forum acusticum 2026
8-12 September · Graz, Austria

# DEVELOPMENT OF THE AUDIOVISUAL ITALIAN MATRIX SENTENCE TEST AND IMPLEMENTATION IN THE SPEECH-IN-NOISE AND LISTENING EFFORT ASSESSMENT (SPiNE) TOOLBOX

Chiara Visentin[1], Massimiliano Masullo[2], Matteo Pellegatti[1], Federico Cioffi[2], Nicola Prodi[1]

[1] *Department of Engineering, University of Ferrara, Italy*

[2] *Department of Architecture and Industrial Design, Università degli Studi della Campania "Luigi Vanvitelli", Italy*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.



## Abstract

This study presents the development of the audiovisual version of the Italian Matrix Sentence Test, and its implementation in a dedicated toolbox (SPiNE: SPeech in Noise and listening Effort assessment) for administering audiovisual speech-in-noise tests and collecting intelligibility and listening effort measurements. The audiovisual material was created using recordings of a real speaker and a corresponding avatar generated through audio-driven facial animation within Unreal Engine. The SPiNE toolbox enables audio-video stimulus presentation and test administration of both constant-stimuli and adaptive procedures. It also allows the collection of behavioral (i.e., response time) and subjective measures of listening effort, alongside speech intelligibility metrics. The material and toolbox support the development of ecologically valid protocols for audiovisual speech testing. Indeed, by enabling precise control over visual speech cues, virtual avatars represent a promising tool for advancing research on listening effort and for supporting future clinical applications.

**Keywords**: *audiovisual perception; matrix sentence test; listening effort; avatar; facial animation*

## 1. Introduction

Listening effort has emerged over the past decade as a key concept in hearing science, attracting increasing research interest and leading to the development of several theoretical frameworks [1,2,3]. In clinical audiology, assessing listening effort is critical for evaluating hearing aid performance and understanding the impact of hearing loss. More broadly, it provides insight into the challenges individuals face when communicating in complex acoustic environments characterized by background noise and reverberation. Traditionally, hearing assessment has focused primarily on speech intelligibility, i.e., the ability to recognize words in isolation or within sentences correctly. However, it is now widely acknowledged that high intelligibility does not necessarily correspond to low listening effort, as successful communication involves cognitive processes that extend beyond simple word recognition. Consequently, there is growing interest in incorporating listening effort measures into standard audiological test batteries, as they can complement traditional metrics and provide a more comprehensive characterization of real-world listening difficulties.

Despite this interest, several barriers hinder the routine adoption of listening effort as a clinical outcome measure. A major challenge lies in the lack of consensus on how listening effort should be defined and on the most appropriate methods for its measurement [1]. A further limitation of standard audiological assessments is their focus on the auditory modality, neglecting the multisensory nature of speech perception. In real-world communication, visual cues aid in predicting upcoming auditory input and reducing the cognitive demands on listeners, increasing their prediction precision [4]. This strongly motivates the need for more ecological tests that incorporate both auditory and visual cues.

On the other hand, while the build-up of speech intelligibility tests based on audio-video recordings can be time- and cost-consuming, novel technologies, i.e., virtual reality or artificial intelligence, can be used to generate avatars which offer precise control over visual speech cues [5]. The growing adoption of immersive audiovisual technologies has further promoted the use of virtual avatars as experimental tools in acoustic

***Corresponding author***: *chiara.visentin@unife.it*

research. These systems enable the creation of controlled yet ecologically valid environments in which complex auditory phenomena, such as speech intelligibility, spatial hearing, and multi-speaker interaction, can be investigated. Compared to traditional laboratory paradigms, which often rely on simplified or static stimuli, avatar-based approaches allow for the simulation of dynamic and socially relevant listening scenarios. A key advantage of virtual avatars lies in their ability to provide precise control over both auditory and visual variables. Avatar-based frameworks have demonstrated the potential to assess listening effort and speech understanding in ecologically realistic yet controlled conditions [6]. Recent work highlights the importance of virtual avatars and visual context in enhancing the ecological validity of acoustic experiments. Audiovisual cues, such as lip movements and facial expressions, were found to enhance speech intelligibility in noisy conditions and improve performance significantly [7,8]. A study conducted in reverberant environments showed that the inclusion of contextual visual cues and lip-synchronized facial information of the avatar can significantly influence speech intelligibility [9].

These findings reinforce the role of virtual avatars not only as visual complements to auditory stimuli but as essential components for reproducing realistic communication conditions, enabling more accurate investigation of human listening behavior in environments characterized by reverberation and competing sound sources. On the other hand, differences in reproduction systems, levels of immersion, and user behavior may still influence perceptual outcomes. Therefore, validation against real-world measurements and the development of standardized experimental protocols remain necessary.

