Speech intelligibility in reverberation based on audio-visual scenes recordings reproduced in a 3D virtual environment

Building and Environment 258 (2024) 111554

Contents lists available at ScienceDirect

Building and Environment

journal homepage: www.elsevier.com/locate/buildenv

Speech intelligibility in reverberation based on audio-visual scenes recordings reproduced in a 3D virtual environment

Angela Guastamacchia [a,*], Fabrizio Riente [b], Louena Shtrepi [a], Giuseppina Emma Puglisi [a], Franco Pellerey [c], Arianna Astolfi [a]

[a] Politecnico di Torino, Department of Energy, Corso Duca degli Abruzzi, 24, Turin, 10129 TO, Italy
[b] Politecnico di Torino, Department of Electronics and Telecommunication, Corso Duca degli Abruzzi, 24, Turin, 10129 TO, Italy
[c] Politecnico di Torino, Department of Mathematical Sciences, Corso Duca degli Abruzzi, 24, Turin, 10129 TO, Italy

ARTICLE INFO

Keywords:
Speech intelligibility
Self-motion
Visual cues
Audio-visual recordings
3D virtual environment
Spatial release from masking

ABSTRACT

Audio-visual scenes were collected in a medium-sized reverberant conference hall through in-field 3rd-order ambisonics impulse response recordings and 360-degree stereoscopic videos. The visual scenes included cues of the room and the location of the sound sources, without lip-sync-related cues. Speech intelligibility tests based on seven audio-visual scenes were administered inside an immersive virtual 3D environment reproduced through a spherical 16-speaker array synched with a head-mounted display. Forty normal-hearing subjects were engaged to test the effects on speech intelligibility of a talker in front of the listener and amplified by two lateral symmetrical loudspeakers, in the case of (i) different listener-to-talker distances, (ii) one-talker noise at various azimuth angles around the listener, (iii) high reverberation with –5 dB signal-to-noise ratio, (iv) self-motion, and (v) visual cues. We conducted tests in four configurations, that is, audio-visual and audio-only, both with self-motion and in the static condition. The static audio-only tests scored the highest speech intelligibility, followed by a tie between audio-visual with self-motion and in the static condition. Speech intelligibility decreased as the target-to-listener distance increased in all the noisy scenes. Additionally, speech intelligibility increased when the noise azimuth was at 120° compared to both 180° and 0° , with the talker at approximately 8 m from the listener. The advantage of the spatial separation of the noise signal in reverberation is evident in the case of the audio-visual with self-motion test. This suggests a spatial release from masking in the presence of reverberation, one-talker-interfering noise and within an more ecological scene.

1. Introduction

Speech Intelligibility (SI) is the primary acoustic objective in small and large classrooms, conference and court halls, but also in eating establishments, tube and rail stations, airport hallways, etc., where the people's task is speech communication. SI tests are typically performed in laboratories, which should accurately reproduce real-life acoustic scenes to ensure ecological validity of the outcomes [1]. Thus, the challenge is to recreate Audio-Visual (AV) scenes in which participants feel fully immersed in the virtual space and behave as if they were actually present in the environment (i.e., recalling natural eyes, head, and torso movements that help the listener maximize speech recognition [2]). This holds even more when participants are hearing-aid users, being the directional filtering embedded in hearing devices strongly dependent on the listener's head orientation [3]. In order to improve the realism of the provided scenes, these should preferably be composed starting from audio and video recordings of real-life communication scenarios rather than from simulations. Although simulations would be the best fit for research, allowing to quickly and frequently modify the implemented scene at need, some studies have pointed out that subjects prefer real videos over renderings of virtual characters [4,5]. In fact, while video recordings are less flexible than simulations, they prove to be more efficient in situations demanding a high degree of realism, especially when the scene is relatively simple (few actors, vehicles, etc.) [6].

Nevertheless, fostering a life-like listener's Self-Motion (SM) by reproducing realistic immersive AV environments might be insufficient to ensure true ecological listening tests. Indeed, although Grimm et al. [2] found that SM is essential to ensure greater ecological validity, they also stated that SM might not be relevant for the actual completion of the task, as it depends on the environment, age, noise level, task, and instructions. Thus, in the future, the ecological validity of the SI tests should be further boosted with the inclusion of real-time social interactions between the speaker and the listener. Still, it was found

* Corresponding author.
E-mail address: angela.guastamacchia@polito.it (A. Guastamacchia).

https://doi.org/10.1016/j.buildenv.2024.111554
Received 12 February 2024; Received in revised form 12 April 2024; Accepted 21 April 2024
Available online 23 April 2024

that head orientation effectively contributes to the Signal-to-Noise ratio (SNR) enhancement, as participants oriented themselves in a way that resulted in higher SNRs. Furthermore, in this context, visual cues might play a dual role, supporting sound source localization and acting as potential distractors, as it is common to look away from the target talker at times during multi-talker conversations [7].

However, beyond SM, visual cues have been proven to affect ecological validity to different extents: contextual and source-related visual cues affect localization [8], acceptation of the auditory illusion [9], and SM [5], while seeing face and mouth movements of the talker heavily supports speech comprehension [10,11]. These reasons are at the base of recent studies that explored the role of visual cues in SI tests. In [12,13], virtual renderings of contextual and source-related were proposed. In [12], one anechoic and one reverberant scenario were simulated that showed a surrounding virtual ring of loudspeakers to indicate the possible multi-talker noise locations, while the target speech was presented without any visual correspondence. Results showed worsened SI scores for the reverberant condition, longer target-to-receiver distances, and a higher number of masking noise sources, but no significant differences were found between Audio-Only (AO) and Audio-Visual (AV) tests. Similarly, in [13], a reverberant AV scenario was proposed, with a fixed frontal interfering talker and a target talker changing among four positions around the listener. The interfering source had no visual counterpart, while a static avatar picture represented the target sound. The speech-in-noise test was conducted for three different administration conditions, i.e., AV and AO conditions, both allowing SM, and AO in the static condition (S), to investigate whether, at first, people would rotate to improve SI independently on the provided visual cues, and, in case, if seeing the target location would further contribute to the SI enhancement. Significant differences were only found in the case of the target talker either at 90° or -90° azimuth, where the S condition led to the best SI ratings followed by the AV one. In this case, the SI improvement brought by the AV condition w.r.t. the AO one suggests that participants were likely to use the visual cue of the location to alter rotation patterns for their benefit exploiting SM.

With the aim of fostering ecological auditory research while facilitating exchange between laboratories, an open-source database of AV environments was recently published [14]. In [1], the first contributions are presented, which involve in-field multi-channel recordings of Room Impulse Responses (RIR) to auralize SI tests coupled with virtual renderings of the visual scenes for three environments. Nevertheless, only a few studies attempted to address SI measurements exploiting real recordings of the visual scene, also accounting for the effect of lip-reading the target source. In particular, in [15], SI tests were carried out displaying a 360° video of a café scenery showing a frontal conversational partner and chatting customers in the background. The target speech was reproduced by a frontal loudspeaker, while a generic café background noise was emitted by four surrounding loudspeakers. Participants improved their speech recognition with the provision of visual cues, and the results were in line with the literature. However, it seems that this study did not account for the true acoustical conditions of the displayed environment, only presenting anechoic speeches with unmatched background noise. In [16], a one-talker video recording was blended inside a 360° video to account for the effect of the lips movement for the target source. Nevertheless, the inserted masking noises (either 2-talker or speech-shaped masking noise) had no visual counterpart. By comparing the same SI test in AO and AV conditions, the SRT50 scores showed an improvement of up to about 9 dB in the AV case when the speaker and the noise source were co-located. The Speech Reception Threshold (SRT) is the signal to noise ratio (SNR) yielding a certain percentage of correct recognition scores, such as 50% speech intelligibility which is denoted as SRT50 (in dB SNR). Despite these efforts, researchers need to address self-motion and visual counterparts more deeply, especially when virtual reality is concerned, involving body movement and all kinds of possible visual cues to recreate more and more ecological scenes. As far as SM is concerned, no definitive conclusions have been drawn about its advantage in spatialised SI tests. On one side, head orientation towards the higher SNRs could effectively contribute to SI enhancement, but in the presence of visual cues, these could be distractive and result in decreased SI. Conversely, visual cues could be positive for the localization of sound sources and auditory immersion but could be even more distractive, especially without the target talker's lip-sync cues. To the Authors' knowledge, there are no systematic studies that addressed the effects of SM and AV conditions, both separately and together, to assess their effect on SI. It is expected that for a normal hearing subject, in a common listening condition in which the listener is facing the target and hence maximizes the SNR, SM can reduce SI when only audio is presented to the listener due to the possible shifting of the focus from the target. In this listening condition, it is either expected that SM with AV could improve SI compared with AO with SM, due to the relevance of visual cues in connection with SM to recreate a more natural scene, or it could worsen SI because visual cues might increase the distractibility.

In complex listening scenarios, with a target speaker and several speech sources at different azimuths and distances from the target, it is often investigated the effect of Spatial Release from Masking (SRM). The SRM in the context of speech perception denotes the improvement of SI in noise when the speech and the noise sources are spatially separated [17]. SRM with high reverberation has been investigated since the Seventies, but its advantage in these challenging conditions has not been fully explored and only a few studies have considered very high reverberation times. Furthermore, the investigation of SRM within immersive virtual reality needs to be deepened. The influence of visual cues and self-motion in SRM has not been considered so far and this work aims to explore the advantage of the spatial separation of the noise signal for SI in a more ecological setting, which represents a frequently attended environment for normal hearing and hearing impaired persons [18]. Plomp (1976) [19] found that with 2.5 s of reverberation time and a listener 2 m away from the speaker, the SI did not change for informational competing noise (connected discourse) coming from azimuth varying from 45° to 180°. Conversely, Kidd et al. [20] found an SRM of up to 15–17 dB in both dead and reverberant environments when the masker was informational. This last result suggests that the process responsible for improving SI in the spatially separated condition does not depend on the corruption of binaural information. What seems is that listeners use the "precedence effect" to perceive the target as distinct from the masker in reverberant conditions. According to this effect, localization information is derived from signal transients in a very insensitive way to reverberation [21]. Hui et al. [22] conducted SI tests using an Ambisonic-based sound reproduction system with a 16-channel loudspeaker array and examined the benefits from SRM in two environments with a reverberation time of 0.7 s and 1.8 s, respectively. They also considered the listener at 2 m and 5 m from the target speaker. The noise source was babble noise played from eight azimuth angles. For shorter target speech distance, SI was higher at −45° and 135° azimuth and lower at 0° and 180° azimuth, for both the reverberation conditions. Puglisi et al. [23] investigated the effect of very short and high reverberation times on SI in realistic classroom acoustic scenarios with tests administered via headphones and found that the SRM when the noise source was 1 m far from the listener and at 120° azimuth resulted in significant improved SI by up to about 3 dB SNR in case both of energetic masker in low reverberation and of informational masker in high reverberation, suggesting a perceptual segregation mechanism which sorts out competing voices according to their directions in the least favorable listening situations. From the previous findings, it is clear that the perceptual segregation mechanism allows the detection of the target signal within reverberant and challenging scenarios, but none of the previous studies have explored if the distraction caused by visual cues and self-motion could affect this detection.

