audilu: Eine App zur Erstellung von Audio-Deskriptionen für Kurzvideos

Casey Kreer

Dresden, den 22. September 2023

Contents

Einleitung

Im Folgenden präsentiere ich "audilu", eine innovative App zur Erstellung von Audio-Deskriptionen (AD) für Kurzvideos in sozialen Netzwerken. Die Entwicklung eines Prototypen wurde im Rahmen des Prototype Fund durch das Bundesministerium für Bildung und Forschung (BMBF) gefördert. Der Quellcode des Prototyps ist unter einer freien Lizenz auf https://audilu.de verfügbar. Ich beschreibe in diesem Whitepaper die Motivation hinter dem Projekt, dessen technische Umsetzung und einige Insights und Erkenntnisse aus der Entwicklung und User-Tests.

Audilu kombiniert einige bereits bestehende Techniken und Prinzipien, um bei der Erstellung von qualitativ hochwertiger AD für ein blindes oder sehbehindertes Publikum zu unterstützen. Die App selbst richtet sich hauptsächlich an sehende Content Creator, die ihre Videos barrierefreier gestalten und diese mit AD versehen möchten. Insbesondere sollen allgemeine Nachrichtenmedien und Privatpersonen mit sehbehinderten Kontakten angesprochen werden.

Motivation

In sozialen Netzwerken und auf Messengern werden täglich Millionen von Videos hochgeladen. Die überwiegende Mehrheit dieser Videos sind jedoch für Menschen mit Behinderungen nicht zugänglich, da sie keine Untertitel oder Audiodeskriptionen respektive Textbeschreibungen enthalten. Dies bedeutet eine substantielle Einschränkung der Teilhabe von Menschen mit Behinderungen an der Gesellschaft, insbesondere vor dem Hintergrund, dass Online-Video auch in politischen und wirtschaftlichen Zusammenhängen immer relevanter wird.

In den letzten Jahren ist die Verwendung von Untertiteln in sozialen Netzwerken stark angestiegen. Dies ist hauptsächlich auf die Entwicklung von präziser Spracherkennungs-Software zurückzuführen, die es ermöglicht, Untertitel und Closed Captions vollständig automatisch zu generieren und in Videos einzubetten, ohne dass dafür langwierige manuelle Nachbearbeitung durch die Creators notwendig ist. Zusätzlich dazu lässt sich beobachten, dass stark stilisierte Closed Captions, die beispielsweise intensive Text-Effekte enthalten, auch als Erkennungsmerkmal eingesetzt werden, um die Aufmerksamkeit des Publikums länger zu halten. Diese Entwicklung betrifft sämtliche sozialen Netzwerke, die Video-Content anbieten, wie beispielsweise YouTube, Instagram und TikTok, und ist sehr positiv zu bewerten, da die Barrierefreiheit somit gesellschaftlich normalisiert wird. Abgesehen davon, dass diese Gestaltung der Untertitel selbst nicht barrierefrei ist, ermöglichen sie dennoch in einem ersten Ansatz für eine kleine Gruppe von Menschen den Zugang.

Ebenso stark angestiegen ist die Verwendung von Alternativtexten. Diese können von Screenreadern vorgelesen werden und ermöglichen es blinden und sehbehinderten Menschen, die Inhalte von Fotos zu verstehen. Die sozialen Netzwerke verfügen inzwischen häufig über eine Funktion, die es ihren Nutzer:innen ermöglicht, Alternativtexte für ihre Fotos zu verfassen. Einige Plattformen experimentieren auch mit Funktionen, die es ermöglichen, Alternativtexte mittels Machine Learning-Modellen automatisch zu generieren. Auch die Microsoft Office-Suite bietet eine solche Funktion an.

