Toward Six Degrees of Freedom Audio Recording and Playback Using Multiple Ambisonics Sound Fields

Audio Engineering Society

Convention Paper

Presented at the 146th Convention

2019 March 20–23, Dublin, Ireland

Toward Six Degrees of Freedom Audio Recording and Playback Using Multiple Ambisonics Sound Fields

Eduardo Patricio[1], Andrzej Rumiński[1], Adam Kuklasiński[1], Łukasz Januszkiewicz[1], Tomasz Żernicki[1]

[1] *Zylia Sp. z o.o., Umultowska 85, 61-149 Poznań, Poland*

Correspondence should be addressed to Tomasz Żernicki (tomasz.zernicki@zylia.pl)

ABSTRACT

This paper describes a strategy for recording sound and enabling six-degrees-of-freedom (6DoF) playback, making use of multiple simultaneous and synchronized higher-order ambisonics (HOA) recordings. For the evaluation of the proposed approach a 3D audio-visual navigable playback system was implemented. Subjective listening tests were conducted presenting three distinct scenarios, one using spatialized mono sources and the other two interpolated listening points from 1st and 3rd order multiple ambisonics sound fields. The obtained results demonstrate that HOA recordings are suitable for reproduction of 6DoF immersive audio scenes.

1 Introduction

Immersive media has had a significant increase in popularity and, as related technologies are developed, its usefulness has also seen growth with potential applications in entertainment, research, commerce and education. Six-degrees-of-freedom (6DoF) usually refers to the physical displacement of a rigid body in space. It combines 3 rotational (roll, pitch and yaw) and 3 translational (up-down, left-right and forward-back) movements. The term is also used to refer to the freedom of navigation in immersive / VR environments.

While 6DoF has long been a standard in computer gaming, with widely available tools to implement both immersive audio and video, the same cannot be said about cinematic audio and video scenarios. Most VR content available nowadays presents a 3DoF (three-degrees-of-freedom) scenario, in which the user occupies a single, fixed point of view allowing rotation, but not translation movements. There have been noticeable advancements in 6DoF volumetric videography [1] (e.g. Facebook RED Manifold camera and Orbx formats for OctaneRenderer™ [2]; Intel® Realsense™ [3]) which are relevant to VR/AR applications. On the other hand, there is still much to be done regarding live recorded 6DoF audio solutions. Cinematic (i.e. live captured / recorded) 6DoF audio implies capturing entire sound fields at once in a common physical space as opposed to encoding sound objects in a synthetic manner.

There is growing interest in 6DoF audio [4][5], but the solutions for live recorded scenarios are still very limited. Live recorded 6DoF audio can be particularly useful in scenarios in which it is of relevance to capture the acoustic characteristics of a specific space (e.g. concert room) or synchronized spatially spread sound sources (e.g. performing arts; sports events). It is possible to indicate two main approaches to live recorded 6DoF audio rendering. The first makes use of single ambisonics recordings and simulated off-center listening perspectives. That is, perspectives other than the center of the spherical sound field which is the standard in ambisonics technology [6][7][8]. The second one relies on performing interpolation of simultaneous spatially adjacent recordings [9][10][11].

The research endeavor that led to the present paper is founded in the hypothesis that an interpolation strategy based on loudness crossfading of spatially adjacent ambisonics recordings could be successfully used to generate an audio medium that allowed 6DoF interaction. An experimental interpolation strategy was devised to investigate whether viable perceptual results could be achieved.

This paper is structured as follows. In Section 2, we describe the recording and playback components of the proposed method. Section 3 brings information regarding an interactive implementation developed to enable subjective evaluation of the proposed interpolation method. Section 4 describes the conducted adapted MUSHRA test and its results can be found in Section 5. Following the test results, a short discussion is done in Section 6. To conclude, Section 7 presents a summary of this research effort.

2 Proposed method

The method proposed in this paper is composed of a recording system and a playback system.

