are developer-oriented alternatives that do not yet offer a high-level composition environment for composers.

The WebVR API allows VR production on the browser and can be used to wrap and play back 360 video recordings as well. WebVR provides a platform to develop VR experiences using HTML5, CSS3, and JavaScript.[5] Various WebVR frameworks can abstract the complex scripting related to webGL and JavaScript and enable VR production with HTML5 code.

[5] https://www.w3.org/TR/webxr/

2.6 Selected Tools

We have discussed the workflow to consider when documenting spatial works. This general framework is now applied toward a spatial multimedia project with a number of changes and expansions to it. The following subsections will discuss the tools used specifically for the composition and development of the project, *od*.

2.6.1 Space—CRAIVE-Lab

The Collaborative-Research Augmented Immersive Virtual Environment Laboratory, or CRAIVE-Lab, is a state-of-the-art interactive immersive environment facility operated by Rensselaer Polytechnic Institute in Troy, NY [24]. The lab features an immersive surround video screen shaped in a rounded rectangle, with combined resolution of 15,360 x 1,200 pixels. The panoramic image is front-projected by eight short-throw projectors mounted on the grid above the lab. These projectors are calibrated and warped to create a smooth continuous image along the entire projection surface, with special consideration given to the rounded corners. The unique projection screen was used to play an original panoramic video composition to be recorded. Behind the full length of the screen is a 128-channel speaker array with additional ones hung from the above grid for HOA support; it should be noted that the on-site sound setup was not employed for our project due to the recording setup to be discussed. In the context of the project *od*, the goal was to export the CRAIVE-Lab itself from its real-world location and enable audiences to virtually experience works made within it, as opposed to its main function, where audiences walk into transplanted or hyper-real sites.

2.6.2 Recording—GoPro Omni

GoPro's Omni was our 3D camera of choice due to its availability. The multi-camera rig used for *od* contains six GoPro Hero 4 cameras on each face of its cube-shaped design in order to record spherical scenes. By uploading a special firmware provided by the manufacturer, all cameras can be synced and controlled through a single remote control. Each camera records in 4k resolution and 4:3 aspect ratio; these are calibrated and stitched together using proprietary software. As mentioned, *od* does not incorporate live 3D HOA recording for the project due to budget limitations and the pursuit of virtual ambisonic production with the tools described in SEC. 2.6.4. But if audio recording was required in this case, the CRAIVE speaker system would be employed and the microphone would be placed right above or below the camera rig as close to the center as possible and its image removed in the visual processing stage. The main strength of recording HOA is that any spatial diffusion method can be captured from a steady point of observation such as acousmonium, vector base amplitude panning (VBAP), or wavefield systems.

2.6.3 Codec—FFmpeg

FFmpeg is a complete, cross-platform command-line tool capable of recording, converting, compressing, and streaming audio and video in a wide range of current and ob-

solete digital formats. This tool will provide for virtually all encoding and decoding needs for any audio or video aside from ambisonics; encoding ambisonics for the project used the SPAT tools discussed below. Although FFmpeg cannot perform ambisonic encoding, it does offer multi-channel audio file compression useful to decrease large HOA file sizes and load more efficiently.

2.6.4 Audio Processing—Max/MSP and SPAT

IRCAM's SPAT is a library of Max/MSP objects written in C++ capable of spatializing sound in real-time for ambisonics, binaural synthesis, and artificial reverberations [25]. For ambisonic composition and binaural monitoring SPAT has proved to be a productive setup for musical exploration; without a 3D recording these tools allow composers and engineers to create interesting spatial soundfields and output using HOA. SPAT offers a dynamic 3D environment for organizing and manipulating sound sources, in addition to modeling loudspeaker arrangements for real-time synthesis and diffusion. SPAT can be used for live performances, mixing, post-production, installations, VR, and other applications.

