Influence of stand configurations on ecological validity of audiovisual recording systems

Influence of stand configurations on ecological validity of audiovisual recording systems

Angela Guastamacchia[1], Giuseppina Emma Puglisi[1], Andrea Bottega[1], Louena Shtrepi[1], Fabrizio Riente[2], Arianna Astolfi[1]

[1] *Department of Energy, Politecnico di Torino, 10129 Turin, Italy, Email:angela.guastamacchia@polito.it*
[1] *Department of Energy, Politecnico di Torino, 10129 Turin, Italy, Email:giuseppina.puglisi@polito.it*
[1] *Department of Energy, Politecnico di Torino, 10129 Turin, Italy, Email:andrea.bottega@polito.it*
[1] *Department of Energy, Politecnico di Torino, 10129 Turin, Italy, Email:louena.shtrepi@polito.it*
[2] *Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Turin, Italy, Email:fabrizio.riente@polito.it*
[1] *Department of Energy, Politecnico di Torino, 10129 Turin, Italy, Email:arianna.astolfi@polito.it*

Introduction

In recent years, the strong development of virtual reality (VR) and related technologies have led to their use in various fields [1], among which the auditory research one, which has found in VR the means for empowering hearing-impairment diagnostical procedure and hearing devices fitting. In particular, VR is being exploited to reproduce ecological listening tests based on the spatial auralization of everyday-life scenes inside complex auditory environments, further coupled with the related visual information [2,3]. Whether these AudioVisual (AV) scenes are based on simulations or in-field shootings, the attempt is to achieve scenes that come closer and closer to authenticity, that is, to reproduce ordinary scenarios that are indistinguishable from reality [4]. In this regard, it is necessary that both visual and sound fields are generated (or acquired) and reproduced properly so as to recreate a sense of immersion by recalling the complex AV interaction typical of real-life auditory perception [1,4]. Specifically regarding in-field recordings, devices that stereoscopically capture the 360° visual scene with a good resolution are already commercially available [2]; yet, although they can simultaneously acquire spatial audio, the maximum available resolution is limited to the first ambisonic order, making it necessary to use a separate additional Spherical Microphone Array (SMA) for high-order ambisonics recording [4-6] when enhanced spatial audio resolutions are needed (i.e., better perceived sound localization). However, the use of two different devices, placed one on top of the other, composing the final recording system for simultaneous audio-video acquisition leads to discrepancies in the recorded scenes from the real ones, as (i) the SMA falls within the visual field captured by the camera, undermining the authenticity of the recorded scene, (ii) the presence of the camera near the SMA influences the recorded sound field, (iii) the non-coincidence of the centers of the two devices leads to a mismatch between the acoustical and visual fields, affecting the coherence between the audio scene and the video scene [4]; that means the listener does not perceive the spatial origin of the sound source as coincident with the location of the source image he sees.

Nevertheless, to the authors' knowledge, no in-depth considerations have been published on the usage of these kinds of composed AudioVisual Recording Systems (AVRSs) in sight of evaluating and preserving the ecological validity of the produced AV scenes. Thereupon, the presented work proposes a method to analyze the influence of the raised three issues for two different placements on the same stand of an audio and a video device in order to evaluate, depending on the end use, which configuration leads to the most plausible AV recording (i.e., the one that comes closest to the human perception of reality) and to which extent these recordings can still be considered as ecologically valid.

Experimental method

The study was conducted for two examples of AVRSs comprising either the Insta360 ONE X2 or the Insta360 Pro camera to acquire 360° video shootings at up to 5.7K and 8K resolution (at 30fps), respectively, and the 19-channel Zylia ZM-1 SMA (flat frequency response from 28 Hz to 20 kHz) for capturing ambisonic tracks up to the third order. Two possible placements of audio and video devices were compared for each AVRS, as shown in Figure 1. For the first AVRS, the configuration with the ONE X2 on top of the ZM1 (X2-ZM1) was compared with the one with ZM1 on top of ONE X2 (ZM1-X2). Similarly, for the other AVRS, the configuration with the Pro on top of the ZM1 (PRO-ZM1) was compared with the reciprocal one (ZM1-PRO).

