DOI: 10.61782/fa.2025.0855

FORUM ACUSTICUM EURONOISE 2025

THE AVATARS SPEECH INTELLIGIBILITY: COMPARISON BETWEEN TWO DIFFERENT SPEECH GENERATION METHODOLOGIES FOR FACIAL ANIMATION

Federico Cioffi[1*] **Massimiliano Masullo**[1] **Aniello Pascale**[2]
Luigi Maffei[1]
[1] Department of Architecture and Industrial Design, Università degli Studi della Campania "Luigi Vanvitelli", Aversa (CE), Italy
[2] Immensive s.r.l.s., Parete (CE), Italy.

ABSTRACT

Research on speech intelligibility has shown that visual cues, such as facial movements synchronized with acoustic cues, significantly affect listeners' efforts during communication tasks. The mismatch in these elements can adversely affect speech intelligibility outcomes in terms of cognitive load and correct comprehension. This task is even more critical in noisy environments where listeners must discern speech against challenging background noise. In even more interactive virtual environments, communication with the avatars becomes increasingly prevalent, requiring a comprehensive understanding of their dynamics to ensure effective interactions between the avatars involved. Utilizing Unreal Engine's MetaHuman technology, the present study compares two different speech generation methodologies (synthesised text-to-speech vs human voice recording) for testing automatic facial animation generations through a laboratory experiment that investigated how these can affect avatars' speech intelligibility under adverse acoustic conditions. Thirty-six words from the Diagnostic Rhyme Test (DRT) were recorded by a human voice and generated through text-to-speech software to drive the animations. Participants were presented with 72 animations with an adversarial babble noise with a fixed signal-to-noise ratio of -13 dB. The study showed that animations driven by the human voice, in comparison with the synthesized one, significantly improved the avatars' speech intelligibility.

Keywords: virtual reality, facial animation, metahuman, speech intelligibility, unreal engine.

1. INTRODUCTION

Communication is a key component of human interaction. When communicating, individuals do not rely solely on auditory signals but also interpret visual cues, such as lip movements, to enhance their understanding and distinction of speech [1]. Due to the brain's limited cognitive capacity, challenging listening environments demand increased mental effort for speech processing, which in turn reduces the resources available for tasks like memory retention and critical thinking [2]. This challenge is particularly significant in educational settings, where excessive listening strain can contribute to cognitive fatigue, diminishing focus and academic performance. Speech intelligibility (SI) is essential for ensuring that verbal communication remains clear, effective, and easily understood. However, adverse acoustic conditions frequently hinder SI [3]. Since learning requires considerable cognitive effort, poor SI can disrupt information retention, engagement, and overall educational success, emphasizing the importance of developing

[*]***Corresponding author***: *federico.cioffi@unicampania.it*
Copyright: *©2025 First author et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*

innovative evaluation techniques and enhancement strategies. Recent developments in human-computer interaction and audiovisual speech synthesis have facilitated the creation of virtual avatars capable of mimicking human speech. For instance, virtual tutors have been integrated into educational environments to aid in teaching science to children [4] and to enhance language learning experiences [5]. Advances in virtual reality (VR) and artificial intelligence offer new possibilities for refining assessment methods through the use of digital human avatars. With these avatars is possible to create controlled experimental settings for studying speech perception, featuring precise lip-sync animations and realistic facial movements that closely replicate natural speech patterns [6].
Leveraging Unreal Engine's MetaHuman technology [7], this study examines two different speech generation approaches (synthetic text-to-speech voice versus human voice recordings) to generate facial animation and explore how these methods influence avatars' speech intelligibility in acoustically challenging conditions.

2. MATERIALS

2.1 The virtual scene and avatar creation

A virtual avatar was created using the MetaHuman Creator platform [7]. Once customized, the avatar can be directly exported into an existing Unreal Engine project and is ready to be animated (see Fig 1). Furthermore, a simple, empty white room was modelled in the 3D Studio Max software to serve as a virtual scenario for the experiment.

