Technology
Arts Sciences
TH Köln

Transkription wissenschaftlicher Podcasts unter Verwendung von automatischen Spracherkennungssystemen

Fakultät für Informations- und Kommunikationswissenschaften
der Technischen Hochschule Köln

Abschlussarbeit

zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.)

vorgelegt von

Robin Hilbrecht

Abgabetermin: 23.10.2023

Erstprüfer:	Prof. Dr. rer. nat. Konrad Förstner
Zweitprüfer:	Benjamin Wolff (M.Sc.)

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

API Application Programming Interface

ASR Automated Speech Recognition

CNN Convolutional Neural Network

E2E End-to-End

GELU Gaussian Error Linear Unit

GMM Gaussian Mixture Model

HMM Hidden Markov Model

JSON JavaScript Object Notation

NLP Natural Language Processing

NLTK Natural Language Toolkit

OOV Out-of-Vocabulary

POS Part of Speech

RDF Ressource Description Framework

RNN Recurrent Neural Network

RSS Really Simple Syndication

Seq2seq Sequence to Sequence

STT Speech-to-Text

TSV Tab-separated Values

VAD Voice Activity Detection

WER Word Error Rate

1 Einleitung

Gender-Hinweis

In dieser vorliegenden Arbeit wurde keine explizite Verwendung einer gendergerechten oder genderneutralen Sprache verfolgt. Die gelegentliche Verwendung des generischen Maskulinums in dieser Arbeit bringt keine Präferenz für ein Geschlecht zum Ausdruck und dient auch nicht dazu, bestimmte Geschlechter oder Identitäten zu betonen oder zu diskriminieren. Es handelt sich vielmehr um eine sprachliche Konvention, die darauf abzielt, die Lesbarkeit zu fördern, ohne die Bedeutung und die Würdigung aller Geschlechter und Identitäten zu beeinträchtigen.

1.1 Hintergrund

Die moderne Informationslandschaft ist durch eine kontinuierliche Flut von Daten geprägt, die nahezu jeden Aspekt unseres Alltags durchdringt. Im Zuge dieser Entwicklung haben sich Podcasts zu einem bedeutsamen Medium entwickelt, welches eine einzigartige Möglichkeit bietet, Inhalte zu präsentieren und zu konsumieren. Im Gegensatz zu schriftlichen oder visuellen Medien ermöglichen Podcasts den Zuhörern, sich in verschiedene Themen zu vertiefen, während sie gleichzeitig anderen Aktivitäten nachgehen. Diese Flexibilität hat dazu beigetragen, ein großes Publikum anzusprechen, das möglicherweise weniger geneigt ist, traditionelle schriftliche Quellen zu nutzen.

Die thematische Vielfalt von Podcasts erstreckt sich von wissenschaftlichen Entdeckungen über philosophische Diskurse bis hin zu sozialen Debatten. Podcasts bieten eine Plattform zur Diskussion von Ideen und zur Anregung kritischer Dialoge. Die Einbindung von Experteninterviews, Debatten und persönlichen Perspektiven verleiht den Inhalten eine lebendige Authentizität. Besonders beeindruckend ist die zügige Expansion von Podcasts in den akademischen Bereich. Forschende und Fachkundige aus diversen Fachgebieten erkennen zunehmend das Potenzial von Podcasts als Medium zur Verbreitung ihrer Erkenntnisse und Ideen. Dank der direkten und persönlichen Natur der gesprochenen Sprache können komplexe Konzepte auf verständliche Weise vermittelt werden, ohne dabei an intellektuellem Anspruch einzubüßen. Eine Eigenschaft die es Fachleuten ermöglicht, ein breiteres Publikum zu erreichen, ohne auf wissenschaftliche Präzision verzichten zu müssen. Die Bedeutung von Podcasts für die Verbreitung wissenschaftlicher Erkenntnisse wird zusätzlich durch die zunehmende Interaktivität des Mediums gestärkt. Hörer können oft direkt auf Inhalte reagieren, sei es über soziale Medien, Diskussionsforen oder Live-Frage-Antwort-Sitzungen. Diese interaktive Dynamik zwischen Produzierenden und Publikum schafft eine lebendige Gemeinschaft, in der Ideen ausgetauscht und vertieft werden können.

Während Podcasts zweifellos einen besonderen Platz als Medium einnehmen, stellen ihre vorwiegend auditiven Eigenschaften Herausforderungen für ihre Sichtbarkeit im digitalen Raum dar. Im Unterschied zu textbasierten oder visuellen Medien, die durch Texte, Bilder und Meta-

daten leicht durchsuchbar sind, setzen Podcasts auf menschliche Sprache. Diese Abhängigkeit erschwert die automatisierte Indexierung und Wiederfindung spezifischer Inhalte innerhalb von Podcasts. Zusätzlich beeinträchtigt das Fehlen visueller Hinweise in herkömmlichen Suchergebnissen die Identifikation und Unterscheidung relevanter Podcasts. Dies wird besonders deutlich, wenn versucht wird, bestimmte Themen, Schlüsselwörter oder Zitate in Podcasts zu lokalisieren. Während geschriebener Inhalt vollständig durchsucht werden kann, bleiben Podcasts im Allgemeinen auf akustischen Inhalt beschränkt. Das Fehlen von Transkriptionen oder detaillierten Metadaten könnte dazu führen, dass wertvolle Podcasts mit einsichtsvollen Informationen im digitalen Rauschen untergehen, was die Verbreitung von wissenschaftlichen Kenntnissen beeinträchtigt. Trotz ihrer zunehmenden Bedeutung im akademischen Bereich könnten wertvolle Einsichten und Fachmeinungen unbemerkt bleiben, es sei denn, sie werden angemessen indexiert und transkribiert.

Um die Auffindbarkeit und Durchsuchbarkeit von Podcast-Inhalten im digitalen Raum zu optimieren, kann auf automatisierte Transkriptionsverfahren zurückgegriffen werden. Die jüngsten Fortschritte haben die Fähigkeiten der automatischen Spracherkennung in bemerkenswerter Weise weiterentwickelt. Insbesondere durch die rasanten Entwicklungen im Bereich des Deep Learning wurden Spracherkennungssysteme auf ein neues Niveau gehoben. Für die wissenschaftliche Community ergeben sich aus der Integration von KI-gestützter Transkription einige positive Aspekte. Die technologische Anwendung steigert die Effizienz des Transkriptionsprozesses erheblich, was die langwierige manuelle Arbeit reduziert. Dieser Effekt minimiert nicht nur den Arbeitsaufwand, sondern trägt auch dazu bei, eine konsistente und zuverlässige Transkriptionsqualität sicherzustellen. Über die rein operative Effizienz hinaus bietet die Implementierung von KI-gestützten Transkriptionssystemen (auch: automatische Spracherkennungsysteme; engl. Automated Speech Recognition (ASR)) auch einen weitaus tieferen Mehrwert. Die generierten Transkripte erzeugen eine strukturierte Textgrundlage, die als Nährboden für innovative Technologien wie quantitative Analysen und Natural Language Processing (NLP) dient. Darüber hinaus trägt die Einbindung von automatischen Spracherkennungssystemen auch dazu bei, die Zugänglichkeit der Inhalte zu erhöhen. Insbesondere Personen mit Hörbeeinträchtigungen können von dieser Entwicklung profitieren. Die Bereitstellung von Transkriptionen ermöglicht diesen Personen den Konsum der Podcast-Inhalte und die Teilnahme am Informationsaustausch. Dieser inklusive Aspekt stellt einen bedeutenden Schritt in Richtung einer barrierefreien Informationsgesellschaft dar.

Zusammenfassend lässt sich festhalten, dass Podcasts in der modernen Informationslandschaft als bedeutsames Medium fungieren, das eine einzigartige Möglichkeit bietet, Wissen zu vermitteln und aufzunehmen. Trotz ihrer Vorzüge im audiobasierten Format stehen Podcasts jedoch vor Herausforderungen hinsichtlich ihrer Auffindbarkeit und Durchsuchbarkeit im digitalen Raum. Die Integration von KI-gestützten Transkriptionsverfahren eröffnet vielversprechende Wege, diese Hürden zu überwinden. Durch die optimierte Zugänglichkeit, die gesteigerte Effizienz und die inklusive Ausrichtung tragen diese technologischen Fortschritte dazu bei, das volle Potenzial von wissenschaftlichen Podcasts auszuschöpfen und den Dialog sowie die Verbreitung von Wissen auf eine neue Ebene zu heben.

1.2 Zielsetzung und Motivation

Diese Bachelorarbeit befasst sich mit der Untersuchung und Bewertung der Anwendung automatischer Spracherkennungssysteme zur Transkription wissenschaftlicher Podcasts. Dabei liegt der Fokus auf der Analyse der Vorzüge, die sich aus der Nutzung von KI-gestützten Transkriptionsverfahren ergeben. Die Arbeit beleuchtet die technologischen Grundlagen, die zur Entwicklung dieser Systeme geführt haben, und untersucht die Auswirkungen auf die Qualität der generierten Transkripte. Durch eine eingehende Auseinandersetzung mit diesem Thema soll ein tieferes Verständnis für die Potenziale und Herausforderungen der Kombination von Podcasts und KI-gestützter Transkription gewonnen werden. Die These die sich daraus ergibt und der in dieser Arbeit nachgegangen wird lautet wie folgt:

> *Die Integration von KI-gestützten Transkriptionsverfahren in wissenschaftliche Podcasts ermöglicht nicht nur eine effiziente und qualitativ hochwertige Transkription, sondern eröffnet auch neue Möglichkeiten zur Verbesserung der Auffindbarkeit, Zugänglichkeit und Verbreitung von wissenschaftlichen Inhalten.*

Ziel dieser Arbeit ist es, das Potenzial von Podcasts und KI-gestützten Transkriptionssystemen zu verbinden, den Austausch von Wissen zu fördern und inklusive wissenschaftliche Diskussionen zu ermöglichen. Damit soll nicht nur das Beste aus beiden Welten vereint, sondern auch ein Beitrag zu einer reichhaltigen Informationsgesellschaft und offenen Wissenschaft geleistet werden.

1.3 Aufbau der Arbeit

Die vorliegende Arbeit ist in sechs Kapitel gegliedert. Der Aufbau der Arbeit folgt einer klaren Struktur, die den Lesenden durch das Thema führen soll. Von der einleitenden Exposition bis zu der abschließenden Schlussbetrachtung.

1. **Einleitung**: Dieser Abschnitt erläutert den Kontext und die Relevanz des Forschungsthemas. Die Einleitung stellt die Grundlagen des Themas dar, nennt die Ziele der Untersuchung und betont deren Bedeutung. Die Zielsetzung und die Motivation werden präzise dargelegt.

2. **Literaturübersicht und Theorie**: Dieses Kapitel beginnt mit einem Überblick über die moderne Informationslandschaft, gefolgt von den Vorteilen und Herausforderungen von Wissenschaftspodcasts. Im zweiten Teil werden KI-gestützte Methoden zur automatischen Transkription vorgestellt und die Modellstruktur von Speech Transformern erläutert.

3. **Datensatz**: Dieser Abschnitt soll die verwendeten Datensätze für die Untersuchung detailliert beschreiben. Dabei werden die Quellen und Inhalte der Datensätze näher erläutert.

4. **Methodik**: Dieses Kapitel bietet einen Einblick in die wesentlichen Werkzeuge und Metriken, die im Rahmen der Forschung verwendet wurden. Zudem wird die speziell ent-

wickelte Pipeline vorgestellt, die als zentrales Instrument zur Erstellung und Analyse der Podcast-Transkripte dient.

5. **Ergebnisse und Diskussion**: In diesem Teil werden die Ergebnisse der durchgeführten Untersuchung präsentiert. Hierbei werden die Erkenntnisse, die aus der Anwendung der KI-gestützten Transkription gewonnen wurden, detailliert dargelegt und diskutiert.

6. **Fazit und Ausblick**: Im abschließenden Kapitel erfolgt ein kurzes Fazit der Arbeit, in dem die Hauptpunkte und Erkenntnisse der Untersuchung nochmals kompakt dargestellt werden. Des Weiteren werden offene und weiterführende Fragen thematisiert.

2 Literaturübersicht und Theorie

2.1 Podcasts in der modernen Informationslandschaft

2.1.1 Einblick in die moderne Informationslandschaft

Die moderne Informationslandschaft ist geprägt von dem Begriff „New Media“. Er wird in der Medienwissenschaft verwendet, um digitale Technologien zur Erstellung, Verbreitung und Konsumierung von Inhalten zu beschreiben. Beispiele für New Media umfassen soziale Medien, Podcasts, Streaming-Dienste, Websites, Blogs, Apps, digitale Kunstwerke und mehr. Der Aufstieg der New Media verändert die Art der Kommunikation und hat einen transformierenden Einfluss auf verschiedene Aspekte der Gesellschaft (D. B. Dhiman, 2023). Diese Transformation wird durch die folgenden, digitalen Eigenschaften der Medien vorangetrieben:

- **Interaktivität**: Im Vergleich zu traditionellen Medien bieten New Media oft die Möglichkeit zur Interaktion. Benutzer können auf Inhalte reagieren, kommentieren, sie teilen und selbst generieren, was zu einer aktiven Beteiligung führt.
- **Multimediale Merkmale**: Die moderne Medienlandschaft bietet vielfältige Formate, darunter Texte, Bilder, Videos, und interaktive Grafiken, wodurch es Menschen ermöglicht wird, Informationen auf die für sie am besten geeignete Weise zu konsumieren.
- **Globale Zugänglichkeit**: Ereignisse in einem Teil der Welt können heute rasch globale Auswirkungen haben. Dies kann verschiedene Bereiche betreffen, darunter Finanzmärkte, politische Entwicklungen, die öffentliche Meinung und die internationale Zusammenarbeit. Nachrichtenereignisse wie Naturkatastrophen, politische Krisen oder wissenschaftliche Durchbrüche können nun weltweit Aufmerksamkeit erregen und die Interessen zahlreicher Nationen beeinflussen.

Die Demokratisierung von Inhalten durch New Media hat erhebliche Auswirkungen auf die Verbreitung von Informationen in der heutigen Gesellschaft. Die Verfügbarkeit kostengünstiger oder sogar kostenloser Werkzeuge und Plattformen ermöglicht es Menschen heute, Inhalte von praktisch überall aus zu erstellen, sei es in Form von Texten, Videos, Podcasts oder anderen Medienformaten. Dies hat zu einer breiten Palette von Meinungen, Kulturen und Lebenserfahrungen geführt, die zu einer dynamischeren und vielfältigeren Medienlandschaft beitragen (D. B. Dhiman, 2023). Während im Unterhaltungssektor Streaming-Dienste wie Netflix, Amazon Prime oder Youtube herkömmliche Vertriebsmodelle für Filme und Fernsehen auf den Kopf gestellt haben (Budzinski et al., 2021), hat New Media auch den Journalismus grundlegend verändert. Die Entstehung von Online-Nachrichtenportalen und Bürgerjournalismus hat die herkömmliche Rolle der Nachrichtenvermittelnden in Frage gestellt und neue Modelle für die Produktion und Verbreitung von Nachrichten hervorgebracht (Y. Wu, 2018).

Aus einer vom Reuters Institut für Journalismus[1] in Auftrag gegebenen Studie bezüglich des

[1] https://www.digitalnewsreport.org/

Nachrichtenkonsumverhaltens weltweit geht hervor, dass das Informationsverhalten maßgeblich von den Verhaltensgewohnheiten der jüngeren Generation beeinflusst wird, welche mit sozialen Medien aufgewachsen sind und ihre Aufmerksamkeit im Hinblick auf Nachrichten eher Influencern und Prominenten schenken als Journalisten. Auch wurde erkannt, dass junge Leute im Kontrast zu vorherigen Nutzergruppen im Allgemeinen eher eine schwache Verbindung zu etablierten Nachrichtenmagazinen und Portalen haben und sich vorzugsweise über alternative Wege wie Soziale Medien oder News-Aggregatoren[2] wie Reddit[3] oder Google News[4] informieren. Traditionelle Medien wie Fernsehen und Print hingegen spielen trotz des Onlineangebots eine immer kleiner werdende Rolle (Newman et al., 2023). Dieser Strukturwandel hin zu plattformdominierten Medienumgebungen wurde durch das einschlägige globale Ereignis der COVID-19 Krise verstärkt, welche zu einem erheblichen Digitalisierungsschub geführt hat. Bedingt durch globale Lockdowns zur Eindämmung des Virus, zwang die Pandemie Unternehmen, Bildungseinrichtungen, Regierungen und viele andere Organisationen dazu, sich schnell an neue Umstände anzupassen und verstärkt auf digitale Technologien und Lösungen umzusteigen (Wamba et al., 2023). Zusätzlich greifen Menschen in Zeiten der Krise vornehmlich auf Nachrichtenquellen zurück, welche leicht zugänglich sind und eine unmittelbare Berichterstattung bieten (Aelst et al., 2021). Auch die wissenschaftliche Kommunikation hat sich in den letzten Jahren stark verändert. Dieser Wandel ist auf die verstärkte Nutzung digitaler Technologien, die Zunahme von Open-Access-Publikationen und die vermehrte Nutzung sozialer Medien für den Austausch und die Verbreitung von Forschungsergebnissen sowie auf eine verstärkte internationale Zusammenarbeit in der Wissenschaft zurückzuführen. Zudem ergab sich im Rahmen der COVID-19 Pandemie die Problematik, dass der klassische Weg der Publikation nicht mehr mit der Geschwindigkeit der Forschung mithalten konnte. Dies führte zum Aufstieg des Preprints um wissenschaftliche Ergebnisse unmittelbar teilen zu können (Fraser et al., 2021; Chiarelli et al., 2019). Als Folge dessen sind Forschungsinformationen breiter verfügbar, Veröffentlichungen erfolgen schneller und der Dialog zwischen Wissenschaftlern und der Öffentlichkeit ist intensiver.

Neben den vielen positiven Aspekten der modernen Informationslandschaft, ergeben sich gleichzeitig jedoch auch Herausforderungen in Bezug auf die Qualität von Informationen und die Verbreitung von Fehlinformationen (Van Bavel et al., 2020), welche sich durch die Verwendung sozialer Medienplattformen erheblich beschleunigt hat (B. Dhiman, 2022). Nutzer können Informationen und Nachrichten ohne strenge Überprüfung oder Qualitätskontrolle teilen, was dazu führt, dass falsche oder irreführende Informationen schnell viral werden können. Darüber hinaus begünstigen personalisierte Algorithmen sozialer Medien und Suchmaschinen die Verstärkung von bereits vorhandenen Meinungen, in der Medienwissenschaft auch als Filterblase bezeichnet, und können dazu beitragen, dass Menschen Fehlinformationen unbemerkt akzeptieren. Dabei werden interaktionsfördernde Inhalte unabhängig von ihrer Glaubwürdigkeit priorisiert (D. B. Dhiman, 2023). Dies kann im Rahmen von Desinformationskampagnen gezielt ausgenutzt werden, um Verwirrung zu stiften oder politische und soziale Spannungen zu schüren (Wibowo, 2022). Das ohnehin bereits bestehende Misstrauen in der Bevölkerung

[2] Ein News-Aggregator, auch als Feed-Leser oder Nachrichtenleser bezeichnet, ist eine Client-Software oder eine Webanwendung, die öffentliche Webinhalte wie Online-Zeitungen, Blogs, Podcasts und Videoblogs (Vlogs) an einem Ort sammelt und im Stile der eigenen Website übersichtlich darstellt. (Miles, 2009)

[3] https://www.reddit.com

[4] https://news.google.com

gegenüber informativen Medien (Newman et al., 2023) kann sich durch die Verbreitung von Fehlinformationen weiter manifestieren und erhebliche Folgen für das Vertrauen in verlässliche Informationsquellen haben (Hameleers et al., 2022, insbesondere im wissenschaftlichen Bereich, wo vertrauenswürdige Informationen unerlässlich sind (Zarocostas, 2020).

Die Ersteller von Medien, sei es im Bereich des Journalismus, der Unterhaltung oder der Wissenschaft, stehen vor der anspruchsvollen Aufgabe, die Anforderungen an einfach zu konsumierende, ansprechende und dennoch informative Medien zu erfüllen. Dabei ist es von entscheidender Bedeutung, vertrauenswürdige Quellen und Qualitätsstandards aufrechtzuerhalten. Angesichts der kontinuierlichen Veränderungen in der Informationslandschaft wird die Fähigkeit, glaubwürdige und leicht zugängliche Informationen bereitzustellen, zu einer zentralen Verantwortung. Gleichzeitig ist es unerlässlich, dass Medienkonsumenten Medienkompetenz entwickeln, um Informationen kritisch zu hinterfragen und zu überprüfen. In einer Welt, in der die Informationsflut unaufhaltsam wächst, wird die Fähigkeit, Informationen zu filtern, zu verifizieren und auf verlässliche Quellen zuzugreifen, zunehmend wichtiger.

2.1.2 Podcasts als einzigartige Möglichkeit des Konsums

Im Zuge des digitalen Wandels sind Podcasts zu einem faszinierenden und vielseitigen Medium der Audiokommunikation geworden. Dieses Kapitel dient dazu einen kurzen Überblick über Podcasts zu verschaffen, erklärt ihre evolutionäre Entwicklung und hebt schließlich die Merkmale hervor, die Podcasts zu einem einzigartigen, einflussreichen und zugänglichen Kommunikationsmittel machen.

Jemily Rime, Chris Pike und Tom Collins definieren einen Podcast als ein episodisches, herunterladbares oder streambares, hauptsächlich aus gesprochenem Audioinhalt bestehendes Medium, das über das Internet verbreitet wird und zu jeder Zeit, wie an jedem Ort, abgespielt werden und von jeder Person produziert werden kann, die dies wünscht (Rime et al., 2022). Die Wurzeln der Podcasts reichen zurück bis zur Ära des Radios, als Menschen begannen, Audioinhalte über Rundfunksendungen zu verbreiten (Madsen, 2009). In den späten 1990er Jahren entwickelten einige Blogger die Idee des „Audiobloggings“ und begannen, Audioaufnahmen in ihre Blogs zu integrieren. Dies waren die ersten Ansätze zu Audio-On-Demand-Inhalten (Rime et al., 2022). Der Name „Podcast “ wurde von Ben Hammersley im Jahre 2004 geprägt - eine Kombination aus „pod“, kurz für „play on demand“, und „Broadcast“(Hammersley, 2004). Der Durchbruch für Podcasts kam jedoch im Jahr 2005, als Apple in iTunes 4.9 die Unterstützung für Podcasts einführte (Apple, 2005). Seitdem ist der Podcast-Markt rapide gewachsen (Grand View Research, n. d.). Die Etablierung des Smartphones als Massenmedium (Berry, 2015) und sinkende Kosten für mobiles Internet, trugen erheblich zur Popularität von Podcasts bei. Mit erschwinglicheren Datenplänen und Tarifen wurde es für immer mehr Menschen möglich, große Mengen an digitalen Inhalten standortunabhängig herunterzuladen und zu streamen. Parallel dazu führte die steigende Qualität von Podcast-Produktionen und die Verschiebung traditioneller Podcast-Aggregatoren zu etablierten Plattformen zu einem weiteren Anstieg an Beliebtheit (Katzenberger et al., 2022; Newman et al., 2023). Die COVID-19-Pandemie führte zu einem signifikanten Anstieg in der Podcast-Branche, insbesonders Wissenschaftspodcasts (Dernbach, 2022), angetrieben durch weltweite Lockdown-Maßnahmen und die wachsende

Beliebtheit von Remote-Arbeit. Da Menschen mehr Zeit zuhause verbrachten, stieg die Nachfrage nach leicht zugänglichen Unterhaltungsinhalten deutlich an (Grand View Research, n. d.). Zusätzlich haben Nachrichten- und Medienorganisationen, spezielle Podcasts gestartet, wie beispielsweise „Das Coronavirus-Update“[5] des deutschen Fensehnetzwerks NDR, um zeitnahe COVID-19 Informationen bereitzustellen. Die Produktionskosten und Hürden die zur Erstellung von Podcasts überwunden werden müssen sind vergleichsweise niedrig, weshalb sie bei Verlägen immer beliebter werden um den neuen Vorlieben des modernen Publikums gerecht zu werden (Newman et al., 2023). Zudem wird ein wachsender Trend zu nachrichtenorientierten Podcasts mit Video-Elementen, auch Vodcast genannt, beobachtet (Newman et al., 2023). Aus einer im Jahre 2019 durchgeführten Studie über die Nutzung von Podcasts im amerikanischen Raum des Marktforschungsunternehmen Edison Research geht hervor, dass Podcasts hautpsächlich zur Unterhaltung, zum Informieren oder zur Entspannung gehört werden (Edison Research, 2019). Insbesondere die jüngere Generation, welche bereits als Haupttreiber des starken Wandels der Medienindustrie identifiziert wurde, bildet die Haupthörerschaft von Podcasts (Edison Research, 2023). Heutzutage ist der Podcast-Markt stark fragmentiert. Key Players sind Anbieter wie Spotify, Amazon, Apple, Sound Cloud und Audacy, welche ein breites Angebot aus verschiedensten Bereichen vorweisen, darunter Kultur und Gesellschaft, Bildung, Kunst, Comedy, Technik, Wissenschaft, Nachrichten und viele weitere (Listen Notes Inc., n. d.). Dabei wird vermehrt auf intelligente Lösungen gesetzt um die Reichweite von Inhalten und Werbungen zu erhöhen. Spotify beispielsweise setzt machinelle Lösungen ein um Podcasts einerseits zu transkribieren und andererseits durch aktive Übersetzung einem globalen Publikum zur Verfügung zu stellen und die Reichweite zu erhöhen (Grand View Research, n. d.). Der Erfolg des Podcasts begründet sich auf die besonderen Attribute, die ihn in ihrer Gesamtheit von anderen Medienformen abgrenzen:

Flexibilität: Die primär audiobasierte Natur von Podcasts gestattet den Hörern, Informationen durch das auditive Medium zu konsumieren, ohne auf visuelle Reize angewiesen zu sein. Dies erzeugt eine vielseitige Erfahrung, bei der die Konzentration auf den reinen Klang und die Inhalte im Vordergrund steht. Ein weiterer Vorteil ist der mobile und globale Zugang von Podcasts. Da sie über das Internet verbreitet werden, können Hörer weltweit auf eine breite Palette von Inhalten zugreifen, solange eine Internetverbindung besteht. Dies erweitert den potenziellen Hörerkreis und ermöglicht es, Podcasts unabhängig von ihrem geografischen Standort zu konsumieren. Der „On-Demand“-Aspekt von Podcasts gestattet Hörern die Freiheit, Podcasts nach eigenem Ermessen abzurufen und anzuhören. Im Gegensatz zu Live-Radio oder Fernsehen können sie ihre Hörerfahrung individuell gestalten und Podcasts dann hören, wenn es in ihren Zeitplan passt. Podcasts bieten die Möglichkeit zur Integration verschiedener Medienelemente wie Videos, Webseiten und mehr, wodurch sie eine vielseitige und multimediale Kommunikationsplattform darstellen (Dernbach, 2022). Abschließend ist die Fähigkeit, Podcasts zu jeder Zeit und parallel zu anderen Aktivitäten zu hören, ein entscheidender Vorteil dieses Mediums. Da Podcasts ausschließlich auf akustischem Inhalt basieren, können Hörer sie während des Autofahrens, Kochens, Sporttreibens oder anderer täglicher Aufgaben genießen. Dies optimiert die Zeitnutzung und erlaubt es den Menschen, Wissen und Unterhaltung nahtlos in ihren Alltag

[5] https://www.ndr.de/nachrichten/info/podcast4684.html

zu integrieren (Donnelly und Berge, 2006; Jowitt, 2007; Quintana und Heathers, 2020).

Persönliche Ebene: Podcasts zeichnen sich durch ihren persönlichen Charakter aus. Anders als bei Massenmedien können Hörer Podcasts nach ihren eigenen Interessen und Vorlieben auswählen, was eine maßgeschneiderte Hörerfahrung ermöglicht. Da Podcasts oft über Kopfhörer gehört werden, entsteht eine unmittelbare Verbindung zwischen dem Sprecher und dem Hörer. Mündliche Sprache bietet im Vergleich zu schriftlicher Sprache eine umfassendere Kommunikationsmöglichkeit, da sie verschiedene linguistische Aspekte wie Dialektvariationen und individuelle Eigenheiten einschließt, die in schriftlicher Form oft geglättet oder standardisiert werden. Diese Feinheiten in der gesprochenen Sprache tragen zu einem tieferen Verständnis des Hintergrunds, der Kultur und der persönlichen Eigenschaften des Sprechers bei und machen sie zu einem reichen und vielschichtigen Kommunikationsmittel (Jones et al., 2021). Darüber hinaus werden Podcasts häufig in persönlichen Haushalten oder in vertrauten Umgebungen aufgenommen. Anders als in einem Studio oder vor einer großen Produktionseinheit können Podcaster ihre Gedanken und Geschichten in einem gemütlichen und vertrauten Raum teilen. Dies schafft ein Gefühl der Nähe und Authentizität, das in anderen Medien schwer zu erreichen ist. Hörer fühlen sich oft, als würden sie an einer privaten Konversation teilnehmen, was den persönlichen Charakter von Podcasts verstärkt. Schlussendlich handelt es sich bei Podcast-Sprechern oftmals um Personen, die bereits eine Online-Präsenz oder soziale Medien nutzen, um mit ihrem Publikum in Kontakt zu treten. Dadurch können Hörer bereits eine Verbindung zu den Podcastern aufgebaut haben, bevor sie überhaupt ihren ersten Podcast anhören, was ein Gefühl von Vertrautheit und persönlicher Bindung schafft (Berry, 2016; Katzenberger et al., 2022).

