HOCHSCHULE DER MEDIEN

Bachelorarbeit

im Studiengang Audiovisuelle Medien

Sprachsynthese durch Künstliche Intelligenz als Alternative für das nachträgliche Ersetzen von Dialogen (ADR):

Eine Untersuchung basierend auf einer Wahrnehmungsstudie

vorgelegt von

Torben Barth

Matrikel-Nr. 40994

an der Hochschule der Medien Stuttgart am 14.11.2024

zur Erlangung des akademischen Grades eines Bachelor of Engineering

Erstprüferin: Prof. Dr. Heike Adel-Vu

Zweitprüfer: Prof. Oliver Curdt

Ehrenwörtliche Erklärung

Hiermit versichere ich, Torben Barth, ehrenwörtlich, dass ich die vorliegende Bachelorarbeit mit dem Titel: „Sprachsynthese durch Künstliche Intelligenz als Alternative für das nachträgliche Ersetzen von Dialogen (ADR): Eine Untersuchung basierend auf einer Wahrnehmungsstudie“ selbstständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Ebenso sind alle Stellen, die mit Hilfe eines KI-basierten Schreibwerkzeugs erstellt oder überarbeitet wurden, kenntlich gemacht. Die Arbeit ist noch nicht veröffentlicht oder in anderer Form als Prüfungsleistung vorgelegt worden.

Ich habe die Bedeutung der ehrenwörtlichen Versicherung und die prüfungsrechtlichen Folgen (§ 24 Abs. 2 Bachelor-SPO, § 23 Abs. 2 Master-SPO (Vollzeit)) einer unrichtigen oder unvollständigen ehrenwörtlichen Versicherung zur Kenntnis genommen.

Stuttgart, den 13.11.2024

Ort, Datum

Unterschrift

Kurzfassung

Ziel dieser Arbeit ist es, zu untersuchen, inwiefern ein KI-TTS-System das klassische ADR-Verfahren während einer Filmtonpostproduktion ersetzen kann. Nach einer Analyse der Grundlagen des ADRs und der künstlichen Sprachsynthese werden aktuelle Entwicklungen in beiden Bereichen näher beleuchtet. Im Rahmen einer Wahrnehmungsstudie in Form einer Umfrage wurde die Leistung zweier KI-TTS-Systeme hinsichtlich ihrer Fähigkeiten ähnliche, natürliche und emotionale Sprache zu generieren, evaluiert. Die Ergebnisse zeigen, dass KI-generierte Sprache derzeit den Anforderungen eines Ersatzes des ADR-Verfahrens nicht gerecht werden kann und insbesondere bei deutschsprachigen Modellen weiterer Entwicklungsbedarf besteht.

Abstract

The goal of this study is to explore whether a TTS system based on artificial intelligence can replace the traditional ADR process during film sound postproduction. After reviewing the basics of ADR and artificial speech synthesis, current developments in both areas are examined in more detail. In a perception study conducted through a survey, the performance of two AI-TTS systems was evaluated based on their ability to produce similar, natural, and emotional speech. The results show that AI-generated speech currently does not meet the standards required to replace the ADR process, with further improvements needed, especially for German-language models.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

ADR ..... *Automated Dialogue Replacement*
DAW..... *Digital Audio Workstation*
KI ..... *Künstliche Intelligenz*
MFCC..... *Mel Frequency Cepstral Coefficient*
MOS..... *Mean Opinion Score*
SMOS..... *Similarity Mean Opinion Score*
TTS..... *Text-To-Speech*

1. Einleitung

Von der Idee bis zum finalen Produkt gibt es im Prozess einer Filmproduktion einige wichtige, zeitintensive und auch kostenintensive Schritte. Neben dem Dreh selbst ist die Postproduktion einer solchen Produktion meistens die Phase, für die am meisten Zeit eingeplant werden muss (Roberts & Backstage, 2023). Während der Postproduktion werden sowohl die Bild- als auch die Tonaufnahmen bearbeitet, abgerundet und aufgewertet. Um Zeit und Kosten bei diesen wichtigen Schritten zu sparen, gibt es immer wieder neue Innovationen und neu entwickelte Tools, durch die ein effizienteres Arbeiten ermöglicht werden soll.

