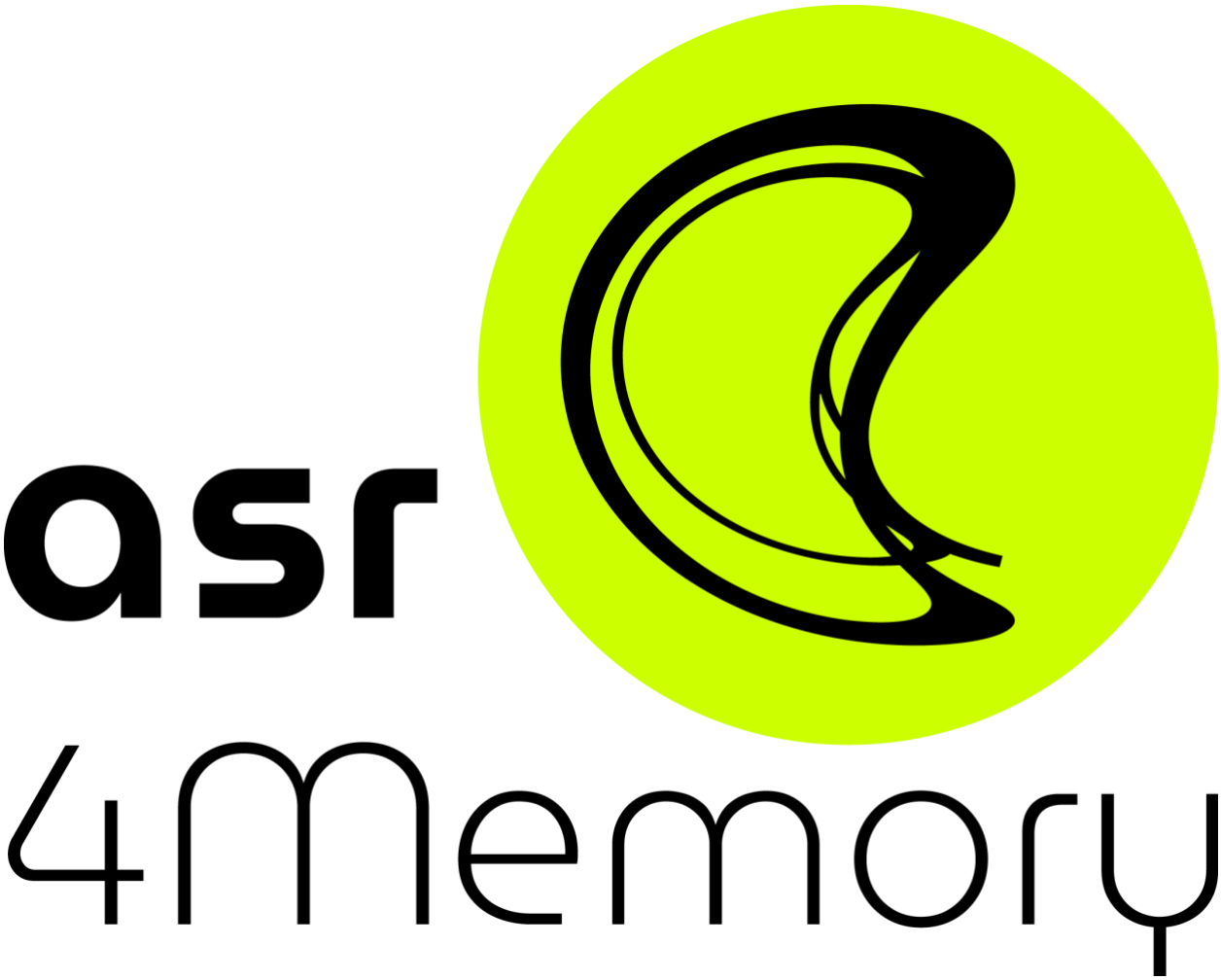
Blogbeitrag 4Memory

Vom gesprochenen Wort zum Text: KI-gestützte Transkription audiovisueller Forschungsdaten

Im Rahmen der „Incubator Funds“ 2024, die vom NFDI-Konsortium „4Memory“ der historisch arbeitenden Geisteswissenschaften ausgeschrieben werden, hat das Projekt „ASR4Memory“ einen KI-gestützten Transkriptionsservice für historische audiovisuelle Forschungsdaten aufgebaut, der u.a. in dem Forschungsfeld der Oral History zur Anwendung kommt.



