DOI: 10.61782/fa.2025.0451

FORUM ACUSTICUM
EURONOISE 2025

IMPACT OF REAL AND SYNTHETIC LIP-SYNC-RELATED VISUAL CUES ON SPEECH INTELLIGIBILITY IN A HIGHLY REVERBERANT CONFERENCE HALL

Andrea Galletto[1] **Angela Guastamacchia**[2] **Fabrizio Riente**[3]
Louena Shtrepi[2] **Giuseppina Puglisi**[2] **Andrea Albera**[4]
Franco Pellerey[5] **Arianna Astolfi**[2*]
[1] Department of Control and Computer Engineering, Politecnico di Torino, Italy
[2] Department of Energy, Politecnico di Torino, Italy
[3] Department of Electronics and Telecommunication, Politecnico di Torino, Italy
[4] Department of Surgical Sciences, Università degli Studi di Torino, Italy
[5] Department of Mathematical Sciences, Politecnico di Torino, Italy

ABSTRACT

Recent hearing research has advanced through Virtual Reality systems, exploiting immersive Audio-Visual (AV) environments based on acoustic simulations and 3D video rendering to conduct ecological listening tests and explore factors influencing Speech Intelligibility (SI). Usually, immersive AV scenes derive from simulations using avatars, but investigations are needed to detect possible differences compared to recorded real-world scenarios with real speakers. This work aims to evaluate, within immersive tests developed with 360° 3D video coupled with third-order ambisonic audio recordings, the impact on SI of (i) lip-sync-related visual cues compared with the absence of lip-sync and (ii) lip-sync-related visual cues of a real person compared with photorealistic avatars in the same scenario. SI tests on normal-hearing subjects were conducted for different auditory scenarios representing a highly reverberant conference hall with a frontal target speaker, either in a quiet situation or with an interferer talker from 120° and 180° azimuth. Results confirm the importance of lip-sync-related visual cues for speech intelligibility and using highly realistic avatars to come closer to the real speaker SI scores.

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

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

1. INTRODUCTION

Only in recent years speech intelligibility (SI) has been studied exploiting the latest Virtual Reality (VR) technologies, providing a higher ecological validity for the tests. Investigations on visual cues effects on SI, particularly those related to lip-sync, could benefit from this innovative approach as more natural test environments could be implemented, obtaining a more natural subject response during tests. Previous research has demonstrated the importance of facial and lip-sync movement on SI [1, 2]. Recently, Grimm et al. [3] found that a better realism of head movement and facial expression brings towards better speech comprehension. Investigations on SI through more ecological tests have been performed by Guastamacchia et al. [4, 5], exploiting an immersive VR environment derived by real recordings in field, both for the audio and visual part. In both studies, the audio is of 3rd-order ambisonics administered by a homogeneous spherical array of 16 loudspeakers, synced with the immersive visual stimulus provided by the Oculus Quest 2 Head Mounted Display (HMD). In the first study, the VR environment only contextualized the situation, showing the listeners the highly reverberant conference hall of

the Egyptian Museum of Turin. The target speaker was represented by a talkbox, located in the speaker's position, helping listeners to clearly identify the target origin. However, the study focused only on SI in a highly reverberant space, without taking into account any visual dynamics cues related to speaker behavior, such as lip-sync-related visual cue effects, as it completely lacked the target speaker. An improvement was made in the second study [5], in which a female target speaker was integrated into the same conference room video recordings, allowing listeners to see the real target speaker and the relative lip-sync movement. This allowed to investigate the contribution of lip-sync movement on SI. In this study, an avatar has been substituted to the real speaker to investigate the differences in affecting the speech comprehension. Avatars have been already used in literature as substituted of real speaker. As example, Grimm et al. [3] used avatars with a different level of visual realism to investigate the communication effort at change of the realism, concluding that under noise condition the rate of communication effort decreases as the level of visual realism increases, while in quiet condition the effort was rated as very easy. Furthermore, Yamada et al. [6] investigated whether the presence or absence of a visual avatar and the lip-synching affect speech comprehension in noisy VR environments, obtaining positive results on comprehension in presence of lip-sync. An avatar can be more easily created and manipulated in a VR environment compared to utilizing a video of a real person. If visual cues provided by a synthetic avatar yield the same SI as that offered by a real speaker, this could open new perspectives for deeper investigations into SI within ecologically valid settings. Avatars can be created and animated with a plethora of different software, each offering different levels of details and realism, both for the appearance and the movement. The mosts exploit neural networks to create lip-sync movement starting from speech audio signals. Wav2Lip [7], freely available [8] on GitHub [9], and its commercial counterpart [10], are both promising software for changing the talker lip-sync movement in already existing video based on new speech audio track. Both manipulate the visemes (expressive movement of the facial musculature corresponding to each phoneme of a language) exploiting a Generative Adversarial Network, but the free version offers a lower resolution for the final videos. Live Speech Portraits, based on [11] is also freely available on the web [12]. Its animation technique is capable of maintaining the talking style of the speaker, due to a preliminary training with a brief video of the real person. However, this software is able to manipulate only the head of a person and not the full body. However, these software are not the only available for this kind of applications. The state of the art in this field evolves day by day and new applications appear in the wild. Generally, these applications are offered by commercial companies as a payment service with different levels of features based on increasing fee. However, for research purposes, relying on freely available open-source tools is essential for achieving rapid and profound advancements in knowledge. Therefore, as an initial approach, this research utilizes freely available software for avatar creation and animation that can adequately fulfill the research objectives by providing satisfactory animation quality.

