FREIE UNIVERSITÄT BERLIN

Oral-History.Digital

formemory

ASR4Memory

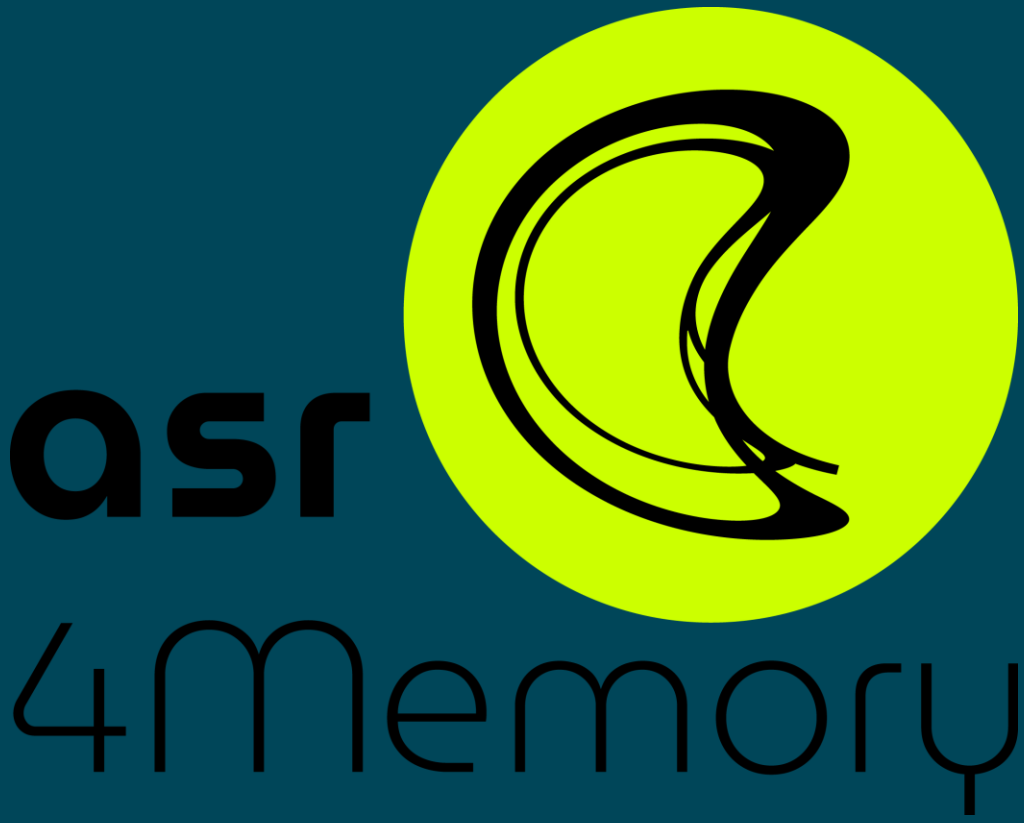
Ein KI-gestütztes Transkriptionsangebot für audiovisuelle Forschungsdaten

Dr. Tobias Kilgus, Verena Nägel | Digitale Interview-Sammlungen | Universitätsbibliothek | Freie Universität Berlin

Projekt

- **Förderung**: NFDI, 4Memory, Incubator Funds 2024
- **Mitarbeiter**: Tobias Kilgus, Peter Kompiel, Marc Altmann (alle FU Berlin), Christian Horvat (FB Mathematik, FHNW)
- **Umsetzung**: Universitätsbibliothek der FU: Abt. Forschungs- & Publikationsservices mit Universitätsarchiv, Team Digitale Interview-Sammlungen
- **Webseite**: https://www.fu-berlin.de/asr4memory
- Anbindung an die Erschließungs- und Recherche-Plattform "**Oral-History.Digital**“ (OH.D)

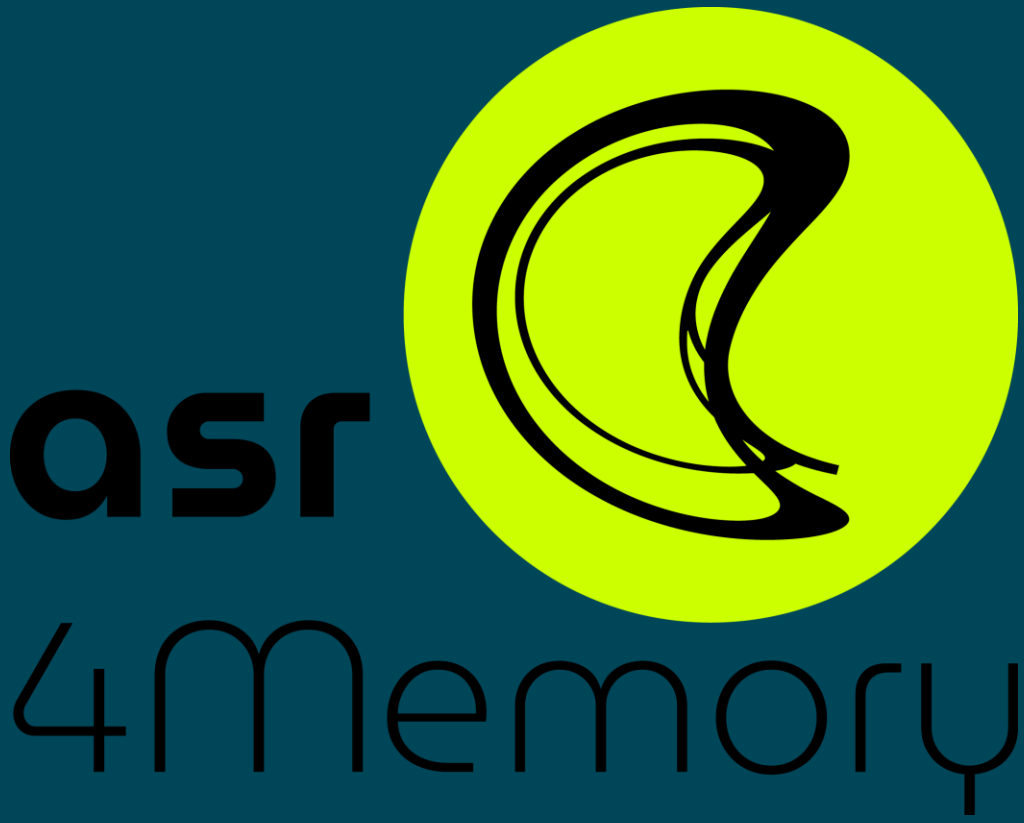
Warum haben wir das Projekt gemacht?



