Composing Spatial Music with Web Audio and WebVR

Cem Çakmak
Rensselaer Polytechnic Institute
110 8th Street, Troy NY
cakmao@rpi.edu

Rob Hamilton
Rensselaer Polytechnic Institute
110 8th Street, Troy NY
hamilr4@rpi.edu

ABSTRACT

Composers have been exploring complex spatialization techniques within multi-channel sound fields since the earliest days of electroacoustic and electronic music. However the reproduction of such works outside of highly specified concert halls and academic research facilities, or even their accurate reproduction within those spaces, is difficult and unpredictable at best. Tools such as Omnitone combine the reach and simplicity of web browsers with the flexibility and power of higher-order ambisonics (HOA) and binaural rendering, ensuring greater accessibility for existing spatial electronic musical works as well as acting as a platform upon which future works for virtual sound fields can be implemented. This paper describes the technical design and artistic conception of one such spatial composition for binaural listening and immersive visuals on the web - *od* - produced in the CRAIVE-Lab, an immersive audio-visual facility.

1. INTRODUCTION

Spatial, or multi-channel, approaches in electronic music composition are as old as the practice itself. As such works furthered experimentation in sound, light and multimedia, they surrounded, and sought to immerse the audience 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. Spatial electronic music often depends on research institutions and patronage due to their experimental nature and funding required, while the audience ranges from highly trained ears to unassuming visitors. Furthermore, such works have always remained a niche and their recreation unfeasible. This paper breaks down a recent multimedia work, *od*, to illustrate a more accessible and inclusive way to experience spatial electronic music based on Web Audio and WebVR technologies. In order to achieve this, the authors seek to bring together a number of online and offline tools together in an artistically meaningful way.

Emmerson identifies two traditions of diffusion in electronic music: the idealist and realist approaches [5]. In short, the idealist approach works toward conveying the composer's vision as accurately as possible to the listener, whereas the realist approach emphasizes the diversity of the hearing experience among individuals within the listening space. This difference in experience is due to a variety of factors that include audience seating, architecture of the space, or external disruptions; since such works focus on timbre and space as opposed to traditional musics that rely predominantly on pitch and rhythm, they are much less tolerant to disturbances, faulty equipment, or lower qualities of production [15].

Regardless of the compositional approach, spatial works remain poorly documented in terms of the audience experience, and their reconstruction remains a challenge. Thus the main goals of the research are as follows:

- Utilize contemporary web technologies to facilitate the composition and dissemination of a novel, spatial music piece.
- Exercise the above-mentioned idealist approach in a way that includes all listeners in an idealized listening situation.
- Enable access to a specific physical site that is typically difficult to visit.
- Augment the immersiveness of a real-world location with VR production techniques.

2. BACKGROUND

Beyond stereophony lies an infinite playground; one where loudspeakers work together not only to form a directional image, but to encompass the listeners and innovate new spatial complexities. While ambisonic systems aim to recreate a spherical soundscape as accurately as possible, composers also use unconventional speaker arrangements, where loudspeakers are treated as instruments that contribute to the music with their characteristics and form a relationship with their surroundings. The "acousmonium", first designed in 1974 at GRM [7], is an asymmetrical approach in multichannel loudspeaker distribution where the focus is on combining loudspeakers that vary in size, shape, response and function together, and showcase not only auditory but also visual novelty. Furthermore, Iannis Xenakis' 1971 polytope *Persepolis* was set in the dark of night outdoors, among the ruins of the ancient Persian city of the same name. In addition to the 8-channel composition diffused over 59 speakers, the performance utilized lasers, spotlights, as well children parading with torches over the hills [10]. Besides real-world locations, specific constructions were designed and built,

Web Audio Conference WAC-2019, December 4–6, 2019, Trondheim, Norway.

bringing together composers, architects, and technicians in order to create unique yet ephemeral structures for spatial performances; notable examples include the Philips Pavilion in EXPO '58 Brussels, or the Pepsi Pavilion in EXPO '70 Osaka [16].

