forum acusticum 2026
8-12 September · Graz, Austria

REAL VS. VIRTUAL SPEECH INTELLIGIBILITY IN A LECTURE ROOM

Angela Guastamacchia[1], Louena Shtrepi[1], Fabrizio Riente[2], Giuseppina Emma Puglisi[3], Arianna Astolfi[1]

[1] *Department of Energy, Politecnico di Torino, Italy*
[2] *Department of Electronics and Telecommunications, Politecnico di Torino, Italy*
[3] *University Sustainability, Research Infrastructures and Laboratories, Politecnico di Torino, Italy*

Abstract

Improving acoustic quality in learning environments is important for students of different ages, especially vulnerable groups, such as hearing aid users. When field assessments are not feasible, virtual reality (VR) provides a valuable alternative for controlled laboratory-based subjective evaluations. In this study, thirteen normal-hearing participants performed speech intelligibility tests in a university lecture room at Politecnico di Torino and repeated the same tests about one month later in a VR system based on Fifth-Order Ambisonics (5OA). The aim was to perceptually validate the VR system, previously assessed using objective measures. Five audio-visual scenarios were selected for systematic comparison, varying in target speech azimuth, distance from the listener, and the presence or absence of frontally localised masking noise. These conditions enabled the evaluation of the spatial performance of the VR system in relation to speech intelligibility. The results showed that spatial performance was preserved; although a small difference in Speech Reception Thresholds (SRTs) of 1 dB was statistically significant, it was likely not perceptible. This suggests that the VR system can realistically approximate real-world listening conditions with good perceptual accuracy.

Keywords: *speech intelligibility, classroom acoustics, virtual acoustics, lecture room, speech reception threshold*

1. Introduction

Acoustic quality is essential for comfortable and effective learning environments [1]. High noise levels and poor speech clarity in classrooms raise cognitive demands, making listening harder [2]. Consequently, maximising speech intelligibility (SI) is a primary objective in the acoustic design of the room [3]. This is especially important for vulnerable groups such as hearing-impaired individuals and hearing aid users. Recent advances in virtual reality (VR) enable new ways of studying subjective auditory perception in realistic spatial sound environments for hearing research [4]. It is also an effective way to study acoustics in existing and planned spaces. Ecologically valid virtual acoustic environments (VAEs) are crucial to advance research on auditory perception, assess the real-world impact of hearing loss, and develop effective HA fitting strategies. Based on this, researchers have also compiled shared databases of well-characterised auditory scenes to ensure reproducibility and comparability between studies [5]. VAEs validation is usually performed objectively, comparing monoaural and binaural acoustic parameters obtained both in the field and in the lab [6]. Only a limited number of studies have conducted validation by comparing speech intelligibility tests with actual listeners in both the VAE and real environments [7], [8], but none, to the Authors knowledge, have included audio and visual cues. In this study, a large lecture hall was used as a case study to compare Speech Reception Thresholds (SRTs), obtained from speech intelligibility tests, between real and virtual environments embedded in complex audio-visual scenes. The validation has not yet included vulnerable populations or hearing-aid users and, therefore, represents an initial methodological step toward the development of an inclusive validation procedure. Future validation studies will involve specific, as homogeneous as possible, subgroups of these populations.

***Corresponding author**: arianna.astolfi@polito.it.*

2. Method

2.1. Lecture room

A university lecture room at the Politecnico di Torino, with a volume of 800 m^3 room (11.3 m in width, 9.2 m in length, and 7.7 m in average height), was used as case study for the SI tests. It is located on the ground floor of a renovated building surrounded by a private pedestrian area, primarily used by university students during class intervals. It is characterised by an irregular ceiling, an acoustically reflective linoleum floor, two windows, and two doors. Inside the room, there are six rows of 15 wooden seats with tables, arranged in front of a blackboard and a large acoustically reflective desk, behind which the lecturer typically stands. In addition, the room lacks acoustic treatments, as all walls and ceiling are finished in plaster. Figure 1 shows the 3D model of the room.

