forum acusticum 2026
8-12 September · Graz, Austria

# HOW WELL DO AVATARS SPEAK? FRAMEWORK FOR AVATAR FACIAL ANIMATION CONTROL

Antoine Bourachot[1], Bernhard U. Seeber[1]

[1] *Audio Information Processing, Technical University of Munich, 80333 Munich, Germany*



## Abstract

Successful communication relies, among other factors, on the accurate perception of the acoustic signal. However, speech understanding is not purely an acoustic process: visual information from the speaker's face, especially lip movements, plays a key complementary role. These cues become particularly valuable in challenging or noisy listening conditions. As virtual environments expand in both research and everyday communication, digital avatars are increasingly used not only as a medium for interaction but also as a tool for investigating human communication processes.
This work presents a modular pipeline for animating and controlling realistic avatars in virtual environments. The proposed approach relies on video recordings as reference material for the production of facial animations, allowing for a precise treatment of complex visemes. The pipeline is designed in a modular way to integrate within our virtual environment. Particular attention is paid to audio-visual alignment. Visual animation onset is measured using a photodiode and compared with the corresponding audio onset. The results show that the audio systematically precedes the visual animation, allowing the misalignment to be compensated for by introducing a fixed audio delay.

**Keywords:** *avatar, facial animation, unreal engine, audio-visual synchronization*

## 1. Introduction

The study of human communication in a laboratory setting raises many questions and challenges, as our everyday environment is highly complex and dynamic. Reproducing this environment is therefore crucial to creating research contexts that are consistent with real-life situations. In everyday situations, our sound environment is generally dynamic, featuring background noise, moving sound sources, reverberation, and so on. Our visual environment allows us to confirm or validate certain auditory information. In fact, in complex listening situations, numerous visual cues supplement the auditory information and enhance it.
The real-time Simulated Open Field Environment (rtSOFE) [1]–[3] was developed to provide dynamic soundscapes that include multiple moving sound sources and the reverberation specific to the space represented by each of these sources. It is augmented by a 4-wall video projection system that creates complex audiovisual environments. This CAVE-like audiovisual space is installed in an anechoic chamber at TUM and permits controlled and reproducible audiovisual studies. A virtual reproduction of the Theresienstraße subway station in Munich has been developed and verified [4], [5].
Several studies have been conducted in this virtual environment, characterizing individual communication behavior during 2- or 3-way conversations [6], [7]. Yet, studying the impact of particular visual cues or feedback strategies is difficult, as they may arise seldom or in different contexts in free-flowing communication. This is where avatars can help, as they offer total control over possible interactions, as well as complete reproducibility across sessions. Here, we present a new method developed in our laboratory to achieve highly realistic rendering of facial movements using animation control plug-ins. The animation pipeline is verified by examining synchronization between our visual and audio systems.

## 2. The use of avatars in perceptual studies

The use of avatars and virtual agents has become widespread in research into mediated communication, particularly in the fields of education and training, remote collaboration, and human–machine interaction. These digital representations are used to study various dimensions of interaction, such as social presence and the perceived quality of communication [8]–[10].

***Corresponding author**: antoine.bourachot@tum.de.*

In the context of spoken communication, one of the main challenges concerns the reproduction of facial movements associated with speech, particularly those of the lips and jaw. Sumby and Pollack [11] found in the 1950s that observing the speaker's face significantly improves speech intelligibility, particularly when the acoustic signal is degraded by noise. Several studies have since shown that animated faces and talking avatars can also offer a perceptual advantage over audio-only presentations [12]–[14]. However, the significance of this benefit varies with the quality and methods used to generate the movements.

Various approaches can be used to generate an avatar's facial movements. Audio-driven methods generate movements of the lips, jaw, and sometimes the rest of the face based on the characteristics of the speech signal. Video-driven or performance-capture methods transfer the movements observed on a real speaker's face to the avatar. Other approaches rely on parametric models, visemes, action units, or combinations of predefined facial shapes. These approaches involve different trade-offs in terms of experimental control, production cost, personalization, audio-visual synchronization, and perceived naturalness.

A particularly interesting experimental approach in this context is the "Wizard of Oz" paradigm [15]. In this type of protocol, the participant believes they are interacting with a fully or partially autonomous system, whilst a hidden experimenter controls certain verbal or non-verbal responses of the virtual agent. Utilizing a pre-recorded animation and dialogue script, the experimenter selects the appropriate response to deliver. As responses are pre-recorded, their reproduction quality can be analyzed, and multimodal cues can be carefully controlled and manipulated.

## 3. Avatar animation production

In order to produce high-quality avatars and utilize animation within our virtual environment, we employ a pipeline that combines state-of-the-art commercial products and in-house-developed tools. The avatar animation is divided into two parts: body animation and facial animation. Each has its own acquisition pipeline, but both rely on recording an actor. Once created, these animations can be stored and reused in our perceptual studies (see Fig. 1).