In this study, we collected AV scenes in a medium-sized conference room through in-field 3rd-Order Ambisonics RIRs recordings and 360° stereoscopic video shootings. The visual scenes include cues on the spatial location of the sound sources without lip-sync cues. SI tests were administered through a spherical 16-speaker array and a head-mounted display to a sample of normal-hearing subjects. We tested the effect of high reverberation on the SRM in the case of a realistic situation of a target talker in front of the listener and amplified by two lateral symmetrical loudspeakers, one-talker noise around the listener, different listener-to-target talker distances and the influence of video recordings and SM.

The research questions to which we want to answer are the following:

Q1 : In a reverberant virtual sound environment, does the head SM affect SI when different spatial configurations for the one-talker interfering noise are presented?

Q2 : When contextual and source positional visual cues are presented together with the virtual sound environment, does SI change with and without SM?

Q3 : In a reverberant virtual sound environment, is the SRM detectable in the presence of one-talker interfering noise at different distances from the target source, with and without SM and with and without visual cues?

2. Method

2.1. Participants

Forty normal-hearing naive native Italian speakers (30 males and 10 females) aged 22 to 46 years (average of 28.4 years, standard deviation of 4.2 years) were voluntarily recruited for the ecological SI test and rewarded with candies, water bottles, block notes, and pens. All of them were previously screened through a pure-tone audiometry test to ensure none of them had a hearing loss, potentially invalidating the test results. A maximum threshold of 16 dB for the average hearing loss at each ear from 500 Hz to 4 kHz was chosen as inclusion criteria. The participants had either regular vision or vision corrected to normal, and they did not exhibit any conditions that might have impacted their movement. Indeed, prescription glasses were allowed during the experiments, as it would not compromise the test.

2.2. AV scenes

A highly reverberant conference hall of the Egyptian Museum of Turin was chosen to record the AV scenes. It represents a typical room with adverse acoustics where good speech comprehension is instead highly required. The hall, which is acoustically untreated, has a volume of 1500 m^3 and is furnished with 100 light chairs and two wooden tables, one above a 30 cm high wooden stage in the front for the main talker and one in the back for the control station of the two-loudspeaker amplification system. Fig. 1 shows the 3D model of the hall with the position of the loudspeakers, at a height of 1.7 m measured from the center of the loudspeaker array, and the talker at the front table at 1.5 m from the floor.

Seven scenes representing typical communication situations inside the room were defined and collected, each with a different spatial configuration for the listener, the target talker, and the interfering talker identifying the competitive noise source. Additionally, to faithfully convey the realistic usage of the conference hall, the target speech was always presented as amplified by the two room loudspeakers on the side walls. Fig. 2 outlines the conference hall floor plan showing the locations of the room loudspeakers (LS) and the positions conceived for the listener (L), the target (T), and the interfering talkers (N) in all seven scenes. In particular, two listening locations among the audience (in the sitting positions, 1.2 m above the floor) were selected, one closer and one farther away from the front target talker sat behind the first table. Moreover, for each listening position, one-talker interfering noise was presented alternatively from at least two different directions (always at 1.2 m from the floor) to evaluate how speech recognition changes when noise is co-located with the target at 180° or 0° azimuth and when noise is spatially separated at 120° azimuth. Table 1 shows the details of the different scenes.

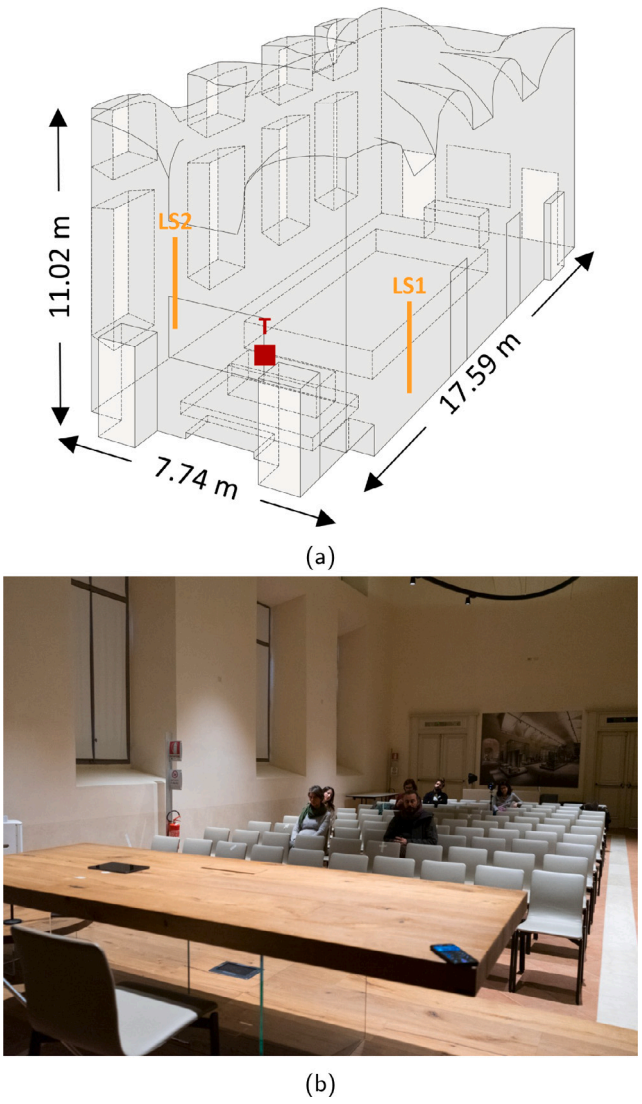


Fig. 1. (a) 3D model of the conference hall where LS are the two room loudspeakers and *T* is the position of the target talker. (b) Picture of the conference hall taken with the same orientation as the 3D model above.

2.2.1. AV scenes acquisition

In order to capture the AV scenes, 4 K 360° stereoscopic videos and 3rd-order ambisonics RIRs were acquired, placing the recording systems in the listening positions and the sound source either in target or noise locations, oriented towards the listening point, except for $N2_{0°}$ where the sound source was rotated of 180°. The sound source was the NTi Audio Talkbox acoustic signal generator, characterized by a flat frequency response from 100 Hz to 10 kHz and an energy distribution featuring the same polar diagram of the human voice. The 19-capsule spherical microphone array Zylia ZM-1, with a nominal flat frequency response from 28 Hz to 20 kHz, was used for the audio recordings, while the Insta360 Pro 360° camera was employed for the video acquisition. During the recordings, a few actors were inside the

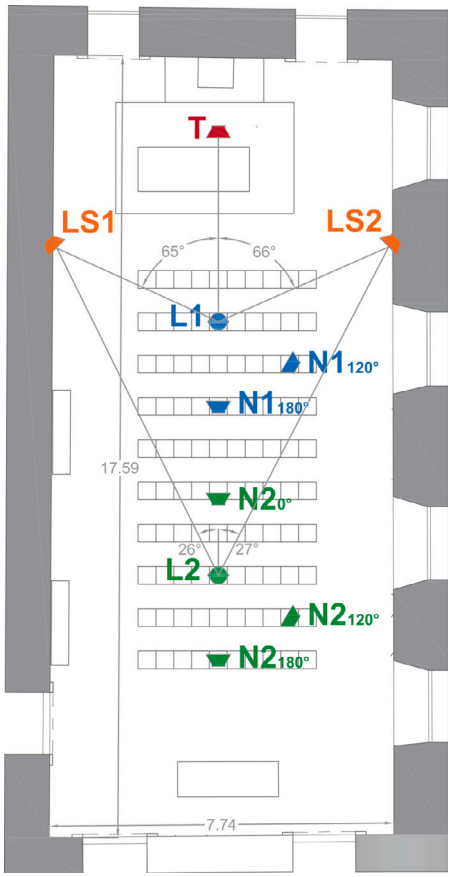


Fig. 2. Conference hall floor plan with locations of loudspeakers (LS1,LS2), target speech (T) and competitive noise ($N1_{120°}$, $N1_{180°}$, $N2_{120°}$, $N2_{180°}$, $N2_{0°}$) sources for all listening positions (L1,L2).

Table 1
List of all scenes with loudspeakers (LS1, LS2), target (T), and interfering talker (N) positions in terms of distance (m) and azimuth angles (clockwise notation) from the listening positions (L1, L2). Noise azimuth and distance fields are signed with N/A (Not Applicable) in case of scenes without masking noise.

Scene number	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Listener location (L)	L1	L1	L1	L2	L2	L2	L2
T azimuth (°)	0°	0°	0°	0°	0°	0°	0°
L-to-T distance (m)	4.1	4.1	4.1	9.8	9.8	9.8	9.8
LS1 azimuth (°)	−65°	−65°	−65°	−26°	−26°	−26°	−26°
L-to-LS1 distance (m)	4.0	4.0	4.0	8.2	8.2	8.2	8.2
LS2 azimuth (°)	66°	66°	66°	27°	27°	27°	27°
L-to-LS2 distance (m)	4.2	4.2	4.2	8.3	8.3	8.3	8.3
N azimuth (°)	N/A	120°	180°	N/A	120°	180°	0°
L-to-N distance (m)	N/A	1.8	1.8	N/A	1.8	1.8	1.8

hall, as it represents a possible worst-case scenario of occupancy as confirmed by the conference hall management.

A total of seven spatial RIRs were collected, with the Talkbox emitting a 5-second-long exponential sine sweep signal from 20 Hz to 20 kHz. Specifically, three RIRs were acquired by placing the receiver in the L1 location and the sound source either in T, $N1_{120°}$, or $N1_{180°}$. Four other RIRs were acquired with the receiver in L2 and the sound source either in T, $N2_{120°}$, $N2_{180°}$, or $N2_{0°}$ positions. For the measurement of the RIRs associated with the Talkbox in the target talker location, the room microphone connected to the 2-loudspeaker system was switched on and placed in front of the Talkbox at a 20 cm distance to include the overall effect of the room amplification system in the sampled RIRs. The directional loudspeakers of the room were the Warray 20 (Webaudio, Govone, CN, Italy), with a frequency range from 100 Hz to 16 kHz. The encoding of the Zylia 19-channel output

Fig. 3. Picture of the RIR recording procedure in case of Zylia placed in the farthest listening location w.r.t. the target speaker and the Talkbox placed in the respective 120° azimuth noise location.