Format: Ein weiteres charakteristisches Element von wissenschaftlichen Podcasts ist das häufig längere Format. Im Gegensatz zu traditionellen, wissenschaftlichen Veröffentlichungen oder kurzen Nachrichtenberichten bieten Podcasts die Möglichkeit, tiefer in ein Thema einzutauchen. Dies ermöglicht es Wissenschaftlern und Experten, ihre Forschung und Erkenntnisse ausführlicher zu präsentieren und zu diskutieren. Hörer haben die Gelegenheit, sich in die Materie zu vertiefen und ein umfassenderes Verständnis für komplexe wissenschaftliche Fragestellungen zu entwickeln (Newman et al., 2023). Durch die episodische Struktur, bei der Inhalte in einzelnen, regelmäßig veröffentlichten Folgen oder Episoden präsentiert werden, wird den Hörern eine verfolgbare Serie geboten bei der komplexe wissenschaftliche Themen in überschaubare Teil runtergebrochen werden. Der Really Simple Syndication (RSS)-Standard oder RSS-Feed, als offenes Vertriebsformat für Podcasts, beinhaltet verschiedene Metadatenfelder (Apple, n. d.), wie beispielsweise den Titel, die Beschreibung und die Sprache und soll eine effiziente und weitreichende Verbreitung von Episoden ermöglichen (Quintana und Heathers, 2020). Metadaten werden sowohl zur Beschreibung des gesamten Podcasts (Feed-Level) als auch zur Beschreibung einzelner Episoden (Episoden-Level) verwendet.

Vielfalt: Podcasts decken eine immense Bandbreite von Themen ab, von Wissenschaft und Technologie über Kunst und Kultur bis hin zu Geschichte, Unterhaltung, Bildung und vielem mehr (Listen Notes Inc., n. d.). Auch Nischenthemen werden aufgrund des flexiblen und persönlichen Charakters, sowie niedrigen Produktionskosten abgedeckt. Diese Vielfalt ermöglicht es Hörern, Podcasts auszuwählen, die ihren persönlichen Interessen und Neigungen entsprechen.

Darüber hinaus bieten Podcasts oft unterschiedliche Perspektiven und Stimmen. Sie dienen als Plattform für vielfältige Meinungen, Hintergründe und Erfahrungen, was die kritische Auseinandersetzung mit verschiedenen Standpunkten fördert und den Horizont der Hörer erweitert. Die Vielfalt der Inhalte von Podcasts zeigt sich auch in verschiedenen Formaten, darunter Interviews, Diskussionen, Geschichten, Nachrichtenberichte und vieles mehr.

Podcasts spielen in der Medienlandschaft eine einzigartige und vielfältige Rolle. Mit ihren flexiblen Hörerlebnissen, dem persönlichen Charakter, längeren Formaten und einer breiten Palette von Inhalten sind sie zu einem wichtigen Medium geworden, das Wissen und Unterhaltung auf eine ganz neue Weise vermittelt. Ihr Potenzial, die Zuhörer in den Bann zu ziehen und komplexe Themen zugänglich zu machen, zeigt, dass Podcasts weiterhin eine wichtige Rolle in der modernen Kommunikation spielen werden.

2.1.3 Vorteile von Wissenschaftspodcasts

In der Wissenschaftskommunikation haben sich Podcasts seit ihrer Entstehung etabliert (Leander, 2020). Die Gesamtzahl der Wissenschaftspodcast-Serien ist zwischen 2004 und 2010 linear gewachsen und hat seit 2010 exponentiell zugenommen (MacKenzie, 2019), was auf eine wachsende Nachfrage nach Wissenschaftsinhalten in Podcast-Form hinweist und Wissenschaftspodcasts als relevantes und beliebtes Medium für die Vermittlung von wissenschaftlichem Wissen und Diskussionen erscheinen lässt. Im Folgenden sollen die Vorteile von Wissenschaftspodcasts für die wissenschaftliche Community aufgezeigt werden. Dazu wird das Feld der Wissenschaftspodcasts zunächst von anderen Themenfeldern der Podcast-Industrie abgegrenzt. Leander (2020) definiert Wissenschaftspodcasts nach MacKenzie, Birch & Weitkamp und Hellermann:

Wissenschaftspodcasts sind Podcasts, die sich mit Themen im Zusammenhang mit dem Forschungsbetrieb und verschiedenen Bereichen der Wissenschaft befassen. Sie dienen der Vermittlung von wissenschaftlichen Kenntnissen und setzen dabei auf wissenschaftliche Glaubwürdigkeit und Rationalität. Sie bieten ein breites Spektrum an Inhalten, die sowohl den wissenschaftlichen Diskurs als Kernelement als auch andere Teilbereiche wie Forschung, Lehre, Infrastruktur und Administration der Wissenschaft abdecken. Dabei werden wissenschaftliche Kenntnisse von praktischem oder Alltagswissen, abgegrenzt, da es rational überprüfbar sein muss. Auch schließen Wissenschaftspodcasts explizit pseudo-wissenschaftliche und nicht-wissenschaftliche Themen aus. Abschließend weißt Leander darauf hin, dass der deutsche Begriff „Wissenschaftspodcast“ über die von MacKenzie mit dem Begriff „science podcast“ definierten Felder Natur- und Technikwissenschaften hinaus auch Geistes-, Sozial- und Kulturwissenschaften mit einschließt (Leander, 2020; MacKenzie, 2019; Birch und Weitkamp, 2010; Hellermann, 2015).

Selbst in Zeiten des wissenschaftlichen Fortschritts und der immer spezialisierteren Forschung sind einige der drängendsten globalen Herausforderungen von multidisziplinärer Natur. Themen wie Klima, Energie, Gesundheit und Migration sind Paradebeispiele dafür. Diese Komplexität verlangt nicht nur nach naturwissenschaftlichen Antworten, sondern wirft gleichzeitig eine Vielzahl von sozialen, ethischen und rechtlichen Fragen auf (Dernbach, 2022). Das Verständnis und die Bewältigung dieser Themen erfordern daher einen interdisziplinären Ansatz, bei dem verschiedene wissenschaftliche Disziplinen zusammenarbeiten müssen. Die Aufgaben der Wissenschaft umfassen nicht nur die Generierung von Wissen, sondern auch, die Kommunikation

von relevantem Wissen in einer verständlichen und zugänglichen Form an die Öffentlichkeit (Weingart und Schulz, 2014). Es geht darum, eine aufgeklärte Meinungsbildung zu ermöglichen und sicherzustellen, dass wissenschaftliche Erkenntnisse für die Gesellschaft nutzbar sind. Eine wesentliche Erkenntnis ist, dass Wissenschaftskommunikation im Dialog stattfinden sollte, sowohl zwischen Wissenschaftlern als auch zwischen der Wissenschaft und der Öffentlichkeit. Dieser Dialog sollte nicht auf traditionelle Kommunikationskanäle beschränkt sein, sondern auch digitale Medien und soziale Netzwerke umfassen (Dernbach, 2022). Hierbei besteht jedoch die Herausforderung, dass viele Wissenschaftler soziale Medien nur zögerlich verwenden und die Möglichkeiten der Öffnung und Partizipation noch nicht ausreichend nutzen (Dernbach, 2022).

In diesem Kontext bieten Podcasts eine vielversprechende Möglichkeit für die Wissenschaft, komplexe Themen sowohl innerhalb der wissenschaftlichen Gemeinschaft als auch an das außerakademisches Publikum zu vermitteln (Dernbach, 2022). Podcasts sind in der Regel frei zugänglich und bieten somit eine offene Plattform für den Wissensaustausch. Dies unterstützt das Prinzip der Open Science[6] und ermöglicht es der Zivilgesellschaft und Forschenden, auf wissenschaftliche Inhalte zuzugreifen, ohne auf kostenpflichtige Veröffentlichungen angewiesen zu sein. Durch den informellen und dialogorientierten Charakter von Podcasts können Forscher komplexe Konzepte anschaulich erklären und in kleinen Paketen episodisch veröffentlichen. Sie erfordern im Vergleich zu anderen Medienformen relativ geringe Ressourcen und haben somit niedrigere Einstiegshürden für die Produktion, was es jungen Wissenschaftlern und unterrepräsentierten Gruppen in der Wissenschaft, sei es aufgrund von Fachgebiet, Geschlecht, Ethnie oder anderen Faktoren, ermöglicht, ihre Forschung und Perspektiven einem breiten Publikum vorzustellen (Quintana und Heathers, 2020). Damit könnten Podcast dazu beitragen, den so genannten „Matthew-Effekt“[7] in der Wissenschaft zu mildern, indem sie eine gerechtere Chance für alle bieten, ihre Arbeit zu präsentieren und gehört zu werden (Quintana und Heathers, 2020).

Somit sind Wissenschaftspodcasts eine vielversprechende Plattform für die wissenschaftliche Community, um komplexe Themen interdisziplinär zu behandeln, einen offenen Dialog mit der Gesellschaft zu führen, gleichzeitig den Zugang zu Forschungsinhalten für ein breiteres Publikum zu erleichtern und die Grundsätze der Open Science zu fördern.

2.1.4 Herausforderungen für Wissenschaftspodcasts

Insgesamt stellen Wissenschaftspodcasts ein wertvolles Mittel zur Wissenschaftskommunikation dar, welches jedoch mit einigen problematischen Aspekten einher geht, die sorgfältig angegangen werden müssen, um sicherzustellen, dass sie ihr Potenzial als effektives Kommunikationsmittel ausschöpfen können. Im Folgenden werden zunächst einige generelle Aspekte

[6] Open Science ist eine wissenschaftliche Praxis, die auf Offenheit, Transparenz und Zusammenarbeit basiert. Sie umfasst die Freigabe von Forschungsdaten, Methoden und Ergebnissen, um einen breiten Zugang zur wissenschaftlichen Gemeinschaft und der Öffentlichkeit zu ermöglichen. Dies fördert die Reproduzierbarkeit von Studien, beschleunigt den wissenschaftlichen Fortschritt und verbessert die Qualitätssicherung, da eine breite Gemeinschaft von Forschern die Ergebnisse prüfen kann (Vicente-Saez und Martinez-Fuentes, 2018)

[7] Im Kontext von Wissenschaft und Podcasts bedeutet der Matthew-Effekt, dass bereits renommierte Wissenschaftler oder etablierte wissenschaftliche Institutionen in der Regel leichter Zugang zu traditionellen Medien und Kommunikationskanälen haben, um ihre Forschung zu präsentieren. Sie können sich aufgrund ihrer Bekanntheit leichter Gehör verschaffen (Petersen et al., 2011).

beleuchtet. Anschließend werden die Aspekte der Auffindbarkeit von Podcasts und der Qualität genauer betrachtet, da diese im Kontext der Wissenschaftskommunikation eine besondere Hürde darstellen.

Wissenschaftspodcasts sehen sich mit einer wachsenden Konkurrenz konfrontiert, insbesondere von etablierten und finanzstarken Medienunternehmen, welche Podcasts zunehmend in ihr Portfolio aufnehmen, und über die personellen, technischen und finanziellen Ressourcen verfügen, um qualitativ hochwertige Podcasts zu produzieren (Dernbach, 2022; Newman et al., 2023). Dies stellt eine Gefahr für unabhängige Podcaster dar, die möglicherweise nicht über die gleichen Mittel verfügen, um mit dieser Konkurrenz Schritt zu halten. Des Weiteren verfügen Podcasts normalerweise über keine direkten Interaktionsmöglichkeiten, was den wissenschaftlichen Diskurs und die Möglichkeit, Fragen zu klären oder Missverständnisse auszuräumen, einschränken kann. Um den dialogorientierten Charakter von Podcasts voll auszuschöpfen, müssten Podcast-Produzenten und -Moderatoren alternative Wege zur Interaktion mit ihrem Publikum entwickeln, wie beispielsweise die Integration von Online-Foren oder sozialen Medienplattformen in ihren Podcast-Communitys, um Diskussionen anzuregen und Fragen von Hörern in Echtzeit zu beantworten. Darüber hinaus könnten Live-Podcast-Aufzeichnungen mit interaktiven Q&A-Sitzungen oder Interviews mit Experten organisiert werden, um den Zuhörern die Möglichkeit zu geben, direkt mit den Produzenten und anderen Experten in Kontakt zu treten. Die auditive Natur von Podcasts stellt zudem ein Problem für die Zugänglichkeit von Podcasts dar. Nicht alle Menschen haben gleichermaßen Zugang zu dieser Form der Wissenschaftskommunikation, insbesondere diejenigen, die gehörlos beziehungsweise schwerhörig sind oder keinen Zugang zu den erforderlichen technischen Geräten haben oder schlichtweg das Lesen von Texten bevorzugen. Dies könnte zu einer ungleichen Verteilung von Wissenschaftsinformationen führen und einige Bevölkerungsgruppen ausschließen. Alleine in Deutschland ist in etwa jeder fünfte schwerhörig (Schöne, 2021a).

Neben den generellen Herausforderungen, spielt auch die Qualität von Informationen in der Welt der Wissenschaftspodcasts eine entscheidende Rolle, welche von verschiedenen Faktoren beeinflusst wird. Eins der Hauptprobleme besteht darin, qualitativ hochwertige und glaubwürdige Informationen von der Fülle der verfügbaren Podcasts zu unterscheiden (Besser et al., 2010). Dies wird besonders schwierig, wenn die Podcaster keine etablierten Wissenschaftler oder Institutionen repräsentieren und es schwer ist, ihre Expertise und Verlässlichkeit zu überprüfen (Fähnrich et al., 2023). Dieses Problem kann zu Fehlinformationen führen, da Hörer möglicherweise unwissenschaftlichen Behauptungen Glauben schenken. Auf der anderen Seite können Podcasts von bestimmten Unternehmen oder Organisationen finanziert werden, die ihre Unabhängigkeit und Objektivität beeinträchtigen könnten. Dies kann dazu führen, dass Wissenschaftspodcasts eine gewisse Agenda verfolgen oder eine Voreingenommenheit entwickeln, die ihre Berichterstattung und Interpretation wissenschaftlicher Informationen beeinflusst (Dernbach, 2022). Ein weiteres Problem ist die Tatsache, dass nicht alle Podcasts, die sich als wissenschaftlich ausgeben, tatsächlich auf wissenschaftlichen Erkenntnissen basieren. Einige könnten pseudowissenschaftliche Ideen oder Verschwörungstheorien fördern, was zu einer Verbreitung von irreführenden oder falschen Informationen führen kann. Im Kontrast zu wissenschaftlichen Veröffentlichungen mangelt es Podcasts in der Regel an einer formalen

Peer-Review. Dies bedeutet, dass die Informationen, die in Podcasts präsentiert werden, nicht den gleichen Prüfprozessen durch Fachleute unterzogen werden, wie es bei wissenschaftlichen Fachzeitschriften der Fall ist. Das Fehlen etablierter Instanzen, die Podcast-Inhalte mit der notwendigen Expertise überprüfen können, stellt einen zentralen Kritikpunkt dar und kann die Qualität und Genauigkeit der vermittelten Informationen weiterhin beeinträchtigen. Die Grundlage einer vertrauenswürdigen Wissenschaft liegt in der Evidenz und den wissenschaftlichen Erkenntnissen (Dernbach, 2022). Das Vertrauen in wissenschaftliche Informationen ist von entscheidender Bedeutung, um die Akzeptanz von wissenschaftlich belegten Lösungen für globale Probleme wie den Klimawandel, die Gesundheitsversorgung und die Bewältigung von Krisen sicherzustellen. Jedoch kann das Vorhandensein von Wissenschaftspodcasts, die fragwürdige Informationen verbreiten, das Vertrauen in die Wissenschaft untergraben. Wenn Hörerinnen und Hörer unwissenschaftlichen Behauptungen Glauben schenken oder pseudowissenschaftlichen Ideen ausgesetzt sind, kann dies zu Skepsis gegenüber etablierten wissenschaftlichen Erkenntnissen führen. Diese Verwirrung und Unsicherheit können die Fähigkeit der Gesellschaft, wissenschaftliche Lösungen zu akzeptieren und umzusetzen, erheblich beeinträchtigen (Hameleers et al., 2022). Die Produzenten von Wissenschaftspodcasts sollten sich bewusst sein, dass sie dazu beitragen, das Vertrauen in die Wissenschaft aufzubauen oder zu untergraben. Es liegt in ihrer Verantwortung sicherzustellen, dass ihre Inhalte auf wissenschaftlichen Erkenntnissen basieren, objektiv sind und qualitativ hochwertige Informationen liefern.

Eine weitere Herausforderung stellt die Auffindbarkeit von Podcasts und damit der Zugang zu wissenschaftlichen Informationen dar. Da Podcasts hauptsächlich aus gesprochenem Wort bestehen und Suchmaschinen in der Regel auf Text basieren, ergeben sich mehrere Hürden bei der gezielten Informationssuche (Jones et al., 2021), was die Reichweite von weniger populären Formaten beeinträchtigen kann. Obwohl es Tools und Plattformen gibt, die Podcasts indizieren und durchsuchbar machen, sind diese nicht immer weit verbreitet oder effektiv. Selbst wenn Podcasts auf diesen Plattformen verfügbar sind, kann ihre Auffindbarkeit immer noch eingeschränkt sein (Dziatzko, 2023). Die Suche nach Informationen zu einem bestimmten Thema in verschiedenen Podcasts kann sehr zeitaufwändig sein, da jede Podcast-Plattform ihre eigene Suchfunktion hat und zentrale Suchmöglichkeit für Podcasts wie Listen Notes mit den Daten arbeiten, die vom Podcast-Ersteller zur Verfügung gestellt werden. Aufgrund der Struktur von Podcasts stehen zur Indexierung zwei Möglichkeiten zur Verfügung, welche auch ergänzend eingesetzt werden können. Einerseits können Podcasts über Metadaten wie Autor, Titel, Beschreibung, Schlagwörter (Tags), Kategorie, Veröffentlichungsdatum und mehr abgerufen werden. Die Qualität der Metadaten, die mit Podcasts verknüpft sind, kann jedoch stark variieren (Besser et al., 2010; Jones et al., 2021). Unvollständige oder ungenaue Metadaten können die Suche und den Abruf von Inhalten weiter behindern, da sie nicht ausreichend Informationen über den Inhalt der Podcasts liefern. Hierbei könnte die Etablierung von Ressource Description Framework (RDF)[8] ähnlichen Strukturen hilfreich sein (Celma und Raimond, 2008). Auf der anderen Seite kann der Inhalt als Kernstück der Episode verwendet werden, um inhaltsbasierte

[8]Das RDF-Format ist ein standardisiertes Datenformat zur Darstellung von strukturierten Informationen über Ressourcen im Internet oder in anderen Datenbanken. RDF wird häufig verwendet, um Metadaten, Ontologien, Wissensgraphen und semantische Daten zu modellieren und auszutauschen. RDF ermöglicht es, Beziehungen zwischen Ressourcen auf eine semantische Weise zu beschreiben, wodurch maschinenlesbare Informationen erstellt werden können (Vgl.: https://www.w3.org/RDF/ zuletzt abgerufen am 19. Oktober 2023)

Suchen zu ermöglichen (Besser et al., 2010). Dies ist insbesondere nützlich, wenn der Autor oder der Name des Podcast nicht bekannt ist, thematisch jedoch Überschneidungen zur Suche vorhanden sind. Zur Erschließung des Inhalts von Podcasts ist eine ausführliche Beschreibung des Inhalts notwendig beziehungsweise die Transkription des Inhalts. Obwohl die Marktführer der Podcast-Industrie mittlerweile auf automatische Transkriptionen von Podcast Episoden zur Erhöhung der Reichweite umgestiegen sind, fehlt dies häufig noch bei kleineren Produktionen und im Bereich der Wissenschaftspodcasts, wodurch diese zusätzlich benachteiligt werden. Auch bei Podcasts dominieren einige wenige Shows den Markt, ähnlich wie bei Film und Musik, weshalb alle verfügbaren Informationen zur Indexierung genutzt werden sollten, um die Vielfältigkeit der Podcast-Landschaft zu erhalten und die Reichweite der eigenen Produktionen zu erhöhen (Jones et al., 2021). Ein weiteres Problem ergibt sich aus der Länge von Podcast-Episoden, die oft stundenlang sein können. Dies erschwert auch die gezielte Suche nach Informationen. Hörer müssen möglicherweise eine beträchtliche Zeit investieren, um die gewünschten Informationen in einer Episode zu finden, insbesondere wenn sie nicht genau wissen, wo sie suchen sollen.

Die Bewältigung dieser Herausforderungen erfordert nicht nur die Weiterentwicklung technischer Lösungen zur besseren Indexierung und Durchsuchbarkeit von Podcasts, sondern auch eine verstärkte Zusammenarbeit zwischen Podcast-Produzenten, Plattformen, und Suchmaschinenbetreibern, um die Auffindbarkeit von qualitativ hochwertigen Wissenschaftspodcasts zu verbessern. Darüber hinaus ist die Förderung von Standards für Metadaten und Transkriptionen sowie die Bereitstellung von Ressourcen für Podcast-Ersteller entscheidend, um die Qualität und Zuverlässigkeit der Informationen sicherzustellen.

2.2 Automated Speech Recognition

2.2.1 Grundlagen von ASR

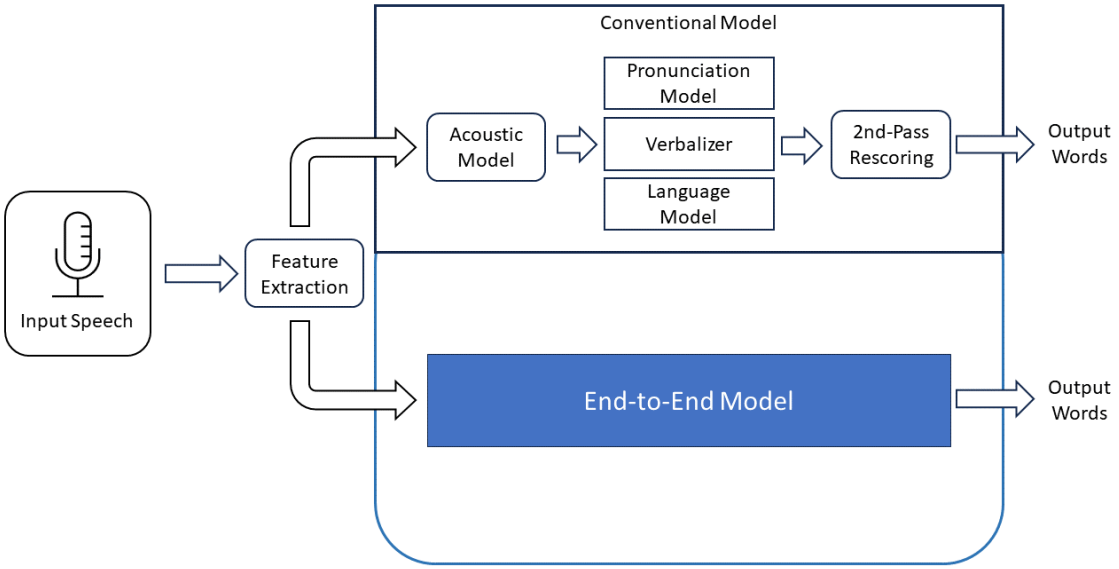
In Anbetracht der wachsenden Bedeutung von Wissenschaftspodcasts und den damit verbundenen Herausforderungen wird klar, dass Podcasts nicht nur als unterhaltsame Informationsquellen dienen, sondern auch einen bedeutenden Beitrag zur Förderung des wissenschaftlichen Diskurses bieten können. Um diese Erkenntnisse jedoch effektiv zu nutzen, ist es entscheidend, die Zugänglichkeit und Integrität dieser Inhalte sicherzustellen. Hier kommen ASR-Systeme ins Spiel, die als vielversprechende Lösungen zur Bewältigung dieser Herausforderungen dienen können. ASR, auch als Speech-to-Text (STT) bezeichnet, ist eine Technologie, die nahtlos die Lücke zwischen menschlichen Sprachwellenformen und geschriebenem Text überbrückt. Sie ermöglicht die Umwandlung gesprochener Sprache in ein Format, das Maschinen verstehen und verarbeiten können, und bietet so ein leistungsstarkes Kommunikationsmittel (Xiong, 2023). ASR bietet nicht nur eine wertvolle Lösung zur Transkription von gesprochenem Inhalt, sondern ermöglicht es auch Menschen, auf die natürlichste und intuitivste Weise mit Maschinen zu interagieren (Xiong, 2023). Es eröffnet eine Welt, in der gesprochene Worte mühelos in digitale Sprache übersetzt werden können, wodurch die Zugänglichkeit verbessert und die Interaktion zwischen Mensch und Maschine erleichtert wird. ASR ist ein komplexer Prozess, der in drei Schritte unterteilt werden kann, um gesprochene Sprache in geschriebenen Text umzuwandeln (Xiong, 2023). Der erste Schritt ist die Sprechpausenerkennung, die auch als Voice Activity

Detection (VAD) bezeichnet wird. Diese Technik identifiziert die Stellen in einem Audioaufzeichnung, an denen tatsächlich gesprochene Worte vorhanden sind, und trennt sie von den Stillephasen, die nicht zur Transkription beitragen (Ramirez et al., 2007). Der zweite Schritt ist die Unterteilung eines gesprochenen Satzes in einzelne Untergruppen, wie Wörter oder phonetische Laute. Dieser Vorgang wird als Segmentation bezeichnet und ist entscheidend, um die richtigen Abgrenzungen zwischen den Worten oder Lauten festzulegen. Es ermöglicht, den Text in verständliche Einheiten zu gliedern und die Zuordnung von Lauten zu den entsprechenden Wörtern sicherzustellen (Huang et al., 2023). Schließlich erfolgt die Klassifikation der einzelnen phonetischen Laute. Hierbei handelt es sich um die Zuordnung der identifizierten Laute zu den entsprechenden Lauten im Alphabet oder zur entsprechenden Schrift, um den Text korrekt zu transkribieren. Diese Phase ist komplex, da Sprachen verschiedene Aussprachen und Betonungen haben können (Karpagavalli und Chandra, 2016).

Traditionell beruhen ASR-Modelle auf linguistischen Regeln um die Wahrscheinlichkeit von Wörtern oder Lauten in einer gesprochenen Sequenz zu schätzen. Diese Systeme verwenden normalerweise Hidden Markov Models (HMMs) in Kombination mit Gaussian Mixture Models (GMMs) für die akustische Modellierung und Sprachmodelle, die auf N-Grammen basieren (Xiong, 2023). Die jüngste Entwicklung in der Welt der automatischen Spracherkennung markiert jedoch einen Paradigmenwechsel. Der erhebliche Fortschritt im Bereich Deep Learning, insbesondere im Bereich neuronaler Netzwerke führte zu einer dramatischen Veränderung der Landschaft der ASR. Neue Systeme verwenden komplexe neuronale Netzwerke, darunter Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs) und Transformer-Netzwerke, um gesprochene Sprache mit beeindruckender Genauigkeit zu transkribieren (Karpagavalli und Chandra, 2016; Vaswani et al., 2017). Es wird zwischen zwei Modell Arten unterschieden: Hybride Modelle (im Folgenden auch als konventionell bezeichnet), die als Mischung aus traditionellen und Deep Learning-Methoden fungieren und End-to-End (E2E) Modelle (Xiong, 2023). Abbildung 2.1 zeigt die grobe Struktur beider Modell Arten. Ein wesentlicher Vorteil von E2E-Modellen ist ihre vergleichsweise simple Struktur. Im Gegensatz zu konventionellen Modellen, die verschiedene Schritte zur akustischen Modellierung und Sprachmodellierung erfordern, eliminieren E2E-Modelle diese Komplexität. Sie vereinen die gesamte ASR-Pipeline in einer einzigen Architektur (Speech und in the Northeast (SANE), 2023). Somit ist die Optimierung des Modells nicht mehr von separaten Komponenten abhängig, sondern nur noch von einer einzigen Ziel- und Verlustfunktion (Li, 2022). Diese Kohärenz in E2E-Modellen verringert den Aufwand und die Expertise, die für die Optimierung notwendig ist, enorm. Die Vereinheitlichung des Optimierungsprozesses macht das Training und die Pflege von E2E-Modellen effizienter und leichter handhabbar. Da alle Teile des Netzwerks darauf ausgerichtet sind, ein gemeinsames Ziel zu erreichen, wird die Fähigkeit des Modells verbessert, den Kontext und die Abhängigkeiten in der gesprochenen Sprache effizienter zu erfassen (Li, 2022). Dies führt zu einer höheren Genauigkeit und Leistung bei der Spracherkennung und ermöglicht die Bewältigung komplexer Sprachmuster, was das Modell insgesamt robuster macht (Speech und in the Northeast (SANE), 2023). E2E-Modelle benötigen auch weniger aufwendige Vorverarbeitung und Vorbereitung der Daten. Dies führt zu einem reduzierten Schulungsaufwand, was Zeit und Ressourcen spart (Speech und in the Northeast (SANE), 2023). Die Flexibilität von E2E-Modellen ist ein weiterer Pluspunkt. Sie können für verschiedene Aufgaben eingesetzt werden, darunter Transkriptionen,

Abbildung 2.1

Konventionelles Modell vs End-to-End Modell



Anmerkung. Adaptiert von: SANE2022 | Tara Sainath - End-to-End Speech Recognition: The Journey From Research to Production von Speech and Audio in the Northeast (SANE). (2023). https://www.youtube.com/watch?v=FvkLYRpBIe0

Übersetzungen und Spracherkennung. Dies macht sie äußerst vielseitig und reduziert den Bedarf an verschiedenen spezialisierten Modellen. Weiterhin konnte die Leistung von E2E-Modellen, durch die Etablierung der Transformer-Architektur verbessert werden. Diese wurde 2017 in dem bahnbrechenden Paper „Attention is All You Need“ vorgestellt (Vaswani et al., 2017). Diese Arbeit führte die Idee der „Self-Attention“ ein, bei der jedes Token in einer Sequenz Aufmerksamkeit auf alle anderen Tokens richten kann, was zu erheblichen Verbesserungen von Deep Learning Modellen führte. Obwohl E2E-Modelle zweifellos viele Vorteile bieten, ist es wichtig zu beachten, dass hybride Modelle in bestimmten Nischen-Disziplinen immer noch relevant sein können. Die Wahl zwischen E2E und hybriden Modellen hängt von den spezifischen Anforderungen eines Projekts und den verfügbaren Ressourcen ab.