Although the works mentioned above are often mythicized as significant achievements in electronic music history, recreating such pieces as they were meant to be listened to is a delicate task, if even possible. Some of these compositions have been released in stereo, but restaging *Persepolis*, for example, a piece riddled with themes of Zoroastrian fire-worshipping, could possibly have extra-musical ramifications in modern day Iran. As for the EXPO pavilions, there were a number of attempts to physically and virtually reconstruct the Philips Pavilion and the Edgard Varèse's piece *Poème Electronique* performed inside the structure in order to assess the works retrospectively [14]. These however, are often expensive undertakings with dubious fidelity to the original.

With the ongoing development of VR and web technologies, researchers are challenged to create new, innovative forms of spatial interactions. As these technologies become more widespread, new spatial artworks can potentially become more accessible and less ephemeral. Fyfe et al. [6] combine web audio streams with motion tracking and binaural audio to create telepresence for networked collaborations. Kermit-Canfield [11] proposes a configurable diffusion tool for building virtual acousmonia and encode the output using ambisonics for transparent speaker arrangements. Çamcı et al. [4] introduce a stylized interactive system with UI controls and visual representations for composing detailed virtual sonic environments with an implementation of Web Audio API. Lastly, personalized head related transfer functions (HRTFs) for spatial sound reproduction are being implemented in Web Audio; Geronazzo et al. propose a framework based on Unity and the Web Audio API for new immersive experiences on mobile VR [8]. Likewise, this research brings together a number of tools and instruments available to construct a contemporary spatial music piece, available to experience through desktop, mobile or head-mounted devices.

3. TOOLS AND INSTRUMENTS

3.1 Omnitone

Written with Web Audio API, Omnitone[1] is a JavaScript implementation for ambisonic decoding up to the third order and binaural rendering on the web browser. Illustrated in Fig. 1, multi-channel ambisonic files are streamed in via AudioBufferSourceNode, and the head position is translated to a rotation matrix via either user interaction or sensor data (or in this project's case, A-Frame camera position). Binaural rendering is handled by Convolver and GainNode interfaces, both native to Web Audio. In line with Google's spatial media specifications, the decoded ambisonic signals for each ear pass through head related transfer functions (HRTFs) based on SADIE filters[2] to simulate binaural hearing.

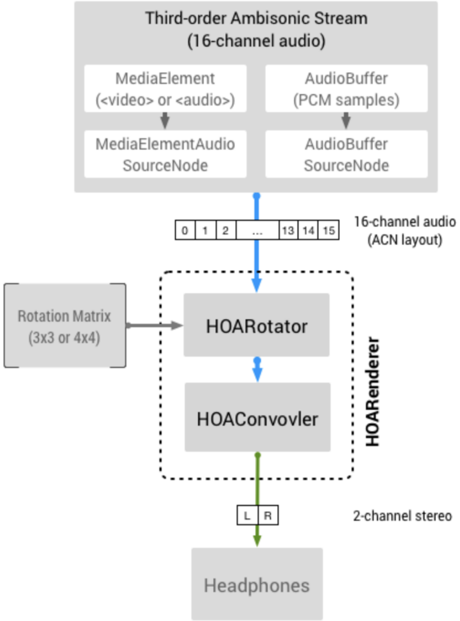

Figure 1: Diagram of Omnitone's third-order ambisonic to binaural rendering

3.2 A-Frame

The A-Frame library[3] is an open source, three.js framework for WebVR. Developed by the Mozilla VR team in order make VR content 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. Furthermore, A-Frame works across a range of mobile, desktop and head-mounted devices [12]. The A-Frame framework has an entity-component system architecture, common for game development applications; most code lives within registered components that are plugged into entities to describe their specific attributes. Unfortunately, A-Frame does not support WebM[4] as of today, an extremely efficient media format for video playback on browsers.

3.3 CRAIVE-Lab

Collaborative-Research Augmented Immersive Virtual Environment Laboratory, or CRAIVE-Lab [13], is a state-of-the-art interactive immersive environment operating in Rensselaer Technology Park. The surround screen, shaped as a rounded rectangle, totals a resolution of 15360x1200 pixels. The panoramic imahe is front-projected by eight Canon REALiS WUX400ST Pro short-throw LCoS projectors mounted on the grid above. These projectors are calibrated and warped with Pixelwix to create a smooth continuous image along the entire projection surface. Furthermore, along the back of the screen is a 128-channel wave field synthesis array set up with additional speakers mounted on the above grid for HOA projection support, although the on-site sound setup is not employed in this research.