Ausgangslage

- **Umfangreiche Bestände/Sammlungen** von AV-Ressourcen (Oral History) liegen vor oder sind im Entstehen → inhaltlich/ wissenschaftlich zu erschließen → Grundlage ist Verschriftung
- **Bisheriges Vorgehen**:
  - Manuelle Transkription
    ➔ sehr zeitaufwändig und kostenintensiv
  - Kommerzielle Transkriptionsdienste
    ➔ (z.T.) datenschutzproblematisch und kostenintensiv
    ➔ Ergebnisse oft nicht zufriedenstellend mit Blick auf Transkriptionsgenauigkeit und Exportformate

Wie sind wir vorgegangen?



Entwicklung

- **Bedarfsermittlung** in der Community
  → Workshop im März 2024 mit Sammlungsinhaber*innen
  → enge Zusammenarbeit mit Historiker*innen (Oral History)
  → 30 Einrichtungen als Pilotnutzende:

    - Welche audiovisuellen Quellen/Formate liegen vor?
    - Welchen Anforderungen und Schwierigkeiten bestehen?
    - Welche Transkriptformate werden benötigt?
    - Welche Infrastrukturen sind erforderlich?

Entwicklung

- Nutzung von Algorithmen der **Künstlichen Intelligenz** (KI) bei der **Automatischen Spracherkennung** (ASR)
  - **„WhisperX“:** Integration der „Whisper“-Reimplementierung (Universität Oxford) → OpenAI (ChatGPT)

- **Open-Source-basierte Software**
  ➔ lizenzkostenfrei
  ➔ Code flexibel nachnutzbar und anpassbar

Entwicklung

- **Definition der automatischen Spracherkennung** (ASR):
  → Umwandlung gesprochener Sprache in Text

In modern times, we expect more of our automatic systems. The task of **automatic speech recognition** (**ASR**) is to map any waveform like this:

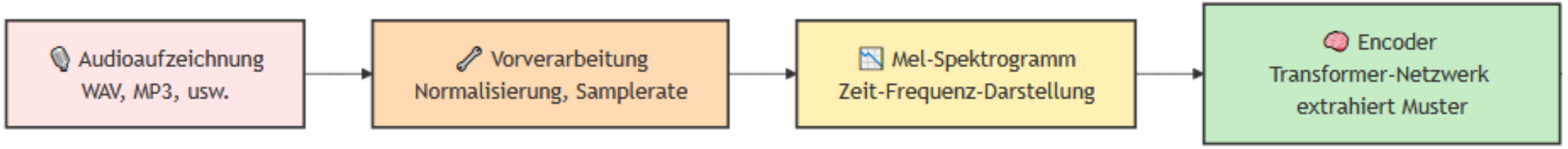
to the appropriate string of words:

```
It's time for lunch!
```

(Jurafsky & Martin, 2023)

Entwicklung

- Funktionsweise der ASR am Beispiel von Whisper:

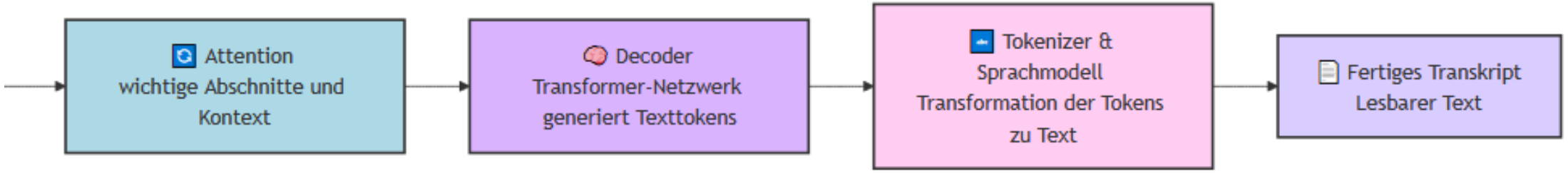


→ INPUT: Eingabe des Audios in die ASR

Vorbereitung des Audio-signals für ASR-Verarbeitung

Umwandlung des Signals in Zeit-Frequenz-Darstellung

Übersetzung des Spektogramms in math. Repräsentation (Embedding), Erkennung von Mustern/Merkmalen