High-resolution facial data (1080p / 60 fps) is captured using a Faceware Mark IV Wireless Headcam system and its associated software suite (Faceware Technologies, Inc., Los Angeles, CA, USA). This solution employs two primary modules: Analyzer, for facial landmark extraction, and Retargeter, which maps movements onto 3D characters within Autodesk Maya (Autodesk, San Francisco, CA, USA). Although this offline process includes automated components, we maintain manual control to adjust individual frames and refine specific visemes that may be problematic during automated extraction, ensuring expressive and faithful performances.

Simultaneously, full-body motion data are captured using 12 OptiTrack Prime 17W cameras and extracted via Motive software (OptiTrack, Corvallis, OR, USA). These movements can be transferred to arbitrary avatars within a game engine through motion retargeting, implemented via our plugin rtMotion (see Section 4.2).

Finally, the exported animation data is processed in Maya using a custom Python script that extracts animation curve values for each articulated element (e.g., bones or rig controllers). The resulting data are stored in MATLAB .mat files, containing a data vector for each animated element and a corresponding timestamp vector for the animation timeline.

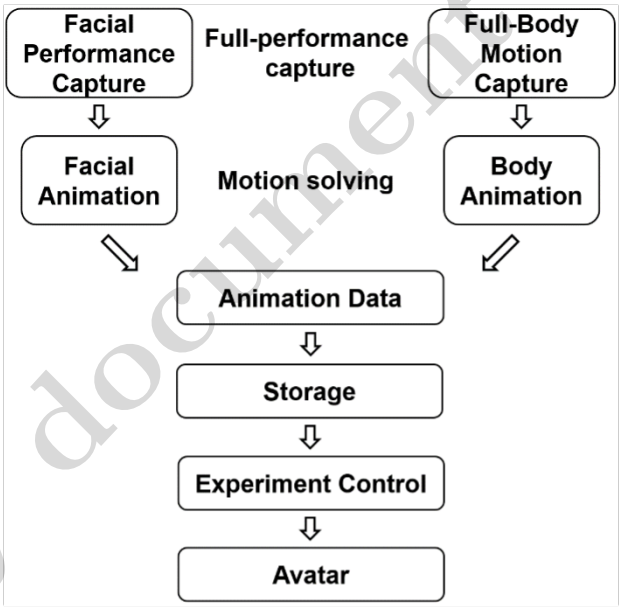


**Figure 1:** General Concept of the Avatar Pipeline

## 4. Experiment control plugins

To conduct our perceptual experiments, we require a system to playback and manipulate the captured animations within a unified framework. Our experimental setup for perceptual studies uses MATLAB (The MathWorks, Inc., Natick, MA, USA) for data acquisition, stimulus triggering, and audio processing. Unreal Engine 5 (Epic Games, Cary, NC, USA) is used for 3D rendering and interactive environments, and rtSOFE [3] for auralization. While previous work [16] focused on integrating auralization for moving sound sources in Unreal Engine, our current contribution extends this framework to support precise animation control. To this end, we developed two in-house plugins: matStreamer, which enables control via MATLAB, and rtMotion, which handles data reception in Unreal Engine. These tools allow us to manage both the deterministic playback of pre-recorded sequences and the real-time manipulation of specific animation components, such as facial expressions, thereby enabling the isolation of specific visual cues.

### 4.1. matStreamer

matStreamer is the MATLAB-side sender in the avatar control pipeline. It loads pre-recorded facial and body animation data and streams them to Unreal Engine in real time using the Open Sound Control (OSC) protocol. rtMotion (see Section 4.2) receives the OSC messages and applies the corresponding motion to the MetaHuman character. Start-up information, such as control ports, destination hosts, and the initial list of characters, is set in a configuration file. After startup, separate runtime components manage assets, character transports, playback sessions, command handling, and so on.

All available animations are preloaded, and OSC messages containing all animation elements and the timestamp are prepared in advance to be triggered at the desired time.

The main motivation behind this plugin is to give experimenters direct control over the animations sent to the avatar from within MATLAB, including the possibility to manipulate particular animation dimensions. Since MATLAB is already widely used in our laboratory for experimental control, data acquisition, and signal processing, integrating avatar animation into the same environment greatly simplifies its incorporation into existing experimental workflows.

### 4.2. rtMotion

rtMotion is an Unreal blueprint plugin containing the elements needed to bring a MetaHuman to life using our pipeline. Its main component, rtMotionParameter, allows users to assign the OSC receive ports for facial and body animations to a Metahuman, along with the necessary settings. This module will receive the data from matStreamer and distribute it amongst the avatar's various animation blueprints.