2.2. Auditory scenes & SI test

As shown in Figure 1, five audio-visual scenes were selected to compare the results for different distances from the target speech, with or without spatially localised masking noise, which could be colocated or separated from the target speech. All scenarios maintained a consistent listener position within the lecture room, with the listener seated towards the back of the audience (position R2 in the floor plan shown in Figure 1), facing the main desk. Depending on the scenario, the target speech was located in front of the listener at a distance of 1.9 m (position S5), 5.6 m (position S1), or at an azimuth of 120° to the right, at a distance of 1.9 m (position S4). All target speech sources were placed at a height of 1.2 m above the floor, emulating a seated speaker orientated toward the listener. When masking noise was present, it was emitted from position S5, facing the listener at a distance of 1.9 m. In summary, the active sound sources for each scenario were as follows: target speech and masking noise from position S5 (scene 1); target speech from position S1 and masking noise from position S5 (scene 2); target speech from position S4 and masking noise from position S5 (scene 3); and target speech from position S5 (scene 4) or position S1 (scene 5) without masking noise. All scenes included diffuse background noise typical of the lecture room in unoccupied conditions with HVAC systems running (37.2 dBA).

The five AV scenes were reproduced within the VR system installed in the Audio Space Lab (ASL) at Politecnico di Torino [6], using the corresponding 5OA RIRs (Fifth-Order Ambisonics Room Impulse Responses) previously used for lab validation based on objective data [6]. The VR system uses commercially available hardware and has two main components: (i) a VAE reproduction system with a 16.2 ambisonics setup; and (ii) a VVE (Virtual Visual Environment) reproduction system using a VR headset. The 5OA RIRs were combined with a 5-minute 5OA recording of background noise at the listener's position in the real lecture room and 2-minute stereoscopic 360° videos recorded in 4K resolution. To ensure consistency between the visual conditions used in the real lecture room tests and those in the VR system, the videos were carefully recorded to replicate exactly the same visual context present during the SI tests in the real lecture room. Specifically, the videos featured one Talkbox speaker at each active sound source location (target speech and masking noise) of the represented auditory scene, identical to those used in the actual tests. No lip-reading information was used, and no extra actors or moving components that could not be exactly reproduced in real-world conditions were incorporated. Only the test operator was present, positioned at the back of the room behind the listener, ensuring the operator remained outside the listener's field of view. The adaptive ItaMatrix speech-in-noise test [9] was used to measure the SRT50 (Speech Reception Threshold at 50% correct speech recognition) values in an open response format, where listeners repeated the words they understood immediately after the completion of each target speech sentence. For each scene, a different list consisting of 20 sentences was used for the target speech. At the beginning of each scene, the speech level was set to be clearly audible and subsequently adapted to achieve 50% word recognition by the end of the sentence list, while maintaining the noise level constant. Specifically, the masking noise was represented by ItaMatrix noise, which is a speech-shaped modulated noise (SSN) that matches the spectral characteristics of the target speech, calibrated to achieve 50.7 dBA at the listener's position. Each target sentence was embedded within segments of this noise, with the noise beginning 2 s before and stopping 1 s after the sentence, applying cosine-shaped rise and fall ramps of 0.6 s and 0.3 s, respectively. In the case of scenes 4 and 5, where no localised masking noise was present, the SRT was calculated with respect to the diffuse Background Noise (BN) level measured in the lecture room, equal to 37.2 dBA.

2.3. Subjects

Thirteen native Italian speakers self-reported normal-hearing (nine males and four females), between 23 and 31 years of age (mean age: 26.9 ± 2.9 years), participated in the study and were compensated with stationery items such as notebooks, pens and water bottles. Of these subjects, 10 had previous experience with spatialised audio and 11 had experience using an HMD for VR (Head-Mounted Display for Virtual Reality). All subjects had normal or corrected-to-normal vision and could comfortably wear corrective lenses or glasses underneath the HMD during the virtual experience. Subjects received written and verbal information about the experiment before both tests were conducted in the real and virtual lecture rooms. A Post-hoc G-power analysis was performed to test the reliability of the number of subjects used for the test [10]. Assuming a Wilcoxon signed-rank test for matched pairs with Laplace parent distribution, effect size of 0.5, one tail p-value of 0.05, and sample size of 13 participants, the statistical power is equal to 0.69.

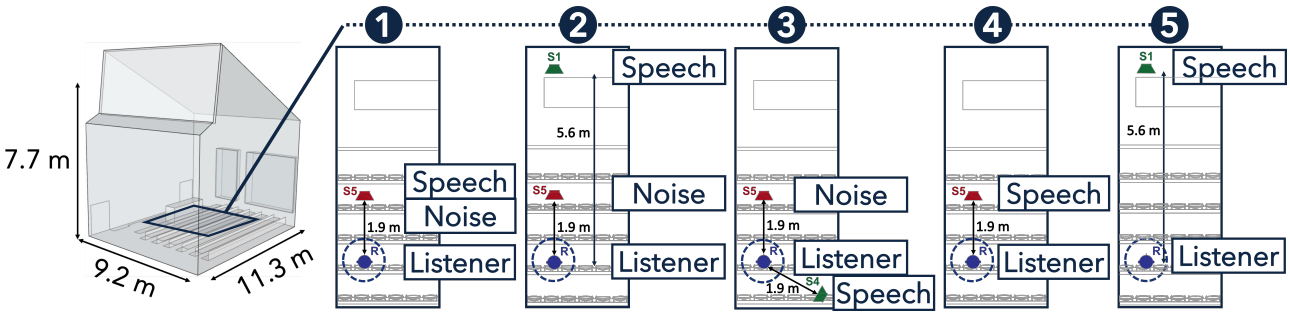

Figure 1: Audio-visual scenes in the lecture room.