Abseits dessen sind Videos in sozialen Netzwerken jedoch noch immer nicht für blinde und sehbehinderte Menschen zugänglich. Audio-Deskription als Konzept wird zuweilen hauptsächlich bei Filmen und Fernsehserien implementiert, die für ein sehr breites Publikum produziert und oft auch noch für viele Jahre relevant bleiben werden. AD für Low-Budget-Produktionen oder auch Inhalte in sozialen Medien ist quasi nicht existent. Dies ist vor allem dem Umstand geschuldet, dass die Erstellung von AD hochspezialisiertes Personal erfordert, das über ein tiefes Verständnis der Inhalte sowie der Kultur und den Erlebnissen des Publikums verfügt. Dieses Personal ist jedoch sehr selten und somit sehr teuer, weshalb eine Audiodeskription oft nicht nur monetär nicht leistbar ist, sondern auch inhaltlich nicht umsetzbar.

Hier setzt “audilu” an. Ziel ist es, Content Creators ein Werkzeug an die Hand zu geben, das sie beim Prozess der Erstellung von Audio-Deskription unterstützt

und einen großen Teil des Prozesses vereinfacht oder gar vollständig automatisiert, ähnlich wie es auch bei Untertiteln der Fall ist. Die Hoffnung ist, dass dadurch die Hürden für die Erstellung und das Teilen von barrierefreien Videos gesenkt und somit mehr gesellschaftliche Teilhabe für blinde und sehbehinderte Menschen ermöglicht werden kann.

Wie funktioniert Audio-Deskription?

In einer Audio-Deskription werden die visuellen Inhalte eines Videos in gesprochener Sprache beschrieben. Die Beschreibungen werden in den Pausen zwischen Dialogen und markanten Geräuschen eingesetzt und ermöglichen somit eine weitestgehend parallele Wahrnehmung von Bild und Ton für blinde und sehbehinderte Menschen. Zusätzlich wird gewährleistet, dass das Video in der gleichen Zeit konsumiert werden kann wie von sehenden Menschen, und auch mit diesen gemeinsam. Bei AD handelt es sich also um eine Art Übersetzung von visuellen Inhalten in gesprochene Sprache.

Dies ist nicht immer einfach, da die visuellen Inhalte eines Videos sehr komplex sein können. Es gibt jedoch einige Prinzipien, die bei der Erstellung von AD hilfreich sind:

- **Priorisierung:** Nicht jedes Detail eines Videos muss beschrieben werden. Es ist wichtig, die wesentlichen Informationen zu priorisieren und sich darauf zu fokussieren, was für die Handlung oder den Kontext des Videos am relevantesten ist.
- **Objektivität:** Die Beschreibungen sollten neutral gehalten und ohne persönliche Interpretation verfasst werden. Ziel ist es, dem Publikum ein klares Bild zu vermitteln, ohne es in eine bestimmte Richtung zu lenken. Dies bleibt weiterhin Aufgabe des restlichen Videos, also der Dialoge, des Tons und der visuellen Inhalte.
- **Kürze und Klarheit:** Da AD in die natürlichen Pausen eines Videos eingefügt wird, sollte sie so prägnant und klar wie möglich sein. Überflüssige Worte können die Beschreibung verlängern und den Fluss des Videos stören.
- **Anschauliche Sprache:** Worte, die klare Bilder im Kopf erzeugen, sind besonders hilfreich. Zum Beispiel ist "eine Frau in einem roten Kleid" aussagekräftiger als "eine Person in Kleidung".
- **Kontinuität:** Es ist wichtig, die Beschreibung konsistent zu halten, insbesondere wenn es sich um wiederkehrende Charaktere oder Elemente in einem Video handelt.

Im Erstellungs-Prozess von AD müssen somit einige Entscheidungen getroffen werden, die sich maßgeblich auf die Qualität des Ergebnisses auswirken.

Wie funktioniert audilu?

Audilu unterstützt diese Entscheidungsfindung durch einige, teils sehr simple Funktionen.

Es hat sich in der Projektphase früh gezeigt, dass Nutzende oft gar nicht bereit sind, die Software überhaupt zu nutzen, wenn dies einen großen zeitlichen Mehraufwand für sie bedeuten würde. Das Design-Ziel war also, die Entwicklung einer vollständigen Audiodeskription in 200% der Laufzeit des Videos zu ermöglichen, also bei einem 60-sekündigen Video innerhalb von 2 Minuten. Dies bedeutete für die Entwicklung einen verstärkten Fokus auf die vollautomatische Bildbeschreibung durch Mittel der künstlichen Intelligenz.

