Impact of contextual and lip-sync-related visual cues on speech intelligibility through immersive audio-visual scene recordings in a reverberant conference room

25-29 August 2024
inter.noise
NANTES FRANCE

Impact of contextual and lip-sync-related visual cues on speech intelligibility through immersive audio-visual scene recordings in a reverberant conference room

MSc. GUASTAMACCHIA Angela[1]
Department of Energy, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129, Turin (Italy)

Mr GALLETTO Andrea[2]
Department of Control and Computer Engineering, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129, Turin (Italy)

Dr. RIENTE Fabrizio[3]
Department of Electronics and Telecommunications, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129, Turin (Italy)

Dr. SHTREPI Louena[4]
Department of Energy, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129, Turin (Italy)

Dr. PUGLISI Giuseppina[5]
Campus Management, Logistics and Sustainability, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129, Turin (Italy)

Dr. ALBERA Andrea[6]
Department of Surgical Sciences, Università degli Studi di Torino
Via Verdi 8, 10124, Turin (Italy)

Prof. PELLEREY Franco[7]
Department of Mathematical Sciences, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129, Turin (Italy)

[1]angela.guastamacchia@polito.it
[2]andrea.galletto@polito.it
[3]fabrizio.riente@polito.it
[4]louena.shtrepi@polito.it
[5]giuseppina.puglisi@polito.it
[6]a.albera@unito.it
[7]franco.pellerey@polito.it

Prof. ASTOLFI Arianna[8]
Department of Energy, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129, Turin (Italy)

ABSTRACT

Recent hearing research has benefitted from the latest Virtual Reality systems that allowed the reproduction of immersive Audio-Visual scenarios to achieve more ecological listening tests. Indeed, efforts have been spent to identify the aspects that convey actual ecological validity, particularly investigating the effects of visual cues and self-motion on Speech Intelligibility through tests mainly based on simulated scenes. However, work must still be addressed when sceneries developed through real recordings inside reverberant environments are concerned. This study used 3rd-order ambisonics recordings and stereoscopic 360° videos inside a reverberant conference hall to create three virtual audio-visual scenes where speech intelligibility tests were performed, introducing informational noise from different angles. A 16-speaker spherical array synced with a head-mounted display was used to administer the immersive tests to 50 normal-hearing subjects. Firstly, tests only composed of the auditory scenes were compared, based on the achieved scores, with tests also providing contextual and positional source-related visual cues, both with and without self-motion, for a total of four different test configurations. Then, to complete the investigation of the visual cues' impact on speech intelligibility, ten normal-hearing subjects were recruited to perform audio-visual tests incorporating lip-sync-related visual cues for the target speech.

1. INTRODUCTION

Speech Intelligibility (SI) serves as the primary acoustic objective in both small and large environments such as classrooms and conference halls. These spaces are where people is mostly engaged in speech communication. SI tests are typically conducted in laboratory settings, where auditory scenes are reproduced. The challenge lies in creating immersive virtual spaces where participants can fully engage and interact. To enhance reliability, real-life situations should ideally be drawn from audio and video recordings of communication scenes rather than relying solely on simulations.

When employing the 3D ambisonics technique for spatial audio reproduction, the listener typically occupies the center of a spherical loudspeaker array. This arrangement allows them to perceive sound naturally through their own ears. This allows an immersive perception of room acoustics to which 360° visual 3D projection can be added [1] that enhances the reliability of the scenes and optimizes participant interaction with the acoustic sound field, allowing subjects to self-move their heads during challenging tests with spatialized audio. Rather than directly facing the sound source, participants might instinctively orient themselves in a manner that improves the signal-to-noise ratio (SNR) [2]. This adjustment results in higher SNRs, contributing to a more accurate and immersive auditory experience. The combination of self-motion and strategic orientation adds depth to the spatialized audio environment, enhancing overall perceptual fidelity.

