Transkription von Sprachaufnahmen

Julian Sartorio, 06.06.2025

A. Beschreibung des Anwendungsfalls

Ziel des Use-Cases ist es, die Fähigkeit und Nutzbarkeit von KI-Anwendungen oder anderer Software zur Transkription von Sprachaufnahmen, also die möglichst wortgetreue Verschriftlichung, zu evaluieren. Als Beispiele seien etwa Aufnahmen einer Lehrveranstaltung oder eines Fachvortrages genannt.

B. Vorgehen

1) Umwandlung von Video-/Tonaufnahmen in WAVE-Format (WAV), welches von allen gängigen Transkriptionsanwendungen gelesen werden kann
2) Auswahl (*Offline-Anwendung*) oder Upload (*Online-Anwendung*) in der jeweiligen Transkriptionsanwendung

C. Test

I. Testszenario

1) Auswahl eines Vorlesungsvideos von Herrn Prof. Dr. Georg Borges (Zivilrechtliche Grundlagen des IT-Rechts)
2) Herausforderungen des Testszenarios

- Tonqualität unter Vorlesungsbedingungen: Hintergrundgeräusche, Hall, niedrige Mikrofonqualität, entfernte Sprecher, mehrere Personen sprechen gleichzeitig oder übereinander
- Schnelles, leises oder undeutliches Sprechen erschwert die Verarbeitung durch KI
- Wechsel zwischen Sprachen und Sprechern verringert die Erkennungsleistung
- Ausgangsdateien: Videos, bei den häufigsten Anwendungen Extraktion der Tonspur erforderlich

II. Ergebnis

Gemini	noScribe
„Ach, so meine Damen und Herren online und vor Ort im letzte Woche hatten wir den paragraph 3 der Vorleistung verwenden können, wir kommen nun zu paragraph 4 Vertragsabschluss per Internet ein wunderschöner Wiederholung und Vertiefung von allgemeinen Stoff des Allgemeinen Teils allgemeines schönste das BMW, außer es haben es Vorbereitung Adventsfest oder ein bisschen spezifische Vertiefung allgemeiner Rechtsfragen von einer gewissen praktischen Bedeutung, ich möchte nur sagen, dass wir in das vorlesungs ein wenig möglicherweise modifizieren werden, ich habe mir überlegt, dass ich obwohl es die Bundesländern aufgebaut habe. Kam mir noch eine Idee, was man noch attraktiver und spannender machen könnte und deswegen während die Gliederung eventuell noch mal anderen ändern zu wissen.	„Meine Damen und Herren, online und vor Ort, letzte Woche hatten wir den Paragraphen 3 der Vorleistung beenden können. Wir kommen nun zu Paragraphen 4, Vertragsabschluss per Internet, eine wunderschöne Wiederholung und Vertiefung von allgemeinem Stoff des Allgemeinheits- und des Allgemeinen Schulzrechts des BNB. Also Examsvorbereitung, NLS-Best, nur ein bisschen ideisch-spezifische Vertiefung allgemeiner Rechtsfragen von einer gewissen praktischen Bedeutung. Ich möchte nur sagen, dass wir das Vorlesungsprogramm ein wenig modifizieren werden. Ich habe mir überlegt, dass ich, obwohl ich die Vorlesung jetzt völlig neu aufgebaut habe, kam mir noch eine Idee, was man noch attraktiver und spannender machen könnte. Und deswegen werden die Gliederung eventuell nochmal ändern. (.)

Nächste Woche findet die Vorlesung nicht statt, wir machen ja immer eine Viertelstunde Zusätze im zwei Termine dadurch aufzuholen. Einer davon wird nächste Woche sein, da bin ich anderen Orts zum Vortrag und kann deswegen nicht hier sein.“	Sie wissen, nächste Woche findet die Vorlesung nicht statt. Wir machen ja immer eine Viertelstunde Zusätze, um zwei Termine dadurch aufzuholen. Eine davon wird nächste Woche sein. Da bin ich andernorts zum Vortrag und kann deswegen nicht hier sein. (..)“