Fig. 4. Picture of the video recording procedure in case of the 360° camera placed in the listening location closest to the target speaker represented by the Talkbox, and one-talker noise at 180° azimuth represented by the dummy head.

into a 3rd-order ambisonics signal was performed in real-time using the A2B-Zylia-3E-Jul2020 19 × 16 filter matrix [24] and controlling the acquisition through a patch of the Plogue Art et Technologie Bidule DAW running on the Notebook PC to which the Zylia was connected. Fig. 3 shows the acquisition of the RIR with the Zylia in the L2 listening position and the Talkbox in the 120° noise source location ($N2_{120°}$).

Concerning the visual scenes, seven 2-minute 3D video recordings were taken placing the 360° camera in the listening positions, while the Talkbox and a Brüel&Kjær 4128 dummy head were placed in the target and interfering talker positions, respectively, to provide the visual reference for the spatial arrangement of the reproduced sounds during the AV SI test. The dummy head was only used during the video recordings to visually differentiate the target talker from the interferer one. Indeed, as aforementioned, for the acoustics RIR acquisition, we chose to use the same sound source, i.e., the Talkbox, in both target speech and noise positions. Videos were first post-produced and then exported in the H.264 format, obtaining, in the end, .mp4 files comprising 4 K 3D 360° videos at 30 fps. The video recording of the scene with the closest target talker and the one-talker interfering noise at 180° azimuth is shown in Fig. 4, while Fig. 5 reports the equirectangular preview of the scene with the closest target talker and the one-talker noise at 120° azimuth.

2.2.2. Acoustical characterization of the AV scenes

In order to acoustically characterize the conference hall in unoccupied conditions, the reverberation time (T_{30}), properly averaged in

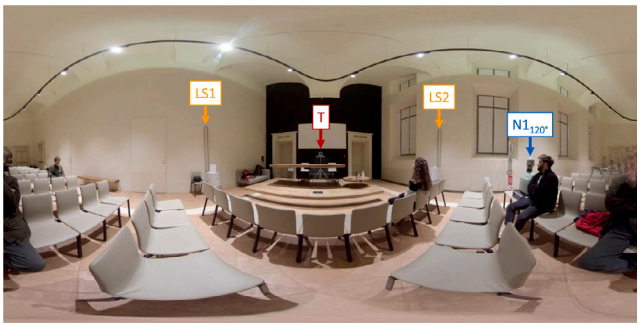


Fig. 5. Equirectangular preview of the visual scene with the listener closer to the target speaker (T) represented by the Talkbox in the front and the one-talker noise at 120° azimuth ($N_{120°}$) represented by the dummy-head.

space and frequency from 250 Hz to 4 kHz, was measured according to the EN ISO 3382-2:2008 standard [25]. The Brüel&Kjær 4292-L omnidirectional sound source driven by the Lab Gruppen LAB300 amplifier and the NTi Audio XL2 omnidirectional class-1 sound level meter were used for the measurements, while the analyses were conducted exploiting the ITA Toolbox [26] open-source MATLAB library. The T_{30} resulted equal to 3.19 s ± 0.44 s, which is 2 s greater than the optimal value for good speech comprehension in small conference halls according to the recent Italian standards on schools [27].

The A-weighted equivalent background noise level was also measured with the sound level meter, which resulted in a value of 39.1 dB, based on an integration time of 3 min.

In order to retrieve an estimation of the SI and target speech levels typically reached in the two listening positions during a conference speech, the STIPA (Speech Transmission Index for Public Address systems) test signal [28] was emitted with an "elevated vocal effort", i.e., measuring 70 dBA at 1 m in anechoic conditions, by the Talkbox placed in the target speech position and amplified by the room loudspeakers (see Section 2.2.1). STIPA values of 0.62 ± 0.01 and 0.55 ± 0.01 and L_{Aeq} of 73.3 dB and 71.8 dB, respectively, were measured in the two listening positions with the sound level meter. STIPA approaches the optimal threshold for conference halls of 0.6 [27], thus showing good speech comprehension in quiet conditions.

Binaural parameters in the listening positions were also provided for all the sound sources, directly derived from the 3rd-order ambisonics RIRs, to gather a clearer picture of the characteristics of the auditory scenes. Specifically, binaural RIRs were taken from the output of the IEM plug-in suite Binaural Decoder inserted in the same Bidule Patch used during the 3rd-order ambisonics RIR recordings. Interaural Level Difference (ILD) and Interaural Time Difference (ITD) were computed to investigate the perceived sound localization, while the Inter-Aural Cross Correlation (IACC) was evaluated to retrieve information on the sound spatial impression. Speech Clarity (C_{50}) and Direct-to-Reverberant energy Ratio (DRR) were also obtained for the two ears and as the average based on the left and right ear RIRs.

The broadband ITD was estimated using the threshold method described in [29]. The broadband ILD was calculated as the energy ratio between the left and right ear. Since in human auditory perception, the ITD and the ILD are used complementarily for the lower and higher frequency range, respectively, the binaural RIRs were low-pass filtered in the first case and high-pass filtered in the second one, using a 10th-order Butterworth filter with a cut-off frequency of 1300 Hz, as in [30]. The broadband DRR values for the left and right ears were calculated by exploiting the open-source MATLAB toolbox in [31], which includes a function that determines the direct sound as the peak of the squared impulse response and returns the DRR value using a time window of 5 ms centered in the peak to select the direct sound [32]. The binaural speech clarity was computed using the ITA Toolbox library [26] in octave bands and provided as average values from 250 Hz to 2 kHz [27,33]. Differences were also obtained between speech clarity from the target source and each noise source at the left and right ears, respectively, and as mean values between the ears. These differences were then used to evaluate the gap in speech clarity between the target and the noise [34].

Fig. 6. Picture of the Audio Space Lab during the execution of the AV SI test.

2.3. Virtual reality system

The tests were conducted in the Audio Space Lab, i.e., a small sound-treated listening room of the Politecnico di Torino, compliant with the ITU-R BS.116-3 recommendation [35], that hosts a 3rd-order ambisonics audio reproduction system synced with the Meta Quest 2 head-mounted display to create an immersive virtual 3D AV environment. The 16.2 ambisonics playback system [36] comprises a 1.2 m radius spherical array of 16 Genelec 8030B 2-way active monitors, properly filtered and delayed to provide a sweet spot equalized in time, amplitude, and phase from 90 Hz to 20 kHz. The loudspeakers are arranged in three rings: one horizontal ring of eight equally spaced speakers at the ear level (first speaker at 0° azimuth) and two 4-speaker rings at +45° and -45° elevation angle (first speaker at −45° azimuth, with 90° spacing). Two more frontal Genelec 8351 A 3-way active monitors placed on the floor are also used as subwoofers to fill the lower frequency range from 30 to 90 Hz. All-round Ambisonic decoding with Max-rE weighting is used to convert 3rd-order ambisonics input signals into the signals driving the ring of 16-speaker, while the omnidirectional channel of the ambisonics tracks, properly filtered, feeds both the subwoofers. All loudspeakers are connected to the Antelope Orion32 32-channel sound card driven by a high-end desktop PC.

In order to run the whole AV reproduction, three software are used, which exchange data through the Open Sound Control protocol to retain the AV sync. The Bidule DAW is used to implement the real-time audio signal processing for the ambisonic decoding and the sweet spot equalization directly driving the multi-channel loudspeaker system, with a sampling frequency set to 48 kHz. The Unreal engine by Epic Games [37] is used to handle the playback of visual scenes by streaming 360° stereoscopic videos (resolution: 3840 × 3840, frame rate: 30 fps, codec: H.264) onto the head-mounted display. Finally, a MATLAB routine is implemented to trigger and keep the AV reproduction in sync and collect the outcomes of the performed SI test.

Fig. 6 shows the Audio Space Lab during the test with a participant.

2.4. Material and generation of the AV si test

The audio tracks of the acoustical scenes for the ecological SI tests were pre-computed using a MATLAB routine starting from the RIRs acquired in the conference hall.

The speech corpus used as target speech was taken from the validated, extended version of the Italian Matrix Sentence Test [38], which comprises lists of 5-word sentences uttered by a female speaker. A standardized phonetically balanced speech, spoken by a female talker, commonly used for speech recognition testing [39] was instead used as interfering noise.

The auralized target signals were properly scaled to achieve in the center of the loudspeaker array, i.e., in the listening position, the same signal level measured in the conference hall in the two listening positions (73 dB(A) for the listening position closest to the target source and 72 dB(A) for the farthest one). The in-noise scenes were generated by summing each auralized target sentence with a different clip of the auralized noise speech, imposing a −5 dB SNR. The SNR value was selected to propose a medium challenging acoustical condition, as SNR values around −5 dB correspond to SRT80 in anechoic conditions [38]. Fig. 7 shows the spectra from 180 Hz to 5.6 kHz for two examples of target sentences and clip of noise speech auralized in case of noise at 180° azimuth and target speech at about 8 m distance. The illustrated spectra were computed from the omnidirectional channel of the 3rd-order ambisonics tracks related to the auralized target and noise with SNR equal to −5 dB. It can be seen that both target and noise speeches have very similar spectra and that in the frequency range between 1 kHz and 2.5 kHz and beyond 5 kHz, they get closer, resulting in an advantage for the target speech intelligibility. These are, in fact, frequency ranges that are very important for speech intelligibility [28]. For the sake of brevity, in Fig. 7, only the spectra for one target-noise source configuration are presented. However, all other configurations followed the same spectral trend.

In addition, to let the participant be ready to listen to the following target sentence, the noise onset was presented a few seconds before the target speech as in [40]. In particular, each track started with 2 s of interfering noise, or silence in the case of in-quiet scenes, after which the 5-word target sentence was presented, and ended with two other seconds of silence or interferer noise, for an overall duration of 6–7 s.

2.5. Experimental procedure

The 40 participants were divided into equal-sized groups of 10 participants characterized by different test administration configurations, which are:

- Audio-Only test with Self-Motion (AO-SM);
- Audio-Only test in the Static condition (AO-S);
- Audio-Visual test with Self-Motion (AV-SM);
- Audio-Visual in the Static condition (AV-S).

Before starting the experiment, participants underwent a training procedure to familiarize themselves with the system used to reproduce the scenes and the SI test. For the S administration conditions, participants were instructed not to turn their heads during the test execution such that the same spatial configuration of target speech and masking noise w.r.t. the listening position was preserved as originally conceived. Conversely, in the case of SM tests, participants were told they were free to move.

For all the test configurations, all seven scenes were presented, auralizing for each scene 20 sentences taken from a different list of the speech-in-noise test. The order of the scenes was randomized and counterbalanced across participants.

SI tests were administered in the open form, that is, with the listener repeating aloud the words she/he understood and the experimenter taking note of the correct ones. In general, the entire test lasted about 35 min per participant, with a 5-minute break after the first 15 min of the test. The experimental procedure used during the tests received ethical approval (reference 100993/2023).

