Vantaggio di intelligibità derivante dalla posizione delle sorgenti sonore per scene acustiche complesse con immersioni audio visive

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

VANTAGGIO DI INTELLIGIBITÀ DERIVANTE DALLA POSIZIONE DELLE SORGENTI SONORE PER SCENE ACUSTICHE COMPLESSE CON IMMERSIONI AUDIO VISIVE

Angela Guastamacchia (1), Louena Shtrepi (1), Giuseppina E. Puglisi (2), Andrea Albera (3), Franco Pellerey (4), Fabrizio Riente (5), Arianna Astolfi (1)

1) DENERG – Dipartimento Energia "Galileo Ferraris", Politecnico di Torino, angela.guastamacchia@polito.it, louena.shtrepi@polito.it, arianna.astolfi@polito.it
2) CALOS – Campus, Logistica e Sostenibilità, Politecnico di Torino, giuseppina.puglisi@polito.it
3) DISMA – Dipartimento di Scienze Matematiche "G. L. Lagrange", Politecnico di Torino, franco.pellerey@polito.it
4) Dipartimento di Scienze Chirurgiche, Università degli Studi di Torino, a.albera@unito.it
5) DET– Dipartimento di Elettronica e Telecomunicazioni, Politecnico di Torino, fabrizio.riente@polito.it

SOMMARIO

La realtà virtuale permette la riproduzione di scenari audio-visivi immersivi per realizzare test di intelligibilità sempre più ecologici. Sono stati fatti passi avanti che includono la riproduzione audio in laboratori ambisonici dove l'ascoltatore può sentire un audio spazializzato realistico e ruotare la testa e il busto, unita a video stereoscopici a 360° che rappresentano l'ambiente. In questo lavoro si riportano i risultati di test di intelligibilità immersivi in una sala conferenza riverberante con audio riprodotto in 3° ordine ambisonico sincronizzato con il visore Meta Quest 2 nell'Audio Space Lab del Politecnico di Torino. Si valuta in particolare il beneficio sull'intelligibilità quando la sorgente e il rumore mascherante informativo sono spazialmente separati.

1. Introduzione

La ricerca in ambito audiologico si è recentemente focalizzata su laboratori nei quali riprodurre con tecniche di realtà virtuale scene audio e video sempre più realistiche, nelle quali effettuare test di intelligibilità sempre più ecologici per valutare la sordità e gli effetti delle protesi acustiche [1]. Tali laboratori riproducono scene di tutti i giorni in cui ai pazienti viene richiesto di riconoscere parole in condizioni acusticamente complesse. In tali laboratori è anche concesso il movimento dell'ascoltatore, seppur limitato [2]. Sono pochi gli studi che ad oggi abbiano testato in modo sistematico gli effetti del movimento e dell'aggiunta del video sull'intelligibilità [3,4]. Inoltre, la somministrazione di un test con il video registrato è ancora più rara, in quanto la gran parte degli studi utilizza simulazioni [5].

In scenari di ascolto complessi, con un parlatore target e diverse sorgenti di rumore informativo a diversi azimut e distanze dall'ascoltatore, viene spesso indagato l'effetto del beneficio sull'intelligibilità quando la sorgente e il rumore mascherante sono spazialmente separati. Nella dizione anglosassone ci si riferisce alla *Spatial Release from Masking* (SRM) [6]. Sono stati studiati vantaggi di intelligibilità derivanti dalla posizione delle sorgenti sonore con alto riverbero dagli anni Settanta, ma in condizioni immersive audio video e in presenza di movimento dell'ascoltatore non sono stati esplorati, e solo pochi studi hanno considerato tempi di riverbero molto elevati.