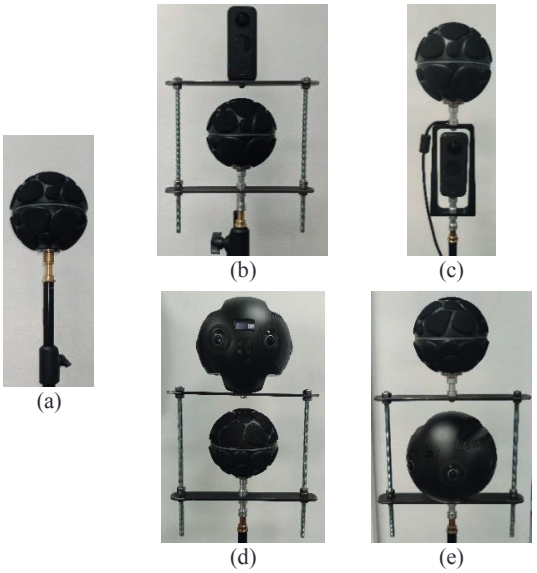


Figure 1: Stand configurations of AVRSs to be compared: (a) baseline (ZM1); (b) X2-ZM1; (c) ZM1-X2; (d) PRO-ZM1; (e) ZM1-PRO.

All the configurations were compared by respectively evaluating: (i) the influence of Zylia on the Insta360's field of view, (ii) the influence of Insta360 on the acquired sound field, and (iii) the level of AV coherence.

Audiovisual coherence

In order to generate coherent AV recordings, the sound source should be spatialized and perceptually localized at the same point in the space where the image of the sound source is displayed in the 360° video. For this to happen, the origin of the acoustical and visual scene should ideally match. However, in cases where the exact coincidence is not possible, to the end of still accomplishing ecological recordings, it is sufficient to ensure that the sound source is placed at a distance from the AVRS such that the difference in the inclination (or elevation) angle between the source seen by the SMA and the camera is lower than the minimum audible angle defining the human auditory spatial resolution, i.e., a maximum difference of 5° [7]. Moreover, this distance depends on the distance between the two centers of the devices placed on the same stand. Specifically, as shown in Figure 2, to perceive AV coherence, the distance between the center of the AVRS and the source (d_{SR}) must be:

$$d_{SR} > \sqrt{\frac{\left(\cos\frac{\alpha_{min}}{2}\right)^2 \cdot \left(\frac{d_{CC}}{2}\right)^2}{1-\left(\cos\frac{\alpha_{min}}{2}\right)^2}} \quad [\text{m}] \qquad (1)$$

where:
d_{CC} is the distance between the SMA center and the footage center of the 360° camera [m];
α_{min} is the minimum audible inclination angle, i.e. the minimum separation between two sources that can be reliably detected [7].

Figure 2: Evaluation of the minimum distance between the AVRS and sound source to obtain perceptually coherent AV scenes.

Influence on the sound field

The influence of both the additional support and the camera on the sampled sound field was evaluated in terms of errors on the recorded unweighted and A-weighted sound pressure levels (L and L_A) compared with the reference condition (ZM1), i.e., the level recorded by the Zylia without any nearby obstacle as in Figure 1(a). In particular, the L_A was computed to point out the level differences for frequencies where the human hearing sensitivity is greater so as to actually evaluate whether the auditory scene, recorded through a given AVRS configuration, would have been perceived as ecologically valid by the human ear. Specifically, for each configuration (ZM1, X2-ZM1, ZM1-X2, PRO-ZM1, and ZM1-PRO), 36 19-channel recordings, on 24-bit with a sampling frequency of 48 kHz, of 10-second pink noise were acquired, each emitted at the same distance but from a different angle around the SMA of the AVRS, through a 3rd-order ambisonic playback system consisting of a spherical array of 16 speakers and 2 subwoofers, with a flat frequency response from 40 Hz to 20 kHz. Two third-octave bands analyses from 50 Hz to 16 kHz were performed to evaluate the L and L_A difference for all AVRS configurations compared with the baseline ZM1 computed as:

$$\Delta L_X^{AVR_i} = L_X^{ZM1} - L_X^{AVR_i} \quad [\text{dB}] \qquad (2)$$

where:
X refers to either the unweighted or A-weighted sound pressure level;
AVR_i refers to each configuration: X2-ZM1, ZM1-X2, PRO-ZM1, ZM1-PRO.

In particular, for $\Delta L_X^{AVR_i}$, average (avg) , standard deviation (std dev), maximum positive (max pos) and negative (max neg) errors are analyzed: (i) as sound source location varies in terms of the average of global equivalent sound pressure level (L_{Aeq}) across all channels, and (ii) as frequency varies in terms of L averaged across all channels and all source positions. During the recordings, the gain of the loudspeaker system was set to achieve a sufficiently high Signal-to-Noise Ratio (SNR > 28 dB) over the entire frequency range of interest.