The impact of visual cues on speech intelligibility is also significant. Specifically, observing the facial expressions and mouth movements of speakers plays a crucial role [3,4]. Studies, such as the work by Neidhardt et al. [5], reveal that visual cues indicating source position affect our ability to localize sounds. Additionally, the acceptance of auditory illusions can be influenced by visual context. When visual cues align with auditory cues, our perception of illusions may change [5]. Hendrikse et al. [6] investigated how visual cues impact self-motion during auditory experiences. When we move (such as turning our head), visual cues play a role in shaping our perception of the sound environment. Nevertheless, a few studies presented the auditory information of the target speech coupled with the visual counterpart to account for the effect of lip movements. Seol

[8]arianna.astolfi@polito.it

et al.'s study [7] and Moore et al. [8] highlight how synchronized visual cues can enhance speech recognition, especially when faced with challenging auditory environments. However, none of these two studies really recreated an audio-visual scenario fully matching real life conditions. The first one did not account for the acoustical effects of the room shown in the video on the reproduced speeches and further presented a generic background noise not truly produced by the interfering talkers visible in the scene. The second study lacked the visual information related to the interfering talkers. Hladek and Seeber [9] included in their tests both contextual and positional source-related visual cues without lip-movement and self-motion in a simulated room with 1.1 s of average reverberation time at mid-frequencies using an array of 36 loudspeakers where participants stood in the center of the array, but they were allowed to move their heads to understand as much as possible from the target speaker that moved around them.

Despite these efforts, the visual counterpart needs to be more deeply addressed by researchers, especially when immersive real visual scenarios are concerned. Therefore, one of the objective of this paper is to study the influence of real contextual visual cues on SI with lip-sync and positional source-related visual cue included. Actually, we cover the unexplored combination of real-environment audio recordings coupled with related 360° videos recordings for visual contextualization. We also investigate the effect of Self-Motion (SM) compared with the Static condition (S) of the listener, the Audio-Only (AO) test provision compared with the AV (Audio-Visual) one and further the inclusion of the Lip-sync-related visual cues (L) inside the AV scene. The case study is a medium sized conference room with very high reverberation time, in which the target talker is in front of the listener, at about 4 m, and amplified by two lateral symmetrical loudspeakers and one-talker noise is around the listener alternatively at two different azimuth angles.

2. METHOD

2.1. Subjects

The tests involved 50 volunteers (13 females, 37 males) recruited from the student and staff population of the Politecnico di Torino. Age ranged from 22 to 46 years, with a mean of 27.8 years and a standard deviation of 4.8 years. All subjects had normal hearing and were native speakers of Italian. Individuals who had been prescribed eyeglasses for vision correction were allowed to wear them also in the case of head-mounted display usage to prevent possible visual problems from affecting the results.

2.2. Audio-Visual scenes

The scenes were recorded in the conference hall of the Egyptian Museum of Turin. Despite its aim, the hall is highly reverberant, being a 1500 m^3-volume room with no acoustical treatment. The room is furnished with 100 light chairs composing the stalls that faces a little wooden stage of 30 cm with on top the main wooden desk behind which the main talker is usually seated during a conference. On both sides of the hall, between the audience and the stage, two loudspeakers at a height of 1.7 m from the center of the array are present to amplify the target talker speech. Figure 1 shows the conference hall viewed from the main desk in the foreground.

Three scenes were recorded, representing communication situations typical for that hall. In particular, each scene presented the same spatial configuration for the listener and the target talker but a different setting for the competitive noise source, a.k.a. interfering one-talker. The two room loudspeakers were always used to amplify the target speech in all the scenes to get a more realistic reproduction of ordinary listening conditions inside the conference hall. Figure 2 shows the hall floor plan with the spatial locations among all the auditory scenes for the Target talker ($T_{0°}$), the Listener (L), the room Loudspeakers (LS1, LS2), and the interfering talkers ($N_{120°}$, $N_{180°}$). The listener is about 4.1 m from the target source, while it is about 4 and 4.2 m from LS1 and

Figure 1: Picture of the conference room viewed from the wooden desk in the foreground.

LS2, respectively. Furthermore, LS1 and LS2 are at about 65° and -66° azimuth angles, respectively, from the listener point of view. Finally, both the interfering noises ($N_{120°}$ and $N_{180°}$) are oriented towards the listener at a the distance of 1.8 m. Details for the listening conditions characterizing the three scenes are in Table 1.