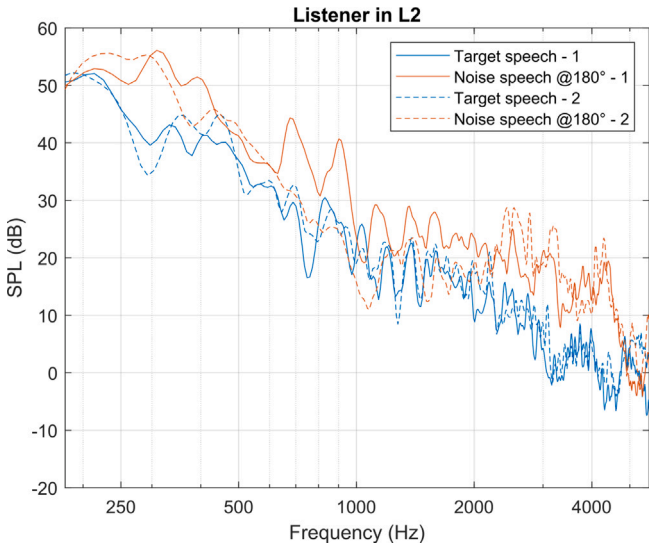


Fig. 7. Examples of auralized target and noise speech spectra with 5 dB SNR for the scene with the listener in L2 location and noise at 180° azimuth.

2.6. Outcome measures and statistical analysis

To evaluate the participants' ability to successfully recognize the target speech inside the proposed scenes the SI scores were collected, defined as the percentage of correctly understood w.r.t. the overall [28]. Furthermore, to account for the ceiling effect, the data were transformed into Rationalized Arcsin Units (RAU) according to the definition in [41] before performing the statistical analysis.

The combined effect of noise azimuth, SM, target distance, and visual cues on the SI outcomes were evaluated through a Linear Mixed Effects model [42], run with IBM SPSS statistics package version 21.0 (Armonk, NY), in which these different conditions and their interactions were assumed to be categorical fixed effects, whereas the subjects were considered as random effects. The term "mixed" refers to the use of both fixed and random effects in the same analysis. In this case, the model for a response variable $Y_{i,j}$, which depends on the ith subject in a scenario j described by the different values assumed by the fixed effects, can be expressed in the form:

$$Y_{i,j} = (a + \alpha_i) + \mu_{1,j} + \mu_{2,j} + \cdots + \mu_{n,j} + \epsilon_{i,j}, \quad (1)$$

where a is the fixed part of the intercept, α_i is the random part of the intercept due to the subject, $\mu_{k,j}$ is the contribution on the value of $Y_{i,j}$ due to the fixed effect k or to the interaction among different fixed effects, in condition j, and $\epsilon_{i,j}$ is the residual or unexplained variation, which is also considered a random effect. The analysis was fitted using Restricted Maximum Likelihood. The importance of each single fixed effect and their interactions was evaluated through the significance of a Type III F test [43]. The standard deviations of α_i and $\epsilon_{i,j}$, i.e., σ_α and σ_ϵ, respectively, were estimated to evaluate the relevance of the subjects in the variability of the SI scores. σ_α represents the general variability between subjects among the four test administration configurations, which include seven scenes for 40 subjects. σ_ϵ represents the variability of the SI score around the regression line for subject i in a condition j, evaluated through the repeated measures for the same subject, which correspond to the 20 sentences for the same scene. The non-parametric Kruskal–Wallis test and Mann–Whitney U-test [44] were applied to compare SI RAU in different auditory conditions since the assumption of normality in the distribution of the scores was violated.

Table 2
Time and level differences between all sound sources (in T, LS1, and LS2 locations) for both listening positions (L1 and L2) used to retrieve the RIRs for the target speech in the proposed auditory scenes.

Sound source		Listening position			
X1	X2	L1		L2	
		$\Delta L_{X2\text{-}X1}$ (dB)	$\Delta t_{X2\text{-}X1}$ (ms)	$\Delta L_{X2\text{-}X1}$ (dB)	$\Delta t_{X2\text{-}X1}$ Δt(ms)
T	LS1	−0.4	2.40		
LS1	LS2	−2.1	1.10	2.1	0.58
LS2	T			−11.8	1.21

3. Results

3.1. Binaural room acoustical parameters

In order to check the direction from which the target speech would be heard during the SI test, being it emitted from three sound sources active at the same time, the time and level difference between all sound sources were measured by analyzing the spatial RIR in time (band-pass filtered from 180 to 5600 Hz) as shown in Fig. 8. The view captured with the 360° camera from both listening positions is illustrated at the time instants when the RIR emitted from T, LS1, and LS2 sound source locations arrive, coupled with the SPL color map [45,46] showing the direction of the incoming sound. As it can be seen from Fig. 8, in the case of the listener in the L1 location, the first sound reaching the listener comes from the Talkbox in location T, while the second RIRs comes from LS1 and arrives after 2.4 ms from the first RIR, and, finally, the last RIR comes from LS2 after 1.1 ms from the LS1 one. Instead, for the L2 listening position, the first approaching RIR comes from LS1, followed by the LS2 source after 0.58 ms, and the T after 1.21 ms from LS2. Table 2 summarizes the values for the time and level differences between all sources, computed from the RIR plots shown below the equirectangular views (see Fig. 8).

In a typical stereophonic configuration, when two identical signals are presented at the same time from the left and right locations, the auditory event is perceived as coming from a single "phantom source" in the front [47]. Instead, with a short delay, i.e., less than 1 ms, and level differences (ΔL) between the two coherent signals reaching the listener's ear, the single sound is perceived at some intermediate locations between the two sources. The precise location of the sound is determined by the kind of signal, the direction of incidence of the two signals, their intensity level difference, and their difference in time of arrival. This is called the "summing localization" effect [47]. When the time difference (Δt) between two signals is from about 1 ms to 5 ms at the listener's position, the location of the auditory event, perceived as a single phantom source, coincides with the position of the sound source whose signal arrives first. This is known as the "localization dominance" [48] effect. Under this effect, in reverberant conditions, humans localize sounds based primarily on the direction of the preceding direct sound instead of the later-arriving reflections. That is the case of the L1 listening position, for which the Δt between the first arriving signal (emitted from the Talkbox from location T) and the second signal (from LS1) differ by 2.4 ms, while the Δt between the first and the last signal (coming from LS2) equals to 3.5 ms. That means that the direction of the target speech is perceived as coming from the front matching the *T* position when the scenes matching the L1 listener's location are proposed. In the case of the listener placed in the L2 location, with an angle of about 60° between LS1 and LS2 emitting the same broadband signal with a 0.58 ms time difference (left earlier) and 2.1 dB level difference (right louder) in the listening position, the direction of the auditory event is perceived as approximately coming from the center between the two loudspeakers [47].

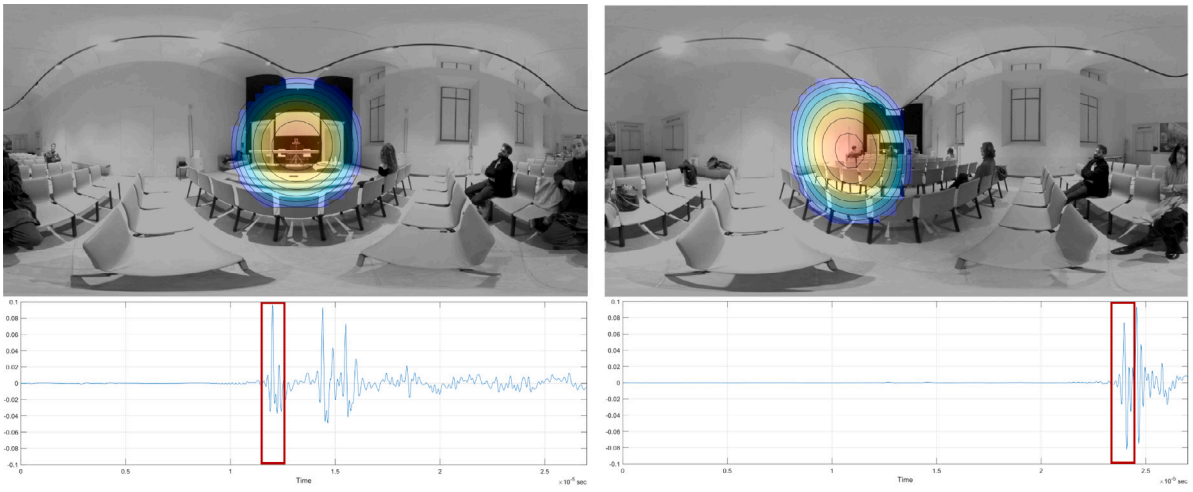
Table 3 collects the values for all the parameters evaluated starting from the binaural RIRs.

Both ITD and ILD values show the expected trend. High values of ITD (540 ms) are found for the noise sources at 120° azimuth, while almost no differences in time delay between the two ears are found in the cases of co-located noise (180° or 0°). While, concerning the binaural RIRs associated with the target, the ITDs confirm the outcome of the analysis on the first arriving sound. In the case of the listener in the L1 location, the ITD is equal to about −20.8 ms, meaning that the first sound reaching the listening position comes from the frontal direction, while for the L2 location, the ITD equals to about -360 ms, pointing out that the first direct sound matches the left loudspeaker location LS1. Similarly to the ITD, the ILD shows higher values in case of separated noise, i.e., ILD equal to about −6 dB for sound coming from 120° on the right. For noise coming from 180° azimuth, the level difference between the two ears is low but slightly higher than the noise at 0° as happens in [30]. However, the ILD for the 120° noise azimuth is half the value of the corresponding ILD measured in [30], which may be due to the higher reverberation time involved in this study (3.2 s versus 1.2 at mid-frequencies). Finally, for both L1 and L2 listening positions, the target binaural RIR shows negligible values for the ILD, further suggesting the listeners would perceive the target speech as coming from the frontal direction.

Concerning the IACC, which is strongly influenced by reverberation, source-to-receiver distance, and angular displacement [30], it presents lower values in the case of spatially separated sound sources, i.e., noise at 120° azimuth, w.r.t. co-located ones, i.e., noise at 180° and 0° azimuth, for the same source-to-receiver distance, i.e., 1.8 m. While, for the binaural RIRs associated with the target source coming from the frontal direction, very low values, i.e., 0.2 and 0.3 for the L1 and L2 listening position, are found, which are due to the combined effect of the high reverberation and longer source-to-receiver distances (about 4 and 8 m).