Ausgangslage

In vielen Bibliotheken, Archiven, Universitäten, Museen und Gedenkstätten existieren bislang nicht erschlossene Sammlungen audiovisueller Quellen, die sich zum Beispiel aus Zeitzeugeninterviews, Fernsehmitschnitten, Radioübertragungen, Dokumentarfilmen, Aufzeichnungen politischer/wissenschaftlicher Vorträge, Mitschnitte von Gerichtsverfahren oder Feldinterviews zusammensetzen. In den Einrichtungen besteht ein großes Interesse, diese

NFDI4Memory - Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 501609550

wertvollen audiovisuellen Ressourcen technisch aufzubereiten, wissenschaftlich zu analysieren, nach den FAIR-Standards zugänglich zu machen sowie bei neu entstehenden („Digital-Born“) Aufzeichnungen die Nachnutzbarkeit mitzudenken. Die Transkription der darin gesprochenen Sprache bildet dafür eine wichtige Voraussetzung. Mit der rasanten Weiterentwicklung der Künstlichen Intelligenz (KI) eröffnen sich vielfältige Möglichkeiten für die Transkription mit Hilfe automatischer Spracherkennung (Automatic Speech Recognition, ASR), die auch in wissenschaftlichen Kontexten wie der Oral History an Bedeutung gewinnen.

Die Nutzung von Open-Source-Spracherkennern bietet zwei wesentliche Vorteile: Zum einen reduziert sie in erheblichem Maße die Datenschutzproblematik, da die audiovisuellen Ressourcen nicht in kommerziellen Cloudumgebungen, sondern ausschließlich auf lokalen Servern der Freien Universität Berlin verarbeitet werden. Zum anderen lassen sich die Kosten, die im Falle umfangreicher Sammlungen für viele Einrichtungen nicht oder nur schwer finanzierbar sind, im Vergleich zu manuellen Transkriptionen und kommerziellen Transkriptionsdiensten deutlich verringern.

Das Angebot zur automatischen Transkription audiovisueller Forschungsdaten wurde sehr gut angenommen und hat gezeigt, dass in der Forschungscommunity ein großer Bedarf daran besteht. Nahezu 30 Einrichtungen stellten Forschungsdaten für eine Pilotnutzung in der Transkriptionspipeline bereit. Die Aufzeichnungen lagen in verschiedenen Sprachen vor, u.a. in Deutsch, Französisch, Portugiesisch, Spanisch, Englisch, Arabisch und Ukrainisch. Durch die Ausrichtung eines praxisorientierten Online-Workshops im März 2024 mit Inhaber*innen von audiovisuellen Datenbeständen und den direkten Austausch mit den Pilotnutzenden wurden die „Bedarfe, Anforderungen und kritische Bewertung des automatisierten Transkriptionsservices“ (Workshop-Titel) sowie die Ausgestaltung eines langfristigen Betriebsmodells beleuchtet. Berücksichtigt wurden technische und funktionale Anforderungen, Schnittstellenbedarfe, Metadatenmodelle, Nutzungsszenarien sowie rechtliche und ethische Aspekte. In diesem Zusammenhang diskutierten wir die Fragen von Diskriminierung sprachlicher Minderheiten und Rassismus sowie generell hinsichtlich der Intransparenz von KI-Methoden. Durch den kontinuierlichen Austausch mit den Pilotnutzenden wurden vielfältige Anwendungsszenarien mit individuellen Bedarfen und Herausforderungen identifiziert, die in die Entwicklungsarbeit einflossen. Im Gegenzug erhielten die Nutzenden qualitativ hochwertige Transkripte ihrer audiovisuellen Ressourcen für die Nachnutzung.

Projektresultate

Die Entwicklungsergebnisse sind seit Januar 2025 als Web-Service der Freien Universität Berlin oder als lokale Installation der Open-Source-Software für Interessierte nutzbar (siehe die Links am Ende des Beitrags). Die technische Grundlage der Transkriptionsstrecke bildet „WhisperX“, eine an der Universität Oxford entwickelte Open-Source-basierte Re-Implementierung des Spracherkenners „Whisper“ der Firma OpenAI, die auch das bekannte Large Language Modell „ChatGPT“ entwickelt hat. „WhisperX“ hat sich in der internen Evaluation als der für die identifizierten Bedarfe geeignetste automatisierte Spracherkenner (ASR) herausgestellt und wurde zur Weiterentwicklung in diesem Projekt ausgewählt. Die auf Grundlage der Programmiersprache „Python“ entwickelte Pipeline, die aus ineinandergreifenden Pre- und Postprocessing-Komponenten aufgebaut ist, ermöglicht eine hohe Transkriptqualität bei einer gleichzeitig hohen Zeiteffizienz in der Datenverarbeitung. Zudem werden Datenschutz und -interoperabilität sichergestellt.

NFDI4Memory - Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 501609550

Für den Datenaustausch wurde eine browserbasierte Weboberfläche (Media Management Tool – MMT) entwickelt, über die die Sammlungsinhaber*innen ihre audiovisuellen Ressourcen auf die an der Freien Universität Berlin betriebenen Transkriptionsplattform laden, in der die Daten sicher und datenschutzkonform ausschließlich auf lokalen Servern verarbeitet werden.