2.2.2 Architektur von Speech Transformern am Beispiel von Whisper

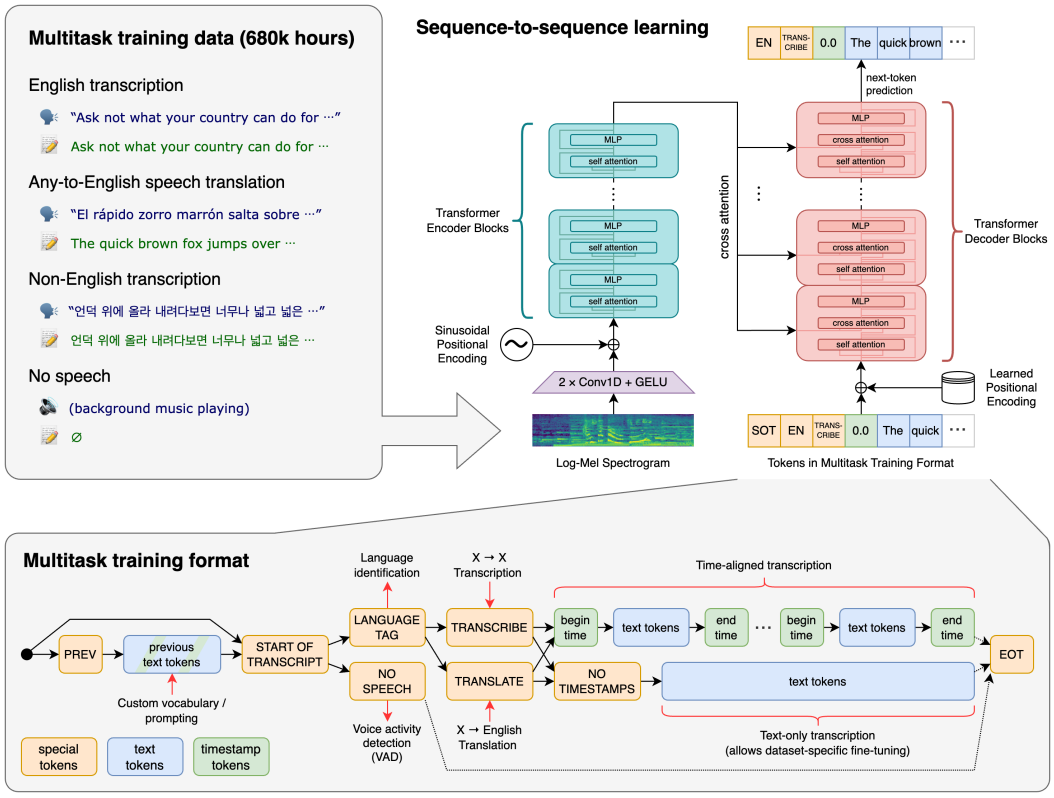
Die Entwicklung von Transformern hat sich als wichtiger Fortschritt erwiesen. Ein herausragendes Beispiel für eine solche Architektur ist das Speech Transformer Modell „Whisper“[9] von OpenAI[10]. Whisper wurde auf einer Trainingsdatenmenge von 680.000 Stunden entwickelt, die aus verschiedenen Quellen im Internet stammt, darunter LibriSpeech, TED-LIUM 3, Common Voice 5.1 und 9, Fleurs und viele Weitere[11]. Whisper ist in erster Linie als ein General-Purpose-Modell für ASR konzipiert. Obwohl Whisper bereits eine beeindruckende Leistung

[9] https://github.com/openai/whisper
[10] https://openai.com/
[11] https://github.com/openai/whisper/tree/main/data

in der Umsetzung von Sprach-zu-Text-Übersetzungen erbringt, ist sein Mehrzweckcharakter ein entscheidender Faktor, der es Anwendern ermöglicht, das Modell an spezifische Anforderungen anzupassen. Da Whisper als Open-Source-Modell verfügbar ist, kann das Modell mit wenigen Daten und etwas Fine-Tuning auf Basis des ursprünglichen Quellcodes auf spezielle Anwendungsfälle, wie etwa bestimmte Dialekte, angepasst werden (Radford et al., 2022). Zudem erweitert es den Anwendungsbereich über die reine Erkennung der englischen Sprache hinaus, indem es 117.000 Stunden in 96 anderen Sprachen sowie Multitasking-Aufgaben wie die Identifikation von Sprachaktivitäten und die Umkehrung der Textnormalisierung einbezieht (Radford et al., 2022). Obwohl Whisper in Teildisziplinen von speziell darauf zugeschnittenen Modellen übertroffen wird, bietet es dennoch eine ausreichend hohe Genauigkeit für die Transkription von Audio-Dateien. Im Test erzielt das Whisper-Modell „large-v2“ eine Word Error Rate (WER) von 4,2% für Englisch und 4,5% für Deutsch (Radford et al., 2022). Zum Vergleich: Die aktuelle Spitze des LibriSpeech-Leaderboards erzielte auf dem *test_clean* Set eine WER von 1,4% („LibriSpeech Test-Clean Benchmark (Speech Recognition)“, n. d.). Weitere Tests ergaben, dass die Spracherkennung auf einem Niveau ist, die in etwa mit der eines Menschen vergleichbar ist (durchschnittliche WER Whisper auf verschiedenen Datensätzen: 8,81% bis 12,2%; durchschnittliche WER manueller Transkripte auf verschiedenen Datensätzen: 7,61% bis 10,5% (Radford et al., 2022)).

Abbildung 2.2

Whisper-Modell Architektur



Anmerkung. Übernommen aus: *Robust Speech Recognition via Large-Scale Weak Supervision* von Radford et al. (2022).

Whisper hat sich durch seine bemerkenswerte Fähigkeit, akustische Signale direkt in Text umzuwandeln, als eines der führenden Modelle auf dem Gebiet der ASR etabliert. Die Architektur von Whisper (siehe Abbildung 2.2) ist geprägt von modernen Deep Learning-Techniken und zeichnet sich durch die Verwendung des Transformer-Modells aus. In diesem Abschnitt werden die Schlüsselkomponenten und das Funktionsprinzip von Speech Transformern am Beispiel von Whisper erläutert.

ASR ist eine Sequence to Sequence (Seq2seq) Aufgabe. Eine Art von maschinellem Lernproblem, bei dem eine Modellsequenz (eine geordnete Abfolge von Datenpunkten) variabler Länge in eine andere Modellsequenz variabler Länge umgewandelt wird. Das Ziel ist es, eine Zuordnung zwischen den Eingabesequenz-Elementen und den Ausgabesequenz-Elementen herzustellen (Sutskever et al., 2014; Zhang et al., 2021). Zu Beginn der Pipeline werden Audio-Daten im Rohformat (siehe Abbildung 2.3) in 30-sekündige Abschnitte unterteilt und in ein Log-Mel Spektrogramm umgewandelt (Radford et al., 2022). Audiosignale sind von Natur aus kontinuierlich, was sie für Transformer-Modelle zu einer Herausforderung macht, da diese auf diskrete Daten spezialisiert sind. Um diese Kluft zu überbrücken, wird oft ein Spektrogramm verwendet (Velardo, 2020). Dieser Schritt kann auch als „Feature-Extraction“ betrachtet werden. Ein Spektrogramm (siehe Abbildung 2.4) ist eine Darstellung des Frequenzspektrums eines Audiosignals über die Zeit. Diese Darstellung ermöglicht es, wichtige Informationen über die Veränderungen

Abbildung 2.3

Audio-Signal

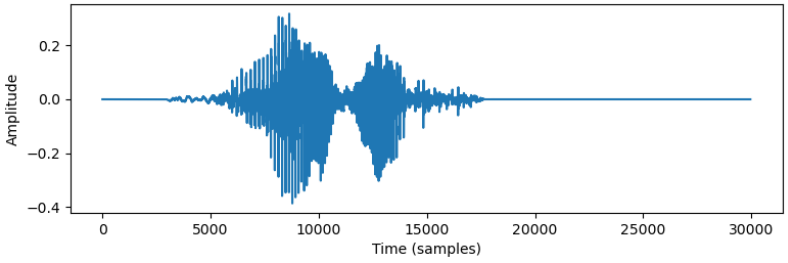


Abbildung 2.4

Spektrogramm

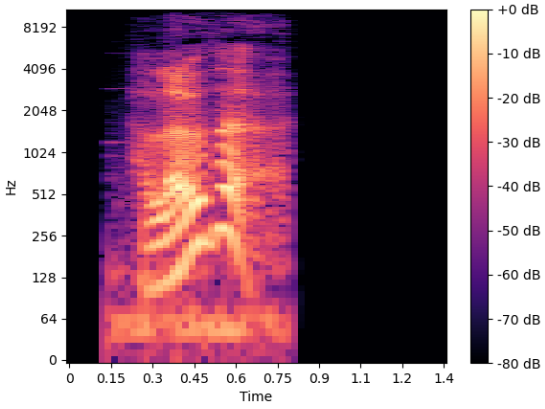
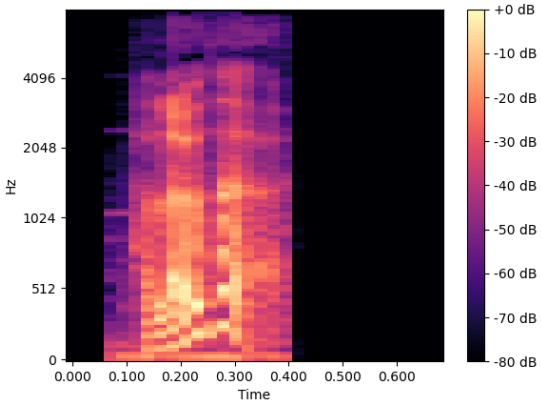


Abbildung 2.5

Mel Spektrogramm

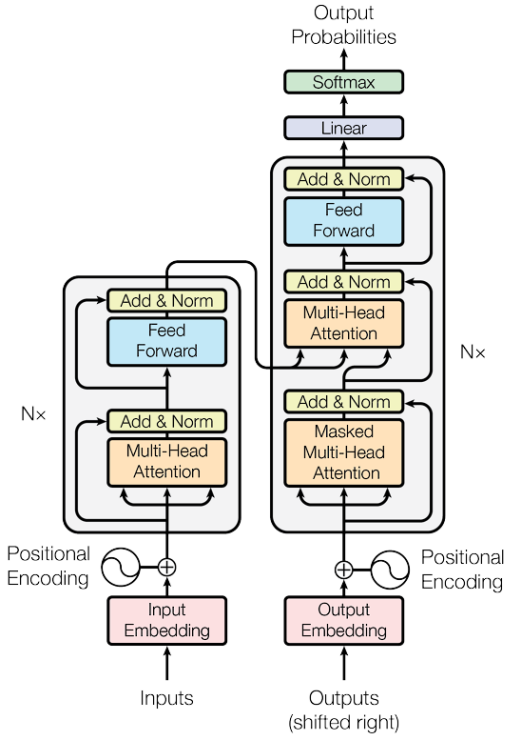


der akustischen Merkmale im Laufe der Zeit zu extrahieren. Ein Spektrogramm ist im Wesentlichen eine Heatmap, in der die X-Achse die Zeit und die Y-Achse die Frequenzen repräsentiert (Velardo, 2020). Die Farbintensität in dieser Darstellung gibt Aufschluss über die Amplitude der jeweiligen Frequenzen zu einem bestimmten Zeitpunkt (Velardo, 2020). Allerdings ist zu beachten, dass Spektrogramme die Frequenzen in linearer Form darstellen. Menschen hingegen nehmen Frequenzen logarithmisch wahr, was bedeutet, dass Unterschiede in den tieferen Frequenzen stärker wahrgenommen werden als in den höheren (Velardo, 2020). Aus diesem Grund wird häufig das Mel-Spektrum verwendet. Das Mel-Spektrogramm (siehe Abbildung 2.5) ist eine spezielle Form des Spektrogramms, bei dem die Frequenzen in eine logarithmische Skala, die als Mels bezeichnet wird, umgewandelt werden. Dies führt zu einer höheren Übereinstimmung zwischen der menschlichen Wahrnehmung und der Darstellung im Spektrogramm (Velardo, 2020). Damit wird das Mel-Spektrum zu einer nützlichen Grundlage für die Verarbeitung von Audiodaten in Transformer-Modellen, die auf solchen diskreten Sequenzen basieren (Velardo, 2020). Das Log-Mel Spektrogramm wird an zwei Conv1D-Layers (Convolutional Layers) übergeben. Conv1D-Layer sind in Anwendungen relevant, bei denen die räumliche Beziehung zwischen Datenpunkten zählt, wie bei der Verarbeitung von Audio-Signalen. Sie ermöglichen dem Netzwerk, Muster und Merkmale in aufeinanderfolgenden Positionen zu erkennen. Während des Trainings erlernt das Netzwerk die relevanten Parameter zur Merkmalserkennung in den Daten (Zhang et al., 2021). Dies wird in CNNs häufig zur Filterung verwendet. Die Dimension der Input-Daten wird auf die Dimension der Output-Daten reduziert (Lin et al., 2014). Als Aktivierungsfunktion wird die Gaussian Error Linear Unit (GELU) Funktion eingesetzt, welche gegenüber anderen Aktivierungsfunktionen eine bessere Approximation von nichtlinearen Funktionen bietet. Die Verwendung von GELU trägt zur Verbesserung der Trainingseffizienz und zur Förderung der Modellgeneralisierung bei (Hendrycks und Gimpel, 2023). Der Output der Convulutional Layers wird als Input Embedding an den Encoder übergeben. Der Encoder hat die Aufgabe Inputsequenzen in eine verborgene Repräsentation umzuwandeln, um Informationen auf einer abstrakten, maschinenlesbaren Ebene darzustellen, in der möglichst alle Informationen der Sequenz kompakt in einem Vektorraum dargestellt werden (Q. Wu et al., 2022).

Whisper verwendet die gleiche Encoder-Decoder-Struktur von E2E-Modellen, wie sie bereits von Vaswani et al. (2017) vorgestellt wurde (Radford et al., 2022). Diese ist in Abbildung 2.6 dargestellt. Die Architektur des Transformer-Modells weist Ähnlichkeiten zu herkömmlichen RNNs auf, unterscheidet sich jedoch in einem entscheidenden Aspekt: die Möglichkeit, Inputsequenzen parallel zu verarbeiten, anstatt sequentiell. Dieser Unterschied erfolgt durch die Einführung von Self-Attention und macht den Transformer besonders effizient (Vaswani et al., 2017). Grundlage für den Encoder des Transformers ist das „Input Embedding“. In diesem Schritt werden die Eingabedaten in eine Form umgewandelt, die vom Modell verarbeitet werden kann (CodeEmporium, 2020; StatQuest with Josh Starmer, 2023). Dies erfolgt durch die Umwandlung von Text oder Symbolsequenzen (im Fall von Log-Mel-Spektrogrammen durch die Umwandlung von Frames in Muster) in numerische Vektoren. Das Ergebnis ist eine hochdimensionale Darstellung der Eingabesequenz, bei der Sequenzen mit ähnlicher Bedeutung räumlich nahe beieinander liegen (CodeEmporium, 2020; StatQuest with Josh Starmer, 2023). Jedes einzigartige Muster wird dabei mit einem eindeutigen numerischen Token versehen. Die Anordnung erfolgt in der Trai-

Abbildung 2.6

Transformer-Modell Architektur



Anmerkung. Übernommen aus: *Attention Is All You Need* von Vaswani et al. (2017).

ningsphase des Modells. Bedeutung ist jedoch oft kontextabhängig. Um dies zu berücksichtigen, wird der Input-Vektor mit Informationen zur Abfolge der Inputsequenz erweitert. Dieser Schritt erfolgt durch einen Positional Encoder, der sicherstellt, dass der Input-Vektor die Position eines Frames innerhalb der Eingabesequenz berücksichtigt (CodeEmporium, 2020; StatQuest with Josh Starmer, 2023). Dadurch entsteht ein Input-Vektor, der sowohl semantische Informationen als auch die räumliche Anordnung innerhalb der Sequenz enthält. Schließlich werden die Input-Vektoren in einer Matrix zusammengeführt und an den Encoder-Block des Transformers weitergegeben. Dies ermöglicht es dem Modell, die Daten zu analysieren und Muster sowie Abhängigkeiten in den Inputsequenzen zu erlernen (CodeEmporium, 2020; StatQuest with Josh Starmer, 2023).

Der Encoder ist in mehrere Blöcke unterteilt, wobei jeder dieser Encoder-Blöcke aus verschiedenen „Layers“ (Layer, zu deutsch: Schicht) besteht. Jede Schicht spielt eine wichtige Rolle bei der Verarbeitung von Eingabedaten und dem Erlernen von relevanten Informationen. Ein zentraler Bestandteil eines Encoder-Blocks ist die „Multi-Head Attention Layer“. Diese Schicht ermöglicht es dem Netzwerk, die Aufmerksamkeit auf relevante Teile der Eingabedaten zu richten und Muster sowie Beziehungen zwischen den Daten effektiv zu erfassen (Vaswani et al., 2017). In den Encoder-Blöcken wird diese Form der Aufmerksamkeit auch als self-attention bezeichnet, da sie Informationen innerhalb derselben Eingabesequenz verwendet (Vaswani et al., 2017). Am Beispiel von Log-Mel Input Daten, vergleicht die Attention Layer jeden einzelnen Frame

mit den Frames derselben Sequenz und berechnet einen Attention Vektor, der anzeigt, wie relevant der Frame im Kontext zu den anderen Frames ist. Dabei ergibt sich jedoch ein Problem: Der Attention Vektor neigt für jedes Element einer Sequenz dazu, die Beziehung zu sich selbst am stärksten zu gewichten, was wenig aussagekräftig ist. Aus diesem Grund werden mehrere Attention Vektoren pro Element erzeugt, was der Schicht den Namen „Multi-Head Attention Layer“ verleiht (Vaswani et al., 2017). Ein gewichteter Durchschnitt wird berechnet, um diesen Effekt abzumildern. Zusammenfassend kann gesagt werden, dass die Attention Layer dazu dient, Beziehungen zwischen einzelnen Elementen einer Sequenz zu erfassen. Jede Attention Layer schließt mit einer Feed Forward Layer ab (Vaswani et al., 2017). Die Gewichtungen und Parameter in dieser Schicht werden während des Trainingsprozesses des Modells optimiert, um die bestmögliche Beziehung zwischen den Eingabedaten und den Ausgabedaten zu lernen (Ojha et al., 2017). Am Ende einer jeden Schicht werden die „Residual Connection“ und die „Layer Normalization“ durchgeführt. Die Residual Connection fügt eine Kopie des ursprünglichen Layer-Inputs zum Layer-Output hinzu. Dies ermöglicht dem Modell, weiterhin mit den Inputsequenzen zu arbeiten, ohne dass diese vollständig neu erstellt werden müssen (Khan et al., 2020). Die Layer Normalization normalisiert jeden Vektor in der Schicht unabhängig voneinander und trägt dazu bei, die Stabilität des Modells während des Trainings zu gewährleisten (Ba et al., 2016). Insgesamt spielen die verschiedenen Schichten innerhalb eines Encoder-Blocks eine koordinierte Rolle, um komplexe Abhängigkeiten in den Daten zu modellieren und die Repräsentation der Eingabedaten schrittweise zu verbessern. Der Output des Encoders wird in Form eines „Hidden Embeddings“ an den Decoder übergeben.

Der Decoder im Transformer-Modell übernimmt die Aufgabe, Informationen, die vom Encoder bereitgestellt werden, zu nutzen, um eine Ausgabesequenz zu generieren (Q. Wu et al., 2022). Während der Decoder einige Ähnlichkeiten mit dem Encoder aufweist, gibt es entscheidende Unterschiede in seiner Funktionsweise. Während der Trainingsphase wird dem Decoder die zugehörige Outputsequenz, auch Zielsequenz genannt, der jeweiligen Inputsequenzen übermittelt. Die Outputsequenz durchläuft ähnlich wie die Inputsequenz im Encoder das Input Embedding und Positional Encoding. Dies ermöglicht dem Modell, die räumliche Position der Elemente in der Sequenz zu berücksichtigen (Vaswani et al., 2017). Eine grundlegende Differenz zum Encoder besteht jedoch darin, dass der erste Decoder-Block eine „Masked Multi-Head Attention Layer“ verwendet. Die Verwendung dieser Maskierung ist von entscheidender Bedeutung, da sie sicherstellt, dass das Modell die Daten nicht auswendig lernt (Overfitting). Da die Elemente der Sequenz parallel betrachtet werden, ohne Kenntnis der Reihenfolge, werden Elemente, die in der Sequenz sequenziell nachfolgen, maskiert, indem sie auf null gesetzt werden (Vaswani et al., 2017). Nach der Bildung der Attention Vektoren wird der Output an den nächsten Decoder-Block weitergeleitet. Während die bisherigen Encoder- und Decoder-Blöcke self-attention verwendeten, bei dem die betrachteten Elemente alle aus derselben Sequenz stammten (im Encoder die Inputsequenz und im Decoder die Outputsequenz), implementiert der nächste Decoder-Block die sogenannte „cross-attention“ oder „Encoder-Decoder Attention“ (Vaswani et al., 2017). Hierbei werden die Attention Vektoren dieses Blocks aus den Vektoren des Encoder-Outputs und des vorherigen Decoder-Blocks zusammengesetzt. Dies ermöglicht es dem Modell, Beziehungen zwischen den Input- und Outputsequenzen zu bilden (Vaswani et al., 2017). Wie bereits beim Encoder werden bei jeder Schicht des Decoders Residual Connection und Layer Normalization

angewendet. Das Endergebnis der letzten Decoder Schicht ist eine Vektorrepräsentation der Zielsequenz. Dieser Vektor wird an eine „Linear Layer“ übergeben, die die hochdimensionale Ausgabe des Decoders in ein konkretes und interpretierbares Ausgabeformat umwandelt. Dabei wird eine „Softmax-Funktion“ verwendet, um Wahrscheinlichkeiten für die verschiedenen Ausgabeelemente zu berechnen (Vaswani et al., 2017). Dies ist entscheidend für die Auswahl des nächsten Ausgabeelements in einer autoregressiven Generierung. Die Pipeline durchläuft diesen Schritt so lange, bis alle Elemente der Inputsequenz in die Outputsequenz überführt wurden.

2.2.3 Anwendung und Mehrwert von ASR

Die menschliche Sprache, als eine der natürlichsten Formen der zwischenmenschlichen Kommunikation, ermöglicht es Menschen, Ideen, Emotionen und Informationen effektiv auszutauschen (Schneider-Stickler und Bigenzahn, 2013). In dieser Hinsicht eröffnet ASR sinnvolle Möglichkeiten und fungiert als Schnittstelle zwischen Mensch und Maschine. Statt Tasten zu drücken oder Text einzugeben, können Benutzer einfach sprechen, um Anfragen zu stellen oder Aufgaben auszuführen (Baig et al., 2012). Dies trägt nicht nur zur Zeitersparnis bei, sondern erhöht auch die Benutzerfreundlichkeit erheblich. Sprachassistenzsysteme wie Siri[12] und Alexa[13], Smart TVs oder sprachgesteuerte Navigationssysteme in Autos sind Beispiele für die Integration von ASR in den Alltag, insbesondere im Zusammenhang mit der Nutzung von Smartphones. Ein weiterer bedeutender Anwendungsbereich von ASR liegt in der Transkription von gesprochenen Inhalten und der automatischen Erstellung von Untertiteln. Plattformen wie Spotify und YouTube nutzen ASR-Technologien, um Transkripte für Podcasts bereit zu stellen (Mauran, 2023), Audio-Dateien in andere Sprachen zu übersetzen (Spotify, 2023) oder um Videounterstützung in Form von automatischer Untertitelung zu erzeugen (Datta et al., 2020). Dies trägt zur Barrierefreiheit bei, insbesondere für Menschen mit Sehbehinderungen oder anderen Einschränkungen. Die Umwandlung von gesprochenem Inhalt in Text oder synthetische Sprache ermöglicht es Personen mit unterschiedlichen Fähigkeiten, den Inhalt zu verstehen und zu genießen, und hilft, digitale Barrieren abzubauen (Schöne, 2021b). In der Industrie, insbesondere im Kundendienst und in Call Centern, spielt ASR eine entscheidende Rolle bei der Verbesserung der Servicequalität und -effizienz. Die automatische Transkription von Telefongesprächen ermöglicht es Unternehmen, wertvolle Daten aus Kundeninteraktionen zu gewinnen, die für Schulungszwecke, Qualitätskontrolle und Echtzeitanalysen genutzt werden können. Darüber hinaus dient ASR als Grundlage für die Implementierung von Bots, die Kundenanfragen automatisch weiterleiten und beantworten (Zweig et al., 2006).

Neben den unmittelbaren Anwendungsfeldern von ASR bildet die automatische Transkription von Audio die Basis für Weiterverarbeitungsmöglichkeiten. Die Transkripte können als Grundlage für weitere Technologien, insbesondere NLP, dienen. Dies könnte ein entscheidender Schritt zur Bewältigung der im vorherigen Abschnitt beschriebenen Herausforderungen von Wissenschaftspodcasts sein. So können, ergänzend zu den Metadaten, weitere Attribute eines Podcasts zur Erhöhung der Reichweite genutzt werden. Dazu gehören beispielsweise der Titel und der Inhalt des Podcasts (Jones et al., 2021). Regulär verwendete Metadaten von Podcasts

[12] https://www.apple.com/de/siri/
[13] https://developer.amazon.com/de-DE/alexa

enthalten oft nur begrenzte Informationen über den Inhalt, die jedoch entscheidend für das Verständnis und die Auffindbarkeit sind. Eine Studie hat gezeigt, dass Retrieval-Anwendungen, die auf den automatisch erzeugten Transkripten des Podcasts basieren, besser abschneiden bei Suchanfragen bezüglich erwähnter Personen, Zitate oder behandelter Themen, während metadatenbasierte Indexierung bessere Ergebnisse für Suchanfragen nach dem Titel des Podcasts oder nach dem Autor erzielte. Interessanterweise war dies der Fall, obwohl die Fehlerquote vergleichsweise hoch ausfiel (Besser et al., 2010). Ähnliche Ergebnisse wurden auch in einer zweiten Studie festgehalten, welche signifikante Verbesserungen durch die Integration von Transkripten in den Metadaten feststellen konnten, trotz Fehlerraten von bis zu 50% (Chelba et al., 2008). Zudem könnte die Integration von Podcasts in vorhandene Workflows und Forschungsdatenbanken die Referenzierung und Zitationen erheblich erleichtern. Insbesondere durch weitere NLP-Methoden wie etwa Topic-Modelling (Vartakavi et al., 2021), bei der dem Text Kategorien zugeordnet werden, welche eine thematische Klassifizierung erlauben. Zusammenfassungen einer Podcast-Episode können durch Text-Summarization (Fetic et al., 2021) erstellt werden und die Zugänglichkeit erhöhen und Entity Recognition (Gundogdu et al., 2018) könnte verwendet werden um im Podcast erwähnte Personen oder wissenschaftliche Artikel zu extrahieren und entsprechend zu referenzieren.

Zusammenfassend ist ASR vielseitig einsetzbar. Neben den unmittelbaren Effekten können die automatisch erstellten Transkripte als solide Grundlage für weitere Textverarbeitungsoperationen dienen, was eine vielschichtige und leistungsstarke Nutzung von ASR in verschiedenen Kontexten ermöglicht.

