Kurzbericht

Die Aufgabenstellung des Teilvorhabens war es, allen Nutzern in Social XR (auch bei unterschiedlichen Endgeräten) Avatare mit hinreichend realistischem Konversationsverhalten zu ermöglichen. Dazu sollten innovative KI-gestützte (Lern-)Verfahren zu entwickeln, um realistische Bewegungen und realistisches Verhalten auch bei eingeschränkten Sensordaten zu erreichen, in dem (Teil-)Verhalten der Avatare autonom synthetisiert und integriert werden, ohne dass es dabei zu Brüchen in der Reproduktionstreue kommt. Anschließend an den aktuellen Forschungsstand wurden dazu Verhaltensdaten (Sprache, Mimik, Blick, Kopf-/Körpergesten, Körperhaltung/-orientierung) in Mehr-Nutzer-Konversationen erhoben und für multimodale maschinelle Lernverfahren eingesetzt, die Hintergrundwissen über Interaktionszustände oder Interaktandenrollen in Form von latenten Faktoren identifizieren und einbeziehen können. Konkret wurden so die Vorhersage von sprachbegleitenden Gesten, Kopfbewegungen und Blickbewegungen anhand sprachlicher Merkmale einzeln und in Kombination betrachtet.

Im Einzelnen wurde eine projektspezifische Laborinfrastruktur aufgebaut bzw. bereitgestellt und ein Datensatz über verschiedene, dyadische Alltagsinteraktionen mithilfe der verfügbaren Aufnahmemodalitäten durchgeführt. Der Stand der Technik im Bereich der Gesten-Generierung hinsichtlich der für das Projekt relevanten Kriterien Qualität und Latenz wurde evaluiert und Konzepte zur Verbesserung der Nutzbarkeit von Gesten-Generierungs-Systemen in Echtzeit-Szenarien wurden entwickelt. Gleichzeitig wurde erforscht, welche Kontextfaktoren die Echtzeit-Vorhersage von Interaktionsverhalten in dyadischer Interaktion in Social XR ermöglichen würden. Neben Dialogrolle und emotionalem Zustand des Sprechers wurde auch versucht, höhere strukturelle Informationen über die Gestenphase zu gewinnen, wozu ein System zur Erkennung verschiedener Gestenphasen auf Basis eines Webcam-Feeds implementiert und evaluiert wurde.

Das nach Analyse des State-of-the-Art bestgeeignete Modell zur Generierung sprachbegleitender Gesten (ZeroEGGS [2]) wurde nachimplementiert und an die projektspezifischen Anforderungen angepasst. Zum einen wurden ein System zur Übertragung der so generierten Gesten auf virtuelle Avatare in Unity entwickelt und ein Netzwerk-Layer zwischen mehreren Unity-Instanzen zur Evaluierung von Multi-User Szenarien eingerichtet. Zum anderen wurde das genutzte Modell erweitert, um nun auch Gesten kontinuierlich auf Basis von Echtzeitsprache generieren zu können. Die Ergebnisse wurden auf einem Avatar in Unity visualisiert und in einer Online-Videostudie evaluiert. Das Modell sowie die Studienergebnisse, die einen starken Trade-off zwischen Qualität und Latenz der Gesten-Synthese zeigten, wurden auf der Fachkonferenz ACM IVA 2023 vorgestellt und veröffentlicht. Das Problem der niedrigen Qualität bei notwendig geringer Latenz wurde durch eine Modell-Erweiterung (latenter Kontextvektor) behoben, so dass das ZeroEEGS-Modell nun Gesten auf Basis einzelner Frames produzieren kann, ohne dabei Kontextinformation zu ignorieren. Dies führte zu einer signifikant erhöhten Qualität bei gleichzeitiger Reduktion der resultierenden Latenz. Das so erweiterte Modell wurde auf der Fachkonferenz IEEE AIxVR 2024 veröffentlicht und dort mit einem Best Poster Award ausgezeichnet.

Eine zweite zentrale Aufgabe bestand darin, zu untersuchen, wie die Kontextabhängigkeit von Interaktionsverhalten ausgenutzt und technisch umgesetzt werden kann. Als Kontext wird hier sowohl die Dialogrolle (Sprecher/Zuhörer), als auch emotionale (z.B. Zufriedenheit, Ausgeglichenheit oder Aufgeregtheit) oder konversationale Zustände (z.B. Zustimmung oder Ablehnung) betrachtet. Die Gestensynthese wurde dafür vorbereitet, indem sogenannte „Stil-Vektoren“ des vortrainierten ZeroEEGs-Modells genutzt werden, um das laufend synthetisierte Verhalten in Echtzeit zu

modulieren. Gleichzeitig wurde das Spektrum des generierbaren Avatarverhaltens erweitert, indem weitere Komponenten zur automatischen Generierung von Zuhörerverhalten entwickelt und in Nutzerstudien evaluiert wurden. Ein Modell analysiert in Echtzeit die Sprache und generiert modulierbares Nickverhalten. Eine andere Komponente erzeugt Posture Mimicry-Verhalten als Teil des automatischen Zuhörerverhaltens. Analog wurde eine Komponente zur Generierung realistischen Blickverhaltens sowie zur direkten Kontrolle der Blickrichtung des Nutzer-Avatars entwickelt.

Für die Nutzung und Integration der entwickelten Modelle wurde Techniken eingesetzt, die es ermöglichen entsprechende Kontextfaktoren in Echtzeit aus dem Nutzerverhalten zu extrahieren. Zum einen wurde ein System zur Vorhersage zukünftiger Sprecher-Zuhörer Rollen in dyadischer Interaktion [3,4] nachimplementiert und angepasst, um in Echtzeit zwischen Rollen-spezifischen Modellen zur Bewegungssynthese zu wechseln. Zum anderen wurde ein Modell zur automatischen Erkennung passender Gesten-Stile aus dem Sprachsignal mittels Emotionserkennung [5,6] implementiert und zur automatischen Anpassung der Stil-Vektoren verwendet (eine alternative Methode zur nutzerseitigen Anpassung der Stil-Vektoren über UI-Elemente wurde ebenfalls implementiert).