2.6.5 Visual Processing—Autopano and MantraVR

To stitch footage together to construct spherical images, we used GoPro Omni's proprietary software: Kolor's AutoPano Giga 4.4 and AutoPano Video Pro 2.6. Unfortunately in September 2018, the week after the site recordings, all Kolor software support for the Omni camera ended, which led us to continue the project with deprecated software. Nevertheless, the stitched footage was exported and further processed in order to augment the spatial qualities of the space with MantraVR, an Adobe After Effects plugin for VR content production.[6]

2.6.6 Audio Playback—Omnitone

Omnitone was our choice for ambisonic-to-binaural rendering due to its availability, ease of use, and how it fits into the workflow. Written with the Web Audio API, Omnitone is a JavaScript implementation for ambisonic decoding and binaural rendering on the web browser for ambisonic streams up to third order.[7] Illustrated in Fig. 2, multi-channel files or buffers are streamed via *AudioBufferSourceNode* and ordered in ACN format. Listener head rotation, tracked using user HID control interaction or head-mounted display-based sensors, is stored and updated in a rotation matrix. Binaural rendering is handled by *ConvolverNode* and *GainNode* interfaces native to Web Audio. In line with Google's spatial media specifications, the decoded ambisonic signals for each ear pass through SADIE HRTFs to simulate binaural hearing.[8]

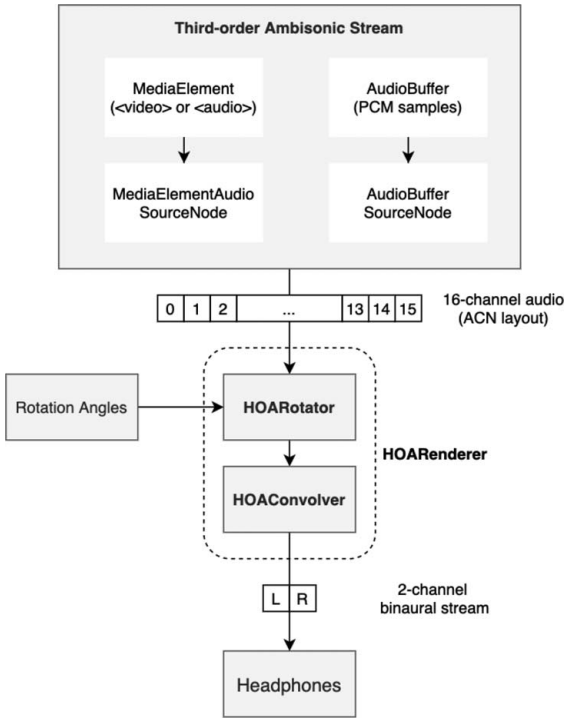

Fig. 2. Omnitone's third-order ambisonic to binaural rendering diagram. *HOARenderer* is linked to a multichannel *SourceNode* and with the speed of native Web Audio features decoded into binaural with rotation data updates. Image adapted from Google Chrome Omnitone GitHub repository, https://github.com/GoogleChrome/omnitone.

2.6.7 Visual Playback—A-Frame

The A-Frame library is an open source, three.js framework for WebVR that enables the writing of VR applications with HTML5.[9] Developed in order to make VR development more accessible, A-Frame provides higher-level coding within a special HTML tag, *a-scene*, to implement and manipulate VR content without having to manage complex WebGL code. A-Frame projects work across a range of mobile, desktop, and head-mounted devices [26]. At this time we anticipate the use of VR headsets primarily with a fixed user-location (3 DOF) as opposed to VR environments allowing user translation and motion with full degrees of freedom. VR headsets provide an opportunity to present HOA applications as they provide a static observer-centered spherical experience, coinciding in their three degrees of freedom.

3 WORKFLOW APPLICATION

During the creation of *od* we applied this workflow as a creative guide for independent production. The composition of *od* involved navigating through different stages in the workflow with the tools described in Sec. 2.6. This task was split into auditory and visual stages due to the distinct media formats used in each. As illustrated in Fig. 3, the composition and revisions of the audio and visual content

[6] https://www.mettle.com/product/mantra-vr/
[7] https://googlechrome.github.io/omnitone/
[8] https://www.york.ac.uk/sadie-project/GoogleVRSADIE.html
[9] https://aframe.io/

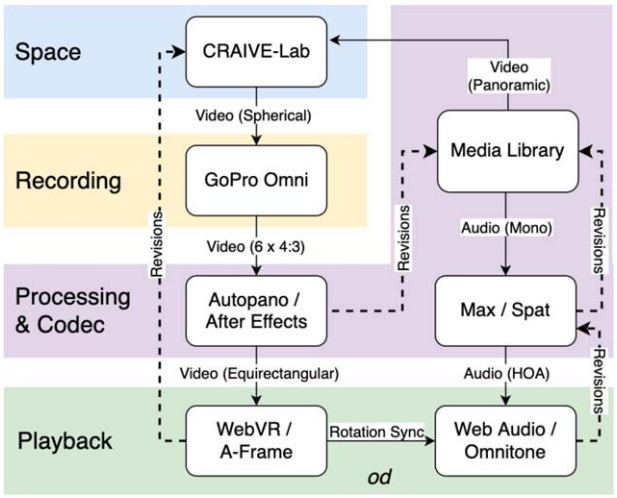

Fig. 3. Production workflow for *od*: the generalized components are detailed, expanded, or reduced in line with project specifics. Since the audio is synthesized rather than captured, Space and Recording components are omitted on the audio side of the production. Technical or aesthetic outcomes of certain stages inform previous components for revisions, indicated by the dashed arrows, and feed back into the workflow.

were mostly discrete from each other, joined and synced together in the playback stage between Omnitone and A-Frame.