2.2.4 Herausforderungen und Entwicklungen in der ASR-Technologie

Die automatische Spracherkennung hat zweifelsohne erhebliche Fortschritte gemacht, dennoch sind noch zahlreiche Herausforderungen zu bewältigen, um die Genauigkeit der Transkription weiter zu verbessern. Diese Herausforderungen sind vielfältig:

Störgeräusche: Hintergrundgeräusche, akustische Umgebungsgeräusche, Nachhall und Störungen im Übertragungskanal sind ständige Begleiter in der Welt der gesprochenen Sprache. Sie können die Qualität der Aufnahmen erheblich beeinträchtigen und ASR-Systeme vor die Herausforderung stellen, relevante Sprachsignale von störendem Rauschen zu unterscheiden. Überlappende Gespräche stellen eine zusätzliche Komplexität dar, da mehrere Sprecher gleichzeitig sprechen können (Bhable et al., 2023).

Darstellung: Die präzise Darstellung der transkribierten Texte ist entscheidend. Während einzelne Orthographiefehler für die technische Verarbeitung von Transkripten kein Problem darstellen (Chelba et al., 2008), leidet jedoch die Lesbarkeit beträchtlich darunter (Putri, 2019). Hierzu gehören die richtige Setzung von Satzzeichen, die Berücksichtigung von Groß- und Kleinschreibung, ordnungsgemäßes Formatieren und die korrekte Abfolge der Worte. Fehler in der Formatierung können die Verständlichkeit der Transkripte beeinträchtigen und die Interpretation erschweren (Putri, 2019). Nicht zuletzt wirkt ein gut formattiertes Transkript wesentlich professioneller. Aktuell bedarf es, in Anbetracht der Fehlerquoten (Radford et al., 2022), jedoch noch Nacharbeit.

Sprachvariation: Die Vielfalt in der natürlichen Sprache ist faszinierend und herausfordernd zugleich. Sie zeigt sich in den unterschiedlichen Akzenten, die Menschen beim Sprechen verwenden, den individuellen Variationen in der Aussprache von Wörtern und Sätzen sowie der Tendenz zu spontaner und informeller Sprache im Alltag. Hinzu kommen regionale Dialekte, die oft eine eigene Grammatik und Wortschatz aufweisen, sowie gelegentliche Sprachfehler, sei es in Form von Versprechern, Wiederholungen oder unklarer Artikulation. Für ASR-Systeme sind all diese Facetten der Sprachvielfalt eine fortwährende Herausforderung, die es zu Lösen gilt (Singh et al., 2023).

Unterscheidung von Sprechern: Die Fähigkeit zur Unterscheidung von Sprechern, auch Diarization genannt, spielt eine entscheidende Rolle bei der Erstellung präziser Transkriptionen (Bhable et al., 2023). Fehlt diese Unterteilung, kann es zu Verwirrungen kommen, insbesondere wenn mehrere Personen gleichzeitig sprechen oder es darum geht, herauszufinden, welcher Sprecher welche Teile des Gesprächs beigetragen hat. Unterbrechen sich die Sprecher dabei noch gegenseitig, wird die Aufgabe für ASR-Systeme noch komplexer. Zudem ändert sich die Beteiligung der Sprecher im Verlauf des Gesprächs, was die Identifizierung zusätzlich erschwert. Die Berücksichtigung dieser dynamischen Faktoren ist entscheidend für die Leistungsfähigkeit des ASR-Systems. In akademischen oder journalistischen Texten werden oft Zitate oder Referenzen auf bestimmte Sprecher oder Experten verwendet (Austria, 2007). Ohne eine klare Identifikation von Sprechern können Zitate und Referenzen nicht korrekt zugeordnet werden. Der Mangel einer eindeutigen Identifikation kann auch weiterführende Recherchen zu den Sprechern behindern.

Mehrsprachige Texte und Code-Switching: In einer globalisierten Welt, in der Menschen aus verschiedenen kulturellen Hintergründen miteinander interagieren, sind multilinguale Gespräche und das Wechseln zwischen verschiedenen Sprachen alltäglich geworden. Da Wissenschaftskommunikation größtenteils in englischer Sprache stattfindet, ist es nicht unüblich in im Laufe eines Diskurses Fachterminologie oder passendere Wörter auf Englisch einzustreuen. Der mühelose Wechsel zwischen zwei Sprachen innerhalb eines Gesprächs wird allgemein als Code-Switching bezeichnet (Myers-Scotton, 2017). Dieses Phänomen stellt eine zusätzliche Herausforderung für ASR-Systeme dar (Xiong, 2023). ASR-Systeme müssen nicht nur in der Lage sein, verschiedene Sprachen zu erkennen, sondern auch die nahtlose Umstellung zwischen ihnen bewältigen.

Out-of-Vocabulary (OOV) und Long-Tail-Effekt: ASR-Systeme stoßen gelegentlich auf Wörter, die nicht in ihren Trainingsdaten enthalten sind, wie beispielsweise seltene Begriffe oder Fachterminologie. Die Erkennung und Behandlung von OOV-Wörtern erfordert eine gewisse Form der Generalisierung, bei der das System auf der Grundlage seiner vorhandenen Kenntnisse versucht, unbekannte Wörter zu interpretieren. Dies kann zu Fehlern führen, wenn die Annahmen des Systems nicht mit dem tatsächlichen Kontext übereinstimmen (Xiong, 2023). Der Long-Tail-Effekt beschreibt die Verteilung von Wörtern in einem Textkorpus, bei der einige Wörter sehr häufig auftreten (der „Kopf“ der Verteilung), während viele andere Wörter selten vorkommen (der „Schwanz“ der Verteilung) (Anderson, 2012; Zhan et al., 2021). Dies

kann bereits beim Training des Modells zu Verzerrungen führen und schlussendlich auch in der Transkription, da seltene, aber wichtige Wörter möglicherweise nicht korrekt erkannt werden (Xiong, 2023).

3 Beschreibung der Datensätze

3.1 Mozilla Common Voice

Das Projekt „Common Voice“[1], erstmalig im Jahr 2017 von Mozilla[2] initiiert, hat sich zum Ziel gesetzt, eine umfangreiche und facettenreiche Datenbank gesprochener Sprache zu etablieren (Ardila et al., 2020). Ein bemerkenswertes Merkmal dieses Projekts ist die partizipative Beteiligung der weltweiten Gemeinschaft. Individuen aus unterschiedlichen geografischen Regionen, kulturellen Kontexten und Sprachgruppen sind eingeladen, kurze Sätze und Texte in ihren jeweiligen Muttersprachen aufzuzeichnen (Mozilla, n. d.). Das Resultat ist eine höchst diverse und repräsentative Sammlung von Sprachdaten, die ein breites Spektrum von Themen und Kontexten abdeckt. Eine entscheidende Komponente dieses Projekts ist die umfassende Qualitätskontrolle der gesammelten Daten (Mozilla, n. d.), zur Gewährleistung einer verlässlichen und fehlerfreien Grundlage.

Der aktuelle Datensatz, „Common Voice Corpus 15.0“[3], wurde am 14. September 2023 veröffentlicht. Er enthält 28.751 Stunden (wovon 19.160 Stunden validiert wurden) in 114 verschiedenen Sprachen. In dieser vorliegenden Arbeit liegt der Fokus jedoch auf den Sprachen Deutsch und Englisch, weshalb andere Sprachen vorerst außer Acht gelassen werden. Der deutschsprachige Datensatz erstreckt sich über eine Größe von 32,49 GB und enthält insgesamt 18.352 Sprachaufnahmen mit einer Gesamtdauer von 1.388 Stunden (wovon 1.301 Stunden validiert wurden). Der englischsprachige Datensatz umfasst 3.347 Stunden (wovon 2.532 Stunden validiert wurden) mit einer Größe von 79,09 GB. Insgesamt sind 88.904 Stimmen in diesen Datensätzen vertreten. Die Datensätze setzen sich aus MP3-Audiodateien zusammen und werden von mehreren Tab-separated Values (TSV)-Dateien begleitet, welche detaillierte Informationen zu den einzelnen Audio-Clips enthalten. Darunter die Dauer, die schriftliche Transkription des gesprochenen Inhalts, Angaben zum Alter, Geschlecht, Herkunft und Akzent der Sprecher sowie Informationen über die Bewertungen (Up- und Down-Votes), die von Rezensenten für die jeweiligen Sprachdateien vergeben wurden. Da das verwendete ASR-System auf dem Common Voice Corpus 5 und 9 trainiert wurde und somit mit diesen Daten bereits vertraut ist, kommen für diese Arbeit Teilmengen aus den Datensegmenten 10 bis 15 für deutsch- und englischsprachige Sätze in Frage, um die tatsächliche Leistungsfähigkeit und Generalisierungsfähigkeiten des Modells in unbekannten Szenarien zu beurteilen. Dies entspricht einer Gesamtdauer von 202 Stunden (wovon 190 Stunden validiert sind) für Deutsch und 347 Stunden (wovon 265 Stunden validiert sind) für Englisch. Tabelle 3.1 zeigt die genaue Aufteilung. Aus Kapazitätsgründen wurde sich jedoch ausschließlich auf die validierten Daten des Delta Segments 13 beschränkt.

Der Common Voice-Datensatz eignet sich ideal als Bewertungsgrundlage für ASR-Modelle. Er zeichnet sich durch vielfältige Sprecher und Akzente aus, repräsentiert natürliche Sprechweise, und unterliegt strengen Qualitätskontrollen. Zudem bietet er präzise schriftliche Transkripte

[1] https://commonvoice.mozilla.org
[2] https://www.mozilla.org
[3] https://commonvoice.mozilla.org/de/datasets

Tabelle 3.1

Common Voice Delta Segmente 10 bis 15, deutsch und englisch

Segment	Erfasste Stunden	Bestätigte Stunden	Anzahl Stimmen
DE-10	46	44	299
EN-10	97	51	2705
DE-12	50	48	272
EN-12	63	64	1152
DE-13	57	53	369
EN-13	48	46	1117
DE-14	37	34	320
EN-14	71	56	1212
DE-15	12	11	165
EN-15	68	48	750

und Metadaten, die Einblicke in die Leistung in verschiedenen Sprechergruppen ermöglichen. Es ist jedoch zu beachten, dass der Common Voice-Datensatz hauptsächlich Alltagssätze enthält und möglicherweise nicht ausreichend wissenschaftliche oder spezifische Fachthemen abdeckt. Dies könnte seine Eignung für die Evaluierung von ASR-Modellen in Bezug auf wissenschaftliche Podcasts einschränken. Darüber hinaus handelt es sich bei den Aufnahmen im Datensatz um kurze skriptbasierte Sätze, was dazu führt, dass das Auftreten von Füllwörtern, die in natürlichen Gesprächssituationen häufig vorkommen, minimiert ist. Dies sollte bei der Bewertung der ASR-Modellleistung berücksichtigt werden.

3.2 GigaSpeech

„GigaSpeech“ ist ein umfangreiches Spracherkennungskorpus, das zur Entwicklung und zum Training automatischer Spracherkennungssysteme dient (Chen et al., 2021). Die Sammlung umfasst insgesamt 40.000 Stunden englischsprachiger Audioaufnahmen, wovon 10.000 Stunden sorgfältig transkribiert und gekennzeichnet sind. Die Audiodaten stammen aus verschiedenen Quellen wie Audiobüchern, Podcasts und YouTube-Videos und behandeln eine breite Palette von Themen aus insgesamt 28 Kategorien, darunter Kunst, Wissenschaft, Sport und vieles mehr (Chen et al., 2021). Das Korpus bietet Trainingsuntergruppen in verschiedenen Größen, von 10 Stunden bis zu 10.000 Stunden. Es wurde sorgfältig darauf geachtet, dass der umfangreichste Datensatz, der als „XL“ bezeichnet wird, eine WER von nicht mehr als 4% aufweist. Für alle kleineren Trainingsdatensätze, wurde eine strengere Maßnahme angewendet. Hier wurde eine maximale WER von 0% festgelegt, was bedeutet, dass in diesen Datensätzen keinerlei akzeptierte Fehler in der automatischen Transkription toleriert wurden (Chen et al., 2021). Zusätzlich gibt es spezielle Evaluationssets (DEV und TEST), die von erfahrenen menschlichen Transkribenten überarbeitet wurden, um eine hohe Qualität der Transkriptionen zu gewährleisten (Chen et al., 2021). Tabelle 3.2 zeigt die genaue Verteilung der einzelnen Untergruppen. Die Metadaten werden in einer zusammenhängenden JavaScript Object Notation (JSON)-Datei zur Verfügung gestellt. Zu den Datenfeldern gehören unter anderem die Kategorie, der Text, das Audio in Bytes, die Quelle, Start- und Endzeit, der Titel und der Link zum Original. Zugang zum GigaSpeech

Tabelle 3.2

Gigaspeech subsets

Subset	Audiobook	Podcast	Youtube	Total
XL	2.655h	3.499h	3.846h	10.000h
L	650h	875h	975h	2.500h
M	260h	350h	390h	1.000h
S	65h	87,5h	97,5h	250h
XS	2,6h	3,5h	3,9h	10h
DEV	-	-	-	12h
TEST	-	-	-	40h

Anmerkung. DEV und TEST Sets bestehen aus zufällig ausgewählten Passagen der Gesamtmenge, wobei TEST zum Teil manuell zusammengestellt wurde um eine höhere Abdeckung zu erzielen.

Datensatz kann über GitHub[4] oder Huggingface[5] beantragt werden. In dieser Arbeit wurden Auszüge aus dem „M“ Datensatz gewählt. Insgesamt über 27.700 Proben mit einer Gesamtdauer von über 30 Stunden. Diese wurden nicht über Huggingface direkt importiert sondern aus einzelnen parquet-Dateien[6] zusammen gestellt, welche ebenfalls auf Huggingface vorfindbar sind. Der GigaSpeech-Datensatz stellt eine wertvolle Ergänzung zum Common Voice Datensatz dar, da er nicht nur auf ein breites Spektrum an Sprechern, Dialekten und Themen abbildet sondern auf auch aus Auszügen von realen Beispielen besteht, darunter Podcasts. Die Metadaten des GigaSpeech-Datensatzes ermöglichen es, eine Auswahl an wissenschaftlichen Podcasts zu identifizieren und zu testen. Im Gegensatz dazu erfordert der Common Voice Datensatz zusätzliche Textklassifizierung, um eine thematische Zuordnung vorzunehmen. Es ist jedoch wichtig zu beachten, dass der GigaSpeech-Datensatz ausschließlich englische Sprache enthält und somit keine Rückschlüsse auf wissenschaftliche Podcasts in deutscher Sprache zulässt.

3.3 Open Science Radio

Das „Open Science Radio“[7] ist ein intermittierend veröffentlichtes Podcast-Projekt, das 2013 von Matthias Fromm ins Leben gerufen wurde und sich inhaltlich mit dem Themenfeld „Open Science“ befasst. In diesem Zusammenhang werden diverse Subthemen abgedeckt, darunter „Open Access “, „Citizen Science “, „Öffentliche Wissenschaft“, und „Open Education“. Die primäre Zielsetzung des Projekts besteht darin, als Informationsquelle und Bildungsressource für Personen zu dienen, die ein Interesse an Open Science hegen oder ihr Verständnis zu diesem Thema vertiefen möchten (Fromm, 2013). Darüber hinaus fungiert das Open Science Radio als Plattform für aktuelle Berichterstattung über Entwicklungen und Neuigkeiten im Bereich von Open Science (Fromm, 2013). Hierbei werden Hörerinnen und Hörer über die jüngsten Trends, laufende Projekte und Initiativen innerhalb der Open-Science-Gemeinschaft auf dem Laufenden gehalten. Infolgedessen bietet das Projekt eine zeitgemäße und relevante Informati-

[4] https://github.com/SpeechColab/GigaSpeech

[5] https://huggingface.co/datasets/speechcolab/gigaspeech

[6] Apache Parquet ist ein spaltenorientiertes, open source Dateiformat für die Speicherung von Daten. Es wurde entwickelt, um große Mengen strukturierter Daten effizient und komprimiert zu speichern („Apache Parquet“, n. d.).

[7] https://www.opensciencerадio.org/

onsquelle für diejenigen, die an der Open-Science-Bewegung interessiert sind (Fromm, 2013; „Podcast 'Open Science Radio'", n. d.). Die Podcast-Folgen werden in teils deutscher und teils englischer Sprache gehalten.

Das „Open Science Radio" wurde als aus mehreren Gründen als Fallbeispiel für diese Arbeit ausgewählt. Zunächst bietet die Bilingualität des Podcasts die Möglichkeit, das ASR-Modell auf zwei verschiedenen Sprachen zu testen – Deutsch und Englisch. Dies erweitert den Anwendungsbereich des ASR-Modells und erlaubt die Evaluierung der Leistungsfähigkeit in einer multilingualen Umgebung. Darüber hinaus stellt der Podcast eine äußerst wertvolle Ressource für die empirische Forschung dar, da er eine Fülle von realen Audioinhalten bereitstellt. Dies ermöglicht es, die ASR-Modelle unter realen Bedingungen zu testen und die Genauigkeit ihrer Transkriptionen anhand tatsächlicher Podcast-Episoden zu bewerten. Des Weiteren bietet das Open Science Radio eine breite Vielfalt von Sprechern und Diskussionsteilnehmern. Diese Vielfalt ist von erheblichem Wert, da ASR-Modelle oft mit verschiedenen Sprecherakzenten und -stilen konfrontiert werden, insbesondere in Podcasts, in denen verschiedene Gastgeber und Interviewgäste auftreten. Diese Diversität ermöglicht eine umfassendere Evaluierung der ASR-Leistung und deren Fähigkeit, unterschiedliche Sprechweisen erfolgreich zu bewältigen. Schließlich besteht eine thematische Überschneidung zwischen dem Open Science Radio und dem Bachelor-Thema. Die Inhalte des Podcasts behandeln Themen im Zusammenhang mit Open Science, was einen klaren Zusammenhang mit der Forschungsarbeit herstellt. Im Gegensatz zu den vorherigen Datensätzen, ist im Fall der Podcasts des Open Science Radio keine Ground Truth verfügbar. Da es enorm zeitaufwendig wäre, den Text handschriftlich zu transkribieren und sorgfältig zu überprüfen, wird in dieser Arbeit auf die Erstellung einer Ground Truth verzichtet. Stattdessen werden systemgenerierte Transkripte der Podcasts erstellt und hinsichtlich ihrer ihrer Repräsentativität der Original-Datei betrachtet.

4 Methodik

4.1 Metriken

4.1.1 Word Error Rate

Die WER ist eine häufig verwendete Metrik in der ASR, um die Genauigkeit eines Transkriptionssystems zu bewerten. Sie dient dazu, Abweichungen zwischen einer Referenztranskription und einer systemgenerierten Transkription in Bezug auf die Anzahl der Fehler zu quantifizieren (Ali und Renals, 2018). Die WER wird beispielsweise zum Benchmarking auf dem LibriSpeech Datensatz (Panayotov et al., 2015) verwendet („LibriSpeech Test-Clean Benchmark (Speech Recognition)“, n. d.), welcher weithin bekannt in der modernen Spracherkennungsforschung ist (Radford et al., 2022).

Die WER betrachtet typischerweise folgende Parameter:

- **Referenztranskription**: Eine „Referenztranskription“ ist eine schriftliche Darstellung von gesprochenem oder audiovisuellem Inhalt, die als zuverlässiger Bezugspunkt oder „Ground Truth“ für die Evaluierung von Spracherkennungs- und Transkriptionssystemen dient. Sie stellt den Standard dar, an dem die Genauigkeit und Qualität automatischer Transkriptionssysteme gemessen werden. (*„the cat sleeps on the carpet.“*)
- **Substitutionen**: Das sind Wörter in der Referenztranskription, die versehentlich durch andere Wörter in der systemgenerierten Transkription ersetzt wurden (*„the cat sleeps on the cupboard.“*).
- **Einfügungen**: Dies sind zusätzliche Wörter in der systemgenerierten Transkription, die in der Referenztranskription nicht vorhanden sind (*„the black cat sleeps on the carpet.“*).
- **Löschungen**: Hierbei handelt es sich um Wörter in der Referenztranskription, die in der systemgenerierten Transkription fehlen (*„cat sleeps on the carpet.“*).

Die WER wird mit der Formel 4.1 berechnet (Ali und Renals, 2018):

$$\text{WER} = \frac{S + I + D}{N} \cdot 100 \tag{4.1}$$

Wobei:
S = Anzahl der Substitutionen
I = Anzahl der Einfügungen
D = Anzahl der Löschungen
N = Gesamtzahl der Wörter in der Referenztranskription

Das Ergebnis wird üblicherweise in Prozent angegeben, wobei niedrigere WER-Werte eine höhere Genauigkeit anzeigen (Ali und Renals, 2018). Zum Beispiel bedeutet eine WER von 5%, dass durchschnittlich 5% der Wörter in der systemgenerierten Transkription Fehler enthalten

oder nicht mit der Referenztranskription übereinstimmen. Aus der oberen Referenztranskription *„the cat sleeps on the carpet“* und der systemgenerierten Transkription *„cat sleeps on the cupboard“* ergibt sich eine WER von ungefähr 33%:

$$\text{WER} = \frac{1\ (\text{Substitution}) + 0\ (\text{Einfügungen}) + 1\ (\text{Löschungen})}{6\ (\text{Gesamtzahl der Wörter in der Referenztranskription})} \times 100 \approx 33\%$$

Die WER ist eine wichtige Maßzahl, um die generelle Leistung von ASR-Systemen objektiv zu bewerten. Jedoch bringt die WER auch einige Probleme mit sich, die es bei der Evaluierung zu beachten gilt (Wang et al., 2003). Das Hauptproblem liegt darin, dass die WER alle Fehler gleich gewichtet. Wenn beispielsweise „ihre“ und „Ihre“ (in Groß- und Kleinschreibung) vertauscht werden, kann die WER, trotz der semantischen Ähnlichkeit, dies als Substitution erfassen. Ebenso werden Füllwörter wie „äh, uh, oh“ als Einfügungen erkannt, obwohl diese keinen großen Einfluss auf die Bedeutung des Inhalts haben und im natürlichen Sprachfluß häufig angewandt werden (O'Connor, 2023). Um dem entgegen zu wirken ist es notwendig das Referenztranskript und das systemgenerierte Transkript auf ein Grundlevel zu bringen. Dies geschieht mithilfe eines „Normalisierers“ (O'Connor, 2023). Ein Normalisierer (engl. Normalizer) bezieht sich in diesem Kontext auf einen Algorithmus, der verwendet wird, um die Transkripte in eine standardisierte oder einheitliche Form zu bringen und den Vergleich zwischen ihnen zu erleichtern. Zu den Aufgaben eines Normalizers gehören die Textbereinigung, bei der unnötige Zeichen oder Formatierungen entfernt werden, die Kapitalisierung, um sicherzustellen, dass Groß- und Kleinschreibung konsistent sind, und die Entfernung von Zeitstempeln oder anderen nicht relevanten Informationen. Darüber hinaus kann er Füllwörter oder Wiederholungen entfernen und Abkürzungen in ihre Vollformen umwandeln, um die Konsistenz und den Fokus auf den eigentlichen Inhalt des Transkripts zu verbessern. Selbst nach einer Normalisierung der Transkripte kann die WER immer noch ein falsches Bild produzieren, da sie schlecht Ähnlichkeiten abbilden kann, was insbesondere bei Substitutionen problematisch wird (O'Connor, 2023). Dies wird bei der Betrachtung von Tabelle 4.1 deutlich: Es ist bemerkenswert festzustellen, dass sowohl System A

Tabelle 4.1

Beispiel Transkriptionen Substitutionen WER

Art	Transkript	WER
Referenz	Johns favorite restaurant is 'La Trattoria'.	
System A	Jons favorite restaurant is 'La Tratoria'.	33,3%
System B	Garfields favorite restaurant is 'La Tratoria'.	33,3%

als auch System B in Bezug auf die WER äquivalente Werte aufweisen, obwohl System A deutlich näher an der Referenztranskription liegt als System B. Womöglich würde System A im Vergleich zur Audiodatei ohne Berücksichtigung der Referenztranskription als korrekt klassifiziert werden. Dieses Phänomen unterstreicht die Limitationen der WER als isolierte Metrik zur Bewertung der Transkriptionsqualität und verdeutlicht die Notwendigkeit einer tiefergehenden Analyse, um subtile Unterschiede in der Transkriptionsgenauigkeit zu erfassen. Daher sollte sie immer im Kontext anderer Metriken betrachtet werden, um eine möglichts genaue Bewertung der Leistung eines Transkriptionssystems zu ermöglichen.

4.1.2 Jaro-Winkler-Ähnlichkeitsmaß

Ergänzend zur WER kann das Jaro-Winkler-Ähnlichkeitsmaß (engl.: Jaro-Winkler-Similarity) eingesetzt werden. Das Jaro-Winkler-Ähnlichkeitsmaß ist eine Maßzahl zur Bewertung der Ähnlichkeit zwischen zwei Zeichenketten oder Wörtern. Es wurde entwickelt, um den Grad der Übereinstimmung zwischen diesen Zeichenketten zu messen, insbesondere wenn es um das Erkennen von Tippfehlern oder geringfügigen Abweichungen geht. Grundlage bildet das Jaro-Ähnlichkeitsmaß (engl. Jaro-Similarity) von Matthew A. Jaro (Jaro, 1989), welches von von William E. Winkler im Jahre 1990 zum Jaro-Winkler-Ähnlichkeitsmaß modifiziert wurde (Winkler, 1990), sodass ein Bonus für gemeinsame Präfixe in den Zeichenketten gegeben wird. Bei der Berechnung dieser Metrik werden folgende Faktoren berücksichtigt:

- Die Anzahl der Zeichen, die in beiden Zeichenketten übereinstimmen.
- Die Anzahl der Transpositionen zwischen den Zeichenketten.
- Die Längen der Zeichenketten, um zu prüfen, ob sie eine ähnliche Länge aufweisen.

Für die zu vergleichenden Zeichenketten $|s_1|$ und $|s_2|$ wird das Jaro-Ähnlichkeitsmaß wird mit der Formel 4.2 berechnet (Cohen et al., 2003):

$$\text{sim}_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases} \tag{4.2}$$

Wobei:
sim_j = Jaro-Ähnlichkeitsmaß
m = Anzahl der übereinstimmenden Zeichen
$|s_i|$ = Längen der Zeichenketten $|s_i|$
t = Anzahl der Transpositionen

Die Übereinstimmung von Zeichen in den Zeichenketten s_1 und s_2 wird unter der Bedingung definiert, dass diese Zeichen identisch sind und ihren Abstand zueinander nicht weiter als $\left\lfloor \frac{\max(|s_1|,|s_2|)}{2} \right\rfloor - 1$ Zeichen überschreiten dürfen (Cohen et al., 2003).

Das Jaro-Winkler-Ähnlichkeitsmaß erweitert die Formel 4.2 und wird mit folgender Formel 4.3 berechnet (Cohen et al., 2003):

$$\text{sim}_w = \text{sim}_j + p \cdot l \cdot (1 - \text{sim}_j) \tag{4.3}$$

Wobei:
sim_w = Jaro-Winkler-Similarity
p = Konstante, normalerweise 0.1
l = Länge des gemeinsamen Präfixes der Zeichenketten

Das Jaro-Winkler-Ähnlichkeitsmaß liefert einen Wert, der zwischen 0 und 1 liegt. Ein Wert von 0 zeigt keine Übereinstimmung an, während ein Wert von 1 eine perfekte Übereinstimmung bedeutet. Je höher der Wert, desto ähnlicher sind die Zeichenketten (Cohen et al., 2003). Diese Metrik findet Anwendung in verschiedenen Bereichen wie Datenbereinigung, Zeichenkettenvergleichen, Rechtschreibprüfung und bei der Bewertung der Ähnlichkeit von Namen oder Wörtern,

insbesondere wenn es darum geht, kleine Abweichungen zu erkennen. Daher ist sie bestens geeignet um die WER zu ergänzen. In Tabelle 4.2 wurde das Jaro-Winkler-Ähnlichkeitsmaß für die beiden Beispieltranskripte (System A und System B) aus Tabelle 4.1 berechnet. In diesem Fall erzielt System B, obwohl immer noch mit einer recht hohen Ähnlichkeitszahl, eine deutlich schlechtere Bewertung im Vergleich zu System A.

Tabelle 4.2

Beispiel Transkriptionen Substitutionen Jaro-Winkler-Similarity

Art	Transkript	WER	sim_w
Referenz	Johns favorite restaurant is 'La Trattoria'.		
System A	Jons favorite restaurant is 'La Tratoria'.	33,3%	98,8%
System B	Garfields favorite restaurant is 'La Tratoria'.	33,3%	90,5%

4.2 Verwendete Tools und Bibliotheken

In diesem Abschnitt werden die wichtigsten Tools und Bibliotheken vorgestellt, die zur Erzeugung der Ergebnisse benötigt wurden. Der Code, sowie die Daten können auf GitHub[1] eingesehen werden.