Alle technischen Komponenten wurden in die HiAvA-Gesamtarchitektur integriert und eingepflegt. Das entwickelte Modell zur Generierung sprachbegleitender Gesten wurde in einer Studie mit den Projektpartnern in Würzburg in asymmetrischen dyadischen Interaktionsszenarien zwischen Teilnehmern verwendet und evaluiert. Eine gemeinsame Veröffentlichung der Ergebnisse über die Auswirkungen der so generierten Gesten in dyadischer Kommunikation befindet sich in Vorbereitung.

Schlussbericht

Beschreibung der erzielten wissenschaftlich-technischen Ergebnisse

In dem Projekt „*Skalierbare hybride Avatar-Agenten-Technologien für alltagstaugliche soziale Interaktion in XR* (HiAvA)“ hat die Universität Bielefeld die Teilvorhaben zum Bereich KI-unterstützte Verhaltensmodelle in den Arbeitspaketen 5 und 6 wie im Projektplan vorgesehen umgesetzt. Dabei wurden im Einzelnen folgende Ergebnisse erzielt.

AP 5: KI-unterstützte Verhaltensmodelle für Sprecher-Avatare/Agenten

Aufnahme von Trainingsdaten (AP 5.1): Um Trainingsdaten zu erheben und Evaluationsstudien durchführen zu können, wurde ein Konzept eines Social XR-Setups entwickelt. Dazu wurde festgelegt, wie die mittels verschiedener Modalitäten aufgenommenen Verhaltensdaten von zwei Personen (die per Avatar in einer Social XR-Umgebung interagieren) zur Entwicklung eines Gesten-Generierungs-Systems genutzt werden können und welche Parameter abseits der reinen Datenaufnahme von den Proband*innen erhoben werden sollten. Zur Durchführung dieser Interaktionsstudie und zur späteren Ansteuerung der Avatare im HiAvA-Kontext wurden dann die notwendigen Beschaffungen getätigt. Dies umfasste die Aufrüstung des Motion Capture Systems OptiTrack in Form eines Hand-Kits, welches Finger-Tracking ermöglicht, sowie die Ausstattung des Labors mit einer vollständigen Virtual-Reality Ausrüstung, also einem HTC Vive Pro Eye Headset inklusive Zubehör und einem VR-fähigen Rechner. Außerdem wurde ein iPhone 12 Pro Max angeschafft, welches der Aufnahme von Gesichtsbewegungen während der Interaktionsstudie dienen sollte. Die vorgesehene Studie zur Aufnahme von Trainingsdaten zur Gestensynthese wurde in Zusammenarbeit mit den anderen Projektpartnern durchgeführt (im Umfang von 8 dyadischen Alltagsinteraktionen). Die Aufnahmen geschahen mittels dreier Systeme zur Aufnahme von Ganzkörperbewegungen (OptiTrack Motive 3, Captury, HTC Vive Tracker) sowie eines vor dem Gesicht montierten iPhones, welches unter Verwendung von Apple’s ARKit die Gesichtsausdrücke aufnahmen.

Online-Sprachanalyse zur sprachbasierten Gestensynthese (AP 5.2 und 5.3): Ein Hauptvorhaben war es zu untersuchen, wie nonverbales (insbesondere gestisches) Verhalten der Avatare aus dem Sprachsignal synthetisiert werden kann, wenn keine entsprechenden technische Mittel zur direkten Aufnahmen (Motion Tracking) des Nutzenden gegeben sind (z.B. bei Nutzung eines mobilen Endgerätes). Es wurde eine ausführliche Evaluation des derzeitigen Standes der Technik im Bereich der automatischen Gestengenerierung durchgeführt. Ansätze wurden hinsichtlich ihrer Qualität untersucht, aber vor allem mit Blick auf die für das Projekt HiAvA relevanten Aspekte betrachtet. Die im Projekt vorgesehen Eingabemodalitäten führten zur Auswahl rein auf Sprachbasis arbeitender Systeme, da bei der direkten Ansteuerung der Avatare keine Text-Transkription vorhanden sein würde. Auch wurden die Systeme auf ihre Echtzeit-Fähigkeit hin geprüft, da eine schnelle und kontinuierliche Vorhersage der richtigen Geste für eine gegebene sprachliche Äußerung im Projekt von essenzieller Bedeutung war.

Als ein möglicher Schritt zur Verbesserung der Qualität und Geschwindigkeit der Gestenvorhersage wurde die Einbindung gegebener visueller Information erforscht. Ein Kamerabild wie es beispielsweise bei der Bedienung eines HiAvA-Avatars über einen Desktop-PC entstehen könnte wurde verwendet, um Informationen über die aktuelle „Gestenphase“ (also z.B. Vorbereitung, expressive Phase oder Rückführungsphase) in Echtzeit vorherzusagen. Die Annahme war, dass diese Informationen in einer inkrementell aufgebauten Gesamtarchitektur für die Gestensynthese

informativ und zudem schneller erkennbar sein würden, als bei ausschließlicher Nutzung des Sprachsignals. Um dieses zu untersuchen, wurde ein Modell zur Erkennung von Gestenphases entwickelt. Dabei wurde auf die Nutzung effizienter Architekturen geachtet. So kamen anstelle der häufig verwendeten LSTM-basierten Modelle, Convolutional Neural Networks zum Einsatz, was abermals zu einer erhöhten Vorhersagegeschwindigkeit führen sollte.

