CHALMERS
UNIVERSITY OF TECHNOLOGY

Case Study of Equipping a High-Fidelity 360 Camera with a 4th-Order Equatorial Ambisonic Microphone Array

Audio Engineering Society

Convention Express Paper 81

Presented at the 154th Convention
2023 May 13–15, Espoo, Helsinki, Finland

Case Study of Equipping a High-Fidelity 360 Camera with a 4th-Order Equatorial Ambisonic Microphone Array

Jens Ahrens[1] and Karolina Jaruszewska[2]

[1]*Chalmers University of Technology, 412 96 Gothenburg, Sweden*
[2]*KFB Acoustics, 51-502 Wrocław, Poland*

Correspondence should be addressed to Jens Ahrens (jens.ahrens@chalmers.se)

ABSTRACT

We present a case study of a commercial 360 camera that we equipped with an experimental 9-channel equatorial microphone array that uses the camera as the baffle and thereby becomes invisible for the video capture. The microphone array produces a 4th-order ambisonic audio recording that we render binaurally. The setup as a whole produces 360 audio-visual content that does not require any manual post-processing. The camera model that we chose has a perfectly circular horizontal cross-section, but its overall shape departs from a sphere. We demonstrate that approximating this baffle as a sphere in the signal processing causes negligible deviation of the resulting ear signals. The standard equatorial array processing approaches can therefore be employed, and costly calibration measurements are avoided. All resources are provided for download.

1 Introduction

360 cameras can be considered the video equivalent of compact microphone arrays. When the captured video is played back on virtual reality goggles, the user sees the captured scene from a first-person perspective (also referred to as point of view (POV)). Similarly, when the signals from the microphone array are rendered binaurally, then the user hears the captured scene from a first-person perspective. Head tracking is commonly available for the video playback so that the user can "look around" in the scene. Head tracking in the audio playback is also state-of-the-art. The video perspective and the audio perspective therefore align even if the user rotates their head.

The spatial audio capabilities of commercial 360 cameras are often very limited. This makes it essentially impossible to record visual content from a first-person perspective with matching high-fidelity audio without having external audio equipment in the field of view of the camera and without manual post-processing. Promising concepts for overcoming this were presented in the scientific literature in the recent years. We identified a promising combination of commercial 360 camera and experimental spatial audio capture method that we evaluate in this paper.

Applications that the tested technology may be useful for include capture of all scenarios that are intended to be experienced from a first-person view and where the use of spot microphones or boom microphones is either not possible or not desired. Examples for this are sports events, events in public space, and music and dance performances.

2 Overview of the Relevant Spatial Audio Capture Methods

A variety of methods for capturing spatial audio content have been proposed in the literature. Most popular are arrays that provide a representation of the captured sound scene that is compatible with the ambisonic framework. We also focus on this setting in the present work because of its flexibility, most notably the possibility of employing arbitrary rotational head tracking during binaural playback.

One of the first capture solutions that became available commercially was the Soundfield microphone (or B-format microphone) that provides 1st ambisonic order [1]. Spherical microphone arrays (SMAs) with rigid baffles constitute the subsequent evolutionary step. While there is no theoretical limit for the ambisonic order of SMAs, most arrays do not provide an order higher than approx. 4-7. This does not constitute a fundamental limitation as there are only selected scenarios where an even higher order may be desirable [2, 3] – if this is possible to be implemented in practise at all.

A variety of variations of the SMA concept exist. Microphone arrays that use non-spherical baffles typically use numerical solutions for obtaining an ambisonic representation of the microphone signals [4, 5, 6, 7, 8, 9, 10]. Measurements of the array transfer functions (ATFs) for free-field sound incidence from different directions are usually required by these methods, which are resource-intensive to obtain. This makes such methods less attractive although they may exhibit considerable potential in the present use case [11].

SMAs do not require ATFs to be measured because the acoustic properties of the baffle can be described analytically. The SMA concept has been combined with 360 video capture in a single housing by some manufacturers. A downside with SMAs is the circumstance that the number of microphones that they require is high and that those microphones need to be distributed over the entire surface of the baffle. Nth-order capture requires at least $(N+1)^2$ microphones, which is 64 microphones for 7th order.