The plugin also provides animation blueprints for the face and body, ready to receive animation values from matStreamer. Body animations are also provided with a generic skeleton derived from the OptiTrack model, along with the retargeting files required to link these skeletons to the MetaHuman.

The facial animation Blueprint also contains a trigger to rtSOFE to trigger audio file playback when Unreal actually receives and processes the animation data.

## 5. Audio-visual synchronization

To ensure a visually and aurally coherent scene, audiovisual synchronization must be verified. By separating the animation, audio, and video pipelines to address our various research challenges, these elements may become out of sync. In this section, we will examine the framework within which our environment is set up, as well as the method we developed to evaluate audiovisual delay.

### 5.1. Audiovisual environment

The CAVE-like SOFE setup described in the introduction, in which we conduct our experiments, consists of four video projectors controlled by one computer and 61 speakers controlled by another. It also includes motion capture (MOCAP) management and audio simulation (calculation of audio impulse responses). A description of the system can be found in [2].

The delays between receiving the animation, triggering the audio, and the final display on the screen can therefore be relatively significant.

### 5.2. Measurement method

We measure the time lag between the image projected by a video projector and the audio signal triggered during animation processing. To do this, we use a photodiode (PD333-3C/H0/L2) connected via a resistor network to a memory oscilloscope (Tektronix MSO4014B, Tektronix, Inc., Beaverton, OR, USA), receiving also the audio signal. The data sent by matStreamer for facial animation contains all facial control parameters, but only one value changes: "C_jaw_Y", that is, the parameter controlling the opening and closing of the jaw, which changes from 1 (mouth open) to 0 (mouth closed) after one second. (see Fig. 2)

The audio signal used is a 10 kHz sinusoidal signal (sample frequency of 44.1 kHz), played back via the rtSOFE convolver on the frontal loudspeaker with loudspeaker EQ active. The sound plays for one second as well, and then returns to zero. The recording is made at the center of the speaker array with a Microtech MM 210 measurement microphone (Microtech Gefell GmbH, Gefell, Germany) connected to an RME Micstasy Preamp (RME/Audio AG, Haimhausen, Germany).

The signals recorded are then analyzed in MATLAB. The time delay is determined from the relative shift between the envelopes of the two signals.

### 5.3. Results

A series of 10 measurements was taken. The measured delay between the audio and video is 15.89 ± 5.83 ms (mean ± SD), with the audio leading the video. Beyond the observed audio lead over the video, the measured jitter remains consistent with the 16.7 ms of a single frame in Unreal Engine running at 60 fps.

## 6. Discussion and future work

The measurements reveal a consistent lead of the audio over the visual output, which can be compensated by introducing a constant delay to achieve audio-visual synchronization. While some jitter is present, it remains in the frame time. This confirms the framework's readiness for more complex tasks.

The next stage of this work is to evaluate the full use of this pipeline for avatar creation with a view to conducting 'Wizard of Oz'-style experiments. This evaluation will focus on the quality of the avatars' visual cues for speech perception in noisy environments.

We present a pipeline that enables the creation and manipulation of avatar animations based on real-world recordings, with modular integration via various plugins for avatar control within the experimental frameworks developed in our laboratory. Measuring the delay of our signal allows us to synchronize the audio and video with frame-level precision.

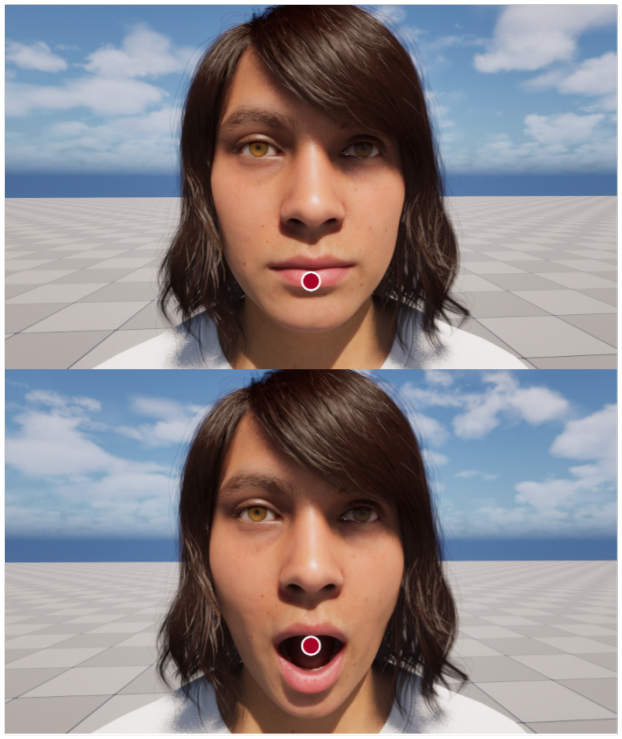

**Figure 2:** Animation signal used to measure the delay. The red dot represents the position of the photodiode in front of the screen.