Encoder: Identifikation der Relevanz und des Kontexts von Wörtern und Sätzen

Vorhersage und Bildung von Texteinheiten, die Wörter und Satzzeichen repräsentieren

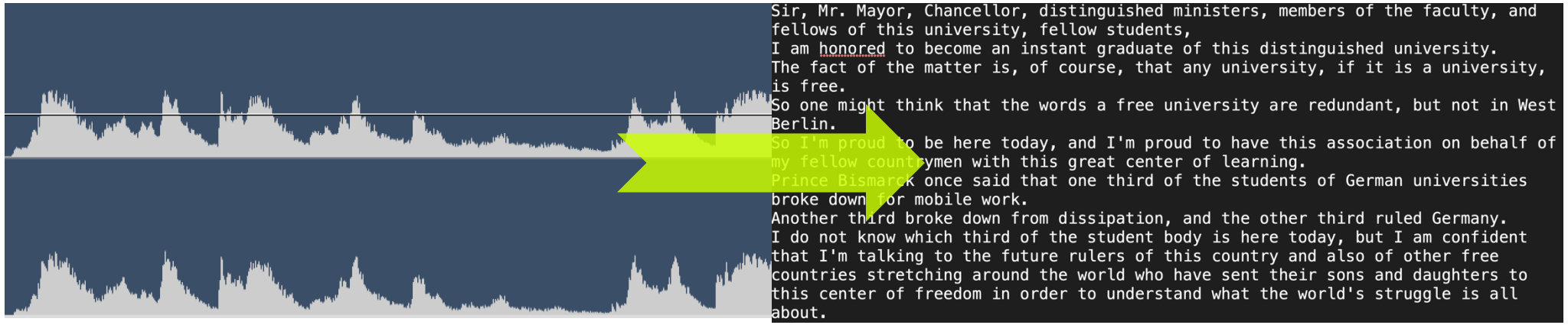
Übersetzung der Einheiten in Wörter, Aufbau der Satzstrukturen

→ OUTPUT: Ausgabe des fertigen Transkripts

Entwicklung

- **Option 1: Browserbasierter Webservice**
  → Transkriptions-Pipeline

Upload der AV-Ressourcen → Prüfung und Optimierung der Tonspur → Automatische Transkription → Generierung der Exportformate → Bereitstellung der Exportformate

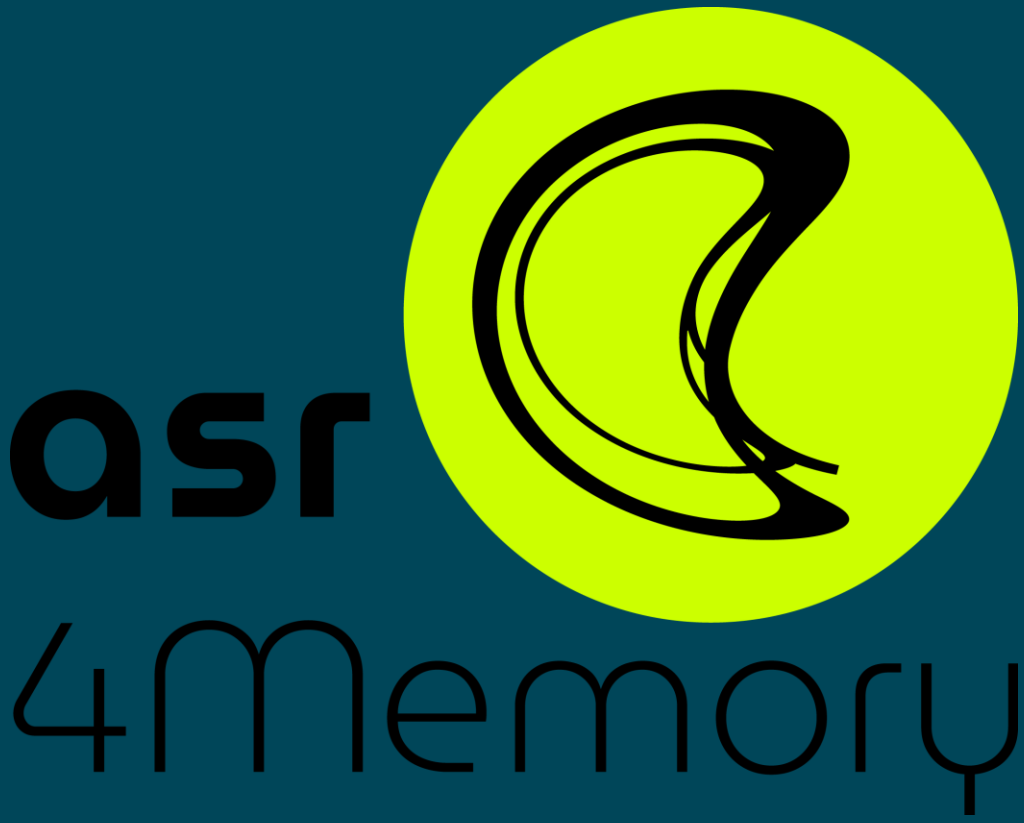


Entwicklung

- **Option 2: Open-Source-Code**
  - ermöglicht lokale Installation der Software
  - Weiterentwicklung und -verteilung in der Community

- **Github-Repositorien**: https://github.com/asr4memory
  - Transkription: https://github.com/asr4memory/asr-transcribe
  - Evaluation: https://github.com/asr4memory/asr-evaluate
  - Tonoptimierung: https://github.com/asr4memory/asr-optimize
  - Upload-Tool: https://github.com/asr4memory/mmt-py

Welche Ergebnisse haben wir erreicht?