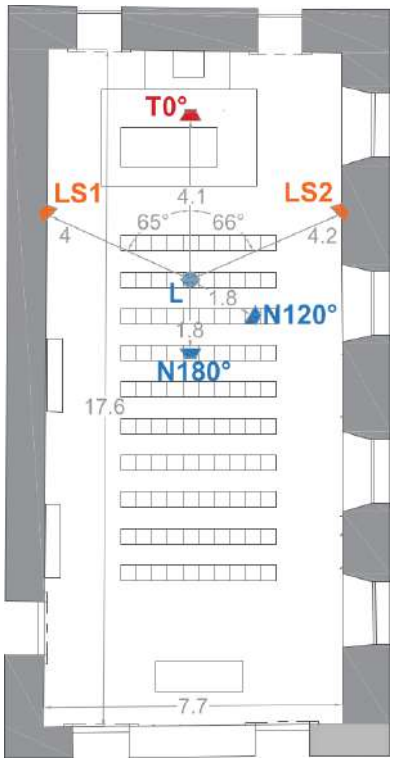

Figure 2: Floor plan of the conference hall with the positions of the listener (L), the target speech ($T_{0°}$), the two loudspeakers (LS1, LS2), and the interfering noise ($N_{120°}$, $N_{180°}$).

Table 1: Distances and azimuth angles between the listener (L), target talker ($T_{0°}$), room loudspeakers (LS1, LS2), and interfering talkers (N). N/A (Not Applicable) where the scene does not involve masking noise.

Scene number	**1**	**2**	**3**
Listener location (L)	L	L	L
$T_{0°}$ azimuth (°)	0	0	0
L-$T_{0°}$ distance (m)	4.1	4.1	4.1
LS1 azimuth (°)	65	65	65
L-LS1 distance (m)	4.0	4.0	4.0
LS2 azimuth (°)	-66	-66	-66
L-LS2 distance (m)	4.2	4.2	4.2
N azimuth (°)	N/A	-120	-180
L-N distance (m)	N/A	1.8	1.8

There was no noise in scene one, just the target talker and the listener. Scene two involved the interfering talker at 120° ($N_{120°}$) azimuth w.r.t. the listener; finally, in scene three, the interfering talker was positioned at 180° azimuth ($N_{180°}$), i.e., it was positioned behind the subject. Listeners and interfering talkers were seated at 1.2 m high from the floor, while the target talker was on a chair above the wooden stage, i.e., at a total height of 1.5 m from the floor.

2.3. AV scenes acquisition

The AV scenes providing contextual and positional source-related visual cues were recorded using the Insta360 Pro 360°-camera and the 19-capsule spherical microphone array Zylia ZM-1. Video and audio were recorded separately by locating the recording devices in the listener's position (L) once a time. The NTi Audio Talkbox was used as the sweep sound source to acquire the three 3^{rd}-order ambisonics Room Impulse Responses (RIRs). For the first RIR, the NTi was placed in the target talker position ($T_{0°}$), and the sound was amplified using the audio system available in the conference hall, i.e., LS1 and LS2, to consider its overall effect on the auditory scene. The other two RIRs were recorded with the NTi in positions $N_{120°}$ and $N_{180°}$, respectively, without using the amplification system, as they represent noise from the stalls. Concerning the visual scenes, three two-minute videos were shot placing the Talkbox in the target talker location and a dummy head in $N_{120°}$ and $N_{180°}$ to provide the positional source-related visual cues. The visual cues of lip movements for the target talker, i.e., the lip-sync-related visual cues, were filmed afterward inside a studio with the same camera and settings to maintain visual consistency with the three scenes described. An actress seated on a chair with a green screen behind was recorded while she was uttering the target talker's speech. The distance between the actress and the camera was the same as in the three scenes. Specifically, the actress performed to listen and repeat the ITAMatrix sentences test [10]; only the clips with the best-synchronized lip movements were retained. The actress was cut out from the footage, getting rid of the background, and then composited with the videos providing the contextual background of the scenes previously recorded, exploiting the use of some VFX. In the final videos, the actress is on the stage behind the wooden conference table; she substitutes the NTi, and her mouth is in the same position as the talkbox to ensure spatial coherence with the sound source origin ($T_{0°}$). Figure 3 shows the final compositing result, in equirectangular format, with the actress behind the wooden table on the stage.

