Schlussbericht zum Verwendungsnachweis

Verbundprojekt:	**„VoluProf**: Photo und Audiorealistische*r Volumetrische*r Mixed Reality Professor*in für omnipräsente und nutzeroptimierte Lehre“
Teilprojekt:	„Sprachsynthese zur Avatarvertonung und Spracherkennung zur Nutzer*inneninteraktion“
Förderprogramm:	„VR/AR: digitale Gesellschaft“
Förderkennzeichen:	16SV8708
Durchgeführt von:	Aristech GmbH Galileistraße 1-3 69115 Heidelberg
Projektlaufzeit:	01.09.2021 – 28.02.2025
Projektleiter:	Benjamin Körner

Bundesministerium
für Forschung, Technologie
und Raumfahrt

Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung, Technologie und Raumfahrt (BMBFTR) unter dem Förderkennzeichen **16SV8708** gefördert.

Autor:

Benjamin Körner, Aristech GmbH

Inhaltsverzeichnis

1. Kurzdarstellung

1.1. Ursprüngliche Aufgabenstellung

Ziel war die realistische Stimme des Avatars aus einem von dem/r Lehrenden bereitgestellten Text zu generieren. Dies sollte mit möglichst wenig Daten dieser Person machbar sein, sodass auf umfangreiche Sprachaufnahmen im Tonstudio verzichtet werden kann. Weiterhin war geplant zu untersuchen, inwiefern Sprachaufnahmen aus bisherigen Quellen hierzu genutzt werden können, z.B. aus schon aufgezeichneten Vorlesungen oder Vorträgen. Basierend auf einem maschinell trainierten Grundmodell der deutschen Sprache und den sprachlichen Merkmalen, welche anhand einzelner Aufnahmen extrahiert werden, sollte dazu aus textuellem Input "live" oder „offline" die Stimme für den Avatar synthetisiert werden. Für die Lippensynchronität des animierten Avatars war angedacht Ausspracheinformationen für den Inhalt der Vorlesung bereitzustellen werden. Diese könnten genutzt werden, um die Lippenbewegung an die tatsächlich gesprochenen Laute anzupassen.

Das im Projekt entwickelte System sollte weiterhin hinsichtlich der Qualität der technischen Nutzer*innenerfahrung optimiert und evaluiert werden. Dazu sollten psychophysikalische Testverfahren für die Qualitätsbeurteilung von MR-Umgebungen entworfen und validiert werden. Auf Basis der entwickelten Testverfahren war geplant konstituierende Aspekte wahrgenommener Qualität in MR (etwa visueller Detailgrad, räumliches Audio, VR-Krankheit oder Immersionserfahrung) zu identifizieren. Der Einfluss und die Interaktion der Systemparameter (etwa Latenz, Kanalbreite oder Betrachtungsabstand) auf die wahrgenommene Qualität sollte zudem quantifiziert und zur Systemoptimierung genutzt werden. Die in kontinuierlichen Tests gewonnenen Daten (qualitätsannotiertes Videomaterial) sollten dazu genutzt werden, um modell- und datengetriebene Methoden zur automatischen Qualitätsschätzung zu entwickeln. Des Weiteren sollte auch die Authentizität der generierten Stimme des Avatars durch subjektive Testverfahren untersucht werden.

Zusammengefasst ergaben sich folgende Teilziele für die Aristech GmbH in diesem Projekt:

- Natürliche und schnelle Synthese der Originalstimme der/s Lehrenden
- Integration des volumetrischen Abbilds und der nachgebildeten Stimme zur Erstellung eines volumetrischen Avatars der/s Lehrenden
- Möglichkeit der sprach- und gestenbasierten Interaktion mit dem volumetrischen Abbild der/s Lehrenden derart, dass Nach- und Rückfragen ermöglicht und so bildungsrelevante Dialoge und Feedback an Nutzer*innen unterstützt werden

- Technische Auswertung der Nutzer*innenerfahrung im Hinblick auf die wahrgenommene audiovisuelle Qualität

1.2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Das Vorhaben VoluProf entstand vor dem Hintergrund der durch die COVID-19-Pandemie bedingten massiven Umstellung der Hochschullehre auf digitale Formate. Während des Lockdowns in den Jahren 2020 und 2021 wurde der Lehrbetrieb überwiegend über Videokonferenzsysteme und statische Video- und Audioskripte aufrechterhalten. Diese Formen der Online-Lehre ermöglichten zwar die notwendige Grundversorgung, offenbarten jedoch zugleich deutliche Defizite: Eine geringe Interaktivität, das Fehlen sozialer Resonanzräume sowie die eingeschränkte Immersivität führten zu niedrigeren Lernerfolgen und wurden von Lehrenden wie Lernenden als unbefriedigend bewertet.

Aus dieser Problemstellung ergab sich die förderlogische Ausgangslage: Die Notwendigkeit, digitale Lehr- und Lernmedien qualitativ weiterzuentwickeln, um didaktisch wirksame, interaktive und nachhaltige Formen der Online-Lehre zu ermöglichen.