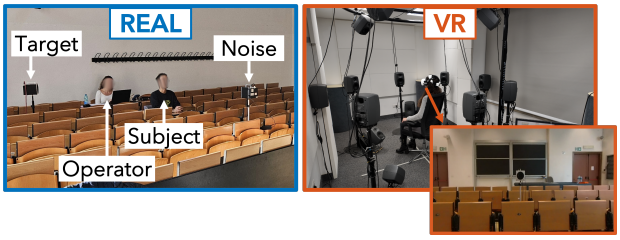

Figure 2: Real vs. VR example for scene 3.

2.4. Experimental procedure

For reasons related to the availability of the lecture room, all subjects first performed the tests in the real lecture room, followed by testing within the VR system approximately one month later. However, the ItaMatrix speech-in-noise test is specifically designed for repeated speech intelligibility assessments, as its randomly generated, semantically unpredictable sentences minimise learning and recall effects across repeated administrations. The test dates in the real lecture room were specifically selected to avoid lecture days, thereby minimising external noise caused by student activity.

During the setup of virtual tests within the VR system, each auditory scene was calibrated to match the masking noise level and the starting level of target sentences previously experienced in the real lecture room. During calibration, the recording of diffuse background noise from the real lecture room was played at the correct level within the VR system. Specifically, for each subject, diffuse background noise reproduction was started at the beginning of the test and continued uninterrupted until the test was complete within the VR system.

Each test session began with two lists of sentences (a total of 40 sentences) presented in scene 1 as training, to account for possible learning effects, as suggested in [9]. A two-minute break was provided at the end of each sentence list, with an additional five-minute break after completing the first three scenes. The test lists were randomised between subjects for each scene, as was the order of the scene. No subject heard the same test list more than once on the same day. However, to ensure consistency between the real and virtual settings, both the order of the scene and the associated test list for each scene were maintained identically between the real and virtual tests for each subject. Furthermore, subjects were instructed to keep their heads still, facing forward throughout each test scene in both real and virtual environments.

Within the VR system (see Figure 2), subjects were fitted with the HMD to stream the visual scene, seated on a swivel chair located in the centre of the speaker array. The chair was individually adjusted so that the head of each topic was precisely centred within the centre of the panel. The use of HMD aimed to replicate the visual context experienced during the tests in the real lecture room and to account for the possible impacts of Head-Related transfer function (HRTF) alterations induced by the use of HMD on the perceived SI [11]. In total, each subject participated in a one-hour testing session in the real lecture room and another one-hour session within the virtual reality system.

3. Results

Figure 3 shows the box plots of the SRT50 values for the speech intelligibility tests of the real and VR conditions in the different audio-visual scenes in the lecture room. Figure 4 analyses the spatial performance of VR system showing the box plots of the SRT50 differences between scenes 1 and 3 (Figure 4 A), between scenes 2 and 1 and between scenes 5 and 4 (Figure 4 B), comparing speech intelligibility tests within the real and VR conditions. Thus, Figure 4 A shows the preservation of the Spatial Release from Masking (SRM), which refers to the improvement in your ability to hear a target sound (such as speech) when it is spatially separated from background noise [12]. In fact, in scene 1 the target speech and the noise are co-located while in scene 3 the target speech and the noise are separated by a 120° azimuth angle. Figure 4 B illustrates the preservation of the effect of the frontal target distance on the speech intelligibility regardless of the presence of the colocated noise.