The aim of this work is to compare speech intelligibility in a highly reverberant environment using immersive AudioVisual (AV) scenes developed through in-field recordings. These scenes present different spatial configurations between listener, target speaker, and interfering speaker. The comparison focuses on varying visual cues associated with the target speaker, specifically: (i) a loudspeaker serving merely as a placeholder to indicate the target speaker direction, (ii) realistic lip-sync visual cues provided through video recordings of a real speaker, and (iii) synthetic lip-sync cues provided by an animated avatar.

2. METHODS

2.1 Avatar development

To develop the avatar for our purposes we chose Avaturn [13] and Audio2Face [14]. Avaturn is a website service that allows anyone to create an avatar with a good photorealistic detail with just three photos. Audio2Face is a production-ready software by NVIDIA, able to animate the facial mesh of a 3D model avatar instead of working directly on video clips of the speaker. This allows a fully digital production without the requirement of any video recording, indeed the software can be used to animate from a cartoon style face till a photorealistic one. Furthermore, the software is able to infer the emotion of the speaker from the audio tracks and consequently can manipulate the visemes in a more coherent manner. The background neural networks are trained with English language, but the software can work with audio of different languages.

2.2 AV scenes

AV scenes were recorded as described in [5]. The reverberant environment of the study is the conference hall of Egyptian Museum of Turin. It has a volume of 1500 m^3 without any acoustical treatment; it is highly reverberant, with a T_{30} reverberation time of 3.19 s $\pm$ 0.44 s, more than two seconds with respect to the optimal value in speech environments. Fig. 1 shows the conference hall in equirectangular format from the subjects' point of view. T0° spots the target speaker located in front of the subject at a distance of 4.1 m. LS1 and LS2 identify the two vertical arrays of loudspeakers of the room audio system, which amplify the target speech being positioned respectively at 65° on the left and 66° on the right azimuth from the listening point at a distance of 4.0 m and 4.2 m, respectively. Three different listening scenarios are proposed, which are: one in-quiet scene, with only the target speaker active, and two in-noise scenes, with both the target speaker and a single interfering speaker active simultaneously. For the in-noise scenes, the interfering speaker was presented at either 120° or 180° azimuth oriented toward the listener, at a distance of 1.8 m. Visual cues for the interfering speaker only included its location represented by a static dummy head, as shown in Fig. 1 by label N120° for the case of interfering speaker at 120°. The videos with the real lip-sync movement were developed as described in [5]. Concerning the scenes with synthetic lip-sync movement, as mentioned, the avatar was generated by using Avaturn and three photos of the same actress that provided the real lip-sync movement. Then, using Blender [15], the avatar was customized and placed in sat pose to fit the existing VR scenarios. Finally, some secondary slight movements for the upper body and the head were added to obtain better realism. Lip-sync movement and facial expressions were then animated by using Audio2Face exploiting audio tracks of the female validate version of the Italian Matrix Sentence Test (ITAMatrix) [16], commonly used for standard SI tests for Italian speakers. Lastly, Blender was used to render the final animation. In post-production the resulting videos were composed with the existing videos displaying the conference hall to obtain the final videos including the synthetic lip-sync-related visual cues suitable for the SI tests.
In the bottom-left corner of Fig. 1 the details of the three different visual cues investigated for the target speaker are shown.