Eine dieser Innovationen, die in den letzten Jahren an Bedeutung gewonnen hat, ist der Einsatz von Künstlicher Intelligenz (KI). Eine Prognose des Marktvolumens für KI von „Next Move Strategy Consulting“ (vgl. Abbildung 1) zeigt wie stark das Interesse an Künstlicher Intelligenz ist und wie stark die Investitionen in diesem Bereich in den nächsten Jahren ansteigen werden, sodass auch in Zukunft mit weiteren Entwicklungen und Innovationen gerechnet werden kann.

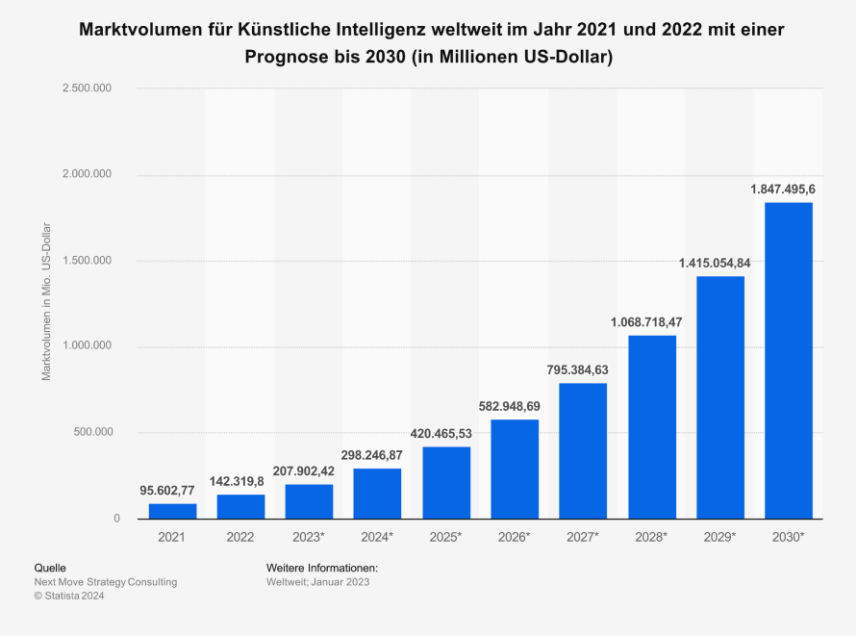

Abbildung 1: Marktvolumen für Künstliche Intelligenz weltweit im Jahr 2021 und 2022 mit einer Prognose bis 2030 (Schreibfehler korrigiert) (Next Move Strategy Consulting, zitiert nach de.statista.com, 2023)

Doch schon bereits heute kommen KI-gestützte Tools in verschiedenen Phasen der Postproduktion zum Einsatz. So kann eine KI etwa die Rohdaten eines Films nach visuellen Inhalten sortieren, sodass weniger Zeit in das Dateimanagement investiert werden muss (Shade Inc., 2024), oder aber bei der Bearbeitung der Bildaufnahmen selbst unterstützen. Im Bereich des Color Gradings, beim Entfernen von unerwünschten Objekten, oder auch bei der Erstellung von visuellen Effekten gibt es bereits KI-basierte Systeme, die vollautomatisch Aufgaben übernehmen können oder zumindest unterstützend zum Workflow und einer effizienteren Arbeitsweise beitragen können (Blackmagic Design, 2024).

