ASR4Memory. Automatische Transkription und domänenspezifisches Fine-Tuning von Spracherkennungsmodellen für die Oral History

In zahlreichen Archiven, Universitäten, Museen, Gedenkstätten und Bibliotheken liegen bislang unerschlossene Sammlungen audiovisuellen Forschungsdaten, wie z.B. Oral-History-Interviews. Die Einrichtungen zeigen großes Interesse daran, diese wertvollen Quellen technisch aufzubereiten, multimodal zu erschließen und nach wissenschaftlichen Standards zugänglich zu machen (vgl. Apel et al. 2022). Für eine nachhaltige Nachnutzung und Zugänglichmachung gemäß der FAIR-Prinzipien ist die Transkription dieser Quellen von zentraler Bedeutung. Bislang wurde diese jedoch meist manuell und mit erheblichem Aufwand durchgeführt (Wilkinson et al. 2016). Das durch das NFDI-Konsortium 4Memory geförderte Projekt „ASR4Memory“[1] hat eine datenschutzkonforme und KI-gestützte Anwendung auf Basis von „WhisperX“ entwickelt (Kilgus/Kompiel 2025), die in das Dienstportfolio von NFDI4Memory aufgenommen wurde.[2] Darüber hinaus wurde ein erstes prototypisches Fine-Tuning des Modells mit kuratierten, anonymisierten Oral-History-Daten auf einem Hochleistungsrechner umgesetzt.

Anwendung

Die Anwendung „ASR4Memory“ ermöglicht einen einfachen, datenschutzkonformen Upload audiovisueller Ressourcen über eine browserbasierte Oberfläche. Nach Prüfung der Dateiintegrität wird die Tonspur extrahiert, technisch optimiert und mit WhisperX (Modell „large-v3“) in höchster Qualität transkribiert. Die automatische Spracherkennung umfasst bis zu 30 Sprachen, inklusive Sprecherdiarisierung sowie Wort- und Satzalinierung (vgl. Bain et al. 2023). Ein angepasstes Postprocessing-Skript sorgt für satzbasierte Alinierung und Segmente von maximal 120 Zeichen, die sich für Untertitel und inhaltliche Erschließung eignen.

Die generierten Transkriptformate umfassen:

- Text- und PDF-Dateien für die Langzeitsicherung,
- VTT- und SRT-Dateien mit Timecodes für Medienplayer,
- CSV-Dateien mit und ohne Sprecherangaben zur manuellen Korrektur,
- JSON-Dateien mit Timecodes und Sprecherauszeichnungen

[1] Projektwebseite „ASR4Memory“: https://www.fu-berlin.de/asr4memory (zugegriffen: 15.07.2025).
[2] https://4memory.de/dienste-ressourcen/

Fine-Tuning

Trotz enormer Fortschritte erzeugt das Modell „large-v3“ für Oral History-Interviews noch immer teils fehlerhafte und nicht vollständig wortgetreue Transkripte (vgl. Wollin-Giering et al. 2024). So werden historische Begriffe und Eigennamen schlecht erkannt und Häsitationen („ähm“, „äh“) sowie Wort- und Satzabbrüche geglättet. Daher wurde untersucht, ob ein Fine-Tuning des Modells mit domänenspezifischen Trainingsdaten die Soll-Qualität verbessern kann. Grundlage bildeten 95 lektorierte Oral-History-Interviews aus der Sammlung Erlebte Geschichte[3] (über 300 Stunden Audiomaterial).

Zur Vorbereitung des Trainingsdatensatzes wurden automatisierte Workflows entwickelt, die Aufgaben wie die textuelle und akustische Anonymisierung (LLM-gestützte NER), Aufteilung von ca. 200.000 Audio- und dazugehörigen Transkriptsegmenten in Trainings-, Validierungs- und Test-Split (80% / 10% / 10%) sowie Konvertierung ins HDF5-Format übernehmen.

Das Fine-Tuning erfolgte über Hyperparameter-Optimierung mit bayes’schen Statistikmethoden mithilfe des Python-Pakets „Ray Tune“, um im Rahmen eines populationsbasierten Ansatzes Parameter wie Learning Rate, Warm-Up Steps und Weight Decay iterativ zu optimieren.

Die Anpassung des Modells führte zu einer signifikant verbesserten Transkriptqualität: Die Wortfehlerrate sank um rund 10 %, historische Begriffe und Eigennamen wurden etwa 15 % besser erkannt. Zudem reduzierten sich die typischen Sprachglättungen Im Vergleich zum Ausgangsmodell deutlich.

Der Vortrag präsentiert die Funktionsweise der ASR-Anwendung und den Stand der Fine-Tuning-Ergebnisse. Beides wird in einem iterativen Austausch mit der geschichtswissenschaftlichen Community weiterentwickelt. Der Quellcode ist open source in einem GitHub-Repositorium veröffentlicht.[4]

Literatur

Apel, Linde, Almut Leh und Cord Pagenstecher (2022): Oral History im digitalen Wandel. Interviews als Forschungsdaten. https://zeitgeschichte-hamburg.de/files/public/FZH/PDF/apel_erinnern_ebook_offen.pdf (last access: 15.10.2025).

Bain, Max et al. (2023): WhisperX: Time-Accurate Speech Transcription of Long-Form Audio, https://doi.org/10.48550/arXiv.2303.00747.

Kilgus, Tobias and Peter Kompiel (2025): Das Transkript im Zeitalter der Künstlichen Intelligenz. Automatische Spracherkennung in der Oral History, in: BIOS - Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen 38 (1), 2025 (accepted).

[3] https://archiv.erlebte-geschichte.fu-berlin.de/ (zugegriffen: 19.10.2025).
[4] https://github.com/asr4memory (zugegriffen: 19.10.2025).

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18.

Wollin-Giering, Susanne, Markus Hoffmann, Jonas Höfting und Carla Ventzke (2024): Automatic Transcription of English and German Qualitative Interviews, in: Forum Qualitative Sozialforschung 25 (1), https://doi.org/10.17169/fqs-25.1.4129.