This audiovisual environment is often used in an architectural context and research in interaction, where a specific

[1] https://googlechrome.github.io/omnitone/

[2] https://www.york.ac.uk/sadie-project/GoogleVRSADIE.html

[3] https://aframe.io/

[4] https://www.webmproject.org/

site is reproduced visually as well as aurally; given that impulse responses taken from the real site modeling its reverberance is possible. Furthermore, some projects add in new designs within the site footage, creating hyper-real immersive landscapes. But, in the context of this project, the goal is to export CRAIVE-Lab from its real-world location and enable people to virtually experience works made within it, as opposed to its main function where real-world sites are placed within the immersive space.

3.4 Max & SPAT5

IRCAM's *Spatialisateur*, or SPAT [3], is a library of Max objects written in C++ for spatializing sound in real-time for ambisonics or nearfield binaural synthesis, and generating artificial reverberations for room effect. SPAT offers a dynamic 3D environment for organizing and manipulating sound sources, in addition to modeling loudspeaker arrangements for real-time synthesis and diffusion. The SPAT library can be used for live performances, mixing, post-production, installations, VR and other applications. The 5th version of SPAT [2] implements the open sound control (OSC) protocol for the processors throughout the library, along with more detailed documentation, improved HOA features, cross-operability with VR SDKs and other aspects.

3.5 GoPro Omni

GoPro's Omni rig encapsulates and powers six GoPro Hero 4 cameras on each side of its cube-shaped structure 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. To stitch all the footage together to construct a spherical image, we used Kolor softwares AutoPano Giga 4.4 and AutoPano Video Pro 2.6, all deprecated since September 2018. Finally, the stitched footage is further processed with MantraVR[5], an Adobe After Effects plugin for VR content production.

4. COMPOSITION

The composition process involves navigating through different environments using the tools described in section 3. This task is mainly split into two parts due to their respective media types: the audible and the visual. As illustrated in Fig. 2, the design of the audio and visuals are mostly independent from each other during composition, joined together and synced between A-Frame and Omnitone.

4.1 Auditory Design

Despite the strong influence of 20th century spatial works on the aethetics of the project as mentioned in 2, the central musical idea comes from a stereo drone installation by La Monte Young, *The Base 9:7:4 Symmetry in Prime Time When Centered above and below The Lowest Term Primes in The Range 288 to 224 with The Addition of 279 and 261 in Which The Half of The Symmetric Division Mapped above and Including 288 Consists of The Powers of 2 Multiplied by The Primes within The Ranges of 144 to 128, 72 to 64 and 36 to 32 Which Are Symmetrical to Those Primes in Lowest Terms in The Half of The Symmetric Division Mapped below and Including 224 within The Ranges 126 to 112, 63 to 56 and 31.5 to 28 with The Addition of 119*, exhibited at the Dream House, located in downtown Manhattan [17].

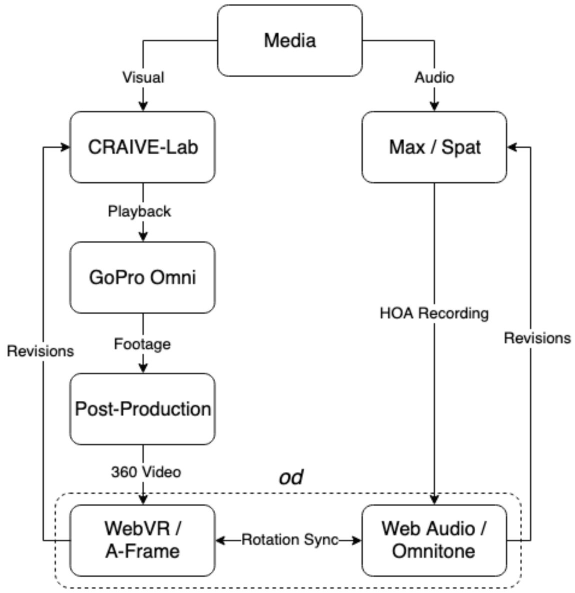

Figure 2: Composition workflow for *od*

The sonic properties of the drone are strictly related to the listener's body, position and movement; as long as the individual stands still, the drone is constant, but when the head, or the body is displaced, certain partials diminish and new one emerge, activating the drone. The artistic motivation was thus composing a spherical drone in a virtual space, where the user's change in viewpoint activates the uniform soundscape and sets timbre in motion.