In einem ersten Schritt wird die Integrität des audiovisuellen Digitalisats automatisch geprüft. Um eine höchstmögliche Transkriptionsqualität zu gewährleisten, wird die Tonspur aus der Mediendatei in einem für die automatische Transkription passenden Audioformat extrahiert und das darin enthaltene Sprachsignal für die nachfolgende Spracherkennung automatisch tontechnisch optimiert (Anpassung der Lautstärke, Hervorhebung der Sprachfrequenzen). Danach folgt die Transkription in der Originalsprache (aktuell sind 30 Sprachen möglich) mit passgenauen Einstellungen, woraus schließlich die Transkriptformate erzeugt und im letzten Schritt den Nutzer*innen online über eine sichere Datenübertragung bereitgestellt werden.

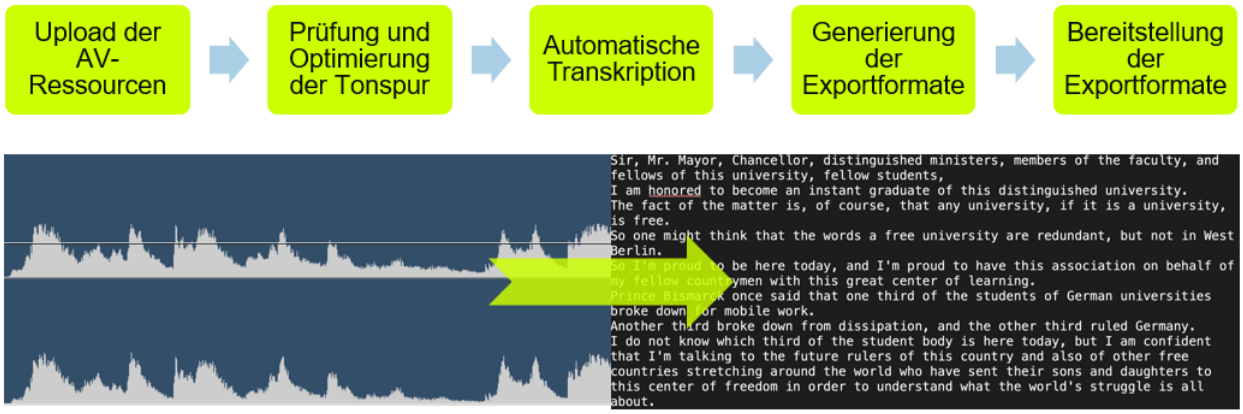
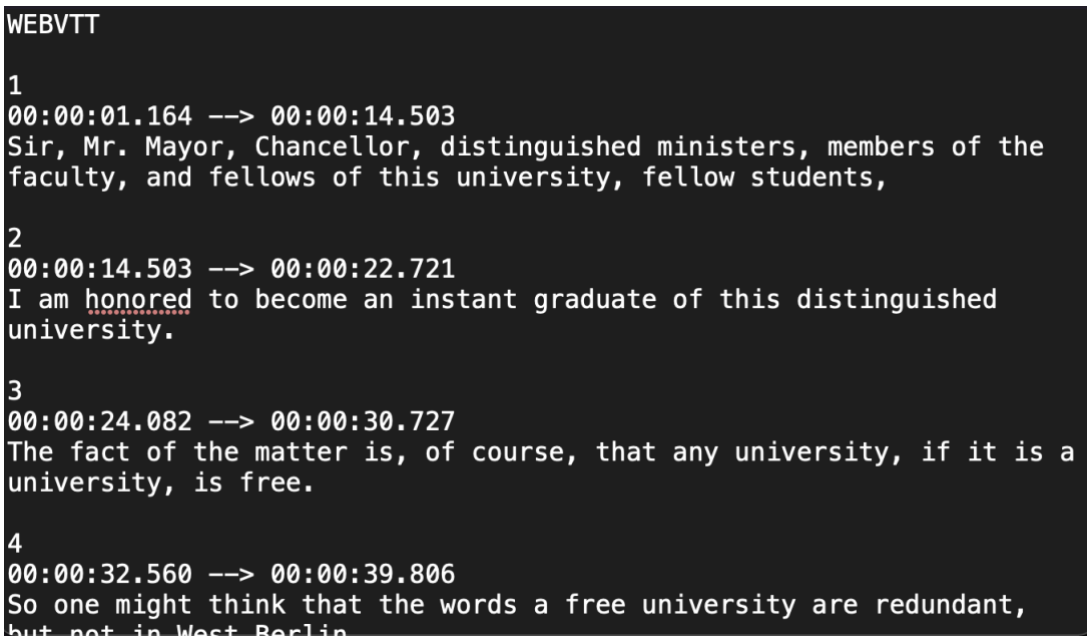


Abbildung 1: Schematischer Workflow der Transkriptionspipeline mit den einzelnen Arbeitsschritten.