Table 1 shows the median of the SRT50 values for the real and VR speech intelligibility tests in the lecture room for the different scenes and the P-value of the Wilcoxon signed-rank test with One-tail [13] for H0 equal to X1≤X2 or X1≥X2, where X1 is the Real score and X2 is the VR score. The Cohen's d-value

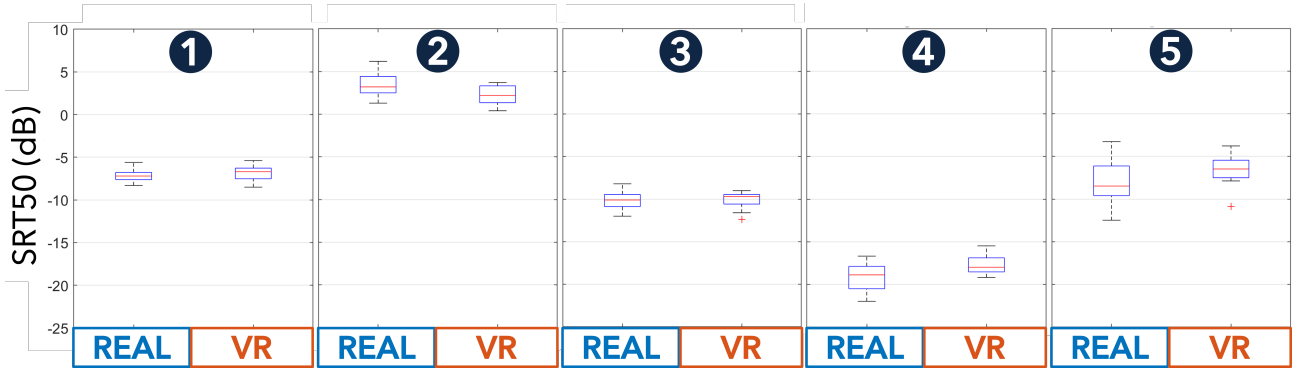

Figure 3: Box plots of the STR50 values for real and VR speech intelligibility tests of the different audio-visual scenes in the lecture room.

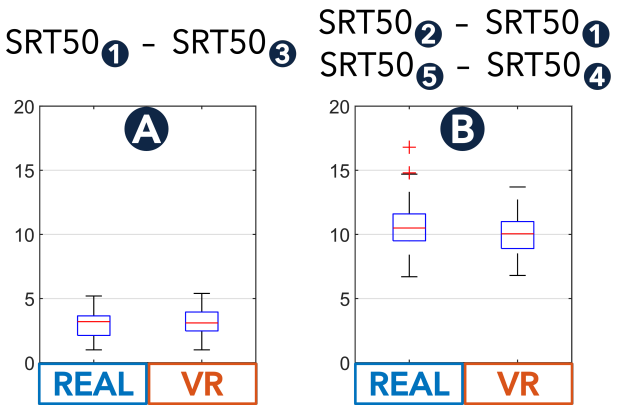

Figure 4: Box plots of the difference between SRT50 values across audio-visual scenes in the lecture room for both real and VR speech intelligibility tests. A) Difference in SRT50 values between the scenes 1 and 3. B) Difference in SRT50 values between the scenes 2 and 1 and the scenes 5 and 4, together.

for the comparison between Real and VR conditions is also shown. The P-values <0.01 and the P-values <0.05 indicate the rejection of H0, but we consider the former to be the most significant to comment on the searched results. Cohen's d-values >0.8 indicate a substantial shift between the distributions. A P-value less than 0.01 is found for the comparison of scene 2 in Real vs VR settings, with better performance in VR settings. P-values less than 0.05 are found for the comparisons of scenes 4 and 5, with better performance in Real settings. No significant differences are found concerning scene 1 and 3, the SRM in angle, and the effect of the target distance, suggesting spatial performance are correctly preserved in the VR system, while still some slight differences are detected concerning scene-specific speech intelligibility tests. The P-values of the Two-tail Wilcoxon signed-rank test for the null hypothesis H0: X1 = X2, confirm the results of the One-tail tests with values of 0.005, 0.041 and 0.097 for the comparison of scenes 2, 4, and 5, respectively, in Real vs VR settings.

Table 1: Median of the SRT50 for Real and Virtual SI tests in the lecture room for the different scenes, P-value of the One-tail Wilcoxon signed-rank test and Cohen's d-value for the comparison between Real and VR conditions. Scene (Sc), Median (Mdn) values. One-tail Wilcoxon signed-rank test for either H0 equal to X1≤X2 or X1≥X2. P-values<0.01 are reported in **bold**, p-values<0.05 in *italics*, both indicating the rejection of H0. Cohen's d-values>0.8 indicate a substantial shift between the distributions.