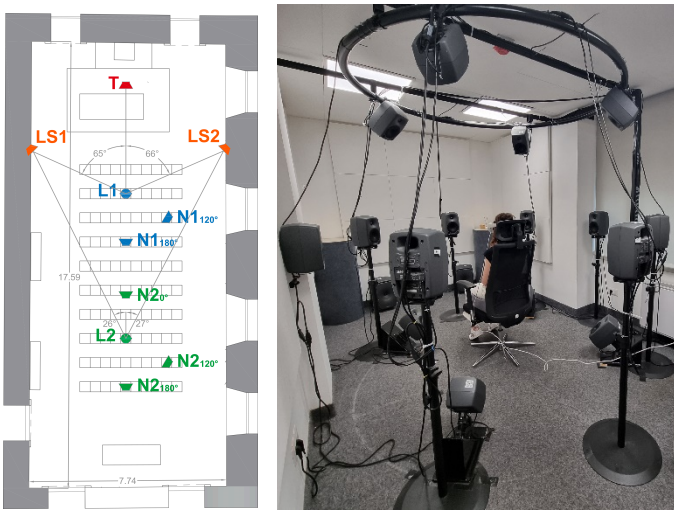
In questo lavoro si valuta l'influenza sull'intelligibilità del parlato (i) dell'aggiunta all'audio del contesto visivo rappresentante l'ambiente (AV) e (ii) della rotazione della testa e del busto dell'ascoltatore (*self-motion*, SM), in scene immersive registrate nella sala conferenza riverberante del Museo Egizio di Torino. Inoltre, (iii) si valuta il beneficio spaziale sull'intelligibilità con sorgente e rumore mascherante informativo spazialmente separati, a diverse distanze dalla sorgente. I soggetti sono 40 volontari normoudenti. I test sono effettuati nell'Audio Space Lab del Politecnico di Torino, dove l'audio è riprodotto in 3° ordine ambisonico, sincronizzato con il visore Meta Quest 2, nelle seguenti condizioni:

1) scene uditive, in condizioni statiche (AO-S);
2) scene uditive, permettendo il movimento (AO-SM);
3) scene AV in condizioni statiche (AV-S);
4) scene AV permettendo il movimento (AV-SM).

2. Metodologia

2.1 Caso studio

Il caso studio è la sala conferenze del Museo Egizio di Torino che ha un volume di 1500 m^3 e un tempo di riverberazione alle medie frequenze di 3.2 s. Le posizioni di ascolto L1 e L2 a 1,2 m da terra, del parlatore target T a 1,5 m da terra, degli altoparlanti laterali che amplificavano il segnale target LS1 e LS2 a 1,7 m da terra, delle sorgenti di rumore informativo $N1_{120°}$, $N1_{180°}$, $N2_{0°}$, $N2_{120°}$, $N2_{180°}$, a 1,2 m da terra, sono mostrati in Figura 1(a). Il rumore interferente informativo era di genere femminile e raccontava una storia foneticamente bilanciata. Gli ascoltatori erano rivolti verso il parlatore target e si trovavano alla distanza di 4 m e 10 m circa. La distanza dei due altoparlanti LS1 e LS2 era 4 m e 8 m circa, per i due ascoltatori L1 e L2, rispettivamente.



(a) (b)

Figura 1 – (a) Pianta della sala conferenze del Museo Egizio con le posizioni degli ascoltatori L1 e L2, del parlatore target T, degli altoparlanti LS1 e LS2, delle sorgenti di rumore $N1_{120°}$, $N1_{180°}$, $N2_{0°}$, $N2_{120°}$, $N2_{180°}$. (b) Esecuzione del test audio-visivo all'interno dell'Audio Space Lab.

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

2.2 Scene audio video (AV)

Le scene AV erano sette per ogni soggetto. Tre scene con l'ascoltatore a 4 m e quattro scene con l'ascoltatore a circa 10 m dalla sorgente. Per la distanza di ascolto 4 m sono state considerate la condizione senza rumore e due condizioni di rumore interferente spazializzato a 1,8 m dall'ascoltatore, i.e., $N1_{120°}$, $N1_{180°}$. Per la distanza 10 m le condizioni di rumore considerate sono state 3, cioè, $N2_{0°}$, $N2_{120°}$ e $N2_{180°}$, a cui si è aggiunta la condizione in quiete. L'audio per le singole scene è stato acquisito tramite misure di risposte all'impulso ambisoniche del 3° ordine tramite l'array microfonico Zylia ZM-1 in L1 e L2 e la sorgente NTi Talkbox in T e amplificata da LS1 e LS2. Il rumore di parlato interferente proveniva sempre da una Talkbox. Il contesto visivo 3D è stato registrato in 4K con la videocamera 360 Insta360 Pro in L1 e L2. Per ricreare visivamente una condizione realistica è stata posizionata la Talkbox in T e il manichino Brüel&Kjær 4128 nelle posizioni della parlatrice interferente.

