Audiovisual recording and reproduction of ecological acoustical scenes for hearing research: a case study with high reverberation

DOI: 10.61782/fa.2023.0666

forum acusticum 2023

AUDIOVISUAL RECORDING AND REPRODUCTION OF ECOLOGICAL ACOUSTICAL SCENES FOR HEARING RESEARCH: A CASE STUDY WITH HIGH REVERBERATION

Angela Guastamacchia[1*] **Giuseppina Emma Puglisi**[1] **Andrea Albera**[2]
Louena Shtrepi[1] **Fabrizio Riente**[3] **Marco Carlo Masoero**[1]
Arianna Astolfi[1]

[1] Department of Energy, Politecnico di Torino, Italy
[2] Department of Surgical Sciences, Università degli Studi di Torino, Italy
[3] Department of Electronics and Telecommunication, Politecnico di Torino, Italy

ABSTRACT

Recent research has focused on the validation of methods and procedures to perform ecological tests for the assessment of hearing sensitivity under the complex acoustical conditions of everyday life environments. Virtual Reality (VR) has been extensively used to reproduce immersive acoustical scenes in combination with visual cues in order to account for the multisensory perception of the physical environment that happens in real-life situations. However, due to the complexity of recording and reproduction procedures, the main studies focus on either audiovisual rendering of simulated scenarios or in-field audio recordings without real visual contextualization. This work proposes a pilot case study involving a challenging listening environment (a conference hall with 3.2 s of reverberation time at mid-frequencies), where 360° audiovisual scenes were recorded and then reproduced in laboratory using a 16-loudspeakers array and a VR headset. Multiple scenarios involving different target- and noise-source positions were acquired through 3rd-order-ambisonics recordings of room impulse responses and 360° stereoscopic video footage. Speech intelligibility tests were auralized for these scenarios, considering informational masking noise at different signal-to-noise ratios, and administrated to a panel of 5 normal-hearing subjects to validate the proposed VR methodology that will be applied for future studies involving hearing-impaired listeners too.

Keywords: *ecological audiovisual scenes, hearing-impaired, speech intelligibility.*

1. INTRODUCTION

Guaranteeing a high degree of Speech Intelligibility (SI) is the primary objective that acoustic design should focus on in classrooms [1] and conference halls [2]. This is even more needed for hearing-impaired subjects who wear hearing-aids or cochlear implants, who are more challenged by the acoustical environment than normal-hearing listeners. Indeed, complex listening scenarios may occur in everyday life situations, being mainly due (i) to high reverberation [3], which also affects voice production [4], (ii) to noise [5], and (iii) to the influence of the mutual position of the target source and the receiver [6]. SI tests should be administered within these listening scenarios to account for this complexity of the real-life.

SI tests are typically performed reproducing auditory scenes in laboratory. The test presentation may be via headphones or via loudspeakers. These tests, however, suffer from the lack of completeness in replicating a complex auditory scene, as they miss the contribution of visual cues, which is proved to support SI to a significant extent. Indeed, the effect of seeing the face and mouth movements of the talkers is a well-documented effect and known to contribute to SI [7], [8] and, more in general, source-related visual

[*] ***Corresponding author***: angela.guastamacchia@polito.it.

cues indicating source position can affect localization, acceptation of the auditory illusion [9], and self-motion [10]. Visual cues can be included in SI tests performed in laboratory through ecological AudioVisual (AV) scenes, which represent challenging auditory scenarios encountered in real life. To account for the effect of lip movements, a few studies presented the auditory information of the target speech coupled with the visual counterpart represented either by an avatar [10] or a video of the real speaker [11]. More in general, as far as the influence of visual cues is concerned, some papers already focused on: (i) the influence of visual impression on the room-related and source-related sound expectation [9], (ii) the perception of room acoustics with and without visual cues and mismatching visual cues [12], (iii) the influence of matching and mismatching visual cues on localized speech comprehension [13]. However, none of these focused on the effect of room-related and source-related visual cues on SI.