3.1 Auditory Design

The central musical idea for *od* came from a stereo drone installation by La Monte Young, *Dream House*, exhibited in an apartment located in downtown Manhattan [27]. The timbre quality of Young's drone is strictly related to the listener's body positioning and movement; it remains static as long as the individual stands still, but certain partials diminish and new ones emerge when position is changed, activating timbre through motion and the superimposition of prime frequencies. The main spatial strategy was thus to compose a spherical drone in a virtual space that activates and fluctuates in timbre when the users change their rotational position. In the context of *od*, Max/MSP with SPAT externals offered an audio workflow embedded in the general workflow wherein (1) an ambisonic scene with spatialized sound sources is specified in SPAT; (2) the scene is encoded in 3rd order HOA, following the ACN/SN3D scheme for compatibility with Omnitone; (3) the ambisonic stream is then recorded into a 16-channel wave file and simultaneously transcoded to binaural for headphone monitoring; and (4) the scene is transformed in yaw, pitch, and roll and monitored binaurally within SPAT.

The ambisonic scene consists of 20 virtual sound sources arranged as the vertices of a regular dodecahedron around the listener, as seen in Fig. 4. Instead of utilizing prime frequencies as in Young's piece, *od* employed three spectral clusters of randomized low, mid, and high frequencies around 150, 1,500, and 6,000 Hz. Although initial tests featured pure tones, triangle waves ultimately served as more suitable sound sources, since they contain overtones

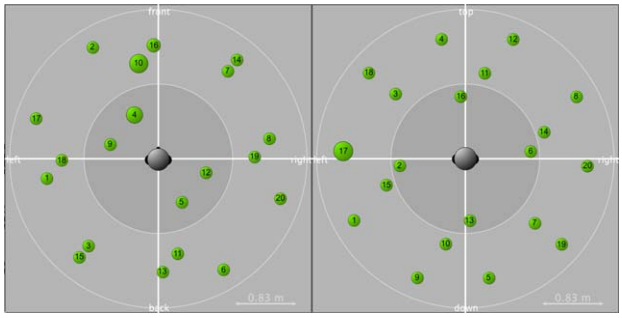

Fig. 4. Snapshot from the spat5.viewer interface with 20 sound sources distributed in a dodecahedral pattern around the listener. The composition intention is to create pitch and phase differences between the sources that will contribute to timbral variations when the head/camera is rotated.

that enhanced the timbral quality without weighing down the fundamentals as much as sawtooth or square waves. Sampling audio files with pitch shifts spread across the dodecahedron produced more interesting timbral and spatial qualities and is implemented in the current audio version.

3.2 Visual Design

As mentioned in Sec. 2.1, the goal of the visual design is to render the space and experience virtually accessible through the browser. The camera array was placed at the center of the room in three axes, or at the visual "sweet-spot." Recordings were then stitched together to construct an 8,000 x 4,000 pixel equirectangular image from the center of CRAIVE-Lab. The recording took place while a pre-composed high definition panoramic video played on the projection screen. Extending from one of the author's previous acousmonium performances, the composed video consisted of concert footage stretched, reproduced, and manipulated for the 15,360 x 1,200 screen.

On the horizontal axis of the 360 footage, the screen is continuous apart from the room entry but large areas remained blank on the vertical axis. In order to augment the immersive qualities of the room, we strategically placed virtual mirrors on the spherical image, as seen in Fig. 5, to expand the projected image in all directions using effects plugins described in Sec. 2.6.5. In order to ensure a smooth spherical scene, we applied feathering while stitching the raw footage together. Along with accurately stitched videos, unorthodox methods during the workflow produced time-offset versions that play with differences up to a few seconds in order to create additional movement. Overlapping of the various time periods act as bizarre temporal fractures that appear and disappear spatially. Accidental clusters of color streaks in this augmented space create contrasting elements and emerging kaleidoscopic patterns. With audio-reactive features of MantraVR, we further manipulated certain parameters with a stereo mixdown of the composed audio where the artificial patterns continue to revolve in response to the audio dynamics. The final, augmented spherical video was converted to .mp4 and specified as an A-Frame asset to be wrapped for online playback.