Recently, equatorial microphone arrays (EMAs) were presented [12], which employ microphones along the equator of a spherical baffle. As with SMAs, the acoustic properties of the baffle are described analytically. This reduces the required number of microphones to $2N+1$ for Nth order capture (i.e., 15 microphones for 7th order). This is not only less expensive to produce, but also the placement of the microphones is easier to realize. The prize to pay for this reduction of the required number of microphones is the circumstance that EMAs capture a horizontal projection of the impinging sound field. Interaural elevation cues are still correct [10], but monaural cues are lost. It is unclear at this point in how far this is actually a limitation because it has not been proven that practical SMAs are able to preserve monaural elevation cues. Refer to [1] for a demonstration of an EMA.

3 Prototype Design

Higher-order ambisonic capture requires a baffle of sufficient size. Small 360 cameras like the GoPro Max or similar are so small that, when used as a baffle, the spatial resolution is very limited at low frequencies [11]. We chose to employ an Insta360 Pro camera, which has a radius of 71 mm and is therefore only slightly smaller than a typical human head so that all perceptually relevant information can be extracted at low frequencies [13, 14]. Another advantage of the Insta360 Pro is the fact that the shape of its housing deviates only slightly from spherical so that there is a chance that analytical solutions can be employed for the signal processing. Given that it is very difficult – if possible at all – to distribute microphones evenly over the surface area of the camera housing, we opted for an equatorial layout. Refer to Fig. 1 for a photograph of the prototype.

The Insta360 Pro has a housing that has a perfectly circular equator along which 6 cameras are distributed equi-angularly. Given that the microphones of the EMA that we envision need to be positioned equi-angularly, too, as well as such that they do not cover any of the cameras, a total number of microphones that is an integer multiple of 3 seems most convenient. We chose to employ 9 microphones, which allows for capture

[1] https://youtu.be/95qDd13pVVY

Fig. 1: Photograph of the prototype

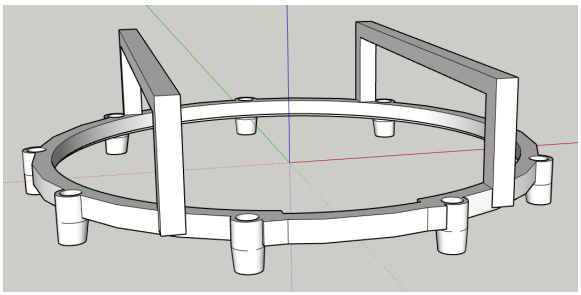

Fig. 2: 3D model of the microphone mount

with 4th ambisonic order. This appears to be a useful comprise between perceptual quality and practical convenience [2].

We used Røde Lavalier GO omnidirectional lapel microphones together with Røde VXLR+ adapters that allow for using a standard multichannel audio interface (Antelope Orion 32) and 48 V phantom power supply. The microphone holder depicted in Fig. 2 was custom designed and 3D printed. The design is available for download. Refer to Sec. 7 for further details.

4 Signal Processing

As we demonstrate in Sec. 5, the prototype can be used exactly like an EMA with a perfectly spherical baffle. We used the MATLAB implementation of the standard EMA processing that is provided in [15] to encode the raw microphone signals linearly into 4th-order ambisonics and to produce the binaural renderings. The ambisonic representation that we compute from the microphone signals is compatible with standard ambisonic tools like SPARTA[2] and the IEM Plugin Suite[3] so that the reader can experience the content that we provide in Sec. 7 with the head tracking system of their choice.

Measurements showed that the sensitivities of the employed microphones varied by up to 4 dB which we compensated for in a frequency-independent manner. The binaural output was usable also without the calibration but head rotations or moving source produced a perception that was less smooth compared to the case with calibration.

Order-limited ambisonic recordings/renderings usually have to equalized because primarily the order truncation and the spatial aliasing alter the spectral balance of the signals. A variety of equalization methods has been proposed in the literature, most of which showed to be similarly effective [2] so that there is not the one and only candidate. For convenience, we provide only non-equalized data in this paper. The qualization that we employed in our demonstration recordings is explained on the project website provided in Sec. 7.

Fig. 3 summarizes the processing pipeline used in this paper. The binaural rendering is performed according to [15, Eq. (16)].