2.1 Recording system

The recording system is composed of multiple higher-order ambisonics (HOA) microphones distributed across the recorded sound scene. The proposed playback method is an interpolator and as such does not support extrapolation of the listening point beyond the grid of microphones. Thus, the entire area or volume that is to be made navigable in the resulting recording needs to be covered by the grid. Although the proposed playback method does not depend on it, in order to ensure even coverage, we propose the use of a uniform grid composed of equilateral triangles or squares. Such system would be sufficient in cases where the height dimension is not needed, i.e. for 5DoF recordings. In cases where full 6DoF *with height* is to be recorded, several layers of the grid may be stacked one above the other, possibly with an offset. Orientation of each HOA microphone in the grid should be the same, i.e. the “front” and “top” of all microphones should point to the same directions, respectively.

In the work reported here, the 19-channel ZYLIA ZM-1 spherical microphone array [12][13] was used together with a software A-B converter [14] capable of producing ambisonics B-Format of up to the 3rd order.

Additionally, as the ZYLIA ZM-1 does not support external synchronization through a word clock or USB input, a dedicated synchronization method was implemented. The method is based on a hardware and a software component:

- piezoelectric buzzers driven by a common synchronization signal, attached at the base of each ZM-1 microphone,
- a software tool detecting a synchronization impulse played by the buzzers near the beginning and end of each recording.

This synchronization method allows the beginning of the recording from each HOA microphone to be time-aligned as well as the sample clock drift to be estimated. This operation allows for linear interpolation of audio samples. Similar synchronization method was also successfully used in previous authors’ work with multiple ZM-1 microphones [15][16].

2.2 Playback system

The second component of the proposed method is a playback system capable of ambisonics signal interpolation at locations between physical microphones used during the recording stage.

The proposed method of ambisonics sound field interpolation operates on time-domain ambisonics signal components which we denote $y_{m,p}(n)$, where m is the number of the HOA microphone, p is the ambisonics component index, and n is the sample index. The interpolated ambisonics signal $x_p(n)$ is calculated as a sum of contributions from all HOA microphones in the recording grid. These contributions are calculated by a distance-dependent filtering and scaling of the original ambisonics components. Denoting the number of HOA microphones in the recording grid by M, the distance between the point of interpolation and the m-th microphone by d_m, the scaling function by $a_p(d_m)$, the filter by $h(d_m)$, and using the convention that $(a*b)(n)$ is the convolution of signals $a(n)$ and $b(n)$, the interpolated signal can be expressed by:

$$x_p(n) = \sum_{m=1}^{M} a_p(d_m)\big(h(d_m) * y_{m,p}\big)(n) \quad (1)$$

The scaling function $a_p(d_m)$ applies a gradual attenuation of contributions from far-away microphones. Additionally, a re-balancing of the ratio between the 0^{th} order omnidirectional ambisonics component (p=0) and the directional components of higher orders (p>0) is applied. Thanks to the distance-dependent attenuation, the closest ambisonics sound fields have the greatest contribution to the output signal. The ambisonics component re-balancing removes directional information from the signal of microphones other than the nearest one. This decreases the chance of ambiguous localization of sources when moving across the mid-point between two microphones. Attenuation $l(d_m)$ and component re-balancing $k_p(d_m)$ are progressively applied only when d_m exceeds corresponding threshold values t_l and t_k. Beyond these distances $l(d_m)$ and $k_p(d_m)$ change linearly in dB: the greater the distance the stronger the attenuation and the greater the dominance of the omni-directional component over the directional ones. Mathematical formulation of the above follows:

$$a_p(d_m) = 10^{[l(d_m)+k_p(d_m)]/20} \quad (2)$$

$$l(d_m) = \begin{cases} 0, & \text{if } d_m \le t_l, \\ s_l(d_m - t_l), & \text{if } d_m > t_l \end{cases} \quad (3)$$

$$k_p(d_m) = \begin{cases} 0, & \text{if } d_m \le t_k, \\ s_{k,p}(d_m - t_k), & \text{if } d_m > t_k \end{cases} \quad (4)$$