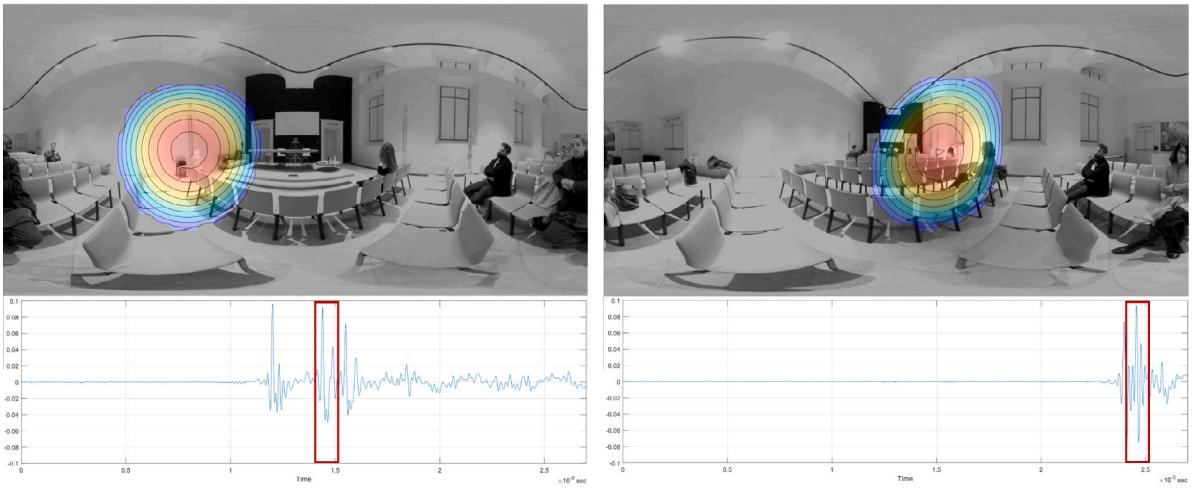
In order to identify the impact of the reverberated components for each sound (target or noise) approaching the listening positions, the left and right ear DRR values are analyzed. In the case of noise at 120° azimuth, the direct component is predominant on the right ear, showing values equal to about 6.8 dB, while, for the noise at 180°, small differences are found between the DRR at the left and right ear. Here, the mean DRR equals about −1 dB, meaning the energy associated with the reverberant sound is slightly higher than the one of the direct sound. Concerning the binaural RIRs for the target sound, similar values are found between the two ears, but different values of mean DRR are found between the L1 and L2 listening positions. Particularly, when the listener is closer to the target, the energy of the direct sound is almost equal to the energy of the reverberated part, while, in the case of the farthest location, the reverberated component results in being 5 dB higher than the direct one. A similar result holds for the noise at 0° azimuth, for which the mean DRR is about −7.8 dB, despite the noise being presented from the same distance as the noise at 120° azimuth. This is due to the orientation of the noise source that is turned 180° from the listener. Thus, the sound reaching the listening position is almost entirely made of the reverberant component.

Finally, the binaural values computed for the C_{50} relate with the DRR ones, following the same trend. Higher C_{50} values are achieved at the right ear in case of noise source at 120° azimuth, with a high difference between the right and left ear. At the same time, lower values are found for the noise at 180° and 0° and the target sound. In particular, a higher mean C_{50} value is found for the listening position closer to the target compared with the farthest one. The worst C_{50} corresponds to the noise as 0° azimuth with a value of −3.5 dB. The differences between the C_{50} values for the target and the noise presented in each in-noise scene are outlined in Table 4. In the case of scenes 2 and 5, with the noise from 120° azimuth, the clarity of the noise at the right ear is 5 and 7 dB higher than the target for the L1 and L2 listening positions, respectively. For all other scenes, the difference in clarity is roughly the same between the two ears. In the case of the listener in the L1 location, the target clarity is higher than the noise at 180°, while the inverse occurs for the L2 listening position. Lastly, the higher target clarity compared with the noise is achieved in the L2 location when the noise is presented from the 0° azimuth.



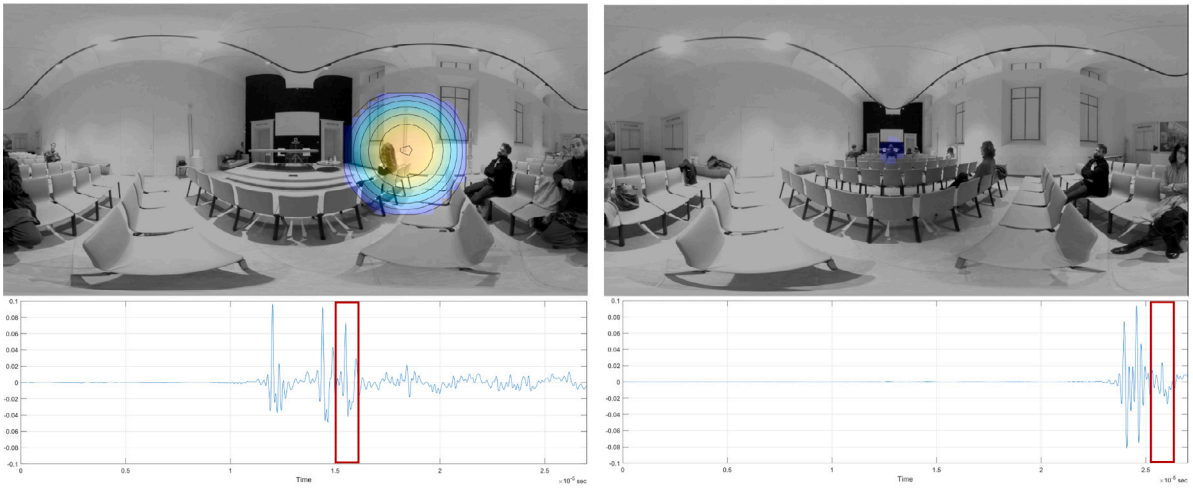
(a) L1 - First RIR coming from T.

(d) L2 - First RIR coming from LS1



(b) L1 - Second RIR coming from LS1

(e) L2 - Second RIR coming from LS2



(c) L1 - Third RIR coming from LS2

(f) L2 - Third RIR coming from T

Fig. 8. SPL color map on the equirectangular view from the listening positions L1 and L2, showing the time history of the direction of arrival of the normalized RIR emitted from the Talkbox in position T and amplified by LS1 and LS2. Below the equirectangular view, the first 27 ms of the RIR captured by the omnidirectional channel of the ambisonics track is shown.

Table 3
ITD (low-pass filtered with a cut-off frequency of 1.3 kHz), ILD (high-pass filtered with a cut-off frequency of 1.3 kHz), broadband IACC, left ear, right ear and within-ear mean DRR (broadband) and C_{50} (average from 250 Hz to 2 kHz) values computed from the binaural RIRs derived from the 3rd-order ambisonics recordings for the auditory scenes collection.

	Active sound source positions						
	T@4.1 m,0° LS1@4 m,-65° LS2@4 m,66°	N@1.8 m,120°	N@1.8 m,180°	T@9.8 m,0° LS1@8.2 m,-26° LS1@8.3 m,-27°	N@1.8 m,120°	N@1.8 m,180°	N@1.8 m,0°
ITD (μs)	−20.8	540.0	0.0	−360.0	540.0	20.8	0.0
ILD (dB)	0.7	−6.0	−1.1	0.5	−6.3	1.3	0.7
IACC (–)	0.2	0.4	0.7	0.3	0.4	0.7	0.6
Left C_{50} (dB)	1.8	1.5	−0.1	−1.4	0.6	−0.2	−3.6
Right C_{50} (dB)	1.1	6.7	0.2	−1.1	6.0	0.0	−3.5
Mean C_{50} (dB)	1.5	4.9	0.0	−1.0	4.2	−0.1	−3.5
Left DRR (dB)	0.0	−0.4	−2.0	−5.0	0.1	−0.4	−7.7
Right DRR (dB)	−1.1	6.7	−0.8	−4.6	6.8	−1.6	−8.0
Mean DRR (dB)	−0.5	4.5	−1.3	−4.8	4.6	−0.9	−7.8

Table 4
Difference between target and noise binaural RIRs speech clarity averaged from 250 Hz to 2 kHz.

Active sounD Sources	Left C_{50} (dB)	Right C_{50} (dB)	Mean C_{50} (dB)
T @ 4.1 m, 0° LS1 @ 4 m, −65° LS2 @ 4 m, 66° N @ 1.8 m, 120°	0.3	−5.6	−3.4
T @ 4.1 m, 0° LS1 @ 4 m, −65° LS2 @ 4 m, 66° N @ 1.8 m, 180°	1.9	0.9	1.5
T @ 9.8 m, 0° LS1 @ 8.2 m, −26° LS2 @ 8.3 m, 27° N @ 1.8 m, 120°	−2.0	−7.1	−5.2
T @ 9.8 m, 0° LS1 @ 8.2 m, −26° LS2 @ 8.3 m, 27° N @ 1.8 m, 180°	−1.2	−1.1	−0.9
T @ 9.8 m, 0° LS1 @ 8.2 m, −26° LS2 @ 8.3 m, 27° N @ 1.8 m, 0°	2.2	2.4	2.5

3.2. Speech intelligibility in the different scenes

Fig. 9 shows the mean and the standard deviation for the SI percentage scores achieved in each test administration condition (AO-SM, AO, AV-SM, AV-S) for each scene.

Significances resulting from the Type III F-test on the fixed effects and their interactions [43] pointed out that the target distance *T* (sig. = 0.000) and noise azimuth NA (sig. = 0.004) are significantly predictive of SI, while visual cues AV (sig. = 0.648) and self-motion SM (sig. = 0.111) are not significant themselves but in their interaction, such as SM*T (sig. = 0.072), T*NA (sig. = 0.001), AV*SM*T*NA (sig. = 0.001).

The standard deviation σ_α, which represents the general variability between subjects among all the test configurations, is equal to 9.7, while the standard deviation σ_ε, evaluated through the 20 sentences for the same subject, is equal to 29.2. The former is lower than the latter, and this reveals that the inter-subject variability is significantly lower than the intra-subject variability. Based on this outcome, the effect of each single subject and her/his variability in repeated measures has not been considered in the comparison among the different test configurations.