To begin addressing these gaps, the present study was conceived with both methodological and applied research objectives. From a methodological perspective, the first goal was the development of the audiovisual version of the Italian Matrix Sentence Test (ITAMatrix), a widely used tool in both clinical and research settings. This material will be made available at the end of the project as standardized material for use across studies, following similar developments in other languages [10]. The second objective was the development of a toolbox for the administration of audiovisual speech intelligibility tests, designed to be easily accessible in clinical contexts and to enable the collection of listening effort measures alongside traditional intelligibility metrics [11].

Following the development of these tools, their feasibility and applicability will be evaluated through two experiments aimed at investigating the differences between a real speaker and an avatar. This comparison is motivated by the potential future use of avatars as a valid alternative, or substitute, for recorded video, allowing audiovisual tests to be administered without the need for new audio–video recordings while maintaining a visual speech component. While the methodological developments have been completed, the experimental phase will be carried out in a subsequent stage of the project.

## 2. Materials and Methods

### 2.1 Recordings of the audio-video materials

The audiovisual material was recorded in the anechoic chamber of the University of Ferrara, Italy. The recorded material consisted of a predetermined subset of the Italian Matrix Sentence Test [12], comprising 140 sentences randomly selected from the pool of over 100000 possible sentences. The selected sentences were further checked to ensure a balanced occurrence of all 50 words of the base matrix.

The sentences were recorded by an adult, native Italian, female speaker, with a trained voice and expertise in stage reading. She was instructed to maintain a neutral facial expression, pronounce the sentences clearly, speak at a conversational rate (four to five syllables per second), and maintain a consistent vocal output throughout the recording. During the recording session, she viewed the sentences to be recorded via an off-camera screen. The session lasted around 3 hours.

The speaker was seated on a chair positioned in front of a green screen to isolate her from the surrounding environment. To minimize shadows and ensure uniform illumination of the speaker's face, the scene was lit using the anechoic chamber lighting supplemented by three LED spotlights (two positioned at the front and one at the rear). The videos were recorded with a camera (Panasonic Lumix S5M2X) equipped with a 24-105 optic lens, placed approximately 3 m away, at a height of 85 cm from the ground.

The speech was recorded using a shotgun microphone (Sennheiser ME66) positioned above the camera, alongside the camera's internal audio acquisition system. The recording setup is shown in Figure 1.

**Figure 1**. Audiovisual recording set-up.

### 2.2 Avatar development

A virtual avatar was developed at the Sens i-Lab of the Università degli Studi della Campania, using the MetaHuman Creator [13]. The platform allows users

deep character customization and the possibility to import them directly into an existing Unreal Engine project. In parallel, a minimal grey room was created in Autodesk 3ds Max to serve as the virtual background.
The recorded videos with the real speaker were used as stimuli to drive the Audio-Driven Animation [7] for MetaHuman plugin. This approach enabled the processing of each audio track to automatically generate facial motion within Unreal Engine, producing synchronized, voice-driven animations mapped onto the MetaHuman character's facial rig. Additionally, the body rig of the avatar was manipulated to match the body pose of the human counterpart. Each animation was exported as a video stimulus ready for playback in the toolbox (Figure 2). The video materials featuring the real speaker were used as a reference to match the aspect ratio in the generated stimuli (Figure 3).

**Figure 2**. The avatar speaker image of the experiment.

**Figure 3**. The real speaker image of the experiment.

### 2.3 SPiNe Toolbox

The SPiNE (SPeech in Noise & Listening Effort Assessment) toolbox was developed in-house at the Department of Engineering of the University of Ferrara. It is based on a LabVIEW (version 21.0.1f6) script and enables both the presentation of experimental stimuli and the collection of response data.
The toolbox communicates via Open Sound Control (OSC) commands with an audio-rendering engine implemented in Reaper, which manages audio–video playback. Within Reaper, the X-MCFX plug-in is used to convolve the audio signal with the equalization filter of the headphones or the impulse response of an environment (either measured or simulated).
The toolbox allows the administration of tests using either a constant-stimulus paradigm (fixed signal-to-noise ratio, SNR) or an adaptive procedure. In the constant-stimulus paradigm, the SNR is directly set within the audio management software, and the outcome measure of the toolbox is the intelligibility score obtained in each trial. In the adaptive procedure, once the target intelligibility level is defined, the speech reception threshold SRT is calculated using the staircase procedure described in study [14]. The overall audio level is adjusted by modifying only the level of the target speech signal, starting from a predefined SNR value. The outcome measures in this case are the intelligibility score and the SNR value obtained for each trial.
The toolbox additionally allows collecting listening effort measures, including both behavioral and subjective metrics. Behavioral listening effort is quantified through response time. When the intelligibility test is administered in a closed-set format (i.e., participants are presented with the base matrix on a tablet and are required to select the perceived words directly from it; Figure 4), the response time is defined as the time interval between the end of the audio–video playback and the selection of the word in the first column [15].
Subjective listening effort is collected using a visual analogue scale (Figure 5). Participants answer the question "How demanding was it to understand the words?" ("Quanto è stato impegnativo capire le parole?") by moving a slider along a continuous horizontal scale whose numerical output range from 1 to 10. The left endpoint is labelled "Very little" ("Pochissimo") and the right endpoint "Very much" ("Moltissimo"). Higher scores, therefore, indicate greater perceived listening effort. In addition, participants rate their confidence in the responses provided using the same 10-point visual analogue format. Depending on the experimental protocol, subjective ratings can be collected either at the end of each list of 20 sentences, corresponding to an experimental block with a fixed listening condition, or on a trial-by-trial basis after each sentence.