Um die vielfältigen Bedürfnisse der Nutzenden abzudecken, werden verschiedene Transkriptformate generiert, darunter TXT- und ODS-Dateien zur manuellen Nachbearbeitung in Textverarbeitungsprogrammen, CSV- und JSON-Dateien zur automatischen Datenverarbeitung in weiteren Systemen (z.B. Repositorien), VTT- und SRT-Dateien zur Untertitelung von AV-Medien in Software-Playern sowie PDF-Dateien zur Bereitstellung und Langzeitsicherung der Transkripte. Zum Teil enthalten die Exportdateien Sprecher*innenauszeichnungen sowie wort- oder satzbasierte Zeitmarken (Timecodes). Auch eignen sich die Formate für den Import in die Erschließungs- und Recherche-Plattform „Oral-History.Digital“, wofür aktuell Schnittstellen zum 4Memory Data Space entwickelt werden. So können Transkripte mit entsprechenden Nutzungsrechten der 4Memory-Community zugänglich gemacht werden.

NFDI4Memory - Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 501609550



```
WEBVTT

1
00:00:01.164 --> 00:00:14.503
Sir, Mr. Mayor, Chancellor, distinguished ministers, members of the
faculty, and fellows of this university, fellow students,

2
00:00:14.503 --> 00:00:22.721
I am honored to become an instant graduate of this distinguished
university.

3
00:00:24.082 --> 00:00:30.727
The fact of the matter is, of course, that any university, if it is a
university, is free.

4
00:00:32.560 --> 00:00:39.806
So one might think that the words a free university are redundant,
```

Abbildung 2: Beispiel des Untertitelformats VTT mit satzbasierten Zeitmarken und englischer Transkription.

Die exakte Zeitkodierung der Transkripte mit millisekundengenauen Zeitmarken (für jeden Satz und jedes Wort) ermöglicht ihre Synchronisierung mit den audiovisuellen Medien und führt somit zu einer verbesserten Auffindbarkeit, Zugänglichkeit, Interoperabilität und Nachnutzbarkeit (FAIR) der audiovisuellen Ressourcen in digitalen Anwendungen. Schließlich eröffnen sich durch die Verschriftung der Ton- und Filmaufnahmen plattformübergreifende Analyse-, Kontextualisierungs- und Metadatenoptionen, u.a. im Bereich der Textanalyse und des Natural Language Processings.

1	WORD	START	END	SCORE
2	Ja.	00:00:02.077	00:00:02.398	616
3	Darf	00:00:02.438	00:00:02.637	455
4	ich	00:00:02.677	00:00:02.758	329
5	rein?	00:00:02.959	00:00:03.738	0.65
6	Ja.	00:00:03.778	00:00:04.099	764

Abbildung 3: Beispiel der CSV-Ausgabedatei mit wortbasierten Timestamps und deutscher Transkription.

Neben dem Service-Angebot an der Freien Universität Berlin ist der Quellcode open source in öffentlichen GitHub-Repositorien veröffentlicht. Somit können die Nutzenden selbständig und dezentral die Pipeline auf lokalen Maschinen aufsetzen, datenschutzkonform betreiben, den Quellcode eigenständig weiterentwickeln und in der Forschungscommunity weiter verteilen.

Schwachstellen der Spracherkennung

Auch wenn die neue Generation der KI-basierten Spracherkenner eine deutlich höhere Qualität bei der automatischen Transkription audiovisueller Daten ermöglicht, weisen auch diese Schwächen auf. Der schwerwiegendste (wenn auch eher seltene) Fehler ist die sog. Halluzination, bei der nicht gesprochene Inhalte, die ursprünglich aus den KI-Trainingsdaten stammen, fälschlicherweise generiert (sprich erfunden) werden und keinerlei inhaltlichen Bezug zum Gesprochenen haben. Ein weiteres Defizit besteht in der Falscherkennung von wichtigen

NFDI4Memory - Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 501609550

Entitäten wie Personennamen, Organisationen, Orten oder Ereignissen. Nachfolgend ein Beispiel:

- Gesprochener Text: „Ja, das war Sommer 89, wo die dann alle da in Prag, da in der Botschaft gesessen haben und der **Genscher** hat sie dann da rausgekloppt."
- Erkannter Text: „Ja, das war Sommer 89, wo die dann alle da in Prag da in der Botschaft gesessen haben und der **Kenja** hat sie dann da rausgeklappt."

Zudem haben die ASR-Tools zum Teil Probleme damit, einen gesprochenen Satz dem/der jeweiligen Sprecher*in korrekt zuzuordnen, was besonders bei schnellen Sprecherwechseln, parallelem Sprechen und im Falle vieler Sprecher*innen auftritt.

Des Weiteren „glätten" die Spracherkenner das Transkript, in dem Füllwörter, Wortwiederholungen, Satzabbrüche und Verzögerungen oft nicht transkribiert werden. Beispiele:

- Gesprochener Text: „Wir wohnten damals in Berlin-Köpenick. In_ in_. Äh, ähm, na, eine Hütte war es ja eigentlich gewesen."
- Erkannter Text: „Wir wohnten damals in Berlin-Köpenick. In einer Hütte war es ja eigentlich."