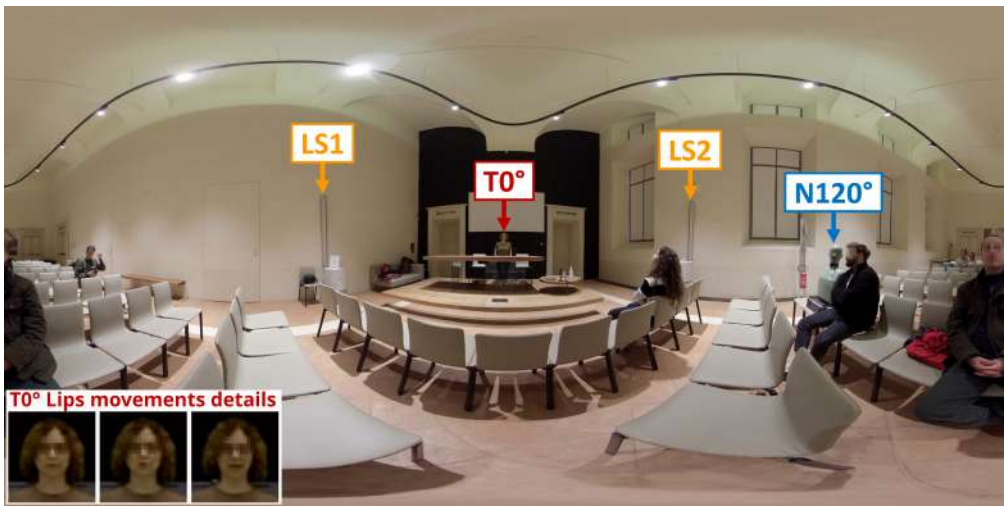

Figure 3: Equirectangular preview of the AV scene with lip movements as viewed from the listener's position. $T_{0°}$ indicates the target speaker, LS1 and LS2, the two loudspeakers, $N_{120°}$ the noise at 120° azimuth represented by the dummy-head. In the low left corner, lip movements details are shown.

2.4. Acoustical characterization of the hall

The conference hall acoustical characterization was carried out following the EN ISO 3382-2:2008 standard [11]. The relief was carried out under unoccupied conditions and the reverberation time was averaged uniformly across frequencies ranging from 250 Hz to 4 kHz octave-bands and in space. The resulting reverberation time T_{30} was equal to 3.19 s ± 0.44 s. This value is two seconds over the optimal value for good comprehension, based on Italian standards for educational environments [12]. The amplifier Lab Gruppen LAB300 was used to drive the dodecahedral omnidirectional loudspeaker Brüel&Kjær 4292-L as the sound source. SPL measurements were made using NTi Audio XL2 omnidirectional class-1 sound level meter. The A-weighted equivalent background noise level was below 40 dBA. The open-source MATLAB library ITA Toolbox [13] was used for data analysis.

2.5. Virtual reality system

Tests were conducted in the Audio Space Lab (ASL), a small sound-treated room at Politecnico di Torino, compliant with ITU-R BS.116-3 recommendations [14]. The lab features a 3^{rd}-order ambisonics system synced with the Meta Quest 2 head-mounted display for immersive 3D AV reproduction. The 16.2 ambisonics system [15] includes a spherical array of 1.2 m radius of 16 Genelec 8030B monitors and two Genelec 8351A monitors on the floor in the front used as subwoofers. All speakers are connected to the Antelope Orion32 sound card driven by a high-end workstation. Bidule DAW handles real-time audio processing on the workstation, while Unreal Engine [16] streams visual scenes to the head-mounted display. A MATLAB routine syncs the AV reproduction, exploiting the Open Sound Control protocol to communicate with Bidule DAW and Unreal Engine, and collects the test outcomes. Figure 4 shows the ASL during a test session.