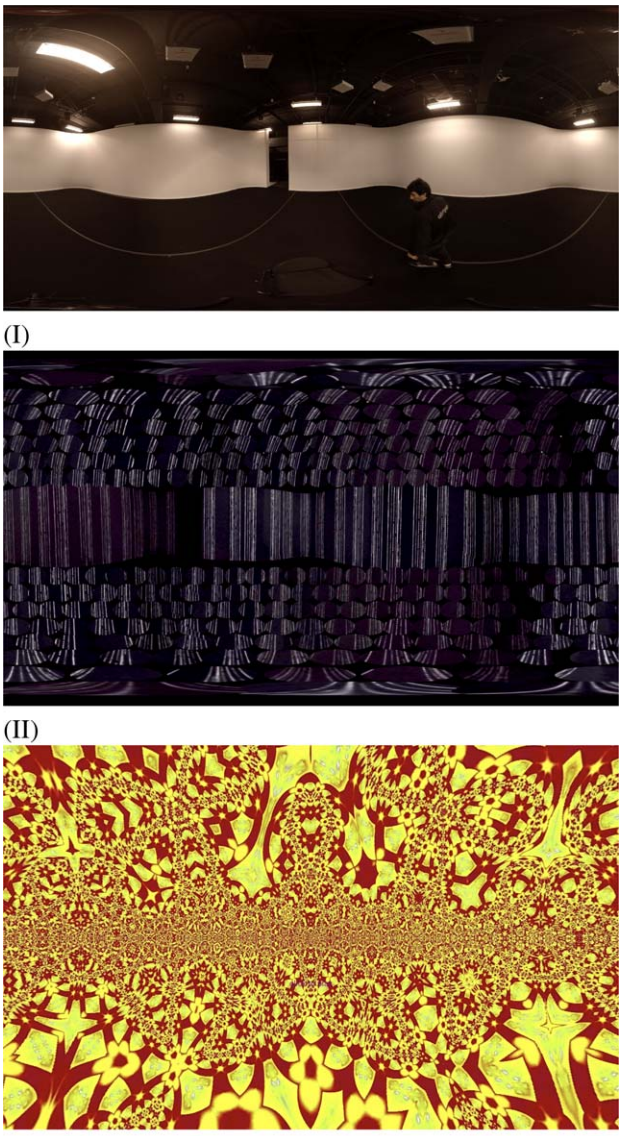

Fig. 5. Visual design stages of *od*: (I) equirectangular projection of CRAIVE-Lab panoramic screen, (II) panoramic video augmented with virtual mirrors, and (III) additional visual processing.

3.3 Interaction Design

The recorded composition, a fixed 16-channel wave file, is then piped into the Omnitone HOA decoder as an audioContext via *AudioBufferSourceNode* and rendered as binaural signals in the browser with reference to the rotation matrix updates from A-Frame. Communication between the video sphere (A-Frame) and binaural rendering (Omnitone) is achieved by updating values from registered A-Frame camera components that can be expanded; in our case, an entity attached with a tick handler grabs the current rotation values at every frame in Euler angles. The azimuth and elevation values are converted to a 3 x 3 matrix by a standard Open-GL "View" Matrix calculation. This matrix is then set as Omnitone's rotation matrix, updating the *HOARenderer* with the current rotation data as illustrated in Fig. 2. Finally, other HCI elements such as volume control, playback start/stop, keyboard interaction, and navigation smoothing were incorporated as registered A-Frame components.

4 CONCLUSION AND DISCUSSION

New spatial compositions continue to emerge from research institutions, studios, and art exhibitions, yet there are still no established or standardized methods to archive and distribute them. In this paper we have proposed a workflow scheme to build and present spatial multimedia. This workflow seeks to bring perspective to and cope with changing tools and technologies so that such practices remain accessible and inclusive. Web Audio and WebVR technologies reduce the need for expensive tools and performance spaces for development, offering a foundation for future high-level, web-based compositions. Furthermore, we applied the workflow to create *od*, a spatial audiovisual composition for binaural listening on the web. A number of technologies were employed for the project to not only document a space accurately, as suggested by the workflow, but to further process audiovisual media with an artistic approach. Our project for example began with recording spatial events in an immersive space and ended with an augmented version using VR effects tools. Since there are no standard methods of documentation for such practices, with this workflow and project we hope to demonstrate an approach that may be of use to creatives and researchers alike.