Im Rahmen des Projekts wurde intensiv ein User Experience-Konzept getestet, um diese Technologien zum Zweck der Erstellung von Audiodeskription miteinander zu kombinieren. Das Projekt wurde dazu bei der Projekt-Messe OUTPUT.DD Dresden präsentiert, um potentielle Nutzende über die Möglichkeiten von Audio-Deskription aufzuklären. Dort hat sich unter anderem herausgestellt, dass nur etwa 5 Prozent des Laufpublikums überhaupt wussten, was eine Audiodeskription ist. Ein zusätzlicher Fokus für die Entwicklung war also, sämtliche Erklärtexte soweit anzupassen, dass auch ohne dieses Wissen qualitative Ergebnisse erzielt werden konnten.

Sprechpausen-Erkennung

Durch die Integration von Spracherkennung, die auf die Tonspur des Videos angewendet wird, können Sprechpausen automatisch identifiziert werden, die sich für die Einbettung von AD eignen. Die Faustregel ist hier, dass eine Pause mindestens 1,5 Sekunden lang sein sollte, um ausreichend Zeit für die Beschreibung zu bieten. Selbst bei einer sehr schnell gesprochenen Beschreibung ist dies die untere Grenze, um das Publikum nicht zu überfordern. Die Grund-Annahme ist, dass die durchschnittliche Sprech-Geschwindigkeit in der deutschen Sprache ca. 90 bis 120 Wörter pro Minute beträgt. In diesen sehr kurzen Sprechpausen können somit maximal 3 Wörter untergebracht werden. Audilu berechnet diese Richtwerte und zeigt sie pro Sprechpause an.

Während des User-Testings hat sich gezeigt, dass diese Anzeige sehr hilfreich für die Content Creators ist, da sie ihnen eine greifbare Orientierung bietet, wie viel Platz sie tatsächlich für die Beschreibung haben und wie sie **priorisieren** können, welche Elemente sie beschreiben. Gleichermaßen führte das aber auch dazu, dass Beschreibungen größtenteils nur noch als Stichpunkte verfasst wurden, um die Sprechpausen nicht zu überziehen. Dies ist jedoch nicht immer hilfreich, da die Beschreibung dadurch sehr abgehackt und unflüssig klingen kann.

Diese Nachteile konnten abgemildert werden, indem statt einer fixen Wortanzahl ein Wertebereich angezeigt wird, der knapp unter bis knapp über der durchschnittlichen Sprechgeschwindigkeit liegt. Dies ermöglicht es, die Beschreibung etwas flüssiger zu gestalten, ohne dabei die Sprechpause zu überziehen. Viele Wörter, die Sätze flüssig klingen lassen, sind sehr kurz und fallen deshalb nicht sehr ins Gewicht.

Eine vollständige Spracherkennung wurde mittels des Machine Learning-Modells *Whisper* von OpenAI getestet, jedoch aufgrund der hohen Kosten

und der langen Bearbeitungszeit der API nicht in den Prototypen integriert. Stattdessen wird nun das Paket https://github.com/ricky0123/vad von Ricky Samore genutzt, welches das Modell *Silero VAD* von Alexander Veysov (https://github.com/snakers4/silero-vad) verwendet. Dieses Voice Activity Detection (VAD) Modell ist sehr schnell und in einer Vielzahl von Sprachen ausreichend präzise. Es kann in unter Echtzeit auf eine Audiodatei angewendet werden und erkennt Zeitabschnitte, in denen eine menschliche Stimme zu hören ist, und gibt diese als Liste von Start- und End-Zeitpunkten zurück. Diese Liste wird anschließend invertiert, um die Sprechpausen zu erhalten. Direkt in diesem Schritt werden auch alle Sprechpausen aus der Liste entfernt, die kürzer als 1,5 Sekunden sind.