Zudem werden nicht-sprachliche Lautäußerungen (wie Lachen oder Weinen), Sprechpausen und direkte Rede nicht transkribiert:

- Gesprochener Text: „Ich sage: ‚Weißt du was, jetzt fahre ich mal schnell rüber, hole uns ein paar Zigaretten.' <s(lachend) Fahre die Schönhauser runter, bieg in die Brunnenstraße ein,> da standen sie, einer nebeneinander."
- Erkannter Text: „Ich sage, weißt du was, jetzt fahre ich mal schnell rüber, hol uns ein paar Zigaretten. Ich fahre da schön runter und da kriege ich eine Brunnenstraße hin. Da standen sie jeder nebeneinander."

Die Spracherkenner führen grammatikalische Korrekturen, insbesondere beim Satzbau, durch. Zudem werden Dialekte und Akzente ins Hochdeutsche transformiert:

- Beispiele: „ick" → „ich" / „jewesen" → „gewesen" / „kriech' ta" → „kriegte er" / usw.

Aufbereitung der Trainingsdaten

Die KI-basierte Spracherkennung weist einige Schwächen auf, erzeugt Transkriptionsfehler und ermöglicht keine rein wortgetreue Transkription. Daher wurde im Projekt geprüft, ob über das Feintuning des Spracherkennungsmodells mit domänenspezifischen Trainingsdaten die Transkriptqualität erhöht werden kann. Dafür ausgewählt wurden bereits optimal transkribierte Oral-History-Interviews mit hoher Audioqualität. Da die für das Training erforderliche Aufbereitung ein sehr zeitaufwändiger Arbeitsschritt ist, wurden im Projekt effiziente Workflows entwickelt, in denen mehrere KI-unterstützte Komponenten die technischen Aufgaben übernehmen. Zur Aufbereitung gehören, erstens, die Klärung der rechtlichen Bedingungen für die Nutzung der Forschungsdaten im KI-Training. Zweitens wird ein technisch komplexes Preprocessing durchgeführt, das die textuellen und audiovisuellen Daten in einzelne Segmente aufsplittet (insgesamt etwa 190.000) und in das für Deep-Learning optimierte HDF5-Format konvertiert. Drittens stellt die textuelle und akustische Anonymisierung der Zeitzeugeninterviews sicher, dass die in der Aufzeichnung genannten sensiblen Informationen sowohl aus dem Transkript als auch aus der Tonspur automatisiert entfernt werden. Mithilfe einer LLM-gestützten Named Entity Recognition (NER) werden die zu anonymisierenden Entitäten wie Namen und Adressen erkannt, die wiederum in der Tonspur mit dem FFmpeg-Tool millisekundengenau unkenntlich gemacht

NFDI4Memory - Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 501609550

werden. Der Anonymisierungsvorgang erfolgt zum einen aufgrund rechtlicher Vorgaben des High-Performance Computers sowie aus ethischen Gründen mit Blick auf die Zeitzeugen und ihre aufgezeichneten Erfahrungsberichte. Zudem verhindert die Anonymisierung der Trainingsdaten unerwünschte Halluzinationen von sensiblen Informationen aus den Trainingsdaten, die in der Anwendung des trainierten Modells bei anderen audiovisuellen Ressourcen auftreten könnten.

Feintuning der Spracherkenner

Im Rahmen eines *Proof-of-Concepts* konnten wir nachweisen, dass sich über ein Deep-Learning-basiertes Finetuning des Spracherkennungsalgorithmus auf dem High-Performance Computer der Freien Universität Berlin eine signifikante Verbesserung der Transkriptqualität erzielen lässt. Das Feintuning wird maßgeblich von den Hyperparametern beeinflusst, welche vorab festgelegt werden und den Lernprozess des ASR-Modells steuern. Um einen möglichst hohen Trainingseffekt zu erzielen, werden aus einer Vielzahl möglicher Hyperparameter die wirkungsstärksten Parameter mit entsprechenden Einstellungen identifiziert (z.B. Learning Rate, Anzahl der Trainingsepochen). Für die Optimierung der Hyperparameter setzen wir bayessche Statistikmethoden und die Bibliothek „Ray Tune“ ein, um verschiedene Kombinationen von Hyperparametern parallel und in iterativen Durchläufen (Epochen) über die Trainingsdaten zu prüfen und die besten Trainingsmodelle auszuwählen.