In the context of our project, Max/MSP with SPAT externals offered a composition workflow where:

- An ambisonic scene is composed and visually monitored via spat5.viewer object, as seen in Fig. 3,
- The scene is normalized with the N3D scheme for spherical harmonic components and streamed in the HOA encoder of third order,
- The HOA stream is then recorded into a 16-channel sound file and simultaneously transcoded to binaural for headphone listening,
- Finally, the scene is transformed in yaw, pitch, and roll in order to monitor and interact with the soundscape binaurally

This ambisonic scene consists of 20 virtual sound sources arranged as the vertices of a dodecahedron surrounding the listener, as seen in Fig. 3. Instead of utilizing prime frequencies as in Young's piece, we have randomly distributed three spectral clusters of low, mid and high frequencies that are centered around 150, 1500, and 6000 Hz. Although starting with pure tones, we have found triangle waves more suitable as sound sources, since they contain overtones that enhance to the timbral quality without weighing down the fundamentals as much as sawtooth or square waves. Finally, for further effects, we attached Perlin noise generators to these sound sources to add smooth subtle vibrations, and make the auditory scene feel more organic.

[5] https://www.mettle.com/product/mantra-vr/

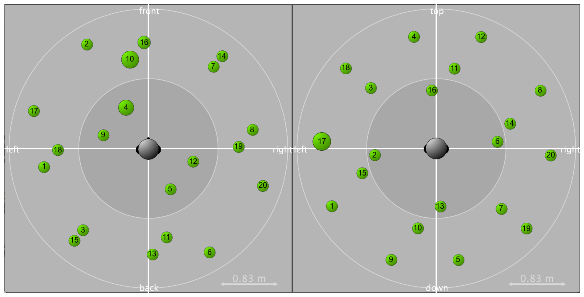

Figure 3: Snapshot from the spat5.viewer interface with 20 sound sources distributed in a dodecahedral pattern around the listener

4.2 Visual Design

As mentioned in section 3.3 above, the main goal of the visual design is to render the CRAIVE space accessible through the browser to experience virtually. In order to achieve this, we shot a number of 360°video recordings using a GoPro Omni rig consisting of six GoPro Hero 4 cameras set to 4:3 4k resolution. The camera array was placed at the center of the room in all three axes, or the “sweet-spot”. Extending from the author’s previous 43-channel acousmonium performance held at EMPAC Studio 1[6], the projected video screen consisted of footage from the concert, stretched, reproduced and manipulated for the 15360x1200 screen. While a number of formats were experimented with for the projection screen of unusual dimensions, we have found the .webm format significantly more efficient than other encodings in terms quality, of file size and smooth playback, despite the fact that it meant for media online.

The individual camera recordings were then stitched together with proprietary software to construct a 8000x4000 spherical image from the center of CRAIVE-Lab. On the horizontal, the projection screen is continuous apart from the entrance; but above and below the screen remained large blank areas. In order to augment the immersive qualities of the room, we strategically placed virtual mirrors on the spherical image, as seen in Fig. 4, to expand the projected image in all directions using Mantra VR plugins. The final, augmented spherical video is then converted to .mp4 and specified as an A-Frame asset to be played played back on the web browser.

4.3 Implementation

The HRTF filters for SPAT’s binaural rendering were default KEMAR filters. These differ from the SADIE filters for Omnitone’s rendering in terms of a number of attributes; subjects comparing HRTF qualities in a recent research have rated KEMAR as brighter, richer, better externalized and overall more preferable compared to SADIE database samples [1]. While this did not have a significant effect on the timbral variations emerging upon perspective change, it nontheless demands the composer to be mindful of the differences in overall timbre during the compositional process and the final web Audio implementation. This is one of the causes for revisions and repetition of the procedure in a loop as illustrated in Fig. 2. The recorded composition, a fixed multi-channel wave file, is then piped into the Omnitone HOA decoder as an audioContext via *AudioBufferSourceNode* and rendered as binaural signals on the browser with reference to the rotation matrix updates from A-Frame.