Das Projekt knüpfte an mehrere günstige Voraussetzungen an:

- Technologische Reife: Durch Fortschritte im Bereich der Mixed-Reality-Hardware sowie in der volumetrischen Erfassung und Animation von Personen standen bereits Basistechnologien zur Verfügung, die für den Einsatz in immersiven Lernumgebungen adaptiert werden konnten.
- Entwicklungen in der Sprachsynthese: Neue Deep-Learning-Ansätze erlaubten die realistische Generierung von Stimmen aus Textvorgaben, womit ein zentrales Element für glaubwürdige Avatare verfügbar war.
- Gesellschaftlicher Bedarf: Die Erfahrungen aus den „Pandemie-Semestern" machten die Relevanz innovativer, digitaler Lehrformate evident und führten zu einer breiten Akzeptanz und Offenheit gegenüber neuen Lehrtechnologien.

Die Projektumsetzung erfolgte im Einklang mit den Zielen der Förderung:

- Stärkung der digitalen Hochschullehre durch Entwicklung neuer Methoden, die über reine Wissensvermittlung hinaus interaktive und immersive Lernprozesse unterstützen,
- Förderung technologischer Innovation durch die Verknüpfung von Mixed Reality, volumetrischer Avatar-Technologie und KI-basierter Sprachsynthese in einem nutzungsorientierten Gesamtsystem,

- Integration ethischer, rechtlicher und sozialer Implikationen (ELSI-Aspekte), um die Technologieentwicklung verantwortungsvoll zu gestalten und Akzeptanzhürden frühzeitig zu adressieren.

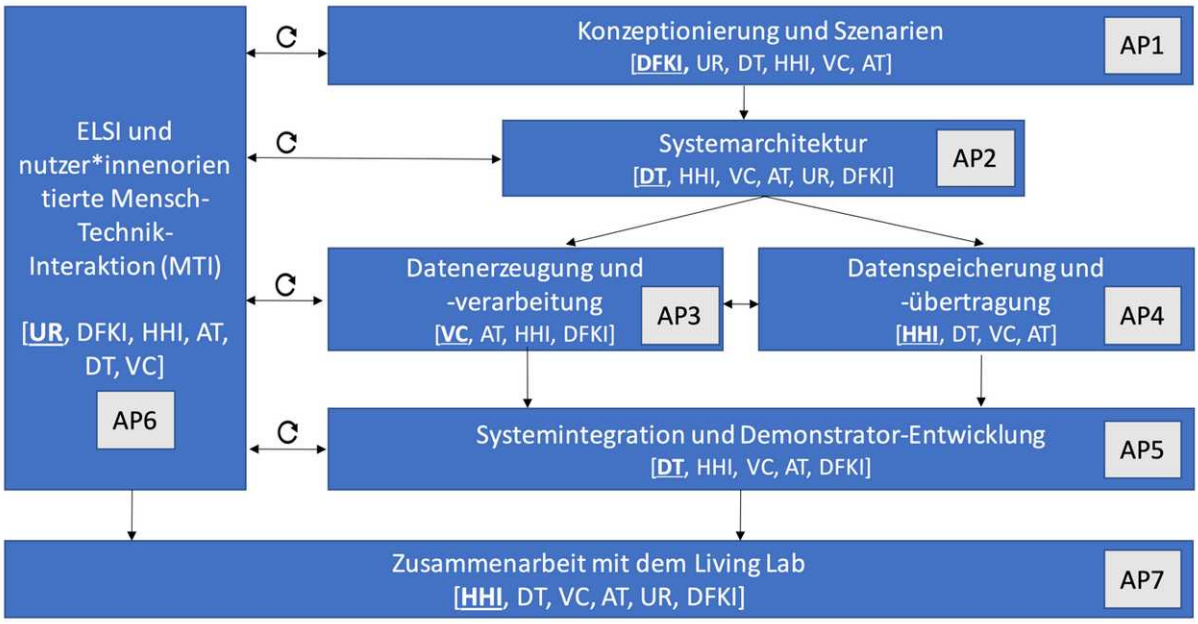
Damit konnte das Vorhaben auf einer Kombination aus gesellschaftlicher Relevanz, technologischer Machbarkeit und hoher Innovationshöhe aufsetzen und trug zugleich zur Erreichung der förderpolitischen Zielsetzungen im Bereich der digitalen Bildung bei.

1.3. Planung und Ablauf des Vorhabens

Das Konsortium bestand aus folgenden Partnern:

- Fraunhofer HHI,Video Coding and Analytics (VCA) & Vision and Imaging Technologies (VIT)
- Deutsche Telekom AG, Innovation Laboratories
- Volucap GmbH
- Aristech GmbH
- Universität Rostock, Philosophische Fakultät / Institut für Medienforschung & Theologische Fakultät /Technikethik
- Deutsches Forschungszentrum für Künstliche Intelligenz, Educational Technology Lab

Die Arbeiten wurden in sieben Arbeitspakete aufgeteilt, welche sich wiederum in Unterarbeitspakete gliederten:



Für jedes Arbeitspaket gab es einen hauptverantwortlichen Partner welcher die Zusammenarbeit der weiteren beteiligten Partner koordinierte.

Aristech war dabei in folgenden Arbeitspaketen mit insgesamt 42 Personenmonaten involviert.

