POLITECNICO DI TORINO
Repository ISTITUZIONALE

IMPATTO DELL'IMMERSIONE AUDIOVISIVA SULL'INTELLEGIBILITÀ DEL PARLATO

Original
IMPATTO DELL'IMMERSIONE AUDIOVISIVA SULL'INTELLEGIBILITÀ DEL PARLATO / Guastamacchia, A., Albera, A.. - ELETTRONICO. - (2024). (50° Convegno Nazionale AIA Taormina (ITA) 29-31 maggio 2024).

Availability:
This version is available at: 11583/2990319 since: 2024-07-03T13:36:42Z

Publisher:
Associazione Italiana di Acustica (AIA)

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

26 November 2026

Associazione Italiana di Acustica
50° Convegno Nazionale
Taormina, 29-31 Maggio 2024

IMPATTO DELL'IMMERSIONE AUDIOVISIVA SULL'INTELLEGIBILITÀ DEL PARLATO

Angela Guastamacchia (1), Andrea Albera (2)

1) Dipartimento Energia - Politecnico di Torino, Torino, angela.guastamacchia@polito.it
2) Dipartimento di Scienze Chirurgiche - Università degli Studi di Torino, Torino, andrea.albera@unito.it

SOMMARIO

Tre scene audiovisive sono state registrate all'interno di una sala conferenza riverberante per ricreare test di intellegibilità immersivi con diverse condizioni di rumore all'interno di un sistema di riproduzione audiovisivo spazializzato. Gli effetti della provvisione della scena visiva, con e senza labiale dell'interlocutore target, e del movimento naturale dell'utente sull'intellegibilità del discorso target sono stati studiati su un gruppo di 50 soggetti, evidenziando come solo nel caso di test con labiale venga raggiunta un'intellegibilità maggiore del test standard, ovvero test in condizioni statiche senza inclusione del video.

1. Introduzione

Negli ultimi tempi, la ricerca sull'udito ha beneficiato degli enormi sviluppi nel campo della realtà virtuale che hanno reso possibile la creazione e riproduzione immersiva di scenari audiovisivi al fine di ottenere test di ascolto più ecologici [1], ovvero test che mirino a ricreare in laboratorio condizioni e scenari di ascolto tipici della vita reale. A tal fine, alcuni studi hanno iniziato ad esplorare quali siano i fattori che maggiormente contribuiscono alla validità ecologica, investigando il ruolo della scena visiva associata a quella uditiva e gli effetti del movimento naturale del soggetto in ascolto, detto Self-Motion (SM) [2], sull'intellegibilità del parlato. Il contributo della scena visiva può essere inserito nei test, e quindi analizzato, per differenti gradi di complessità (o realismo). In particolare, la somministrazione del contesto visivo, inteso come presentazione visiva del luogo in cui avviene l'ascolto e della posizione delle sorgenti sonore, ha dimostrato influire sulla capacità di localizzazione dei suoni [3], sull'accettazione dell'illusione uditiva e sul SM [4], mentre la possibilità di vedere i movimenti della faccia e delle labbra dell'interlocutore target ha dimostrato supportare fortemente la comprensione del parlato [5]. Tuttavia, pochi studi hanno esaminato analiticamente l'effetto dei singoli fattori visivi e della loro combinazione con il SM sull'intellegibilità. Fichna et al. [6] e Hládek et al. [7], hanno valutato il contributo del contesto visivo parziale, ovvero mostrando solo alcune delle posizioni delle sorgenti sonore, combinato al SM tramite simulazioni AudioVisive (AV) di ambienti mediamente riverberanti, confrontando diverse condizioni di somministrazione del test, quali: solo scena uditiva in condizioni statiche, ovvero senza SM, e scena AV sia con che senza SM. L'esclusione, però, di parte dei contributi visivi e l'impiego di simulazioni di ambienti fittizi al posto di registrazioni di ambienti reali, considerati più efficaci in termini di realismo [8] e preferenza degli utenti [4], evidenzia il bisogno di investigare più a fondo il tema.