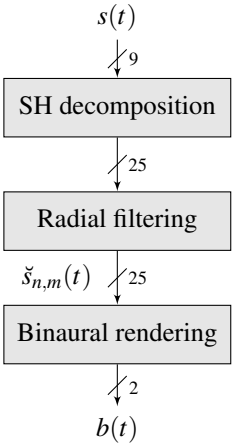

Fig. 3: Block diagram of the signal processing pipeline. $s(t)$ are microphone signals, $\breve{s}_{n,m}(t)$ are the ambisonic signals ($(N+1)^2 = 25$ channels), and $b(t)$ is the binaural output.

[2] https://leomccormack.github.io/sparta-site/
[3] https://plugins.iem.at/

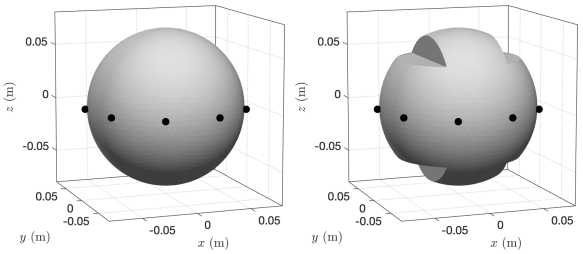

Fig. 4: Simulated baffle shapes. Left: Sphere. Right: Insta360 Pro. The black dots denote the microphone positions.

5 Evaluation

The ultimate way of evaluating a capture technology is, of course, listening to it. We provide links to a variety of demos of our prototype (as well as to other resources) in Sec. 7 for this purpose. In this section, we provide numerical data on the array's performance.

A significant amount of literature is available on the performance of SMAs and EMAs, for example, [2, 12, 13, 14, 16]. We therefore focus on the peculiarity of the present setup, which is the fact that the shape of the Insta360 Pro camera is almost spherical, but no perfectly as it is the case in the SMA and EMA literature. We analyse what the consequences of this deviation are.

We do this by comparing the performance of a perfectly spherical baffle vs. the present almost-spherical baffle of the exact same size ($R = 71$ mm), with the exact same microphone placement, and the exact same signal processing that was designed for ideal EMAs. We used the *mesh2hrtf* implementation of the boundary element method (BEM) from [17, 18] to simulate the ATFs for both cases for different sound incidence angles in the frequency range from 50 Hz to 16 kHz. The baffle shapes are depicted in Fig. 4.

The ground truth binaural transfer function (BTF) of the processing pipeline for a plane wave that impinges on the arrays under free-field conditions is simply the HRTF (of the HRTF set that is employed for the rendering) for the given incidence direction. We use this scenario in the remainder of this section. The HRTF set we use is of a Neumann KU100 dummy head [19].

We chose to present only data for horizontal sound incidence in this paper. Our prototype turned out to exhibit properties for non-horizontal sound incidence that are identical to those of other equatorial arrangements [10, 12]. The demonstration recordings that are provided in Sec. 7 comprise non-horizontal sound incidence, too.

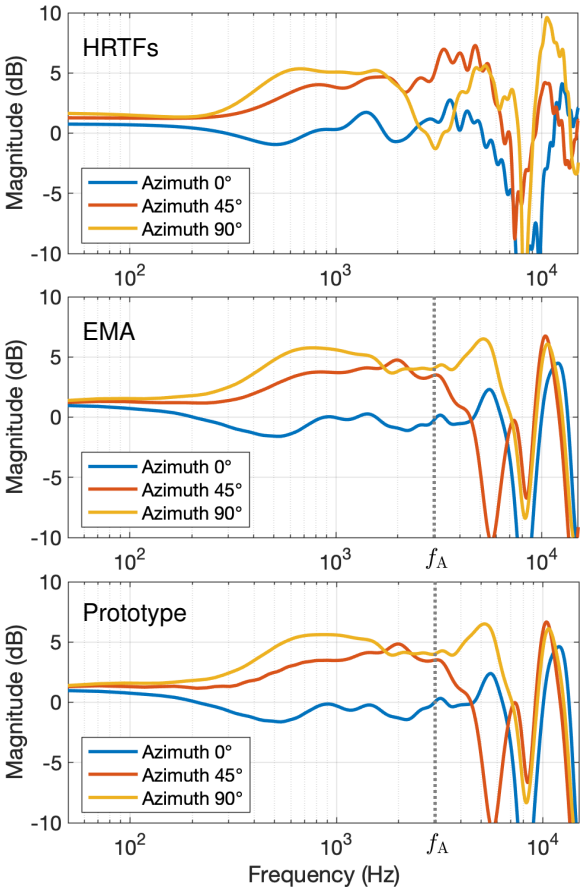

Fig. 5: Left ear signals for horizontally propagating plane waves with different incidence azimuths. Top: HRTFs. Middle: BTFs of an ideal EMA. Bottom: BTFs of the prototype (EMA with a baffle that deviates from a sphere).