Influence on the visual field

When an undesired static object falls within the upper or lower field of view captured by the camera, as it occurs for the analyzed AVRSs, well-grounded video post-production techniques can be applied to mask the presence of the object in the scene by reconstructing the background (floor or ceiling) behind the object, without yielding any visible artifacts [8]. However, some measures need to be applied when shooting: (i) in the case of ZM1-X2 and ZM1-PRO, the camera should be mounted so as to take advantage of the blind spots between the lenses to hide the sidebars of the stand, (ii) moving objects should be prevented to appear within the portion of the field of view to be edited.

Results

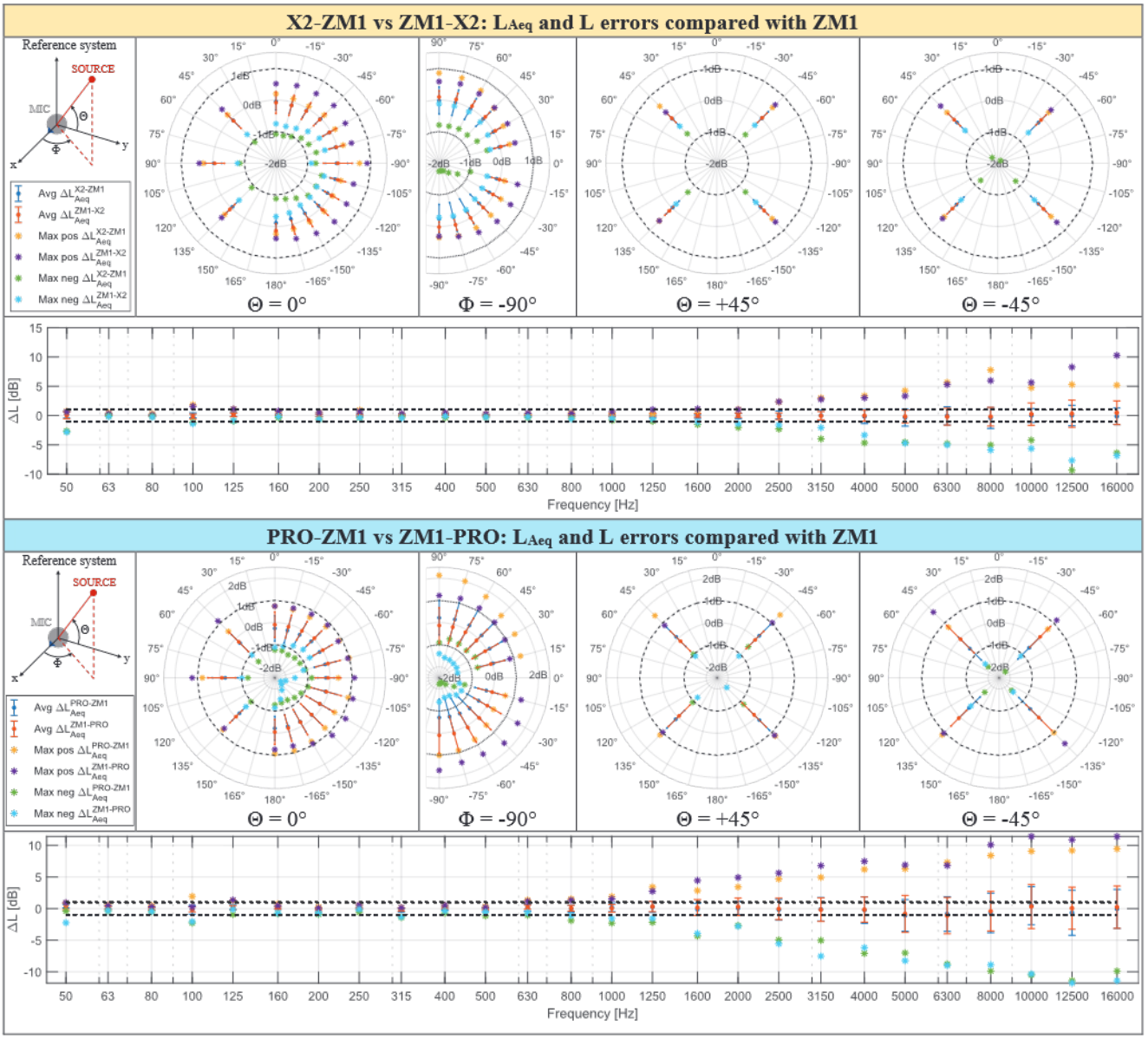
As aforementioned, it is necessary to have a minimum distance between the sound source and the AVRS to obtain perceptually consistent AV recordings. Table 1 summarizes the d_{SR} values for all AVRSs configurations given their d_{CC}.