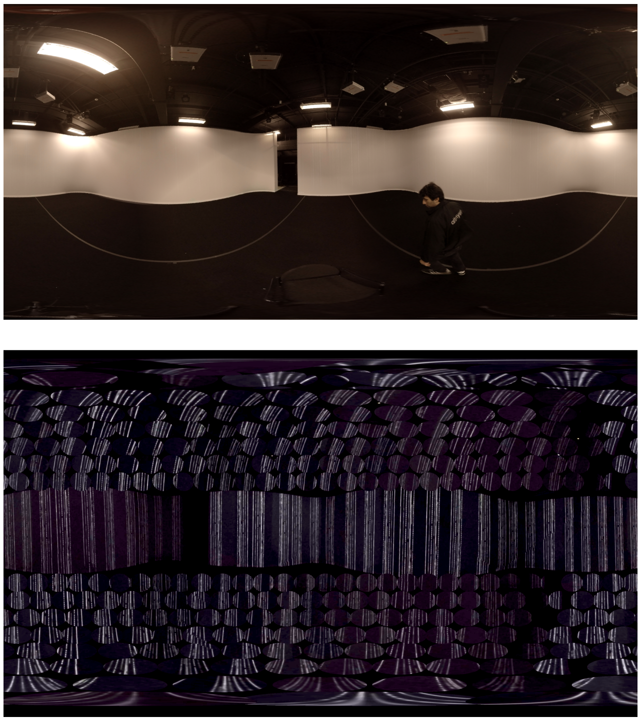

Figure 4: Spherical image of the CRAIVE panoramic screen (above) and the room augmented with virtual mirrors in post-production (below)

Communication between A-Frame and Omnitone is achieved by updating values from registered components; in our case, an entity attached as a camera component with a .tick() handler grabs the current rotation values at every frame in Euler angles. The azimuth and elevation values are converted to a 3x3 matrix by a standard Open-GL ”View” Matrix calculation; the matrix is then set as the Omnitone rotation matrix, updating the HOARenderer as illustrated in Fig. 1.

5. CONCLUSION AND DISCUSSION

This paper documented a contemporary research in electronic music composition and multimedia design that resulted in *od*, a spatial composition for binaural listening on the web. We have used a number of online and offline technologies to not only document a physical space accurately, but to further augment its immersive audiovisual qualities with an artistic approach. New spatial compositions continue to emerge at research institutions, studios and art exhibitions, yet there are still no standard methods to archive and share such works. Since there are no standard methods of documentation for such works, we hope to demonstrate a favorable approach to document spatial music that may be of use to other composers. The tools exist and are available; with the use of ambisonic microphones and Web Audio-based renderers like Omnitone, composers can easily and accurately share their recorded works online for binau-

[6] https://vimeo.com/308695695

ral listening. Making these accessible to an online audience could facilitate healthier communication among composers and listeners and expand the community. Furthermore, Web Audio and WebVR technologies will hopefully reduce the need for expensive equipment and performance spaces; this will likely evolve the practice into a more inclusive and accessible one, enabling more people to contribute.

With these thoughts in mind, we can address some issues we faced for future improvements. Firstly, we have found Google Chrome to be the only reliable browser during the development phase for both Web Audio and WebVR components. Moreover, 360°videos and ambisonic files are large in size due to the high amount of information they contain; thus, it takes considerable time on the user side to load and play the composition. Utilizing other formats (such as .webm videos instead of .mp4) for smoother and longer playback with better spatial fidelity would be a significant performance improvement. Streaming the spatial media rather than downloading is another possible future development; as of today A-Frame only supports a component for streaming with the Vimeo API, but this is available only for paying members.[7] HRTF filters differ from one another in terms of their spectral qualities and spatial perception, as mentioned in section 4.3. Providing multiple HRTF libraries to select and compare within the Web Audio renderer would facilitate further musical experimentation. Methods of modeling one's own ear and creating their personal HRTF filters are possible and available today [9]; so in addition to the standard libraries, implementing individualized HRTFs would also be an interesting feature.