Das Modell selbst lässt in seinen Zeitabschnitten vor und nach der erkannten Stimme etwas Platz. Diese werden von Audilu direkt übernommen, um zu gewährleisten, dass die Beschreibung nicht zu nah an den Dialogen liegt und somit das Publikum nicht zu überfordern.

Die zusätzliche Spracherkennung von *Whisper* bot den Vorteil, dass die Generierung von **Beschreibungsvorschlägen** mit zusätzlichem Kontext über das Video versorgt werden konnte. Dies sorgte für eine höhere Qualität der Vorschläge, da diese nicht nur auf den abstrakten Bildhinhalten der Sprechpause basierten, sondern direkt auf das Video angepasst werden konnten und passende Tags vom verwendeten Sprachmodell ausgewählt werden konnten.

Text-to-Speech und Sprachaufnahmen

In einigen Situationen kann es hilfreich sein, die Beschreibung nicht selbst zu sprechen, sondern von einer Text-to-Speech-Engine (TTS) generieren zu lassen, beispielsweise wenn die Stimme des Content Creators selbst im Video zu hören und es unmöglich ist, die Beschreibung von einer anderen Person einsprechen zu lassen. Somit kann eine klare Trennung zwischen Video-Content und AD geschaffen und die **Klarheit** verbessert werden.

Wichtig ist es zu beachten, dass das Publikum die dafür genutzten Stimmen oft als künstlich erkennt und dass damit häufig einhergeht, dass die Qualität der AD als minderwertig wahrgenommen wird. Content Creators sollten sich also bewusst sein, dass die Verwendung von Text-to-Speech nicht immer die beste Option ist und dass hier auch die Authentizität ihres Inhalts leiden kann. Auch der Eindruck von **Objektivität** kann durch die Verwendung von TTS beeinträchtigt werden, da die Stimmen oft sehr neutral und emotionslos klingen.

In der Praxis zeigt sich aber auch, dass die Qualität von Audio-Deskriptionen deutlich besser ist, wenn diese zuvor getippt und auf die maximale Wortanzahl hin optimiert wurden, die dabei automatisch geprüft und mittels einer Fortschrittsanzeige live dargestellt werden kann. Audilu lässt deshalb das Textfeld und die Fortschrittsanzeige immer eingeblendet. Die Nutzer:innen sollen dennoch dazu ermutigt werden, wenn möglich, die Beschreibung selbst einzusprechen, auch wenn sie diese bereits getippt haben.

Automatische Geschwindigkeitsanpassung

Egal ob die Beschreibung selbst gesprochen oder von einer TTS-Engine generiert wird, es ist wichtig, dass die Geschwindigkeit der Beschreibung für das Publikum angemessen ist. Grundsätzlich gilt, dass die Beschreibung nicht zu schnell gesprochen werden sollte, um das Publikum nicht zu überfordern. Gleichzeitig sollte sie aber auch nicht zu langsam sein. Idealerweise passt sie sich an den Rythmus der bestehenden Tonspur an, um den Fluss des Videos nicht zu stören.

Im User-Testing konnte beobachtet werden, dass sich ungeübte Content Creators oft sehr schwer damit tun, die Geschwindigkeit der Beschreibung selbst einzuschätzen. In den meisten Fällen sprechen sie etwas zu langsam und treffen somit nicht die natürlichen Pausen des Inhalts. Audilu bietet die Möglichkeit, die Geschwindigkeit der Beschreibung automatisch zu beschleunigen oder zu verlangsamen, um die vorgesehene Pausenzeit besser zu treffen. Dies ist jedoch nur eine Notlösung, da die erzeugten Audios dadurch oft sehr unnatürlich und **inkonsistent** klingen. Die Standardwerte für die Verschnellerung sind deshalb 0,8-fache und 1,4-fache Geschwindigkeit. Sollte nur eine geringfügige Änderung der Geschwindigkeit um Faktor 0,1 notwendig sein, wird diese Änderung automatisch und ohne Benutzerinteraktion vorgenommen. Bei größeren Änderungen wird die Nutzer:in darauf hingeweisen und erhält die Möglichkeit, die Beschreibung neu einzusprechen oder neu erzeugen zu lassen.

