Production of six-degrees-of-freedom (6DoF) navigable audio using 30 Ambisonic microphones

Bartłomiej Mróz
Multimedia Systems Department
Gdańsk University of Technology
Gdańsk, Poland
bartlomiej.mroz@pg.edu.pl

Marek Kabaciński
Zylia sp. z o. o.
Poznań, Poland
marek.kabacinski@zylia.pl

Tomasz Ciotucha
Zylia sp. z o. o.
Poznań, Poland
tomasz.ciotucha@zylia.pl

Andrzej Rumiński
Zylia sp. z o. o.
Poznań, Poland
andrzej.ruminski@zylia.pl

Tomasz Żernicki
Zylia sp. z o. o.
Poznań, Poland
tomasz.zernicki@zylia.pl

***Abstract*— This paper describes a method for planning, recording, and post-production of six-degrees-of-freedom audio recorded with multiple 3rd order Ambisonic microphone arrays. The description is based on the example of recordings conducted in August 2020 with the Poznan Philharmonic Orchestra using 30 units of Zylia ZM-1S. A convenient way to prepare and organize such a big project is proposed – this involves details of stage planning, collecting needed equipment, management, and running the recording. Additionally, the placement of 2D and 360 cameras as well as the design of moving camera paths are described. Two use cases of the collected Ambisonic audio are proposed. The first one is a production of a virtual reality application using Unreal Engine 4 and Wwise. The second one uses a combination of the virtual scene from the previously mentioned application and a footage from a mobile, moving camera to create a so-called "walkthrough movie". The proposed approach for six-degrees-of-freedom Ambisonic interpolation was evaluated on listening tests. Those tests demonstrate that it is possible to recreate a navigable 3D audio taking into account Ambisonic only input and given assumptions for microphone placement.**

Keywords—virtual reality, spatial audio, Ambisonics, sound source localization, six-degrees-of-freedom

I. Introduction

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 and virtual reality (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 rotational, but not translational movements.

One of the newest topics in the audio-video industry is volumetric video. The first laboratories dedicated to this topic appeared only a few years ago [1], [2]. Volumetric video is a technique that captures a three-dimensional space, such as a location or performance. Gilbert et. al. proposed a convolutional autoencoder that enables high fidelity volumetric reconstructions of human performance to be captured from multi-view video comprising only a small set of camera views [3]. Another technique comprising a small set of camera views was proposed by Huang et. al. [4]. In their method, a deep-learning-based approach for performance capture using a passive and highly sparse multi-view capture system is presented.

Such techniques are parallel to what is called in audio as multi-point Ambisonics. There were numerous attempts at creating navigable virtual Ambisonic-encoded sound fields, also by the authors of this paper. Most notably, Tylka and Choueiri examined a linear extrapolation methods for virtual sound field navigation [5]. They also proposed a parametric method for virtual navigation within an array of Ambisonic microphones [6], [7]. In their method, they propose an interpolation between the microphones that are nearest to the desired listening position using a regularized least-squares matrix of filters. Furthermore, a detailed comparison between practical domains resulted in establishing guidelines of applicability of such parametric interpolation methods is presented [8]. Moreover, Tylka in his PhD dissertation evaluates a virtual navigation of Ambisonic sound fields, as well as proposes a method for applications with distant sources and sparsely distributed microphones [9]. Another method for 6DoF binaural audio reproduction is proposed by Plinge et al. [10]. In their method, they focus on a parametric model for first order Ambisonics (FOA) recordings. Also, Patricio et al. focused specifically on classical music recordings [11]. In this paper, an approach for utilizing non-adjacent spherical microphone arrays and audio source separation algorithms is presented. Subsequently, Rumiński et al. participated in MPEG standardization meetings, reporting on recording test material for 6DoF sound scenes [12], [13]. Consequently, small-scale and large-scale 6DoF recordings were conducted, resulting in VR applications and a "walkthrough movie" productions [14], [15].

II. Poznan Philharmonic Orchestra 6DoF Recordings

This chapter describes a recording session with Poznań Philharmonic Orchestra, which resulted in multi-point third-order Ambisonic (3OA) recordings. The recordings were further processed and incorporated into 6DoF productions.