Recently, some studies on ecological SI tests have been however proposed. In [14], SI has been assessed across two AV scenarios simulating an anechoic and a reverberant environment. Inside both environments, a ring of eight equally-spaced virtual loudspeakers surrounding the listening position was rendered to visually represent the location of multi-talker noise sources. Similarly, in [15], a reverberant AV scenario was proposed, with a frontal one-talker noise and a target speech signal changing among 4 positions. However, conversely to [14], in [15] only the target sound was visually represented displaying an avatar picture. Furthermore, with the aim of fostering ecological auditory research while facilitating exchange between laboratories, an open-source database of audiovisual environments was recently published [16]. In [17], SI tests carried out with virtual renderings of the visual scenes for three environments of the database are described, which involve in-field multi-channel recordings of Room Impulse Responses (RIR).

Nevertheless, only a few studies attempted to address SI measurements exploiting real recordings of both the acoustical and visual scene. In particular, in [3] one 360° video of a café scenery was shot, placing the conversational partner in the front while other customers were chatting in the background. The video was then proposed to each subject through an Head-Mounted Display (HMD), while an anechoic target signal was reproduced through a frontal loudspeaker and a generic café background noise was coming from four other loudspeakers. One step forward was taken in [11], where a one-talker video recording was blended inside a 360° video to account for the effect of the lip-movement for the target source. Anyway, the inserted masking noises had no visual counterpart. The SRT50 scores (i.e., the signal-to-noise ratio needed to yield to 50% of SI) showed an improvement of up to about 9 dB in the AV case. However, both studies did not account for the acoustical effects of the displayed environment, including anechoic speeches with unmatched background noise.

Despite these steps, the visual counterpart needs to be more deeply addressed by researchers, especially for real visual scenarios. Indeed, although video recordings are less malleable than simulations, which can be easily modified even after the creation of the scene, they usually provide improved realism and quality than simulated environments and, when the production of the scene is not overly complicated (e.g., few actors, few vehicles), shootings become more convenient than simulations, for which reaching the same level of realism would require more effort [18]. In addition, some studies have already pointed out that subjects favor real videos over simulations with virtual characters [10], [19].

This paper is a preliminary contribution to fill the lack in the available literature on the study of the influence of real contextual visual cues on SI. Although if it is planned for future studies, in this paper no lip-sync related visual cues were included due to technical issues, as this feature requires substantial effort and time to be implemented. Nevertheless, in the case of target speaker located far away from the listener position the face and mouth movements may not be visible and it may be possible not to include them in the video, as it happens in this work that can thus be considered as highly reliable. For situations where the target speaker is close to the listener, the inclusion in the 360° video will be performed blending a video recording of the speaker in the 360° video, as in [11].

To the aim of making a step forward in the available knowledge, in this work we will cover the unexplored combination of real-environment audio recordings coupled with related 360° videos recordings for visual contextualization, very high reverberation, and informational masking noise on SI. Indeed, the effect of high reverberation on SI still needs to be deepened. Among the studies relying on in-field measurements, only a few of them have considered the effect of very high reverberation in challenging auditory scenes. In particular, a mid-frequency reverberation time of 3.1 s, 1.2 s, and 2.03 s were considered in [3], [4], and [5], respectively, but none of them included visual cues and [4] and [5] involved only energetic and not informational noise, which is actually the one hindering the most speech comprehension [1].

AV scenes collected through in-field 3rd-Order Ambisonics (3OA) RIRs recordings and 360° stereoscopic video footage of a conference hall are used for ecological SI tests

administration to a small sample of five normal-hearing subjects. High reverberation time, co-located and separated informational noise and different listener-to-talker distances were considered, as well as the influence of video recordings on SI. The visual scenes include cues on the spatial location of the sound sources, which are useful for localization and acceptation of the auditory illusion, without lip-sync cues. This work represents a first step to validate an ecological protocol for hearing-aid users that will be used in clinics by ENT doctors, who have been involved in the set-up and design of experiments of the study.

2. METHODOLOGY

2.1 Audiovisual scenes collection

2.1.1 Ecological scenes selection

The environment where the AV scenes were recorded was chosen to represent a typical room with adverse acoustics where good speech comprehension is highly required, i.e., a conference hall with a high reverberation time [14]. The hall is located on the first floor of the 17th century-building hosting the Egyptian Museum of Turin. The room has a volume of 1500 m^3 and is furnished with 100 chairs, one table above a 30 cm high stage where the speaker is usually located, and one table hosting the control station of a 2-loudspeaker amplification system in the back.

Seven scenes with different listening positions, target speech- and masking noise-source locations representing typical communication situations inside the room were defined. Fig.1 shows the hall plan with noise and target source locations, loudspeakers and listening positions.