Wenn es nicht anders möglich ist, kann auch ein Standbild im Video eingefügt werden, um die Beschreibung zu verlängern. Dies ist jedoch nur in sehr speziellen Fällen sinnvoll, da es den Fluss des Videos unterbricht und somit ebenfalls die **Kontinuität** stört. Zusätzlich wird das Video somit nicht mehr zeitsynchron zur Version ohne AD abspielbar.

Beschreibungsvorschläge

Um die Content Creators bei der schnellen Erstellung der Beschreibung zu unterstützen, bietet audilu die Möglichkeit, **Beschreibungsvorschläge** zu generieren. Diese werden mittels eines Machine Learning-Modells erzeugt. Es handelt sich dabei um ein sogenanntes *Large Language Model*: **gpt-3.5-turbo** bzw. **gpt-4** von OpenAI, welche über eine API eingebunden werden. Zusätzlich dazu wird ein Machine Learning-Modell für die Bilderkennung verwendet, um einzelne Standbilder zu Beginn von Sprechpausen aus dem Video zu analysieren und einige Objekte und Konzepte zu identifizieren und sogenannte *Tags* zu extrahieren. Getestet wurden hier das *Image AI*-Modell von Microsoft sowie der Standard-Service von Imagga, welche ebenfalls über eine API eingebunden werden. Beide produzieren gemischte Ergebnisse.

Das Modell führt eine Art "Unterhaltung" mit audilu, in der es die Beschreibungsvorschläge generiert. Dazu werden die folgenden *Prompts* verwendet:

> The following tags are automatically recognised from a still image in a video and your task is to reconstruct and interpret an extremely

> short image description of the scene. Use audio description language. Respond with just your description and refrain from just listing the tags. Try to capture mood and ambience if possible in the word constraint. Use words at most for this new segment. You absolutely MUST respect this limit as your output must be speakable by a human in seconds.
>
> tag1 tag2 ...

Das Modell antwortet dann mit einem Beschreibungsvorschlag, der in der App angezeigt wird. Die Nutzer:innen können diesen Vorschlag per Klick sofort übernehmen oder ignorieren. Wenn der Vorschlag übernommen wird, kann dieser bearbeitet werden, beispielsweise um die Länge der Pause besser zu treffen oder einen kleinen Fehler zu korrigieren. Es wurde sich bewusst dagegen entschieden, die Vorschläge automatisch zu übernehmen, da dies mit dem Unwillen der User einherging, große Veränderungen vorzunehmen und somit ebenfalls die Qualität der Beschreibung verschlechterte. Ebenfalls ist es aus diesem Grund auch nicht möglich, einen Beschreibungsvorschlag neu generieren zu lassen. Die Modelle generieren häufig sehr ähnliche Texte, die sich nur minimal unterscheiden und es somit keinen Mehrwert bietet, einen neuen Vorschlag zu generieren.

Für jede weitere Sprechpause im Video wird die bisherige Kommunikation erneut an das Sprachmodell gesendet und mit dem folgenden Prompt und den Tags des nächsten Bildes ergänzt:

> The following tags have also been detected in this scene, so they may also be described very quickly if they add anything. Please add these new details in a single bullet point. It is crucial to not repeat any information you have already provided in your description until now. Do not refer to the previous description. Rephrasing it is not allowed. Use words at most for this new segment. You absolutely MUST respect this limit as your output must be speakable by a human in seconds.
>
> tag1 tag2 ...

Dies sorgt idealerweise dafür, dass die Beschreibung für das Publikum nicht redundant wird und auf den bereits zuvor gelieferten Informationen aufbauen kann. Auch hier wird der generierte Vorschlag wieder angezeigt und kann übernommen oder ignoriert werden.