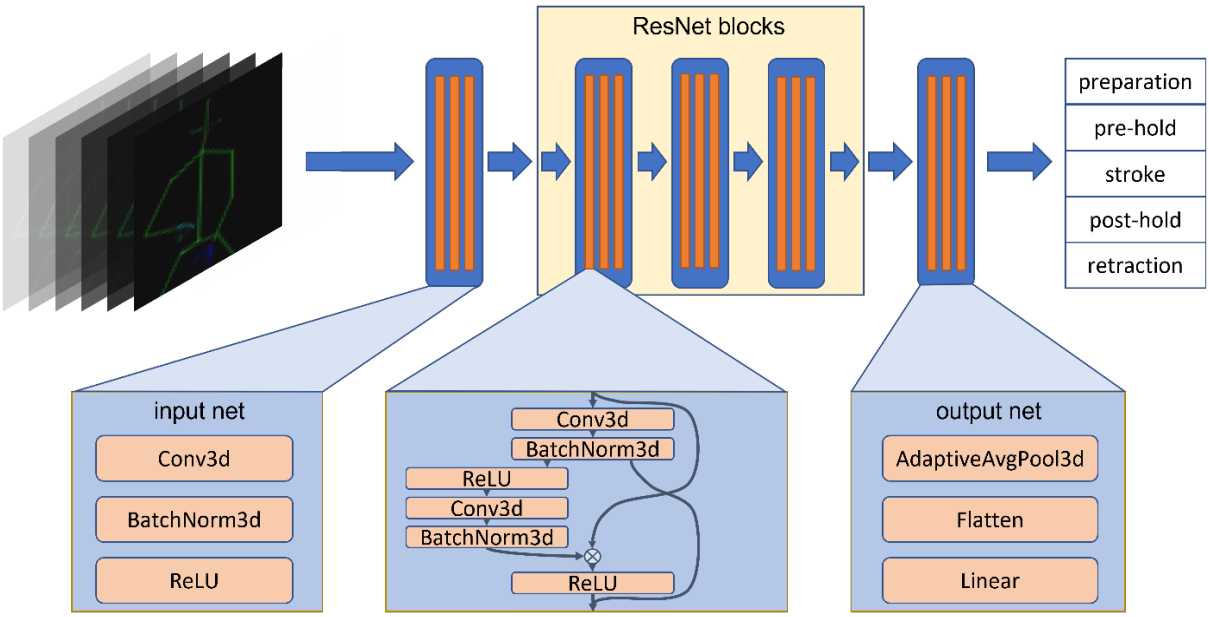

Abbildung 1: Schematischer Aufbau des Modells zur Erkennung von Gestenphasen aus Videosequenzen zur Vorhersage passender Gesten. Frame-Pakete, welche Skelett-Informationen enthalten, werden analysiert, um die darin enthaltene Gestenphase vorherzusagen.

Bei dem verwendeten Modell (siehe Abbildung 1) handelt es sich um ein dreidimensionales ResNet, welches sich als eine effiziente Art der Verarbeitung von Bildstapeln bewährt hat [1]. Die der Architektur zur Verfügung gestellten Bilddaten sind einzelne Frames eines Kamera-Feeds, welche vorverarbeitet werden, um lediglich eine Skelett-Repräsentation der zu analysierenden Person zu zeigen. Ein Stapel von 10 Bildern (herunterskaliert auf 64x64 Pixel) wird jeweils eingelesen und anschließend von einem Neuronalen Netz verarbeitet. Dieses besteht wiederum aus einem Eingabenetz, einem Ausgabenetz und einem sequenziellen Stapel sogenannter „residual blocks“. Das Eingabenetz besteht aus einer dreidimensionalen Faltung, gefolgt von einer dreidimensionalen Normalisierungsschicht und einer ReLU-Aktivierungsschicht. Die residual blocks enthalten jeweils kleinere sequenzielle Netze aus einer Faltungsschicht, einer Normalisierungsschicht, einer ReLU-Aktivierungsschicht, sowie einer weiteren Faltung und Normalisierung. Anschließend werden die Ausgaben, sowie die ursprüngliche Eingabe des Blocks summiert und erneut durch eine ReLU-Aktivierungsschicht geschickt. Dieser Block wird in dreimal drei Gruppen wiederholt, nach dem vierten und siebten Block wird jeweils eine Unterstichprobe der Daten genommen, um das Netz zu verkleinern. Das Ausgabenetz besteht aus einer dreidimensionalen adaptiven Zusammenlegungsschicht, einer Glättung, sowie einer linearen Zuordnung auf die fünf möglichen Ausgabewerte, welche den gesuchten Gestenphasen entsprechen. Bei der Erprobung dieses Ansatzes erwies sich das Modell allerdings als zu ungenau. Im weiteren Projektverlauf wurde diese Technologie daher verworfen, da sich gezeigt hat, das Gestensynthese auch ohne diese Zusatzinformationen in guter Qualität und mit einer ausreichend niedrigen Latenz erreicht werden kann.

Basierend auf einer detaillierten Evaluation moderner Gestengenerierungssysteme, wurde eine Aufstellung der Anforderungen erarbeitet, welche die Anwendung in einem Echtzeitkommunikationsszenario in Social XR ermöglichen: inkrementelle Gestensynthese, keine Verwendung von Information über zukünftige Sprache, Sprache nicht in Textform, effiziente Inferenz, Kontrolle über Startpose der resultierenden Geste. Diese Anforderungen sind neuartig und werden von vorhandenen Ansätzen (die für die Offline-Synthese aus kompletten Sprachäußerungen von mehreren Sekunden Länge entwickelt wurden) nicht erfüllt. Anhand dieser Anforderungen wurde das bestgeeignete aktuelle System ausgewählt, wobei es sich um die ZeroEGGS-Architektur von Ghorbani et. al. handelte [2]. Das Modell von Ghorbani et. al. wurde technisch verfügbar gemacht und in die HiAvA-Umgebung integriert. Zusätzlich musste das Modell nachtrainiert und für unsere Zwecke technisch erweitert werden. Diese Anpassungen ermöglichen erstmal die inkrementelle Generierung kleiner Gestenfragmente, basierend auf Echtzeit-Sprachinput. Diese Gestenfragmente können anschließend über eine Websocket Verbindung an eine VR/XR-Umgebung gesendet und dort zu einer kontinuierlichen, unterbrechungsfreien Bewegung kombiniert sowie visualisiert werden.