Two listening locations in the audience, 1.2 m from the floor, were selected, one closer and one farther away from the target speaker sat behind the table at 1.5 m from the floor. For each listening position interfering noise at 1.2 m from the floor was considered from two different directions, that is, when noise is co-located with the target at 180° or 0° azimuth, and when noise is spatially separated at 120° azimuth [20]. One-talker informational noise was chosen as competitive noise. To faithfully represent typical communication inside the conference hall, amplified target speech from the loudspeakers on the lateral walls was considered. Tab. 1 shows details of the configurations for the 7 selected scenes.

2.1.2 Audiovisual recording procedure

To capture the AV scenes, 4K stereoscopic (3D) omnidirectional videos and 3OA RIR signals were acquired, placing the recording systems in the listening positions and the sound source either in the target or in the noise locations, oriented towards the listener, except for $N2_{0°}$ where the source was rotated of 180°. The recording procedure was performed first by collecting all audio recordings and then shooting all 360° videos, since we did not require any audio-video synchrony. To ensure a proper match between visual sceneries and reproduced sound fields, actors were asked to remain quiet and keep the same position, even if some movements were allowed to enhance the immersivity of the scene. Inside the hall, few actors were present as this represents the worst-case scenario of occupancy, but still possible as confirmed by the museum management.

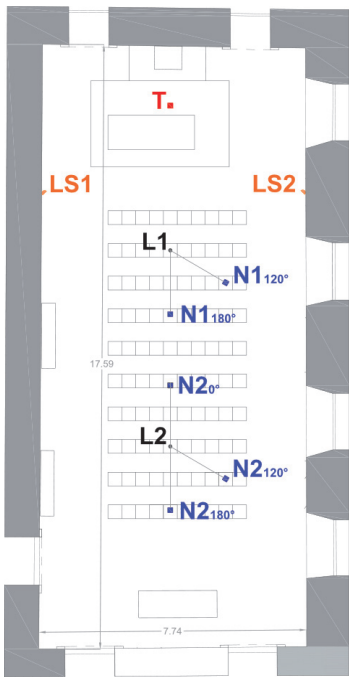

Figure 1. Conference hall floor plan with locations of loudspeakers (LS1/LS2), target (T) and noise (N1/N2) sources for all listening positions (L1/L2).

The spatial RIRs were computed starting from exponential sine sweep recordings (32-bit float at 48 kHz sample rate) acquired through the Zylia ZM-1 spherical microphone array (flat frequency response from 28 Hz to 20 kHz) and emitted by the NTi Audio Talkbox acoustic signal generator (flat frequency response from 100 Hz to 10 kHz) characterized by energy distribution featuring the same polar diagram of the human voice. The room microphone connected to the 2-loudspeaker system was switched on and

placed in front of the Talkbox at 20 cm to include the effect of the amplification in the sampled RIRs.
The 2-minute omnidirectional video recordings were taken using the Insta360 Pro 360° camera placed in the listening positions, while the Talkbox and a dummy head were placed in the target and noise positions, respectively, to provide the visual reference for the spatial arrangement of the reproduced sounds during the AV SI test. Furthermore, videos were first post-produced, removing from the visual field the tripod holding the 360° camera and then exported in the H.264 format, obtaining, in the end, .mp4 files comprising 4K 3D 360° videos at 30fps. Fig. 2 shows the equirectangular preview of the scene 2 video.

Table 1. List of all selected scenes with loudspeakers (LS1/LS2), target and noise source positions in terms of distance (m) and azimuth angles (counterclockwise notation) from the listening position (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
Target azimuth	0°	0°	0°	0°	0°	0°	0°
Target 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°
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°
LS2 distance (m)	4.2	4.2	4.2	8.3	8.3	8.3	8.3
Noise azimuth	N/A	120°	180°	N/A	120°	180°	0°
Noise distance (m)	N/A	1.8	1.8	N/A	1.8	1.8	1.8