Neben der steigenden Anwendung im Bildbereich, kommen KI-Tools auch immer mehr bei der Tonpostproduktion zum Einsatz. Neben gängigen Anwendungsfällen, wie Rauschunterdrückung, das Entfernen von Störgeräuschen oder das Isolieren von Stimmen (Waves Audio Ltd., 2022), gibt es mittlerweile auch weitere Entwicklungen. So ist es z.B. möglich das sog. „Panning“, also das Positionieren der Tonspuren im Stereobild, von einer KI durchführen zu lassen. Die KI analysiert hierbei die sich bewegenden Personen und Objekte im Bild und positioniert deren Tonspur passend dazu im Stereobild (Blackmagic Design, 2024). Diese zeitintensive Arbeit, die zuvor durch ein*e Toningenieur*in erledigt werden musste, kann nun also vollständig von einer KI übernommen werden, wodurch der Fokus in der Tonpostproduktion auf die kreativen und wichtigeren Dinge gesetzt werden kann.

Ein Bereich der Tonpostproduktion, der bislang jedoch nur wenig Innovation erfahren hat, ist der Prozess des „Automated Dialogue Replacements“ (ADR), also dem automatischen Ersetzen des Dialogs. Während dieses Schrittes werden Dialogausschnitte, die aufgrund von Störgeräuschen oder minderwertiger Qualität nicht benutzt werden können, von den entsprechenden Schauspieler*innen erneut in einem Tonstudio eingesprochen (Kizer, 2024, S. 1). Dieser Prozess ist sehr zeit- und kostenintensiv und bedarf ein hohes Maß an Vorbereitung und Organisation (Kizer, 2024, S. 105). Mit zunehmender Entwicklung und Verbesserung von KI-basierten Text-To-Speech (TTS) - Modellen, stellt sich die Frage, ob, und mit welcher Qualität, ein solches Modell diesen Prozess ersetzen kann.

Die vorliegende Arbeit soll untersuchen, mit welcher Qualität die Stimmen zweier Schauspieler*innen geklont und auf den Filmausschnitt angepasst werden können, und

was für Auswirkungen die durch eine KI generierte Stimme auf die Wahrnehmung von Natürlichkeit und Emotionen hat.

1.1. Problemstellung

Um zu verdeutlichen, was für Vorteile sich durch den Ersatz des ADR-Verfahrens mit einem KI-Sprachmodell ergeben, muss zunächst geklärt werden, welche Probleme und Herausforderungen beim Postproduktionsschritt des ADRs auftreten und inwiefern eine KI-basierte Lösung dabei zu Verbesserungen führen kann.

Grundsätzlich sind Aufnahmen im Rahmen von ADRs immer mit zusätzlichem Aufwand, sowie zusätzlichen Kosten verbunden. Im Prozess sind zahlreiche Akteur*innen involviert, von den Schauspieler*innen selbst, über Tontechniker*innen, bis hin sogar zu den Regisseur*innen (Kizer, 2024, S. 105). Die Arbeit, die durch ADR zusätzlich anfällt, beläuft sich dabei jedoch nicht nur auf die Aufnahmen selbst. Es ist schon im Vorfeld nötig viele Vorbereitungen zu treffen, da die nachträglichen Aufnahmen heutzutage in den meisten Fällen immer noch vor Ort in einem Tonstudio durchgeführt werden (Kizer, 2024, S. 105–110). Dadurch ergeben sich weitere logistische und auch finanzielle Herausforderungen, die zum einen nicht vernachlässigt werden dürfen, und zum anderen eine rechtzeitliche Planung erfordern, damit es letztendlich zu keiner Qualitätseinbuße im Bezug auf das fertige Endprodukt kommt. Neben der Organisation im Vorfeld, ist auch eine zeitintensive Nachbearbeitung der Aufnahmen nötig. Die Aufnahmen müssen nicht nur qualitativ bearbeitet werden (z.B. das Entfernen jeglicher Störgeräusche), sondern müssen auch von einer tontechnisch erfahrenen Person klanglich an die Originaltonaufnahmen der Schauspieler*innen und deren Umgebung in den jeweiligen Filmszenen angepasst werden, um den Zuschauer*innen damit ein qualitativ möglichst hochwertiges und immersives Erlebnis zu liefern (Kizer, 2024, S. 199).