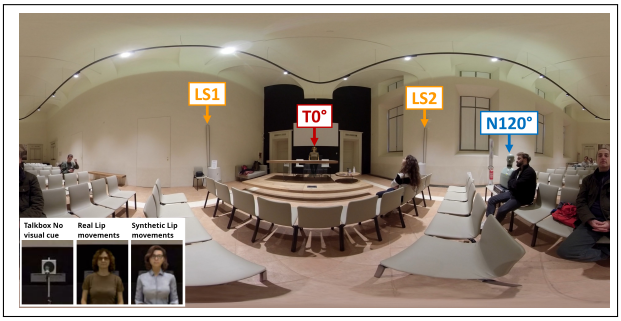

Figure 1. Image of the conference hall in equirectangular format. Bottom-left an examples of the 3 different target speakers.

2.3 Speech Intelligibility tests

To investigate lip-sync movement contribution on SI, three visual conditions for the target source representation were used: (i) a loudspeaker, i.e., no lip-sync movement; (ii) the video of a real person, i.e., subjects could see real lip-sync movement; (iii) an avatar, i.e., subjects could see synthetic lip-sync movement provided by the avatar. Each of these three conditions was administered to a different group of ten normal hearing participants.
Tests were conducted in a VR environment where the visual part was displayed by a HMD device Oculus Quest 2, while the audio part was delivered by a homogeneous spherical array of 16 loudspeakers with 3rd-order ambisonics. Subjects were sitting in the center of the array, with head in the sweet-spot. In the sweet spot, that is, in the center of the spherical array of the Audio Space Lab, the sound pressure level of the target speech was set to 73 dB(A), corresponding to the same level measured in the listening position within the real conference hall. The Signal-to-Noise Ratio (SNR) between the target and the interfering speech was set to -5 dB. The 30 participants were Italian native speakers (20 males and 10 females) aged 22 to 46 years (average 26.2 years). All of them were trained to become confident with the test system and with their task through the trial test sessions. The main participants' task was to listen to the target speech signal, which was a 5-word sentence from the ITAMatrix Sentence Test [16], and then repeat the words understood. For all subjects the same list of 20 sentences was presented for the given listening scenarios, while different scenarios were proposed each with a different test list. During the test, participants were required to remain seated without moving their heads to ensure that the auditory scenes,

specifically referring to the spatial locations of the sound sources, remained consistent for all participants. This design choice avoided introducing the confounding variable of participants' self-motion into the investigation, which, at this stage, specifically aimed at examining the impact of different visual cues without the additional complexity arising from varying participant orientations. The experimental procedure obtained ethical approval (reference 128194/2023).

3. RESULTS

The SI was evaluated in terms of percentage of correctly understood words within a set of 20 sentences. Fig. 2 shows the results of SI obtained for each test condition and listening scenario. However, to correct for the floor

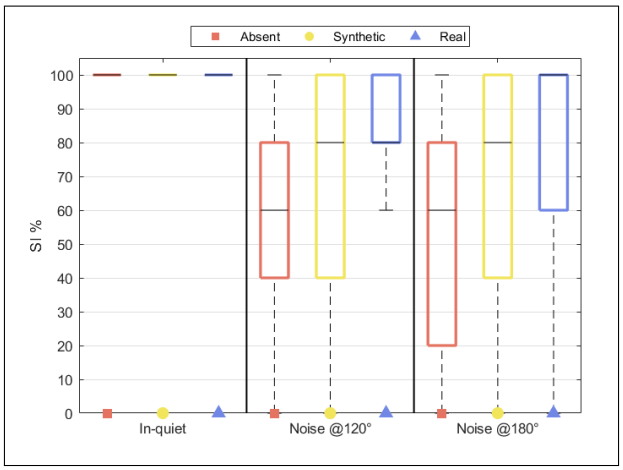

Figure 2. SI scores for each listening scenario: In-quiet, Noise @120°, Noise @180°, and test condition: no lip-sync movement (Absent), real lip-sync movement (Real), synthetic lip-sync movement (Synthetic).

and ceiling effects [17] and according to the definition in [18], the resulting SI scores were converted in Rationalized Arcsin Units (RAU) before carrying out the statistical analyses. Due to the violated assumption of normality in the difference score distributions to be compared, checked by applying the Shapiro-Wilk test, the non-parametric U-Mann-Whitney test was used for the statistical analyses [19].