2.2 Room acoustical characterization

The acoustics of the conference hall was characterized in unoccupied conditions through measurements of reverberation time (T_{20}), Speech Transmission Index for Public Address systems (STIPA), and background noise level, referring to the UNI 11532-2 [21] Italian standard. To assess both T_{20} and STIPA, the Talkbox was used as sound source, always placed in the target speech position, while the NTi Audio XL2 calibrated omnidirectional class-1 sound level meter was used as a receiver, placed in different positions of the audience. T_{20} was computed as the spatial average across 5 receiver positions starting from RIRs. The obtained T_{20} value averaged from 250 kHz to 2 kHz equals 3.2 s (standard deviation equal to 0.5 s), about 2 s more than the optimal value (1 s ± 0.2 s) according to [21]. Moreover, the STIPA was measured in the two listening positions (L1 and L2) with the proper signal emitted from the Talkbox (70 dBA @ 1m) and amplified by the room loudspeakers. The corresponding global A-weighted sound pressure level of the STIPA signal (called L_{Aeq} signal) was also measured in both the listening positions in order to keep the same target speech level value during the SI test. At last, also the background noise level (L_{Aeq} quiet) was measured. Tab.2 reports STIPA and L_{Aeq} values with the corresponding optimal values from [21].

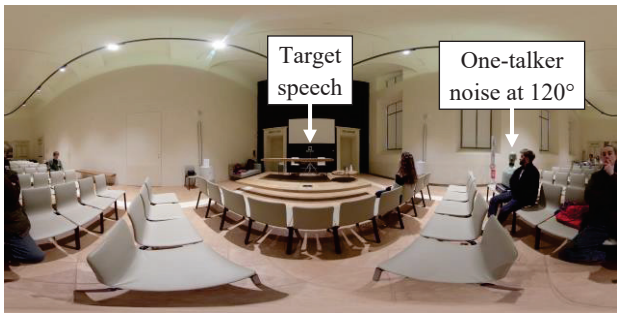

Figure 2. Example of equirectangular video preview for scene 2 (target at 4.1 m, noise at 120° azimuth).

Table 2. STIPA and L_{Aeq} signal, background noise level and critical distance measured values and optimal ranges in unoccupied conditions. Values in brackets represent the standard deviation, while N/A stands for Not Applicable. Values in bold indicate measured values compliant with the optimal ones. *Note that the STIPA optimal value refers to 80% occupied conditions.*

	STIPA (-)		L_{Aeq} quiet (dB)	L_{Aeq} signal (dB)		r_C (m)	$5*r_C$ (m)
	L1	L2		L1	L2		
Measured values	**0.62** (0.01)	0.55 (0.01)	**39.1** (N/A)	73.3 (N/A)	71.8 (N/A)	1.2	6.12
Optimal values [21]	≥ 0.60		≤ 41.0	N/A			

Both STIPA and background noise L_{Aeq} are compatible with the optimal references. Furthermore, also the critical radius (r_c), identifying the source-to-receiver distance for which the intensity of the direct and diffused field is equal and the corresponding 5-time distance (i.e., the distance beyond which SI fully depends on the reverberant field) [1], are detailed in Tab.2. It follows that target and masking noises are all placed beyond the r_c for both listening positions and that, in the farthest case, the target even exceeds the 5-time distance, pointing out the predominance of the reverberated field in about all prepared acoustical scenes.

2.3 Ecological audiovisual Speech Intelligibility test

2.3.1 Subjects

Five normal-hearing native Italian speakers (3 males and 2 females) aged 24 to 35 (average of 28 years, standard deviation of 5 years) were recruited on a voluntary basis for carrying out the ecological SI test. 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.

2.3.2 Experimental set-up

The tests were conducted in Audio Space Lab, i.e., a small sound-treated listening room of the Politecnico di Torino where the background noise level is below 38 dB for all third octave bands from 100 Hz to 10 kHz. It hosts a 3OA audio reproduction system synchronized with the Meta Quest 2 HMD to create a virtual AV 360° immersive environment. The audio playback system comprises 16 Genelec 8030B 2-way active monitors, homogeneously placed to form a spherical array surrounding the listening position at a 1.2 m distance, equalized to have a flat frequency response from 40 Hz to 20 kHz in the sweet spot. Two more Genelec 8351A 3-ways active monitors are also used as subwoofers to fill the lower frequency range. All loudspeakers are connected to the Antelope Orion[32] 32-channel sound card directly driven by a high-end desktop PC, running the Unreal engine controlling both the visual test reproduction and the audio test playback through the integration of WWise audio engine further connected to the Reaper digital audio workstation.