A. Planning of the recording

As for preparations, measurements of the venue were collected (Aula of Adam Mickiewicz University in Poznań). The Aula is a very large room: it is approximately 20 m wide, 55 m long, 15 m high, and the RT60 decay time is around 2.5 seconds. Subsequently, a map of the place was designed. This was necessary in order to create a microphone grid among which the musicians were seated. Special rules had to be obeyed, due to COVID-19 pandemic restrictions –

a distance between instrumentalists had to be at least 2 m. 30 ZYLIA ZM-1S microphone arrays were placed in a regular triangular grid – in this way, the distance between the adjacent arrays was always 2 m. There are a few exceptions though – three ones in the seating area were spaced about 4 m. Also, two microphones on the stage are a bit closer to their neighbor because of the physical limitations of the venue and instrument placement. Last layer of the map was the cable routing. The venue is big, therefore it was crucial to have the cables long enough. USB extenders, USB hubs and 3,5 mm jack extenders were used for such long connections of multiple ZYLIA ZM-1S microphone arrays.

Last thing to consider is the preparation of computers for the recording. Having a backup computer is obligatory. Both recording systems should have a lot of free space available. One minute of the recording from single ZYLIA ZM-1S is 172,8 MB (20 channels with 48 kHz at 24 bits).

B. *The recording process*

The microphones were placed with as much accuracy as possible in accordance to the prepared map. As mentioned earlier, some microphones had to be placed slightly different because of the physical limitations. In case of the three microphones in the seating area, such distances between them were chosen to capture more space in front of the stage. After the recordings were completed, measurements of the exact microphone placement were collected, as shown in Fig. 1.. The venue where the recording took place is shown in Fig. 2..

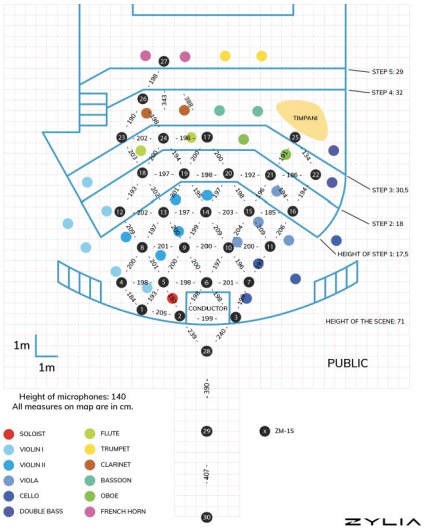

Fig. 1. The plan of the recording venue, including the positions of musicians and microphones

C. *Post-production of recorded material*

After the conversion from A-Format to B-Format of Ambisonics, the recordings were processed by the ZYLIA 6DoF rendering algorithm (as described in the section III). The algorithm parameters were obtained in a subjective manner.

D. *Virtual-reality application*

According to the shape of the stage and the measurements, a 3D graphical representation of the venue was prepared. This was used as a visual component of the VR application developed in Unreal Engine 4. The game engine was combined with the Wwise version of the ZYLIA 6DoF renderer, which served as the audio component. Models of microphones and musicians were placed on the virtual scene accordingly to the measurements performed after the recording. In this way, the virtual representation mirrored the real scene in a very high detail. Such an application is an immersive way to interactively experience the recorded materials, as it can be viewed through a VR headset. It is also worth noting that the application is already freely available in the Oculus App Lab [16].

Fig. 2. Picture of the recording event

E. *A "walkthrough movie"*

During the recording session, a mobile 2D camera has been used to dynamically change position while recording, thus creating a recorded path. The path of this camera was designed beforehand, and it included the stage and the seating area, as to recreate both musicians and audience perspective. This path was recreated on the virtual stage using a virtual camera in the Unreal Engine 4. The two recorded paths – real and virtual – served as building blocks of the walkthrough movie in which the two realities are connected. Such an approach is not interactive, but it showcases a novel way of storytelling – with the use of real-venue-recorded volumetric audio over a regular 2D video material.