Ergebnisse

- Tool als **Web-Service** (in Forschungsszenarien) oder als **Open-Source-Software** (Eigeninstallation) nutzbar

- **Datenschutzkonform**:
  - **Web-Service**: Datenverarbeitung ausschließlich in der Infrastruktur von „Oral-History.Digital“ (Server an der Freien Universität)
  - **Open-Source-Software:** Installation und Betrieb auf eigenem Rechner (keine Clouds, externe Dienste, Datenabflüsse)

Ergebnisse

- **Performance**:
  - schnelle Verarbeitung der Quellen (~ RTF 0,02, Bsp.: 4h in 4 min)
  - qualitativ hochwertige Transkription → niedrige Wortfehlerrate bei verbreiteten Sprachen

- **Diarisierung**: satzbasierte Erkennung und Annotation der Sprecher*innen

- **Segmentlänge**: intelligente, dynamisch anpassbare Begrenzung der Zeichen pro Segment

Ergebnisse

- **Mehrsprachigkeit → Unterstützung von etwa 30 Sprachen**:
  - deutsch, englisch, spanisch, ukrainisch, usw. (Gesamtliste)
  - automatische Sprachdetektion möglich
- **Alignierung → Exakte Zeitkodierung:**
  - Transkripte mit Millisekunden-genauen Zeitmarken
  - ermöglicht ihre Synchronisierung mit AV-Medien
  - führt u.a. zu verbesserter Auffindbarke
  - weitere Nutzungsmöglichkeiten, u.a. in der Textanalyse (NER, TM), Anonymisierung/Pseudonymisierung

	WORD	START	END
2	Ja.	00:00:02.077	00:00:02.398
3	Darf	00:00:02.438	00:00:02.637
4	ich	00:00:02.677	00:00:02.758
5	rein?	00:00:02.959	00:00:03.738
6	Ja.	00:00:03.778	00:00:04.099

Ergebnisse

- **Exportformate**: TXT, CSV, PDF VTT, SRT

```
Sir, Mr. Mayor, Chancellor, distinguished ministers, members of
the faculty, and fellows of this university, fellow students,
I am honored to become an instant graduate of this distinguished
university.
The fact of the matter is, of course, that any university, if it
is a university, is free.
So one might think that the words a free university are
redundant, but not in West Berlin.
So I'm proud to be here today, and I'm proud to have this
association on behalf of my fellow countrymen with this great
center of learning.
Prince Bismarck once said that one third of the students of
German universities broke down for mobile work.
Another third broke down from dissipation, and the other third
ruled Germany.
I do not know which third of the student body is here today, but
I am confident that I'm talking to the future rulers of this
country and also of other free countries stretching around the
world who have sent their sons and daughters to this center of
```

→ zur manuellen Nachbearbeitung in Text-programmen und zur Langzeitsicherung

```
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,
```

→ Untertitelung von AV-Medien in Playern

Ergebnisse

- Exportformate: JSON CSV


```
{
    "start": 1.164,
    "end": 14.503727722772277,
    "sentence": "Sir, Mr. Mayor, Chancellor, distinguished ministers, members of
},
{
    "start": 14.503727722772277,
    "end": 22.721,
    "sentence": "I am honored to become an instant graduate of this distinguishe
},
{
    "start": 24.082,
    "end": 30.727,
    "sentence": "The fact of the matter is, of course, that any university, if it
},
{
    "start": 32.56,
    "end": 39.806,
    "sentence": "So one might think that the words a free university are redundar
},
{
    "start": 41.067,
    "end": 50.454,
    "sentence": "So I'm proud to be here today, and I'm proud to have this associ
```


IN	TRANSCRIPT
00:00:01.164	Sir, Mr. Mayor, Chancellor, distinguished ministers, members of the faculty, and fellows of this university, fe
00:00:14.504	I am honored to become an instant graduate of this distinguished university.
00:00:24.082	The fact of the matter is, of course, that any university, if it is a university, is free.
00:00:32.560	So one might think that the words a free university are redundant, but not in West Berlin.
00:00:41.067	So I'm proud to be here today, and I'm proud to have this association on behalf of my fellow countrymen w
00:00:52.256	Prince Bismarck once said that one third of the students of German universities broke down for mobile wo
00:01:01.570	Another third broke down from dissipation, and the other third ruled Germany.
00:01:08.713	I do not know which third of the student body is here today, but I am confident that I'm talking to the future
00:01:35.037	I know that when you leave this school, you will not imagine that this institution was founded by citizens of
00:01:48.189	and was developed by citizens of West Berlin, that you will not imagine that these men who teach you hav
00:02:10.448	This school is not interested in turning out merely corporation lawyers or skilled accountants.
00:02:18.774	What it is interested in, and this must be true of every university, It must be interested in turning out citizen
00:02:28.925	men who comprehend the difficult, sensitive tasks that lie before us as free men and women, and men wh
00:02:44.141	That's why you're here, and that's why this school was founded, and all of us benefit from it.
00:03:08.126	It is a fact that in my own country, in the American Revolution, that revolution and the society developed th