Table 1: d_{CC} and d_{SR} values for all AVRSs configurations.

	X2-ZM1	ZM1-X2	PRO-ZM1	ZM1-PRO
d_{CC}	16 cm	12 cm	15 cm	18 cm
d_{SR}	1.85 m	1.40 m	1.75 m	2.10 m

All d_{SR} values result equal to or greater than the minimum distance (d_{min}) required to avoid artifacts due to the near-field effect related to the SMA size ($d_{min} \geq 1.40\ m$). Moreover, if the proper d_{SR} distance is ensured also from any other moving object inside the scene, the stand and the ZM1 can be successfully removed from the visual scene for all AVRSs.
Furthermore, Table 2 presents all outcomes of the acoustical analyses related to the effects of the camera presence nearby the SMA. The main findings for both AVRSs are listed below.

Table 2: Outcomes from the spatial (polar diagrams) and frequency acoustical analyses of all AVRSs configurations. All graphs display the level difference (avg, std dev, max pos, max neg) compared with the baseline configuration (ZM1), and the JND in black dashed lines. The first column illustrates the reference system for the polar coordinates and the legend for spatial and frequency analyses.



For the AVRS with the ONE X2:

- the avg and std dev values of ΔL_{Aeq}^{X2-ZM1} and ΔL_{Aeq}^{ZM1-X2} fall within the Just Noticeable Difference (JND) of 1 dB for all surrounding sound source locations;
- all avg values, as source location varies, are centered at 0 dB and are associated with std dev values lower than 0.5 dB;
- the max pos and neg values of ΔL_{Aeq}^{X2-ZM1} and ΔL_{Aeq}^{ZM1-X2} reach in the worst case 1 dB, except for all sources in the lower hemisphere for which the ΔL_{Aeq}^{X2-ZM1} max neg value (i.e., higher L_{Aeq} captured by the AVRS compared with the ZM1 configuration) progressively deviates from 1 dB to 2 dB at very negative Θ, where the sound impinging at middle-high frequencies on the camera above is reflected and adds up, leading to higher L_{Aeq} values;
- the avg values of ΔL^{X2-ZM1} and ΔL^{ZM1-X2}, as frequency varies, are centered at 0 dB, while the related std dev values fall within the JND for all frequencies up to 4 kHz, after which they start to slowly increase, reaching the maximum value of 2 dB at 12.5 kHz in the ZM1-X2 case;
- similarly, all max pos and neg values, as frequency varies, gradually diverge, till exceeding 10 dB at 16 kHz in the ZM1-X2 case, starting from 2 dB at 2.5 kHz, for which the wavelength is comparable with the camera height (11.3 cm), confirming the outcomes from the spatial analysis.

Similar observations can be made about the other AVR system, for which, in general, higher error values are

reported due to the impact of a bigger spherical 360° camera (diameter of 14.3 cm).
In particular, for the AVRS with the Pro:

- $\Delta L_{Aeq}^{PRO-ZM1}$ and $\Delta L_{Aeq}^{ZM1-PRO}$ maximum errors tend to exceed the JND limits for almost all source positions;
- all avg values, as source location varies, are centered at 0 dB, except for sound sources located in the lower hemisphere in case of PRO-ZM1 configuration, for which the avg value is shifted towards -0.5 dB, making the std dev to exceed the lower bound of JND;
- all std values, as source location varies, show values reaching 0.5 dB in the worst case;
- as for the X2-ZM1 case, the $\Delta L_{Aeq}^{PRO-ZM1}$ max neg at negative Θ angles reaches 2dB. Conversely, for the same angles, the $\Delta L_{Aeq}^{ZM1-PRO}$ shows max pos values (i.e., lower L_{Aeq} captured by the AVRS compared with the ZM1 configuration) that hit almost 2 dB, meaning that the presence of the camera below the ZM1 masks part of the middle-high frequency content;
- for all the positive Θ angles, the exact reverse trend occurs. In case of PRO-ZM1, the camera covers the SMA leading to higher $\Delta L_{Aeq}^{PRO-ZM1}$ max pos values (up to 2 dB), while in case of ZM1-PRO, the camera below the ZM1 reflects part of the sound back to the SMA, leading to higher $\Delta L_{Aeq}^{PRO-ZM1}$ max neg values (up to 1 dB);
- all avg values of $\Delta L^{PRO-ZM1}$ and $\Delta L^{ZM1-PRO}$, as frequency varies, fall within the JND and are centered at 0 dB up to 4 kHz, while the $\Delta L^{ZM1-PRO}$ and $\Delta L^{PRO-ZM1}$ std dev values fall within the JND up to 1.25 kHz and 2 kHz, after which they begin to deviate, hitting almost 4 dB at 16 kHz.
- similarly, maximum errors begin strongly deviating from the JND from 1.25 kHz (3 dB) onward, till exceeding 10 dB at 10 kHz and 16 kHz, in the ZM1-PRO case and PRO-ZM1, respectively.