Table 5 shows the mean and the standard deviation of the mean SI expressed in RAU for the four test configurations, i.e., AV-SM, AV-S, AO-SM, and AO-S. The Kruskal–Wallis test for independent samples

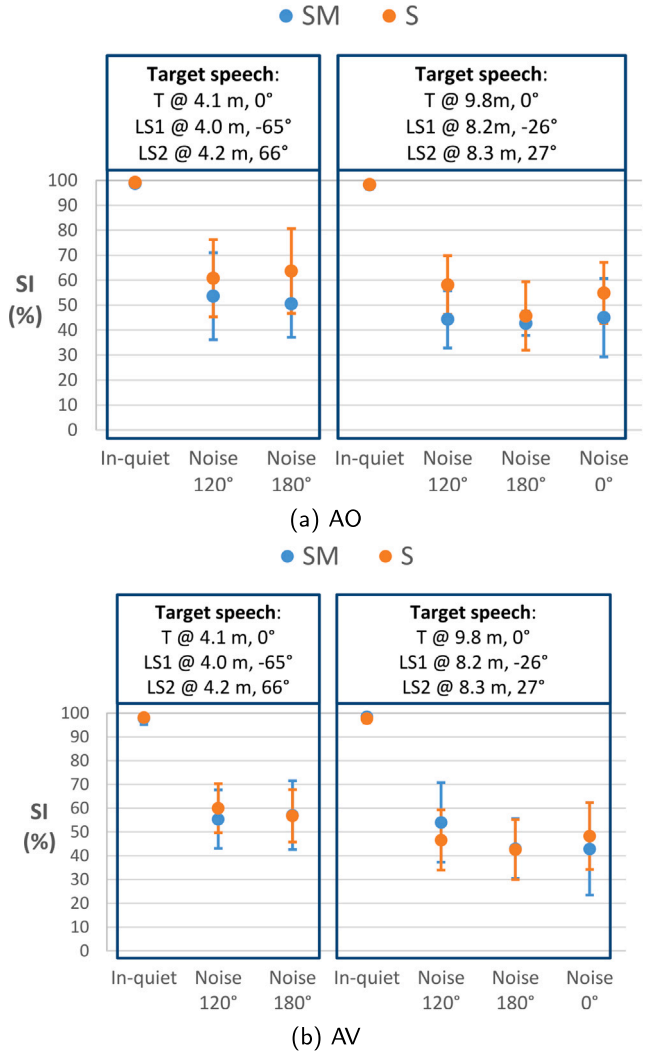


Fig. 9. Mean and standard deviation values of the percentage SI scores for each scene for the comparison between the Self-Motion (SM) and the Static condition (S) in case of (a) Audio-Only and (b) Audio-Visual tests.

refuses the null hypothesis of the same distribution across the cases (*p*-value = 0.00), and from Table 5, it is evident that the highest score is for the AO-S condition. When the test is carried out for the other three conditions, i.e., AV-SM, AV-S, and AO-SM, the null hypothesis of the same distribution across the cases is refused again (*p*-value = 0.031).

Table 5
Mean and Standard Deviation (SD) of the SI scores in RAU for each test configuration (AV-SM, AV-S, AO-SM, AO-S).

Test configuration	N	Mean	SD
AV-SM	1000	50.99	32.64
AV-S	1000	51.18	31.09
AO-SM	1000	47.81	29.88
AO-S	1000	**57.16**	30.59

Indeed, looking at Table 5, the AO-SM condition achieves a lower mean SI score than the other two test configurations. Finally, when the same test is applied only on AV-SM and AV-S tests, the null hypothesis of the same distribution across the cases cannot be rejected (p-value = 0.737), pointing out that, overall, the AO-S test condition leads to the best SI mean score, followed by the AV-SM and AV-S conditions in a tie, and by the AO-SM as test leading to the worst mean SI score. The aforementioned analyses were carried out involving all auditory scenes except for the ones in in-quiet auditory conditions.

Table 6(a) shows the results from the U-Mann Whitney analyses where, for each scene, the effect of SM has been investigated, with and without visual cues. In particular, the comparison between the test configurations for each scene is reported, i.e., AV-SM versus AV-S, AO-SM versus AO-S, and AV-SM versus AO-S. Table 6(b) explores the differences between the auditory scenes (different listener-to-target distance and noise conditions) for each test configuration. Cases of p-values lower than 0.05 are reported in bold and indicate the rejection of H_0: $M_{X_1} \geq M_{X_2}$ in favor of the alternative hypothesis H_1: $M_{X_1} < M_{X_2}$, where M_{X_1} and M_{X_2} are the medians of RAU distributions in the conditions X_1 and X_2, respectively. Values lower than 0.05 are reported in bold and italic style and indicate the rejection of H_0: $M_{X_1} \leq M_{X_2}$ in favor of the alternative hypothesis H_1: $M_{X_1} > M_{X_2}$.

Table 6(a) shows that the SI scores with SM were lower than without SM, i.e., in S condition, for scenes 1, 2, 4, and 6, while for scenes 3, 5, and 7 no significant differences were found. Conversely, in the AV case, the SM scored better than the S condition in scene 5, while no significant differences were found for all other scenes. From the comparisons between the AO-S and the AV-SM in Table 6(a), it results that the former performs better than the latter in 4 out of 7 scenes, i.e., for scenes 1, 2, 3 and 7.

From the comparison between scenes 6 and 5, as shown in Table 6(b), when the target is at about 8 m from the listener, the SI scores increase when the noise azimuth is 120° compared to 180° in AV-SM and AO-S tests. This does not hold when the target is at about 4 m from the listener, as it is pointed out from the comparison between scenes 3 and 2. The same SRM is true in the case of scene 7 versus 5, that is, when the target is farther away from the listener, the SI increases for the noise azimuth at 120° compared to 0° in AV-SM test configurations.

Table 6(b) also shows a significant increase in SI from scene 6, with noise at 180° azimuth, to scene 7, with noise at 0°, with the target at about 8 m from the listener. This can be explained by the difference between the binaural C_{50} related to the target at 0° and the noise at different azimuths, as shown in Table 4. When the target and the noise are at 0° azimuth, the average speech clarity difference between the two conditions is 2.5 dB, while when the target is at 0° azimuth and the noise and 180°, the average speech clarity difference is −0.9 dB. This means that, with noise at 0° azimuth, the target is clearer than the noise, while the opposite occurs when the noise is at 180° azimuth. It should be underlined that the noise source, when placed at 0° azimuth, was directed towards the target, with the listener behind. In this way, the listener did benefit from a lower direct sound level and higher diffuse level of the noise source (DRR equal to −7.8 dB) due to the directivity of the Talkbox source [49].

Always in Table 6(b), from the comparison of the SI scores between scenes 4 and 1 with farther and closer distances from the target, no significant differences were found in quiet conditions as suggested by the slight change in the STIPA values. From the comparison between scenes 5 and 2, with the same noise azimuth of 120° and different target-to-listener distances, and between scenes 6 and 3, with the same noise azimuth of 180° and different target-to-listener distances, as expected, SI scores were worse for the farthest listening position, with a more marked effect for 180° noise. This can be explained by looking at Table 4, where the target mean C_{50} gets worse and worse than the noise mean C_{50} as the listener's distance from the target increases. This agrees with Fichna et al. [12] and Puglisi et al. [23]. In particular, Puglisi et al. [23] found higher SRT80s as the target-to-listener distance increased in the case of reverberant and informative noise conditions.

4. Discussion

4.1. Q1

In a reverberant virtual sound environment, does the head SM affect the SI when different spatial configurations for the one-talker interfering noise are presented?

Table 5 shows that in AO, SM brought a decrease in SI compared to the S condition. This is confirmed by looking at Table 6(a), which shows that, for the AO condition, the SM scored less than the S condition for 4 out of 7 scenes.

Head movements may be used to maximize the signal level received at one ear [50] or to maximize the difference between the signal level and the noise level [51]. The latter strategy would generally be more effective and is expected to lead directly to an increase in speech intelligibility, but in our cases, for the AO tests, the SM might have negatively impacted the spatial unmasking and the head orientation benefits. Other studies did not find any significant effect of SM on speech intelligibility. Hladek and Seeber [13] did not show significant differences in SI when the speech noise and the target were at 0° for the three conditions AV-SM, AO-SM, and AO-S, while the AO-S condition determined higher SI when the target was at 90° and -90° and the speech noise at 0°, followed by the AV-SM condition. Frissen et al. [52] studied the effect of speech-irrelevant head movements on speech intelligibility with multiple maskers in the acoustic scene and did not find any significant positive effect of the head movement. Shen et al. [53], when the target and the noise were behind the listener, observed that the head movements from the head turners listeners were unlikely to be initiated to optimize SRM. Furthermore, data suggest a slight improvement in speech intelligibility for non-head turners relative to head turners. The SI increase when the head SM is not allowed can be explained by the increased focus of the subjects towards the target sound rather than when they can move. In support of this, the subjects involved in this study who were not allowed to move their head during the test reported they closed their eyes to increase the focus on the target sound.

4.2. Q2

When contextual and source positional visual cues are presented together with the virtual sound environment, does the SI change with and without SM?

As drawn from the results in Table 5, the AO-S tests perform better than AV tests. This is also confirmed by the comparisons between the AO-S and the AV-SM in Table 6(a), where the former performs better than the latter in 4 out of 7 scenes, i.e., for scenes 1, 2, 3 and 7. Furthermore, from Table 5, it seems that results from AV are equivalent in SM and S conditions, as found in [12], although one could expect that allowing SM during the AV tests should lead to better SI scores than the S case, being the AV-SM the test configuration that gets closer to the real-life listening experience. Indeed, Hladek and Seeber [13] did show that the AV-SM condition performed better than the AV-S one. According to Neidhardt et al. [9], visual cues are relevant in connection with SM. Within Virtual Reality in six degrees of freedom,

Table 6
U-Mann Whitney statistical analyses results for the comparison between: (a) the different test configurations (AV-SM vs AV-S, AO-SM vs AO-S, AV-SM vs AO-S) for each scene, and (b) the different acoustical conditions of the scenes for the same test configuration. Cases of p-values lower than 0.05 are reported in bold and indicate the rejection of H_0: $M_{X_1} \geq M_{X_2}$ in favor of the alternative hypothesis H_1: $M_{X_1} < M_{X_2}$, where M_{X_1} and M_{X_2} are the medians of RAU distributions in the conditions X_1 and X_2, respectively. Values lower than 0.05 are reported in bold and italic style and indicate the rejection of H_0: $M_{X_1} \leq M_{X_2}$ in favor of the alternative hypothesis H_1: $M_{X_1} > M_{X_2}$.

(a)	COMPARISON BETWEEN TEST CONFIGURATIONS (X_1 vs X_2) FOR EACH SCENE						
Scene number	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Active sound source position	T@4.1 m,0° LS1@4 m,−65° LS2@4 m,66°	T@4.1 m,0° LS1@4 m,−65° LS2@4 m,66° N@1.8 m,120°	T@4.1 m,0° LS1@4 m,−65° LS2@4 m,66° N@1.8 m,180°	T@9.8 m,0° LS1@8.2 m,−26° LS1@8.3 m,−27°	T@9.8 m,0° LS1@8.2 m,−26° LS1@8.3 m,−27° N@1.8 m,120°	T@9.8 m,0° LS1@8.2 m,−26° LS1@8.3 m,−27° N@1.8 m,180°	T@9.8 m,0° LS1@8.2 m,−26° LS1@8.3 m,−27° N@1.8 m,0°
AV-SM vs AV-S					***0.011***		
AO-SM vs AO-S	**0.010**	**0.000**		**0.000**		**0.001**	
AV-SM vs AO-S	**0.015**	**0.042**	**0.029**				**0.000**
(b)	COMPARISON BETWEEN SCENES (X_1 vs X_2) FOR EACH TEST CONFIGURATION						
Scene number	*3* (T,LS~4 m, ... N@180°) *vs* *2* (T,LS~4 m, ... N@120°)	*6* (LS~8 m, ... N@180°) *vs* *5* (LS~8 m, ... N@120°)	*7* (LS~8 m, ... N@0°) *vs* *5* (LS~8 m, ... N@120°)	*7* (LS~8 m, ... N@0°) *vs* *6* (LS~8 m, ... N@180°)	*4* (LS~8 m, ...) *vs* *1* (T,LS~4 m, ...)	*5* (LS~8 m, ... N@120°) *vs* *2* (T,LS~4 m, ... N@120°)	*6* (LS~8 m, ... N@180°) *vs* *3* (T,LS~4 m, ... N@180°)
AV-SM		**0.000**	**0.000**				**0.000**
AV-S				***0.048***		**0.000**	**0.000**
AO-SM						**0.000**	**0.012**
AO-S		**0.000**		***0.001***			**0.000**

spatial auditory illusions are effective if they support an interactive listener motion. These expectations are also based on the experiences from everyday listening. However, a reason for which, in this study, the AV-SM tests imply no improvement in the SI scores compared with the AO-S condition might be that one should be able to see lips movements to gain substantial benefit from visual cues [11,54,55] (not included in this study). Nevertheless, the accuracy of speech reading decreases rapidly as a function of the distance from the speaker, and, in our cases, with a distance of about 4 m and 8 m, it is unlikely that it could have brought a significant improvement.