2.3 Test di intelligibilità e procedura di somministrazione

I test di intellegibilità sono composti da frasi di cinque parole facenti parte della versione femminile dell'Italian Matrix Sentence Test [7]. Il livello del segnale nelle posizioni di ascolto era come nella condizione reale. Il rapporto segnale rumore con il parlato femminile interferente proveniente da un diverso azimut attorno all'ascoltatore era pari a – 5 dB. I test sono stati somministrati nell'ASL (Fig. 1(b)), che include un sistema di riproduzione AV 3D composto da un array sferico di 16 altoparlanti GENELEC 8030B [8], per la riproduzione audio in 3° ordine ambisonico, sincronizzato con il visore Meta Quest 2. Hanno partecipato volontariamente 40 ascoltatori normoudenti madrelingua italiani (30 maschi e 10 femmine) di età compresa tra i 22 e i 46 anni, che sono stati suddivisi in quattro gruppi da dieci, corrispondenti a una delle quattro condizioni di test: AO-S, AO-SM, AV-S, AV-SM. La prima fase di test è consistita in un training e successivamente ad ogni soggetto sono state fatte ascoltare 20 frasi per ognuna delle 7 scene, per una durata complessiva di 35 minuti circa. Il test era in forma aperta, cioè l'ascoltatore doveva ripetere le parole comprese all'operatore che le annotava il numero di parole corrette su un computer. La percentuale di parole comprese, definita intelligibilità, è stata trasformata in Rationalized Arcsin Units (RAU) per correggere l'effetto "pavimento" e "soffitto" dei punteggi [9].

3. Risultati

In Tabella 1 sono riportate le medie e le deviazioni standard dell'intelligibilità del parlato trasformata in RAU per ognuna delle quattro condizioni di test. Dal confronto fra le medie si evince che la condizione AO-S, cioè quella con solo audio e in posizione statica, determina la maggiore intelligibilità, mentre le condizioni con video, sia con movimento (AV-SM) che senza movimento (AV-S) sono in seconda posizione in parità. La peggiore condizione è quella con solo audio e movimento (AO-SM).

Tabella 1 – Media e Deviazione Standard delle percentuali di intelligibilità del parlato trasformate in RAU per ogni scena. AV (Audio-Video), SM (Self-Motion), S (static), AO (Audio-Only).

Scena	N	Media	Dev. St.
AV-SM	1000	51	33
AV-S	1000	51	31
AO-SM	1000	48	30
AO-S	1000	57	31

In Tabella 2 si riportano i risultati dell'applicazione del test di Mann-Whitney a una coda ai punteggi di intelligibilità trasformati in RAU fra le condizioni a coppie di rumore co-locato con l'ascoltatore o spazialmente separato. Il vantaggio di intelligibilità per sorgenti spazialmente separate si riscontra solo quando il target è più lontano dall'ascoltatore. In particolare, i punteggi RAU aumentano quando l'azimut del rumore è 120° rispetto a 180° nei test AV-SM e AO-S. Lo stesso si verifica quando l'azimut del rumore a 120° rispetto che a 0° in condizione AV-SM. Il vantaggio della separazione spaziale nel riverbero è più evidente nell'impostazione AV-SM, che è la condizione più ecologica tra i nostri test. Ciò suggerisce che il vantaggio di intelligibilità è evidente anche nelle condizioni di ascolto più ecologiche e vicine al mondo reale.

Tabella 2 Confronto fra l'intelligibilità in RAU nelle scene con parlatore target di fronte e altoparlanti laterali e sorgente di rumore informativo a diversi azimut e distanze dall'ascoltatore. Sono riportati i *p-value* inferiori a 0,05 del test di Mann-Whitney a una coda che indica il rifiuto dell'ipotesi nulla H0: MX1≥MX2 in favore dell'ipotesi alternativa H1: MX1< MX2, dove M sono le mediane delle due distribuzioni X1 e X2, rispettivamente.

Scena	T, LS~8m, N@180° vs T, LS~8m, N@120°	T, LS~8m, N@0° vs T, LS~8m, N@120°
AV-SM	0.000	0.000
AV-S		
AO-SM		
AO-S	0.000	

4. Ringraziamenti

Gli autori ringraziano il personale tecnico del Museo Egizio per la disponibilità accordataci ad effettuare indagini AV.

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] 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.

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

[5] 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.

[6] G. E. Puglisi, A. Warzybok, A. Astolfi, and B. Kollmeier. Effect of reverberation and noise type on speech intelligibility in real complex acoustic scenarios. *Building and Environment*, 2021, pp. 108137.

[7] 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.

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

[9] R. Cueille et al., "Effects of reverberation on speech intelligibility in noise for hearing-impaired listeners" *Royal Societ Open Science*, vol. 9(8), pp. 210342, 2022.