Die Nutzung eines KI-basierten Sprachmodells beim ADR-Verfahren, um die Stimmen der Schauspieler*innen zu klonen, könnte einige dieser zeitintensiven Arbeitsschritte eliminieren, und dadurch mehr Zeit für andere wichtige Postproduktionsschritte schaffen, was letztlich zu einer Qualitätssteigerung des Endprodukts führen würde. Vor allem, was die Aspekte der Vorbereitung und Organisation betrifft, wäre eine große Zeiteinsparung durch die Nutzung eines KI-Systems möglich. Da die benötigten Dialogausschnitte von einer KI generiert werden würden, gäbe es für die jeweiligen

Schauspieler*innen keine Notwendigkeit mehr, sich extra in ein Tonstudio begeben zu müssen, was zusätzlich auch noch die Abhängigkeit von deren Verfügbarkeit minimiert und somit auch die zeitliche Flexibilität des Postproduktionsprozesses erhöhen würde. Ein weiterer Vorteil, den die Nutzung einer KI mit sich bringen würde, wäre, dass an der Erstellung und Nachbereitung der Aufnahmen weniger Arbeitskräfte beteiligt sein müssten. Dies würde nicht nur die Koordination erheblich vereinfachen, sondern könnte zusätzlich auch noch Kosten sparen, wodurch vom Budget einer Produktion letztendlich mehr Geld in andere wichtige Produktionsschritte investiert werden kann.

Das Ersetzen des klassischen ADR-Prozesses mit einem KI-System und das damit verbundene Klonen der Stimmen könnte also einige Vorteile im Hinblick auf zeitliche Flexibilität, finanzielle Mittel und Organisationsaufwand mit sich bringen. In dieser Arbeit soll daher evaluiert werden, ob und mit welcher Qualität ein KI-TTS-Modell dies bewerkstelligen kann. Um dies durchzuführen, wurden deshalb drei zentrale Anforderungen an die Fähigkeiten definiert, die ein KI-TTS-Modell haben sollte, um den ADR-Prozess ersetzen zu können. Diese lauten wie folgt:

1) Das Modell sollte dazu in der Lage sein, menschliche Stimmen anhand eines kurzen Ausschnitts zu klonen, wobei die geklonten Ausgaben eine sehr hohe Ähnlichkeit mit den originalen Stimmen aufweisen sollten.
2) Das Modell sollte dazu in der Lage sein, natürlich klingende Ausgaben in Bezug auf Prosodie und Audioqualität zu erzeugen.
3) Das Modell sollte dazu in der Lage sein, emotionale Sprache kontrolliert zu erzeugen.

1.2. Zielsetzung und Fragestellung

Ziel der vorliegenden Arbeit ist es, zu untersuchen und zu bewerten, ob KI-Sprachmodelle das klassische ADR-Verfahren während der Postproduktion ersetzen, bzw. in welchem Ausmaß und mit welcher Qualität ersetzen können. Dabei soll zum einen die generelle Umsetzbarkeit dieses Vorhabens mit einem KI-basierten Sprachmodell, das lokal auf dem Computer arbeitet, untersucht werden und zum anderen soll mithilfe einer Studie in Form einer Umfrage evaluiert werden, ob ein von einer KI generierter Dialog sowohl den audiotechnischen als auch den künstlerischen Anforderungen gerecht werden kann.

Die zentrale Forschungsfrage lautet dabei:

Inwiefern kann künstliche Sprachsynthese durch ein KI-TTS-Modell die menschliche Stimme während des ADR-Prozesses einer Filmtonpostproduktion ersetzen?