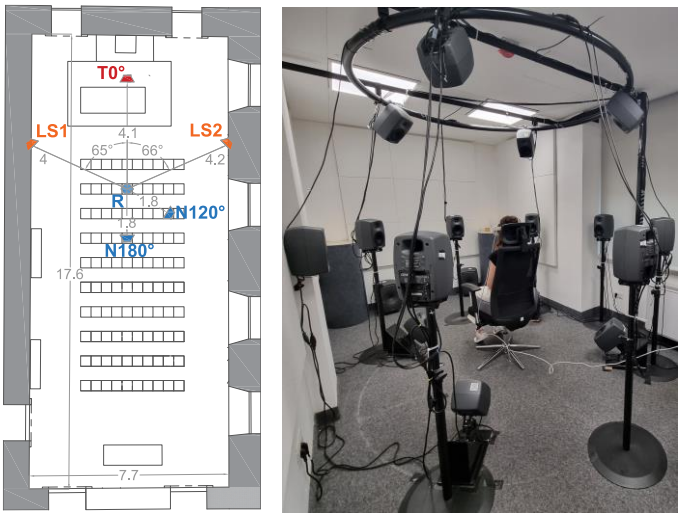
2. Metodologia

Il presente studio si propone, quindi, di valutare l'influenza dei singoli fattori visivi e del SM, utilizzando scene AV immersive registrate all'interno di una reale sala conferenza altamente riverberante e implementando test di intelligibilità per tre scenari di ascolto con un interlocutore target e diverse condizioni di rumore. L'analisi avviene tramite diverse somministrazioni delle tre scene create, incrementando di volta in volta il grado del realismo, ovvero presentando:

1) unicamente le scene uditive, in condizioni statiche (AO-S);
2) unicamente le scene uditive, permettendo il SM (AO-SM);
3) le scene AV fornenti il completo contesto visivo, in condizioni statiche (AV-S);
4) le scene AV con completo contesto visivo e SM (AV-SM);
5) le scene AV mostranti sia il completo contesto visivo che il labiale dell'interlocutore target, senza SM (AV-L).

2.1 Materiali: scene AV e test di intellegibilità

Le scene AV sono state collezionate all'interno della sala conferenze del Museo Egizio di Torino, con volume di 1500 m^3 e tempo di riverbero alle medie frequenze di 3.2 s. Tre scenari di ascolto, conformi al reale utilizzo della sala, sono stati selezionati, aventi lo stesso interlocutore target (T0°), a 4 m di fronte all'ascoltatore seduto in platea (R) e amplificato dai due altoparlanti laterali della sala (LS1 e LS2), e diverse condizioni di rumore, ovvero o in assenza di rumore o in presenza di un parlatore interferente seduto a 120° (N120°) o 180° (N180°) rivolto verso R a 1.8 m di distanza, come mostrato in figura 1(a).



(a) (b)

Figura 1 – (a) Pianta della sala conferenza con posizioni di R, T0°, N120°, N180°, LS1 e LS2. (b) Esecuzione del test AV all'interno dell'ASL.

Per l'acquisizione audio delle scene, tre misure ambisoniche del 3° ordine di risposte all'impulso sono state effettuate registrando un segnale sweep tramite l'array microfonico Zylia ZM-1 in R e la sorgente NTi Talkbox amplificata da LS1 e LS2 in T0°, a 1.5 m di altezza, e in N120° e N180° a 1.2 m da terra.

Associazione Italiana di Acustica, 2024
ISBN: 978-88-88942-69-8

Per ricreare le scene visive fornenti esclusivamente il contesto visivo, video 3D in 4K muti della durata di due minuti sono stati girati posizionando la videocamera 360 Insta360 Pro in R, la Talkbox in T0° e il manichino Brüel&Kjær 4128 in N120° e N180°. Per creare, invece, le scene visive mostranti anche il labiale dell'interlocutore target, sono state effettuate riprese a parte, con l'utilizzo di uno schermo verde di sfondo, di una persona pronunciante le frasi del test di intellegibilità associate al parlatore target. Infine, tramite tecniche di post-processamento, l'intera figura del parlatore target è stata inserita all'interno dei video girati in sala conferenze. In figura 2 è mostrata l'anteprima equirettangolare della scena con l'interlocutore reale in T0° e il parlatore interferente a 120° azimuth rappresentato dal manichino.



Figura 2 – Anteprima equirettangolare della scena visiva con interlocutore reale frontale e parlatore interferente fittizio a 120°.

Come materiale audio per implementare i test di intellegibilità nelle scene previste, sono state utilizzate, per l'interlocutore target, frasi da cinque parole provenienti dalla versione femminile dell'Italian Matrix Sentence Test [9] (imponendo 73 dBA nella posizione di ascolto) e, per la parlatrice interferente, diversi spezzoni di un brano standard foneticamente bilanciato, settando un SNR pari a -5 dB.

2.2 Soggetti, setup e procedura sperimentale

50 normoudenti madrelingua italiani (37 maschi e 13 femmine) di età compresa tra i 22 e i 46 anni (media = 27.8, deviazione standard = 4.8) sono stati reclutati volontariamente.

I test sono stati condotti nell'Audio Space Lab (ASL) del Politecnico di Torino (vedi Fig. 1(b)) che ospita un sistema di riproduzione 3D AV composto da un array sferico di 16 altoparlanti GENELEC 8030B [10], per la riproduzione audio in 3° ordine ambisonico, sincronizzato con il visore Meta Quest 2.