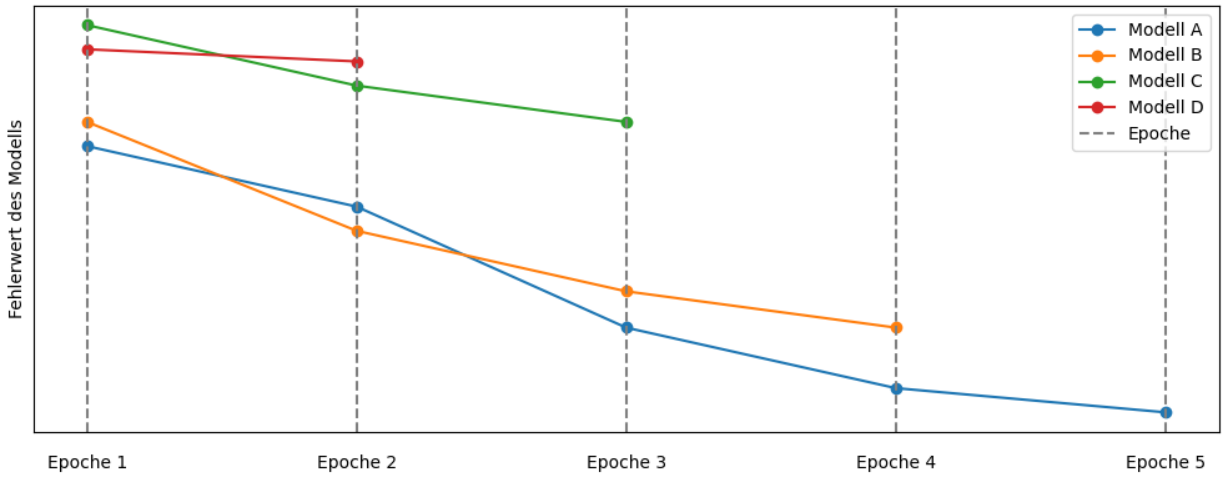


Abbildung 4: Schematische Darstellung der Hyperparameter-Optimierung, die verschiedene Hyperparameter-Kombinationen (farbige Linien) testet und schrittweise (je „Epoche“) die schlechteren Modelle mit höheren Fehlerwerten verwirft, sodass nur die besten bis zum optimalen Ergebnis weitertrainiert werden (in diesem Beispiel das Modell A).

Durch das Hyperparameter-Finetuning konnte die Spracherkennung systematisch verbessert und eine höhere Transkriptqualität erreicht werden. In der Auswertung des feingetunten Modells zeigte sich, dass durch das domänenspezifische Training zum einen die quantitative Wortfehlerrate signifikant reduziert wurde (für das kleinste Modell um etwa 10 Prozent). Zum anderen wurden auch historische Begriffe und Eigennamen deutlich besser erkannt und somit qualitative Verbesserungen des Spracherkenners erzielt (etwa 15 Prozent). Jedoch konnten wir den Machbarkeitsnachweis aufgrund der begrenzten Rechenkapazitäten nur bei den kleineren Sprachmodellen durchführen. So beanspruchten die Trainingsprozesse auf dem High-Performance Computer viel Zeit und erzeugten große Datenmengen. Dennoch lässt sich aus den Ergebnissen folgern, dass – sofern ausreichende Rechenkapazitäten und fachlich kuratierte Trainingsdaten vorliegen – die Entwicklung domänenspezifischer Spracherkennungsmodelle machbar ist. Indem sie schwerwiegende Fehler reduzieren (insbesondere bei Personennamen,

NFDI4Memory - Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 501609550

Orten und Ereignissen), erzielen sie eine signifikant höhere Transkriptgenauigkeit und schaffen so einen erheblichen Mehrwert für die jeweilige Fachdisziplin.

Evaluation der Transkription

Bislang wird zur Bewertung von Spracherkennern meist die Wortfehlerrate herangezogen, die eine wichtige quantitative und vergleichbare Metrik darstellt. Die Wortfehlerrate ist allerdings nur bedingt aussagekräftig: Die fehlerhafte Transkription von wichtigen historischen Begriffen und Eigennamen ist deutlich problematischer als die Transkription eines falschen Artikels oder Füllworts. Jedoch bewertet die Wortfehlerrate beide Fehlerarten gleich. Um die diese Defizite systematisch zu analysieren (und künftig zu reduzieren), wurde ein Evaluationsverfahren mit der Zielsetzung entwickelt, die quantitativen Evaluationsergebnisse über eine qualitative Bewertung der Transkriptionsgüte zu ergänzen.

Zentraler Baustein in diesem Verfahren ist ein lokal betriebenes, offlinefähiges und Open-Source-basiertes Large Language Modell (LLM), das mit 70 Milliarden Parametern komplexe Aufgabenstellungen bewältigt, allerdings große Rechenressourcen bei der Anwendung benötigt. Zunächst definierten wir trennscharfe Fehlerkategorien und gewichteten diese nach ihrer Relevanz für die Transkriptqualität (unter Einbezug der bisher mit den automatischen Transkripten arbeitenden Historiker*innen an der FU Berlin). Das LLM kategorisiert (auf Grundlage eines spezifischen Prompts) die auftretenden Fehler und berechnet die Fehlerverteilung im erzeugten Transkript. Die erzeugte Auswertung zeigt die Häufigkeit der auftretenden Fehlertypen auf und gibt Rückschluss darüber, wie schwerwiegend die Fehler für die Transkriptqualität sind. Dieser gemischte Ansatz hat sich als sehr hilfreich in der Bewertung KI-unterstützter Spracherkenner erwiesen.