Diese Frage bildet den Ausgangspunkt für die Untersuchung und wird für die Umsetzung der Wahrnehmungsstudie nochmals durch weitere spezifischere Forschungsfragen ergänzt, welche wie folgt lauten:

1) Wie hoch wird die Ähnlichkeit einer KI-generierten geklonten Stimme im Vergleich zu der menschlichen Originalstimme wahrgenommen?
2) In welchem Ausmaß kann eine KI-generierte geklonte Stimme in Kombination mit einer visuellen Komponente die audiotechnischen und sprachlichen Anforderungen hinsichtlich der wahrgenommenen Natürlichkeit erfüllen?
3) Wie verlässlich werden Emotionen eines KI-generierten Dialogausschnitts im Vergleich zu der menschlichen Originalstimme von Testhörer*innen erkannt?

1.3. Aufbau der Arbeit

Im folgenden Kapitel wird der strukturelle Aufbau der vorliegenden Arbeit vorgestellt, um so einen Überblick über die Vorgehensweise und die Themen, die zum Beantworten der Forschungsfrage nötig sind, zu geben.

Zunächst werden in Kapitel 2 die Grundlagen des ADR-Verfahrens in der Filmpostproduktion näher erläutert und welche Arten von KI-basierten Tools in der Tonpostproduktion bereits Anwendung finden. Die Grundlagen schließen mit einem Überblick über die Funktionsweise eines KI-basierten Sprachmodells und der damit verbundenen Sprachsynthese ab. Anschließend präsentiert Kapitel 3 den aktuellen Stand der Forschung, wobei insbesondere auf aktuelle Themen im Bereich der Sprachsynthese und des ADRs eingegangen wird und die Fortschritte der letzten Jahre erläutert werden. Vor der eigentlichen Durchführung der Studie wird zunächst in Kapitel 4 noch auf die Methodik und die Evaluationskriterien der Umfrage eingegangen. Hierbei wird zudem die genaue Vorgehensweise bei der Sprachsynthese mit den gewählten KI-Systemen und bei der anschließenden klanglichen Nachbearbeitung erläutert. Im weiteren Verlauf dieser Arbeit (Kapitel 5) wird dann die wissenschaftliche Studie präsentiert. Dabei werden zuerst die benötigten Hypothesen aufgestellt, die Struktur der Studie genauer erläutert und

abschließend die Ergebnisse durch statistische Tests wissenschaftlich ausgewertet. Zum Ende dieser Arbeit, in Kapitel 6, erfolgt die Interpretation der ausgewerteten Ergebnisse im Kontext der Hypothesen und die Beantwortung der Forschungsfrage, bevor das Fazit und ein abschließender Ausblick auf mögliche zukünftige Forschungsbereiche im 7. Kapitel folgen.

2. Grundlagen

In diesem Kapitel sollen die Grundlagen der Filmtonpostproduktion und speziell des ADR-Verfahrens, sowie die Grundlagen der Künstlichen Intelligenz näher erläutert werden. Dabei wird sowohl beispielhaft auf einige KI-Anwendungen in der Tonpostproduktion eingegangen als auch auf die grundlegende Funktionsweise von Sprachsynthese durch eine KI.

2.1. Filmtonpostproduktion und ADR

Da der ADR-Prozess von anderen Prozessen in der Postproduktion abhängt, wird in diesem Kapitel zuerst auf die generellen Schritte eingegangen, die in einer klassischen Filmtonpostproduktion durchlaufen werden.

2.1.1. Dialog-Edit

Der erste Schritt in der Tonpostproduktion ist bereits ein sehr wichtiger Schritt im Bezug auf den ADR-Prozess. Der erste Schritt ist nämlich der sog. „Dialog-Edit“. Ziel dieser Phase in der Postproduktion ist es, die am Filmset aufgenommenen Tonspuren für die abschließende Mischung aller Tonspuren vorzubereiten (Purcell, 2007, S. 1). Den Dialog-Edit übernimmt dabei häufig der*die Settonmeister*in, also die Person, die auch für die Aufnahme am Set verantwortlich war. Es ist allerdings auch üblich, dass für diesen Arbeitsschritt ein*e separate*r Dialog-Editor*in herangezogen wird, um die aufgenommenen Tonspuren für die Mischung anzupassen, da diese keinerlei Kenntnisse über den Filmdreh haben und somit unvoreingenommen arbeiten können. Nachteil hierbei ist jedoch, dass externe Dialog-Editor*innen mehr Einarbeitungszeit benötigen, was daher unbedingt miteingeplant werden sollte. Je nach Budget und zeitlicher Flexibilität muss hier also entschieden werden, wer den Schritt des Dialog-Edits übernimmt. Wichtig ist außerdem, dass beim Dialog-Edit, anders als der Name es vermuten lässt, nicht nur die Dialoge bearbeitet und angepasst werden, sondern auch Umgebungsgeräusche, spezielle Soundeffekte, und generell alles, was direkt am Set aufgenommen wurde (Purcell, 2007, S. 1).