2.3.3 Generation of ecological SI test scenes

The audio tracks of the acoustical scenes auralizing the ecological SI tests were pre-computed using a Matlab routine starting from the RIRs acquired in the conference hall. In particular, anechoic signals containing 3-word sentences taken from the validated Simplified Italian Matrix Sentence Test (SiIMax) [22] (spoken by a female talker) were convolved with the RIRs recorded in the listening positions when the sound source was placed in the target position. The choice of the 3-word sentence test was based on a paper by the authors [22], where it was proved that the 3-word test fits good as well as the 5-word test for normal-hearing adults. Indeed, SRT50 values were close for both the tests, i.e., -7.0 ± 0.6 dB for the 3-word test [22], and -6.8 ± 0.8 dB for the 5-word test [23]. The same applied for SRT80 values, i.e., -4.5 ± 1.1 dB [22] for the 3-word test and about -4.5 dB for the 5-word test [23]. Furthermore, the 3-word test results more efficient when restrictions about the measurement procedure apply as in case of short time available [22].
To auralize the one-talker noise, the RIRs acquired when the sound source was located in the noise positions were convolved with the anechoic recording of a standardized phonetically balanced speech (spoken by a female talker) commonly used for speech recognition testing [24]. Furthermore, the auralized target signals were properly scaled to achieve in the center of the loudspeaker array (i.e., the listening position) the same signal level measured in the conference hall in the two listening positions. Then, the audio tracks to reproduce the ecological SI test in the presence of the masking noise were computed by summing auralized target and noise speeches imposing a -5 dB SNR. The clinicians selected the SNR value to propose a medium challenging acoustical condition. However, SNR around -5 dB corresponds to SRT80 in anechoic conditions [22] [23], and in our case it was plausible to reach 80% of SI scores. Each track started with 2 s of silence or interferer noise, after which the 3-word target speech was presented, and ended with two other seconds of silence or interferer noise, for an overall duration of 5 to 6 s.

2.3.4 Experimental procedure

For each subject, the SI test was split into two different test administration conditions, separating the SI test performed in the only audio condition from the one in the AV condition (i.e., with the subject wearing the HDM) with a 10-minute break in-between. Before starting the whole experiment, each participant underwent a training procedure to familiarize themself with the AV system used to reproduce the scenes and the SI test itself, i.e., to understand which of the two concurrent speeches (target and one-talker noise) he had to listen to. Additionally, as asked by the clinicians, subjects were instructed not to turn their head

during the test execution such that the same spatial configuration of target speech and masking noise with respect to the listening position was preserved as originally conceived. The SI test was conducted in the open form, with the operator taking note of the number of correctly understood words after each auralized 3-word sentence. For both test administration conditions, all 7 scenes were presented, auralizing for each scene 14 different SiIMax sentences. In the case of only audio test, the participants had to repeat the words they heard as soon as there was silence, meaning that the current sentence was finished. While in the case of the AV test condition, a pop-up was shown through the HDM displaying a sentence asking to repeat what the subject just heard. Moreover, both the scenes order and which of the two test conditions was performed first was counterbalanced across participants. On the whole, the entire test lasted for about one full hour per participant.

3. RESULTS AND DISCUSSION

Fig. 3 shows the SI percentage outcomes from the ecological tests for all 7 scenes and for both AV and only audio test conditions.

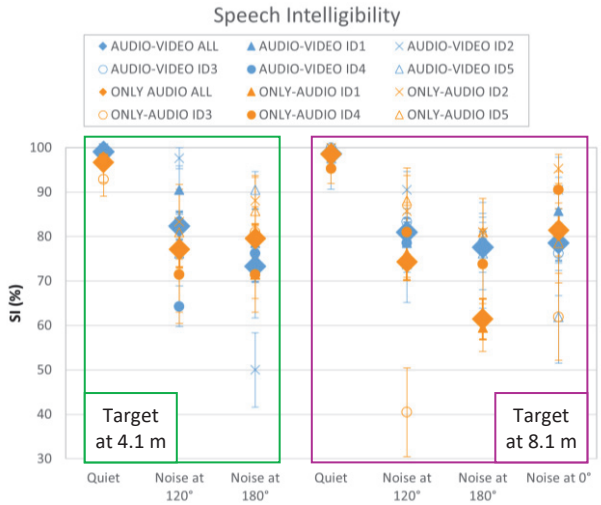

Figure 3. Average and standard error of the SI scores for all 7 scenes in case of AV and only audio test conditions for all subjects and for each single subject. *Note that the mean distance from the 2 loudspeakers is considered as target distance.*