Die folgenden Diagramme zeigen die einzelnen Fehlerkategorien und die Fehlerhäufigkeiten des originalen sowie des feingetunten Modells. Zum Beispiel repräsentiert „d2“ das „Entfernen eines bedeutungsvollen Wortes mit inhaltlicher Relevanz“ oder „i2“ das „Einfügen eines halluzinierten bedeutungsvollen Wortes“. Die „weights“ definieren die Gewichtung des Fehlertyps hins. der Transkriptqualität. Je höher der Wert, desto größer die Relevanz für die Transkriptqualität.

NFDI4Memory - Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 501609550

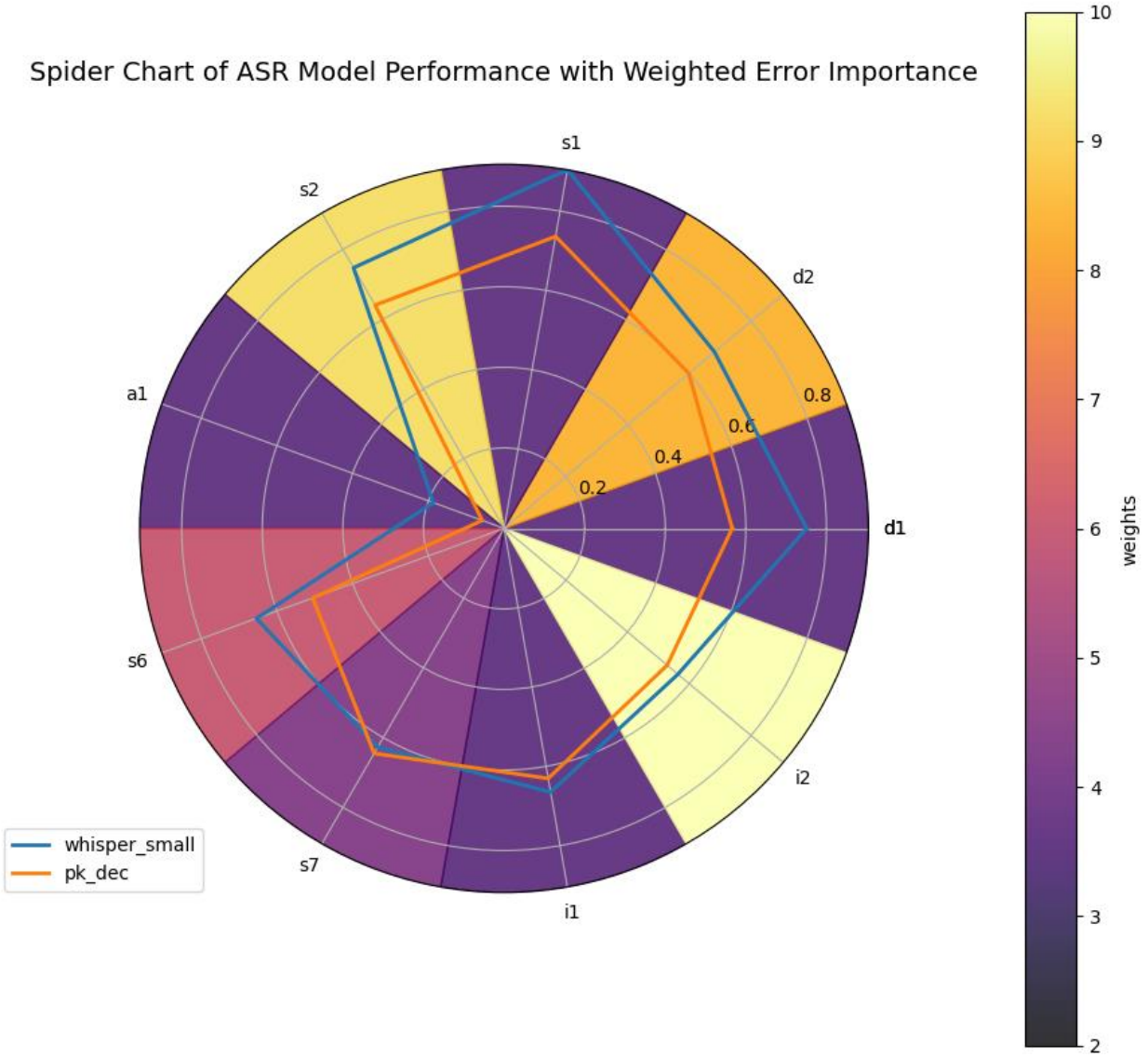


Abbildung 5: Vergleich der (gewichteten) Fehlerhäufigkeiten zwischen dem Original-Modell (blau) und feingetunten Modell (orange).