Wir nutzten für die Prototypenentwicklung Unity 3D und realisierten darin eine lokale Testumgebung, die Avatar-basierte Interaktion mehrerer Personen über das Internet ermöglicht. Die Skalierbarkeit und Portierung in die HiAvA-Umgebung wurde von Anfang an berücksichtigt. In der umgesetzten Umgebung interagieren Avatare via Sprache und können mithilfe der modifizierten ZeroEGGS-Architektur sprachbegleitende Gesten in Echtzeit ausführen. In einer ersten Version wurde ZeroEGGS mit unterschiedlichen großen Sprachsegmenten als Eingabe getestet. Um die Qualität der resultierenden Echtzeit-Gesten zu evaluieren, wurde eine Online-Studie durchgeführt, welche subjektive Bewertungen der generierten Gesten in verschieden langen Inkrementen zum Ziel hatte. Die Länge der Inkremente wirkte sich hierbei auf die minimale Zeit aus, bis infolge eines eingehenden Sprachabschnitts eine zugehörige Geste ausgeführt werden kann. Die Ergebnisse wiesen einen signifikanten Trade-off zwischen Gestenqualität und der Reaktivität des Systems auf Sprachinput auf. Abbildung 2 zeigt eine illustrierende Darstellung der inkrementellen Gestengenerierung, wie sie in unserem System zum Einsatz kommt. Abbildung 3 zeigt die Ergebnisse der Evaluationsstudie. Zu sehen sind die subjektiven Bewertungen der Teilnehmer*innen, nachdem Sie Videos der generierten Gesten unseres Systems sowie zweier Kontrollgruppen gesehen haben. Die Kontrollgruppen waren die unveränderte ZeroEGGS-Architektur sowie die Ground Truth-Bewegungen der Sprecher zur dazugehörigen Sprache.

Aus den Ergebnissen der Evaluation des Modells war eine gute Qualität (vergleichbar mit oder sogar besser als Ground Truth) bei unzureichender Latenz, sowie eine unzureichende Qualität bei akzeptabler Latenz (Segmente von 0.5 Sekunden Länge oder weniger) ersichtlich. Das von uns entwickelte System zur inkrementellen Generierung von Gestensequenzen in Echtzeit und die Ergebnisse der Studie auf der ACM Int. Conference on Intelligent Virtual Agents (IVA) 2023 publiziert und präsentiert [7]. Zudem wurde auf der Konferenz eine Demo-Umgebung vorgeführt, welche es Nutzer*innen erlaubte mittels vollständig verkörperter Avatare zu kommunizieren, deren non-verbales Verhalten allein mithilfe der eingehenden Sprache generiert wurde.

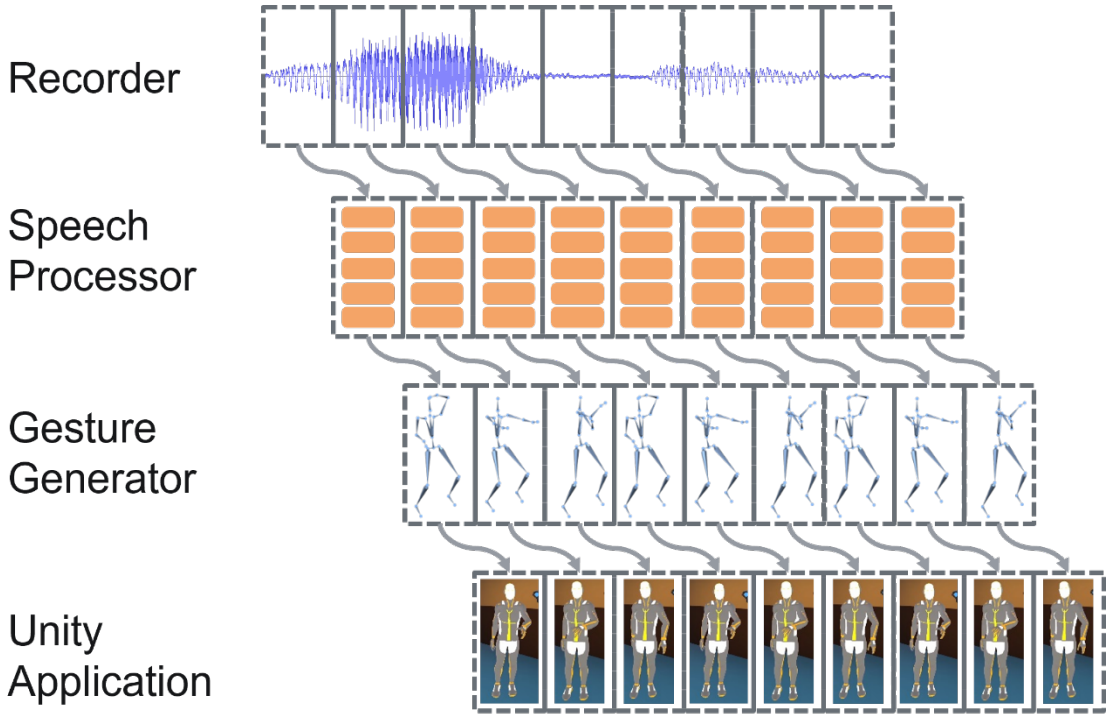

Abbildung 2: Darstellung der inkrementellen Gestengenerierung. Sprache wird stückweise aufgenommen und verarbeitet. Gesten werden generiert (mit Hilfe des ZeroEEGS-Modell) und in Unity 3D zu einer vollständigen Bewegung zusammengefügt.

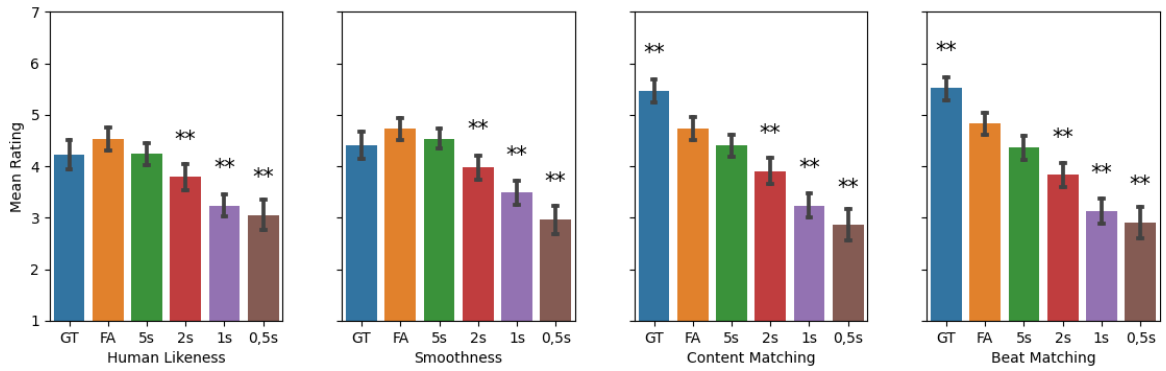

Abbildung 3: Ergebnisse der Evaluationsstudie zur inkrementellen Gestengenerierung durch Nutzerbewertungen verschiedener Versionen unserer Gestengenerierung, basierend auf Spracheingabe-Inkrementen von 5s, 2s, 1s und 0.5s sowie unverändertem ZeroEGGS („FA“) und Ground Truth („GT“).