III. ALGORITHM DETAILS

The proposed algorithm is described by Patricio et al. in [14]. The presented algorithm uses the distance attenuation technique. The interpolated signal is equal to the sum of all higher-order Ambisonic (HOA) signals presented on scene multiplied by an attenuation coefficient. The final equation for p HOA-component of interpolated signal $x(n)$ is as follows:

$$x_p(n) = \sum_{m=1}^{M} a_p(d_m)\, y_{m,p}(n) \qquad (1)$$

Where $a_p(d)$ describes attenuation function, $y_{m,p}(n)$ represents the p-th HOA-component for m-th microphone, and d_m is the distance of the interpolation point to the m-th input signal. The scaling function $a_p(d)$ applies a gradual attenuation of contributions of HOA signals present in the scene. Additionally, it applies the re-balancing of omnidirectional component (p=0) and the other components of HOA signal. Also, this function is responsible for the biggest contribution of the closest microphone. The equation for $a_p(d)$ is:

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

Where $l(d)$ and $k_p(d)$ are linear functions depending on the distance between listening point and m-th microphone. In these functions there are 4 parameters which have impact on the overall shape. They have to be chosen for given microphone grid; it also must be ensured that there is no gap in the interpolation. In the performed experiment the applied volume threshold and range is equal to 0.9, whereas HOA threshold Is 0.9 and HOA range is 1.3. These parameters were chosen by a sound engineer experienced in Ambisonic and 6DoF rendering, and are in line with previous observations in [14] and [15]. Furthermore, the values used in the listening test ensure that the closest microphones have the biggest contribution to the interpolated signal.

IV. Subjective Evaluation

A. Methodology

The algorithm was evaluated on the sound source localization performance. The test was conducted in an acoustically controlled environment. The test was performed with the use of Oculus Quest head-mounted display connected to a PC via Oculus Link feature. The visual scene presented to participants did not include any objects in order to avoid the bias from the visual cues – the scene presented simple blue space with sky-like cloudy ceiling. The sound was played back via open-back high-fidelity headphones – namely Sennheiser HD650 – connected to a Focusrite audio interface. The use of a head-mounted display allowed for head tracking and thus for dynamic binauralization, but it also allowed the participants to immerse into the sound scene, separate from the environment and focus more on the sound source localization task. Furthermore, this approach allowed for some kind of gamification – participants were pointing the sound source with a laser-like pointer, and by pressing the trigger, the red cross appeared on the screen. Most participants expressed a positive and enthusiastic feedback about the procedure. The test environment is presented in Fig. 3..

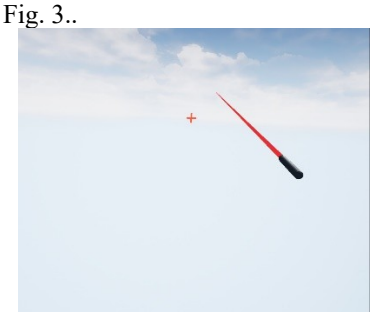

Fig. 3. The view inside the Oculus app – the laser-like pointer and the cross marking participant's perveived localization are visible

B. Audio stimuli

In order to have a full control of the environment, a virtual sound scene was created. Also, in order to achieve a test that is similar in terms of assumptions and conditions to the actual recording situations, a concert hall with virtually placed instruments was designed. It was done with the use of anechoic orchestral recordings done by Pätynen et al. [17]. In order to achieve a high realism of the virtual sound scene, the directivity patterns were included, derived from [18]. The auralization was designed with the use of IEM Plug-in Suite [19] in the 3rd order of Ambisonics. Each of the instruments' recordings were on separate tracks, with *DirectivityShaper* and *RoomEncoder* plugins inserted. These plugins allow for adding a directivity pattern to a track, as well as position the sound source in a virtual room, thus allowing for creating a virtual acoustic scenes. The sound scene was situated in a shoebox-shaped room with the dimensions of 6 m width, 8 m length and 5 m height, imitating a moderate concert venue. The reverberation from early reflections was not too long, but it was further extended with *FDNReverb* plugin in order to match the longer reverberation of a concert hall. The instruments were placed as to reproduce Mozart-era orchestral placement. Then, the position of a listener (or: a virtual microphone) was rendered in multiple positions. The rendered tracks were 16-channel 3rd order Ambisonic signals. There were 27 positions in total, and they were designed as a triangular grid with the distance of 1 m between each microphone. Furthermore, the three positions of virtual microphones were selected for the 6-degrees-of-freedom interpolation: the position of a microphone no. 5 (as a "first row of seats" situation), the position of a microphone no. 16 (as a "the middle of the orchestra" situation), and the position of a microphone no. 24 (as a "boundary, near walls" situation). With this approach, the interpolation was occurring on a radius of 2 m, which corresponds to the distance between microphones in the Poznan Philharmonic recordings. The plan of the virtual sound scene is presented in Fig. 4..