2.6. AV SI test material and generation

The audio scenes for ecological SI tests were pre-computed using MATLAB scripts from the RIRs collected in the conference hall. The target speech was taken from the validated female version of the Italian Matrix Sentence Test (ITAMatrix) [10]. A standardized phonetically balanced speech [17] from another female speaker served as interfering noise. The auralized target signals were scaled to obtain at the sweet spot the same level usually reached inside the conference hall in

Figure 4: Picture of the Audio Space Lab during an AV SI test session.

the listening position L by the amplified target speech in $T_{0°}$, i.e., 73 dB(A). In-noise scenes were created by summing the target sentences with the noise clips, setting a -5 dB SNR. This SNR value corresponds to a moderately challenging acoustical condition akin to SRT80 in anechoic conditions [10]. Noise onset preceded the speech by a few seconds, as in [9, 18], to prepare the participant to listen to the target sentence. Each track began with 2 seconds of noise, or silence for in quiet scenes, followed by the ITAMatrix target sentence, and ended with 2 seconds of silence or noise, totaling 6-7 seconds.

2.7. Experimental procedure

The participants were divided into five groups of 10 people each. A test with different administration configurations was submitted to each group:

- Audio-Only test with Self-Motion (AO-SM);
- Audio-Only test in the Static condition (AO-S);
- Audio-Visual test wth Self-Motion (AV-SM);
- Audio-Visual in the Static condition (AV-S);
- Audio-Visual in the Static condition with lyp-sync-related visual cues (AV-S-L).

A training procedure was administered to make participants familiar with the test. For the S condition, participants were informed to keep their heads still, without turning, to maintain the spatial configuration of target speech and masking noise relative to the listener. In SM tests, participants were informed they could turn around freely but stayed sitting on the swivel chair. In all test configurations, participants experienced three scenes. Each scene featured 20 sentences from a distinct speech-in-noise test list. The sequence of scenes was randomized and balanced across participants. SI tests were conducted in an open format, where listeners verbally repeated understood words and the experimenter recorded correct responses. The test lasted approximately 20 minutes per participant. The experimental procedure received ethical approval (reference 100993/2023).

2.8. Statistical analysis

Speech intelligibility scores were transformed in Rationalized Arcsin Units (RAU) according to the definition in [19] in order to correct the floor and ceiling effects [20]. The non-parametric Kruskal-Wallis test and Mann-Whitney U-test were used to compare SI RAU scores under different auditory

conditions, as the assumption of normality in the score distribution was violated [21].

The combined effect of noise azimuth (NA), self-motion (SM), visual cues, i.e., audio-visual (AV) and lip movement (L) on the SI outcomes was evaluated through a particular multiple regression analysis, that is, the Linear Mixed Effects model (LME) [22], run with IBM SPSS statistics package (version 21.0, Armonk, NY). In the present study, the noise azimuth, visual cues, lip movement, and the interactions among them are the fixed effects, which are considered categorical variables, whereas the subjects are the random effects. The LME was fitted using Restricted Maximum Likelihood, and the importance of both each single fixed effect and their interactions was evaluated through the significance of a F test [23]. The standard deviations of random effect and of the residuals were estimated as well to evaluate the relevance of the subjects in the variability of the SI because of the repeated measures for the same subjects.

3. RESULTS

3.1. Speech intelligibility for the scenes

Figure 5 shows the mean and the standard deviation of the speech intelligibility (SI) scores for each test group and scene, while in Table 2 the mean and the standard deviation for the SI scores achieved in each test administration condition (AV-S-L, AO-S, AV-S, AV-SM, AO-SM) are presented after the transformation of the SI scores in RAU scores. The Kruskal-Wallis test for independent samples refuses the null hypothesis of the same distribution across the cases (p-value equal to 0.00) and, as expected, the presence of visual cues and lip movement in the static condition, that is, AV-S-L, scored the highest, followed by audio-only in the static condition, i.e., AO-S. A tie is reached between visual cues without lip movement with and without self-motion, i.e., AV-S and AV-SM, and the lowest score is for audio only with self-motion, i.e., AO-SM.