Distance-dependent attenuation and ambisonics order re-balancing are formulated nearly identically cf. (3) and (4). However, the attenuation slopes for ambisonics component re-balancing can be different for each ambisonics component index p. In the intended use, this slope will be positive for the 0^{th} order component and negative for higher-order components:

$$s_{k,p=0} > 0, \quad s_{k,p>0} < 0 \quad (5)$$

Moreover, the sum of the 0^{th} order gain slope $s_{k,p=0}$ and the overall gain slope s_l should be negative. This ensures that the gain of the 0^{th} order component decreases as d_m increases.

$$s_{k,p=0} + s_l < 0 \quad (6)$$

The distance-dependent $h(d_m)$ is a first-order low-pass infinite impulse response filter whose cut-off frequency f_c is equal to 20kHz when d_m is below a threshold value t_f, and falls linearly with a slope $s_f < 0$ when d_m is above t_f:

$$f_c(d_m) = \begin{cases} 20\text{kHz}, & \text{if } d_m \le t_f, \\ 20\text{kHz} + s_f(d_m - t_f), & \text{if } d_m > t_f \end{cases} \quad (7)$$

The distance-dependent filtering simulates natural attenuation of high-frequency sounds that would occur over long distances.

While more advanced methods based on physical modeling of the sound field have been proposed in the past [9] [10], the relative simplicity of the proposed method allows real-time interactive systems to be easily implemented.

3 Evaluation platform

An interactive system was developed to test the proposed interpolation method of simultaneous adjacent ambisonics recordings. Its final design choices, regarding functionality and parameter control, were based on the general theoretical proposition and the need to perform interactive subjective evaluations.

The system has two main components:

- an input/control application built as a representational navigable 3D environment built in Unity [17], and
- a playback program (Figure 1) built with Max MSP [18] that executes all the necessary audio transformations based on the navigation input data.

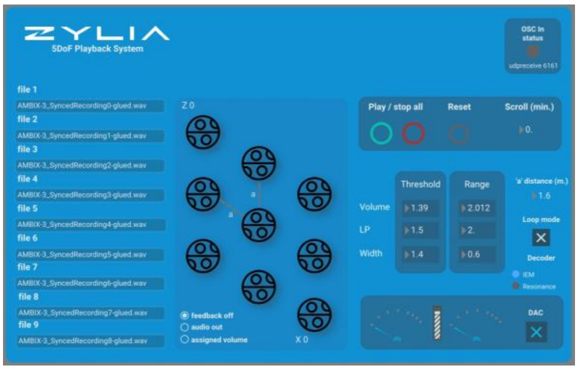

Figure 1. Playback program user interface.

The navigable Unity scene is configured to send the current location and orientation of the virtual view-point to the Max MSP component via OSC protocol. This positioning data is used to calculate the distance d_m between the listener's position and the center of each sound field. This distance is then used to perform the signal transformations outlined in Section 2.2. This playback tool with adjustable parameters made possible to listen to several interpolation configurations. Eventually, based on subjective auditory experience, a specific set of parameter values was chosen as efficient. It is possible to consider parameter adjustments for other scenarios.

The system considers a specific microphone arrangement as seen on the central area of the application's user interface (Figure 1). The distance between microphones, in meters, can be set in the program to match the distance used during recording. This parameter is essential to calculate the position of each microphone in the grid and, consequently, perform the necessary distance-based interpolations.

The output of the interpolated ambisonics sound fields is sent to a binaural decoder (either IEM's Binaural decoder [19] or Google's Resonance binaural decoder [20] and should be listened to on headphones. The standard ambisonics rotation transformations are done by IEM's 'Scene Rotator' VST plug-in.