NFDI4Memory - Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 501609550

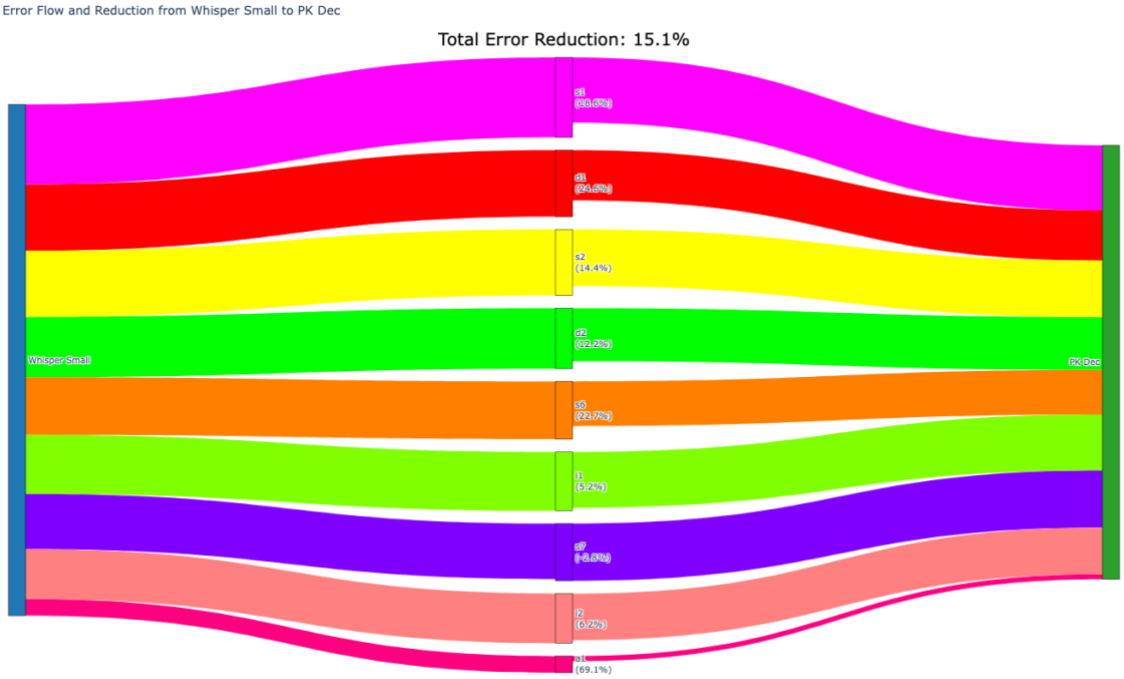


Abbildung 6: Darstellung der Fehlerreduktion zwischen dem Original-Modell (links) und dem feingetunten Modell (rechts).

Nächste Schritte

Aktuell arbeiten wir daran, ein domänenspezifisches Spracherkennungsmodell für die historisch arbeitenden Geisteswissenschaften zu entwickeln. Dabei verfolgen wir das Ziel, dass in deutschsprachigen Aufzeichnungen historische Begriffe und Zusammenhänge wie Orte, Ereignisse und Eigennamen mit noch höherer Wahrscheinlichkeit korrekt erkannt und die für KI-Anwendungen charakteristischen Halluzinationen verringert werden. Zudem sollen sowohl non-verbale Kommunikationsereignisse und Sprechpausen millisekundengenau erfasst als auch die multilinguale Transkription von mehrsprachigen Quellen ermöglicht werden. Die Zielstellung ist, dieses optimierte und domänenspezifische Modell als open source der Fachcommunity zur Nutzung in Forschungskontexten zur Verfügung zu stellen.

Zum anderen haben wir den Bedarf identifiziert, das Spracherkennungsmodell mit selteneren Sprachen (z.B. Aramäisch, Quechua) zu trainieren und somit deren automatische Transkription zu ermöglichen. Diese angepassten Modelle könnten in unsere Transkriptionspipeline integriert und somit den entsprechenden Communities zugänglich gemacht werden.

Für die Anbindung an weitere Plattformen und Repositorien sollen zusätzliche wissenschaftliche Exportformate über die Transkriptionspipeline zukünftig angeboten werden, so die Austauschformate TEI-XML und/oder IIIF-AV.

Das Projekt wird durchgeführt von Dr. Tobias Kilgus, Peter Kompiel, Marc Altmann und Dr. Christian Horvat.

Weiterführende Links

- Projektwebseite: https://www.fu-berlin.de/asr4memory
- GitHub-Repositorien: https://github.com/asr4memory

NFDI4Memory - Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 501609550

- „Lange Nacht der Wissenschaften“ 2024 in Berlin: https://www.ada.fu-berlin.de/kalender/LNDW2024.html
- Transkription der Zeitzeugeninterviews des Projekts „Erlebte Geschichte“ an der Freien Universität: https://archiv.erlebte-geschichte.fu-berlin.de
- Beispiel der automatischen Spracherkennung: https://www.cedis.fu-berlin.de/services/medien/av-medien/test/kennedy-rede/index.html

NFDI4Memory - Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 501609550