4.3. Q3

In a reverberant virtual sound environment, is the SRM detectable in the presence of one-talker interfering noise at different distances from the target source, with and without SM and with and without visual cues?

Answering this question is possible by observing Table 6(b). As already stated, from the comparison between scenes 6 and 5, when the target is at about 8 m from the listener, (i) the SI scores increase when the noise azimuth is 120° compared to 180° in AV-SM and AO-S tests. The same occurs in the case of scene 7 versus 5, when the target is at the same distance from the listener, (ii) the SI increases for the noise azimuth at 120° compared to 0° in AV-SM condition. The advantage of spatial separation of the noise signal in reverberation is more evident in the AV-SM setting, which is the most ecological condition among our tests. This suggests that the SRM advantage is also evident in settings which are closer to the real world [18].

Looking at the difference between the average binaural C_{50} between the target at 0° azimuth and the noise at 120° azimuth shown in Table 4, it is −5.2 dB, while it is −0.9 dB and 2.5 dB when the noise is at 180° and 0° azimuth, respectively. Thus, results (i) and (ii) are not explained by C_{50} differences between the target and noise.

The process responsible for improving speech intelligibility in the spatially separated condition depends on a perceptual segregation mechanism that sorts out competing voices from the target stream in reverberation, which is identified as the "precedence effect" [18,48,56,57]. Biberg and Ewert [58], observed a SRM of 3 dB with an informational masker, a reverberation time of 3 s, a target-to-listener distance of 6 m and an average binaural DRR of the talker in front to the listener of about −7 dB. The room acoustical conditions were very similar to this study, where we had the target-to-listener distance of 8 m, the average binaural DRR of the talker in front of the listener was about −5 dB and the reverberation time was 3.2 s. Kidd et al. [20] found a spatial release from masking up to 15–17 dB in a reverberant environment when the masker was informational. The signals were played from two loudspeakers at 1.5 m from the listener at the same height. The target source was at 0° azimuth, and the masker was played both at 0° and 90° azimuth. They obtained almost the same SRM for three different rooms where the reverberation increased and the monoaural DRR averaged across the 0° and 90° speaker's locations changed from 16.9 dB, 6.3 dB and −0.9 dB. They thus proved that unfavorable consequences of reverberation on binaural hearing were not relevant to the results and the advantage of spatial separation was preserved. Puglisi et al. [23] found an SRM of about 3 dB for informational masker in high reverberation with the target at 0° azimuth and 1.5 m from the listener and the noise source at 120° azimuth and 1 m far from the listener. Hui et al. [22] examined the benefits of spatial release from masking in one reverberant environment with babble noise played from eight azimuth angles. For shorter distances from the source, speech intelligibility was higher at −45° and 135° azimuth and lower at 0° and 180° azimuth. For longer distances from the source, no significant difference emerged for the different azimuths. Conversely, in our study, the SRM occurs only for the longest distance from the target source, i.e., about 8 m (see scene 6 versus 5 in Table 6(b)), while it does not appear for the shortest one, i.e., about 4 m, (see scene 3 vs 2 in Table 6(b)).

Our study is a step forward in detailing the precedence effect out of traditional laboratories. To the authors' knowledge, none of the previous studies have investigated the effect of SRM with video and self-motion within scenes in which audio and video were recorded from real settings. Furthermore, no study were found that explored the SRM in realistic scenes with one target talker in front of the listener and amplified by two symmetrical lateral loudspeakers. This is a typical listening condition found in conference halls and classrooms where speech is usually amplified. In the review by Brown et al. [18], they underlined the need to have an investigation on precedence effect in more ecological settings with the final aim of understanding the preservation of accurate sound localization in highly reverberant environments which are frequently attended by normal hearing and hearing impaired [48]. Of course, the stimuli used in laboratories have proven to be useful in measuring the connection between temporal and spatial aspects related to auditory perception, which are at the base of the functioning of auditory brainstem circuits. It seems that given the prevalence of reflected sound in the natural world evolution has stimulated precedence effect specifically to overcome the auditory challenges of reflected sound with its suppression [59]. Indeed, even

in strongly reverberant spaces, different sound sources are perceived as separate auditory events in their respective spatial locations, even though head shadow and binaural interaction advantages are reduced, and this occurs especially in reducing the informational type of masking [56]. Results also show that the more distinct competing streams are from one another, the more complete the suppression of the stream in the background is [57]. However, the precise stages of auditory processing involved in this benefit are not fully understood [60].

5. Conclusions

Audio-Visual (AV) scenes were collected in a medium-sized reverberant conference hall through in-field 3rd-order ambisonics impulse response recordings and 360° stereoscopic video shootings. The visual scenes included cues of the room and the spatial location of the sound sources, which are useful for localization and acceptance of the auditory illusion, without lip-sync-related cues. Speech Intelligibility (SI) tests based on those AV scenes were administered through a 3rd-order ambisonics loudspeaker-based audio reproduction system synced with an head-mounted display to reproduce an immersive virtual 3D environment. Forty normal-hearing subjects were engaged to test the effects on SI of a talker in front of the listener and amplified by two lateral symmetrical loudspeakers, in the case of (i) different listener-to-talker distances, (ii) one-talker noise at various azimuth angles around the listener, (iii) high reverberation with -−5 dB Signal-to-Noise Ratio (SNR), (iv) self-motion, and (v) visual cues.

Four test configurations were involved: Audio-Visual tests with Self-Motion (AV-SM) and in the Static condition (AV-S), and Audio-Only tests with Self-Motion (AO-SM) and in the Static condition (AO-S). For each test configuration, seven scenes were proposed with the listener closer or farther from the amplified target speech, either in quiet conditions or with separated (120° azimuth) or co-located masking noise at (180° or 0° azimuth).

The main results are the following:

- the AO-S tests determined the highest SI scores, followed by the AV-SM and AV-S in a tie, and by the AO-SM test that led to the worst SI score.
- SM did not increase SI to a large extent in the AV tests compared to S condition, contrary to what it could have been expected due to the relevance of visual cues in connection with SM to recreate a more natural scene.
- SM reduced SI in the AO condition, thus proving its negative effect due to the shifting of the focus from the target.
- Visual cues without lip-sync-related cues did not increase SI compared to the AO in the static condition, which is the listening condition that brings more focus to the target speech.
- A decrease in SI as the target-to-listener distance increased was found for all in-noise scenes.
- An increase in SI with the noise azimuth at 120° compared to both 180° and 0°, with the target speech at about 8 m from the listener, was evident with Audio-Visual tests with Self-Motion (AV-SM), thus implying a spatial release from masking in the presence of reverberation, one-talker interfering noise and within a more ecological setting. These results suggest that the process responsible for improving SI in the spatially separated condition and reverberation depends on a perceptual segregation mechanism that sorts out competing voices according to their directions.

The results of this study are steps forward in the direction of understanding the complex mechanism at the base of speech comprehension in frequently attended environments, such as conference, court halls and classrooms, but also in eating establishments, tube and rail stations, airport hallways, where high reverberation is detrimental towards the task of guaranteeing a high degree of speech intelligibility. The study outcomes are also useful in the field of hearing research where laboratories for real-life acoustic reproduction are needed to tune hearing devices and to ensure high speech intelligibility to hearing-impaired persons.

CRediT authorship contribution statement

Angela Guastamacchia: Conceptualization, Methodology, Software, Data analysis, Data curation, Investigation, Visualization, Writing – original draft. **Fabrizio Riente:** Writing – review & editing, Supervision. **Louena Shtrepi:** Methodology, Writing – review & editing. **Giuseppina Emma Puglisi:** Conceptualization, Methodology, Writing – review & editing. **Franco Pellerey:** Statistical analysis. **Arianna Astolfi:** Conceptualization, Methodology, Writing – review & editing, Statistical analysis, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Fabrizio Riente reports was provided by Polytechnic of Turin. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors thank the Museo Egizio di Torino and the extras, Stefano Rovera, Ignazio Ligani, Luca Bagetto, and Andrea Albera, for their contributions to the audio-visual scenes.

References

[1] Steven Van De Par, Stephan D Ewert, Lubos Hladek, Christoph Kirsch, Julia Schütze, Josep Llorca-Bofí, Giso Grimm, Maartje ME Hendrikse, Birger Kollmeier, Bernhard U Seeber, Auditory-visual scenes for hearing research, Acta Acust. 6 (2022) 55.
[2] Giso Grimm, Maartje M.E. Hendrikse, Volker Hohmann, Review of self-motion in the context of hearing and hearing device research, Ear Hearing 41 (2020) 48S–55S.
[3] Roohollah Abdipour, Ahmad Akbari, Mohsen Rahmani, Babak Nasersharif, Binaural source separation based on spatial cues and maximum likelihood model adaptation, Digit. Signal Process. 36 (2015) 174–183.
[4] Gerard Llorach, Maartje ME Hendrikse, Giso Grimm, Volker Hohmann, Comparison of a head-mounted display and a curved screen in a multi-talker audiovisual listening task, 2020, arXiv preprint arXiv:2004.01451.
[5] Maartje ME Hendrikse, Gerard Llorach, Giso Grimm, Volker Hohmann, Influence of visual cues on head and eye movements during listening tasks in multi-talker audiovisual environments with animated characters, Speech Commun. 101 (2018) 70–84.
[6] Gerard Llorach, Giso Grimm, Maartje ME Hendrikse, Volker 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.
[7] Maartje ME Hendrikse, Theda Eichler, Volker Hohmann, Giso Grimm, Self-motion with hearing impairment and (directional) hearing aids, Trends Hearing 26 (2022) 23312165221078707.
[8] Axel Ahrens, Kasper Duemose Lund, Marton Marschall, Torsten Dau, Sound source localization with varying amount of visual information in virtual reality, PLoS One 14 (3) (2019) e0214603.
[9] Annika Neidhardt, Christian Schneiderwind, Florian Klein, Perceptual matching of room acoustics for auditory augmented reality in small rooms-literature review and theoretical framework, Trends Hearing 26 (2022) 23312165221092919.
[10] Ken W. Grant, The effect of speechreading on masked detection thresholds for filtered speech, J. Acoust. Soc. Am. 109 (5) (2001) 2272–2275.
[11] Alison MacLeod, Quentin Summerfield, Quantifying the contribution of vision to speech perception in noise, Br. J. Audiol. 21 (2) (1987) 131–141.
[12] Stefan Fichna, Thomas Biberger, Bernhard U Seeber, Stephan D Ewert, 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.
[13] L'uboš Hládek, Bernhard U. Seeber, Speech intelligibility in reverberation is reduced during self-rotation, Trends Hearing 27 (2023) 23312165231188619.
[14] L. Hládek, S. van de Par, S.D. Ewert, B.U. Seeber, audio-visual scenes repository: How to contribute, 2021, http://dx.doi.org/10.5281/zenodo.5532673.