The playback system is capable of 5-degrees-of-freedom playback. Vertical translation movement (up and down) is not included, but it could be implemented in a future iteration for playback of recording grids with microphone arrays placed in different elevations.

4 Subjective evaluation

4.1 Methodology

The main goal of the conducted evaluation experiment was to assess how well the spatial attributes of a recorded acoustic scene are preserved when using the proposed strategy for interpolation of multiple ambisonics sound fields. The following aspects were of particular interest:

- naturalness and realism of the perceived direction and distance of sound sources,
- naturalness and smoothness of auditory image evolution when moving across the scene.

To this end a modified MUSHRA [21] methodology was adopted with audio-visual stimuli presented by means of a computer screen and stereo headphones. Modifications of the MUSHRA methodology were inspired by [9] and mainly included the addition of visual component to the stimuli. This allowed the test subjects to have a visual reference regarding the true placement of sound sources in the scene.

4.2 Audio stimuli

Audio component of the stimuli was prepared as follows. An acoustic scene comprising three sound sources was recorded in a room measuring 4.5 x 6.5 x 2.8 m and exhibiting an average reverberation time of 0.26 s. The sources were chosen to have different tonal and temporal characteristics. The first source was a floor-standing fan that was switched on throughout the recording session. Strips of foil were attached to it in order to make the airflow more audible. Two 5-inch loudspeakers were used as the second and third sources. A sound of a phone ringing intermittently was played through one the loudspeakers and a cartoon soundtrack through the other one. The three sources were arranged in a triangle around the center of the room, 2.5 to 3.5 meters from one another.

The above-mentioned sources were recorded by a system made up of nine ZYLIA ZM-1 HOA microphones arranged in an equilateral triangular grid forming a diamond shape encompassing nearly the entire room (Figure 2). The distance between neighboring microphones in the grid was 1.6 m and the height of all the microphones above the floor was 1.7 m. Since the HOA microphone grid was two-dimensional (without height), the resulting recording did not contain full 6DoF information. This was deemed sufficient for the purpose of this evaluation. In addition to the HOA microphones, three large-diaphragm condenser microphones were used to record each of the sources from a short distance. Directional characteristic of these microphones was set to cardioid which resulted in a high degree of separation between the recorded sources.

The signals registered by the HOA microphones were time-aligned using the system described in Section 2.1 and subsequently transformed to the ambisonics domain using the A-B converter. The ambisonics-encoded signals were processed by the proposed interpolation method and subsequently binauralized by IEM rotator and binaural decoder plugins within the Max MSP described in Section 3.

Since the recording took place in a relatively small room, the low-pass filtering functionality of the proposed method was not used. The remaining parameters of the interpolator were set as follows:

$$t_l = t_k = 1.4\text{m}, \qquad s_l = -38\text{dB/m},$$
$$s_{k,p=0} = 10\text{dB/m}, \; s_{k,p>0} = -126\text{dB/m}$$

Three different renderings of the ambisonics signals were prepared as stimuli for the test:

- The 0th order ambisonics (0OA) interpolated by cross-fading according to listener position. This was included as the hidden anchor in the test.
- The 1st order ambisonics (1OA) interpolated by using the proposed method.
- The 3rd order ambisonics (3OA) interpolated by using the proposed method.

The 0OA signal contained no spatial clues apart from loudness changes according to distance from a given source.

The fourth stimulus condition was prepared by spatializing signals of the cardioid microphones at the original positions of the sound sources in the room using Google Resonance decoder and room reverberation simulator (ResonanceAudioRoom Unity audio component). This stimulus was used as the reference in the MUSHRA test.

4.3 Visual stimulus

The visual component of the stimuli was prepared in Unity 3D engine and consisted of an interactively navigable virtual recreation of the room where the sound signals were recorded. The fan and the phone were represented by 3D objects of a fan and a phone, respectively. At the position of the third source playing a cartoon soundtrack, a TV receiver object was placed. The dimensions of the room and positions of the sources within it corresponded to the physical room dimensions and source positions. A top view of the virtual room is shown in Figure 2. The virtual camera was controllable by means of a keyboard and mouse in a way similar to computer games with first person perspective.