Fig. 5 (top) depicts ground truth example data (i.e. HRTFs) for horizontal sound incidence. Fig. 5 (middle) depicts the BTFs of the ideal EMA from Fig. 4 (left) with a perfectly spherical baffle for the same scenario, and Fig. 5 (bottom) depicts the BTFs of the prototype (Fig. 4 (right)).

The spatial aliasing frequency f_A is approx. 3 kHz for the present setup ($N = 2\pi f_A / c \cdot R$). Below f_A, even

the BTFs of the ideal EMA deviate slightly from the HRTFs. We have no ultimate explanation for this, but we assume that this is due to the combination of order limitation and choice of parameters of the BEM simulation. Using an analytical model for the baffle and a higher order produces lower deviations in this frequency range [12, Fig. 5].

What is most important in the present context is the fact that the BTFs of the prototype and the BTFs of the ideal EMA are almost identical. We found slightly larger deviations above f_A for non-horizontal sound incidence, which we do not show here for convenience as this frequency range is difficult to interpret in terms of perception (Many times, large numerical deviations produce small or no perceptual difference at all.). We provide audio examples of the data from Fig. 5 as well as for simulations of non-horizontal sound incidence on the project website.

Note that the geometrical differences between the ideal EMA and the prototype are in the order of centimeters, which corresponds to the wavelength of sound of a frequency of several kHz. It is not surprising that the acoustic responses of the two baffles are very similar.

The BTFs of both the ideal EMA and the prototype exhibit slightly too low magnitude above f_A, which is typical for order-limited signals and can be equalized [2].

Our main conclusion from the above evaluation is that there seems no advantage in the present context in applying any of the methods that were mentioned in Sec. 2 that take deviations of the array baffle from a sphere into account explicitly. In other words, the reader may experiment with their own (equatorial) microphone placements and can rely on not requiring measured ATFs of their setup.

6 Other Remarks

Our experiment confirms that all required technological solutions are available conceptually. A good amount of work appears to still be required for being able to deploy the concepts in practise. The heat production inside camera arrays requires most high-fidelity systems to comprise a fan that produces a considerable amount of noise. Most camera models – like the one that we used – allow for switching off the fan for a limited duration in the order of minutes for capturing content.

Fig. 6: Screenshot from one of our demo videos. The dark object at the top is the ring of the microphone mount that is visible for the cameras. This can be avoided with a more elaborate design of the mount.

Live streaming, for example of sports events, would be an interesting application. Yet, this requires real-time stitching of the individual camera streams, which is a computationally expensive task. This seems to be producing so much heat that even the camera model that we tested does not allow for switching off the fan when real-time stitching is being performed. Similarly, the data rates that are required for transmitting the individual camera streams so that stitching can be performed externally seem to be a challenge.

Last but not least, the infrastructure for publishing content is very limited. Online 360 video players often support only low orders, which are difficult to record with high fidelity (YouTube supports only 1st ambisonic order, for example). An exception to this limitation is HOAST[4], which supports up to 4th order.

7 Resources

This section summarizes the resources that accompany the present paper:

- Project website:

  `https://github.com/AppliedAcousticsChalmers/ambisonics-for-insta360-pro`

- Audio signal processing:

  `https://github.com/AppliedAcousticsChalmers/ambisonic-encoding`

[4] `https://hoast.iem.at/`

Visit the project website to find a variety of demonstrations of the prototype. Fig. 6 depicts a screenshot from one of our demo videos.

8 Conclusions

The prototype that we presented demonstrates that the technology for joint 360 audio-visual content capture with high fidelity is available. The equatorial microphone placement that we chose allows for employing the established analytical approaches for the signal processing despite the fact that the camera's housing, which serves as the microphone-array baffle, departs from a perfect sphere. There is no advantage in employing any of the previously published approaches that use costly calibration measurements in order to account for non-spherical baffles. A classical spherical microphone array approach is very difficult to apply with the camera model that we chose because the microphone placement is very challenging.