Python Python[2] ist eine weit verbreitete Programmiersprache und eine sogenannte General-Purpose-Sprache. General-Purpose-Sprachen sind vielseitig einsetzbar und flexibel, was bedeutet, dass sie in einer breiten Palette von Anwendungen verwendet werden können, von Webentwicklung über Datenanalyse und wissenschaftliche Berechnungen bis hin zu Automatisierung und maschinellem Lernen. Python zeichnet sich durch seine Benutzerfreundlichkeit, eine große und aktive Entwickler-Community sowie eine umfangreiche Bibliothek für Datenanalyse und wissenschaftliche Berechnungen aus (Balreira et al., 2023) und wurde daher als Programmiersprache für diese Arbeit ausgewählt.

jiwer Jiwer[3] ist eine Python-Bibliothek, die unter anderem für die Berechnung des WER in der automatischen Spracherkennung entwickelt wurde. Sie bietet eine benutzerfreundliche Schnittstelle, um die WER zwischen Transkriptionen und Ground Truth zu berechnen und ist ein wertvolles Werkzeug für die Evaluation von Spracherkennungssystemen.

Matplotlib Matplotlib[4] ist eine Python-Bibliothek zur Erstellung von Grafiken und Diagrammen. Sie bietet umfangreiche Funktionen zur Visualisierung von Daten in verschiedenen Formaten und wird in verschiedenen Anwendungsbereichen, einschließlich Wissenschaft und Datenanalyse, eingesetzt.

NLTK Natural Language Toolkit (NLTK)[5] ist eine Python-Bibliothek für die Verarbeitung und Analyse natürlicher Sprache. NLTK bietet eine breite Palette von Tools und Ressourcen für die

[1] https://github.com/DrBilboArriba/ASR_and_science_podcasts/releases/latest
[2] https://www.python.org/
[3] https://pypi.org/project/jiwer/
[4] https://matplotlib.org/
[5] https://www.nltk.org/

Verarbeitung von Textdaten, einschließlich Tokenisierung, Stemming, Lemmatisierung und Part of Speech (POS)-Tagging. NLTK erweist sich als äußerst nützlich bei Aufgaben wie Textklassifikation, Sentimentanalyse, Named Entity Recognition und ähnlichen Anwendungen. Darüber hinaus enthält es umfangreiche Korpora und Ressourcen, wie beispielsweise das semantische Lexikon WordNet[6], die Forschern und Entwicklern bei der Arbeit mit natürlicher Sprache helfen.

num2words num2words[7] ermöglicht die Konvertierung numerischer Werte in deren textuelle Darstellung Sie bietet Flexibilität bei der Anpassung von Formatierungsoptionen und unterstützt verschiedene Sprachen, was sie zu einem flexiblen Werkzeug macht.

Pandas Pandas[8] ist eine leistungsstarke und weit verbreitete Open-Source-Bibliothek für die Datenanalyse in der Programmiersprache Python. Sie bietet eine breite Palette von Funktionen zur effizienten Verarbeitung und Manipulation von Daten in tabellarischer Form. Pandas erleichtert den Import und Export von Daten aus verschiedenen Dateiformaten wie CSV, Excel, SQL-Datenbanken und mehr.

Seaborn Seaborn[9] ist eine Python-Datenvisualisierungsbibliothek, die auf Matplotlib aufbaut und die Erstellung ansprechender und informativer statistischer Grafiken erleichtert. Die Bibliothek bietet eine Vielzahl von vorgefertigten Diagrammtypen und eine benutzerfreundliche Schnittstelle zur Anpassung von Farbpaletten und Stilen, um visuell ansprechende Darstellungen von Daten zu generieren.

spaCy spaCy[10] ist eine Python-Bibliothek, die sich auf die effiziente Verarbeitung von Textdaten spezialisiert hat. Im Gegensatz zu NLTK wurde spaCy entwickelt, um die Geschwindigkeit und Ressourceneffizienz bei NLP-Anwendungen zu optimieren. Ein besonderes Merkmal von spaCy ist die Bereitstellung von vortrainierten Modellen für verschiedene Sprachen, was die nahtlose Integration in NLP-Projekte erleichtert und gleichzeitig die Entwicklungszeit erheblich verkürzt. In dieser Studie wird spaCy ausschließlich für das POS-Tagging verwendet, da NLTK keine vortrainierten Modelle für die POS-Erkennung in der deutschen Sprache bereitstellt. Darüber hinaus bietet spaCy jedoch auch eine Vielzahl an weiteren leistungsfähigen NLP-Funktionen, darunter Named Entity Recognition, syntaktische Analysen, und die Möglichkeit, benutzerdefinierte Entitäten zu trainieren.

wave Die Bibliothek wave[11] ist in der Standard-Library von Python mit enthalten und ist eine nützliche Ressource zur Verarbeitung von WAV-Audiodateien. Die Bibliothek ermöglicht die Manipulation von Audio-Signalen, Extraktion von Metadaten und Erstellung neuer WAV-Dateien. Sie ist besonders hilfreich für Projekte, die Audioverarbeitung und -analyse erfordern, wie beispielsweise in der Musik- und Spracherkennung oder in der Audiobearbeitung.

Whisper Zur Verwendung von Whisper muss zunächst das Modul importiert und das entsprechende Modell geladen werden. Es existieren verschiedene Whisper Modelle, welche aufgrund

[6] https://wordnet.princeton.edu/
[7] https://pypi.org/project/num2words/
[8] https://pandas.pydata.org/
[9] https://seaborn.pydata.org/
[10] urlhttps://spacy.io/
[11] https://python.readthedocs.io/en/v2.7.2/library/wave.html

unterschiedlich hoher Anzahl an Parametern in Geschwindigkeit und Leistung variieren. Tabelle 4.3 zeigt eine Übersicht über die Attribute der Modelle. Für den

Tabelle 4.3

Whisper Modelle im Vergleich

Model	Parameters	Required VRAM	Relative Speed
tiny	39 M	≈ 1	$\approx 32x$
base	74 M	≈ 1	$\approx 16x$
small	244 M	≈ 2	$\approx 6x$
medium	769 M	≈ 5	$\approx 2x$
large	1550 M	≈ 10	1x

4.3 Pipeline

Abbildung 4.1

Übersicht Pipeline

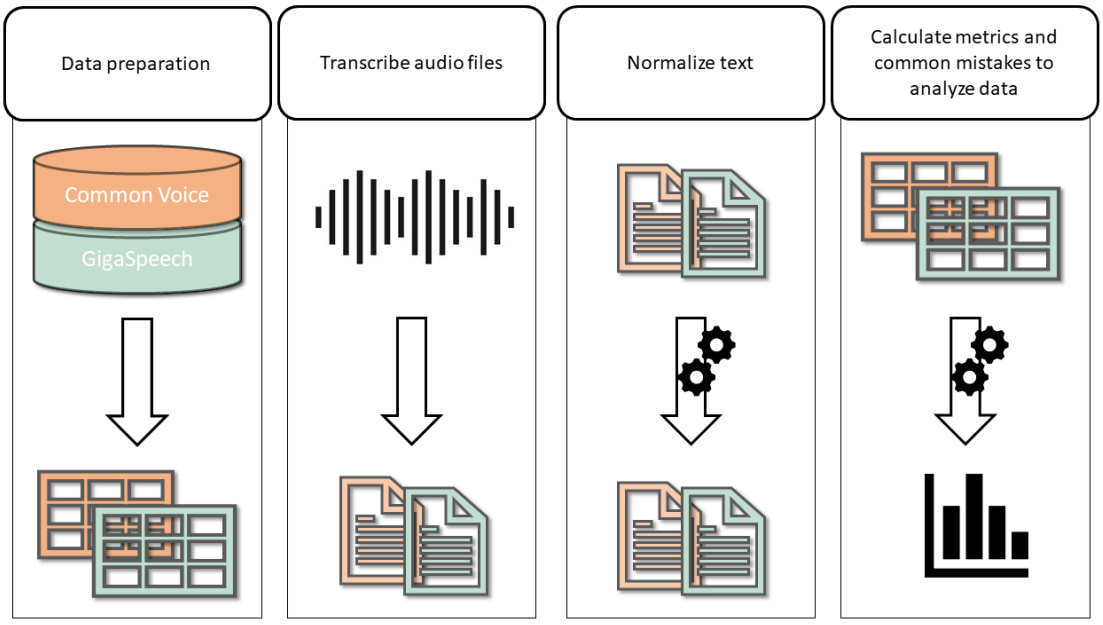


Abbildung 4.1 zeigt die grobe Struktur der Pipeline. Im folgenden Teil werden die einzelnen Stationen der Pipeline erläutert. Vor der Datenverarbeitung wurden die Datensätze sorgfältig in ein übersichtliches Ordnerverzeichnis strukturiert abgelegt.

1. **Daten Vorbereitung**:

   Zunächst werden die Datensätze in einen pandas Dataframe geladen. In diesem Prozess werden die Spalten *segment* und *variant* aus dem Datensatz entfernt, da sie keine relevanten Daten enthalten. Zudem wird die Anzahl der Label in Spalte *accents*, von ursprünglich

115 auf 29 reduziert, um die Handhabung der Daten zu erleichtern. Im GigaSpeech Datensatz werden die Spalten *speaker*, *source*, *original_full_path*, *url*, *audio_id* und *title* entfernt, da sie entweder redundant sind, leer bleiben oder keinen Mehrwert für die geplante Analyse bieten. Darüber hinaus wird die Dauer der einzelnen Audio-Dateien ermittelt, indem die Spalten *begin_time* und *end_time* miteinander verrechnet und dann ebenfalls entfernt werden. Zusätzlich werden sämtliche Transkripte im GigaSpeech Datensatz, die als „garbage-tags " (beispielsweise <OTHER> oder <MUSIC>) deklariert sind und keine natürliche Sprache enthalten, entfernt, da das verwendete Whisper-Modell nur Speech oder No Speech erkennt, jedoch keine Hintergrundgeräusche. Um die Daten weiter zu optimieren, werden die Werte in der Spalte *category*, die Nummern von 0 bis 28 enthält, den entsprechenden Kategorien zugeordnet. Anders als beim Common Voice Datensatz, liegen die Audiosegmente in tabellarischer Form vor, statt als separate Dateien.

2. **Transkription der Audio-Dateien**:

   Zur Erzeugung der Transkripte wurde das tiny-Whisper-Modell eingesetzt, das sich durch seine im Vergleich zu größeren Modellen deutlich schnellere Transkription auszeichnet. Dieser Unterschied ist von großer Bedeutung, insbesondere angesichts der beträchtlichen Menge von 67.535 Audio-Dateien, die für die Transkription verarbeitet werden mussten. Bei der Transkription wurde keine spezifische Zielsprache im Vorfeld festgelegt. Als Ergebnis dieses Prozesses generiert Whisper ein „Dictionary"[12], das verschiedene Parameter enthält, darunter den transkribierten Text, eine Aufschlüsselung in Abschnitte, was bei Audio-Dateien mit einer Länge von über 30 Sekunden relevant wird, und die erkannte Sprache. Für die vorliegende Auswertung wurde ausschließlich der transkribierte Text extrahiert und mit den entsprechenden Dataframes für die Gigaspeech- und Common Voice-Datensätze zusammengeführt.

3. **Textnormalisierung**:

   Bevor der Vergleich zwischen den Ground Truth-Texten und den erstellten Transkripten durchgeführt werden kann, ist die Normalisierung beider Texte ein essenzieller Schritt, um eine sinnvolle Vergleichbarkeit sicherzustellen. Die Normalisierung gewährleistet, dass gleiche Wörter und Phrasen, unabhängig von Unterschieden wie Groß- und Kleinschreibung, als identisch betrachtet werden. Ebenso wird durch die Normalisierung das Rauschen innerhalb der Textdaten minimiert, welches in Form von Sonderzeichen, Interpunktion und Stoppwörtern auftreten kann. Für die speziellen Anforderungen der Analyse wurde ein eigens entwickelter Normalizer eingesetzt. Zuerst wurden spezielle Interpunktions-Tags aus dem GigaSpeech-Datensatz entfernt, die ursprünglich für die Untersuchung von End-to-End-Endpointing und Interpunktionswiederherstellung vorgesehen waren. Dies ist entscheidend, da diese Tags potenziell die WER negativ beeinflussen könnten. Des Weiteren wurden Füllwörter wie 'Uhh' oder 'Eh' aus dem GigaSpeech-Datensatz entfernt, um die Texte weiter zu bereinigen. Anschließend erfolgte die Tokenisierung der zu vergleichenden Texte, wodurch die Zeichenketten in einzelne Wörter aufgeteilt wurden, um die weitere Verarbeitung zu erleichtern. Um das Rauschen im Text weiter zu reduzieren,

[12] Ein Dictionary ist eine Datenstruktur in Python. Es handelt sich um eine Sammlung von Schlüssel-Wert-Paaren, wobei jedem Schlüssel ein bestimmter Wert zugeordnet ist. Diese Datenstruktur ermöglicht es, auf effiziente Weise auf Werte zuzugreifen und sie zu speichern (Real Python, 2023).

wurden sämtliche Satzzeichen entfernt. Zur Eliminierung von Unterschieden in der Groß- und Kleinschreibung wurden alle Buchstaben in Kleinbuchstaben konvertiert. Zusätzlich wurden numerische Zahlen in ihre ausgeschriebenen Formen umgewandelt, da Whisper die Zahlen in numerischer Form ausgibt, während sie in der Ground Truth ausgeschrieben sind. Das Umwandeln von ausgeschriebenen Zahlen zu numerischen Zahlen wird auch als inverse Textnormalisierung bezeichnet und soll die Lesbarkeit der Transkripte erhöhen. Weitere Beispiele für inverse Textnormalisierung sind Geldbeträge, Uhrzeiten und Interpunktion. Die normalisierten Texte wurden abschließend den jeweiligen Dataframes hinzugefügt, um die Weiterverarbeitung und den Vergleich der Transkripte mit den Ground Truth-Texten zu ermöglichen. Auf weitere Methoden der Textnormalisierung wie Stemming, Lemmatisierung oder die Entfernung weiterer Stoppwörter wurde hier verzichtet, da in der Analyse zuerst die Erhaltung der ursprünglichen Aussagen im Vordergrund steht. Die Anwendung von Lemmatisierung, Stemming oder umfangreichen Stoppwortlisten würde dazu führen, dass der ursprüngliche Text entfremdet wird. Dennoch bietet der Normalisierer die Möglichkeit, die Stoppwortliste zu erweitern und auch Lemmatisierung durchzuführen, falls gewünscht.

4. **Datenanalyse**:

   Nach abgeschlossener Textnormalisierung kommen nun die Metriken zum Einsatz, die als Bewertungsgrundlage der Genauigkeit des ASR-Modells dienen. Für diesen Zweck kommt die Bibliothek „jiwer“ zum Einsatz, um die WER zwischen der normalisierten Ground Truth und dem normalisierten Transkript zu ermitteln und eine quantitative Bewertung der Unterschiede zwischen den Texten zu ermöglichen. Ergänzend wird das Jaro-Winkler-Ähnlichkeitsmaß ermittelt, um die Ähnlichkeit zwischen den normalisierten Texten zu quantifizieren. Zur Identifikation der Wortgruppen, in denen die häufigsten Fehler auftreten, wird zunächst die Levenshtein-Distanz erhoben. Die Levenshtein-Distanz ist eine Metrik zur Messung der Unterschiede zwischen zwei Zeichenketten. Sie berechnet die minimale Anzahl von Bearbeitungsschritten, die erforderlich sind, um eine Zeichenkette in eine andere zu transformieren (Kutuzov, 2013). Diese Bearbeitungsschritte können Einfügungen, Löschungen und Substitutionen von Zeichen beinhalten. Hierfür ist es notwendig, eine Tokenisierung der Transkripte durchzuführen, um die Wörter anstelle von Zeichen zu betrachten. Nachdem alle fehlinterpretierten Wörter den Kategorien Einfügungen, Löschungen und Substitutionen zugeordnet wurden, können die Wortgruppen ermittelt werden. Hierbei werden die vorab trainierten POS-Tagger der Pipelines *en_core_web_sm*[13] und *de_core_news_sm*[14] aus der NLP-Bibliothek spaCy geladen. Die Tagger ermöglichen die Klassifizierung der Wortgruppen in den Sprachen Deutsch und Englisch. Zur abschließenden Visualisierung der Ergebnisse kommen die Bibliotheken Matplotlib und Seaborn zum Einsatz ,welche die grafische Darstellung der erhobenen Metriken und Wortgruppen ermöglichen, um die Analysen und Vergleiche der Transkripte übersichtlich und verständlich darzustellen.

[13] https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-3.7.0
[14] https://github.com/explosion/spacy-models/releases/tag/de_core_news_sm-3.7.0

5 Ergebnisse und Diskussion

5.1 Ergebnisse

In diesem Abschnitt werden die erzielten Ergebnisse der vorliegenden Studie präsentiert. Tabelle 5.1 zeigt eine Übersicht der Performance des Modells, sortiert nach Datensätzen. Bei der

Tabelle 5.1

Vergleich der Datensätze Common Voice DE, Common Voice EN und Gigaspeech

Datensatz	Gesamt	WER=0	WER<=0.1	sim_w=1	sim_w>=0.95	WER gesamt
CV-DE	32852	5083	7529	5095	28700	0.31
CV-EN	6983	2144	3046	2148	6198	0.19
GS	27700	14749	18038	14753	24884	0.11

Untersuchung der Leistungen der automatischen Spracherkennung ist zu erkennen, dass der deutschsprachige Teil des Common Voice Datensatzes (im Folgenden CV-DE) mit einer WER von 31% abschnitt, was im Vergleich zu den beiden englischsprachigen Datensätze (CV-EN für Common Voice English und GS für GigaSpeech), als drittbeste Leistung betrachtet werden kann. Die niedrigste WER wurde im GS-Datensatz mit etwa 11% erreicht. Von der Gesamtmenge des CV-DE-Datensatzes konnten insgesamt 5083 Audio-Dateien fehlerfrei transkribiert werden. Dies entspricht einem Anteil von ungefähr 15,5% der Daten. Im Vergleich dazu konnte das Modell den Datensatz CV-EN mit einer Null-Fehlerquote von 30,7% und GS mit 53,2% erkennen. Zu beachten ist, dass es einige Transkripte mit extrem hohen Fehlerraten gibt. Die höchste liegt bei 6200% und stammt aus dem Gigaspeech Datensatz. Diese Ausreißer haben möglicherweise negative Auswirkungen auf die durchschnittliche WER.

Abbildung 5.1

Datensatz Vergleich WER

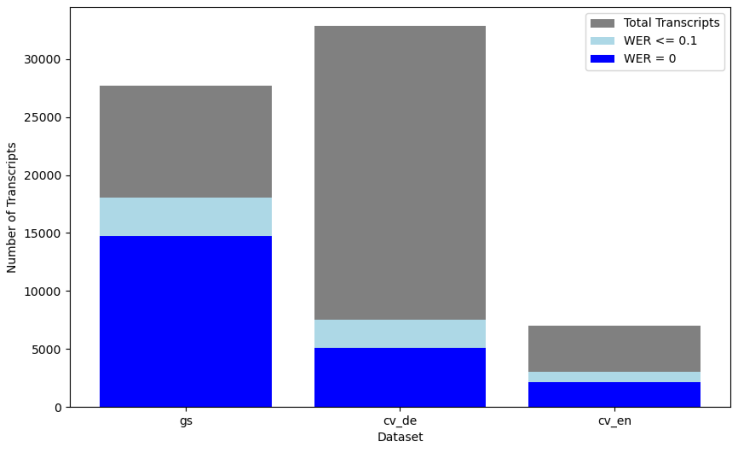
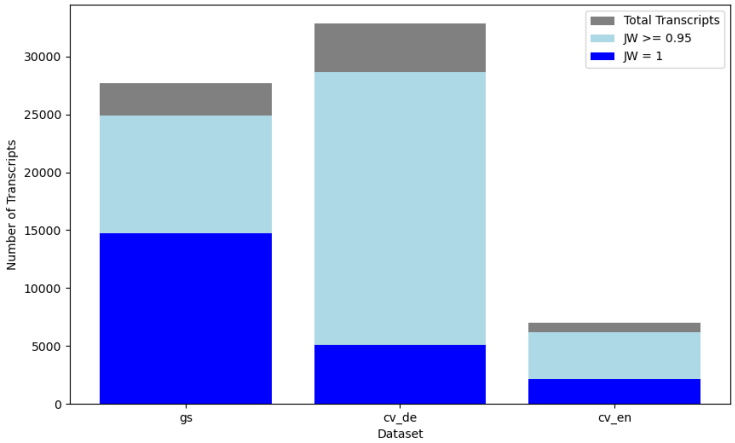


Abbildung 5.2

Datensatz Vergleich sim_w



Interessanterweise liegen beim Schwellenwert von sim_w>=0.95 89,8% der GS-Transkriptionen, 88,6% der CV-EN-Transkriptionen und, trotz der vergleichsweise hohen WER der Gesamtmenge, 87,3% der CV-DE-Transkriptionen vor. Dies zeigt, dass ein Großteil der erkannten Transkriptionen eine hohe Ähnlichkeit zur tatsächlichen Ausgangsdaten aufweist, auch wenn Fehler bei der Erkennung aufgetreten sind (siehe Abbildung 5.1 und Abbildung 5.2).

Im Detail ist zu erkennen, dass die vorherrschenden Unregelmäßigkeiten überwiegend in Form von Inkonsistenzen in der orthografischen Präzision und der Wortreihenfolge auftreten, was jedoch nur einen begrenzten Einfluss auf die inhaltliche Ähnlichkeit der beiden zu vergleichenden Texte hat. Tabelle 5.2 zeigt die charakteristischen Abweichungen zwischen der Ground Truth und dem systemgenerierten Transkript, wobei trotz Fehler die semantische Kohärenz und die inhaltliche Übereinstimmung weitgehend erhalten bleiben.

Tabelle 5.2

Vergleich Ground Truth und systemgeneriertes Transkript

Transkript	WER	sim_w
nach dem überlebten selbstmordversuch arbeitete daran weiter nach dem überlieb ein selbstmodversuch aber te der daran weiter	0.86	0.98
my very best concert bonnet my very best concert on it	0.4	0.96
so i really wan na thank you for the opportunity so i really want to thank you for the opportunity	0.2	0.98

Anmerkung. Jeweils die erste Zeile pro Zelle stellt die Ground Truth dar, die zweite Zeile das systemgenerierte Transkript

Hinsichtlich der Qualität der automatischen Spracherkennung des Common Voice Datensatzes in verschiedenen Altersgruppen (siehe Abbildung 5.3 und Tabelle 5.3) fällt auf, dass die Altersgruppen „teens“ und „twenties“ mit einer durchschnittlichen WER von 23,8% bzw. 25,1%

leicht unter dem Durchschnitt der WER aller Altersgruppen liegen, welche unter Ausschluss der Werte ohne Altersangaben 29,8% beträgt. Die Altersgruppe „thirties“ liegt knapp über dem Durchschnitt, mit einer WER von 30%. Die Altersgruppe „fourties“ stellt die größte Stichprobe innerhalb der vorliegenden Daten dar, mit einer durchschnittlichen WER von 28,0%. Die höchste durchschnittliche WER verzeichnen hingegen die Probanden in der Altersgruppe „fifties“ mit 35,1%. Bemerkenswert ist, dass die Altersgruppe „sixties“ im Vergleich zu anderen Altersgruppen eine vergleichsweise niedrige Fehlerquote von 12,5% aufweist.

Abbildung 5.3

WER nach Altersgruppe des CV Datensatzes

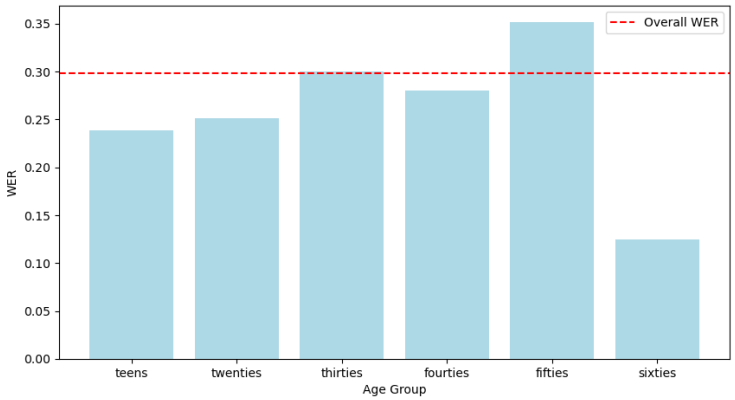


Tabelle 5.3

Verteilung CV nach Alter

Altersgruppe	Anzahl
teens	808
twenties	3310
thirties	6271
fourties	10359
fifties	7190
sixties	83
NaN	11814

Die Verteilung der WER nach Altersgruppe könnte möglicherweise mit der Verteilung der Daten nach Herkunft zusammenhängen. Eine nähere Betrachtung der Abbildungen 5.4 und Abbildung 5.5 zeigt, dass die Altersgruppen „teens“ und „sixties“ hauptsächlich aus englischsprachigen Daten bestehen, während „thirties“, „fourties“, und „fifties“ überwiegend deutsche Daten aufweisen. Zudem fällt auf, dass die WER englischsprachiger Texte in allen Altersgruppen unterhalb der WER deutschsprachiger Texte liegt. In den ursprünglichen Daten zeigte sich eine Korrelation zwischen dem Alter der Probanden von ’teens’ bis ’fifties’ und dem Anstieg der WER. Altersgruppe „sixties“ könnte aufgrund der geringen Anzahl an Daten als nicht repräsentativ angesehen werden. Allerdings offenbarte sich nach Aufteilung der Daten in englisch- und deutschsprachige Segmente, dass die WER in den englischsprachigen Daten stetig abnahm, während deutschsprachige Daten eine gewisse Schwankung aufweisen. Dieses Phänomen wird als Simpson-Paradoxon[1] bezeichnet und verdeutlicht, dass die aggregierten Daten nicht immer das vollständige Bild wiedergeben. Bei näherer Betrachtung der Verteilung des Jaro-Winkler-Ähnlichkeitsmaßes in Abbildung 5.6 wird erneut deutlich, dass die Transkripte in Bezug auf ihre Ähnlichkeit zur Ground Truth eine geringe Spannweite aufweisen. Dabei zeigen sich einige Unterschiede zwischen den Altersgruppen. Besonders in den Altersgruppen „thirties“ und „fifties“ fällt auf, dass die Spannweite der Ähnlichkeitswerte im Gegensatz zu den anderen Altersgruppen breiter ist, was auch die WER wiederspiegelt. Die Altersgruppe „sixties“ weist die geringste Spannweite auf. Insgesamt gibt es in den Daten eine beachtliche Anzahl von Ausreißern, die sich außerhalb der Hauptcluster befinden. Diese Ausreißer weisen Ähnlichkeitswerte im Bereich von

[1] Das Simpson-Paradoxon ist ein statistisches Paradoxon, bei dem eine Beziehung oder Korrelation zwischen zwei Variablen in aggregierten Daten scheinbar vorhanden ist, aber sich umkehrt oder verschwindet, wenn die Daten in Subgruppen aufgeschlüsselt werden. (Henze, 2021)

Abbildung 5.4

CV Datensatz: Vergleich Altersgruppen nach Herkunft

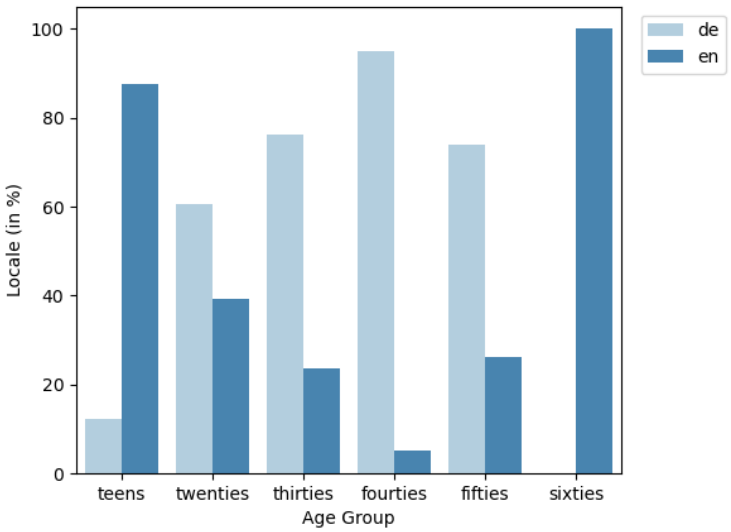
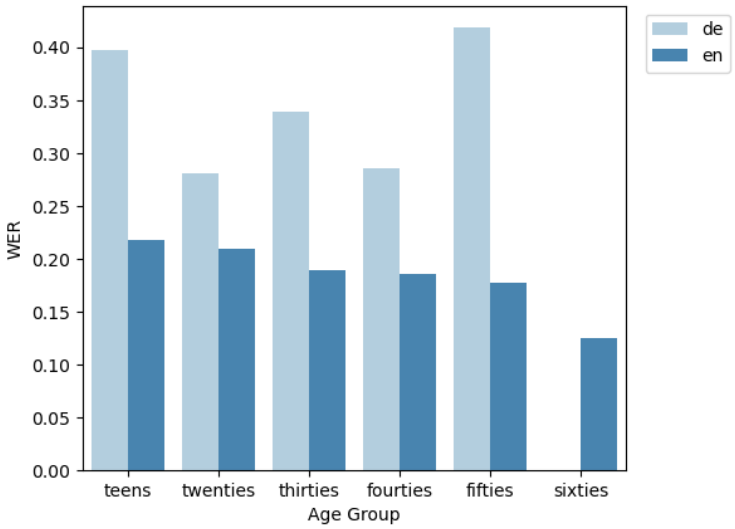