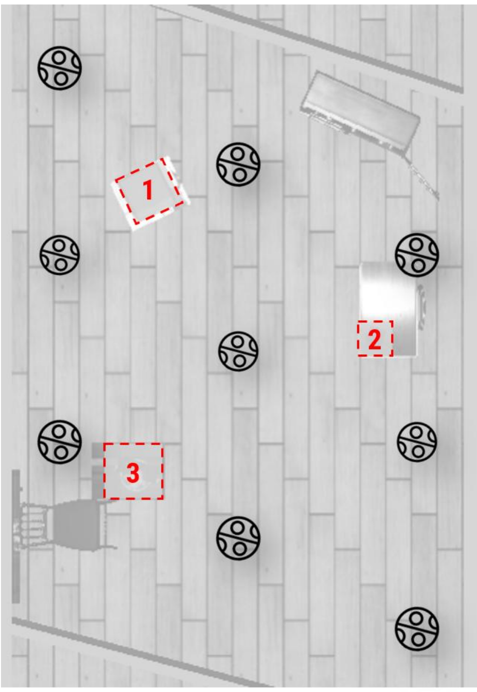

Figure 2. Top view of the virtual room with sound sources and microphone placement indications: (1) TV set, (2) phone and (3) fan.

In Section 3 a 5DoF interactive presentation system was detailed. However, in order to better control the evaluation experiment, a pre-rendered video of the 3D interactive scene was prepared where the virtual viewer and listener move on a predefined path around the room. The movement trajectory in the pre-rendered video included two translation dimensions (front-back and left-right) and one rotation dimension (yaw), necessary and sufficient for verifying the research hypothesis. By removing the interactive aspect during the MUSHRA test and using a pre-rendered cinematic one instead, we were able to ensure that all participants of the experiment experience the same stimuli. The visual component of the stimulus was rendered once and was used for all four audio stimuli described above.

4.4 Test interface

The presentation system consisted of a personal computer with a player application enabling gapless playback switching between the various audio stimuli included in the test while at the same time displaying the visual component which was common between all conditions. The sound was played back through an Audinst HUD-mini USB audio interface and Sennheiser HD 650 headphones with dedicated correction filters provided in IEM Binaural decoder. The test interface was presented to test subjects on a separate computer from the one used for stimuli presentation. Two questions were asked:

- Test 1: In a scale from 0 to 100 how natural and realistic is the acoustic localization of sound sources with respect to their position in the video?
- Test 2: In a scale from 0 to 100 how natural and smooth is the evolution of distance and position of sound objects during changing the listening point in the scene (translation and rotation)?

Additionally, participants were asked to write notes regarding the general listening impression.

5 Results

The listening tests were done with 15 trained subjects with the average age of 29.5 years (with standard deviation of 5.1). 4 subjects were female. 12 subjects had an experience in MUSHRA listening tests before. Most of the subjects were familiar with the acoustics of the room in which the test item was recorded. All of the subjects scored the Reference system over 90 in both tests, however 2 of them scored the 1OA-based systems lower than the Anchor. Therefore, the scores of those subjects were removed from statistical analysis of the results.

Figure 3 shows the absolute scores with 95% confidence intervals for Test 1 and Test 2. For both tests the Reference system performed significantly better than other assessed systems. Still, the performance of 3OA-based systems was rated as “Excellent” in the MUSHRA scale, with average scores of 79.5 for Test 1 and 79.8 for Test 2. The confidence intervals of 1OA- and 3OA-based systems are overlapping by 4-5 MUSHRA points. However, in the differential scores (Figure 4) it can be noticed that for both Tests 3OA-based system performed better than the 1OA-based one, showing statistically significant improvement.

