Minimal Latency Speech-Driven Gesture Generation for Continuous Interaction in Social XR

Niklas Krome & Stefan Kopp, Bielefeld University

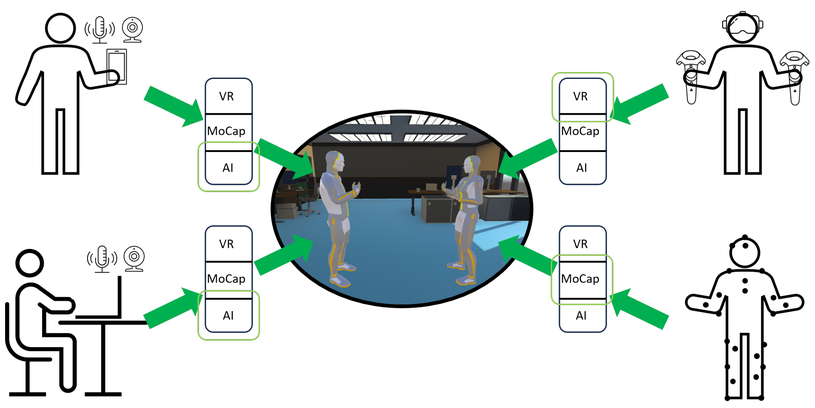
Motivation

Social XR applications usually require advanced tracking equipment to control one's own avatar. We explore if AI-based co-speech gesture generation techniques can be employed to compensate for the lack of tracking hardware that many users face. One main challenge is to achieve convincing behavior quality without introducing too much latency.

Approach

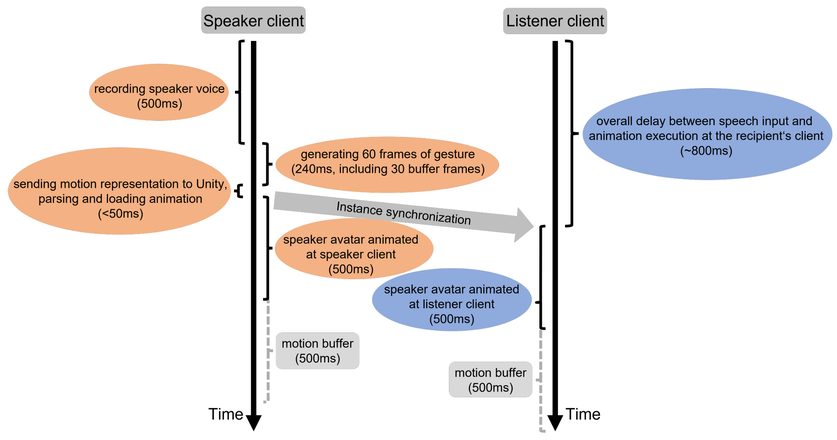
Overview

- Shared Social XR environment
- Realistic behavior, independent of input device
- Speech-driven gestures when no tracking is available
- No impact on the immersion of other users



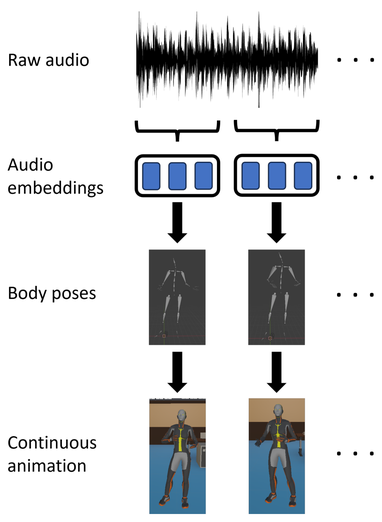
The Latency Problem

- Chunk size determines largest portion the resulting latency
- Overall lag of more than 200ms impacts interaction quality
- Lower chunk sizes decreased motion quality significantly



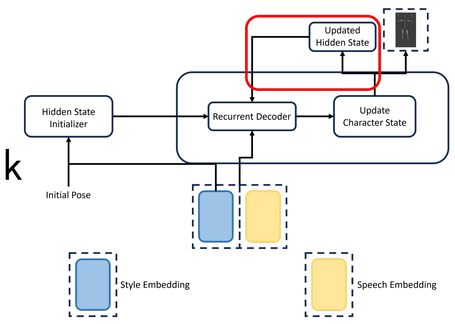
Incremental Generation

- Provide audio in increments (chunks)
- Calculate speech embeddings of sufficient length
- Generate partial gestures (gesture chunks)
- Visualize in Unity as a continuous motion sequence
- Synchronize over network for multiplayer interaction



Frame-wise Generation

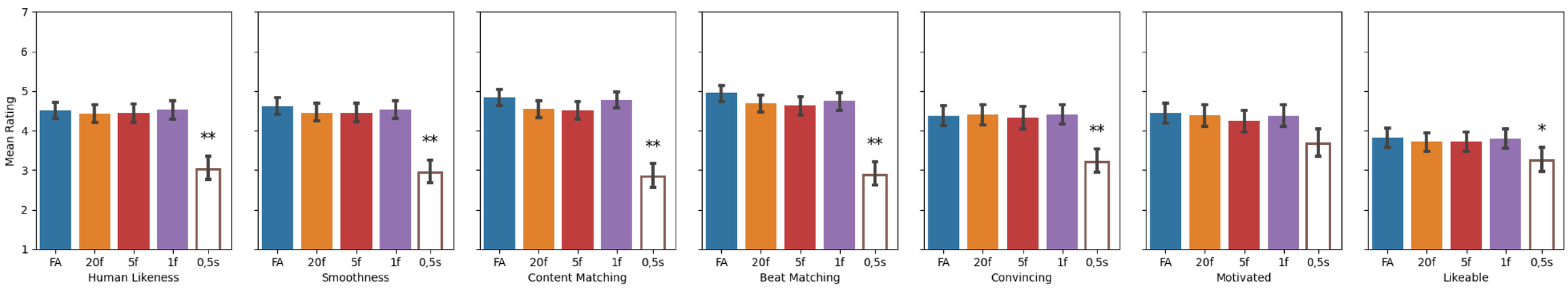
- Recurrent decoder architecture, adapted from ZeroEGGS [1]
- Chunk size reduced to a single frame
- State information preserved over chunk borders
- Minimized information generated for each frame



[1] Ghorbani, S., Ferstl, Y., Holden, D., Troje, N.F. and Carbonneau, M.-A. (2023), ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. Computer Graphics Forum, 42: 206-216. https://doi.org/10.1111/cgf.14734

Results

- Motion quality no longer decreases with smaller chunk sizes and matches that of the baseline model [1]
- Minimum possible chunk size is now 16ms (1 frame at 60fps), down from 500ms with the previous system (hollow bar)
- System now allows for online interaction via speech-driven gestures (latency < 200ms)



Next Steps

- Extension to adapt gesture output in real-time, according to the conversational context
- Study on perception of generated gestures in real interaction

Funded by the German Federal Ministry of Education and Research

Bundesministerium für Bildung und Forschung