A significant practical problem that persists is the heat production in the camera, which requires a fan to be cooling the hardware. Also the availability of compatible playback tools for easy deployment of the content keeps being limited.

9 Disclosure

The authors have no relation, neither direct nor indirect, to the manufacturer of the Insta360 Pro camera that was used in the experiments.

References

[1] Gerzon, M., "The design of precisely coincident microphone arrays for stereo and surround sound," in *50th Conv. of the AES*, 1975.

[2] Lübeck, T., Helmholz, H., Arend, J. M., Pörschmann, C., and Ahrens, J., "Perceptual Evaluation of Mitigation Approaches of Impairments due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data," *JAES*, 68(6), pp. 428–440, 2020.

[3] Lübeck, T., Arend, J. M., and Pörschmann, C., "Binaural reproduction of dummy head and spherical microphone array data—A perceptual study on the minimum required spatial resolution," *The Journal of the Acoustical Society of America*, 151(1), pp. 467–483, 2022.

[4] Moreau, S., Daniel, J., and Bertet, S., "3D sound field recording with higher order ambisonics - objective measurements and validation of a 4th order spherical microphone," in *120th Convention of the AES*, Paris, France, 2006.

[5] Tourbabin, V. and Rafaely, B., "Direction of Arrival Estimation Using Microphone Array Processing for Moving Humanoid Robots," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11), pp. 2046–2058, 2015, doi:10.1109/TASLP.2015.2464671.

[6] Zotkin, D. N., Gumerov, N. A., and Duraiswami, R., "Incident field recovery for an arbitrary-shaped scatterer," in *IEEE ICASSP*, pp. 451–455, 2017, doi:10.1109/ICASSP.2017.7952196.

[7] Politis, A. and Gamper, H., "Comparing modeled and measurement-based spherical harmonic encoding filters for spherical microphone arrays," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 224–228, 2017, doi:10.1109/WASPAA.2017.8170028.

[8] Berge, S., "Acoustically Hard 2D Arrays for 3D HOA," in *Int. Conf. on Immersive and Interactive Audio*, AES, Redmond, WA, USA, 2019.

[9] McCormack, L., Politis, A., Gonzalez, R., Lokki, T., and Pulkki, V., "Parametric Ambisonic Encoding of Arbitrary Microphone Arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–14, 2022, doi:10.1109/TASLP.2022.3182857.

[10] Ahrens, J., Helmholz, H., Alon, D. L., and Amengual Garí, S. V., "Spherical Harmonic Decomposition of a Sound Field Using Microphones on a Circumferential Contour Around a Non-Spherical Baffle," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, pp. 3110–3119, 2022.

[11] Ahrens, J. and Hu, Z., "Evaluation of Non-Spherical Scattering Bodies for Ambisonic Microphone Arrays," in *Proc. of the AES Int. Conf. AVAR*, Redmond, WA, USA, 2022.

[12] Ahrens, J., Helmholz, H., Alon, D., and Amengual Garí, S. V., "Spherical Harmonic Decomposition of a Sound Field Based on Observations Along the Equator of a Rigid Spherical Scatterer," *J. Acoust. Soc. Am.*, (150), 2021.

[13] Bernschütz, B., "Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording," PhD thesis, Technische Universität Berlin, 2016.

[14] Ahrens, J. and Andersson, C., "Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre," *Journal of the Acoustical Society of America*, 145(April), pp. 2783–2794, 2019, doi:10.1121/1.5096164.

[15] Ahrens, J., "Ambisonic Encoding of Signals From Equatorial Microphone Arrays," Technical note v. 1, Chalmers University of Technology, 2022.

[16] Zaunschirm, M., Schörkhuber, C., and Höldrich, R., "Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *The Journal of the Acoustical Society of America*, 143(6), pp. 3616–3627, 2018.

[17] Ziegelwanger, H., Kreuzer, W., and Majdak, P., "Mesh2HRTF: Open-source software package for the numerical calculation of head-related transfer functions," in *22nd ICSV*, Florence, Italy, 2015.

[18] Ziegelwanger, H., Majdak, P., and Kreuzer, W., "Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization," *Journ. of the Acoust. Soc. of America*, 138, pp. 208–222, 2015.

[19] Bernschütz, B., "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," in *Proceedings of AIA/DAGA*, pp. 592–595, DEGA, Meran, Italy, 2013.