Figure 1. The avatar used in the experiment.

2.2 Stimuli acquisition and facial animation generation

To create the stimuli to use in the test, the Italian version of the Diagnostic Rhyme Test (DRT) [8] was taken into consideration as a well-known method used to assess speech clarity. It consists of disyllabic word pairs, distinguished by Trait (Nasalità, Continuità, Stridulità, Coronalità, Anteriorità, Sonorità), given in rhyme, in which the initial consonant is changed. Only the first word of each pair is representative of the specific Trait (see Table 1).
From the DRT words list, 18 pairs were chosen (one per Trait) to serve as auditory stimuli for driving the animations. The word pairs were recorded by both a human voice and generated by the software ElevenLabs [9] for the synthesized version. For the human voice, the word pairs were recorded in a controlled environment using a Rode NTG2 microphone and a ZOOM H5 recorder (see Fig 2). For each pair, two recordings were made separately: one for the first word of the pair and one for the second, for a total of 36 words. All the stimuli words were preceded by a carrier phrase: "Adesso diremo la parola…" ("Now we'll say the word…").

Table 1. Distinctive Traits in the DRT (in bold).

Trait	Description	Example
Nasalità	Whether the air flows through the nasal cavity.	**Nido** / Lido
Continuità	Whether the air flows through the oral cavity in a prolonged way over time.	**Riso** / Liso
Stridulità	Airflow passes through a small slit between two very close surfaces.	**Cina** / China
Anteriorità	Whether the alveolar region is obstructed.	**Nesso** / Messo
Coronalità	The coronal part of the tongue is raised compared to its resting position.	**Sisma** / Scisma
Sonorità	Whether vocal cords are close together and therefore vibrate due to the airflow during sound production.	**Vino** / Fino

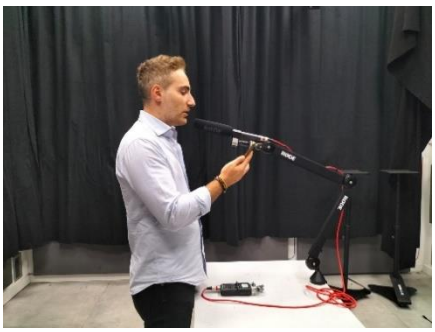

Figure 2. Human voice stimuli recording.

For the recording of the synthesized voice, a 1-minute-long recording from the same person was fed into the software ElevenLabs to clone their voice with AI technology. Afterwards, the written stimuli (carrier phrase plus word) were uploaded into Elevenlabs which output the synthesized voice audio stimuli.
The recorded materials were used to create facial animations using the Audio-Driven Animation for MetaHuman plugin. This feature allows for the processing of audio files into facial animations directly into Unreal Engine. The Audio-

Driven Animation plugin was used to generate the voice-driven animations and save them as video stimuli ready to be played. Finally, a babble noise was mixed in to serve as an adversarial sound. All the sounds were calibrated using an HSU III.2 artificial head. The stimuli were calibrated at 60 dB(A) to simulate normal speech at 1 meter distance [10], whereas the babble noise was calibrated at 73 dB(A), resulting in a signal-to-noise ratio of approximately -13 dB(A). A total of 72 stimuli (36 words per 2 conditions, human versus synthetic) were prepared for the test and loaded into the Psychopy software [11] for the stimuli playback, randomization, and scoring collection (see Figure 3).

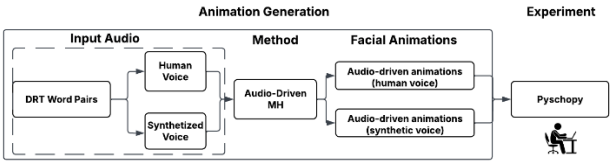

Figure 3. Workflow for generating the facial animations.