Verbesserung der sprachbasierten Online-Gestensynthese (AP 5.3): Aufgrund der Ergebnisse aus der Studie wurde die zugrundeliegende Architektur des Modells zur Gestensynthese überarbeitet, um den trade-off aus Latenz und Qualität zu lösen. Idealerweise würden die resultierenden Bewegungen anstelle von Teilsequenzen in einzelnen Frames generiert werden und somit auch nur einzelne Audio-Frames als Eingabe benötigten, welches bei 60 Frames pro Sekunde etwa 17ms an Audioinformationen entspricht. Im Vergleich dazu benötigt das vorherige System mindestens 500ms,

wobei sich die Reduktion der benötigten Eingabesequenz unmittelbar auf die Latenz des Gesamtsystems auswirkte. Auf der anderen Seite sind für die Generierung einer kohärenten und mit der Sprache konsistenten Geste Informationen über die Zeit hinweg notwendig (Bewegungsverläufe, verbaler Kontext).

Um dieses Ziel zu erreichen, wurde die ZeroEEGs-Architektur erweitert. Dazu wurde ein zusätzlicher, latenter und persistenter „Kontext“-Vektor eingeführt, welcher Informationen über mehrere Frames hinweg speichern kann. Dadurch ist das System in der Lage, Langzeitinformationen über die derzeitige Bewegungssequenz zu speichern und so die Kohärenz und Qualität der Gesten zu erhöhen, aber selber nur einzelne Frames als Eingabe/Ausgabe zu verarbeiten. Abbildung 4 zeigt den Aufbau des entwickelten Modells schematisch. Das resultierende System wurde implementiert und evaluiert und erwies sich als qualitativ ebenbürtig gegenüber dem Ursprungssystem von Ghorbani et al. [2] mit Zugriff auf die vollständigen Audioinformationen (siehe Abbildung 5). Es gelang uns also, ein State-of-the-art Modell so zu adaptieren, dass es echtzeitfähig ist, ohne die Qualität der Gesten negativ zu beeinflussen. Diese Ergebnisse wurden in Form eines Poster-Papers auf der IEEE AIxVR Konferenz 2024 veröffentlicht und dort mit dem Best Poster Award ausgezeichnet [8].

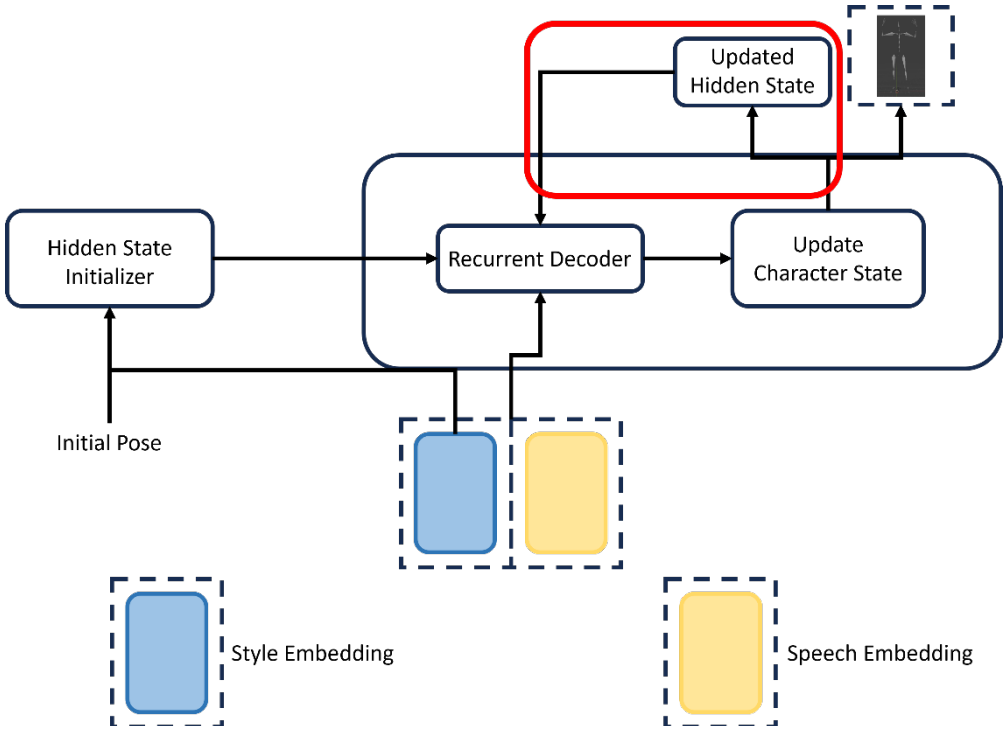

Abbildung 4: Darstellung der sequenziellen Generierung einzelner Posen. Sprache und Stil werden zusammengeführt und als Eingabe verwendet. Die nächste Pose wird auf Basis der Eingabe, sowie der akkumulierten Informationen über vorangehende Posen generiert. Die resultierende Bewegungssequenz ist flüssig und kohärent.