D. Bewertung

1. ChatGPT

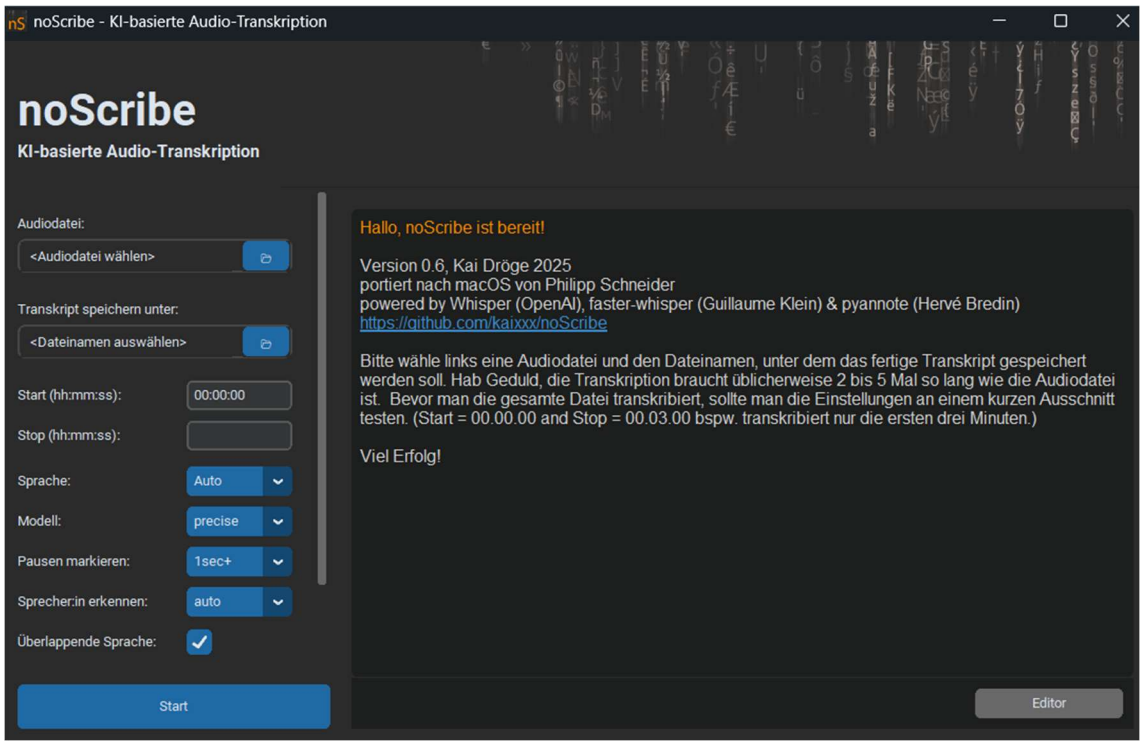
- ChatGPT kann keine Videos oder Audiodateien sofort verarbeiten. Es sind Umwege über verschiedene als GPT integrierte Plattformen (z.B. Whisper Transcriber, Video Insights, TurboScribe) notwendig.
- Dabei müssen die zu transkribierenden Dateien teilweise auf den Plattformen hochgeladen werden, um dann auf diese via ChatGPT zuzugreifen. Dies ist je nach Plattform und Dateiengröße kostenpflichtig.
- Teilweise können diese direkt in ChatGPT hochgeladen werden (z.B. bei Whisper), wobei es aber bei größeren Datenmengen Probleme gibt. Allgemein liegt das Limit bei 512MB, aber auch darunter treten Fehler wie „low memory“ oder Verbindungsfehler auf. Die Verfügbare Leistung seitens OpenAI ist folglich zu gering, um größere Aufzeichnungen zu transkribieren. Das LeistunWhisgsproblem könnte durch eine kleinteiligere Eingabe, also 30 bis 60 Sekunden, gelöst werden, was aber offensichtlich ungeeignet ist.

2. Weitere (kostenpflichtige) Anwendungen

- Whisper (kostenlos): Das auch von NoScribe genutzte Tool Whisper von OpenAI kann auch isoliert über Python genutzt werden, erfordert aber eine wesentlich umständlichere Einrichtung. Zugleich erfolgt im Rahmen von NoScribe eine Zusammenführung mit Faster-Whisperer und Pyannote, die eine schnellere Bearbeitung sowie Sprechererkennung ermöglich. Zugleich ist es nicht erforderlich, die jeweiligen Daten hochzuladen.
- Dragon von Nuance (kostenpflichtig): Professionelle Transkriptions-Anwendung, die auch auf juristische Sprache (Dragon Legal) trainiert ist.

- Gemini 2.0 Flash (kostenpflichtig): Upload von größeren Dateien möglich.
- Microsoft Sharepoint: Kann automatisch Untertitel und Transkriptionsdateien erstellen. Funktioniert schnell und unkompliziert, hat dafür aber eine geringere Genauigkeit.

3. NoScribe



- Graphische Benutzeroberfläche (engl. GUI) für Whisper (OpenAI), Faster-Whisper (Guilaume Klein) und Pyannote (Hervé Bredin)
    - Whisper: Open-Source-Spracherkennungssystem von OpenAI
        - Transkription gesprochener Sprache in Text
        - Unterstützung vieler verschiedene Sprachen
    - Faster-Whisperer: Neuimplementierung des Whisper-Modells von OpenAI mit CTranslate2, die bei gleicher Genauigkeit bis zu 4-mal schneller als Whisper ist und hierbei weniger Speicherplatz benötigt

Implementation	Precision	Beam size	Time	VRAM Usage
openai/whisper	fp16	5	2m23s	4708MB
whisper.cpp (Flash Attention)	fp16	5	1m05s	4127MB
transformers (SDPA)[1]	fp16	5	1m52s	4960MB
faster-whisper	fp16	5	1m03s	4525MB
faster-whisper (`batch_size=8`)	fp16	5	17s	6090MB
faster-whisper	int8	5	59s	2926MB
faster-whisper (`batch_size=8`)	int8	5	16s	4500MB

  - Pyannote: Erkennung von Sprachaktivität, Sprecherwechsel, überlappender Sprache, sowie Einbettung von Sprechern

- kostenlos und quelloffen (GPL-3.0)
- Abhängigkeit von der lokalen Rechenleistung: Dauer der Transkription kann je nach eingestellter Präzision des Programmes und Rechenleistung bis zu mehreren Stunden dauern, ist jedoch nicht limitert. ("*It runs **completely local** on your computer. No data is sent to the internet. No cloud, no worries*")

E. Fazit

Die Anwendung NoScribe ist zur Transkription von Sprachaufnahmen geeignet und liefert unter den verglichenen Anwendungen (Gemini, Transkiptionssoftware des Microsoft Sharepoint, TurboScribe, Whisper) die besten Ergebnisse.