Conclusions

The proposed study evaluates the effect of using a 360° camera and a separate Spherical Microphone Array (SMA) mounted on the same stand on the ecological validity of the recorded 360° AudioVisual (AV) scenes by analyzing: the AV coherence, the influence of the camera on the sound field and of the SMA on the visual field. Two AudioVisual Recording Systems (AVRSs), composed of either the Insta360 ONE X2 or the Pro and the Zylia ZM1, were analyzed, comparing, for each of them, the configuration with the ZM1 on top (ZM1-X2, ZM1-PRO) with the reciprocal one (X2-ZM1, PRO-ZM1). Results show that, for the AVRS involving the X2, the ZM1-X2 configuration achieves coherent ecological AV scenes starting from lower distances from the sound source compared with X2-ZM1 (1.4 m against 1.85 m). Moreover, all ZM1-X2 A-weighted sound pressure level (L_{Aeq}) errors compared with the base ZM1 (ΔL_{Aeq}^{ZM1-X2}) remain within the JND for all surrounding source positions, while, in case of X2-ZM1, when sound sources are located at negative inclination angles, the maximum values of ΔL_{Aeq}^{X2-ZM1} reach 2 dB. However, from 10 kHz up, the ZM1-X2 shows higher error than the X2-ZM1, till hitting a maximum error of 10 dB at 16 kHz. Thus, ZM1-X2 should be preferred in case of closer sources with any inclination angle and when the frequency content of interest is below 10 kHz. While, concerning the AVRS with the Pro, both configurations show L_{Aeq} maximum error values up to 2 dB in case of inclination angles different than 0°. However, the PRO-ZM1 requires minor minimum source-to-AVRS distance (1.75 m against 2.1 m) and effort during the video post-processing procedure and is characterized by a slightly better frequency behavior compared with the ZM1-PRO, even if L maximum error values still exceed 10 dB at 16 kHz.

Reference

[1] A. Hirway, Y. Qiao, and N. Murray, “Spatial audio in 360° videos: does it influence visual attention?,” in *Proceedings of the 13th ACM Multimedia Systems Conference*, Athlone Ireland: ACM, Jun. 2022, pp. 39–51. doi: 10.1145/3524273.3528179.

[2] G. Llorach, G. Grimm, M. M. Hendrikse, and V. Hohmann, “Towards realistic immersive audiovisual simulations for hearing research: Capture, virtual scenes and reproduction,” in *Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia*, 2018, pp. 33–40.

[3] S. Van De Par *et al.*, “Auditory-visual scenes for hearing research,” *Acta Acustica*, vol. 6, p. 55, 2022.

[4] M. Kentgens, S. Kühl, C. Antweiler, and P. Jax, “From Spatial Recording to Immersive Reproduction – Design & Implementation of a 3DOF Audio-Visual VR System,” *New York*, 2018.

[5] R. F. Fela, A. Pastor, P. Le Callet, N. Zacharov, T. Vigier, and S. Forchhammer, “Perceptual Evaluation on Audio-Visual Dataset of 360 Content,” in *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, IEEE, 2022, pp. 1–6.

[6] A. Heimes, M. Yang, and M. Vorländer, “Virtual Reality Environments for Soundscape Research,” 2022, doi: 10.21008/J.0860-6897.2022.1.12.

[7] T. Z. Strybel and K. Fujimoto, “Minimum audible angles in the horizontal and vertical planes: Effects of stimulus onset asynchrony and burst duration,” *The Journal of the Acoustical Society of America*, vol. 108, no. 6, pp. 3092–3095, Dec. 2000, doi: 10.1121/1.1323720.

[8] A. MacQuarrie and A. Steed, “Object removal in panoramic media,” in *Proceedings of the 12th European Conference on Visual Media Production*, 2015, pp. 1–10.