It is noteworthy that, despite the scores of Test 1 and Test 2 of individual subjects varied significantly, the averaged scores of these Tests show high level of correlation.

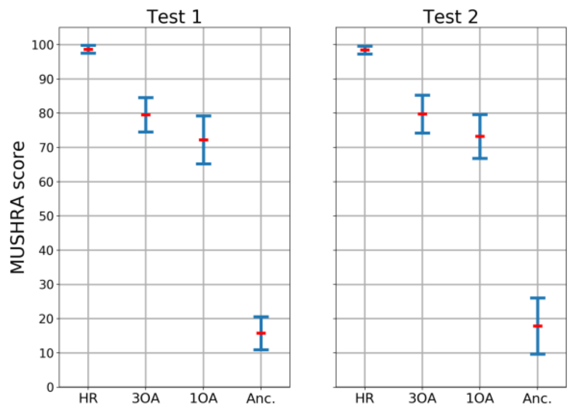

Figure 3. Absolute MUSHRA scores for Test 1 and Test 2. The 95 % confidence intervals (13 listeners) are plotted.

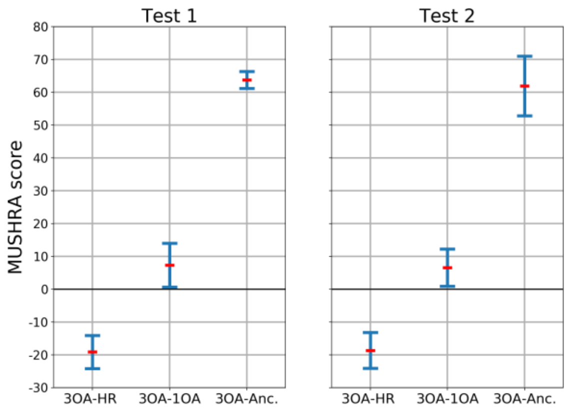

Figure 4. Differential MUSHRA scores (3OA vs other conditions) for Test 1 and Test 2. The 95 % confidence intervals (13 listeners) are plotted.

6 Discussion

As the results of MUSHRA evaluation show, the proposed method can be a viable way to interpolate simultaneous adjacent ambisonics recordings, providing an acceptable level of consistency in terms of sound source localization and perception of the translation movement within the recorded audio scene. Additionally, the experiment indicates that it is preferable to render 6DoF sound fields with high spatial resolution Ambisonics recordings (i.e. the Ambisonics order is an important factor). 1OA-based system was shown to be less precise in recreating point sources in 6DoF space, while the 3OA-based provided results comparable to object based spatialized sound.

During the test subjects also reported that:

- 3OA-based system had more convincing ambient sound than the Reference and 1OA-based systems.
- 1OA- and 3OA-based systems sound more realistic in terms of recreation of the room acoustic properties.
- 3OA-based system provides a better sense of localization and immersion of the sound over the 1OA-based system.
- Acoustic localization of the sound sources in the Reference signal is more obvious but it sounds artificially.

As it is known, individual ambisonics spherical sound fields contain spatial information regarding direction from a single position to reproduce sound transformations based on rotation movements, but it cannot reproduce variable depth or distance between listener and sound sources. In the proposed setup, the sound from neighboring sound fields is projected beyond an established arbitrary radius. Therefore, when interpolated through the present method, multiple ambisonics sound fields provide subtle additional volume and tonal balance variation, enabling the sense of depth a single sonic sphere could not provide. Additionally, this experimental work showed that it is possible to enable smooth transitions between sound fields without clear conflicting rotation issues.

Future work should focus on evaluation the efficiency of proposed method in different scenarios with varied space sizes and reverberation characteristics; sound sources timbres, frequency range and positioning; and addition of non-static sound sources.