## 7. Acknowledgments

Dr. Luboš Hládek, Kaya Cem, Emre Parlak, and Shilai Zhan made a valuable contribution to the development of parts of the animation pipeline.
Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 352015383 – SFB 1330, Project C5.

## References

[1] B. U. Seeber, S. Kerber, and E. R. Hafter, "A system to simulate and reproduce audio–visual environments for spatial hearing research," *Hearing research*, vol. 260, no. 1-2, pp. 1–10, 2010. DOI: 10.1016/j.heares.2009.11.004.

[2] B. U. Seeber and S. W. Clapp, "Interactive simulation and free-field auralization of acoustic space with the rtSOFE," *The Journal of the Acoustical Society of America*, vol. 141, no. 5_Supplement, pp. 3974–3974, 2017. DOI: 10.1121/1.4989063.

[3] B. U. Seeber and T. Wang, *Real-time Simulated Open Field Environment (rtSOFE) software package*, Language: eng, Nov. 2021. DOI: 10.5281/zenodo.5648305.

[4] L. Hládek, S. D. Ewert, and B. U. Seeber, "Communication conditions in virtual acoustic scenes in an underground station," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, IEEE, 2021, pp. 1–8. DOI: 10.1109/I3DA48870.2021.9610843.

[5] L. Hladek and B. U. Seeber, *Underground station environment*, eng, Feb. 2022. DOI: 10.5281/zenodo.6025631.

[6] L. Hladek and B. Seeber, "Effects of noise presence and noise position on interpersonal distance in a triadic conversation.," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 265, Institute of Noise Control Engineering, 2023, pp. 5326–5331. DOI: 10.3397/IN_2022_0776.

[7] L. Hládek and B. U. Seeber, "On speech-hand synchrony during conversations in a virtual underground station," in *Fortschritte der Akustik–DAGA'24*, 2024, pp. 1230–1232. [Online]. Available: https://mediatum.ub.tum.de/doc/1742357/document.pdf.

[8] A. D. Fraser, I. Branson, R. C. Hollett, C. P. Speelman, and S. L. Rogers, "Do realistic avatars make virtual reality better? Examining human-like avatars for VR social interactions," *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 2, p. 100 082, 2024. DOI: 10.1016/j.chbah.2024.100082.

[9] K. L. Nowak and J. Fox, "Avatars and computer-mediated communication: A review of the definitions, uses, and effects of digital representations," *Review of communication research*, vol. 6, pp. 30–53, 2018. DOI: 10.12840/issn.2255-4165.2018.06.01.015.

[10] C. Kyrlitsias and D. Michael-Grigoriou, "Social interaction with agents and avatars in immersive virtual environments: A survey," *Frontiers in Virtual Reality*, vol. 2, 2022. DOI: 10.3389/frvir.2021.786665.

[11] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954. DOI: 10.1121/1.1907309.

[12] A. Devesse, A. Dudek, A. Van Wieringen, and J. Wouters, "Speech intelligibility of virtual humans," en, *International Journal of Audiology*, vol. 57, no. 12, pp. 914–922, Dec. 2018, ISSN: 1499-2027, 1708-8186. DOI: 10.1080/14992027.2018.1511922.

[13] Y. Yu, A. Lado, Y. Zhang, J. F. Magnotti, and M. S. Beauchamp, "Synthetic faces generated with the facial action coding system or deep neural networks improve speech-in-noise perception, but not as much as real faces," *Frontiers in Neuroscience*, vol. 18, p. 1 379 988, 2024. DOI: 10.3389/fnins.2024.1379988.

[14] J. Nirme, B. Sahlén, V. Lyberg Åhlander, J. Brännström, and M. Haake, "Audio-visual speech comprehension in noise with real and virtual speakers," en, *Speech Communication*, vol. 116, pp. 44–55, Jan. 2020, ISSN: 01676393. DOI: 10.1016/j.specom.2019.11.005.

[15] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of Oz studies: Why and how," en, in *Proceedings of the 1st international conference on Intelligent user interfaces - IUI '93*, Orlando, Florida, United States: ACM Press, 1993, pp. 193–200, ISBN: 978-0-89791-556-4. DOI: 10.1145/169891.169968.

[16] F. Enghofer, L. Hládek, and B. U. Seeber, "An'Unreal'Framework for Creating and Controlling Audio-Visual Scenes for the rtSOFE," in *Fortschritte der Akustik-DAGA'21*, 2021. [Online]. Available: https://mediatum.ub.tum.de/doc/1631461/document.pdf.