Finally, the rendered audio stimuli were dynamically binauralized with the IEM's *BinauralDecoder* plugin. This plugin employs Magnitude Least-Squares binauralization method [20], [21], and uses HRTF dataset measured on Neumann KU 100 dummy head [22]. No headphone equalization was applied.

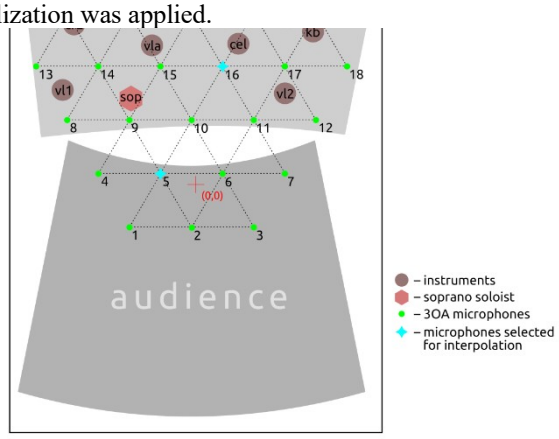

Fig. 4. The plan of the virtual sound scene. The sides of triangles are 1 m long

C. Test procedure

The participants were asked to determine the location of the soprano soloist. There were three test cases and one reference case for each interpolated microphone position: the interpolation in the 1st order of Ambisonics, the interpolation in the 3rd order of Ambisonics, the audio of the 3rd order virtual Ambisonic microphone, and the soprano voice only encoded in 3rd order of Ambisonics as the reference case. This gives 12 stimuli in total, which were presented to the participants in a random order. Furthermore, the initial horizontal rotation of the sound scene was also randomized. There were 20 participants in total, males and females with age ranging from 25 to 50 years – members of Zylia company and members of academic staff of Multimedia Systems Department of Gdańsk University of Technology.

D. Results

Fig. 5. presents the results for different microphone positions. The results are presented with 95% confidence intervals for standard normal distribution. The bars present participants' errors of sound source localization, expressed in degrees of azimuth. No vertical component was taken into account in the error measurement, since the stimuli and listener positions were on the same height, which places them on the same horizontal plane.

For the microphone position no. 5, participants' errors in sound source localization for both interpolations – 1st and 3rd Ambisonic order (1OA, 3OA) – are nearly the same. The mean localization error is 26° and 22°, respectively. The lower and upper boundaries of confidence intervals are 19° and 33° for 1OA interpolation, and 16° and 28° for 3OA interpolation. As for the reference cases, the results are virtually the same, regardless of the orchestra presence in the sound scene. The mean values of localization error for both non-interpolated signals are 6°. Also, the upper and lower boundaries of 95% confidence intervals are the same – the boundaries are 5°-8°.

In the case of microphone position no. 16, the difference between localization error in the cases of 1OA and 3OA interpolation is slightly more pronounced. The mean values are very similar – 27° and 24°, respectively, but boundaries of confidence intervals are 13°-41° for 1OA and 18°-31° for 3OA interpolation. As for the references, again the results are very close to each other: both mean values for non-interpolated cases are 7°, and both boundaries of confidence intervals are 5°-9°.

In the last microphone position no. 24, the sound source localization error is higher for each stimuli. The mean value for 1OA interpolation is 35°, whereas for 3OA interpolation it is 46°. The upper and lower boundaries of 95% confidence intervals are 17°-52° and 25-67°, respectively. Regarding the reference cases, the mean value for 3OA virtual microphone with soprano and orchestra is 25°, and for 3OA virtual microphone with soprano only the mean value is 17°. The boundaries of confidence intervals are 17°-34° and 7°-28°, respectively.