I 50 partecipanti sono stati divisi in cinque gruppi da dieci, ognuno corrispondente a una diversa condizione di test tra: AO-S, AO-SM, AV-S, AV-SM, e AV-L. Dopo una fase di familiarizzazione con il test, ad ogni partecipante sono stati somministrati i test di intellegibilità in forma aperta nelle tre scene, per un totale di 20 frasi target per scena e una durata complessiva di test di 15 minuti circa.

3. Analisi statistiche, risultati e conclusioni

Il test non-parametrico di Mann-Whitney è stato applicato a coppie di condizioni al fine di testare per ogni scena quale condizione di test portasse al miglior punteggio di intellegibilità statisticamente significativo. In figura 3 sono illustrate le medie e le deviazioni standard dei punteggi ottenuti per tutte le scene e le condizioni di test, mentre la tabella 1 riporta i p-value risultanti dalle analisi statistiche. Dai confronti tra i test AO-S, AO-SM, AV-S e AV-SM emerge che i test AO-S portano sempre a migliori punteggi di intellegibilità su un minimo di due scene su tre, come trovato in [7].

Al contrario, un'intellegibilità maggiore è raggiunta con il test AV-L rispetto all'AO-S, evidenziando come il contributo visivo sia di effettivo supporto alla comprensione del parlato solo in caso dell'aggiunta del labiale e come l'introduzione nei test del mero contesto visivo e del SM possa, in realtà, fungere da elemento di distrazione.

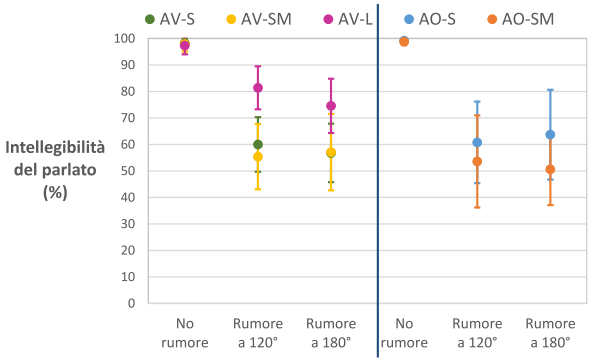


Figura 3 – Medie e deviazioni standard del punteggio di intellegibilità per ogni scenario uditivo e condizione di test.

Tabella 1 – p-value dei confronti per il test di Mann-Whitney. Valori in blu testano l'ipotesi nulla H_0: X1 ≤ X2, i restanti H_0: X1 ≥ X2. p-value ≤ 0.05 rifutano H_0.

X1	X2	No rumore	Rumore a 120°	Rumore a 180°
AO-SM	AO-S	**0.010**	**0.000**	
AV-S	AO-S	**0.046**		**0.022**
AV-SM	AO-S	**0.015**	**0.042**	**0.029**
AV-L	AO-S		**0.000**	**0.000**

4. Ringraziamenti

Gli autori ringraziano il Museo Egizio per la disponibilità e il supporto dimostrati durante le riprese audiovisive.

5. Bibliografia

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

[2] G. Grimm *et al.*, "Review of Self-Motion in the Context of Hearing and Hearing Device Research," *Ear & Hearing*, vol. 41, no. Supplement 1, pp. 48S-55S, Nov. 2020.

[3] A. Ahrens *et al.*, "Sound source localization with varying amount of visual information in virtual reality," *PLoS ONE*, vol. 14, no. 3, p. e0214603, Mar. 2019.

[4] M. M. Hendrikse *et al.*, "Influence of visual cues on head and eye movements during listening tasks in multi-talker audiovisual environments with animated characters," *Speech Communication*, vol. 101, pp. 70–84, 2018.

[5] A. MacLeod and Q. Summerfield, "Quantifying the contribution of vision to speech perception in noise," *British journal of audiology*, vol. 21, no. 2, pp. 131–141, 1987.

[6] S. Fichna *et al.*, "Effect of acoustic scene complexity and visual scene representation on auditory perception in virtual audio-visual environments," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, IEEE, 2021, pp. 1–9.

[7] Ľ. Hládek and B. U. Seeber, "Speech Intelligibility in Reverberation is Reduced During Self-Rotation," *Trends in Hearing*, vol. 27, p. 23312165231188619, Jan. 2023.

[8] G. Llorach *et al.*, "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.

[9] G. E. Puglisi *et al.*, "An Italian matrix sentence test for the evaluation of speech intelligibility in noise," *International journal of audiology*, vol. 54, no. sup2, pp. 44–50, 2015.

[10] A. Guastamacchia *et al.*, "Set up and preliminary validation of a small spatial sound reproduction system for clinical purposes," in *Forum acusticum, 2023*.