Working with new technology has obvious setbacks related to the lifespan of the employed tools. Specific to *od*, our chosen video calibration and stitching software was deprecated soon after the project commenced, and WebVR was standardized together with WebAR into WebXR shortly after project completion. On the other hand, some tools are expected to last over longer periods: accessing the web with browsers, headphones for personal listening, and smart phones in daily communication all demonstrate uses of technology that are embedded in modern life. Finally, we can address some findings and suggestions for future improvements from our experience. At this time we have found Mozilla Firefox and Google Chrome to be the most reliable browsers for both Web Audio and WebVR development. But 360 videos and HOA files are large in size and subsequently it can take considerable time on the client side to load and play. Efficient video formats such as .webm would be a significant performance improvement, but A-Frame currently supports .mp4 playback. In terms of codecs, incorporating ambisonic codecs in the FFmpeg libraries would provide a practical and unified way of dealing with all media involved in these projects. Lastly, HRTF selection in Web Audio applications can improve binaural listening qualities, but this requires further research. Since creative investigations in spatial music would benefit from higher-level workable interfaces to feasibly explore and experiment with new ideas, we have demonstrated a combination of offline ambisonic production and online binaural decoding techniques for audio processing.

5 ACKNOWLEDGMENT

We would like to thank Jonas Braasch for giving us unrestricted access to CRAIVE-Lab, Shawn Lawson for providing information and discussion on VR technologies, Sam Chabot for assistance with CRAIVE-Lab, Görkem Özdemir for crafting and executing the visual design, and Barış Demirdelen for help and assistance with coding.

6 REFERENCES

[1] D. Smalley, "Spatial Experience in Electro-Acoustic Music," *L'espace du son II*, pp. 121–124 (1991).

[2] S. Waters, "Timbre Composition: Ideology, Metaphor and Social Process," *Contemp. Music Rev.*, vol. 10, no. 2, pp. 129–134 (1994), DOI: 10.1080/07494469400640361, URL DOI: 10.1080/07494469400640361.

[3] C. Çakmak and R. Hamilton, "Composing Spatial Music With Web Audio and WebVR," presented at the *Web Audio Conference* (2019).

[4] M. A. Gerzon, "Design of Ambisonic Decoders for Multispeaker Surround Sound," presented at the *58th Convention of the Audio Engineering Society* (1977 Nov.).

[5] É. Gayou, "The GRM: Landmarks on a Historic Route," *Organ. Sound*, vol. 12, no. 3, pp. 203–211 (2007 Dec.), https://doi.org/10.1017/S1355771807001938.

[6] M. A. Harley, "Music of Sound and Light: Xenakis's Polytopes," *Leonardo*, vol. 31, no. 1, pp. 55–65 (1998).

[7] S. Williams, "Osaka Expo '70: The Promise and Reality of a Spherical Sound Stage," presented at the *Proceedings of inSONIC2015: Aesthetics of Spatial Audio in Sound, Music and Sound Art* (2015 Nov.).

[8] S. Sterken, "Reconstructing the Philips Pavilion, Brussels 1958: Elements for a Critical Assessment," in D. Heuvel, M. Mesman, W. Quist, B. Lemmens (Eds.), *Proceedings of the 10th International DOCOMOMO Conference*, pp. 93–98 (IOS Press, Rotterdam, 2008).

[9] E. Kermit-Canfield, "A Virtual Acousmonium for Transparent Speaker Systems," *Proc. 13th Sound and Music Computing Conference (SMC2016)* (2016 Aug.).

[10] M. Gurevich, D. Donohoe, and S. Bertet, "Ambisonic Spatialization for Networked Music Performance," presented at the *17th International Conference on Auditory Display* (2011 Jun.).

[11] M. Geronazzo, J. Kleimola, E. Sikström, A. de Götzen, S. Serafin, and F. Avanzini, "HOBA-VR: HRTF on Demand for Binaural Audio in Immersive Virtual Reality Environments," presented at the *144th Convention of the Audio Engineering Society* (2018 May), convention paper 433.

[12] T. Oberender, "An Architecture of the Dissolution of Boundaries: The Planetarium as a Gallery of the Future," in *Visual Art and Music in Planetariums*, The New Infinity, pp. 53–60 (König, Köln, Germany, 2019).