The first analysis concerned the differences between the SI scores with interfering speaker at 180° and 120° azimuth. Tab. 1 shows that the null hypothesis H_0 for which the SI scores at 180° azimuth are equal to the SI scores at 120° azimuth cannot be rejected. SI scores with interfering speaker at 180° show no statistical differences with SI scores with interfering speaker at 120° as p-values are greater than 0.05 for all the test conditions. Accord-

Table 1. P-value of the comparison for the Mann-Whitney U-test for test condition and interfering speaker position.

H_0	Real Lip	Synthetic Lip	No Labial
180° = 120°	0.174	0.514	0.539

ing to this outcome the SI scores at 180° and 120° azimuth were pooled together regardless of the interfering speaker location, hence the comparisons among the conditions, i.e. the presence of real or synthetic lip-sync movement or its absence, only considered the in-noise tests, as the in-quiet tests all led to the best SI (around 100%), as can be observed in Fig. 2. SI percentage scores averaged between the in-noise listening scenarios for the three test conditions are shown in Fig. 3, along with the corresponding standard deviations. The best SI is achieved for the

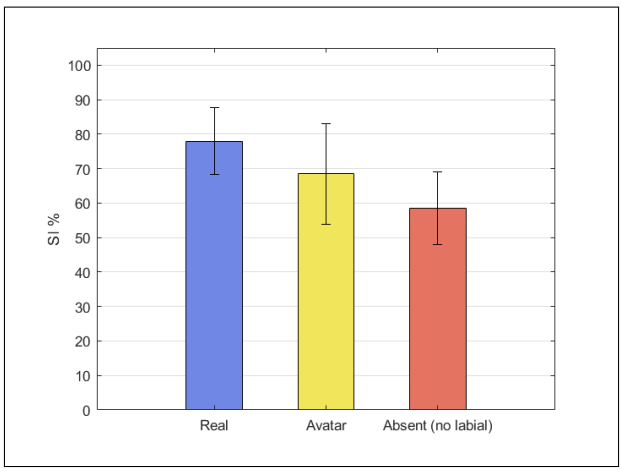

Figure 3. SI scores percentage in noise condition: no lip-sync movement (Absent), real lip-sync movement (Real), synthetic lip-sync movement (Synthetic).

tests with real lip-sync movement, followed by the syn-

thetic one produced by the avatar, while the worst condition is without lip-sync movement. In particular, average SI scores were 78%, 68.5%, and 58.5%, respectively. Tab. 2 shows the average and the standard deviation for the RAU SI scores. Tab. 3 shows the results from the U-

Table 2. Average and Standard Deviation (SD) of the SI scores in RAU for each test conditions. N represents the amount of data.

Test conditions	Mean	SD	N
Real	78.3	28.2	400
Synthetic	68.2	30.0	400
Absent	56.6	32.3	400

Mann Whitney analyses where the effect of the synthetic lip-sync movement, the real one, or its absence have been investigated reporting comparisons among the test conditions, i.e., Synthetic vs Real and Absent vs Synthetic, for the in-quiet and in-noise cases. P-values smaller than 0.05 are underlined and indicate the rejection of the null hypothesis H_0: MX1 $\geq$ MX2 in favor of the alternative hypothesis H_1: MX1 $<$ MX2. It follows that real lip-sync movement statistically significantly improve SI compared to both the absence of lip-sync movement and synthetic lip-sync movement. Nevertheless, the comparison between absent and synthetic lip-sync movement reveals that synthetic lip-sync movement still positively affect SI, as expected, though not to the same extent as real lip-sync movement.

Table 3. P-value of the comparisons for the Mann-Whitney test. Values lower than 0.05 are underlined and indicate the rejection of the the null hypothesis H_0: MX1 $\geq$ MX2 in favor of the alternative hypothesis H_1: MX1 $<$ MX2. MX1 and MX2 are the medians of the distributions (in RAU) in the conditions X1 and X2, respectively.