Abbildung 5.5

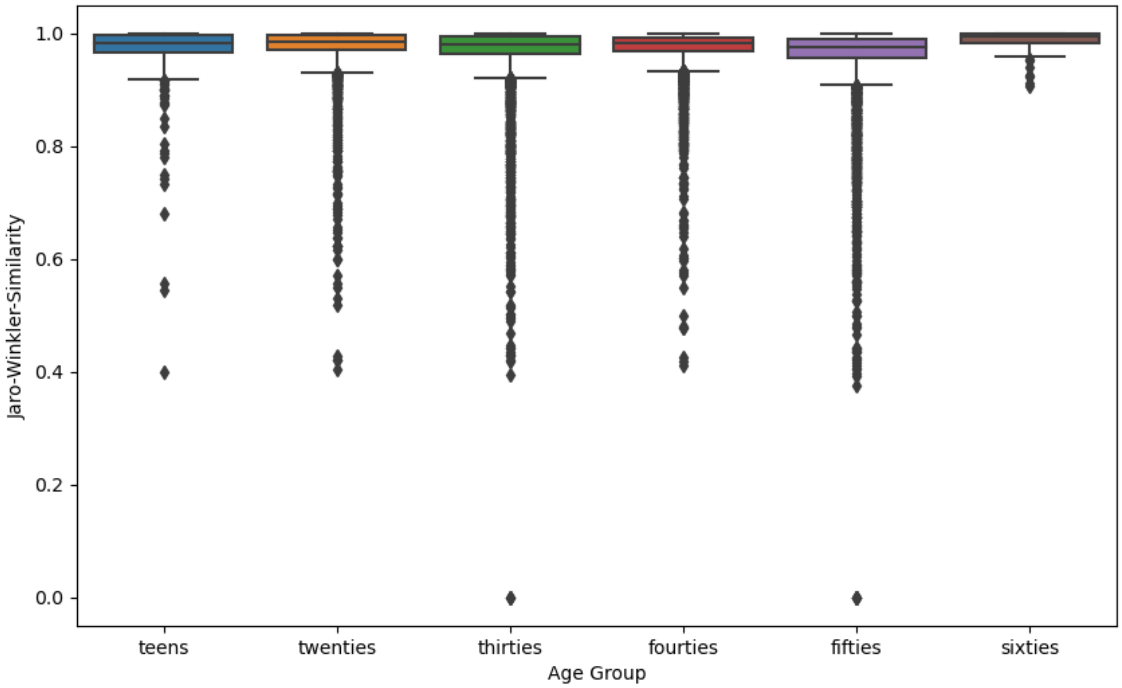
CV Datensatz: Vergleich WER nach Herkunft



0,4 bis 0,9 auf. Insbesondere in den Altersgruppen „thirties“ und „fifties“ finden sich Ausreißer, die eine Ähnlichkeit von nahezu 0 oder exakt 0 aufweisen. Dies ist auf Transkripte zurückzuführen, die in Sprachen übersetzt wurden, die keine lateinischen Buchstaben verwenden und sich gänzlich von der Ground Truth unterscheiden. Es ist wichtig zu betonen, dass diese Ergebnisse in einem breiteren Kontext betrachtet werden sollten. Die hier analysierten Daten sind weisen große Unterschiede innerhalb der Verteilung auf. Darüber hinaus ist eine eindeutige Schlussfolgerung allein auf Grundlage der Altersgruppen herausfordernd, da andere Faktoren, die die Spracherkennung beeinflussen könnten, nicht berücksichtigt wurden. Insbesondere die mögliche Präsenz von Störgeräuschen oder Umgebungsgeräuschen während der Aufnahme kann erheblichen Einfluss auf die Qualität der Spracherkennung haben. Zudem sind sprachliche

Abbildung 5.6

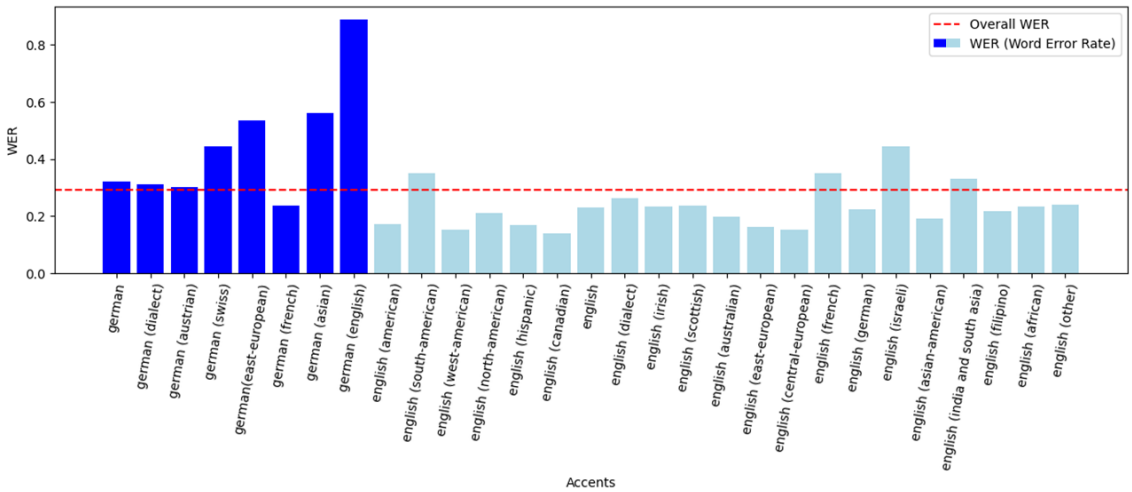
Jaro-Winkler-Ähnlichkeitsmaß verschiedener Altersgruppen im CV Datensatz



Einflüsse wie Dialekte oder Akzente nicht berücksichtigt worden, was ebenfalls eine bedeutende Rolle spielen kann, wie in Abbildung 5.7 dargestellt.

Abbildung 5.7

WER verschiedener Akzent-Gruppen im CV Datensatz

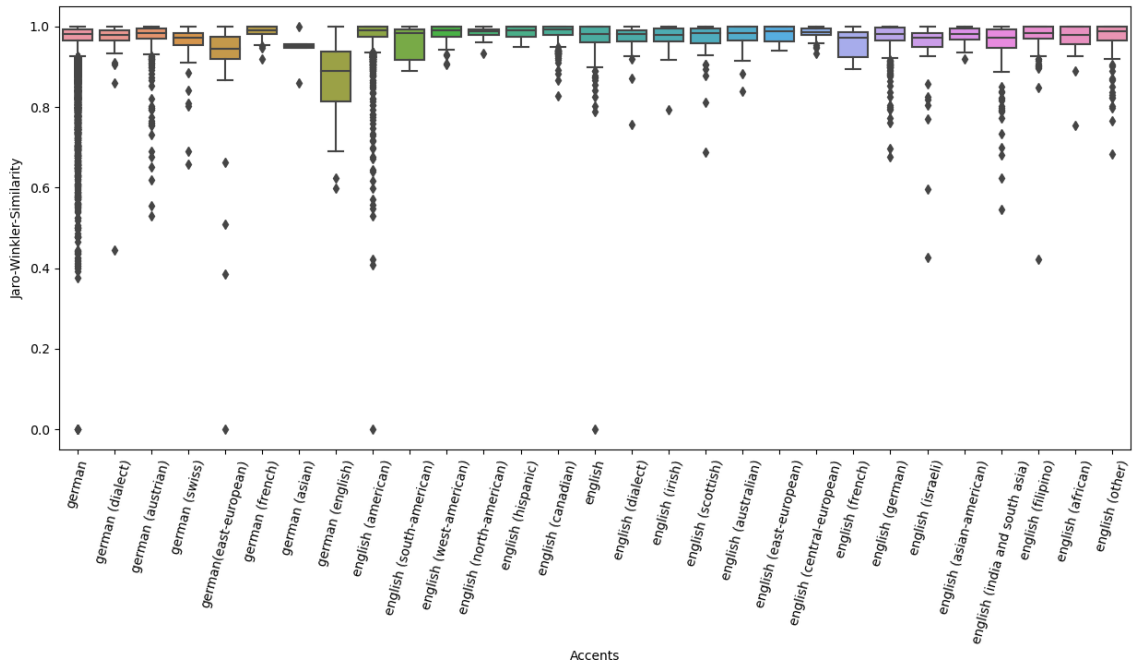


Anhang A liefert eine Übersicht über die Verteilung der verschiedenen Akzente im CV Datensatz. Im Gegensatz zur Unterteilung in Altersgruppen gibt es hier eine weitaus vielfältigere Auswahl an Akzenten, insgesamt acht deutsche und 21 englische Variationen. Auffallend ist, dass die Variationen der deutschen Sprache im Durchschnitt schlechter abschneiden als die Variationen der englischen Sprache. Besonders bemerkenswert ist die Akzent-Gruppe „german (english)“,

die mit einer WER von etwa 90% heraussticht. Ein möglicher Grund dafür könnte die phonologische Ähnlichkeit von „german (english)“ zur englischen Sprache sein, weshalb diese Texte fälschlicherweise als Englisch klassifiziert wurden. Innerhalb der deutschen Akzent-Gruppen erzielt „german (french)“ die geringste WER von 23,5%. Die Gruppe „german“, die den größten Anteil des Datensatzes mit 23.017 Beispielen ausmacht, hat mit 32% eine leicht überdurchschnittliche WER. Der Durchschnitt liegt bei 29,1%. Überrachenderweise liegt die WER von deutschsprachigen Dialekten leicht unterhalb der WER von allgemein deutscher Sprache. Von allen deutschen Variationen liegt lediglich „german (french)“ unterhalb des Durchschnitts. Unter den englischen Variationen weisen vier Akzent-Gruppen eine WER oberhalb des Durchschnitts auf: „english (south-american)“, „english (french)“, „english (israeli)“ und „english (india and south asia)“. Wobei „english (israeli)“ gleichzeitig die höchste Abweichung von 44,5% aufweist. Die geringste WER im gesamten Datensatz verzeichnet die Akzent-Gruppe „english (canadian)“ mit lediglich 14%.

Abbildung 5.8

Jaro-Winkler-Ähnlichkeitsmaß verschiedener Akzent-Gruppen im CV Datensatz



Ergänzend zur WER kann auch hier das Jaro-Winkler-Ähnlichkeitsmaß herangezogen werden, um eine Aussage über die Ähnlichkeit der systemgenerierten Transkripte gegenüber der Ground Truth zu treffen. Zunächst fällt auf, dass, ähnlich wie bei der Verteilung nach Altersgruppen, auch bei der Aufteilung nach Akzenten eine große Anzahl von Ausreißern vorhanden ist. Besonders auffällig ist die Akzent-Gruppe „german (english)“, die nicht nur die höchste WER aufweist, sondern auch die mit Abstand größte Spannweite hinsichtlich des Jaro-Winkler-Ähnlichkeitsmaß hat. Zudem liegt die Ähnlichkeit hier deutlich unter den anderen Akzent-Gruppen, was darauf hin deutet, dass die Fehler in dieser Gruppe nicht auf kleinere Ungenauigkeiten beschränkt sind. Des Weiteren weisen die Kategorien „german (east-european)“, „english (south-american)“ und „english (french)“ ebenfalls recht große Spannweiten auf. Im Gegensatz dazu zeigt die Kategorie „english (canadian)“ eine sehr geringe Spannweite und vergleichsweise wenige starke Ausreißer,

was nochmal unterstreicht, dass die Qualität der Spracherkennung hier stabiler ist. Ein weiterer wichtiger Aspekt ist die Tatsache, dass die Gruppe „german (asian)“ nicht nur generell stark unterrepräsentiert ist im Datensatz, sondern auch zwei Ausreißer aufweist. Diese Ausreißer könnten die durchschnittliche WER und das Ähnlichkeitsmaß dieser Kategorie beeinflussen.

Die Analyse des Datensatzes in Abhängigkeit von verschiedenen Akzenten wirft mehrere wichtige Probleme und Herausforderungen auf, die im Kontext des vorliegenden Datensatzes berücksichtigt werden müssen. Überwiegend deutschsprachige Daten dominieren den Datensatz, während englische Daten unterrepräsentiert sind. Gleichzeitig sind englische Akzente weitaus fragmentierter als deutsche. Darüber hinaus sind einige Akzentgruppen stark unterrepräsentiert. In solchen Fällen können selbst wenige Ausreißer oder schlecht transkribierte Datenpunkte innerhalb dieser Gruppen erhebliche Auswirkungen auf die durchschnittliche Wortfehlerrate haben. Dies macht es schwierig, verlässliche Schlussfolgerungen über die tatsächliche Spracherkennungsqualität für diese Akzente zu ziehen. Ein weiteres Problem besteht darin, dass die gemappten Kategorien (z. B. „german“ und „english“) eine Vereinfachung des tatsächlichen Bildes darstellen. Zum Beispiel kann „german“ viele verschiedene deutsche Akzente und Dialekte umfassen, die in Bezug auf die Verständlichkeit variieren. Eine solche Vereinfachung kann dazu führen, dass Unterschiede innerhalb dieser Gruppen nicht erfasst oder verwaschen werden. Des Weiteren ist Zuverlässigkeit der Angaben zu Akzenten unklar, und sprachliche Nuancen sowie Sprachfehler und Störgeräusche werden nicht berücksichtigt. Eine umfassendere Analyse würde eine differenzierte Berücksichtigung der sprachlichen Vielfalt und der spezifischen Herausforderungen bei der automatischen Spracherkennung erfordern.

Im Gegensatz zum Common Voice Datensatz verfügt der GigaSpeech Datensatz über keine Label zu Altersgruppen oder Akzenten, jedoch über Label bezüglich der Quelle under der thematischen Einordnung der Audio-Datei, welche im Folgenden betrachtet werden.

Abbildung 5.9

WER nach Quelle der Audio-Dateien des GS Datensatzes

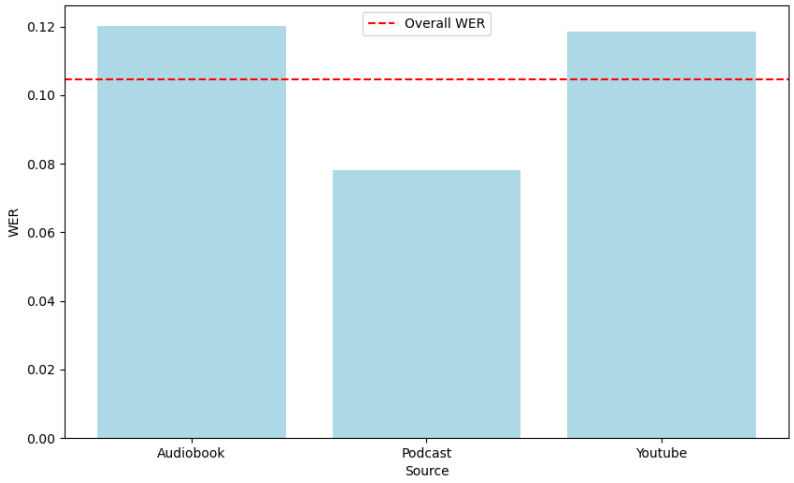


Tabelle 5.4

Verteilung GS nach Quelle

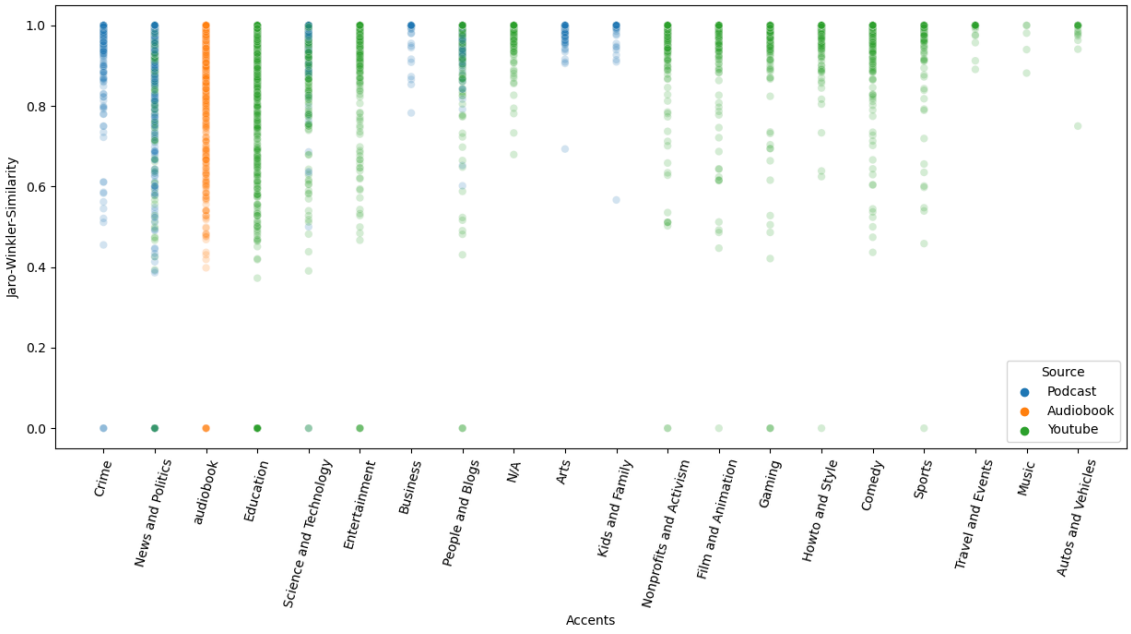
Quelle	Anzahl
Youtube	11591
Podcast	8794
Audiobook	7315

In Abbildung 5.9 ist zu erkennen, dass die Datenquelle „Audiobook“ die höchste WER mit 12% aufweist. Die Datenquelle „Podcast“ mit 7,8% die geringste WER. Die Datenquellen „Audiobook“ und „Youtube“ liegen beide über dem Durchschnitt von 10,4%. Tabelle 5.4 zeigt die Verteilung

der Daten nach Quelle. Da rein von der Betrachtung der Datenquelle her, nicht die thematische Vielfalt der Daten abgebildet wird, ist es hilfreich die Quelle im Kontext der Themen-Kategorien des GigaSpeech Datensatzes zu betrachten. Abbildung 5.10 zeigt die Verteilung des Jaro-Winkler-Ähnlichkeitsmaßes über thematische Kategorien mit farblicher Unterscheidung von der Quelle.

Abbildung 5.10

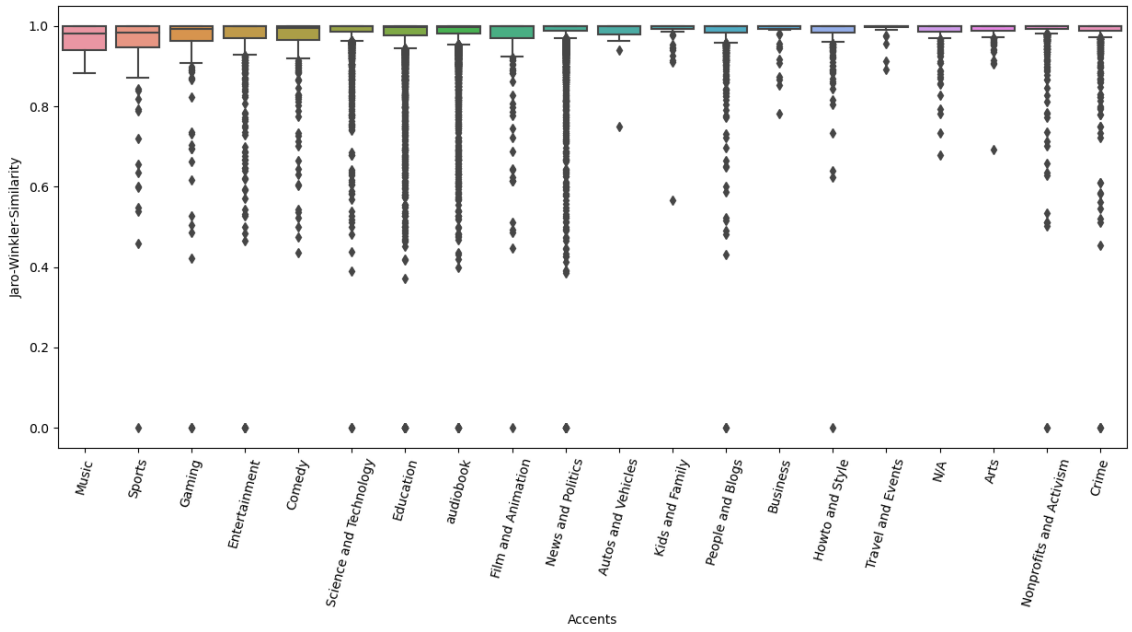
Jaro-Winkler-Ähnlichkeitsmaß verschiedener Kategorien des GS Datensatzes mit farblicher Unterteilung der Herkunftsquelle



Die Audio-Beispiele in den Kategorien „Crime“, „Business“, „Arts“ und „Kids and Family“ stammen hauptsächlich aus Podcasts. „Audiobooks“ bildet eine eigene Kategorie und ist gleichzeitig auch am stärksten vertreten. Die Audio-Dateien der Kategorien „Science and Technology“, „News and Politics“ und „People and Blogs“ sind gemischter Herkunft und stammen sowohl von YouTube als auch von Podcasts. Ebenfalls ist zu erkennen, dass die Kategorien „Kids and Family“, „Business“, „Travel and Events“, „Autos and Vehicles“, und „Music“schwach vertreten sind. Die genaue Verteilung der Daten über die Kategorien ist der Tabelle „Anzahl Daten pro Kategorie im Gigaspeech Datensatz“ im Anhang B zu entnehmen. Insgesamt zeigt die Analyse der Ähnlichkeitsmaße, dass die Transkripte in den meisten Kategorien im oberen Bereich um 0,9 bis 1 liegen, was auf eine hohe Ähnlichkeit zur Ground Truth hinweist. Es gibt jedoch auch viele Ausreißer, die eine geringere Ähnlichkeit aufweisen, was auf eine gewisse Varianz in der Qualität der Transkriptionen hinweist. Eine genauere Analyse von Ausreißern und der durchschnittlichen Ähnlichkeit innerhalb der Kategorien erlaubt der Boxplot in Abbildung 5.11. Die Kategorie „Music“ zeigt im Vergleich zu den anderen Themenkategorien die niedrigste Ähnlichkeit zur Ground Truth. Jedoch sollte berücksichtigt werden, dass „Music“ mit nur fünf Beispielen in dieser Analyse unterrepräsentiert ist, und die Ergebnisse daher möglicherweise nicht die Gesamtheit der Kategorie widerspiegeln. Die Kategorie „Sports“ weist die größte Bandbreite in den Ähnlichkeitsmaßen auf, was darauf hinweist, dass die Qualität der Transkriptionen in

Abbildung 5.11

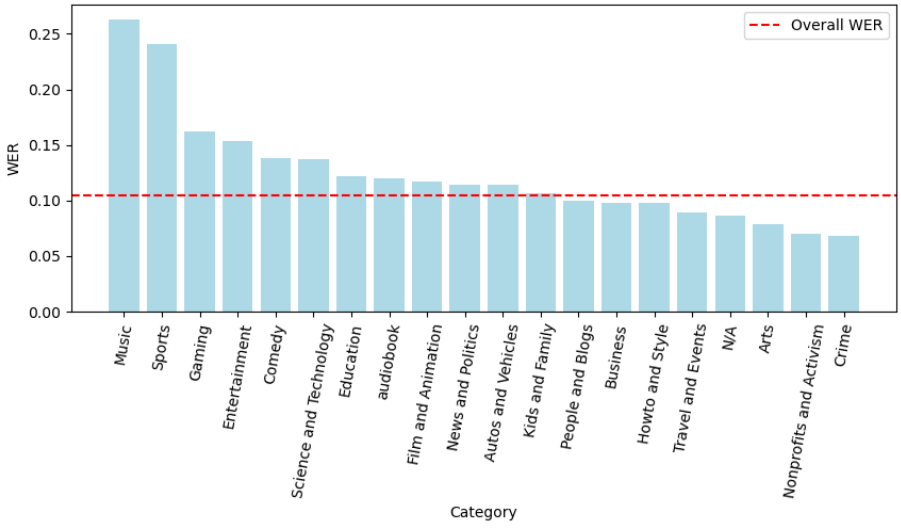
Jaro-Winkler-Ähnlichkeitsmaß verschiedener Kategorien des GS Datensatzes



dieser Kategorie stärker variiert als in anderen. Der Boxplot verdeutlicht erneut, dass es viele Ausreißer gibt, insbesondere im Bereich von 0,4 bis 1. Auch ist festzustellen, dass die Kategorien „Arts“, „Nonprofits and Activism“ und „Crime“ die höchsten Ähnlichkeiten zur Ground Truth aufweisen.

Abbildung 5.12

WER verschiedener Kategorien des GS Datensatzes



Für eine umfassende Analyse der Daten ist es entscheidend, die WER im Kontext verschiedener Themenkategorien zu betrachten. Abbildung 5.12 zeigt, dass es sowohl oberhalb als auch unterhalb des Durchschnitts von 10,5% jeweils zwei Kategorien gibt, die hervor stechen. Die Kategorien „Music“ und „Sports“ zeichnen sich durch eine hohe WER von 26,3% und 25% aus. Im

Gegensatz dazu weisen die Kategorien „Crime“ und „Nonprofits and Activism“ vergleichsweise niedrige WER-Werte von 7% und 6,8% auf.

Aufgrund der Relevanz für diese Arbeit, wird abschließend die Kategorie „Science and Technology“ betrachtet. Tabelle 5.5 liefert einen Überblick über Verteilung der Daten während Abbildung 5.13 Einblicke in die Qualität der Transkriptionen, abhängig von ihrer Herkunftsquelle, liefert.

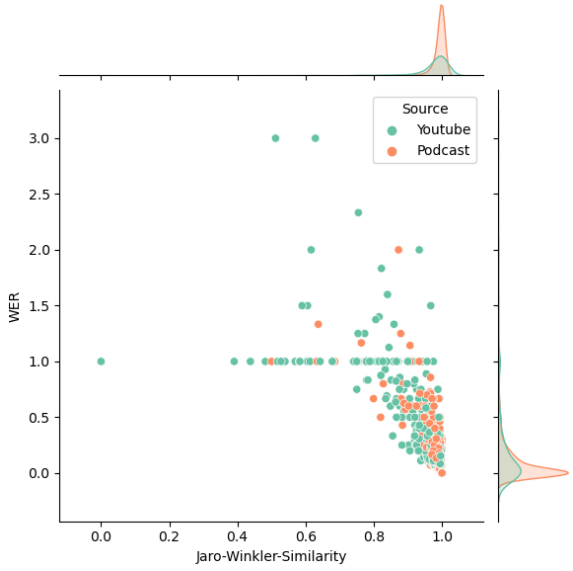
Tabelle 5.5

Überblick „Science and Technology“ aus dem GS Datensatz

Quelle	Anzahl	WER gesamt	WER=0	sim_w=1	sim_w>=0.95
Youtube	954	0.14	460	460	797
Podcast	1425	0.08	791	792	1375

Abbildung 5.13

WER und Jaro-Winkler-Ähnlichkeitsmaß der Kategorie „Science and Technology“ des GS Datensatzes, eingefärbt nach Herkunftsquelle



Dabei fällt auf, dass sowohl die von Youtube stammenden Daten, als auch diejenigen aus Podcasts, hinsichtlich des Jaro-Winkler-Ähnlichkeitsmaß eine geringe Streuung aufweisen. Interessanterweise zeigt sich, dass die WER bei den Youtube-Daten stärker gestreut ist als bei den Podcast-Daten. Dies spiegelt sich auch in der durchschnittlichen WER wider, die bei Podcasts mit 7,8% niedriger ausfällt im Vergleich zu 13,7% bei Youtube. Bei der Betrachtung der Null-Fehlerquote zeigt sich, dass bei Youtube etwa 48% der Transkriptionen fehlerfrei sind, während Podcasts eine Quote von etwa 55% aufweisen. Interessanterweise fällt der Unterschied für die Bedingung sim_w>=0.95 deutlich größer aus. Hier liegen Podcasts mit über 96% vorne, während Youtube auf 83,5% kommt. Auffällig ist auch die enorme Spannweite des Jaro-Winkler-Ähnlichkeitsmaß bei WER = 1. Selbst bei einer WER von 1 oder sogar 2, weisen manche Daten noch eine hohe Ähnlichkeit zur Ground Truth auf. Zudem scheint eine WER von größer 1 weitaus seltener

vorzukommen, als eine WER von 1.