Arbeitspaket	**Leitung**	AT
AP1: Konzeptionierung und Szenarien	**DFKI**	3
AP2: Systemarchitektur	**DT**	5
AP3: Datenerzeugung und -verarbeitung	**VC**	17
AP4: Datenspeicherung und -übertragung	**HHI**	4
AP5: Systemintegration und Demonstrator-Entwicklung	**DT**	7
AP6: Ethische, rechtliche und soziale Implikationen (ELSI) und nutzer*innenorientierte Mensch-Technik-Interaktion (MTI)	**UR**	5
AP7: Zusammenarbeit mit dem Living Lab	**HHI**	1
Summen PM		**42**

Eine geplante anfängliche Nutzerbefragung der Zielgruppe hatte sich coronabedingt verzögert. Da die Entwicklungen des Demonstrators auf dem hieraus abgeleiteten Zielbild aufbauten, haben sich entsprechend auch unsere Entwicklungstätigkeiten nach hinten verschoben.

Weiterhin verzögerte sich das Gesamtprojekt, da die ersten volumetrischen Aufnahmen der Professorin nicht für Animationen geeignet waren und diese wiederholt werden mussten.

Durch die Laufzeitverlängerung wurde der Zeitplan jedoch wieder aufgeholt.

Abgesehen davon wurde jedoch grundsätzlich wurde an der ursprünglichen Arbeits-, Zeit- und Kostenplanung festgehalten.

1.4. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

a) Sprachsynthese

Bis in die frühen 2010er Jahre hinein war die Unit-Selection-Synthese der vorherrschende Ansatz im Bereich der Sprachsynthese. Dabei wird Sprache aus einer umfangreichen Datenbank vorab aufgezeichneter Sprachsegmente (Einheiten) zusammengesetzt. Die Auswahl der passenden Einheiten erfolgt dynamisch anhand linguistischer und akustischer Merkmale des Eingabetextes. Mit ausreichend großen und domänenspezifisch passenden Sprachkorpora konnte so eine hohe Natürlichkeit erzielt werden, die in kontrollierten

Szenarien teilweise kaum vom menschlichen Sprachsignal unterscheidbar war. Allerdings war die Flexibilität dieser Verfahren begrenzt: Die erzielte Qualität hing stark von der Abdeckung der verwendeten Sprachdaten ab, wodurch Übertragbarkeit auf neue Anwendungsbereiche oder variierende Stimmlagen nur eingeschränkt möglich war.

Ab Mitte der 2010er Jahre kam es durch den Einsatz tief neuronaler Netze zu einem Paradigmenwechsel in der Sprachsynthese. End-to-End-Architekturen setzten sich durch, die den Transformationsprozess von Text zu Sprachsignal vollständig datengetrieben abbilden. Typischerweise erfolgt die Generierung in zwei Stufen: Zunächst wird der Text in ein akustisches Zwischenformat, meist in Form eines Melspektrogramms, überführt (z. B. Tacotron, Wang et al. 2017). Anschließend wird das Spektrogramm mit Hilfe eines neuronalen Vocoders in ein hörbares Audiosignal umgewandelt. Hierbei etablierten sich insbesondere Modelle wie etwa WaveNet (van den Oord et al. 2016) die eine ernorme Steigerung der wahrgenommenen Natürlichkeit und Flüssigkeit der erzeugten Sprache ermöglichten.

Darüber hinaus erlauben Deep-Learning-basierte Methoden durch den Einsatz zusätzlicher Trainings-Embeddings eine gezielte Modellierung von Prosodie, Sprachmelodie und Stimmcharakteristika. Damit wurde es erstmals möglich, synthetische Sprache flexibel hinsichtlich Ausdrucksweise, Intonation und Sprecheridentität zu steuern und in einer Qualität zu generieren, die weit über die Grenzen klassischer Unit-Selection-Methoden hinausgeht.

b) Spracherkennung

Die Spracherkennung hat sich seit Beginn der 2010er Jahre stark weiterentwickelt und ist insbesondere durch die Integration in weit verbreitete digitale Assistenten wie Siri (Apple), Cortana (Microsoft), Alexa (Amazon), Bixby (Samsung) und Google Now in den Alltag von Endanwenderinnen und -anwendern vorgedrungen. Diese Systeme machten die Sprachsteuerung von Geräten und Anwendungen massentauglich und etablierten Sprachschnittstellen als relevanten Interaktionskanal.

Allerdings waren diese Dienste bis dato in ihrer Funktionalität noch häufig eingeschränkt: Die Sprachsteuerung erfolgte meist in Form vordefinierter Kommandos, die nicht flexibel an spezifische Anwendungsdomänen oder Benutzerbedürfnisse angepasst werden konnten. Für den Einsatz in spezialisierten Kontexten (z. B. Industrie, Medizin, Jura oder sicherheitskritische Umgebungen) stellte dies eine wesentliche Hürde dar.

Technologisch konnten in den Jahren zuvor jedoch erhebliche Fortschritte erzielt werden. Während noch Anfang der 2010er Jahre Wortfehlerraten (Word Error Rate, WER) von 20–25 % als Stand der Technik galten, wurden durch den Einsatz Deep Neural Networks (DNN) , insbesondere Recurrent neural networks (RNNs), Long Short-Term Memory (LSTM)-Modelle sowie ab etwa 2017 mittels End-to-End-Architekturen deutliche Verbesserungen erzielt. So konnten unter realistischen Bedingungen WER-Werte von 10–15 % erreicht werden.

Darüber hinaus zeigte sich, dass durch eine kontextsensitive Einschränkung des Suchraums (z. B. anwendungs- oder domänenspezifische Vokabulare und Sprachmodelle) noch bessere Ergebnisse möglich sind. Parallel dazu etablierten sich Transformer-basierte Modelle die eine weitere Reduktion der Fehlerraten ermöglichten und den Grundstein für die damals beginnende Ära großskaliger, vortrainierter Sprachmodelle legten.