Sc	Mdn SRT50 (dB)		P-value		Cohen's
	X1:Real	X2:VR	X1≤X2	X1≥X2	d-value
1	-7.2	-6.7	0.927	0.078	0.279
2	3.2	2.2	**0.002**	0.998	0.932
3	-10.1	-9.7	0.612	0.405	0.028
4	-21.9	-21	0.981	*0.020*	0.913
5	-11.5	-9.5	0.954	*0.049*	0.554
A	3.2	3.1	0.757	0.255	0.172
B	10.5	10.1	0.104	0.900	0.371

4. Discussion and Conclusion

Thirteen participants initially completed speech intelligibility tests in an university lecture room at Politecnico di Torino, and then repeated the tests in the VR system employing 3OA reproduction based on 5OA RIR recordings, approximately one month later. Five audio-visual scenes were chosen for comparison, featuring varying distances and azimuth angles of the target speech from the listener, diffuse background noise typical of the room, and either including or excluding a frontally close localised masking SSN. The results show that some SRT differences between the real and virtual tests were statistically significant, but these differences could not necessarily be perceptible. In fact, McShefferty et al. [14] reported a JND of approximately 3 dB for headphone presentation of MST (Matrix Sentence Test) sentences in SSN, applicable to both normal-hearing and hearing-impaired listeners. Although this JND value does not correspond directly to the particular SiN (Speech in Noise) test conditions used in the present study, it can still provide a rough benchmark for determining which SRT differences are more likely to be perceptually meaningful. In view

of this, because the observed SRT most significant difference is 1 dB, the outcomes of the virtual tests can be regarded as perceptually close enough to those of the real tests.

References

[1] G. Minelli, G. E. Puglisi, and A. Astolfi, "Acoustical parameters for learning in classroom: A review," *Building and Environment*, vol. 208, p. 108 582, 2022. DOI: 10.1016/j.buildenv.2021.108582.

[2] C. Visentin, C. Valzolgher, M. Pellegatti, P. Potente, F. Pavani, and N. Prodi, "A comparison of simultaneously-obtained measures of listening effort: Pupil dilation, verbal response time and self-rating," *International Journal of Audiology*, vol. 61, no. 7, pp. 561–573, 2022. DOI: 10.1080/14992027.2021.1921290.

[3] 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*, vol. 204, p. 108 137, 2021. DOI: 10.1016/j.buildenv.2021.108137.

[4] B. U. Seeber, S. Kerber, and E. R. Hafter, "A system to simulate and reproduce audio–visual environments for spatial hearing research," *Hearing Research*, vol. 260, pp. 1–10, 2010. DOI: 10.1016/j.heares.2009.11.013.

[5] S. van de Par et al., "Auditory-visual scenes for hearing research," *Acta Acustica*, vol. 6, p. 55, 2022. DOI: 10.1051/aacus/2022032.

[6] A. Guastamacchia, R. G. Rosso, G. E. Puglisi, F. Riente, L. Shtrepi, and A. Astolfi, "Real and virtual lecture rooms: Validation of a virtual reality system for the perceptual assessment of room acoustical quality," *Acoustics*, vol. 6, no. 4, pp. 933–965, 2024. DOI: 10.3390/acoustics6040052.

[7] J. Cubick and T. Dau, "Validation of a virtual sound environment system for testing hearing aids," *Acta Acustica United with Acustica*, vol. 102, no. 3, pp. 547–557, 2016. DOI: 10.3813/AAA.918972.

[8] J. Schütze, C. Kirsch, B. Kollmeier, and S. D. Ewert, "Comparison of speech intelligibility in a real and virtual living room using loudspeaker and headphone presentations," *Acta Acustica United with Acustica*, vol. 9, p. 6, 2025. DOI: 10.1051/aacus/2024068.

[9] G. E. Puglisi et al., "An italian matrix sentence test for the evaluation of speech intelligibility in noise," *International Journal of Audiology*, vol. 54, no. sup2, pp. 44–50, 2015. DOI: 10.3109/14992027.2015.1061709.

[10] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behavior Research Methods*, vol. 39, no. 2, pp. 175–191, 2007. DOI: 10.3758/BF03193146.

[11] M. Gerken, V. Hohmann, and G. Grimm, "Comparison of 2d and 3d multichannel audio rendering methods for hearing research applications using technical and perceptual measures," *Acta Acustica*, vol. 8, p. 17, 2024. DOI: 10.1051/aacus/2024009.

[12] A. Ihlefeld and B. G. Shinn-Cunningham, "Spatial release from energetic and informational masking in a selective speech identification task," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4369–4379, 2008. DOI: 10.1121/1.2904826.

[13] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945. DOI: 10.2307/3001968.

[14] D. McShefferty, W. M. Whitmer, and M. A. Akeroyd, "The just-noticeable difference in speech-to-noise ratio," *Trends in Hearing*, vol. 19, pp. 1–9, 2015. DOI: 10.1177/2331216515572316.