Bei der Betrachtung der Ergebnisse sollte, wie bereits bei den Ergebnissen des Common Voice Datensatzes, berücksichtigt werden, dass Informationen über eventuelle Störgeräusche in den Audioaufnahmen fehlen, die einen erheblichen Einfluss auf die Qualität der Transkriptionen haben können. Darüber hinaus fehlen dem GigaSpeech Datensatz wichtige Zusatzinformationen zu den Sprechern, wie etwa Akzenten und Dialekten, was eine feinere Analyse der Qualität der Transkriptionen ermöglichen würde. Zudem ist zu beachten, dass auch im GigaSpeech Datensatz einige Themenkategorien unterrepräsentiert sind.

Nach Ausführlicher Betrachtung der Ergebnisse beider Datensätze folgt nun ein kurzer Vergleich der verschiedenen Whisper Modelle *tiny*, *base*, *small*, *medium* und *large*.

Abbildung 5.14

Vergleich Whisper-Modelle auf verschiedenen Datensätzen

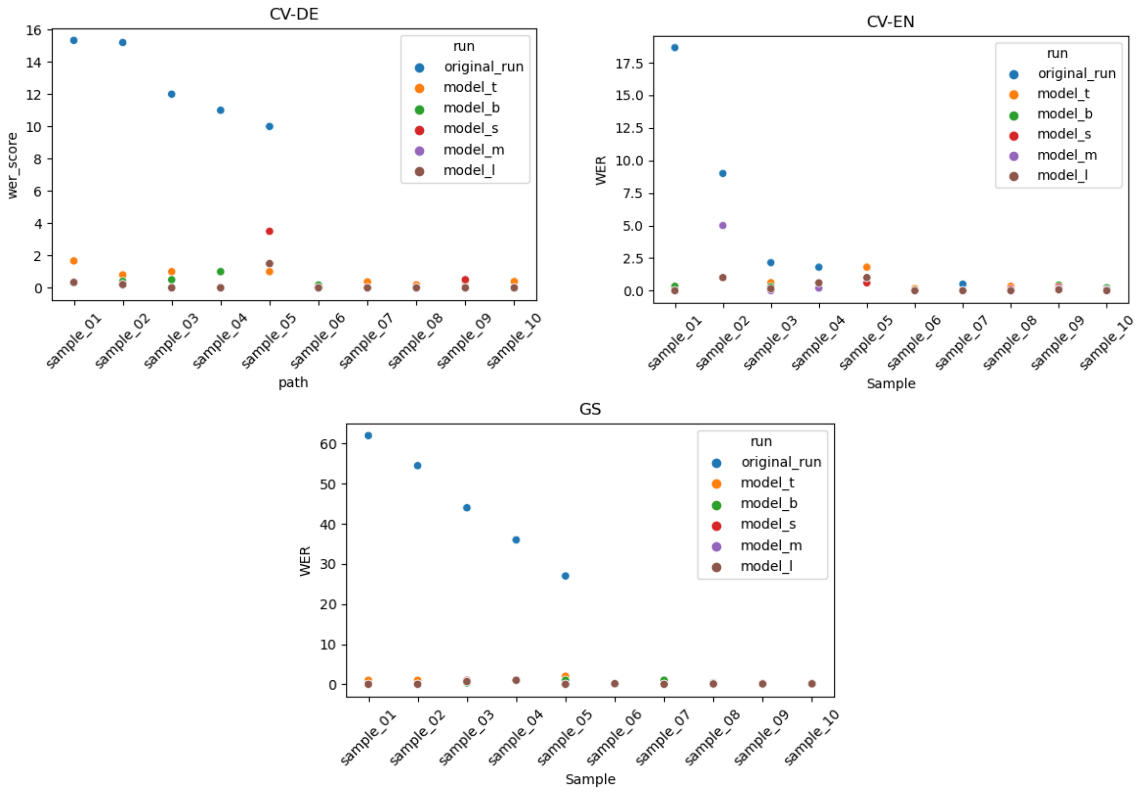


Tabelle 5.6

Whisper-Modell Vergleich auf Basis der WER

Datensatz	original run	tiny	base	small	medium	large
CV-DE	3.55	0.49	0.24	0.18	0.07	0.07
CV-EN	1.47	0.41	0.28	0.15	0.16	0.14
GS	5.47	0.31	0.21	0.17	0.14	0.14

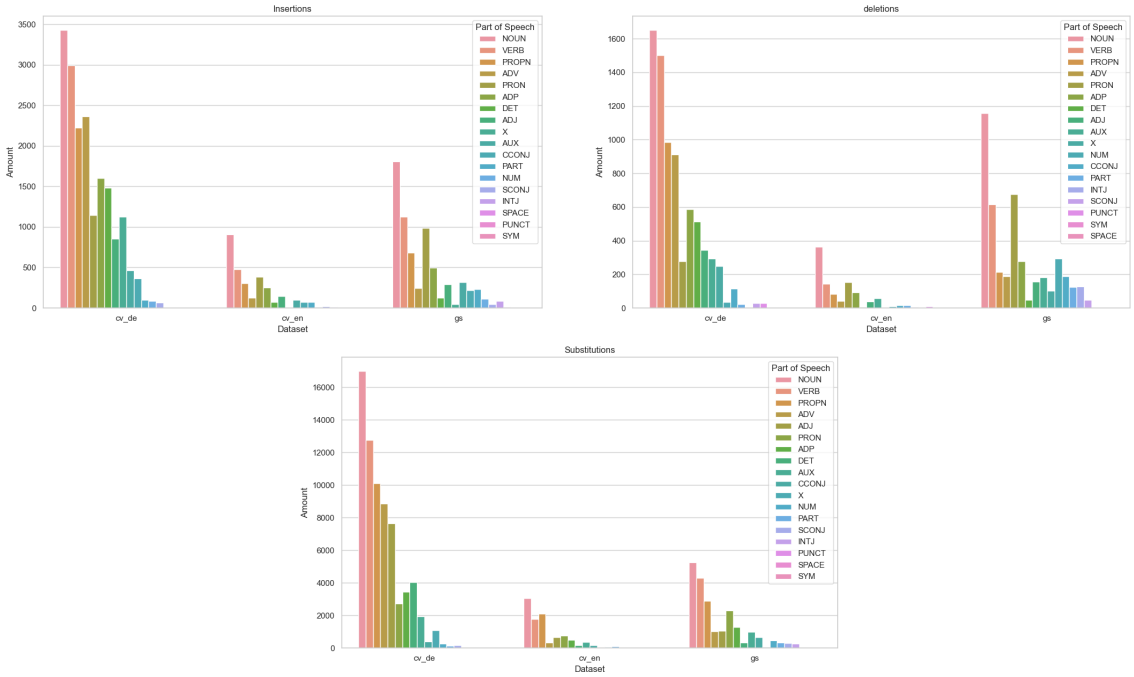
Die Testfälle der in Abbildung 5.14 und Tabelle 5.6 dargestellten Ergebnisse wurden mit je zehn

Proben pro Datensatz durchgeführt, wobei jede Probe aus fünf zufällig ausgewählten Daten und den fünf größten Ausreißern pro Datensatz. Grundlegend verbessert sich die Leistung der Modelle mit zunehmender Größe. Eine weitere wichtige Erkenntnis ist, dass die Voreinstellung der Sprache einen erheblichen Einfluss auf die Ergebnisse hat. Dies wurde besonders deutlich, als ein erneuter Durchlauf mit dem tiny-Modell durchgeführt wurde, wobei die Zielsprache im Voraus festgelegt wurde. Diese Anpassung führte zu einer signifikanten Verbesserung der WER, wenngleich diese noch kein optimales Niveau erreicht. Die detaillierten Plots zeigen, dass dennoch Schwankungen vorhanden sind. Selbst das größte Modell weicht in manchen Fällen stärker ab, als kleinere Modelle, wie etwa in *sample_05* des CV-EN-Datensatzes. Interessanterweise erzielte das medium-Modell auf den Samples des CV-DE-Datensatzes und des GS-Datensatzes jeweils die gleiche durchschnittliche Fehlerquote wie das large-Modell. auf dem CV-EN-Datensatz hingegen schneidet es etwas schlechter ab, als das nächst kleinere small-Modell. Zwar zeigen die Daten einen Zusammenhang zwischen Modell-Größe und WER, jedoch muss beachtet werden, dass es sich hierbei nur um eine kleine Auswahl an Daten handelt. Um eine umfassende Aussage zu treffen, müssten eine größer angelegte Evaluation erfolgen. Es ist jedoch davon auszugehen, dass die durchschnittliche WER auf der Gesamtmenge wesentlich geringer ausfallen könnte, als den hier verwendeten Testfällen, da diese zur Hälfte aus den größten Ausreißern bestehen, welche Whisper vor Herausforderungen zu stellen scheinen. Nur durch das Verständnis der Ursachen von Fehlern können geeignete Lösungen entwickelt und realistische Strategien entwickelt werden, um die Qualität von sprachbasierten Systemen zu optimieren. Die Analyse von Fehlern ist daher ein essenzieller Schritt, um die Leistung von Spracherkennungssoftware und -anwendungen in realen Anwendungsfällen zu verbessern.

Um einen tieferen Einblick in den Ursprung der Fehler zu erhalten, wurden die Fehler nach Löschungen, Einfügungen und Substitutionen kategorisiert und auf Wortgruppen runtergebrochen und in Abbildung 5.15 visualisiert. Die Plots liefern interessante Einblicke in die Verteilung von Fehlern und Fehlerarten. Der CV-DE Datensatz weist im Vergleich zu den anderen Datensätzen eine höhere Fehleranzahl auf. Dies ist vermutlich auf seine Größe und die damit verbundene größere Anzahl von Transkriptionen zurückzuführen. Zudem zeigt dieser Datensatz auch die höchste durchschnittliche Fehlerrate. Ein weiterer bemerkenswerter Punkt ist die Verteilung der Fehlerarten, wobei Substitutionen deutlich häufiger auftreten als Löschungen und Einfügungen. Die Analyse der Wortgruppen, in denen Fehler auftreten, zeigt, dass Nomen am häufigsten von Fehlern betroffen sind, gefolgt von Verben, Eigennamen, Adverben und Adjektiven. Dabei ist jedoch zu beachten, dass auf eine Normalisierung der Wortgruppen hinsichtlich ihrer relativen Häufigkeit aufgrund der recht komplexen Ermittlung verzichtet wurde. Besonders interessant ist, dass die englischsprachigen Datensätze auffällig viele Löschungen und Einfügungen von Nomen aufweisen. Die hier präsentierten Ergebnisse stellen jedoch lediglich eine Annäherung an die tatsächlichen Gegebenheiten dar. Die Verwendung des Levenshtein-Distanzmaßes zur Identifikation von Löschungen, Substitutionen und Einfügungen ist zwar ein nützliches Werkzeug, welches jedoch auch seine Begrenzungen und Schwierigkeiten hat. Zum einen berechnet die Levenshtein-Distanz die minimale Anzahl von Operationen zur Umwandlung eines Textes in einen anderen, ohne die räumliche Anordnung von Buchstaben und Wörtern zu berücksichtigen, was die Identifikation von Löschungen und Einfügungen in komplexen Texten erschweren kann. Zum anderen sind Spacy-Modelle möglicherweise nicht für alle Sprachen und Textarten

Abbildung 5.15

Vergleich der Fehler pro Wortgruppe aufgeteilt nach Datensatz für Einfügungen, Löschungen und Substitutionen



gleichermaßen gut geeignet. Die Qualität der Erkennung kann je nach der Komplexität der Texte und der Sprache variieren.

Abrundend sollen im Folgenden die erstellten Transkripte des Open Science Radio Podcasts kurz analysiert werden, die im Anhang C vorliegen. Diese Transkripte wurden mithilfe des small-Modells generiert, welches einen ausgewogenen Kompromiss zwischen Geschwindigkeit und Leistungsstärke bietet. Jedes Transkript wurde zusammen mit relevanten Metadaten in einer JSON-Datei gebündelt. Bei der Überprüfung der Texte fällt auf, dass beide Transkripte Fehler aufweisen, insbesondere im Bereich der Rechtschreibung, Interpunktion, Grammatik und des Ausdrucks. Diese Fehler beeinflussen die Lesbarkeit und Verständlichkeit der Texte. Besonders grammatikalische Fehler und das Fehlen oder die falsche Verwendung von Interpunktion können die Struktur eines Textes erheblich beeinträchtigen. Rechtschreibfehler hingegen werden oft überlesen und stören den Lesefluss kaum. Auch zeigt sich, dass der deutschsprachige Text im Vergleich zum englischsprachigen Transkript eine höhere Fehleranzahl aufweist. Dies könnte auf die Komplexität der deutschen Grammatik und Rechtschreibung zurückzuführen sein. Es ist jedoch wichtig zu betonen, dass trotz dieser Fehler der eigentliche Sinn und Inhalt beider Transkripte im Wesentlichen erhalten bleiben. Des Weiteren wurde beobachtet, dass Whisper keine Unterscheidung für verschiedene Sprecher vornimmt. Dies könnte insbesondere bei Interviews oder Gesprächen hilfreich sein, um klarzustellen, wer was gesagt hat. Eine solche Unterscheidung könnte die Lesbarkeit und das Verständnis des Transkripts erheblich verbessern.

5.2 Diskussion

Es wurde die Anwendbarkeit des ASR-Modells Whisper auf die Transkription von wissenschaftlichen Podcasts untersucht. Als Grundlage dafür dienten Auszüge aus dem bilingualen Common Voice Datensatz von Mozilla und dem englischsprachigen GigaSpeech Datensatz. Zur Transkription wurde das kleinste Whisper-Modell verwendet. Die Ergebnisse zeigen, dass die WER abhängig von der Sprache und dem verwendeten Datensatz variiert. Es wurde deutlich, dass englische Daten besser verarbeitet wurden als deutsche. Zusätzlich zur WER wurde das Jaro-Winkler-Ähnlichkeitsmaß verwendet. Trotz hoher WER wurde festgestellt, dass die Transkripte eine hohe Ähnlichkeit zur Ground Truth aufwiesen. Dies deutet darauf hin, dass es sich meist um kleinere Orthographiefehler, Wortdreher, Löschungen und Einfügungen handelt, die im Gesamteindruck jedoch wenig ins Gewicht fallen. Im Bezug auf Altersgruppen des Common Voice Datensatzes zeigte sich, dass es eine scheinbare positive Korrelation zwischen Altersgruppen und WER gab. Diese Beziehung erwies sich nach Unterteilung der Daten in englisch- und deutschsprachige Sprecher jedoch als nicht eindeutig. Bei der Unterteilung in Akzentgruppen konnte die Vermutung, dass Akzente und Dialekte tendenziell schwerer zu transkribieren sind, bestätigt werden. Die Untersuchung des Gigaspeech Datensatzes ergab, dass Podcasts besser transkribiert wurden als Audiobooks oder Auszüge von YouTube-Videos. Insbesondere in der Kategorie „Science and Technology“ schnitten Podcasts im Vergleich zu YouTube-Videos gut ab, mit einer hohen Ähnlichkeit zur Ground Truth und einer niedrigen WER von 7,8%. In einem direkten Modellvergleich zeigte sich, dass die WER mit zunehmender Modellgröße abnahm. Dies zeigt, dass die Leistung des ASR-Modells mit mehr Ressourcen und Trainingsdaten verbessert werden kann. Schlussendlich wurden Nomen, Verben, Eigennamen, Adjektive und Adverben als häufigste Fehlerquelle identifiziert.

Die Resultate dieser Studie verdeutlichen, dass die Anwendung von Whisper äußerst geeignet ist, um zuverlässige Transkripte zu erstellen. Selbst das kleinste Modell, das im Vergleich zu seinen größeren Gegenstücken die geringste Leistung aufweist, ist in der Lage, englische Transkripte zu generieren, welche nahezu die Qualität manueller Transkriptionen erreichen. Die Qualität manueller Transkripte ist von verschiedenen Faktoren abhängig und weist eine erhebliche Spannbreite zwischen 0,1% und 11,3% auf, wie in früheren Studien (Stolcke und Droppo, 2017; Zayats et al., 2019) dokumentiert. Mögliche Einflussfaktoren umfassen die Audioqualität, die Erfahrung des Transkribenten, die Vertrautheit mit regionalen Dialekten und Akzenten, Zeitaufwand und Sorgfalt sowie das Fachwissen (Stolcke und Droppo, 2017; Zayats et al., 2019). Es ist besonders bemerkenswert, dass das tiny-Modell bei der Verarbeitung englischsprachiger Daten sogar eine niedrigere Fehlerquote aufweist als die automatisch generierten Transkripte des Spotify-Datensatzes, die unter Verwendung der Google Cloud STT API[2] erstellt wurden und eine hohe WER von 18,1% aufzeigen (Clifton et al., 2020). Es bleibt jedoch unklar, ob diese Fehlerquote repräsentativ für die auf Spotify verfügbaren Transkripte ist.

Deutschsprachige Texte schneiden im Durchschnitt schlechter ab, was wahrscheinlich auf die fehlende Voreinstellung der Zielsprache und dem hohen Anteil englischsprachiger Trainingsda-

[2]Application Programming Interface (API) ist ein Begriff aus der Informatik. Eine API ist eine Sammlung von Regeln und Protokollen, die es verschiedenen Softwareanwendungen ermöglichen, miteinander zu kommunizieren und Informationen auszutauschen (Bloch, 2006).

ten zurückzuführen ist. Daher besteht die Vermutung, dass Whisper ohne Sprachvoreinstellung wesentlich besser in der Lage ist, englische Texte zu identifizieren als Texte in anderen Sprachen. Eine weitere bemerkenswerte Beobachtung ist die erhebliche Variation in der Transkriptionsqualität zwischen verschiedenen Akzenten und Dialekten. Dies legt nahe, dass Whisper Schwierigkeiten hat, mit selteneren Sprachvariationen umzugehen, was in Abschnitt 2.2.4 als „Long-Tail Effekt“ bezeichnet wurde. Ähnliche Herausforderungen könnten sich auch bei Fachausdrücken ergeben, die nur in bestimmten Nischen verwendet werden. Interessanterweise steht diese Beobachtung im Widerspruch zu einer der niedrigsten Fehlerquoten, die in dieser Studie bei englischsprachigen Podcast-Daten aus dem wissenschaftlichen Bereich erzielt wurde. Dies wirft die Frage auf, weshalb Whisper in diesem Kontext so gut abschneidet. Eine mögliche Erklärung für die bessere Leistung bei englischsprachigen wissenschaftlichen Podcasts könnte in der vergleichsweise niedrigeren Anzahl von Phonemen in der englischen Sprache liegen (Xiong, 2023).Infolgedessen könnten seltene oder spezialisierte Begriffe leichter von ähnlichen phonetischen Lauten abgeleitet werden, welche das Modell wiederum als Muster im Log-Mel Spektrogramm identifiziert. Ein anderer Grund könnte in den Trainingsdaten zugrunde liegen, die aus verschiedenen Datensätzen mit akademischem Hintergrund stammen, wie LibriSpeech oder TED-LIUM 3 (Radford et al., 2022). Diese Datensätze könnten dazu beitragen, dass Whisper besser mit den in wissenschaftlichen Podcasts verwendeten Fachtermini umgehen kann.

Neben von der Sprachvariation abhängigen Schwankung gibt es auch einige Ausreißer in den Ergebnissen, bei denen die Fehlerraten extrem hoch sind oder überhaupt keine Ähnlichkeit zur Ground Truth vorhanden ist. Eine WER von über 100% bedeutet, dass das Transkript mehr Fehler enthält, als es überhaupt Wörter im Text gibt. Dies kann auftreten, wenn einzelne Wörter falsch interpretiert und in mehrere kurze Worte aufgeteilt werden. Dennoch wurde im Einzelfall eine ungewöhnlich hohe Fehlerquote von 6200% erzielt. Die Ursache liegt vermutlich in einem Fehler des Modells, da nach sorgfältiger Prüfung des Codes keine offensichtlichen Fehler gefunden werden konnten. Die Empfehlung zur Implementierung von Whisper lautet daher, die Zielsprache vor der Transkription festzulegen und bei extrem hohen Fehlerquoten einen zweiten Durchlauf zu starten, um die Qualität der Transkriptionen zu verbessern.

Der Modellvergleich zeigt, dass bereits die Voreinstellung der Sprache die Fehlerquote erheblich reduzieren kann. Zudem wird angenommen, dass größere Modelle bessere Ergebnisse liefern können. Allerdings hängt die Arbeitszeit stark von den verfügbaren Ressourcen ab und steigt mit der Größe des Modells. Ein gutes Gleichgewicht zwischen Leistung und Geschwindigkeit könnte das small-Modell bieten, das auch zur Erstellung der Beispieltranskripte verwendet wurde. Für Akzente oder Dialekte, welche für das Modell besonders herausfordernd sind, bietet es sich an, dass Model mit einer kleinen Menge an Trainingsdaten zu fine-tunen (Jain et al., 2023; Yang et al., 2023), was dank der offenen MIT Lizenz[3] und dem öffentlichen Quellcode gut umsetzbar ist. Dies könnte sich auch für wissenschaftliche Podcasts als lohnenswert erweisen, um die Fehlerquote noch weiter zu verringern. Die Verwendung von Metadaten könnte ebenfalls zu einer verbesserten Leistung führen (Jones et al., 2021).

Trotz vergleichsweise guter Transkriptionsqualität ist es im Kontext von Wissenschaftspodcasts empfehlenswert, die automatisch erstellten Transkripte weiterhin manuell zu überprüfen und

[3] https://github.com/openai/whisper/blob/main/LICENSE

Fehler zu korrigieren, um die wissenschaftliche Integrität nicht durch fehlerbehaftete Transkripte zu untergraben. Diese wirken im Einzelfall unprofessionell und könnten im schlimmsten Fall zu Missverständnissen führen oder falsche Informationen verbreiten. Auch die Glaubwürdigkeit des Podcasts könnte in Frage gestellt werden. Eine genaue Repräsentation des Gesagten ist möglicherweise nicht immer wünschenswert (Halcomb und Davidson, 2006). Natürliche Sprache in gesprochener Form beinhaltet mündliche Konventionen wie Füllwörter, Slang-Ausdrücke, Wiederholungen und Verzögerungen, die beim Lesen störend wirken können. Es könnte daher sinnvoll sein, eine paraphrasierte Version des Transkripts als Lesebeilage zur Verfügung zu stellen. Ein weiteres Problem ist die Unterscheidung von Sprechern, das im Abschnitt 2.2.4 erläutert wurde. Whisper selbst ist dazu nicht in der Lage. Es gibt jedoch Lösungen wie „whisper-diarization“[4], die versuchen, dieses Problem anzugehen. Allerdings kämpfen sie noch mit der Herausforderung der sich überlappenden Sprache. Eine Alternative besteht darin, Produkte wie Assembly AI[5] oder die Cloud STT API von Google[6] zu verwenden, um die Transkriptionen zu verbessern. Es ist jedoch zu beachten, dass keine dieser Lösungen eine fehlerfreie Transkription garantieren kann. Dies wird jedoch auch selten von Menschen erzielt (Stolcke und Droppo, 2017; Zayats et al., 2019).

Weiterhin deuten die Ergebnisse darauf hin, dass sich die automatischen Transkripte hervorragend zur Verbesserung der Auffindbarkeit von Podcasts eignen. Die Feststellung im Abschnitt 2.2.3, dass selbst automatische Transkripte mit einer WER von bis zu 50% immer noch einen Mehrwert in Retrieval-Anwendungen bieten, ist äußerst bedeutend. Eine Marke die von den Whisper Transkripte im Durchschnitt weit unterboten wird. Die Einbindung von Transkripten in Metadaten ermöglicht nicht nur inhaltsbasierte Suchen, sondern eröffnet auch eine Vielzahl von Möglichkeiten zur Verbesserung der Auffindbarkeit, zur Steigerung der Informationsdichte und zur Erhöhung der Transparenz von Podcasts. Zunächst ermöglicht die Einbindung von Transkripten den Konsumenten, Podcasts anhand von Stichwörtern oder Suchbegriffen zu finden, selbst wenn sie den Titel oder die Beschreibung der Episode nicht kennen (Besser et al., 2010). Des Weiteren dienen sie als Grundlage für weitere NLP Methoden. Beispielsweise um Referenzen, Entitäten, Schlagwörter und Textkategorisierungen zu extrahieren, Episoden zusammenzufassen oder in Abschnitte zu untergliedern (Jones et al., 2021). Auf diese Weise können Podcasts mit bestimmten Themen, wissenschaftlichen Arbeiten, Experten oder Schlagwörtern verknüpft werden. Die aus den Transkripten extrahierten Informationen, wie Schlagwörter, Entitäten und Zusammenfassungen, können in den Metadaten des Podcasts abgespeichert werden und zur Verbesserung der Repräsentation des Podcasts beitragen. Die Einbindung von Transkripten und die Erweiterung der Metadaten erfordern jedoch die Einführung eines einheitlichen Metadatenformats, das auf dem RSS-Format aufbaut. Dies gewährleistet die Konsistenz und Interoperabilität der Metadaten in der Wissenschaft. Ein solches einheitliches Format würde es ermöglichen, die Vielfalt der Informationen, die in Verbindung mit Podcasts verfügbar sind, effizient zu verwalten und zu nutzen (Jones et al., 2021).

Weiterhin eröffnen die extrahierten Entitäten, Textkategorisierungen und Referenzen die Möglichkeit eine Repräsentation und Verknüpfung von Podcasts und Podcast-Episoden innerhalb

[4] https://github.com/MahmoudAshraf97/whisper-diarization
[5] https://www.assemblyai.com/
[6] https://cloud.google.com/speech-to-text?hl=de

eines Knowledge Graphs (Jones et al., 2021) zu realisieren. Dieser Ansatz hat sich bereits in anderen Bereichen, wie beispielsweise bei der Nutzung des NewsGraphs für Nachrichtenempfehlungen, als äußerst effektiv erwiesen (Liu et al., 2019). Ein Knowledge Graph (Wissensgraph) ist eine Datenstruktur, die entwickelt wurde, um Wissen in einer semantisch vernetzten Form darzustellen. Der Graph besteht aus Knoten, die Entitäten oder Konzepte repräsentieren, und Kanten, die Beziehungen zwischen diesen Entitäten darstellen. Im Wesentlichen handelt es sich um eine Wissensdatenbank, die Informationen auf eine Art und Weise organisiert, die maschinenlesbar ist und semantische Bedeutungen trägt (Hogan et al., 2021). Im Kontext von Podcasts können Knowledge Graphs dazu verwendet werden, Podcasts und deren Inhalte zu organisieren. Jede Podcast-Episode, jeder Podcast selbst und seine Metadaten, sowie Entitäten können als Knoten im Knowledge Graph dargestellt werden. Dies ermöglicht eine hierarchische und vernetzte Struktur, die das Auffinden und Verknüpfen von Podcast-Inhalten erleichtert. Podcast-Episoden können semantisch annotiert werden, indem relevante Entitäten und Konzepte aus den Transkripten extrahiert und in den Knowledge Graph aufgenommen werden. Dadurch würden Podcast-Inhalte in einen größeren semantischen Kontext eingebettet werden. Zum Beispiel könnten Zitate von Experten mit ihren Profilen im Knowledge Graph verknüpft werden. Des Weiteren können personalisierte Empfehlungen basierend auf den Interessen und dem Wissen des Hörers erstellt werden. Der Knowledge Graph kann Informationen über Hörerpräferenzen, Fachgebiete, verwandte Konzepte und Podcast-Historie speichern, um maßgeschneiderte Empfehlungen zu generieren. Darüber hinaus kann der Knowledge Graph Podcast-Inhalte auch mit anderen Wissensquellen verknüpfen, wie wissenschaftlichen Artikeln, Büchern oder Online-Enzyklopädien. Dies fördert die interdisziplinäre Wissensvernetzung und ermöglicht es, Podcasts als Informationsquelle in einem breiteren wissenschaftlichen Kontext zu nutzen. Durch die Verankerung von Entitäten und Informationen aus Podcasts können Hörern den Ursprung und die Validität von Podcast-Inhalten nachverfolgen, was die Integrität von Wissenschaftspodcasts stärkt.

Die in dieser Arbeit verfolgte Hypothese, dass KI-gestützte Transkriptionsverfahren eine effiziente Möglichkeit bieten, qualitativ hochwertige Transkripte zu erstellen, welche weiterhin zur Verbesserung der Auffindbarkeit, Zugänglichkeit und Verbreitung von wissenschaftlichen Inhalten beitragen, lässt sich bestätigen. In den letzten Jahren haben die Entwicklungen im Bereich ASR die Einstiegshürden für die Nutzung von Sprachmodellen drastisch reduziert. Die komplexen Strukturen hochoptimierter Systeme, die einst eine hohe Expertise erforderten, wurden in vielen Bereichen durch Transformern ersetzt, die insgesamt weniger komplex in der Implementierung sind. Grundsätzlich kann nun jeder, mit etwas Einarbeitung, diese Systeme verwenden, um solide Ergebnisse zu produzieren, und falls erforderlich, die Pipeline auf eigenen Daten nachtrainieren. Dennoch stehen die Systeme vor einigen Herausforderungen, die es in der Zukunft zu bewältigen gilt. Dazu gehören der Umgang mit Code-Switching und mehrsprachigen Texten, die Verarbeitung von überlappenden Dialogen, die Sprecherdifferenzierung und das Problem des Long-Tail-Effekts. Diese Herausforderungen erfordern weitere Forschung und Entwicklung, um die Leistung und Anwendbarkeit von ASR-Systemen kontinuierlich zu verbessern und sicherzustellen, dass sie auch in komplexen sprachlichen Situationen effektiv arbeiten können.