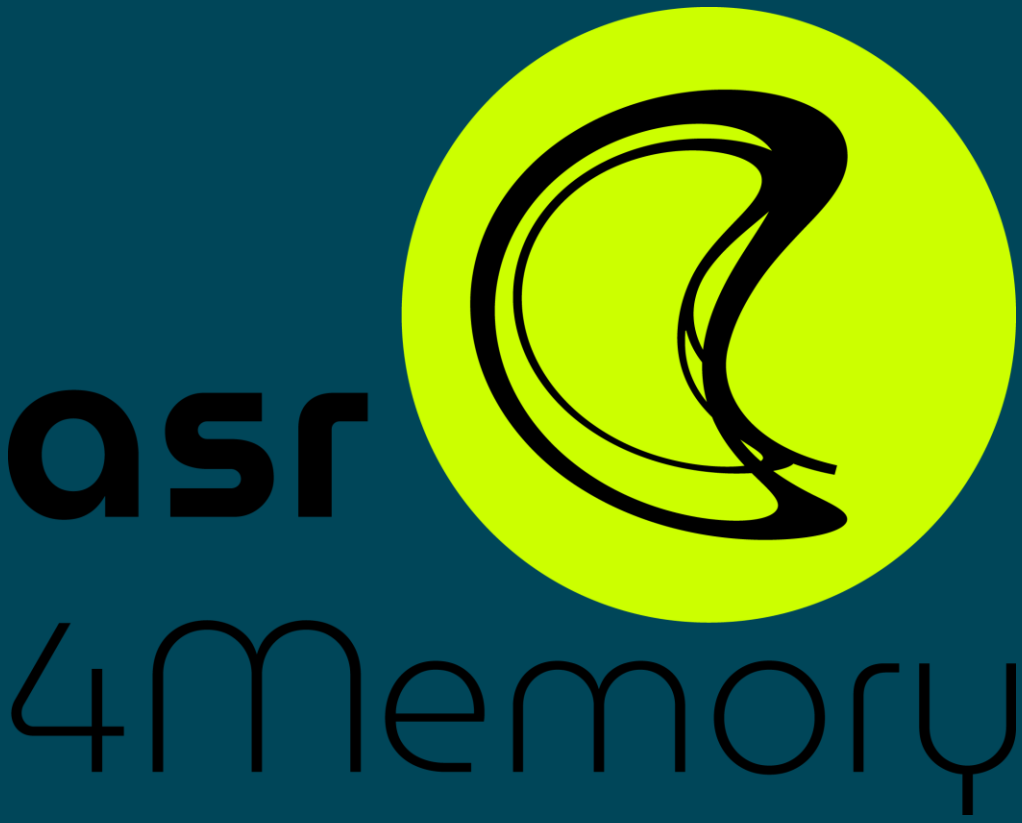
→ zur automat. Datenverarbeitung in Systemen oder Anwendungen (z.B. Information Retrieval)

➔ Weitere Formate: TEI-XML (in Arbeit), IIIF-AV, CMDI, EAD etc. ?

Ergebnisse

- **Beispiel John F. Kennedy 1963 in Berlin**: https://www.cedis.fu-berlin.de/services/medien/av-medien/test/kennedy-rede/index.html
- **GitHub-Repositorien:** https://github.com/asr4memory
- **Media Management Tool (MMT)** https://mmt.oral-history.digital/
- **Projekt-Webseite:** https://www.fu-berlin.de/asr4memory

Welche Schwächen hat die ASR?



Schwächen der ASR

- **Halluzinationen**:
  - Erzeugung nicht gesprochener Inhalte, aus den Trainingsdaten stammend
- **Entitäten-Erkennung**:
  - Falscherkennung von (historischen/aktuellen) Personennamen, Organisationen, Orten, Ereignissen
  - Beispiel:
    - Original: „Und der *Genscher* hat sie dann da rausgekloppt.“
    - ASR: „Und der *Kenja* hat sie dann da rausgeklappt.“
- **Erkennung von Sprecher*innen**: Probleme bei
  - schnellen Sprecherwechseln
  - parallelem Sprechen, vielen Sprecher*innen

Schwächen der ASR

- **Glättungen des Transkripts**
  - Häufig ausgelassen werden:
    - **Füllwörter, Bindewörter** (oft am Satzanfang/-ende)
    - **Wiederholungen, Satzabbrüche, Verzögerungen**

- Beispiel:
  - Original: „*Und* der hat uns da, *äh*, ein paar Sachen, *ähm*, Schriftstücke gezeigt, die, *die*, *die* also eigentlich vollkommen belanglos waren.“
  - ASR: „Der hat uns da ein paar Sachen, Schriftstücke gezeigt, die also eigentlich vollkommen belanglos waren.“