X1	X2	in-quiet	in-noise
Synthetic	Real	0.923	<u>0.000</u>
Absent	Synthetic	0.296	<u>0.000</u>

4. CONCLUSION AND FUTURES PERSPECTIVES

SI was evaluated administering ecological test reproducing a real highly reverberant conference hall for three different spatial configurations of listener, target speech, and interfering speaker. Tests were conducted in an immersive environment with 3rd-order ambisonics audio synced with 360° 3D videos reproduced through an HMD. Three target speaker visual conditions were evaluated: (i) without lip-sync movement, (ii) with lip-sync movement through a video of a real person, (iii) with synthetic lip-sync movement provided by an avatar. Results show that synthetic lip-sync movement contributes to SI, but with a less extent than real one, most likely due to an inaccurate facial animation. The issue could be solved by using more performative animation software, which are currently not free. In addition, to get a deeper understanding of the listening dynamics, it could be worth to investigate the influence on SI when real lip-sync-related visual cues are introduced also for the interfering speaker located within the subjects' field of view, as it could be more disturbing than when outside from the listener's view.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] A. Macleod and Q. Summerfield, "Quantifying the contribution of vision to speech perception in noise," *British Journal of Audiology*, vol. 21, pp. 131–141, Jan. 1987. Publisher: Taylor & Francis.

[2] K. W. Grant, "The effect of speechreading on masked detection thresholds for filtered speech," *The Journal of the Acoustical Society of America*, vol. 109, pp. 2272–2275, May 2001.

[3] G. Grimm, A. Kothe, and V. Hohmann, "Effect of head motion animation on immersion and conversational benefit in turn-taking conversations via telepresence in audiovisual virtual environments," in *Proc. of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, (Turin, Italy), pp. 433–435, European Acoustics Association, Jan. 2024.

[4] A. Guastamacchia, F. Riente, L. Shtrepi, G. E. Puglisi, F. Pellerey, and A. Astolfi, "Speech intelligibility in reverberation based on audio-visual scenes recordings reproduced in a 3D virtual environment," *Building and Environment*, vol. 258, p. 111554, 2024.

[5] A. Guastamacchia, A. Galletto, F. Riente, L. Shtrepi, G. E. Puglisi, A. Albera, F. Pellerey, and A. Astolfi, "Impact of contextual and lip-sync-related visual cues on speech intelligibility through immersive audio-visual scene recordings in a reverberant conference room," in *Inter-noise*, I-INCE - Société Française d'Acoustique (SFA), 2024.

[6] T. Yamada, A. Yamazaki, H. Miyakawa, Y. Mashiba, and K. Zempo, "Visual transition of avatars improving speech comprehension in noisy vr environments," in *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, VRST '21, (New York, NY, USA), Association for Computing Machinery, 2021.

[7] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proc. of the 28th ACM International Conference on Multimedia*, MM '20, (New York, NY, USA), p. 484–492, Association for Computing Machinery, 2020.

[8] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild - source code," 2020. Available at: `https://github.com/Rudrabha/Wav2Lip`. Accessed: 2024-05-31.

[9] GitHub Inc., "Github," 2025. Available at: `https://github.com/`. Accessed: 2024-05-31.

[10] synchronicity labs inc., "Synclab web site." Available at: `https://synclabs.so/`. Accessed: 2024-05-31.

[11] Y. Lu, J. Chai, and X. Cao, "Live speech portraits: real-time photorealistic talking-head animation," *ACM Trans. Graph.*, vol. 40, 12 2021.

[12] Y. Lu, J. Chai, and X. Cao, "Live speech portraits: real-time photorealistic talking-head animation," 2021. Available at: `https://yuanxunlu.github.io/projects/LiveSpeechPortraits/`. Last accessed 2024-06-01.

[13] Avaturn, "Avaturn: Realistic 3d avatar creator," 2023. Available at: `https://avaturn.me/`. Accessed: 2025-02-06.

[14] NVIDIA Corporation, "NVIDIA Audio2Face," 2023. Available at: `https://docs.omniverse.nvidia.com/audio2face/latest/index.html`. Accessed: 2025-02-04.

[15] Blender Foundation, "Blender 4.1," 2023. Available at: `https://www.blender.org/`. Accessed: 2025-02-06.

[16] G. E. Puglisi, A. Warzybok, S. Hochmuth, C. Visentin, A. Astolfi, N. Prodi, and B. Kollmeier, "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.

[17] R. Cueille, M. Lavandier, and N. Grimault, "Effects of reverberation on speech intelligibility in noise for hearing-impaired listeners," *Royal Society Open Science*, vol. 9, no. 8, p. 210342, 2022.

[18] G. A. Studebaker, "A "rationalized" arcsine transform," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 3, pp. 455–462, 1985.

[19] J. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference*. Taylor and Francis, 2003.