Insgesamt zeigt sich das Modell *GPT-4* mit demselben Prompt sehr viel akkurater und es beachtet auch die gegebenen Limits besser. Es erzeugt jedoch auch etwa die zehnfachen Kosten und hat eine längere Ausführungszeit. Deshalb wird in audilu standardmäßig das Modell *GPT-3.5-turbo* verwendet. Alle Prompts können um den Satz "Provide your description in ." ergänzt werden, um die Sprache des Vorschlags zu ändern. Audilu wählt die Sprache anhand der gewählten TTS-Stimme aus.

Die Qualität der Beschreibungsvorschläge ist sehr unterschiedlich. In einigen

Fällen sind sie sehr hilfreich und können direkt übernommen werden. In anderen Fällen sind sie jedoch sehr ungenau und müssen komplett ignoriert werden. Dies ist vor allem dann der Fall, wenn das Modell die Tags nicht korrekt erkennt oder die erkannten Tags einen großen Interpretationsspielraum bieten.

Zukünftige Large Language Models mit multimodalen Fähigkeiten, beispielsweise GPT-4, die dazu in der Lage sind, sowohl Text als auch Bilder zu verarbeiten, könnten eine extreme Verbesserung der Qualität der Beschreibungsvorschläge bedeuten, da diese weniger implizieren müssen und die Übersetzungs-Ebene durch die Tags entfällt. Zum Zeitpunkt der Entwicklung von audilu sind solche Modelle oder Fähigkeiten aber leider noch nicht öffentlich verfügbar.

Technische Umsetzung

Beim Prototypen handelt es sich um eine Web App, deren Frontend in React.js mit JavaScript entwickelt wurde. Um die App auf mobilen Geräten lauffähig zu machen, werden Capacitor und Ionic verwendet. Für die Video-Verarbeitung (Thumbnail-Erstellung, Separation und Vereinigung von Video und Audio, Audio-Geschwindigkeitsanpassungen, ...) kommt ffmpeg zum Einsatz. Zur Erkennung von Sprechpausen nutzt Audilu die Open Source-Bibliothek ricky0123/vad. Zur Generierung der Beschreibungsvorschläge werden die proprietären Dienstleistungen von OpenAI verwendet. Für das Image Tagging wurden Schnittstellen zu proprietären Modellen von Microsoft Azure und Imagga getestet.

Es hat sich gezeigt, dass, obwohl Web-Technologien viele der benötigten Features selbst bereitstellen, diese für Aufgaben der Medienproduktion eher ungeeignet sind. So musste teilweise auf Server-Lösungen umgestiegen werden, die dann wiederrum je nach Internetgeschwindigkeit lange Upload-Zeiten benötigten. Zusätzlich ist die resultierende App sehr langsam und fehleranfällig, was mit vollständig nativen Implementierungen wahrscheinlich ein weniger großes Problem gewesen wäre.

User-Testing

Der Prototyp wurde in mehreren Iterationen mit unprofessionellen Content Creators getestet, die sich größtenteils aber selbst als sehr erfahren im Umgang mit Technologie und Medien einschätzten. Die Testpersonen wurden gebeten, ein Testvideo auszuwählen und dieses mit audilu zu bearbeiten. Bei diesem Video handelt es sich um eine Aufnahme des Twitter-Users @DerFrankyman, dessen Ehefrau mithilfe des Mobilitätsservice der Deutschen Bahn von einem Regionalzug in einen ICE umsteigt. Anschließend wurden die Testpersonen zu ihren Erfahrungen und Erkenntnissen befragt.

Zusätzlich gab es ein Test-Publikum von 3 blinden und sehbehinderten Personen mit unterschiedlichen Sehresten, die das Testvideo mit und ohne AD konsumiert haben. Auch sie wurden anschließend zu ihren Erfahrungen und Erkenntnissen befragt.