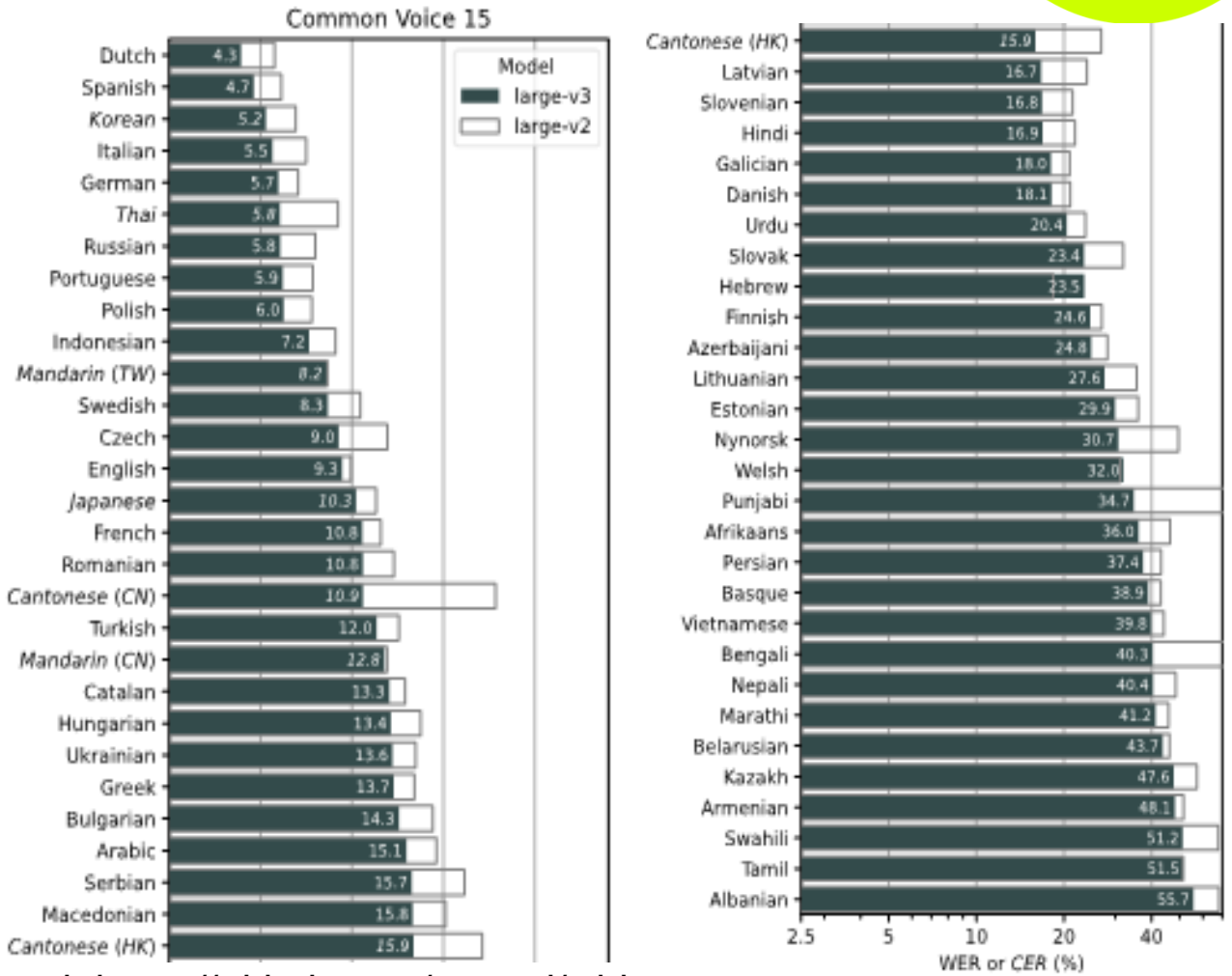
Schwächen der ASR

- **Glättungen des Transkripts**
  - Nicht transkribiert werden:
    - non-verbale Lautäußerungen, Pausen, direkte Rede
  - Grammatikalische Korrekturen durch ASR (Satzbau)
  - Dialekte und Akzente werden ins Hochdeutsche transformiert (öfters fehlerbehaftet)
    - Beispiele:
      - „ick“ → „ich“; „jewesen“ → „gewesen“; „kriech‘ ta“ → „kriegte er“

→ keine „wortgetreue“ Transkription: Optimierte Trainingsdaten, Modell der Wortvorhersage → jedoch gutes „Rohtranskript“