Wie bereits erwähnt, ist der Dialog-Edit bereits ein entscheidender Schritt, der direkt mit dem ADR-Prozess zusammenhängt. Die editierende Person beurteilt während dieser Phase nämlich alle Dialoge nach ihrer Qualität und entscheidet dabei, ob sie für die

Mischung qualitativ ausreichend sind (bzw. ob sie mit technischen Mitteln und Bearbeitung so weit aufgewertet werden können, um die für die Mischung vorausgesetzte Qualität zu erreichen), oder ob Sie durch das ADR-Verfahren nochmals von der schauspielenden Person eingesprochen werden müssen.

Nach Abschluss des Dialog-Edits gibt es einige weitere Schritte, die vor der finalen Ton-Abmischung beendet werden müssen. Dazu gehören unter anderem das Sounddesign, die Geräuschtonaufnahmen (auch „Foley“ genannt), sowie der ADR-Prozess. Da alle diese Schritte unabhängig voneinander sind, werden diese je nach Budget und Zeit parallel durchgeführt. Am Ende der gesamten Tonpostproduktion steht die Abmischung, bzw. auch das Mastering.

2.1.2. Sounddesign

Beim Sounddesign ist die Aufgabe, die visuelle Welt, die der Film erzählt, auch klanglich hörbar zu machen. Sounddesigner*innen erschaffen hierbei neue Klänge, um Orte, Personen, oder Objekte hörbar zu machen und so das Erlebnis für die Zuschauer*innen noch immersiver zu gestalten (Lensing, 2009, S. 37). Zu den Hauptaufgaben gehören hierbei die Gestaltung der Hintergrundgeräusche (auch „Atmo“ genannt), das nachträgliche Anlegen von Soundeffekten, wie z.B. Autos oder Klingeltöne, sowie das künstliche Erzeugen von Klängen, um Objekten oder Personen eine eigene Identität zu verleihen (Lensing, 2009, S. 104–117).

2.1.3. Geräuschtonaufnahmen (Foley)

Die Geräuschtonaufnahmen haben zwar ein ähnliches Ziel wie das Sounddesign, nämlich das nachträgliche Vertonen und Gestalten bestimmter Geräusche, haben dabei jedoch eine völlig andere Vorgehensweise. Der sog. „Foley-Artist“ nimmt dabei nämlich die benötigten Geräusche synchron zum Film auf. Der Grund dafür liegt in der Art der Geräusche, die aufgenommen werden müssen. Während Soundeffekte, wie z.B. Autos, sehr einfach im Nachhinein mithilfe von schon existierenden Klängen aus Bibliotheken und geschicktem Editing produziert werden können, wäre diese Vorgehensweise bei Geräuschen, wie Schritte oder Kleiderrascheln, viel zu aufwändig und zeitintensiv. Stattdessen werden diese Geräusche in Echtzeit von einer Person mithilfe diverser Gegenstände reproduziert und anschließend noch editiert. (Lensing, 2009, S. 124–140)