In order to evaluate the compatibility of the results between two different testing conditions at the same hierarchical level, the normalized error measure (E_N) was applied, i.e., the ratio between the absolute value of the difference between the two average values and the difference expanded uncertainty [25]. E_N values less than 1 indicate compatible results, meaning that the SI value difference between two conditions could be only due to random effects. On the contrary, E_N values greater than 1 result from conditions that are considered incompatible, showing differences that cannot be fully attributed to random effects. E_N was computed for all 7 scenes comparing the AV condition with the only audio one. All E_N values ranged from a minimum of 0, in case of quiet conditions in the farthest listening position from the target speech to a maximum of 0.17 in case of noise at 180° always for the farthest listening position, pointing out no difference between the AV test and the only audio test. A reason for that may be found in test choices. Indeed, since the listener was instructed not to rotate the head during the test and since almost all noises were behind the listener, the visual cues that could have led to improved SI values compared with only audio conditions were notably reduced. Moreover, the E_N was also calculated to evaluate the difference between the two listening positions for 3 noise conditions (in quiet, noise at 120° and at 180°) and for both AV and only audio conditions. Also in this case, all 6 comparisons showed no significant differences, being all E_N values below 0.2 that can be reasonably since both listening positions were beyond the critical distance making the closest and farthest acoustical conditions very similar and only depending on the reverberant field.

Furthermore, apart from the compatibilities already found, below some general considerations are reported:

- The high reverberation time does not undermine the SI in quiet conditions, as, for both listening positions, the average SI percentage value goes from 97 % to 99 %.
- From the quiet condition to in-noise conditions, the SI average seems to get worse as expected, showing differences between average SI in-quiet and in-noise conditions that span from 14 % to 38 %.
- In the case of the AV condition, especially for the listening position closer to the target speech, the SI average value obtained with noise spatially co-located at 180° with the target speech decreases compared with the separated noise configuration, as expected because of the binaural listening.
- In the case of only audio condition, in the farthest listening position, the SI average value seems to get worse from separated to co-located noise configuration showing a difference of the averages equal to 13%.
- Concerning the condition of spatially co-located noise at 0°, contrary to what is expected, the SI does not get

worse. Indeed, this may be due to the orientation of the noise source with respect to the listener, which, contrary to all other noise configurations, is oriented towards the target instead of the listening position, further reducing the amount of direct component reaching the listener.

No significant differences were found between co-located and separated noise configurations, but this could be due to the use of the amplification system for the target speech, which made the target sound more diffused, so coming from a wider spatial range, either from -65° to +66° in the closest conditions or from -26° to +27° in the case of the farthest listening position, instead of the 0° direction.

4. CONCLUSIONS

To the aim of implementing ecological AudioVisual (AV) Speech Intelligibility (SI) tests based on in-field ground-truth scenes measurements for both the acoustical and visual part, this study dealt with (i) the acquisition procedure of 360° AV sceneries, showing different spatial configurations of target speech and one-talker informational masking noise inside a reverberant conference hall and (ii) a preliminary administration of the AV SI tests inside the reproduced virtual scenes, comparing the AV and the only audio condition. SI results for each scene showed no significant difference between the two experimental conditions. However, it should be noted that during the whole experimental test, subjects were instructed to keep the head still in order to maintain the same spatial orientation of the target and noise, so drastically limiting the amount of visual information provided by the 360° recorded scenes (being all noise sources behind the subjects) and impeding the natural listener's movements that are known to influence speech recognition in real-life. In support of this, all participants spontaneously expressed feeling more immersed in the scene when the visual counterpart was shown rather than in the only audio condition. Furthermore, no significant differences were found for each noise condition between the closest and farthest listening locations being both listening positions located in the reverberant field. However, the SI average values from speech-in-quiet to speech-in-noise conditions decreased (spanning from -14 % to -38 %), while the 180° noise condition showed, overall, worse SI averages compared with the 120° noise case, pointing out that the implemented test could still provide results in line with the literature fundings, so confirming the potential validity of the presented methodology. As future improvements, (i) the test should be performed on the extended version of the Italian Matrix Sentence test [23] composed of 5-word target speech sentences, as all subjects affirmed to get used to the target speech after the first trials, (ii) participants will be allowed to naturally turn their head, while a head-tracking system will be added in the experimental test to keep track of the subjects' movements, (iii) other scenes will be acquired excluding the amplification system, (iv) the test will be administered boosting the visual counterpart to further emulate real-life conversation, including lip-reading.

