Associazione Italiana di Acustica
49° Convegno Nazionale
Ferrara, 7-9 giugno 2023

INFLUENZA DEL POSIZIONAMENTO SULLO STESSO SUPPORTO DI VIDEOCAMERA 360 E ARRAY MICROFONICO SULLA PLAUSIBILITÀ DELLE REGISTRAZIONI AUDIOVISIVE

Angela Guastamacchia (1), Giuseppina Emma Puglisi (2), Andrea Bottega (3), Louena Shtrepi (4), Fabrizio Riente (5), Arianna Astolfi (6)

1) Dipartimento Energia - Politecnico di Torino, Torino, angela.guastamacchia@polito.it
2) Dipartimento Energia - Politecnico di Torino, Torino, giuseppina.puglisi@polito.it
3) Dipartimento Energia - Politecnico di Torino, Torino, andrea.bottega@polito.it
4) Dipartimento Energia - Politecnico di Torino, Torino, louena.shtrepi@polito.it
5) Dipartimento di Elettronica e Telecomunicazioni - Politecnico di Torino, Torino, fabrizio.riente@polito.it
6) Dipartimento Energia - Politecnico di Torino, Torino, arianna.astolfi@polito.it

SOMMARIO

Al fine di registrare video immersivi in grado di catturare a 360° sia il campo sonoro che quello visivo, sono state sviluppate videocamere 360 capaci di acquisire in contemporanea audio spaziale fino al primo ordine ambisonico. Tuttavia, quando è richiesta una risoluzione audio spaziale maggiore, è necessario avvalersi di un array microfonico separato a fianco alla camera 360. Per cui, nel lavoro proposto, al fine di ottenere scene audiovisive plausibili, vengono valutati: (i) l'influenza della telecamera sul campo sonoro e del microfono sul campo visivo, (ii) la coerenza audiovisiva delle scene registrate da una coppia microfono-camera 360.

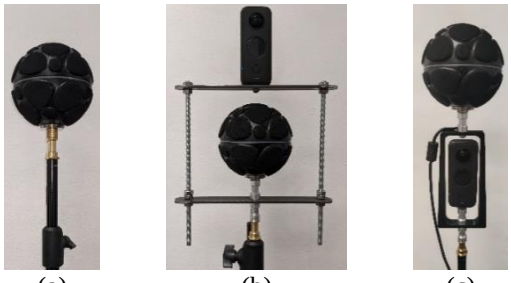
1. Introduzione

Negli ultimi anni, il forte sviluppo della realtà virtuale e delle tecnologie ad essa connesse, hanno portato all'utilizzo della stessa in vari campi [1], dall'intrattenimento, come videogiochi immersivi e riprese audiovisive a 360°, all'applicazione medica, come test di ascolto basati su riprese 360 di scenari di vita reale. In particolare, l'autenticità delle scene, e quindi delle immagini presentate all'usufruente, è il prossimo obiettivo: ovvero riprodurre immagini che siano indistinguibili dalla realtà [2]. A tal proposito, è necessario che non solo il campo visivo, ma anche quello sonoro siano generati o acquisiti, e riprodotti adeguatamente, in modo da ricreare una sensazione di completa immersività richiamando la complessa interazione umana tra percezione visiva e uditiva della vita reale [1,2]. Per quanto riguarda la registrazione in campo, dispositivi in grado di catturare stereoscopicamente con una buona risoluzione la scena visiva a 360° sono già disponibili in commercio, tuttavia, nonostante questi possano già acquisire in contemporanea audio spaziale, la massima risoluzione disponibile è limitata al primo ordine ambisonico, per cui, per catturare il campo sonoro con una risoluzione spaziale maggiore, ovvero migliore localizzazione sonora percepita, è necessario introdurre un ulteriore array microfonico separato in grado di campionare l'ambiente con un ordine ambisonico superiore [2-4]. L'utilizzo di due diversi dispositivi per l'acquisizione contemporanea audio-video, posti uno sopra l'altro sullo stesso supporto, però, porta a delle discrepanze nelle scene registrate da quelle reali, in quanto (i) il microfono rientra nel campo visivo catturato dalla telecamera, inficiando l'autenticità della scena registrata, (ii) la presenza della telecamera nei pressi del microfono influenza il campo sonoro registrato, (iii) la non coincidenza dei centri di ripresa dei due dispositivi porta a un disallineamento tra il campo acustico e quello visivo, inficiando la coerenza tra scena audio e scena video [2]; ovvero, l'origine spaziale della sorgente sonora percepita non coincide con la posizione dell'immagine della sorgente sonora vista dall'ascoltatore.