Insgesamt war der wissenschaftlich-technische Stand der Spracherkennung zu Beginn des Projekts VoluProf durch eine deutliche Steigerung der Erkennungsqualität, die breite Marktdurchdringung in Consumer-Produkten sowie den Übergang von klassischen HMM- und GMM-basierten Verfahren hin zu End-to-End-Deep-Learning-Architekturen geprägt.

1.5. Zusammenarbeit mit anderen Stellen

Neben der Zusammenarbeit mit den Konsortialpartnern sowie deren Mitarbeitenden wurde für die Aufnahmen zum Training von TTS-Stimmen vor allem mit zwei Dozentinnen der Universität Rostock zusammengearbeitet. Dafür fanden im Februar 2023 in den Studios des Projektpartners VoluCap entsprechende Aufnahmen zum Sammeln der für das Training notwendigen Audiodaten statt. Da die gleichen Dozentinnen vom Partner VoluCap auch visuell für die spätere Erzeugung der VoluProfs (hier im Sinne eines volumetrischen Avatars) aufgenommen wurden bot sich ein anschließender Termin für die Audioaufnahmen an. Die beiden Dozentinnen waren neben der Stimmerstellung generell auch für die visuelle „Digitalisierung“ vorgesehen.

2. Eingehende Darstellung

2.1. Erzieltes Ergebnis

Ein zentrales Ziel im Teilvorhaben der Aristech GmbH bestand darin, die Erzeugung realistischer synthetischer Stimmen für Avatare generell zu ermöglichen und diese als Weiterentwicklung zum damaligen Ist-Stand erheblich zu vereinfachen. Ausgangspunkt war die Situation, dass vor Projektbeginn die Erstellung einer vollwertigen Text-to-Speech-(TTS)-Stimme für eine bestimmte Sprecherin oder einen bestimmten Sprecher mit sehr hohem Aufwand verbunden war: Für eine qualitativ hochwertige Synthese waren üblicherweise ca. 40 Stunden Studioaufnahmen notwendig. Diese mussten unter kontrollierten Bedingungen produziert werden, um Rauschen und Störungen zu vermeiden. Ein solcher Aufwand war für den vorgesehenen praktischen Einsatz im Hochschul- und Schulkontext weder realistisch noch skalierbar.

Das Projektziel bestand daher darin, TTS-Modelle auch mit sehr kleinen Datenmengen trainieren zu können, sodass auf aufwendige Sprachaufnahmen im Tonstudio verzichtet werden kann. Dazu wurden zwei Ansätze verfolgt:

- Entwicklung und Adaption von Verfahren, die es ermöglichen, auf Basis von nur wenigen Minuten Audiodaten eine vollwertige synthetische Stimme zu erzeugen.
- Nutzung bestehender Sprachaufzeichnungen und Untersuchung, ob bereits vorhandene Audiomaterialien – etwa aufgezeichnete Vorlesungen oder Vorträge – für das Training genutzt werden können, auch wenn diese qualitativ nicht die Bedingungen professioneller Studioaufnahmen erfüllen.

Beide Ansätze konnten im Rahmen der Projektlaufzeit erfolgreich umgesetzt werden. Konkret zeigte sich, dass bereits mit 1–2 Minuten Sprachmaterial ausreichender Qualität eine synthetische Stimme mit hoher Natürlichkeit und Verständlichkeit trainiert werden konnte. Dies markiert einen entscheidenden technologischen Fortschritt, da der Erstellungsaufwand für personalisierte TTS-Stimmen damit um einen erheblichen Faktor werden konnte.

Die Ergebnisse sind in mehrfacher Hinsicht verwertbar:

- Praktische Anwendbarkeit: Lehrende müssen künftig keine umfangreichen Tonstudioaufnahmen mehr absolvieren, sondern können mit minimalem Aufwand eine eigene digitale Stimme für den Avatar bereitstellen.
- Skalierbarkeit: Durch die massive Reduktion der Datenanforderungen wird die Technologie auch für größere Lehrkontexte realistisch, in denen zahlreiche Lehrende digitalisiert werden sollen.

- Flexibilität: Die Möglichkeit, auf vorhandene Audioquellen zurückzugreifen, eröffnet eine breite Anwendbarkeit auch über den Bildungsbereich hinaus, z. B. bei der Digitalisierung von Rednern in Kultur oder Wirtschaft.
- Wettbewerbsvorteil für Aristech: Die Fähigkeit, hochwertige TTS-Stimmen mit minimalen Trainingsdaten bereitzustellen, verschafft dem Unternehmen einen technologischen Vorsprung in einem stark wachsenden Marktsegment.

Mit diesen Fortschritten konnte ein wesentliches Projektziel erreicht werden: die praktikable und effiziente Erstellung realitätsnaher synthetischer Stimmen, die den Einsatz volumetrischer Avatare in der Lehre überhaupt erst realistisch und wirtschaftlich macht.