7 Summary

In this paper a novel strategy for recording and playback of 6DoF acoustic scenes was proposed and evaluated. A recording setup made up of nine HOA microphone arrays uniformly distributed in a room were configured and used to record an acoustic scene comprising three simultaneously active sound sources. These recordings were integrated into an interactive playback system, implementing the proposed interpolation scheme, allowing for a realistic listening experience with 5DoF-navigable synthetic visual counterpart of the physical recorded room.

References

[1] Huang, Z., Li, T., Chen, W., Zhao, Y., Xing, J., LeGendre, C., & Li, H. (2018). Deep Volumetric Video From Very Sparse Multi-View Performance Capture. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 336-354).

[2] https://home.otoy.com/stream/orbx/

[3] https://realsense.intel.com/

[4] Jot, J. et al. (2017) Group Report: A spatial audio format with 6 Degrees of freedom. The Twenty-Second Annual Interactive Audio Conference - Project Bar-B-Q.

[5] Peters, N., Sen, D., Buffoni, X., Dufour, M. (2018) M44875 - On the application of multiple HOA streams for MPEG-I. International Organization for Standardization ISO/IEC JTC1/SC29/WG11 - Coding of moving pictures and audio, 124 MPEG meeting, Macau, CN.

[6] Tylka, J. G., & Choueiri, E. (2015, October). Comparison of techniques for binaural navigation of higher-order ambisonic soundfields. In Audio Engineering Society Convention 139. Audio Engineering Society.

[7] Schultz, F., & Spors, S. (2013, September). Data-based binaural synthesis including rotational and translatory head-movements. In Audio Engineering Society Conference: 52nd International Conference: Sound Field Control-Engineering and Perception. Audio Engineering Society.

[8] Noisternig, M., Sontacchi, A., Musil, T., & Holdrich, R. (2003, June). A 3D ambisonic based binaural sound reproduction system. In Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality. Audio Engineering Society.

[9] Plinge, A., Schlecht, S. J., Thiergart, O., Robotham, T., Rummukainen, O., & Habets, E. A. (2018, August). Six-Degrees-of-Freedom Binaural Audio Reproduction of First-Order ambisonics with Distance Information. In Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality. Audio Engineering Society.

[10] Tylka, J. G., & Choueiri, E. (2016, September). Soundfield Navigation using an Array of Higher-Order Ambisonics Microphones. In Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality. Audio Engineering Society.

[11] Rivas Méndez, D., Armstrong, C., Stubbs, J., Stiles, M., & Kearney, G. (2018, October). Practical Recording Techniques for Music Production with Six-Degrees of Freedom Virtual Reality. In Audio Engineering Society Convention 145. Audio Engineering Society.

[12] Zamojski, J., Makaruk, P., Januszkiewicz, L., & Zernicki, T. (2017, October). Recording, Mixing and Mastering of Audio Using a Single Microphone Array and Audio Source Separation Algorithms. In Audio Engineering Society Convention 143. Audio Engineering Society.

[13] http://www.zylia.co/zylia-zm-1-microphone.html

[14] http://www.zylia.co/zylia-ambisonics-converter.html

[15] Rumiński, A., Januszkiewicz, Ł., Lejwoda, Ż., Kuklasiński, A., Żernicki, T. (2018) M42571 - Report on Recording a Sound Scene with Multiple HOA Microphones. International Organization for Standardization ISO/IEC JTC1/SC29/WG11 - Coding of moving pictures and audio, 122 MPEG meeting, San Diego, USA.

[16] Patricio, E., Januszkiewicz, Ł., Kuklasiński, A., Rumiński, A., Żernicki, T. (2019) M46067 - Report on Recording of Test Material for 6DoF Audio, 125 MPEG meeting, Marrakech, MA.

[17] https://unity3d.com/

[18] https://cycling74.com/products/max/

[19] https://plugins.iem.at/

[20] https://developers.google.com/resonance-audio/

[21] International Telecommunication Union, "ITU-R BS.1534-3, Method for the subjective assessment of intermediate quality level of audio systems," 2015.