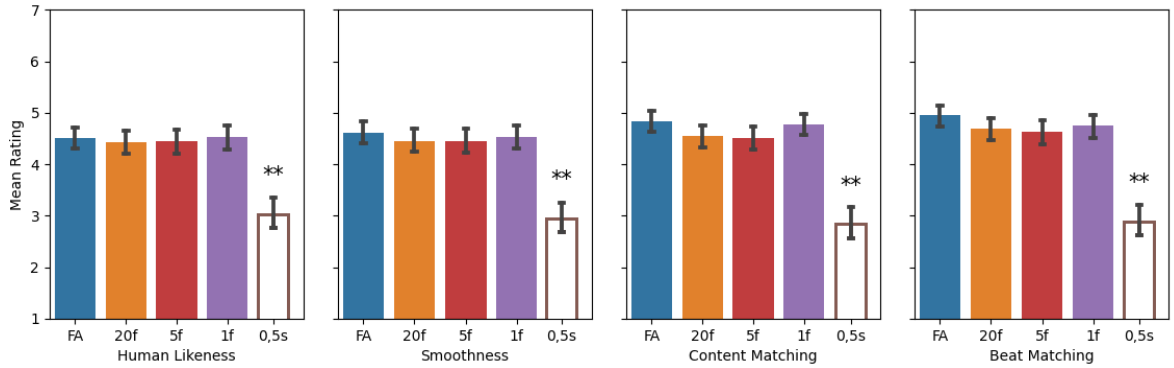

Abbildung 5: Ergebnisse der Nutzerstudie. Bewertungen verschiedener Versionen unserer Gestengenerierung, basierend auf Inkrementen von 20, 5 und 1 Frame bei 60 Frames pro Sekunde, sowie unverändertem ZeroEGGS (FA) und der Variante mit der geringsten Latenz des vorigen Systems (0.5s). Die Länge der Inkremente ist direkt proportional zur Latenz des Gesamtsystems (kleiner ist besser).

Situationsspezifische Verhaltenssynthese (AP 5.4): Ein weiterer Vorteil des Ansatzes basierend auf der erweiterten ZeroEGGs-Architektur ist, dass das Zielt einer situationsspezifischen Verhaltensanalyse mit adressiert wird. In einer dyadischen Interaktion gibt es verschiedene Situationen bzw. Kontexte, die Art, Umfang und Angemessenheit des Verhaltens der Interaktanden beeinflussen. Diese Informationen sind damit auch wichtig, um derartiges Verhalten für Avatare synthetisch zu erzeugen. Das von uns verwendete Modell zur Generierung von Gesten lässt sich mithilfe von Stil-Vektoren parametrisieren, welches in deutlich verändertem, aber trotzdem passenden Verhalten resultiert. Beispiele sind emotionale Stile wie „Happy“, „Sad“, oder „Angry“, aber auch kontext-spezifische Stile wie „Agreement“ und „Disagreement“. Da ein einzelner, unveränderlicher Stil im Laufe einer längeren Interaktion nicht angemessen wäre, wurde ein System zur dynamischen Adaption des Stils entwickelt. Stilwechsel sind technisch dabei extrem kurzfristig (innerhalb eines Frames) möglich und wirken sich direkt auf die folgende Pose aus.

Um angemessene Stile zu erkennen, wurden verschiedene Systeme zur Emotionserkennung aus Sprache adaptiert und getestet [5, 6]. So lassen sich Emotionen in Echtzeit aus der eingehenden Sprache ableiten, um den jeweils passenden Stil anzuwenden. Ein Modell wurde erweitert und auf den Trainingsdaten des ZeroEGGS-Modells nachtrainiert, um eine Erkennung der Modell-spezifischen Stile zu ermöglichen. Erste Ergebnisse sind vielversprechend, aber in echter Interaktion und auf Basis von Echtzeit-Mikrofonsignalen ist die Genauigkeit der Erkennung bisher nicht zufriedenstellend und bedarf weiterer Evaluation.

Integration und Evaluation (AP 5.5): Das im Zuge der letzten Veröffentlichung [8] evaluierte System wurde den Projektpartnern zur Verfügung gestellt und in die HiAvA-Umgebung integriert. Zusammen mit dem Projektpartner Universität Würzburg wurde eine umfangreiche Teilnehmerstudie durchgeführt, um die Qualität der synthetisch generierten Gesten mit real getrackten Gesten in einer Alltagssituation zu vergleichen. Die Teilnehmer*innen sollten in jeweils drei Interaktionen miteinander interagieren. Ein*e Teilnehmer*in wurde hierbei vollständig getrackt, während der*die andere Teilnehmer*in wechselnde Bewegungsmodalitäten verwendete (Tracking, Generierung, Idle). So sollte untersucht werden, wie unser System eine voll verkörperte Interaktion mit Avataren in Social XR beeinflusst und wie diese Beeinflussungen in Relation zu vollständigem Bewegungstracking sowie einem Szenario stehen, indem fehlendes Tracking nicht supplementiert wird und stattdessen nur nonverbales „Idle“-Verhalten gezeigt wird.

Die Auswirkungen der supplementierten Gesten wurden zudem im Kontext verschiedener Endgeräte untersucht. So sollte jeweils die Hälfte der Teilnehmer*innen über ein VR-Headset und die andere Hälfte über einen regulären Bildschirm interagieren, um Einflüsse von Immersion auf die Bewertung des Verhaltens zu testen. Die Effekte und Qualität wurden jeweils durch Nutzer*innenbewertungen mittels etablierter Skalen und Fragebögen erhoben (Appearance and Behavior Plausibility Scale (ABP) des Virtual Human Plausibility (VHP) Fragebogens [10] sowie Uncanny Valley Index (UVI) [11 und Igroup Presence Fragebogens (IPQ) [12]). Die Ergebnisse (siehe Abbildung 6) zeigten keine signifikanten Unterschiede in der Plausibilität der angezeigten Bewegungen des Gegenübers (Abb. 6a). Außerdem waren keine signifikanten Unterschiede bezüglich der „Menschenähnlichkeit“ der Bewegungen zu erkennen. Auch die empfundene „räumliche Anwesenheit“ wurde in unserem Experiment nicht signifikant durch die Animationsquelle der Körperbewegungen des Gegenübers beeinflusst. Wir konnten allerdings signifikante Unterschiede zwischen den verschiedenen Endgeräten ausmachen. Eine Veröffentlichung der Studienergebnisse ist in Arbeit.