Konkret wurden dazu mit zwei Dozentinnen der Universität Rostock im Februar 2023 in den Studios des Projektpartners VoluCap entsprechende Aufnahmen zum Sammeln der für das Training notwendigen Audiodaten durchgeführt. Dies geschah zum einen im Modus der bis dahin notwendigen Tonstudioaufnahmen. Aufgrund der hohen Qualitätsansprüche konnten dabei aber nur 1-2 Stunden verwertbares Audiomaterial gesammelt werden. Viel zu wenig für das Training auf herkömmlicher Basis.
Daher wurde zusätzlich von einer der beiden Dozentinnen auch Audioaufnahmen eines zuvor gehaltenen Webinars und einer Vorlesung ausgewertet und für das TTS-Training aufbereitet. Hierbei fiel die deutlich schlechtere Audioqualität und Klang der aufgenommenen Stimme auf. Dies war aber in soweit auch wenig verwunderlich da diese Aufnahmen nur als Mitschnitt entstanden und der Fokus dabei gar nicht auf der für das TTS-Training notwendigen Audioqualität lag. Als „Nebenprodukt" ließen sich diese aus den Mitschnitten extrahierten Audiosequenzen trotzdem für eine Anreicherung der notwendigen Trainingsdaten nutzen. Hierbei transportierte sich in diesen Mitschnitten anders klingende Stimme der Dozentin jedoch auch in die spätere TTS-Stimme. Die klangliche Abweichung war hierbei durch die weniger optimale Akustik des Seminarraums in welchem die Veranstaltung und deren Mitschnitt stattfand sowie die eingesetzte Aufnahmehardware, vor allem das Mikrofon, zu erklären. Vor diesem Hintergrund haben wir viele verschiedene Trainings der TTS-Stimmen mit Inputdaten aus diesen Mitschnitten und/oder aus den professionell durchgeführten Studioaufnahmen durchgeführt. Am Projekt beteiligte Dozierende der Universität Rostock haben Samples dieser Stimmvarianten dann Studierenden die die „geklonte" echte Dozentin kennen vorgespielt und deren Feedback eingeholt. Bezüglich der Vergleichbarkeit der TTS-Stimme zur echten Dozentin fiel den befragten Studierenden die stimmliche Abweichung durchaus auf. Dieser Punkt wurde im weiteren Verlauf des Projekts von unserer Seite weiter bearbeitet und noch verbessert. Jedoch lässt sich hierzu sagen, dass der Klang und die Qualität der TTS-Stimmen sehr an der Qualität der Trainingsdaten hängt. In einem bisherigen

Aufnahmesetting mit vielen Stunden an Audiomaterial werden qualitative Unsauberkeiten durch die schiere Menge an Trainingsdaten ausgeglichen. Stützt man das Training auf lediglich 1-2 Minuten aus einem Audiomitschnitt fehlt der korrigierende bzw. „glättende" Effekt einer großen Datenmenge für das Training. Letztendlich fand sich aus beiden Quellen jedoch eine passende Kombination von Audioaufnahmen mit denen die notwendige Trainingsmenge von mehreren Tausend aufgenommenen Sätzen jedoch neu erstellt werden konnte. Hierfür notwendig und hilfreich war dafür auch die im Rahmen des Projekts angeschaffte Hardware bzw. vor allem die GPUs. Diese Variante der Stimme war zuletzt im Rahmen des Demonstrators zu hören.
Da die angeschlossenen Videoaufnahmen der zweiten Dozentin von den Projektpartnern nicht für die Verarbeitung im Hinblick auf die volumetrischen Avatare verwendet werden konnten haben wir die Erstellung dieser TTS-Stimme auch nicht weiterverfolgt. Ziel war es ja Stimme und Aussehen einer real existierenden Lehrperson zu „klonen", nur ein Teil ohne den jeweils anderen wäre dabei nicht zielführend gewesen.

Ein weiteres Ziel war den vom DFKI entwickelten Chatbot auch mittels gesprochener Sprache bedienen zu können. Hierzu haben wir auf bestehende Spracherkennungskomponenten aufgesetzt und das HHI bei der Integration dieser in den Gesamtdemonstrator unterstützt. In diesem Zuge kam es auch zu mehreren Updates bzgl. Usability und Performance. Auch dieses Feature findet sich im aktuellen Demonstrator und erfüllt darin und dabei die im Antrag und Projektkonsortium definierten Anforderungen.

Darüber hinaus und neben diesen beiden von Aristech entwickelten Kernkomponenten wurden in Zusammenarbeit mit unseren Partnern auch die anderen Teilarbeitspakete in den folgenden Arbeitspaketen

- **AP 1: Konzeptionierung und Szenarien**
    - AP 1.1: Technikzentrierte Definition der Anwendungsfälle
        - *Für die Verwendung von Sprachsynthese und Spracherkennung in einem echten Setting, müssen die Anforderungen definiert werden. Im Fokus stehen dabei vor allem Rechenlast und Geschwindigkeit von Erkennung und Generierung von Audiodaten.*
    - AP 1.2: Anwenderzentrierte Definition der Anwendungsfälle
        - *Mitwirkung an der Definition und Anpassung der Anwendungsfälle basierend auf den Erwartungen der Nutzer*innen.*

- AP 1.3: Technische und inhaltliche Anforderungsanalyse basierend auf den definierten Anwendungsfällen
  - *Neben Rechenlast und Geschwindigkeit der Sprachsynthese und Spracherkennung, werden Schnittstellen für das Zusammenspiel von Spracherkennung und semantischer Auswertung und Sprachsynthese und visuellem Output definiert.*