[13] M. A. Gerzon, "The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound," presented at the *50th Convention of the Audio Engineering Society* (1975 Mar.), convention paper L-20.

[14] H. Wittek and G. Theile, "Development and Application of a Stereophonic Multichannel Recording Technique for 3D Audio and VR," presented at the *143rd Convention of the Audio Engineering Society* (2017 Oct.), convention paper 9869.

[15] E. Bates, M. Gorzel, L. Ferguson, H. O'Dwyer, and F. M. Boland, "Comparing Ambisonic Microphones — Part 1," presented at the *AES International Conference on Sound Field Control* (2016 Jul.), conference paper 6-3.

[16] T. Carpentier, "A New Implementation of Spat in Max," *15th Sound Music Comput. Conf.*, pp. 184–191 (2018 Jul.).

[17] J. C. Schacher, "Seven Years of ICST Ambisonics Tools for Max/MSP - A Brief Report," presented at the *Proceedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics* (2010 May).

[18] T. Lossius and J. Anderson, "ATK Reaper: The Ambisonic Toolkit as JSFX Plugins," presented at the *Proceedings of the International Computer Music Association Conference* (2014).

[19] T. Carpentier, "ToscA: An OSC Communication Plugin for Object-Oriented Spatialization Authoring," *Proc. 41st Int. Comput. Music Conf.*, pp. 368–371 (2015).

[20] T. Carpentier, "A Versatile Workstation for the Diffusion, Mixing, and Post-Production of Spatial Audio," presented at the *Proceedings of the Linux Audio Conference* (2017 May).

[21] T. Carpentier, "Binaural Synthesis With the Web Audio API," presented at the *1st Web Audio Conference (WAC)* (2015 Jan.).

[22] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database," *Appl. Sci.*, vol. 8, no. 11, p. 2029 (2018 Oct.), https://doi.org/10.3390/app8112029.

[23] M. Binelli, D. Pinardi, T. Nili, and A. Farina, "Individualized HRTF for Playing VR Videos With Ambisonics Spatial Audio on HMDs," presented at the *AES International Conference on Audio for Virtual and Augmented Reality* (2018 Aug.), conference paper P3-5.

[24] G. Sharma, J. Braasch, and R. J. Radke, "Interactions in a Human-Scale Immersive Environment: The CRAIVE-Lab," presented at the *Cross-Surface 2016, in conjunction with the ACM International Conference on Interactive Surfaces and SpacesACM International Conference on Interactive Surfaces and Spaces* (2017 Nov.).

[25] T. Carpentier, M. Noisternig, and O. Warusfel, "Twenty Years of Ircam Spat: Looking Back, Looking Forward," *Proc. 41st Int. Comput. Music Conf.*, pp. 270–277 (2015 Sep.).

[26] S. Neelakantam and T. Pant, "Introduction to A-Frame," in *Learning Web-based Virtual Reality*, pp. 17–38 (Apress, Berkeley, CA, 2017).

[27] J. Grimshaw, "Space Exploration, Part 1: Telos and Stasis in the Dream House," in *Draw a Straight Line and Follow It: The Music and Mysticism of La Monte Young* (Oxford University Press, Oxford, UK, 2011).

THE AUTHORS

Cem Çakmak

Rob Hamilton

Cem Çakmak is a Ph.D. candidate and HASS Fellow at Rensselaer Polytechnic Institute's Electronic Arts program. As a composer and researcher, Cem's music and writings have been presented in international conferences, concert spaces, and music festivals. His research is focused on music and human-computer interaction, spatial electronic music composition, and multimedia design. He also holds an M.Sc. in Music focusing on Sonic Arts from the Istanbul Technical University's Centre of Advanced Studies in Music (MIAM) and a B.Eng. in Civil Engineering from the University of Manchester.

•

Rob Hamilton is an Assistant Professor of Music and Media at Rensselaer Polytechnic Institute. As a composer, performer, researcher, and software designer, his creative and analytical practice explores the cognitive implications of the spaces between interactive game environments, network topographies, and procedurally generated sound and music. He holds a Ph.D. in Computer-based Music Theory and Acoustics as well as an M.A. in Music, Science, and Technology from Stanford University's Center for Computer Research in Music and Acoustics (CCRMA) in the Department of Music, an M.M. in Computer Music Composition from the Peabody Institute of the Johns Hopkins University, and a B.A. in Music and Cognitive Science from Dartmouth College.