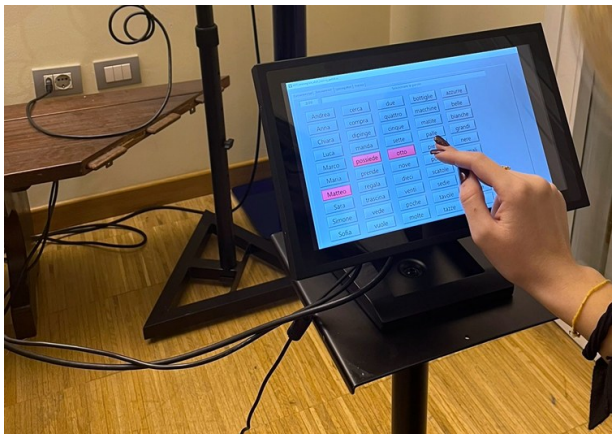

**Figure 4.** Testing with the SPiNE Toolbox.

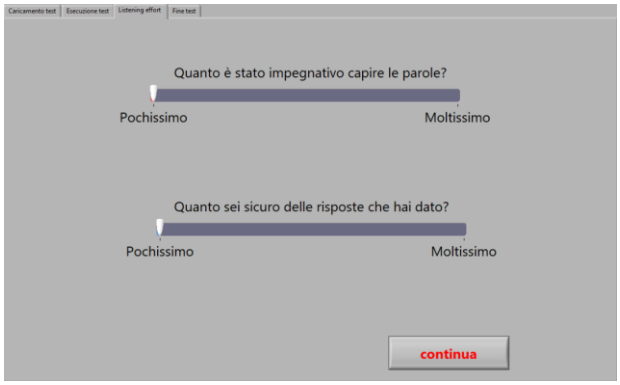


**Figure 5.** Example of the subjective rating scales presented to participants at the end of each test block. The upper scale assessed perceived listening effort with the question "How demanding was it to understand the words?" The lower scale assessed response confidence with the question "How confident are you about the responses you gave?"

## 3. Development of an Italian Audiovisual Matrix Sentence Test

Following the recording phase, the footage was edited to retain only the speaker's upper torso, and the background was replaced with a more neutral gray tone (Figure 3). Adobe Premiere software was used to edit the audio and video signals. For each recording, 5-second video clips were extracted, structured with a period of silence followed by the utterance of the sentence. The audio signals were further processed in Matlab and set to the same root-mean-square level. Finally, all clips were pseudo-randomly recombined to yield 10 equivalent lists, each comprising 20 sentences. The lists were then exported in .mp4 format, with a 48 kHz sampling rate and 24-bit resolution. The stereo audio track of each .mp4 file was further convolved with a frontal head-related transfer function (HRTF), measured in an anechoic chamber using a B&K Type 4100 Head and Torso Simulator, so that the perceived position of the talker was localized in front of the listener.

Additionally, two background noises were developed. The first consisted of long-term speech-shaped noise, derived from a steady-state white noise signal that was spectrally shaped in octave bands to match the speaker's long-term spectrum. This type of noise primarily induces energetic masking of the speech signal. The second type of background noise was designed to reduce intelligibility through informational masking. It was created using anechoic recordings of four narrative stories read by the same speaker featured in the videos, combined into a single stereo file. Specifically, each story was convolved with a different white-noise burst to randomize the phase structure of each speech signal and to make the perceived location of the individual talker less identifiable. The processed stories were then normalized to the same root mean square (RMS) level. Two stories were assigned to the left channel, and two to the right channel, and the signals assigned to each channel were summed to obtain the final stereo masker.

## 4. Planned experiments

Following the development of all materials, two experiments were designed with the aim of comparing an avatar and a real speaker under a variety of listening conditions, as well as establishing a standardized experimental protocol. The experiments will be conducted in parallel at the two collaborating research centers. Before the experimental phase, an alignment between the two centers will be carried out to ensure consistency in the hardware and software setup.