Finally, our general impression from available Web Audio API resources is that they essentially speak to web developers with extra-musical concerns rather than creative investigators. Now supported by a wide range of browsers and devices, Web Audio and WebVR contain much potential for artistic research, but a shared language for spatial media composition is yet unestablished. Therefore with this paper, we hope to demonstrate new possibilities for this emerging field and contribute to its critical discussion.

6. ACKNOWLEDGMENTS

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 assitance with the CRAIVE infrastructure and camera array, Görkem Özdemir for executing the visual design, Büşra Tunç for help with issues on panoramic projections and immersive video, and lastly Barış Demirdelen for help and assistance with coding.

7. REFERENCES

[1] 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. *Applied Sciences*, 8(11):2029, Oct. 2018.

[2] T. Carpentier. A new implementation of Spat in Max. In *Proceedings of the 15th Sound and Music Computing Conference*, pages 184 – 191, Limassol, CY, July 2018.

[3] T. Carpentier, M. Noisternig, and O. Warusfel. Twenty Years of Ircam Spat: Looking Back, Looking Forward. In *Proceedings of the 41st International Computer Music Conference*, pages 270 – 277, Denton, TX, Sept. 2015.

[4] A. Çamcı, P. Murray, and A. G. Forbes. A Web-based System for Designing Interactive Virtual Soundscapes. In *Proceedings of the 42nd International Computer Music Conference*, pages 579–585, 2016.

[5] S. Emmerson. Diffusion-Projection: The Grain of the Loudspeaker. In *Living Electronic Music*, pages 143–170. Routledge, Leicester, UK, Sept. 2017.

[6] L. Fyfe, O. Gladin, C. Fleury, and M. Beaudouin-Lafon. Combining Web Audio Streaming, Motion Capture, and Binaural Audio in a Telepresence System. Berlin, DE, Sept. 2018.

[7] E. Gayou. The GRM: landmarks on a historic route. *Organised Sound*, 12(03), Dec. 2007.

[8] M. Geronazzo, J. Kleimola, E. Sikstroöm, A. de Götzen, and S. SeraïňĄn. HOBA-VR: HRTF On Demand for Binaural Audio in immersive virtual reality environments. In *Audio Engineering Society Convention 144*, Milan, IT, May 2018. Audio Engineering Society.

[9] S. Ghorbal, R. Séguier, and X. Bonjour. Process of HRTF individualization by 3d statistical ear model. In *Audio Engineering Society Convention 141*, Los Angeles, CA, 2016.

[10] M. A. Harley. Music of sound and light: Xenakis's polytopes. *Leonardo*, 31(1):55–65, 1998.

[11] E. Kermit-Canfield. A Virtual Acousmonium for Transparent Speaker Systems. Hamburg, DE, Aug. 2016.

[12] S. Neelakantam and T. Pant. Introduction to A-Frame. In *Learning Web-based Virtual Reality*, pages 17–38. Apress, Berkeley, CA, 2017.

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

[14] S. Sterken. Reconstructing the Philips Pavilion, Brussels 1958: Elements for a Critical Assessment. In D. v. d. Heuvel, M. Mesman, W. Quist, and Bert Lemmens, editors, *Proceedings of the 10th International DOCOMOMO Conference*, pages 93–98, Rotterdam, NL, Sept. 2008. IOS Press.

[15] S. Waters. Timbre composition: Ideology, metaphor and social process. *Contemporary Music Review*, 10(2):129–134, 1994.

[16] S. Williams. Osaka Expo '70: The promise and reality of a spherical sound stage. In *Proceedings of inSONIC2015, Aesthetics of Spatial Audio in Sound, Music and Sound Art*, Karlsruhe, DE, November 2015.

[17] L. M. Young and M. Zazeela. Dream House Opens for the 2016-2017 Season - Our 24th Year. url-http://www.melafoundation.org/DHpressFY17.html, 2016.

[7] https://github.com/vimeo/aframe-vimeo-component