3. METHODOLOGY

Thirty-five participants (15 females, Mage= 29,1; SD= 6,6) were recruited among the students and the personnel of the Univeristà degli Studi della Campania "Luigi Vanvitelli". Participants were tested for normal hearing capabilities before taking the test. All participants gave their written consent to take part in the study. The experiment was carried out in the test room of the Sens-i Lab at the Architecture Department of the Università degli Studi della Campania "Luigi Vanvitelli".

A laptop was positioned in the center of the room. First, participants were tested for their auditory capabilities by means of the Sennheiser Hearing Test app. Once the normal hearing was verified, they sat one meter away from the laptop's screen, wore headphones (Sennheiser HD 200) and started the experiment. Participants were asked to evaluate all 72 stimuli utilizing a keyboard in front of them. After each stimulus was played, the screen prompted a choice between two possible words (see Figure 4). Participants could choose their response by tapping the left or right key on the keyboard according to what they thought was the word the avatar pronounced. Each session lasted about 25 minutes.

Figure 4. Experimental setup (left); Choice prompt example (right).

4. RESULTS

A paired-sample t-test was conducted to compare the Intelligibility Score (IS) obtained by the use of animations generated with human speech (HSA) and generated with synthesized speech (SSA). The formula provided by Bonaventura [8] was used to calculate IS and account for random choices:

$$S = \frac{100 \times (T - 2\,W)}{T} \qquad (1)$$

The results showed a significant difference between the two conditions, $t(34)=5.40$, $p<0.001$. Specifically, the scores obtained by the synthesized speech animation were significantly lower than those obtained by using human speech animation. The use of synthesized speech for generating animations negatively affects their intelligibility, leading to a significant reduction in scores compared to human speech.

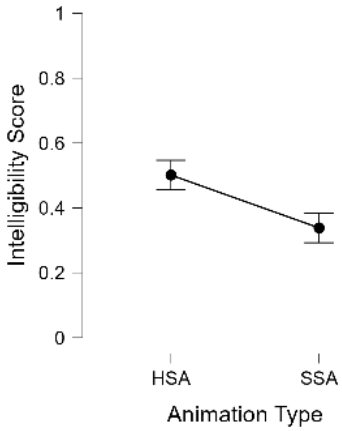

Figure 5. Average IS scores of the two methodologies.

Furthermore, a Repeated Measures ANOVA was conducted to test for interaction effects between words traits and the

type of animation, and how these would affect the intelligibility scores. The analysis featured two within-subjects variables: Animation Type (Human Speech Animation – HAS; and Synthetized Specch Animation - SSA); and Trait (Nasalità, Continuità, Stridulità, Coronalità, Anteriorità, Sonorità). The dependent variable was the mean score per each Trait, meaning only the first word of the pair is considered for the calculations (see Table 1).

Results showed the main effects of Animation Type: $F(1,34)=14.339$, $p\leq0.001$, $\eta^2_p=0.297$; and Trait: $F(5,170)=8.899$, $p\leq0.001$, $\eta^2_p=0.207$; and Animation x Trait interaction: $F(5,170)=5.987$, $p\leq0.001$, $\eta^2_p= 0.150$.

Considering the Animation Type, the Bonferroni *post-hoc* test showed HSA to be more intelligible than SSA condition, ($M=0.094$, $SE=0.025$, $p\leq0.001$).

Considering the Traits, the Bonferroni *post-hoc* test on the *Nasalità*, showed this Trait was significantly less intelligible than all other Traits, except *Sonorità*.

Considering the Animation X Trait interaction, the Bonferroni *post-hoc* test showed the HSA condition to be more intelligible than the SSA condition only when compared to the Traits *Nasalità* ($M=0.166$, $SE=0.077$, $p=0.038$), *Coronalità* ($M=0.131$, $SE=0.063$, $p=0.043$), and *Sonorità* ($M=0.369$, $SE=0.071$, $p\leq0.001$) (see Figure 6).