- **AP 2: Systemarchitektur**

  - AP 2.1: Analyse und Verfeinerung der Anforderungen
    - *Im Bereich der Sprachsynthese werden hier Anforderungen an das Transkript der lehrenden Person ermittelt und geprüft, mit welcher Datenmenge die Stimmeigenschaften optimal abgebildet werden können.*
  - AP 2.2: Definition von KPIs (technisch, sozial, rechtlich, ethisch)
    - *Insbesondere im Bereich der Sprachsynthese wird hier einerseits versucht die Stimme der dozierenden Person möglichst natürlich klingen zu lassen, andererseits müssen auch die Gefahren besprochen und berücksichtigt werden, die das klonen einer Stimme mit möglichst wenig Daten mit sich bringen.*
  - AP 2.4: Evaluation neuer sprachgesteuerter Chatbot-Technologien
    - *Neue und potenziell relevante Technologien zur effizienten Generierung und Übertragung sprechergetreuer Sprachsynthese sowie neuer Chatbot-Technologien werden hinsichtlich ihrer Anwendbarkeit evaluiert. Hierbei wird außerdem ein besonderer Fokus auf Ressourcen und Zeiteffizienz der Algorithmen gelegt.*

- **AP 3: Datenerzeugung und -verarbeitung**

  - AP 3.1: Ermittlung der Anforderungen an animierbare interaktive volumetrische Videos
    - *Ermittlung exakter Dauer einzelner Laute um Informationen für ein Lippensynchrones Video zur Verfügung stellen zu können. Diese sorgt zusätzlich für eine natürliche Animation des Avatars.*
  - AP 3.8: Aufbau einer Datenbasis für Training und Evaluation von Sprachsynthese- und Spracherkennungsmodellen
    - *Datenbasis für Training und Evaluation besteht auf Datenpaaren gesprochener Sprache und Transkription.*
  - AP 3.9: Implementation effizienter Trainingsalgorithmen zur Erzeugung von Sprachsynthese- und Spracherkennungsmodellen
    - *Entwicklung von Modellen zur Unterscheidung von Sprecher-Merkmalen sowie Erzeugung und Erkennung von Mel-Spektrogrammen.*

- AP 3.10: Automatisierte Produktion von synthetisiertem Audio
  - *Implementierung einer Schnittstelle zur Inferenz von Text zu Audio unter Verwendung der Sprachsynthesemodelle.*
- AP 3.11: Entwicklung einer Schnittstelle zwischen Nutzer*innen und Lehrenden mittels Spracherkennung und Sprachsynthese
  - *Die gesprochene Sprache der nutzenden Person wird in Text umgewandelt und über eine Schnittstelle an eine Chatbot-Komponente übermittelt. Die textuelle Antwort wiederum wird über Sprachsynthese generiert und mit der Stimme der lehrenden Person wiedergegeben.*
- AP 3.12: Entwicklung einer Schnittstelle zwischen Chatbot und Nutzer*innen-/Lehrenden-Schnittstelle
  - *Implementierung eines Chatbot mit Schnittstellen zum Spracherkennungs- und Sprachsynthesesystem aus AP 3.11.*
- AP 3.13: Evaluation der wahrgenommenen audiovisuellen Qualität und der Qualität der Immersion
  - *Subjektive Auswertung der generierten Stimme des Avatars mittels üblicher Sprachverständlichkeitsmaße.*

- **AP 4: Datenspeicherung und -übertragung**
  - AP 4.8: Optimierung des Audiooutputs (1PM)
    - *Optimierung von Format und Kompression der Audiodaten im Hinblick auf effizientes Streaming ohne Einbußen der Sprachqualität.*

- **AP 5: Systemintegration und Demonstrator-Entwicklung**
  - AP 5.2: Integration der Sprachtechnologie-Komponenten
    - *Die Integration der Sprachtechnologiekomponenten erfordert eine Anbindung der Schnittstellen in das bestehende System. Diese müssen einen Einsatz im Livebetrieb ermöglichen. Die Audiodaten werden in einem möglichst effizient zu übertragenden Zielformat als Ausgabe zur Verfügung gestellt und aufgenommen. Um Mundbewegungen an das wiedergegebene Audiosignal anpassen zu können, werden über eine Schnittstelle Informationen über Lautdauern übermittelt.*
  - AP 5.6: Integration des Chatbots und iterative Optimierung des integrierten Chatbots
    - *Der in AP 2 entwickelte Chatbot wird in den Gesamtdemonstrator integriert und hinsichtlich der Performanz der Chatbots-Diskurse getestet, die Ergebnisse dokumentiert und in mehreren Iterationen optimiert.*