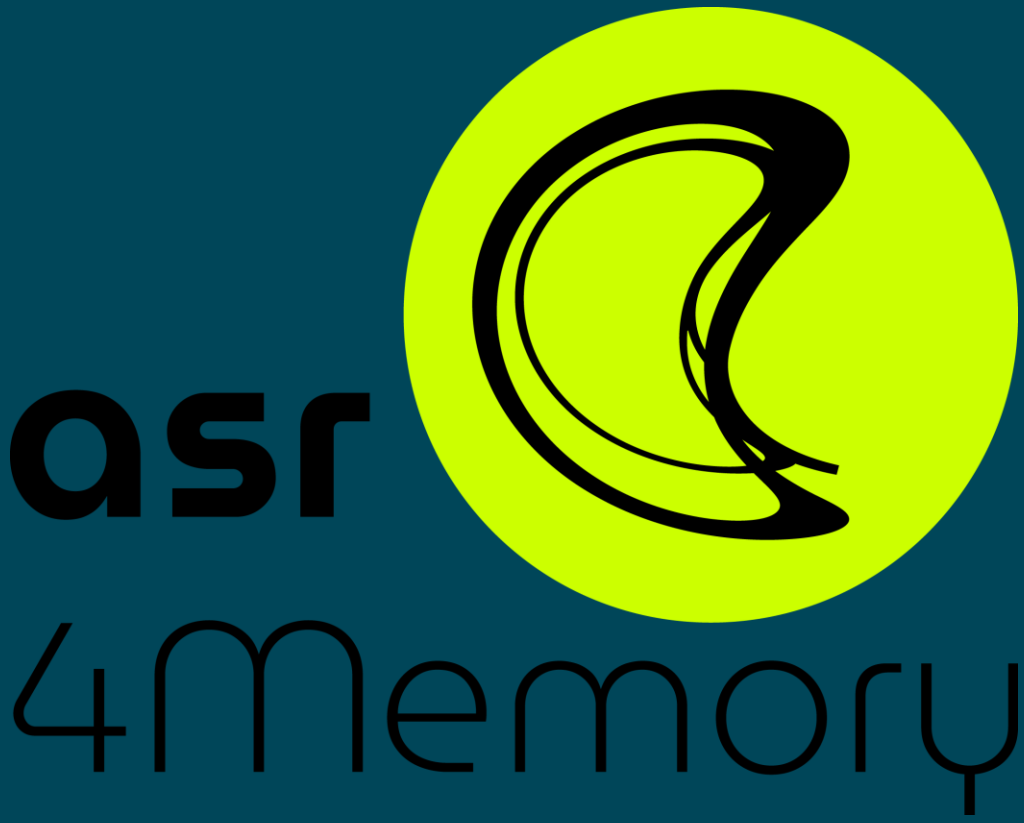
Schwächen der ASR

- **Audioqualität**:
  - Je schlechter die Aufzeichnung der gesprochenen Sprache (Lautstärke, Störgeräusche), desto fehleranfälliger auch die ASR.
- **Sprachen-Unterstützung**:
  - Für das Training von Whisper wurden 680.000 Stunden Audiomaterial mit z.T. Referenztranskripten aus dem Internet verwendet, ~ 100 Sprachen, überwiegend englisch
- **Mehrsprachigkeit**:
  - Whisper kann nur eine Sprache pro Quelle durchgehend verarbeiten



Vgl. https://github.com/openai/whisper

Wie geht es weiter?