Il lavoro presentato propone un metodo per analizzare l'influenza di queste tre problematiche per due diversi posizionamenti sullo stesso supporto di due dispositivi audio e video, al fine di valutare, a seconda dell'utilizzo finale, quale configurazione porti ad avere la registrazione audiovisiva più plausibile, ovvero quella che si avvicini di più alla percezione umana della realtà.

2. Metodo sperimentale

Lo studio è stato condotto per una coppia esempio di dispositivi comprendente l'array microfonico a 19 canali Zylia ZM-1 (risposta piatta in frequenza da 20 Hz a 20 kHz), per l'acquisizione di tracce ambisoniche fino al terzo ordine, e la videocamera Insta360 ONE X2, per la ripresa video a 360° con una risoluzione fino a 5.7K (a 30fps). Due supporti, uno in metallo e uno stampato 3D in plastica, sono stati realizzati per montare entrambi i dispositivi sullo stesso treppiedi in due configurazioni diverse: Insta360 sopra Zylia (X2-ZM1) come mostrato in Figura 1(b) e Zylia sopra Insta360 (ZM1-X2) come in Figura 1(c).



(a) (b) (c)

Figura 1 – Configurazioni di ZM1 e ONE X2 da confrontare: (a) base (ZM1); (b) ONE X2 su ZM1 (X2-ZM1); (c) ZM1 su ONE X2 (ZM1-X2).

I due sistemi di ripresa audiovisiva creati sono stati messi a confronto valutando per ognuno: (i) l'influenza dello Zylia sul campo visivo della Insta360, (ii) l'influenza dell'Insta360 sul campo sonoro acquisito e (iii) il livello di coerenza audiovisiva.

Associazione Italiana di Acustica, 2023
ISBN: 978-88-88942-66-7

2.1 Coerenza audiovisiva

Al fine di generare registrazioni audiovisive coerenti è necessario che il suono venga spazializzato e quindi percettivamente localizzato nello spazio in coincidenza con l'immagine della sorgente sonora mostrata nel video 360. Per far sì che questo avvenga, idealmente i centri di riferimento del campo acustico e visivo dovrebbero coincidere. Nei casi in cui la corrispondenza esatta non sia possibile, per ottenere registrazioni audiovisive plausibili, è sufficiente che la sorgente sonora sia a una distanza dal sistema di registrazione tale da avere una differenza di angolo di elevazione tra la sorgente vista dall'array e dalla videocamera minore del minimo angolo percepibile dall'orecchio umano, ovvero una differenza massima di 5° [5]. Questa distanza, inoltre, dipende dalla distanza tra i due centri dei dispositivi posti sullo stesso supporto. In particolare, come mostrato in Figura 2, per percepire una coerenza audiovisiva la distanza tra il centro del sistema di registrazione e la sorgente (d_{SR}) deve essere:

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

dove:
d_{CC} è la distanza tra il centro dell'array microfonico e il centro focale della videocamera 360 [m];
α_{min} è la minima risoluzione percepibile dall'orecchio umano lungo l'angolo di elevazione.

Figura 2 – Calcolo della distanza minima tra sistema di registrazione e sorgente sonora per ottenere scene audiovisive percettivamente coerenti.