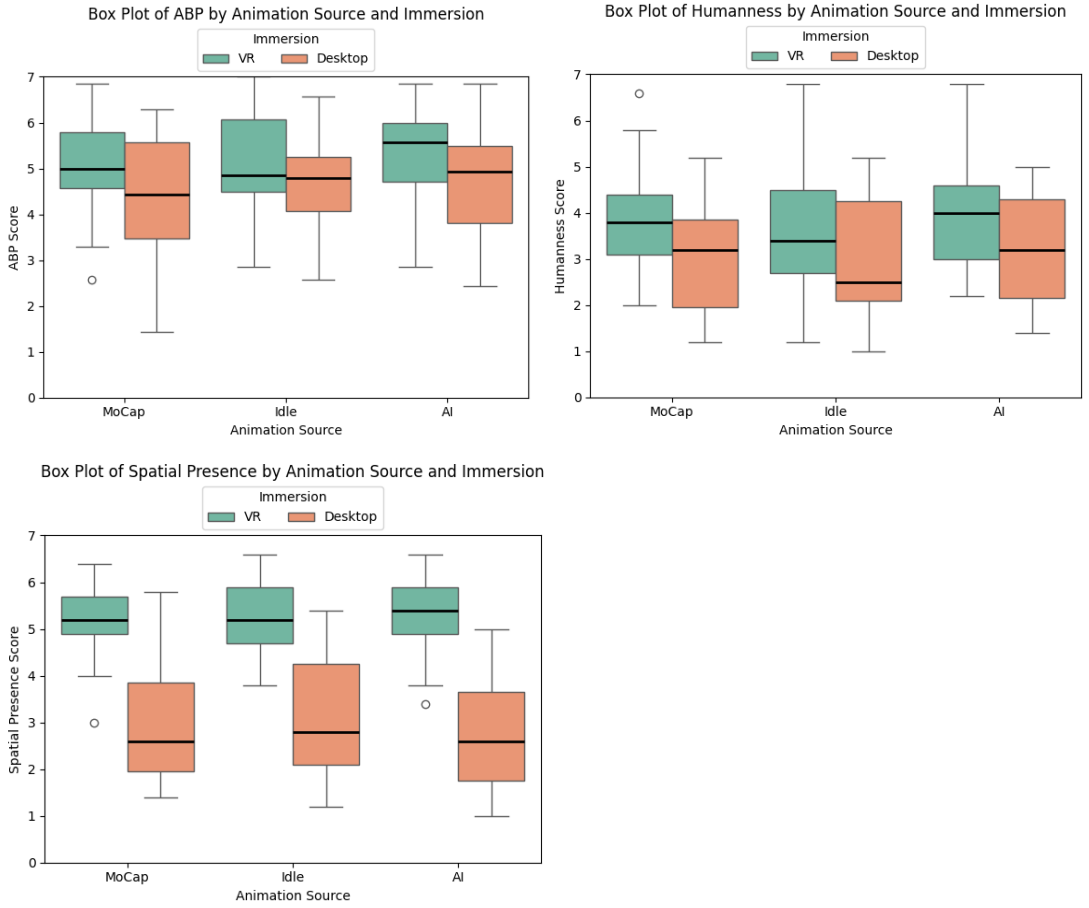

Abbildung 6 a-c: Ergebnisse der Teilnehmerstudie, aufgeteilt nach Animationsquelle und verwendetem Endgerät. a) zeigt die Bewertungen der Plausibilität des erlebten Verhaltens des Gegenübers auf Basis der Appearance and Behavior Plausibility Scale (ABP) des Virtual Human Plausibility (VHP) Fragebogens [10]. b) zeigt die Bewertungen der „Menschenähnlichkeit“ des erlebten Verhaltens des Gegenübers auf Basis des Uncanny Valley Index (UVI) [11]. c) zeigt die Bewertungen des Gefühls der „Räumlichen Anwesenheit“ während der Interaktion auf Basis des Igroup Presence Fragebogens (IPQ) [12].

AP 6: KI-unterstützte Verhaltensmodelle für Multiparty-Konversationen

Aufbauend auf den substanziellen und umfangreichen Arbeiten in AP 5 wurde in AP6 untersucht, in wie weit sich die Verhaltensmodelle durch Erkennung und Anpassung an den situativen Kontext in Social XR-Interaktionen verbessern und verallgemeinern lassen. Dazu wurden exemplarisch der Einbezug von dialogischen Rollen (Sprecher/Zuhörer) sowie die Adressierung einzelner Gesprächspartner mittels gerichteten Blickverhaltens betrachtet. Insgesamt sind die Arbeiten in diesem AP eher explorativ und exemplarisch, da das vollständige HiAwA-Setup zu Umsetzung von längeren Interaktionen zwischen zwei oder sogar mehreren Nutzenden nicht bis zum Ende des Teilvorhabens zur Verfügung stand.

Interaktionsanalyse und Online-Erkennung von Dialogrollen (AP 6.1): Um die Dialogrolle einzelner Nutzenden erkennen und vorhersagen zu können, wurde ein System zur Erkennung von Sprachaktivität [4] integriert, ebenso wie ein System zur Vorhersage zukünftiger Sprecheraktivität [3] in dyadischer Interaktion. Abhängig von der erkannten Dialogrolle werden verschiedene Modelle angewandt, um jeweils passendes Sprecher- oder Zuhörerverhalten für die Avatare zu generieren. Die Generierung von synthetischem Verhalten von Sprecher-Avataren wurde bereits umfassend in AP 5 bearbeitet. Als ein potenzielles Modell zur Generierung von Zuhörerverhalten wurde ein Ansatz basierend auf Posture-Mimicry (Verhaltensspiegelung) entwickelt. Das System bedient sich der Bewegungen des Sprechers und wendet diese in abgeschwächter Form und mit einer Verzögerung auf den Zuhörer an. Die Auswirkungen solcher Avatare auf einen dritten Beobachter, hinsichtlich der empfundenen Beziehung zwischen den Interaktanden, wurden mithilfe einer Online-Studie erforscht.