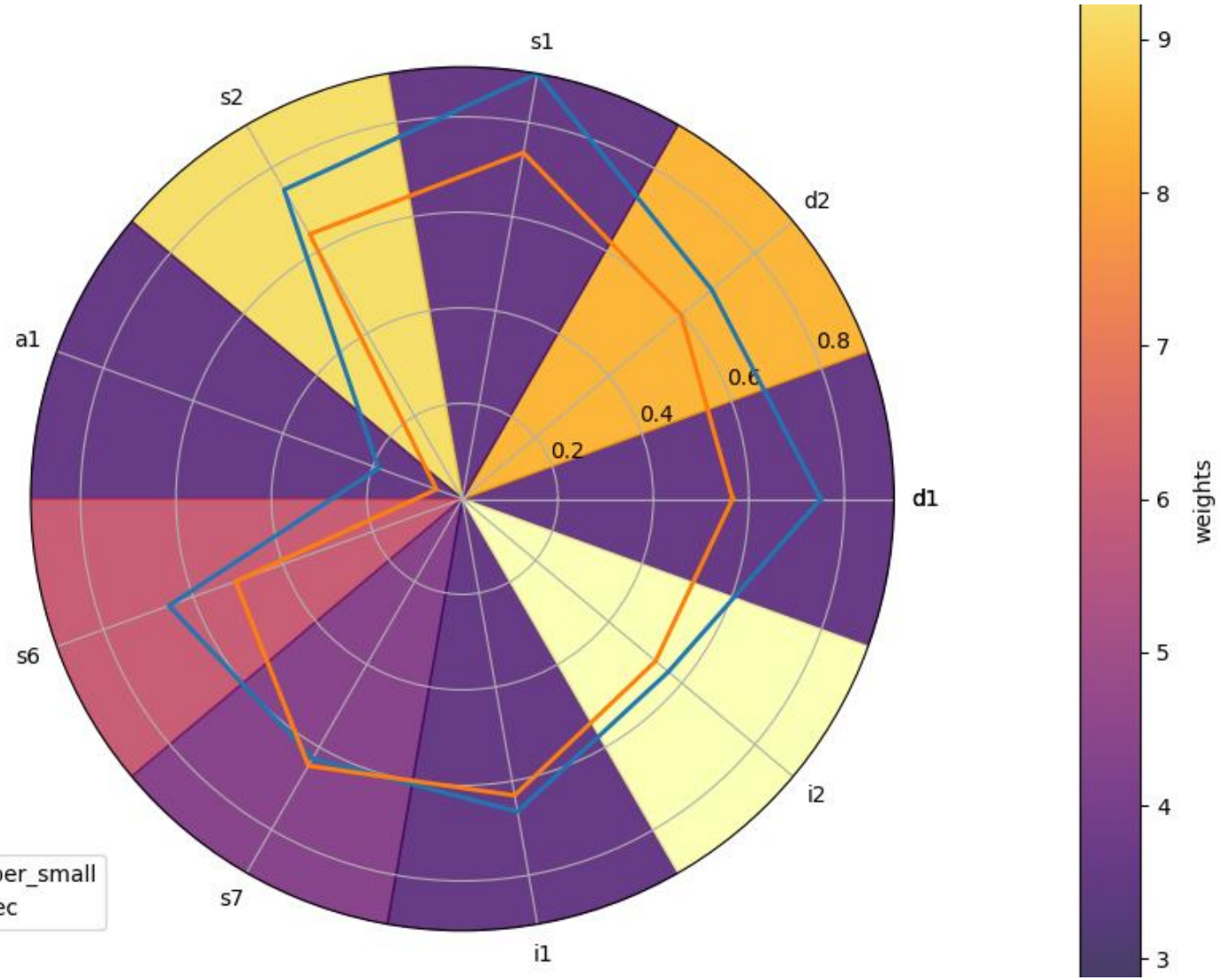
Perspektive

- **Schrittweise Integration der ASR-Funktionalität in OH.D**
  - Start der Pilotphase für ASR-Nutzung ab 2025
- **Weiterentwicklung** der Transkriptions-Pipeline
  - Ausbau der Hardware, z.B. weitere GPUs
  - weitere Austauschformate wie TEI/XML ermöglichen
  - Glättungen und Halluzinationen reduzieren
  - Sprecherauszeichnungen verbessern
  - Mehrsprachigkeit in den AV-Ressourcen unterstützen
  - Anbindung an KI-Komponenten wie NER, Topic Modelling

Perspektive

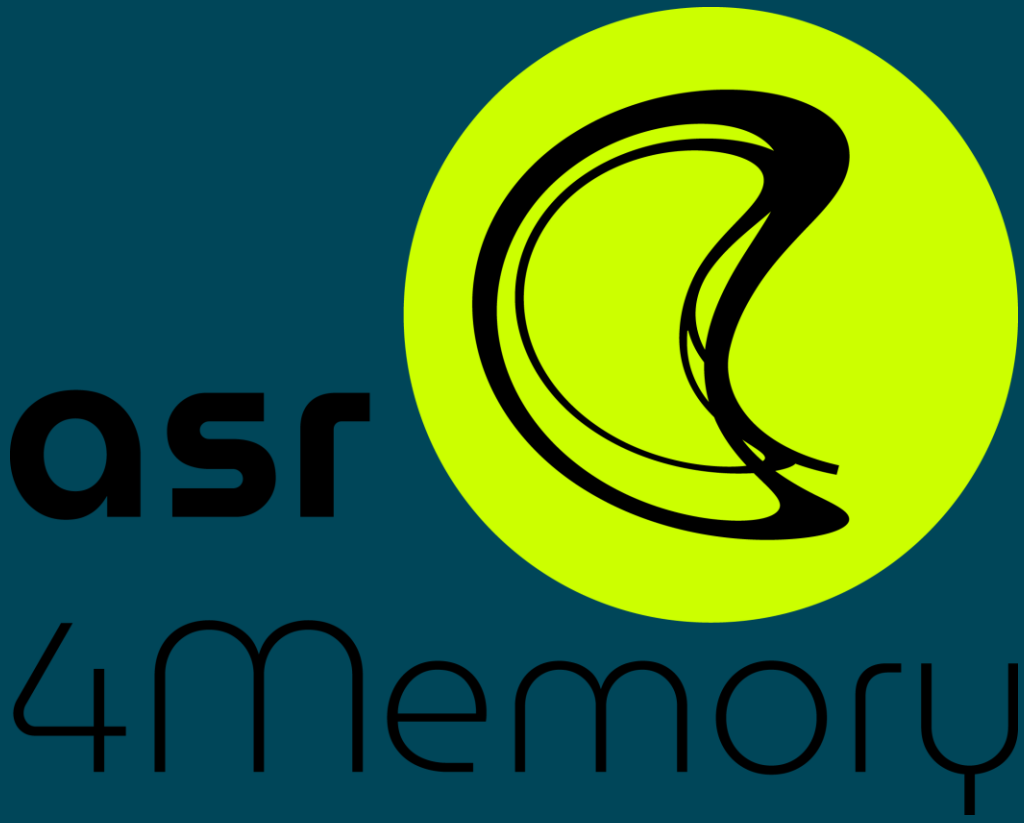
- **Anpassung** der Sprachmodelle:
  - Training und Feintuning von Whisper auf HPC-Cluster
  - ➔ Ziel: domänen-spezifisches ASR-Modell
  - Trainingsdaten: 90 deutschsprachige Oral-History-Interviews, 300 Stunden, etwa 190.000 Segmente/Sätze, optimal aufbereitete Referenztranskripte
  - Zwischenergebnis: Sowohl quantitativ als auch qualitativ bessere Transkriptqualität (weniger schwerwiegende Fehler)

Perspektive



Perspektive

Welche Bedarfe und Anforderungen liegen vor?



Rückmeldung

- **Bedarfe?**
  - Welche und wie viele AV-Ressourcen könnten/sollten transkribiert werden?
- **Feedback!**
  - Wichtig ist für uns die Rückmeldung:
    - welche Fehler oder Auffälligkeiten bei der ASR auftreten,
    - welche Transkriptformate für die Nachnutzung wichtig sind,
    - welche weiteren Funktionen gewünscht werden,
    - etc.

Vielen Dank!