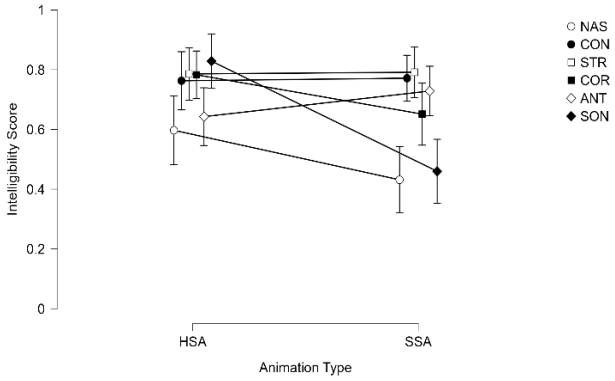

Figure 6. IS for the different word Traits.

5. DISCUSSION AND CONCLUSIONS

The present study aimed to understand the effects of different speech generation methodologies (human voice versus synthesized voice) for animating avatars' facial movements and assessing their intelligibility under challenging listening conditions. To this aim, the DRT and Epic Games' Audio-Driven plugin were used to create animations. The results provided preliminary evidence that the animations generated with the human voice enhanced intelligibility. On the other hand, considering the Traits showed that not all the stimuli benefit equally from the methodology used. This might suggest that certain phonetic characteristics, which require more or less pronounced articulatory movements, may have an important effect on the transformation of the audio input in a more refined facial animation and that some of them still can't be well represented by synthesized voices. Another consideration concerns the use of disyllabic stimuli from the DRT. Real-world communication involves continuous speech with varying intonations, which may interact differently with visual cues. It is conceivable that the observed results might vary when tested with more complex linguistic materials or different background conditions. Future research should investigate considering continuous speech and more diverse acoustic conditions to determine whether the observed results can be generalized to more complex virtual communication scenarios.

These results provide an initial step for future research aimed at developing more effective avatar-based communication.

6. REFERENCES

[1] L. E. Bernstein, E. T. Auer Jr, and S. Takayanagi, "Auditory speech detection in noise enhanced by lipreading," Speech Communication, vol. 44, no. 1-4, pp. 5–18, 2004.

[2] P. Anderson Gosselin and J. P. Gagné, "Older adults expend more listening effort than young adults recognizing speech in noise," J. Speech Lang. Hear. Res., vol. 54, no. 3, pp. 944–958, Jun. 2011. doi: 10.1044/1092-4388(2010/10-0069).

[3] C. Visentin, N. Prodi, F. Cappelletti, S. Torresin, and A. Gasparella, "Speech intelligibility and listening effort in university classrooms for native and non-native Italian listeners," Building Acoustics, vol. 26, no. 4, pp. 275–291, 2019.

[4] W. Ward, R. Cole, D. Bolaños, C. Buchenroth-Martin, E. Svirsky, and T. Weston, "My science tutor: A conversational multimedia virtual tutor," J. Educ. Psychol., vol. 105, no. 4, pp. 1115, 2013.

[5] X. Peng, H. Chen, L. Wang, and H. Wang, "Evaluating a 3-D virtual talking head on pronunciation learning," Int. J. Hum.-Comput. Stud., vol. 109, pp. 26–40, 2018.

[6] A. Devesse, A. Dudek, A. van Wieringen, and J. Wouters, "Speech intelligibility of virtual humans," Int. J. Audiol., vol. 57, no. 12, pp. 914–922, 2018.

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

[8] P. Bonaventura, A. Paoloni, F. Canavesio, and P. Usai, "Realizzazione di un test diagnostico di intelligibilità

per la lingua italiana," Fondazione Ugo Bordoni, Roma, Mar. 1986.
[9] ElevenLabs. (2025). ElevenLabs: AI text-to-speech platform. Retrieved from https://elevenlabs.io
[10] ISO 9921:2003. Ergonomics — Assessment of speech communication.
[11] J. Peirce, J. Gray, Y. Halchenko, D. Britton, A. Rokem, and G. Strangman, "PsychoPy-Psychology software for Python," Integration The VLSI Journal, 2010.