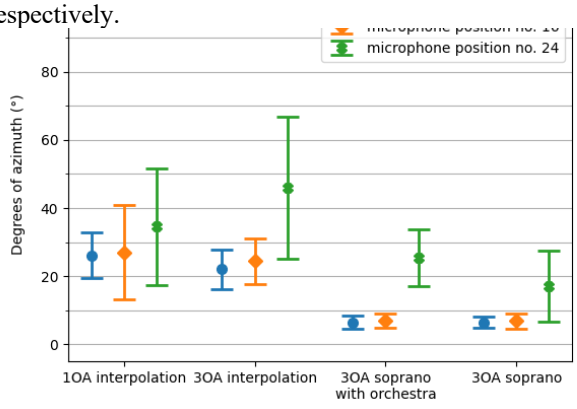

Fig. 5. The results of localization perception task

E. Discussion

The results show an interesting consequence of the interpolation algorithm. Since essentially it is a mix of surrounding Ambisonic microphones, the perceived width of a sound source becomes as wide as width between nearest microphones. Consequently, widening of the sound source yields a resolution loss, which is somewhat counterintuitive when increasing the Ambisonic order. This is especially pronounced in the "first row of seats" scenario – position of microphone no. 5. When it comes to the "middle of an orchestra" scenario – position of microphone no. 16 – the obfuscation resulting from stronger presence of other instruments in the sound scene adds to the loss of spatial resolution between 3OA and 1OA, making the difference between the Ambisonic orders more pronounced. However, in both 5 and 16 microphone positions the error and its spread is rather small, being roughly 20° higher than the error of the reference cases. Apparently, there are two references – 3OA soprano with the orchestra and the soprano voice only. Since the whole audio scene was virtually designed, the stimuli with the solo soprano voice seemed as a reasonable differentiation between other cases with orchestra present. However, the soprano voice occurred as perceptually distinctive enough even in the "middle of the orchestra" scenario. Interestingly, for the acoustically most difficult situation – near the walls of the concert hall, position of microphone no. 24 – each case was prone to a high localization error; even the case with the soprano voice only. Nevertheless, the localization error of 6DoF interpolation cases was not significantly higher than the reference cases – for both 1OA and 3OA interpolations. This suggests that acoustic reflections derived from 6DoF interpolations still provide localization cues and are not blurring the sound scene too much. However, the localizability in such case requires a more, in-depth look, preferably with a higher-precision acoustic model or a recording from a non-virtual venue. It is also worth mentioning that previous tests of this algorithm conducted by the authors included a visual input, which comprised localization cues – the image from the camera or a virtual concert hall with musicians. Those tests concluded with a good efficiency and usefulness of the 6DoF algorithm, as well as decent audio quality and audio-video correlation. The localizability of sound sources increases dramatically with visualization, and the 6DoF algorithm is designed mostly with virtual reality applications in mind. Furthermore, this test consisted of very symmetric case – with the interpolation over equally distant microphones. Such a synthetic, symmetric situation rarely occurs in a VR environment; the listener position and path in a six-degrees-of-freedom movement is much more irregular. As a result, the listener's position is always closer to some microphones. The relationship of distance to the closest microphone from the interpolated area, localizability of sound sources, and presence of localization cues will be examined in detail in the future experiments by the authors.

F. Comparison to small venue recordings

It is important to confront these results to previous work of the authors. In the article [14], the recording comprised 9 microphone arrays and 3 sound sources placed in a short-reverberated room. The testing procedure included a one audio-visual path presented to the participants with 4 different stimuli: recorded signal downmixed to 0th Ambisonic order (0OA), 1OA, 3OA and an object-based representation of the same scene. The results have shown that on the absolute scores, 6DoF rendering based on 1OA and 3OA had comparable scores to the object-based reference representation. Moreover, the differential tests showed that 6DoF rendering based on 3OA outperform the 1OA significantly. This is an expected result due to the highest spatial resolution used in 3OA vs. 1OA. This experiment also demonstrates that the proposed 6DoF rendering and

interpolation can be used successfully for small and large venues in terms of the number of sound sources.

V. Summary

This paper presents a detailed procedure for large-scale multipoint Ambisonic recording, with further extension to creating navigable, six-degrees-of-freedom virtual reality environments. This paper also investigates the performance of the proposed interpolation algorithm in a very synthetic conditions, providing important information regarding localization performance in different acoustic conditions. This study shows a promising utilization of the proposed production workflow; one that enables large events and venues to provide a more immersive, engaging experience for their audiences. The capacity to offer such experiences has been strongly highlighted by recent, pandemic-impeded time. Moreover, due to the ongoing advancements in delivering personalized, user-oriented content, being able to provide a fully interactive high-fidelity entertainment might soon become essential.

Acknowledgments

This work was supported by National Centre for Research and Development (NCBiR), Poland, "Fast Track" programme (project POIR 01.01.01-00-1278/17).