5. ACKNOWLEDGMENTS

The Authors acknowledge the contributions and support of the Museo Egizio di Torino during the audiovisual recordings inside the conference hall.

6. REFERENCES

[1] 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. 108137, 2021.

[2] M. Mickaitis, A. Jagniatinskis, and B. Fiks, "Case study of acoustic comfort improvement in conference room," 2021.

[3] G. E. Puglisi, A. Prato, T. Sacco, and A. Astolfi, "Influence of classroom acoustics on the reading speed: A case study on Italian second-graders," *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. EL144–EL149, 2018.

[4] A. Astolfi, A. Castellana, G. E. Puglisi, U. Fugiglando, and A. Carullo, "Speech level parameters in very low and excessive reverberation measured with a contact-sensor-based device and a headworn microphone," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2540–2551, 2019.

[5] N. Prodi and C. Visentin, "Impact of background noise fluctuation and reverberation on response time in a speech reception task," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 11, pp. 4179–4195, 2019.

[6] A. M. Kubiak, J. Rennies, S. D. Ewert, and B. Kollmeier, "Prediction of individual speech recognition performance in complex listening conditions," *The Journal of the Acoustical Society of America*, vol. 147, no. 3, pp. 1379–1391, 2020.

[7] K. W. Grant, "The effect of speechreading on masked detection thresholds for filtered speech," *The Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2272–2275, 2001.

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

[9] A. Neidhardt, C. Schneiderwind, and F. Klein, "Perceptual matching of room acoustics for auditory augmented reality in small rooms-literature review and theoretical framework," *Trends in Hearing*, vol. 26, p. 23312165221092920, 2022.

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

[11] A. H. Moore, T. Green, M. Brookes, and P. 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.

[12] M. Schutte, S. D. Ewert, and L. Wiegrebe, "The percept of reverberation is not affected by visual room impression in virtual environments," *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. EL229–EL235, 2019.

[13] A. Ahrens and K. D. Lund, "Auditory spatial analysis in reverberant multi-talker environments with congruent and incongruent audio-visual room information," *The Journal of the Acoustical Society of America*, vol. 152, no. 3, pp. 1586–1594, 2022.

[14] S. Fichna, T. Biberger, B. U. Seeber, and S. 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.

[15] L. Hladek and B. U. Seeber, "Behavior and Speech Intelligibility in a Changing Multi-talker Environment," in *Proc. 23rd International Congress on Acoustics, ICA 2019*, 2019, pp. 7640–7645.

[16] L. Hladek, S. van de Par, S. D. Ewert, and B. U. Seeber, "AUDIO-VISUAL SCENES REPOSITORY: HOW TO CONTRIBUTE," Zenodo, Sep. 2021. doi: 10.5281/zenodo.5532673.

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

[18] G. Llorach, G. Grimm, M. M. Hendrikse, and V. 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.

[19] G. Llorach, M. M. Hendrikse, G. Grimm, and V. Hohmann, "Comparison of a Head-Mounted Display and a Curved Screen in a Multi-Talker Audiovisual Listening Task," *arXiv preprint arXiv:2004.01451*, 2020.

[20] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[21] UNI/CT 002, "UNI 11532-2:2020 - UNI Ente Italiano di Normazione." https://store.uni.com/uni-11532-2-2020 (accessed Nov. 29, 2022).

[22] G. E. Puglisi *et al.*, "Evaluation of Italian Simplified Matrix Test for speech-recognition measurements in noise," *Audiology research*, vol. 11, no. 1, pp. 73–88, 2021.

[23] 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.

[24] A. Castellana *et al.*, "Intra-speaker and inter-speaker variability in speech sound pressure level across repeated readings," *The Journal of the Acoustical Society of America*, vol. 141, no. 4, pp. 2353–2363, Apr. 2017, doi: 10.1121/1.4979115.

[25] G. E. Puglisi *et al.*, "Assessment of indoor ambient noise level in school classrooms," 2015.