2.1.4. Tonabmischung und Mastering

Nach diesen Schritten und dem ADR-Prozess, auf den im darauffolgenden Abschnitt näher eingegangen wird, folgt die finale Tonabmischung und das Mastering, wobei ein*e Toningenieur*in alle vorhandenen Tonspuren klanglich aufeinander abstimmt, um so ein stimmiges Gesamtprodukt zu erschaffen (Lensing, 2009, S. 173). Zu den Aufgaben gehören dabei die dynamische Gewichtung, die Filterung und das Positionieren aller Spuren im Stereobild. Zusätzlich sind Mischtonmeister*innen auch dafür verantwortlich geeignete Hallräume auszuwählen, sodass die Größe und die Beschaffenheit von Räumen auch klanglich wahrgenommen werden kann (Lensing, 2009, S. 175). Im Anschluss wird der abgemischte Film noch gemastert, indem das Material für diverse Formate angepasst wird. Ein Film, der für das Kino produziert wurde, wird beispielsweise anders gemastert als ein Film, der für das Fernsehen produziert wurde. Grund hierfür sind unterschiedliche Vorgaben bezüglich Lautheit und auch Dateiformat (Waddell, 2013, S. 29-34).

2.1.5. ADR

Nachdem nun alle Schritte einer klassischen Tonpostproduktion näher beleuchtet wurden, soll nun der ADR-Prozess näher beschrieben und in seinen Grundlagen erklärt werden.

Bevor die eigentliche Aufnahme der Dialogszenen beginnt, müssen noch einige Vorbereitung getroffen werden. Am wichtigsten ist hierbei das sog. „Cueing“, bzw. das Programmieren und Anlegen von Listen, die alle aufzunehmenden Dialogausschnitte enthalten (Kizer, 2024, S. 51). Neben den exakten Zeitstempeln müssen hier auch noch andere Daten, wie z.B. der sprechende Charakter, eine Transkription des gesprochenen Satzes, sowie die Szenennummer angegeben werden. Die für diese Aufgabe verantwortliche Person erhält die zu programmierenden Dialogausschnitte meistens von dem*der Dialog-Editor*in oder dem*der Mischtonmeister*in, welche*r in den häufigsten Fällen aufgrund der Expertise die Entscheidung trifft, ob eine nachträgliche Aufnahme durch ADR notwendig ist (Kizer, 2024, S. 29).

Das Anlegen dieser Listen ist notwendig, um sie während der eigentlichen Aufnahme in Verbindung mit einer Software zu nutzen, die den ADR-Prozess in vielfacher Hinsicht beschleunigt und vereinfacht. Durch das Nutzen einer speziell für den ADR-Prozess entwickelten Software ist es möglich automatisch Hilfsmittel, wie Piepstöne oder Balken zu generieren, die den Schauspieler*innen den Zeitpunkt ihres Einsatzes signalisieren. Viele Softwarelösungen, wie z.B. Nuendo (Steinberg, 2023), bieten auch Funktionen an, mit denen per Knopfdruck zwischen dem originalen Dialogausschnitt und dem neu aufgenommenen gewechselt werden kann, womit eine effizientere Bewertung der Aufnahmen ermöglicht wird. Ohne das vorherige Cueing wären diese Funktionen nur bedingt nutzbar.

Nachdem alle Vorbereitungen getroffen wurde können die nachträglichen Dialogaufnahmen beginnen. Die Aufnahmen werden dabei häufig nicht mit einem klassischen Studiomikrofon gemacht, sondern mit einem Ansteckmikrofon und einer Tonangel. Dieses Verfahren wird genutzt, um möglichst nah an den originalen am Set aufgenommenen Klang heranzukommen, weshalb in den meisten Fällen ebenfalls mit den am Set genutzten Mikrofonen gearbeitet wird. Dies vereinfacht das klangliche Einbetten der Aufnahmen in den Dialog und reduziert den benötigten Einsatz von Software in der Nachbearbeitung (Kizer, 2024, S. 126).

Nachdem die Aufnahmen fertiggestellt wurden, müssen diese noch von einem*r Toningenieur*in bearbeitet werden, um einerseits die Lippensynchronität und andererseits die klangliche Immersion in den Film und den Dialog zu gewährleisten (Kizer, 2024, S. 199). Die fertigen Dialoge werden abschließend für die finale Tonabmischung an die verantwortliche Person weitergeleitet.