[15] Hye Yoon Seol, Soojin Kang, Jihyun Lim, Sung Hwa Hong, Il Joon Moon, Feasibility of virtual reality audiological testing: Prospective study, JMIR Serious Games 9 (3) (2021) e26976.

[16] Alastair H Moore, Tim Green, Mike Brookes, Patrick A Naylor, Measuring audio-visual speech intelligibility under dynamic listening conditions using virtual reality, in: Audio Engineering Society Conference: AES 2022 International Audio for Virtual and Augmented Reality Conference, Audio Engineering Society, 2022.

[17] Gerald Kidd Jr., Christine R Mason, Tanya L Rohtla, Phalguni S Deliwala, Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns, J. Acoust. Soc. Am. 104 (1) (1998) 422–431.

[18] Andrew D. Brown, G. Christopher Stecker, Daniel J. Tollin, The precedence effect in sound localization, J. Assoc. Res. Otolaryngol. 16 (2015) 1–28.

[19] Reiner Plomp, Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of Azimuth of a single competing sound source (speech or noise), Acta Acustica United Acustica 34 (4) (1976) 200–211.

[20] Gerald Kidd, Christine R Mason, Andrew Brughera, William M Hartmann, The role of reverberation in release from masking due to spatial separation of sources for speech identification, Acta Acustica United Acustica 91 (3) (2005) 526–536.

[21] William M. Hartmann, Localization of sound in rooms, J. Acoust. Soc. Am. 74 (5) (1983) 1380–1391.

[22] CT Justine Hui, Yusuke Hioka, Hinako Masuda, Catherine I Watson, Differences between listeners with early and late immersion age in spatial release from masking in various acoustic environments, Speech Commun. 139 (2022) 51–61.

[23] Giuseppina Emma Puglisi, Anna Warzybok, Arianna Astolfi, Birger Kollmeier, Effect of reverberation and noise type on speech intelligibility in real complex acoustic scenarios, Build. Environ. 204 (2021) 108137.

[24] http://pcfarina.eng.unipr.it/Public/Xvolver/Filter-Matrices/Aformat-2-Bformat/Zylia-Jul-2020/.

[25] EN ISO 3382-2, Acoustics - Measurement of Room Acoustic Parameters – Part 2: Reverberation Time in Ordinary Rooms, International Organization for Standardization, Genève, 2008.

[26] Pascal Dietrich, Martin Guski, Johannes Klein, Markus Müller-Trapet, Martin Pollow, Roman Scharrer, Michael Vorländer, Measurements and room acoustic analysis with the ITA-toolbox for MATLAB, in: 40th Italian (AIA) Annual Conference on Acoustics and the 39th German Annual Conference on Acoustics, DAGA, 2013, p. 50.

[27] UNI 11532-2, Caratteristiche Acustiche Interne Di Ambienti Confinati - Metodi Di Progettazione E Tecniche Di Valutazione - Parte 2: Settore Scolastico (Acoustic Characteristics of Indoor Environments - Design Methods and Evaluation Techniques - Part 2: School Sector), Ente Italiano di Normazione, 2015.

[28] BS EN IEC 60268-16:2020, Sound System Equipment - Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index, BSI Standards Publication, 2020.

[29] Brian F.G. Katz, Markus Noisternig, A comparative study of interaural time delay estimation methods, J. Acoust. Soc. Am. 135 (6) (2014) 3530–3540.

[30] L'uboš Hládek, Stephan D. Ewert, Bernhard U. Seeber, Communication conditions in virtual acoustic scenes in an underground station, in: 2021 Immersive and 3D Audio: From Architecture to Automotive, I3DA, IEEE, 2021, pp. 1–8.

[31] Christopher Hummersone, Impulse response acoustic information calculator, GitHub, 2023, URL https://github.com/IoSR-Surrey/MatlabToolbox.

[32] Pavel Zahorik, Direct-to-reverberant energy ratio sensitivity, J. Acoust. Soc. Am. 112 (5) (2002) 2110–2117.

[33] UNI 11532-1, Caratteristiche Acustiche Interne Di Ambienti Confinati - Metodi Di Progettazione E Tecniche Di Valutazione - Parte 1: Requisiti Generali (Acoustic Characteristics of Indoor Environments - Design Methods and Evaluation Techniques - Part 2: General Requirements), Ente Italiano di Normazione, 2018.

[34] Mathieu Lavandier, Sam Jelfs, John F Culling, Anthony J Watkins, Andrew P Raimond, Simon J Makin, Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources, J. Acoust. Soc. Am. 131 (1) (2012) 218–231.

[35] ITU BS.1116-3, Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems, Int. Telecommun. Union, Geneva, Switzerland, 2015, https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-RECBS.1116-3-201502-I!!PDF-E.pdf.

[36] Angela Guastamacchia, Michele Ebri, Andrea Bottega, Enrico Armelloni, Angelo Farina, Giuseppina Emma Puglisi, Fabrizio Riente, Louena Shtrepi, Marco Carlo Masoero, Arianna Astolfi, Set up and preliminary validation of a small spatial sound reproduction system for clinical purposes, in: Forum Acusticum, 2023, pp. 4991–4998, http://dx.doi.org/10.616782/fa.2023.0698.

[37] Epic Games: Unreal Engine 5, URL www.unrealengine.com.

[38] Giuseppina Emma Puglisi, Anna Warzybok, Sabine Hochmuth, Chiara Visentin, Arianna Astolfi, Nicola Prodi, Birger Kollmeier, An Italian matrix sentence test for the evaluation of speech intelligibility in noise, Int. J. Audiol. 54 (sup2) (2015) 44–50.

[39] Antonella Castellana, Alessio Carullo, Arianna Astolfi, Giuseppina Emma Puglisi, Umberto Fugiglando, Intra-speaker and inter-speaker variability in speech sound pressure level across repeated readings, J. Acoust. Soc. Am. 141 (4) (2017) 2353–2363.

[40] Jens Cubick, Torsten Dau, Validation of a virtual sound environment system for testing hearing aids, Acta Acustica United Acustica 102 (3) (2016) 547–557.

[41] Gerald A. Studebaker, A" rationalized" arcsine transform, J. Speech Lang. Hearing Res. 28 (3) (1985) 455–462.

[42] Donald Hedeker, Generalized linear mixed models, in: Encyclopedia of Statistics in Behavioral Science, Wiley Online Library, 2005.

[43] Brady T. West, Kathleen B. Welch, Andrzej T. Galecki, Linear Mixed Models: A Practical Guide Using Statistical Software, Crc Press, 2022.

[44] Jean Dickinson Gibbons, Subhabrata Chakraborti, Nonparametric Statistical Inference, Taylor and Francis, London, 2003.

[45] Daniel Pinardi, Lorenzo Ebri, Costante Belicchi, Angelo Farina, Marco Binelli, Direction Specific Analysis of Psychoacoustics Parameters Inside Car Cockpit: A Novel Tool for Nvh and Sound Quality, Technical report, SAE Technical Paper, 2020.

[46] Angelo Farina, Daniel Pinardi, Marco Binelli, Michele Ebri, Lorenzo Ebri, Virtual reality for subjective assessment of sound quality in cars, in: Audio Engineering Society Convention 144, Audio Engineering Society, 2018.

[47] Jens Blauert, Localization and the law of the first wavefront in the median plane, J. Acoust. Soc. Am. 50 (2B) (1971) 466–470.

[48] Ruth Y Litovsky, H Steven Colburn, William A Yost, Sandra J Guzman, The precedence effect, J. Acoust. Soc. Am. 106 (4) (1999) 1633–1654.

[49] NTi Audio, NTi Audio TalkBox operating manual, URL https://www.ntiaudio.com/Portals/0/data/en/TalkBox-Manual.pdf.

[50] W. Owen Brimijoin, David McShefferty, Michael A. Akeroyd, Undirected head movements of listeners with asymmetrical hearing impairment during a speech-in-noise task, Hearing Res. 283 (1–2) (2012) 162–168.

[51] Mathieu Lavandier, John F. Culling, Prediction of binaural speech intelligibility against noise in rooms, J. Acoust. Soc. Am. 127 (1) (2010) 387–399.

[52] Ilja Frissen, Johannes Scherzer, Hsin-Yun Yao, The impact of speech-irrelevant head movements on speech intelligibility in multi-talker environments, Acta Acustica United Acustica 105 (6) (2019) 1286–1290.

[53] Yi Shen, Monica L. Folkerts, Virgina M. Richards, Head movements while recognizing speech arriving from behind, J. Acoust. Soc. Am. 141 (2) (2017) EL108–EL114.

[54] William H. Sumby, Irwin Pollack, Visual contribution to speech intelligibility in noise, J. Acoust. Soc. Am. 26 (2) (1954) 212–215.

[55] Jean-Luc Schwartz, Frédéric Berthommier, Christophe Savariaux, Seeing to hear better: Evidence for early audio-visual interactions in speech identification, Cognition 93 (2) (2004) B69–B78.

[56] Richard L Freyman, Karen S Helfer, Daniel D McCall, Rachel K Clifton, The role of perceived spatial separation in the unmasking of speech, J. Acoust. Soc. Am. 106 (6) (1999) 3578–3588.

[57] Barbara G. Shinn-Cunningham, Object-based auditory and visual attention, Trends Cogn. Sci. 12 (5) (2008) 182–186.

[58] Thomas Biberger, Stephan D. Ewert, The effect of room acoustical parameters on speech reception thresholds and spatial release from masking, J. Acoust. Soc. Am. 146 (4) (2019) 2188–2200, http://dx.doi.org/10.1121/1.5126694.

[59] Ira J. Hirsh, The relation between localization and intelligibility, J. Acoust. Soc. Am. 22 (2) (1950) 196–200.

[60] Benjamin H. Zobel, Richard L. Freyman, Lisa D. Sanders, Spatial release from informational masking enhances the early cortical representation of speech sounds, Audit. Percept. Cogn. 5 (3–4) (2022) 211–237.