Influence of stand configurations on ecological validity of audiovisual recording systems

Influence of stand configurations on ecological validity of audiovisual recording systems

Original
Influence of stand configurations on ecological validity of audiovisual recording systems / Guastamacchia, Angela; Puglisi, Giuseppina; Bottega, Andrea; Shtrepi, Louena; Riente, Fabrizio; Astolfi, Arianna. - ELETTRONICO. - (2023), pp. 88-91. (Intervento presentato al convegno Proceedings of the 1st AUDICTIVE Conference tenutosi a Aachen , Germany nel June 19-22, 2023) [10.18154/RWTH-2023-05549].

Availability:
This version is available at: 11583/2980450 since: 2023-07-17T18:16:55Z

Publisher:
RWTH Publications

Published
DOI:10.18154/RWTH-2023-05549

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:

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

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:

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

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

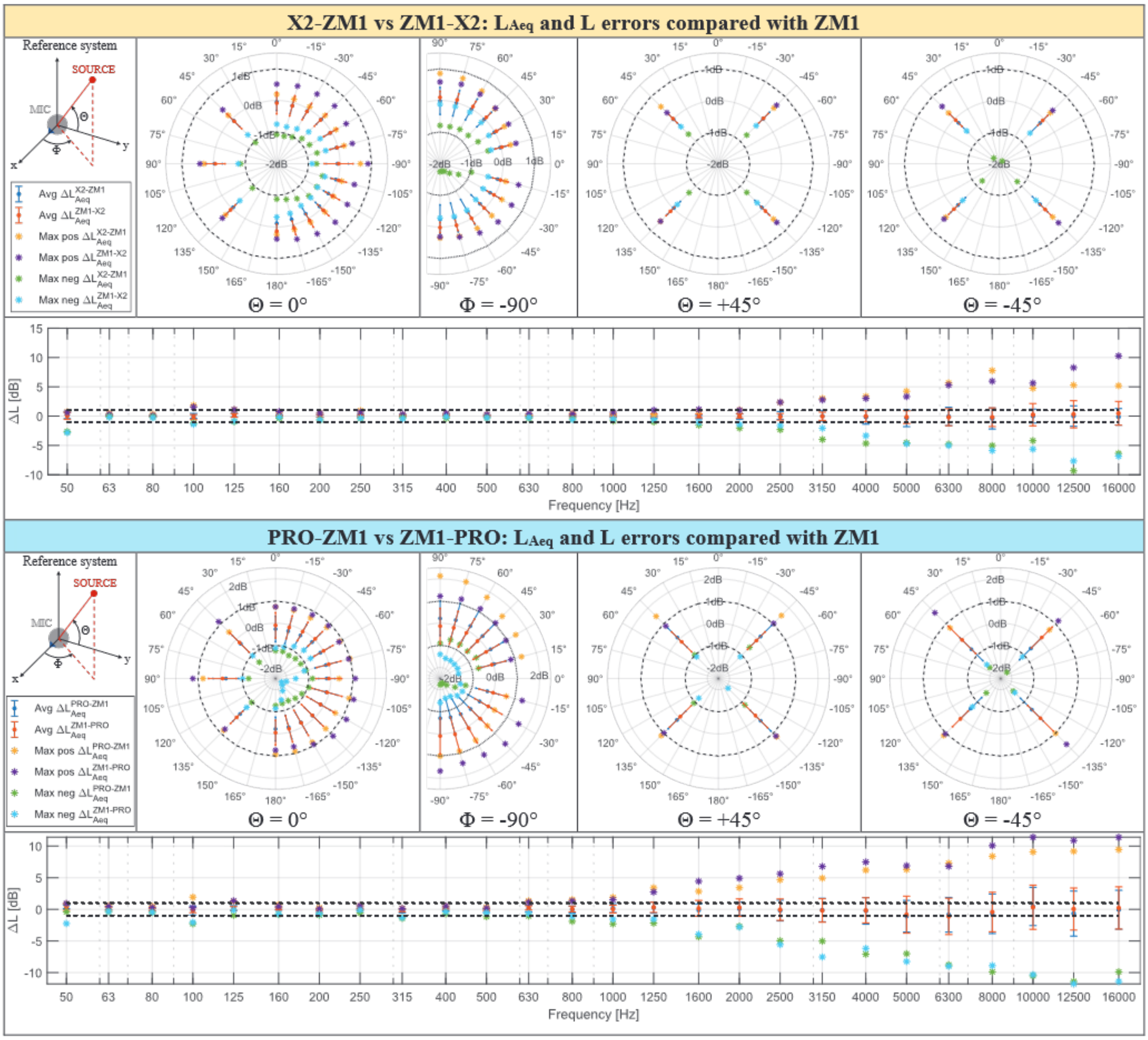
As aforementioned, it is necessary to have a minimum distance between the sound source and the AVRS to obtain perceptually consistent AV recordings. Table 1 summarizes the d_{SR} values for all AVRSs configurations given their d_{CC}.

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