Authors would like to thank musicians of Poznan Philharmonic Orchestra, its director Wojciech Nentwig, Maestro Łukasz Borowicz and soprano Agnieszka Adamczak for cooperation on these recordings.

References

[1] https://www.hhi.fraunhofer.de/en/press-media/news/2018/fraunhofer-hhi-put-into-operation-first-volumetric-video-studio-on-the-european-mainland.html (accessed 21.06.2021)

[2] http://www.vvow.eu/ (accessed 21.06.2021)

[3] A. Gilbert, M. Volino, J. Collomosse, and A. Hilton, (2018) “Volumetric Performance Capture from Minimal Camera Viewpoints”, in: Ferrari V., Hebert M., Sminchisescu C., Weiss Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11215. Springer, Cham. https://doi.org/10.1007/978-3-030-01252-6_35

[4] Z. Huang, T. Li, W. Chen, Y. Zhao, J. Xing, C. LeGendre, L. Luo, C. Ma, and H. Li, (2018) “Deep Volumetric Video From Very Sparse Multi-view Performance Capture”, in: Ferrari V., Hebert M., Sminchisescu C., Weiss Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11220. Springer, Cham. https://doi.org/10.1007/978-3-030-01270-0_21

[5] J. G. Tylka and E. Y. Choueiri, “Performance of Linear Extrapolation Methods for Virtual Sound Field Navigation”, J. Audio Eng. Soc., March 2019.

[6] J. G. Tylka and E. Y. Choueiri, “Soundfield Navigation using an Array of Higher-Order Ambisonics Microphones”, in Audio Engineering Society Conference: 2016 International Conference: Audio for Virtual and Augmented Reality, September 2016.

[7] J. G. Tylka and E. Y. Choueiri, “Fundamentals of a Parametric Method for Virtual Navigation Within an Array of Ambisonics Microphones”, J. Audio Eng. Soc., March 2020.

[8] J. G. Tylka and E. Y. Choueiri, “Domains of Practical Applicability for Parametric Interpolation Methods for Virtual Sound Field Navigation”, J. Audio Eng. Soc., November 2019.

[9] J. G. Tylka, “Virtual Navigation of Ambisonics-Encoded Sound Fields Containing Near-Field Sources”, Doctoral dissertation, Princeton University, June 2019.

[10] A. Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. A. Habets (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.

[11] E. Patricio, M. Skrok, and T. Zernicki, "Recording and Mixing of Classical Music Using Non-Adjacent Spherical Microphone Arrays and Audio Source Separation Algorithms," Engineering Brief 525, (2019 October.)

[12] A. Rumiński, Ł. Januszkiewicz, Ż. Lejwoda, A. Kuklasiński, and T. Żernicki, (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.

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

[14] E. Patricio, A. Ruminski, A. Kuklasinski, L. Januszkiewicz, and T. Zernicki, "Toward Six Degrees of Freedom Audio Recording and Playback Using Multiple Ambisonics Sound Fields," Paper 10141, (2019 March.)

[15] T. Ciotucha, A. Ruminski, T. Zernicki, and B. Mróz, "Evaluation of Six Degrees of Freedom 3D Audio Orchestra Recording and Playback using multi-point Ambisonics interpolation," Paper 10459, (2021 May.)

[16] https://www.oculus.com/experiences/quest/3901710823174931/ (accessed 21.06.2021)

[17] J. Pätynen, V. Pulkki, and T. Lokki, "Anechoic recording system for symphony orchestra," Acta Acustica united with Acustica, vol. 94, nr. 6, pp. 856-865, November/December 2008

[18] J. Pätynen and T. Lokki, (2010). Directivities of symphony orchestra instruments. Acta Acustica united with Acustica, 96(1), 138-167. https://doi.org/10.3813/AAA.918265

[19] https://plugins.iem.at/ (accessed 10.08.2021)

[20] C. Schoerkhuber, M. Zaunschirm, and R. Hoeldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares", Fortschritte der Akustik, DAGA, 2018

[21] F. Zotter and M. Frank, (2019). “Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality”, pp. 89-90. 10.1007/978-3-030-17207-7.

[22] B. Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100", Proceedings of the 40th Italian (AIA) Annual Conference on Acoustics and the 39th German Annual Conference on Acoustics (DAGA) Conference on Acoustics. 2013. http://audiogroup.web.th-koeln.de/ku100hrir.html