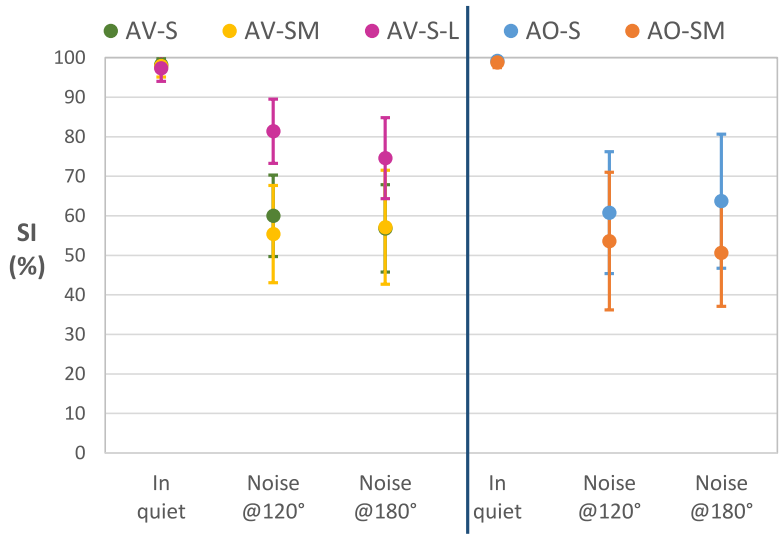

Figure 5: Means and standard deviations of the speech intelligibility scores for each listening scenario and test condition.

Table 3 shows the significance of the fixed effects parameters and their interactions. The dependent variable is the RAU score. Lip movement (L) is significantly predictive of SI as well as self-motion (SM), while noise azimuth (NA), self-motion (SM) and visual cues (AV) are only significant in combination among them. The standard deviation of the random part of the intercept, due to the subject, which represents the general variability between subjects among all the test configurations, is equal to 10.3, while the standard deviation of the residual or unexplained variation, evaluated through the 20 sentences for the same subject, is equal to 28.9. 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 2: Mean and Standard Deviation (SD) of the SI scores in RAU for each test configuration (AV-S-L, AO-S, AV-S, AV-SM, AO-SM).

Test Configuration	**N**	**Mean**	**SD**
AV-S-L	400	78.28	28.22
AO-S	400	62.50	30.12
AV-S	400	58.42	30.46
AV-SM	400	56.61	32.33
AO-SM	400	52.32	30.7

Table 3: F-values and corresponding significances resulting from a test on fixed effects and their interactions; p-values with a value less than 0.1, indicating strong evidence of the fixed effect on the RAU, are in bold. NA refers to the effect of the the noise azimuth, AV to the effect of the visual cues, SM to the effect of the self-motion and L to the effect of lip movement.

Fixed effects	**F**	**Sig.**
Intercept	1371.740	**.000**
VC	.001	.976
L	15.610	**.000**
SM	2.847	**.098**
NA	1.196	.274
AV*SM	1.382	.246
AV*NA	.129	.719
L*NA	.659	.417
SM*NA	.120	.729
AV*SM*NA	2.918	**.088**

3.2. Comparison among scenes

Table 4 shows the results from the U-Mann Whitney analyses where, for each scene, the effect of self-motion or the static condition, SM or S, the audio with visual cues or the audio-only test provision, AV or AO, and the lip movement, L, have been investigated. In particular, the comparisons between the test administration configurations for each scene are reported, i.e., AO-SM versus AO-S, AV-S versus AO-S, AV-SM versus AO-S and AV-S-L versus AO-S. The table provides p-values smaller than 0.05; overlined those that indicate the rejection of the null hypothesis H0: MX1 $\geq$ MX2 in favor of the alternative hypothesis H1: MX1 $<$ MX2, and underlined those that indicate the rejection of null hypothesis H0: MX1 $\leq$ MX2 in favor the alternative hypothesis H1: MX1 $>$ MX2, where MX1 and MX2 are the medians of the RAU distributions in the conditions X1

and X2, respectively.