Ein weiteres System wurde entwickelt, das automatisch Verhalten in der Zuhörerrolle generiert um aufmerksames Zuhören („listening behavior“) zu signalisieren. Neben separat gesteuertem Blickverhalten und sonstigem vordefinierten „idle listening“-Animationen, wurde ein Modell implementiert, das Nickbewegungen für eine Gruppe an Zuhörern generiert, wenn der*die Nutzer*in redet. Die Auswirkungen von negativem und positivem Feedback während eines Vortragsszenarios wurden mithilfe dieses Systems in einer VR-Studie evaluiert und zeigten gute Ergebnisse.

Adressierung und Orientierung (AP 6.2): Zur Generierung realistischer Augenbewegung wurde ein bestehendes prozedurales Modell integriert, welches automatisch Mikro-Sakkaden und Blinzelverhalten ausführt und so zu einem realistischen Aussehen der Avatare führt. Auch eine realistische Animation beim Wechsel zwischen Fokuspunkten wird durch das System ermöglicht. Aufbauend darauf wurde eine automatische Parametrisierung des Modells mittels unseres generativen Systems ermöglicht. So kann beispielsweise basierend auf der Sprecheraktivität des Nutzers der Blick auf das Gegenüber gerichtet werden und abseits davon zufällig zwischen Objekten im Raum wechseln. Ebenso kann in einer Multi-Party-Interaktion gezielt der aktuelle Sprecher im Raum fokussiert werden. Die Erweiterung und Anwendung des Modells auf Multi-Party Szenarien wurde allerdings nicht vollständig abgeschlossen.

Integration in Gesamt-Plattform und Client-Anwendungen (AP 6.3): Die Systeme zur Erkennung der Dialogrollen sind in die Gesamt-Plattform integriert worden. Die Adressierung und Orientierung in Multi-Party Interaktion wurde nicht in die Gesamt-Plattform integriert, steht aber als Komponenten in entsprechender Modularisierung zur Verfügung.

Verwertbarkeit der Ergebnisse

Die innerhalb des Teilvorhabens erlangten wissenschaftlichen und technischen Ergebnisse haben ein großes Potenzial hinsichtlich ihrer Verwertbarkeit. Wissenschaftlich sind die technischen und empirischen Ergebnisse des Teilvorhabens in den beiden Publikationen [7] und [8] veröffentlich. Eine weitere, größere Publikationen der finalen Evaluationsstudie ist gemeinsam mit den anderen Projektpartnern in Vorbereitung. In der Lehre wurden und werden aktuell zahlreiche Bachelor- und Masterarbeiten unter Verwendung der Projektergebnisse durchgeführt und haben umgekehrt auch zu den Ergebnissen beigetragen (z.B. die Implementierung des Nick-Verhaltens von Zuhörer-Avataren).

Technisch liegen die Ergebnisse allesamt in Form von implementierten und modularisierten Software-Paketen vor, die in die HiAwA-Plattform bereits integriert sind oder ohne großen Aufwand integriert werden können. Dies gilt auch für andere Social XR-Umgebungen, da in großem Umfang auf Standard-Schnittstellen und -Technologien zurückgegriffen wurde. So wird der Output der Gestensynthese bspw. als Stream über eine Netzwerkschnittstelle bereits gestellt und kann leicht angebunden werden. So werden diese Ergebnisse an der Universität Bielefeld bereits in Folgeprojekten (z.B. zur VR-basierten Diagnose von erworbenen Kommunikationsstörungen) weiterverwendet.

Innerhalb der Projektlaufzeit haben sich signifikante Fortschritte auf dem Gebiet der ML-basierten Verhaltensergebnisse ergeben. Diese sind aber immer noch weitestgehend auf die Offline-Analyse fokussiert und konzentrieren sich daher auf Qualität anstelle von Latenz. Der Fortschritt in Richtung interaktiver VR-Anwendungen bei den kommerziellen Umgebungen (z.B. NVIDIA Holodeck für VR-basierte Kollaboration), bei den frei verfügbaren Umgebungen (z.B. VR2Gather) sowie bei der verfügbaren Hardware macht die Relevanz der erzielten Ergebnisse dabei noch größer.

Referenzen

[1] Gaurav Menghani: “Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better”. In: CoRR abs/2106.08962 (2021). arXiv: 2106.08962. url: https://arxiv.org/abs/2106.08962.

[2] Ghorbani, S., Ferstl, Y., Holden, D., Troje, N., F., & Carbonneau, M-A. (2022). ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech (arXiv:2209.07556). http://arxiv.org/abs/2209.07556 arXiv:2209.07556.

[3] Ekstedt, E. & Skantze, G. (2022). Voice Activity Projection: Self-supervised Learning of Turn-taking Events (arXiv:2205.09812). http://arxiv.org/abs/2205.09812.

[4] Silero Team. (2021). Silero-VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. https://github.com/snakers4/silero-vad.

[5] Ma, Z., Zheng, Z., Ye, J., Gao, Z., Zhang, S., & Chen, X. (2023). Emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation (arXiv:2312.15185). http:://arxiv.org/abs/2312.15185

[6] Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10745–10759.

[7] Krome, N., & Kopp, S. (2023). Towards Real-time Co-speech Gesture Generation in Online Interaction in Social XR. *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, 1–8.

[8] Krome, N., & Kopp, S. (2024). Minimal Latency Speech-Driven Gesture Generation for Continuous Interaction in Social XR. *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*, 236–240.
[9] https://assetstore.unity.com/packages/tools/animation/realistic-eye-movements-29168
[10] D. Mal, E. Wolf, N. Döllinger, M. Botsch, C. Wienrich, and M. E. Latoschik. Virtual human coherence and plausibility – towards a validated scale. In 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), pp. 788–789, 2022. doi: 10.1109/VRW55335.2022.00245 3, 6
[11] C.-C. Ho and K. F. MacDorman. Measuring the uncanny valley effect: Refinements to indices for perceived humanness, attractiveness, and eeriness. International Journal of Social Robotics, 9:129–139, 2017. doi: 10.1007/s12369-016-0380-9 6
[12] T. Schubert, F. Friedmann, and H. Regenbrecht. The experience of presence: Factor analytic insights. Presence: Teleoperators & Virtual Environments, 10(3):266–281, 2001. doi: 10.1162/105474601300343603 6