- **AP6: Ethische, rechtliche und soziale Implikationen (ELSI) und nutzer*innenorientierte Mensch-Technik-Interaktion (MTI)**
    - AP 6.2: MEESTAR-Workshop: Fokusdiskussion mit Projektbeteiligten
        - *Die ethischen, rechtlichen, sozialen und nutzer*innenspezifischen Dimensionen der zu entwickelnden Technologie werden mit allen Projektbeteiligten mit Hilfe eines strukturierten und angeleiteten MEESTAR-Workshops gemeinsam diskutiert und bewertet.*
    - AP 6.5: Erarbeitung eines Leitfadens zu ELSI und Nutzer*innenerwartungen sowie Auswahl der Testpersonen bzw. Merkmalsträger für die Akzeptanz- und Nutzer*innenstudien
        - *Unterstützung bei der Erstellung des Leitfadens zur Akzeptanz-Abfrage der Sprachsynthese*
    - AP 6.6: Akzeptanz- und Nutzungsstudie I mit Demonstrator
        - *Unterstützung bei der Durchführung der Interviews zur Akzeptanz-Abfrage der Sprachsynthese.*
    - AP 6.7: Akzeptanz- und Nutzungsstudie II mit Demonstrator in der Systemarchitektur
        - *Unterstützung bei der Durchführung der Interviews zur Akzeptanz-Abfrage der Sprachsynthese.*
    - AP 6.8: Akzeptanz- und Nutzungsstudie III: Fokusgruppendiskussionen
        - *Unterstützung bei der Durchführung der Gruppendiskussion zur Akzeptanz-Abfrage der Sprachsynthese.*
- **AP7: Zusammenarbeit mit dem Living Lab**
    - AP7.2: Einbringung der technischen Ergebnisse in das Living Lab
        - *Entwicklung gemeinsamer Spezifikationen, Schnittstellen und Testprozeduren, welche die Einzellösungen aus VoluProf und den anderen geförderten Projekten in einem übergreifenden Kontext zusammenführen und evaluieren.*

erfolgreich bearbeitet und deren Ziele erreicht.

2.2. Voraussichtlicher Nutzen sowie Verwertbarkeit der Ergebnisse und Erfahrungen

Das im Projekt VoluProf entwickelte Gesamtsystem konnte im Rahmen eines Demonstrators erfolgreich umgesetzt werden. Eine direkte Vermarktung wäre jedoch gegenwärtig nur im

Rahmen einer gemeinschaftlichen Weiterentwicklung nahezu aller Projektpartner realisierbar. Während der Projektlaufzeit hat sich gezeigt, dass insbesondere die für eine Anpassung an spezifische Lehrumgebungen und Lehrpersonen erforderlichen zeitlichen und personellen Aufwände derzeit noch zu hoch sind, um für Universitäten oder Schulen eine unmittelbare Nutzungsperspektive zu bieten.
Für eine breite Akzeptanz und Verwertbarkeit ist daher eine weitere Automatisierung und Verschlankung der Trainingsalgorithmen erforderlich, sodass der Anpassungsaufwand seitens der Bildungseinrichtungen erheblich reduziert wird. Erst unter diesen Bedingungen ist eine nachhaltige Integration in den Bildungsalltag realistisch.

Neben den unmittelbaren technischen Ergebnissen führte VoluProf auch zu einer nachhaltigen Kompetenzsteigerung innerhalb des Konsortiums. Wir als Aristech GmbH, haben im Zuge des Projekts unsere Expertise im Bereich Künstliche Intelligenz und deren praktischer Anwendung im Bildungsumfeld deutlich ausbauen konnte. Dies stärkt die Wettbewerbsfähigkeit des Unternehmens und eröffnet zusätzliche Verwertungsperspektiven über den Bildungsbereich hinaus, etwa in kulturellen oder industriellen Anwendungsfeldern.
Hierbei sind vor allem die Erfahrungen und Fortschritte im Bereich der TTS-Entwicklung bzw. dessen Trainings im Hinblick auf die erhebliche Reduktion der erforderlichen Datenmenge zu nennen.

2.3. Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Außerhalb des Konsortiums sind uns keine vorhabenrelevanten Ergebnisse durch Dritte bekannt.

2.4. Veröffentlichungen

Direkt und initiiert von unserer Seite gab es keine Veröffentlichungen. Wir waren jedoch am Inhalt des gemeinsamen Buchprojekts bzw. Kapitels „VoluProf – A system to bring your professor in your Living Room“ bzw. „VoluProf – Virtuelle Lehre tauscht Hörsaal gegen Wohnzimmer“ der Förderlinie „Interaktive Systeme in virtuellen und realen Räumen – Innovative Technologien für die digitale Gesellschaft (VAR2)“ beteiligt.
Nachfolgend findet sich eine Liste aller im Berichtszeitraum verfassten Arbeiten, die im Rahmen des Voluprof-Projekts entstanden sind.

Konferenzen, Journals, Buchkapitel:

1. Serhan Gül, Cornelius Hellge, Peter Eisert:
   Latency Compensation Through Image Warping for Remote Rendering-based Volumetric Video Streaming
   Proceedings of the IEEE International Conference on Image Processing (ICIP), October 2022
2. Jangwoo Son, Yago Sanchez, Christian Hampe, Dominik Schnieders, Thomas Schierl, Cornelius Hellge L4S congestion Control algorithm for interactive low latency applications over 5G, 2023 IEEE International Conference on Multimedia and Expo (ICME), Brisbane, Australia, July 2023.
3. Esther Greussing, Franziska Gaiser, Stefanie Helene Klein, Carolin Straßmann, Carolin Ischen, Sabrina Eimler, Katharina Frehmann, Miriam Gieselmann, Charlotte Knorr, Angelica Lermann Henestrosa, Andy Räder & Sonja Utz (2022): Researching interactions between humans and machines: methodological challenges. In: Publizistik. Vierteljahreshefte für Kommunikationsforschung. DOI: 10.1007/s11616-022-00759-3.
4. Peter Eisert, Oliver Schreer, Ingo Feldmann, Cornelius Hellge, Anna Hilsmann, Volumetric Video – Volumetric video–acquisition, interaction, streaming and rendering, pp. 289-326, Academic Press, UK, ISBN: 978-0-323-91755-1, September 2022.
5. Wegner, Juliane / Räder, Andy (2023): "Zwischen Virtualität und Realität. Mixed Reality in der Hochschulbildung." Communicatio Socialis, H4(56), S. 493-505.
6. Wolfgang Paier, Paul Hinzer, Anna Hilsmann, Peter Eisert, Video-Driven Animation of Neural Head Avatars, Vision, Modeling, and Visualization, The Eurographics Association, Braunschweig, Germany, September 2023
7. Wolfgang Paier, Anna Hilsmann, Peter Eisert, Unsupervised learning of style-aware facial animation from real acting performances, Graphical Models, Science Direct, vol. 129, September 2023
8. Paul Knoll, Wieland Morgenstern, Anna Hilsmann, Peter Eisert, Animating NeRFs from Texture Space: A Framework for Pose-Dependent Rendering of Human Performances, submitted to VISAPP 2024
9. Wieland Morgenstern, Paul Knoll, Anna Hilsmann, Peter Eisert, Animatable Virtual Humans: Learning pose-dependent human representations in UV space for interactive performance synthesis, IEEE VR 2024
10. Jangwoo Son, Yago Sanchez, Thomas Schierl, Cornelius Hellge, "Low-latency Cloud-based Streaming through L4S Congestion Control over 5G," ICIP 2024.
11. Jangwoo Son, Yago Sanchez, Thomas Schierl, Cornelius Hellge, "L4S Congestion Control for XR Remote-Rendering over 5G," IEEE Communications Magazine: eXteneded Reality 2024.
12. David Moreno-Villamarin et al., Multi-Resolution Generative Modeling of Human Motion from Limited Data, ACM SIGGRAPH CVMP, London, Nov. 2024, best paper runners up

Messen:

1. MWC 2024, Mixed Reality Video Streaming with L4S
2. MWC 2023, Mixed Reality Video Streaming with L4S
3. *MWC 2022, Mixed Reality Streaming over 5G
4. IBC 2022, Interactive Volumetric Video Streaming over 5G

Präsentationen:

1. Cornelius Hellge, Anna Hilsmann, Thomas Buchholz, Andy Räder, VoluProf Panel, 3IT Summit, 5.Mai 2022, Berlin, Germany.
2. Cornelius Hellge, Interactive Volumetric Video Streaming over 5G using L4S, IEEE BTS Pulse, July 14 2022 (online). (https://www.5g-mag.com/post/12-14-07-22-ieee-bts-pulse)
3. Cornelius Hellge, Holoportation & Avatars, FutureHotels Innovation Breakfast, June 2022. https://www.linkedin.com/company/futurehotel/ (online)
4. Anna Hilsmann, "Von digitalen Avataren zu virtuellen Menschen" Go-Visual 2022, Berlin, Germany.
5. Cornelius Hellge, "Streaming of Immersive Media", 2023 IEEE SPS/EURASIP Summer School on Metaverse Technologies, Cagliari, Italy, 18-22 September 2023.
6. Peter Eisert & Anna Hilsmann: From volumetric Video to Realistic Avatars, Technology Innovation Days 2023; 27. + 28. Juni 2023
7. Peter Eisert & Anna Hilsmann, From volumetric Video to Realistic Avatars, SMPTE Webcast, 2. März 2023
8. Cornelius Hellge, Thomas Buchholz, "Kommunikation Deluxe: Plaudern im 22. Jahrhundert – Von Avataren bis zu Hologrammen", T-Labs Spatial Insights Edition 1, October 2023
9. Johann-Christian Põder: Interdisziplinäre Ringvorlesung Digitales Lehren und Lernen „Mixed Reality Avatare in der Hochschullehre", Universität Rostock, 22.06.2023
10. Johann-Christian Põder, Andy Räder: Universität im Rathaus „Avatare in der Hochschullehre? Die Zukunft des Lernens mit Datenbrille und Künstlicher Intelligenz", Universität Rostock, 28.09.2023
11. Andy Räder: Ringvorlesung Digital Humanities im Fokus „Lehren und Lernen in der Spatial Reality: Chancen und Herausforderungen für die Hochschullehre", Universität Rostock, 20.11.2023
12. Dominik Schnieders, Cornelius Hellge, Philipp Landgraf, "Managed Latency with L4S", DTAG, Technology Innovation: Deep Dive, MWC 2022.
13. Johann-Christian Põder: Diskussion Schulleiter:innen "Tag der Schulleiterinnen und Schulleiter in Mecklenburg-Vorpommern, Ministerium für Bildung und Kindertagesförderung, VoluProf-Workshop/MEESTAR, 01.03.2024.
14. Benjamin Körner: DIZcover „Innovationsbooster durch Fördermittel" am Beispiel des Projekts „VoluProf", European Digital Innovation Hub applied Artificial Intelligence and Cybersecurity (EDIH-AICS), Digital Hub angewandte Künstliche Intelligenz Karlsruhe
15. Johann-Christian Põder: Diskussion Studierende Blockseminar "Balance: Einführung in interdisziplinäres Denken"(Fakultät für Maschinenbau und Schiffstechnik, Theologische Fakultät, Hochschule Wismar: Studierende aus Maschinenbau, Architektur, Kunst und Theologie) und Seminar "Ethik der medizinischen Informationstechnik" (Fakultät für Informatik und Elektrotechnik/Universitätsmedizin Rostock; Thema: Ethisches Design, Medizindidaktik), 05.06.2024.