Table 4: P-value of the comparisons for the Mann-Whitney test. Here, p-values lower than 0.05 are shown overlined and indicate the rejection of the the null hypothesis H0: MX1 ≥ MX2 in favor of the alternative hypothesis H1: MX1 < MX2. P-values lower than 0.05 are shown underlined and indicate the rejection of H0: MX1 ≤ MX2 in favor the alternative hypothesis H1: MX1 > MX2. MX1 and MX2 are the medians of the RAU distributions in the conditions X1 and X2, respectively.

X1	**X2**	**In quiet**	**Noise @120°**	**Noise @180°**
AO-SM	AO-S	$\overline{0.010}$	$\overline{0.000}$	
AV-S	AO-S	$\overline{0.046}$		$\overline{0.022}$
AV-SM	AO-S	$\overline{0.015}$	$\overline{0.042}$	$\overline{0.029}$
AV-SM	AV-S			
AV-S-L	AO-S		$\underline{0.000}$	$\underline{0.000}$

In general, the RAU scores among the three auditory scenes with self-motion were either lower or equal than the scores in the static condition, both for audio-only and audio-visual tests. Audio-visual tests with self-motion were not different from the Audio-visual ones in the static condition. In the case of in-noise scenes, the audio-visual tests without self-motion but with the lip movement led to higher RAU scores than the best condition without lip-movement, that is, the audio-only in the static condition, pointing out that lip-sync-related visual cues truly make the difference in real-life auditory challenging situations.

4. CONCLUSIONS

Audio-Visual (AV) scenes with and without lip-sync-related cues were collected in a medium-sized reverberant conference hall through in-field 3rd-order ambisonics impulse response recordings and 360° stereoscopic video shootings. 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. Fifty normal-hearing subjects were engaged to test the effects on SI of a talker in front of the listener at about 4 m and amplified by two lateral symmetrical loudspeakers at about the same distance, in the case of (i) one-talker noise at two azimuth angles around the listener, (ii) high reverberation with –5 dB Signal-to-Noise Ratio (SNR), (iii) self-motion, (iv) visual cues and (v) lip-movement.

Five test configurations were involved: Audio-Visual tests with Self-Motion (AV-SM) and in the Static condition (AV-S), Audio-Only tests with Self-Motion (AO-SM) and in the Static condition (AO-S) and Audio-Visual tests in the Static condition with Lip-movement (AV-S-L). For each test configuration, three scenes were proposed either in quiet or with separated (120° azimuth) or co-located at (180° azimuth) informative masking. The main results are the following:

- the AV-S-L tests scored the highest SI followed by the AO-S tests, and then by the AV-SM and AV-S in a tie, and by the AO-SM test that led to the worst SI score;
- SM scored the same as the static condition S for the AV tests;
- SM reduced SI in the AO condition.

The findings from this study represent significant progress in unraveling the intricate mechanisms underlying speech comprehension in frequently visited settings. These environments include classrooms and conference halls, where excessive reverberation poses an acoustic challenge. Beyond these practical applications, the study outcomes hold relevance for hearing