2.2 Influenza sul campo sonoro

L'influenza del supporto aggiuntivo e della videocamera sul campo sonoro campionato è stata valutata in termini di errore sul livello di pressione sonora (L_Z) registrato rispetto alla condizione base (ZM1), ovvero sull'L_Z registrato dallo Zylia senza alcun ostacolo vicino come in Figura 1(a). In particolare, per ogni configurazione (ZM1, ZM1-X2 e X2-ZM1), sono state acquisite 36 registrazioni a 19 canali su 24-bit a 48 kHz di rumore rosa da 10 secondi, ognuna emessa alla stessa distanza ma da un'angolazione diversa attorno al sistema di registrazione, sfruttando un sistema di riproduzione ambisonica del terzo ordine composto da un array sferico di 16 altoparlanti più 2 subwoofer, con risposta in frequenza piatta dai 40 Hz ai 20 kHz. Due analisi, in bande di terzi d'ottava nell'intervallo di frequenza da 50 Hz a 16 kHz, sono state eseguite per valutare l'errore medio, la sua deviazione standard e i valori massimi e minimi delle registrazioni ottenute con ZM1-X2 e X2-ZM1 rispetto a ZM1 sul: (i) livello di pressione sonora equivalente globale (L_{eq}) mediato su tutti i canali al variare della posizione della sorgente sonora e (ii) L_Z mediato su tutti i canali e tutte le posizioni al variare della frequenza. Durante le registrazioni, il guadagno del sistema di altoparlanti è stato impostato in modo da ottenere un rapporto segnale-rumore (SNR) sufficientemente alto (SNR > 28 dB) su tutto il range di frequenze di interesse.

2.3 Influenza sul campo visivo

Quando un oggetto statico indesiderato rientra nel campo visivo superiore o inferiore registrato dalla videocamera, come accade in questi casi con l'array microfonico, è possibile applicare delle tecniche di post-produzione video che permettano di mascherare la presenza dell'oggetto ricostruendo lo sfondo (pavimento o soffitto) sopra l'oggetto stesso. È necessario però applicare delle accortezze durante la ripresa: (i) nel caso di ZM1-X2, la camera deve essere montata in modo da sfruttare i punti ciechi tra le lenti e nascondere le barre laterali del supporto, (ii) oggetti in movimento non devono rientrare nella porzione di campo visivo da modificare.

3. Risultati

Per ottenere registrazioni audiovisive percettivamente coerenti è necessario avere $d_{SR} \geq 1.40\,m$ nel caso di ZM1-X2 ($d_{CC} = 12\,cm$) e $d_{SR} \geq 1.85\,m$ nel caso di X2-ZM1 ($d_{CC} = 16.2\,cm$). Inoltre, entrambe le distanze sono maggiori o uguali alla distanza minima d_{min} necessaria a evitare artefatti dovuti all'effetto di campo vicino, legati alle dimensioni dell'array microfonico ($d_{min}^{ZM-1} \geq 1.40\,m$). Per quanto riguarda, invece, l'influenza della presenza della Insta360 sul campo sonoro acquisito, questa risulta all'incirca comparabile per entrambe le configurazioni. In generale, nell'analisi dell'errore su L_{eq} al variare delle posizioni della sorgente, valore medio, deviazione standard, massimi e minimi rientrano nella Just Noticeable Difference (JND), eccetto per alcune posizioni dove massimi e minimi sforano la JND di meno di 0.5 dB. In particolare, è visibile l'effetto della Insta360 posta sopra lo Zylia nel caso di sorgenti sonore presenti nella semisfera superiore, per le quali l'errore massimo è di 1.4 dB. Dall'andamento dell'errore in frequenza, invece, è visibile come, per frequenze superiori a 4 kHz, per le quali la lunghezza d'onda è comparabile con le dimensioni della camera (11 cm), deviazione standard e valori massimi e minimi aumentino progressivamente, in genere per tutte le posizioni della sorgente, fino a raggiungere un errore massimo di 10 dB a 16 kHz sul canale 15 nel caso di ZM1-X2. In particolare, ZM1-X2 presenta errori minori di X2-ZM1 fino agli 8 kHz.

4. Conclusioni

Il lavoro proposto valuta l'effetto dell'utilizzo di una camera 360 e di un array microfonico separato, montati sullo stesso supporto, per registrare scene audiovisive 360 sulla loro plausibilità. Due configurazioni per lo Zylia ZM1 e la camera Insta360 ONE X2 sono state confrontate: ZM1-X2 vs X2-ZM1. X2-ZM1 è preferibile quando la sorgente sonora è a più di 1.85 m di distanza dal sistema di registrazione, mentre per distanze inferiori, fino a un minimo di 1.4 m, ZM1-X2 è più indicata, soprattutto nel caso in cui il contenuto in frequenza di interesse sia inferiore a 8 kHz.

5. Bibliografia

[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] 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.

[3] 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)*, Taipei City, Taiwan: IEEE, Jul. 2022, pp. 1–6. doi: 10.1109/ICMEW56448.2022.9859426.

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

[5] 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.