Feature	Kommentare
Sprechpausen-Erkennung	Die Sprechpausen-Erkennung wurde als sehr hilfreich und effizient für die Nutzenden empfunden. Sie erlauben ein schnelles Voranschreiten im Prozess der AD-Erstellung.
Text-to-Speech	Die Verwendung von Text-to-Speech wurde als sehr angenehm empfunden, da die Nutzenden in der Test-Umgebung eher selten bereit waren, selbst Sprachaufnahmen durchzuführen. Die Anzeige der geschätzten Wortanzahl wurde ebenfalls als sehr hilfreich empfunden, jedoch häufig als Anreiz verstanden, diese auszureizen, was in manchen Fällen für eine verminderte Qualität sorgte.
Automatische Geschwindigkeitsanpassung	Die automatische Geschwindigkeitsanpassung wurde im Test sehr wenig genutzt. Sie degradiert allerdings die Qualität des Ergebnisses und die Bereitschaft des Publikums, die Audio-Deskription anzuhören.
Beschreibungsvorschläge	Die Beschreibungsvorschläge wurden von Nutzenden als eher hilfreich empfunden, da sie einen Struktur-Vorschlag enthalten. Die Nutzenden waren sich bewusst, dass sie die Vorschläge nicht einfach übernehmen können, sondern diese noch bearbeiten müssen. Bearbeitete Beschreibungsvorschläge wurden durch das Testpublikum allerdings in allen Fällen negativer bewertet als von Beginn an selbst geschriebene Beschreibungen. Das Bearbeiten von Vorschlägen sorgte allerdings immer für eine Zeit-Ersparnis von ca. 30 Prozent.

Desweiteren wurde im User-Testing immer eindeutiger, dass die meisten kurzen Videos auf Social Media entweder überhaupt keine Stimme enthalten, oder diese während der gesamten Laufzeit des Videos zu hören ist. Letzteres ist

ein Ausschlusskriterium für eine Audiodeskription, während bei Ersterem in vielen Fällen eine Textbeschreibung das bessere Mittel ist, um Barrierefreiheit herzustellen. In diesem Fall eignet sich auch das Voice Over-Tool in bestehenden mobilen Videoschnittsoftwares, weshalb Audilu leider in den meisten Fällen neben der schlechten Qualität der Ergebnisse auch für die Content-Erstellenden keinen Mehrwehrt bietet.

Fazit

In den meisten Fällen ist es für Content Creators nicht sinnvoll, Audilu zu verwenden, da sich nur wenige ihrer Inhalte für AD eignen. AD-Erfahrene Creators erleben durch Audilu auch keine Effizienzsteigerung im Erstellungsprozess, da die meisten Funktionen eher unflexibel und die Beschreibungsvorschläge selten zutreffend sind.

Auch das Publikum der durch Audilu erzeugten Audiodeskriptionen ist mehrheitlich unzufrieden mit der abgelieferten Qualität, da diese nicht den Standards entspricht, die sie von professionellen AD gewohnt sind. Die AD als solche wurde in den meisten Fällen abgelehnt und es wurde bevorzugt, das Video ohne AD zu konsumieren.

Weiterentwicklung

Eine Weiterentwicklung der App wird aufgrund der Vielzahl der Defizite nicht angestrebt. Während der Projektlaufzeit sind zusätzlich neue KI-Modelle unter Anderem von der US-amerikanischen Firma OpenAI veröffentlicht worden, die großes Potenzial für vollautomatisierte Audiodeskription zeigen. Ich gehe davon aus, dass das Konzept hinter Audilu, und in großen Teilen auch menschlich erstellte Audiodeskription als Solche, innerhalb der nächsten 5 bis 10 Jahre nicht mehr häufig anzutreffen sein werden, wenn Large Language Models neben den bereits jetzt sehr guten Bildbeschreibungsfähigkeiten auch mit zeitbasierten Medien umgehen werden können.

Dies bietet für die Barrierefreiheit und gesellschaftliche Teilhabe sehr großes Potential, da Audio-Deskriptionen damit individuell und in Echtzeit unter Berücksichtigung von Vorlieben jeder einzelnen Person erstellt werden können. Dies ist mit menschlich erstellten Audiodeskriptionen nicht möglich.