research. Specifically, they inform the design of real-life acoustic reproduction laboratories, which play a crucial role in fine-tuning hearing devices and enhancing speech intelligibility for individuals with hearing impairments.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Bernhard Seeber and Samuel Clapp. Interactive simulation and free-field auralization of acoustic space with the rtSOFE. *The Journal of the Acoustical Society of America*, 141:3974–3974, 05 2017.
2. Giso Grimm, Maartje Hendrikse, and Volker Hohmann. Review of self-motion in the context of hearing and hearing device research. *Ear & Hearing*, 41:48S–55S, 11 2020.
3. Ken Grant. The effect of speechreading on masked detection thresholds for filtered speech. *The Journal of the Acoustical Society of America*, 109:2272–5, 06 2001.
4. A MacLeod and Q Summerfield. Quantifying the contribution of vision to speech perception in noise. *Br J Audiol*, 21(2):131–141, May 1987.
5. Annika Neidhardt, Christian Schneiderwind, and Florian Klein. Perceptual matching of room acoustics for auditory augmented reality in small rooms - literature review and theoretical framework. *Trends in Hearing*, 26:23312165221092919, 05 2022.
6. Maartje Hendrikse, Gerard Llorach Tó, Giso Grimm, and Volker Hohmann. Influence of visual cues on head and eye movements during listening tasks in multi-talker audiovisual environments with animated characters. *Speech Communication*, 101:70–84, 06 2018.
7. Hye Yoon Seol, Soojin Kang, Jihyun Lim, Sung Hwa Hong, and Il Joon Moon. Feasibility of virtual reality audiological testing: Prospective study. *JMIR Serious Games*, 9(3):e26976, August 2021.
8. Alastair H. Moore, Tim Green, Mike Brookes, and Patrick A. Naylor. Measuring audio-visual speech intelligibility under dynamic listening conditions using virtual reality. In *Audio Engineering Society Conference: 2022 AES International Conference on Audio for Virtual and Augmented Reality*, Aug 2022.
9. Ľuboš Hládek and Bernhard U. Seeber. Speech intelligibility in reverberation is reduced during self-rotation. *Trends in Hearing*, 27:23312165231188619, 2023. PMID: 37475460.
10. Giuseppina Emma Puglisi, Anna Warzybok, Sabine Hochmuth, Chiara Visentin, Arianna Astolfi, Nicola Prodi, and Birger Kollmeier. An italian matrix sentence test for the evaluation of speech intelligibility in noise. *International Journal of Audiology*, 54(sup2):44–50, 2015.
11. EN ISO 3382-2. Acoustics — Measurement of room acoustic parameters — Part 2: Reverberation time in ordinary rooms. Standard 3382-2:2008, ISO: the International Organization for Standardization, 2008.
12. 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). Standard 11532-2:2015, UNI: Ente Italiano di Normazione, 2015.
13. Pascal Dietrich, Martin Guski, Johannes Klein, Markus Müller-Trapet, Martin Pollow, Roman Scharrer, and Michael Vorlaender. 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)*, page 50, 2013.

14. ITU BS.1116-3. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. Standard BS.1116-3, ITU: International Telecommunication Union, 2015.
15. Angela Guastamacchia, Michele Ebri, Andrea Bottega, Enrico Armelloni, Angelo Farina, Giuseppina Emma Puglisi, Fabrizio Riente, Louena Shtrepi, Marco Carlo Masoero, and Arianna Astolfi. Set up and preliminary validation of a small spatial sound reproduction system for clinical purposes. In *Forum Acusticum*, page 4991–4998, 2023.
16. Epic games. Unreal Engine 5. www.unrealengine.com. Last accessed 2024-04-02.
17. Antonella Castellana, Alessio Carullo, Arianna Astolfi, Giuseppina Puglisi, and Umberto Fugiglando. Intra-speaker and inter-speaker variability in speech sound pressure level across repeated readings. *The Journal of the Acoustical Society of America*, 141:2353–2363, 2017.
18. J. Cubick and Torsten Dau. Validation of a virtual sound environment system for testing hearing aids. *Acta Acustica united with Acustica*, 102:547–557, 2016.
19. Gerald A. Studebaker. A "rationalized" arcsine transform. *Journal of Speech, Language, and Hearing Research*, 28(3):455–462, 1985.
20. Raphael Cueille, Mathieu Lavandier, and Nicolas Grimault. Effects of reverberation on speech intelligibility in noise for hearing-impaired listeners. *Royal Society Open Science*, 9(8):210342, 2022.
21. J.D. Gibbons and S. Chakraborti. *Nonparametric Statistical Inference*. Taylor and Francis, 2003.
22. Donald Hedeker. *Generalized Linear Mixed Models*. John Wiley and Sons, Ltd, 2005.
23. B.T. West, K.B. Welch, and A.T. Galecki. *Linear Mixed Models: A Practical Guide Using Statistical Software*. CRC Press, 2022.