5.3 Limitationen und Begrenzungen

Die zentrale Limitation der vorliegenden Arbeit ergibt sich aus den begrenzten Rechenkapazitäten, aufgrund derer die Menge an verfügbaren Daten, die für die Analyse genutzt werden konnte, erheblich eingeschränkt war. Dies führte dazu, dass ein vollständiger Durchlauf des mittleren oder großen Whisper-Modells nicht möglich war, was die Möglichkeit eines belastbaren Vergleichs der Modelle erheblich einschränkte. In diesem Zusammenhang trat die Herausforderung unregelmäßig verteilter Daten auf, die aufgrund der begrenzten Rechenkapazitäten weiter verstärkt wurde. Eine gleichmäßigere Verteilung der Daten wäre wünschenswert gewesen, um vergleichbare Schlussfolgerungen ziehen zu können. Dennoch wurde in dieser Arbeit die Darstellung der Daten in all ihren Facetten der Exklusion unterrepräsentierter Gruppen vorgezogen. Zusätzlich limitierte die begrenzte Anzahl verfügbarer Datenfelder die Aussagekraft der Analyse. Zusätzliche Informationen bezüglich der Audio-Dateien, wie beispielsweise Hinweise auf Störgeräusche, Sprachfehler oder Versprecher, wären wertvoll gewesen und hätten eine tiefgreifendere Analyse ermöglicht. Zukünftige Forschung mit erweiterten Kapazitäten und Datenquellen könnte dazu beitragen, die Einschränkungen zu überwinden und eine umfassendere Studie durchzuführen. Des Weiteren können erste Erkenntnisse aus dieser Studie als Ausgangslage für zukünftige Forschung dienen. In dieser Arbeit wurde die Bedeutung einer transparenten und evidenzbasierten wissenschaftlichen Kommunikation im Zusammenhang mit Wissenschaftspodcasts diskutiert. Die Ergebnisse dieser Untersuchung bieten jedoch keine konkreten Maßstäbe zur Bewertung von Qualität oder Transparenz. Stattdessen konnten lediglich Anregungen und Hypothesen aufgestellt werden, die in zukünftigen Forschungsprojekten näher untersucht werden müssen.

6 Fazit und Ausblick

Zusammenfassend lässt sich feststellen, dass die in dieser Studie untersuchte Anwendbarkeit von KI-gestützten Transkriptionsverfahren auf die Transkription von wissenschaftlichen Podcasts vielversprechende Ergebnisse erzielt hat. Es wurde festgestellt, dass die Ergebnisse je nach Sprache variieren. Englischsprachige Daten erzielten weitaus bessere Ergebnisse als deutschsprachige. Einige Akzente und Dialekte stellten jedoch eine Herausforderung dar, was teilweise auf die fehlende Vorseinstellung der Sprache zurückzuführen ist. Auch gab es im Einzelfall Ausreißer mit extrem hoher WER. Des Weiteren weisen Transkripte trotz teilweise hohen Fehlerquoten in den meisten Fällen eine sehr hohe Ähnlichkeit zur Ground Truth auf. Daher wird vermutet, dass es sich bei den Fehlern meistens um kleinere Rechtschreibfehler oder Wortdreher handelt. Es wurde festgehalten, dass die Implementierung von Whisper die Festlegung der Zielsprache vor der Transkription und bei hohen Fehlerquoten einen zweiten Durchlauf einschließen sollte. Im Modellvergleich zeigte sich, dass mit steigender Modellgröße auch die Transkriptionsqualität zunimmt. Die Qualität der Transkripte ist ausreichend um inhaltsbasierte Suchen zu ermöglichen, da selbst bei einer hohen WER von bis zu 50% noch ein Mehrwert für Retrieval-Anwendungen geboten wird. Überraschenderweise erzielte eine Teilmenge des GigaSpeech Datensatzes, die aus der Kategorie „Science and Technology“ und der Quelle „Podcasts“ stammte, eine durchschnittliche WER von 7,3%. Dies lässt zunächst vermuten, dass Whisper für den skizzierten Anwendungsfall dieser Arbeit bestens geeignet ist und die Eingangshypothese somit bestätigt werden konnte, jedoch ist eine eingehendere Analyse auf umfassenderen Daten erforderlich ist, um diese Ergebnisse zu validieren.

Die Ergebnisse dieser Untersuchung könnten durch eine tiefergehende Analyse, unter Verwendung einer größeren und gleichmäßiger verteilten Datengrundlage, ergänzt durch umfangreichere Informationen über die Merkmale der Audiodaten, verfeinert werden. Dies würde eine breitere Grundlage für die Beurteilung der Leistung von Whisper bieten. Insbesondere eine Analyse von lemmatisierten Texten würde es ermöglichen, den Erhalt der semantischen Bedeutung in den generierten Transkripten auf einer deutlich substanzielleren Ebene zu erforschen. Die Integration zusätzlicher Bewertungsmetriken, wie die Erkennung von Entitäten, könnte einen entscheidenden Fortschritt darstellen, um die Kapazitäten von Whisper in Bezug auf die Verarbeitung spezialisierter Informationen zu evaluieren. Damit würde die Eignung des Systems für die Bewältigung komplexer, fachspezifischer Daten unterstrichen und die Relevanz für eine breite Palette von Anwendungen gesteigert werden. Darüber hinaus könnte eine eingehendere Analyse der POS-Tags bei der Identifizierung und gezielten Bewältigung von Schwachstellen in der Transkriptionsqualität von Whisper hilfreich sein. Dies würde dazu beitragen, die Präzision und Kohärenz der generierten Texte weiter zu steigern, was insbesondere in wissenschaftlichen und akademischen Kontexten von großer Bedeutung ist.

Für zukünftige Forschung bieten sich auf diesem Gebiet spannende Möglichkeiten. Speziell könnten weitere NLP-Methoden wie Entity Recognition, Topic-Modelling und Text-Summarization dazu beitragen, die Zugänglichkeit, Durchsuchbarkeit und Indexierung von Podcast-Inhal-

ten zu verbessern. Gleichzeitig sollte die kontinuierliche Verbesserung von ASR-Technologien im Fokus stehen, um die Genauigkeit von Transkripten weiter zu steigern. Darüber hinaus könnte die Erforschung von Empfehlungssystemen und semantischer Suche in Podcasts den Nutzern helfen, relevante Inhalte leichter zu finden und zu entdecken. Diese Entwicklungen könnten dazu beitragen, den wissenschaftlichen Diskurs zu fördern und die Integration von Podcasts in Forschung und Wissenschaft zu erleichtern. Insgesamt eröffnet die vorliegende Studie interessante Perspektiven für die Zukunft, um die Potenziale von ASR und NLP-Technologien voll auszuschöpfen und die Kommunikation und Nutzung von wissenschaftlichen Podcasts weiter zu verbessern.

Literatur

Aelst, P. V., Toth, F., Castro, L., Štětka, V., de Vreese, C., Aalberg, T., Cardenal, A. S., Corbu, N., Esser, F., Hopmann, D. N., Koc-Michalska, K., Matthes, J., Schemer, C., Sheafer, T., Splendore, S., Stanyer, J., Stępińska, A., Strömbäck, J., & Theocharis, Y. (2021). Does a Crisis Change News Habits? A Comparative Study of the Effects of COVID-19 on News Media Use in 17 European Countries. *Digital Journalism*, *9*(9), 1208–1238. https://doi.org/10.1080/21670811.2021.1943481

Ali, A., & Renals, S. (2018). Word Error Rate Estimation for Speech Recognition: e-WER. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 20–24. https://doi.org/10.18653/v1/P18-2004

Anderson, C. (2012). 10. The Long Tail. In M. Mandiberg (Hrsg.), *The Social Media Reader* (S. 137–152). New York University Press. https://doi.org/doi:10.18574/nyu/9780814763025.003.0014

Apache Parquet. (n. d.). https://parquet.apache.org/ Zuletzt abgerufen am 19. Oktober 2023.

Apple. (n. d.). *A Podcaster's Guide to RSS*. https://help.apple.com/itc/podcasts_connect/#/itcb54353390 Zuletzt abgerufen am 19. Oktober 2023.

Apple. (2005). *Apple Takes Podcasting Mainstream*. https://www.apple.com/newsroom/2005/06/28Apple-Takes-Podcasting-Mainstream/ Zuletzt abgerufen am 19. Oktober 2023.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 4211–4215.

Austria, J. L. (2007). Developing evaluation criteria for podcasts.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer Normalization.

Baig, F., Beg, S., & Khan, M. F. (2012). Controlling Home Appliances Remotely through Voice Command. *International Journal of Computer Applications*, *48*(17), 1–4. https://doi.org/10.5120/7437-0133

Balreira, D. G., Silveira, T. L. T. d., & Wickboldt, J. A. (2023). Investigating the impact of adopting Python and C languages for introductory engineering programming courses. *Computer Applications in Engineering Education*, *31*(1), 47–62. https://doi.org/https://doi.org/10.1002/cae.22570

Berry, R. (2015). Serial and Ten years of Podcasting: Has The Medium Finally Grown Up.

Berry, R. (2016). Part of the establishment: Reflecting on 10 years of podcasting as an audio medium. *Convergence*, *22*(6), 661–671. https://doi.org/10.1177/1354856516632105

Besser, J., Larson, M., & Hofmann, K. (2010). Podcast search: user goals and retrieval technologies. *Online Information Review*, *34*(3), 395–419. https://doi.org/10.1108/14684521011054053

Bhable, S. G., Deshmukh, R. R., & Kayte, C. N. (2023). Comparative Analysis of Automatic Speech Recognition Techniques. *Proceedings of the International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022)*, 897–904. https://doi.org/10.2991/978-94-6463-136-4_79

Birch, H., & Weitkamp, E. (2010). Podologues: conversations created by science podcasts. *New Media & Society*, *12*(6), 889–909. https://doi.org/10.1177/1461444809356333

Bloch, J. (2006). How to Design a Good API and Why It Matters. *Companion to the 21st ACM SIGPLAN Symposium on Object-Oriented Programming Systems, Languages, and Applications*, 506–507. https://doi.org/10.1145/1176617.1176622

Budzinski, O., Gaenssle, S., & Lindstädt-Dreusicke, N. (2021). The battle of YouTube, TV and Netflix: an empirical analysis of competition in audiovisual media markets. *SN Business & Economics*, *1*, 116. https://doi.org/10.1007/s43546-021-00122-0

Celma, Ò., & Raimond, Y. (2008). ZemPod: A semantic web approach to podcasting [Semantic Multimedia]. *Journal of Web Semantics*, *6*(2), 162–169. https://doi.org/https://doi.org/10.1016/j.websem.2008.01.003

Chelba, C., Hazen, T. J., & Saraclar, M. (2008). Retrieval and browsing of spoken content. *IEEE Signal Processing Magazine*, *25*(3), 39–49. https://doi.org/10.1109/MSP.2008.917992

Chen, G., Chai, S., Wang, G., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., Jin, M., Khudanpur, S., Watanabe, S., Zhao, S., Zou, W., Li, X., Yao, X., Wang, Y., Wang, Y., … Yan, Z. (2021). GigaSpeech: An Evolving, Multi-domain ASR Corpus with 10,000 Hours of Transcribed Audio.

Chiarelli, A., Johnson, R., Pinfield, S., & Richens, E. (2019, September). Accelerating scholarly communication: The transformative role of preprints. https://doi.org/10.5281/zenodo.3357727

Clifton, A., Reddy, S., Yu, Y., Pappu, A., Rezapour, R., Bonab, H., Eskevich, M., Jones, G., Karlgren, J., Carterette, B., & Jones, R. (2020). 100,000 Podcasts: A Spoken English Document Corpus. *Proceedings of the 28th International Conference on Computational Linguistics*, 5903–5917. https://doi.org/10.18653/v1/2020.coling-main.519

CodeEmporium. (2020). *Transformer Neural Networks - EXPLAINED! (Attention is all you need)* [Video]. https://www.youtube.com/watch?v=TQQlZhbC5ps Zuletzt abgerufen am 19. Oktober 2023.

Cohen, W. W., Ravikumar, P., Fienberg, S. E., et al. (2003). A Comparison of String Distance Metrics for Name-Matching Tasks. *IIWeb*, *3*, 73–78.

Datta, P., Jakubowicz, P., Vogler, C., & Kushalnagar, R. (2020). Readability of Punctuation in Automatic Subtitles. In K. Miesenberger, R. Manduchi, M. Covarrubias Rodriguez & P. Peňáz (Hrsg.), *Computers Helping People with Special Needs* (S. 195–201). Springer International Publishing.

Dernbach, B. (2022). Hineinhören in die wunderbare Welt der Wissenschaft. Podcasts als Medium der Wissenschaftskommunikation. In V. Katzenberger, J. Keil & M. Wild (Hrsg.), *Podcasts: Perspektiven und Potenziale eines digitalen Mediums* (S. 307–332). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-38712-9_12

Dhiman, B. (2022). Condition of Women Prisoners in Model Jail, Chandigarh: A Communication Study. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4205096

Dhiman, D. B. (2023). Key Issues and New Challenges in New Media Technology in 2023: A Critical Review. *Journal of Media & Management*, *5*(1), 1–4.

Donnelly, K., & Berge, Z. (2006). Podcasting: Co-Opting MP3 Players for Education and Training Purposes. *Online Journal of Distance Learning Administration*, *9*(3). https://www.learntechlib.org/p/193222/

Dziatzko, J. (2023). *Podcast Hosting Vergleich: Die 7 Besten im ultimativen Test 2023*. https://solobusinesstribe.de/podcast-hosting/ Zuletzt abgerufen am 19. Oktober 2023.

Edison Research. (2019). *The Podcast Consumer 2019*. http://www.edisonresearch.com/wp-content/uploads/2019/04/Edison-Research-Podcast-Consumer-2019.pdf Zuletzt abgerufen am 19. Oktober 2023.

Edison Research. (2023). *The Podcast Consumer 2023*. http://www.edisonresearch.com/wp-content/uploads/2023/03/The-Podcast-Consumer-2023-1.pdf Zuletzt abgerufen am 19. Oktober 2023.

Fähnrich, B., Weitkamp, E., & Kupper, J. F. (2023). Exploring 'quality' in science communication online: Expert thoughts on how to assess and promote science communication quality in digital media contexts. *Public Understanding of Science*, *32*(5), 605–621. https://doi.org/10.1177/09636625221148054

Fetic, R. A., Jordahn, M., Lima, L. C., Egebæk, R. A. F., Nielsen, M. C., Biering, B., & Hansen, L. K. (2021). Topic Model Robustness to Automatic Speech Recognition Errors in Podcast Transcripts. *CoRR*, *abs/2109.12306*. https://arxiv.org/abs/2109.12306

Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálfy, M., Nanni, F., & Coates, J. A. (2021). The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLoS biology*, *19*(4), e3000959. https://doi.org/10.1371/journal.pbio.3000959

Fromm, M. (2013). *OSR000 Open Science Radio*. https://www.opensciencerадio.org/2013/01/02/osr000/

Grand View Research. (n. d.). *Podcasting Market Size, Share & Trends Analysis Report by genre (News & Politics, Society & Culture, Comedy, Sports), by format (Interviews, panels, solo), by region, and segment Forecasts, 2023 - 2030*. https://www.grandviewresearch.com/industry-analysis/podcast-market Zuletzt abgerufen am 19. Oktober 2023.

Gundogdu, A. S., Sanghvi, A., & Harrigian, K. (2018). Recognizing Film Entities in Podcasts.

Halcomb, E. J., & Davidson, P. M. (2006). Is verbatim transcription of interview data always necessary? *Applied Nursing Research*, *19*(1), 38–42. https://doi.org/https://doi.org/10.1016/j.apnr.2005.06.001

Hameleers, M., Brosius, A., & de Vreese, C. H. (2022). Whom to trust? Media exposure patterns of citizens with perceptions of misinformation and disinformation related to the news media. *European Journal of Communication*, *37*(3), 237–268. https://doi.org/10.1177/02673231211072667

Hammersley, B. (2004). Audible Revolution. *The Guardian*. https://www.theguardian.com/media/2004/feb/12/broadcasting.digitalmedia Zuletzt abgerufen am 19. Oktober 2023.

Hellermann, M. (2015). *Wissenschaft in Film und Fernsehen: die mediale Morphologie audiovisueller Wissenschaftskommunikation*. LIT Verlag.

Hendrycks, D., & Gimpel, K. (2023). Gaussian Error Linear Units (GELUs).

Henze, N. (2021). *Das Simpson-Paradoxon*. https://doi.org/10.5445/IR/1000132557

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge Graphs. *ACM Comput. Surv.*, *54*(4). https://doi.org/10.1145/3447772

Huang, W. R., Zhang, H., Kumar, S., Chang, S.-y., & Sainath, T. N. (2023). Semantic Segmentation with Bidirectional Language Models Improves Long-form ASR.

Jain, R., Barcovschi, A., Yiwere, M., Corcoran, P., & Cucu, H. (2023). Adaptation of Whisper models to child speech recognition.

Jaro, M. A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, *84*(406), 414–420. https://doi.org/10.1080/01621459.1989.10478785

Jones, R., Zamani, H., Schedl, M., Chen, C.-W., Reddy, S., Clifton, A., Karlgren, J., Hashemi, H., Pappu, A., Nazari, Z., Yang, L., Semerci, O., Bouchard, H., & Carterette, B. (2021). Current Challenges and Future Directions in Podcast Information Access. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1554–1565. https://doi.org/10.1145/3404835.3462805

Jowitt, A. (2007). Perceptions and usage of library instructional podcasts by staff and students at New Zealand's Universal College of Learning (UCOL). *Reference Services Review*, *36*. https://doi.org/10.1108/00907320810895396

Karpagavalli, S., & Chandra, E. (2016). A review on automatic speech recognition architecture and approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, *9*(4), 393–404.

Katzenberger, V., Keil, J., & Wild, M. (2022). Mehr als die Summe seiner Teile: Entwicklung, Forschungsstand und Definition von Podcasts. In V. Katzenberger, J. Keil & M. Wild (Hrsg.), *Podcasts: Perspektiven und Potenziale eines digitalen Mediums* (S. 1–19). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-38712-9_1

Khan, T. M., Alhussein, M., Aurangzeb, K., Arsalan, M., Naqvi, S. S., & Nawaz, S. J. (2020). Residual Connection-Based Encoder Decoder Network (RCED-Net) for Retinal Vessel Segmentation. *IEEE Access*, *8*, 131257–131272. https://doi.org/10.1109/ACCESS.2020.3008899

Kutuzov, A. (2013). Improving English-Russian sentence alignment through POS tagging and Damerau-Levenshtein distance. *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, 63–68.

Leander, L. (2020). Wissenschaft im Gespräch: Wissensvermittlung und -aushandlung in Podcasts. *kommunikation @ gesellschaft*, *21*(2), 24. https://doi.org/https://doi.org/10.15460/kommges.2020.21.2.621

Li, J. (2022). Recent Advances in End-to-End Automatic Speech Recognition.

LibriSpeech Test-Clean Benchmark (Speech Recognition). (n. d.). https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-clean Zuletzt abgerufen am 19. Oktober 2023.

Lin, M., Chen, Q., & Yan, S. (2014). Network In Network.

Listen Notes Inc. (n. d.). *Podcast stats: How many podcasts are there?* https://www.listennotes.com/podcast-stats/ Zuletzt abgerufen am 19. Oktober 2023.

Liu, D., Bai, T., Lian, J., Zhao, X., Sun, G., Wen, J.-R., & Xie, X. (2019). News Graph: An Enhanced Knowledge Graph for News Recommendation. *KaRS@CIKM*, 1–7.

MacKenzie, L. E. (2019). Science podcasts: analysis of global production and output from 2004 to 2018. *R. Soc. open sci.*, *6*, 180932–180932. https://doi.org/10.1098/rsos.180932

Madsen, V. (2009). Voices-cast: a report on the new audiosphere of podcasting with specific insights for public broadcasting. *ANZCA09 Conference Proceedings*, 1191–1210.

Mauran, C. (2023, September). *Spotify now transcribes podcasts so you can read along. Here's how it works*. https://mashable.com/article/spotify-transcription-podcasts-new-feature Zuletzt abgerufen am 19. Oktober 2023.

Miles, A. (2009). RIP RSS: Reviving Innovative Programs through Really Savvy Services. *Journal of Hospital Librarianship*, *9*(4), 425–432. https://doi.org/10.1080/15323260903253753

Mozilla. (n. d.). *Mozilla Common Voice*. https://commonvoice.mozilla.org/de/about Zuletzt abgerufen am 19. Oktober 2023.

Myers-Scotton, C. (2017). Code-Switching. In *The Handbook of Sociolinguistics* (S. 217–237). John Wiley Sons, Ltd. https://doi.org/https://doi.org/10.1002/9781405166256.ch13

Newman, N., Fletcher, R., Eddy, K., Robertson, C. T., & Kleis Nielsen, R. (2023). *Reuters Digital News Report 2023*. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf

O'Connor, R. (2023). How to evaluate speech recognition Models. https://www.assemblyai.com/blog/how-to-evaluate-speech-recognition-models/ Zuletzt abgerufen am 19. Oktober 2023.

Ojha, V. K., Abraham, A., & Snášel, V. (2017). Metaheuristic design of feedforward neural networks: A review of two decades of research. *Engineering Applications of Artificial Intelligence*, *60*, 97–116. https://doi.org/https://doi.org/10.1016/j.engappai.2017.01.013

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

Petersen, A. M., Jung, W.-S., Yang, J.-S., & Stanley, H. E. (2011). Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proceedings of the National Academy of Sciences*, *108*(1), 18–23. https://doi.org/10.1073/pnas.1016733108

Podcast 'Open Science Radio'. (n. d.). https://wissenschaftspodcasts.de/podcasts/open-science-radio/ Zuletzt abgerufen am 19. Oktober 2023.

Putri, T. A. (2019). An analysis of types and causes of translation errors. *Etnolingual Journal*, *3*(2), 93–103.

Quintana, D. S., & Heathers, J. (2020). How podcasts can benefit scientific communities. https://doi.org/10.1016/j.tics.2020.10.003

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via Large-Scale Weak Supervision. *arXiv preprint arXiv:2212.04356*. https://doi.org/10.48550/arxiv.2212.04356

Ramirez, J., Gorriz, J. M., & Segura, J. C. (2007). Voice Activity Detection. Fundamentals and Speech Recognition System Robustness. In M. Grimm & K. Kroschel (Hrsg.), *Robust Speech*. IntechOpen. https://doi.org/10.5772/4740

Real Python. (2023). *Dictionaries in Python*. https://realpython.com/python-dicts/ Zuletzt abgerufen am 19. Oktober 2023.

Rime, J., Pike, C., & Collins, T. (2022). What is a podcast? Considering innovations in podcasting through the six-tensions framework. *Convergence*, *28*(5), 1260–1282. https://doi.org/10.1177/13548565221104444

Schneider-Stickler, B., & Bigenzahn, W. (2013). Kommunikation in der modernen Gesellschaft. In *Stimmdiagnostik: Ein Leitfaden für die Praxis* (S. 1–8). Springer Vienna. https://doi.org/10.1007/978-3-7091-1480-3_1

Schöne, A. (2021a). Mit Audiotranskription die Reichweite von Podcasts steigern. *Media Lab Bayern*. https://www.media-lab.de/de/blog/artikel/mit-audiotranskription-die-reichweite-von-podcasts-steigern Zuletzt abgerufen am 19. Oktober 2023.

Schöne, A. (2021b). *Mit Audiotranskription die Reichweite von Podcasts steigern*. https://www.media-lab.de/de/blog/artikel/mit-audiotranskription-die-reichweite-von-podcasts-steigern Zuletzt abgerufen am 19. Oktober 2023.

Singh, A., Mehta, A. S., S, A. K. K., G, D., Date, G., Nanavati, J., Bandekar, J., Basumatary, K., P, K., Badiger, S., Udupa, S., Kumar, S., Savitha, Ghosh, P. K., V, P., Pai, P., Nanavati, R., Saxena, R., Mora, S. P. R., & Raghavan, S. (2023). Model Adaptation for ASR in low-resource Indian Languages.

Speech & in the Northeast (SANE), A. (2023, März). *SANE2022 | Tara Sainath - End-to-End Speech Recognition: The Journey From Research to Production* [Video]. https://www.youtube.com/watch?v=FvkLYRpBIe0 Zuletzt abgerufen am 19. Oktober 2023.

Spotify. (2023, September). *Spotify's AI voice translation pilot means your favorite podcasters might be heard in your native language*. https://newsroom.spotify.com/2023-09-25/ai-voice-translation-pilot-lex-fridman-dax-shepard-steven-bartlett/ Zuletzt abgerufen am 19. Oktober 2023.

StatQuest with Josh Starmer. (2023, Juli). *Transformer Neural Networks, ChatGPT's foundation, clearly explained!!!* [Video]. https://www.youtube.com/watch?v=zxQyTK8quyY Zuletzt abgerufen am 19. Oktober 2023.

Stolcke, A., & Droppo, J. (2017). Comparing Human and Machine Errors in Conversational Speech Transcription. *Interspeech 2017*. https://doi.org/10.21437/interspeech.2017-1544

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks.

Van Bavel, J., Boggio, P., Capraro, V., Cichocka, A., Cikara, M., Crockett, M., Crum, A., Douglas, K., Druckman, J., Drury, J., Ellemers, N., Finkel, E., Gelfand, M., Han, S., Haslam, S., Jetten, J., Kitayama, S., mobbs dean, d., Napper, L., & Willer, R. (2020, März). *Using social and behavioural science to support COVID-19 pandemic response*. https://doi.org/10.31234/osf.io/y38m9

Vartakavi, A., Garg, A., & Rafii, Z. (2021). Audio Summarization for Podcasts. *2021 29th European Signal Processing Conference (EUSIPCO)*, 431–435. https://doi.org/10.23919/EUSIPCO54536.2021.9615948

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need.

Velardo, V. (2020). *MEL Spectrograms Explained Easily* [Video]. https://www.youtube.com/watch?v=9GHCiiDLHQ4 Zuletzt abgerufen am 19. Oktober 2023.

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, *88*, 428–436. https://doi.org/https://doi.org/10.1016/j.jbusres.2017.12.043

Wamba, S. F., Gumbo, S., Twinomurinzi, H., Bwalya, K., & Mpinganjira, M. (2023). Digital transformation under COVID-19: A Bibliometric Study and future research agenda. *Procedia Computer Science*, *219*, 271–278.

Wang, Y.-Y., Acero, A., & Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, 577–582. https://doi.org/10.1109/ASRU.2003.1318504

Weingart, P., & Schulz, P. (2014). Einleitung: Das schwierige Verhältnis zwischen Wissenschaft, Öffentlichkeit und Medien. In *Wissen, Nachricht, Sensation*.

Wibowo, A. T. (2022). Hoax and fake news by saracen syndicate and the problems for national cyber security. *Indonesian Journal of Counter Terrorism and National Security*, *1*(1), 91–108.

Winkler, W. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods*.

Wu, Q., Wang, P., Wang, X., He, X., & Zhu, W. (2022). Deep Learning Basics. In *Visual Question Answering: From Theory to Application* (S. 15–26). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-0964-1_2

Wu, Y. (2018). Social media engagement in the digital age: Accountability or threats. *Newspaper Research Journal*, *39*. https://doi.org/10.1177/0739532918796236

Xiong, X. (2023). Fundamentals of speech recognition.

Yang, H., Zhang, M., Tao, S., Ma, M., & Qin, Y. (2023). Chinese ASR and NER Improvement Based on Whisper Fine-Tuning. *2023 25th International Conference on Advanced Communication Technology (ICACT)*, 213–217. https://doi.org/10.23919/ICACT56868.2023.10079686

Zarocostas, J. (2020). How to fight an infodemic. *The Lancet*, *395*(10225), 676. https://doi.org/https://doi.org/10.1016/S0140-6736(20)30461-X

Zayats, V., Tran, T., Wright, R., Mansfield, C., & Ostendorf, M. (2019). Disfluencies and Human Speech Transcription Errors.

Zhan, Z., Zhao, J., Zhang, Y., Gong, J., Wang, Q., Shen, Q., & Zhang, L. (2021). Grabbing the Long Tail: A data normalization method for diverse and informative dialogue generation. *Neurocomputing*, *460*, 374–384. https://doi.org/https://doi.org/10.1016/j.neucom.2021.07.039

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into Deep Learning. *arXiv preprint arXiv:2106.11342*.

Zweig, G., Siohan, O., Saon, G., Ramabhadran, B., Povey, D., Mangu, L., & Kingsbury, B. (2006). Automated Quality Monitoring in the Call Center with ASR and Maximum Entropy. *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, *1*, I–I. https://doi.org/10.1109/ICASSP.2006.1660089