In both experiments, the playback level of the speech signal will be fixed at 60 dB(A), corresponding to a normal vocal effort [16]. The auditory stimuli will be presented through headphones that will be equalized and calibrated using an artificial head.

The first experiment will investigate whether, at a fixed intelligibility level, speech perception with an avatar produces listening effort comparable to that observed with a real speaker, and whether any difference is modulated by participants' lip-reading abilities or by the type of background noise. The second experiment will examine whether speech perception differs across real speaker, avatar, and audio-only conditions at different levels of acoustic degradation.

## 5. Concluding remarks

This study presents the development of an audiovisual Italian Matrix Sentence Test and a dedicated toolbox for the assessment of speech intelligibility and listening effort in ecological conditions. The integration of virtual avatars enables precise manipulation of visual speech cues while reducing the need for time- and cost-intensive audio–video recordings. The experimental framework, currently under implementation, is designed to systematically compare real and avatar-based intelligibility tests under different listening conditions. The results will provide insights into the extent to which avatars can reliably reproduce audiovisual speech perception and listening effort observed with real speakers.

The study contributes to the advancement of standardized methodologies for audiovisual testing and highlights the potential of immersive and AI-based technologies in acoustic research and clinical audiology.

## 6. Acknowledgments

The authors gratefully acknowledge Andrea Trevisani (e-Learning and Multimedia Services Office, University of Ferrara, Italy) for his technical support in the recording and editing phases. They also thank Riccardo Segala for implementing the OSC communication protocol for the adaptive procedure, as part of his Bachelor's Thesis.

## References


[1] M. K. Pichora-Fuller, S. E. Kramer, M. A. Eckert, B. Edwards, B. W. Hornsby, L. E. Humes, M. Lemke, A. Lunner, and M. Matthen: "Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL)," *Ear and Hearing*, vol. 37, pp. 5S–27S, 2016.

[2] J. Rönnberg, J. Rudner, T. Lunner, and A. Zekveld: "When cognition kicks in: Working memory and speech understanding in noise," *Noise & Health,* vol. 12, pp. 263–269, 2010.

[3] S. L. Mattys, R. M. O'Leary, R. A. McGarrigle, and A. Wingfield: "Reconceptualizing cognitive listening," *Trends in Cognitive Sciences*, 2025.

[4] J.E. Peelle and M.S. Sommers: "Prediction and constraint in audiovisual speech perception," *Cortex*, vol. 68, pp. 169-181, 2015.

[5] A. Hussain, A.M. Goman, M. Gogate, K. Dashtipour, J. Kirton-Wingate, Z. Hussain, et al.: "Audio-visual speech-in-noise tests for evaluating speech reception thresholds: A scoping review," *PLoS One*, vol. 21, no. 1, pp. 1-19, 2026.

[6] A. Devesse, A. van Wieringen, and J. Wouters: "AVATAR assesses speech understanding and multitask costs in ecologically relevant listening situations," *Ear and Hearing*, vol. 41, no. 3, pp. 521-531, 2020.

[7] F. Cioffi, M. Masullo, A. Pascale, and L. Maffei: "Speech Intelligibility in Virtual Avatars: Comparison Between Audio and Audio–Visual-Driven Facial Animation," *Acoustics*, vol. 7, no. 2, 30, 2025.

[8] J. K. Cooper, J. Vanthornhout, A. van Wieringen, and T. Francart: "Objectively Measuring Audiovisual Effects in Noise Using Virtual Human Speakers," *Trends in Hearing*, vol. 29, 23312165251333528, 2025.

[9] A. Guastamacchia, A. Galletto, F. Riente, L. Shtrepi, G. Puglisi, A. Albera, 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 *Proc. of INTER-NOISE and NOISE-CON Congress*, (Nantes, France), pp. 9174-9185, 2024.

[10] G. Llorach, F. Kirschner, G. Grimm, M. A. Zokoll, K. C. Wagener, and V. Hohmann: "Development and evaluation of video recordings for the OLSA matrix sentence test," *International Journal of Audiology*, vol. 61, pp. 1–10, 2022.

[11] 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

[12] 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(sup2), pp.44-50, 2015.

[13] International Epic Games: Epic Games metahuman creator. https://metahuman.unrealengine.com (2021).

[14] T. Brand, and B. Kollmeier: "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," *The Journal of the Acoustical Society of America*, vol. 111, pp. 2801-2810, 2002.

[15] C. Visentin, and N. Prodi: "A matrixed speech-in-noise test to discriminate favorable listening conditions by means of intelligibility and response time results," *Journal of Speech, Language, and Hearing Research*, vol. 61, pp. 1497-1516, 2018

[16] ISO 9921:2003. Ergonomics — Assessment of speech communication. International Organization for Standardization, Geneva, Switzerland.