2.2. KI in Postproduktion

In diesem Kapitel soll dargestellt werden, welche Arten von KI heutzutage bereits in der Tonpostproduktion Anwendung finden, bzw. welche Tools es bereits gibt, und wie diese in ihren Grundzügen funktionieren.

2.2.1. Denoising

Die wohl gängigste Anwendung findet KI in der Aufgabe des sog. „Denoisings“. Ziel dabei ist eine Audioaufnahme qualitativ aufzuwerten, indem vorhandenes Hintergrundrauschen entfernt wird. Neben bekannten Audiotools, wie „Clarity Vx“ (Waves Audio Ltd., 2022), gibt es mittlerweile auch zahlreiche Webseiten, auf denen mit nur wenigen Klicks Hintergrundrauschen mithilfe von KI entfernt werden kann. Die grundlegende Funktion basiert dabei auf der Nutzung eines neuronalen Netzes.

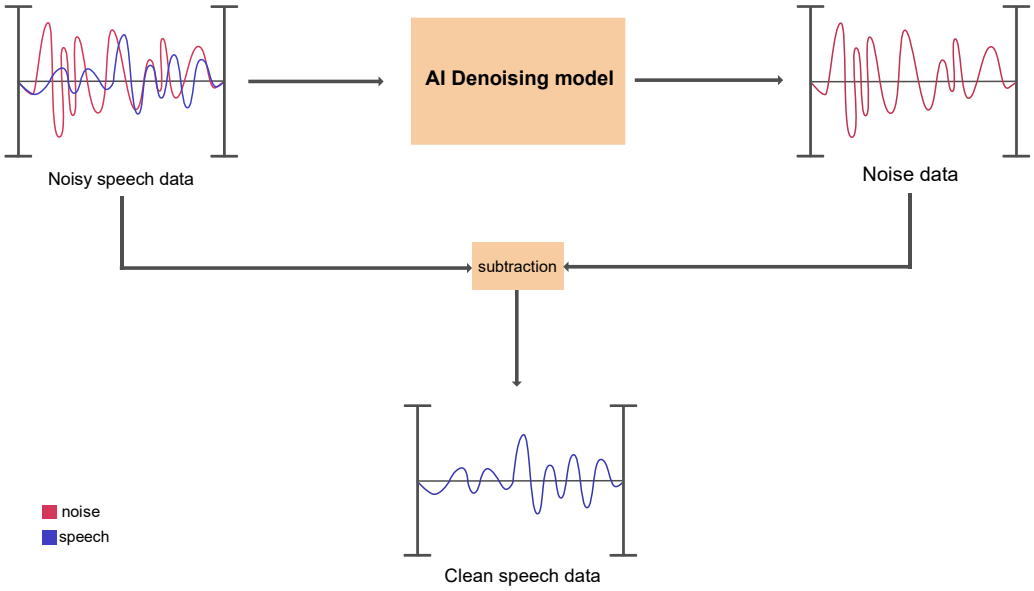

Abbildung 2: Arbeitsprozess eines KI-basierten Audio Denoising Modells (in Anlehnung an Mohammed & Radhika, 2022, S. 2)

Die KI analysiert dabei die eingegeben Audiodaten und trifft daraus eine Vorhersage über das vorhandene Rauschen, welches abschließend von der eingegebenen Audioaufnahme subtrahiert, bzw. entfernt wird (vgl. Abbildung 2). Das Training eines solchen Modells erfolgt mit Eingabe- und Ausgabepaaren, von denen ein Datenpunkt Rauschen enthält, während der andere kein Rauschen enthält (Mohammed & Radhika, 2022, S. 8).

Vor dem Training und der Nutzung müssen die Audiodaten jedoch in ein Spektrogramm gewandelt werden, das neben den Zeit- und Amplitudeninformationen auch Informationen über die Frequenz enthält (vgl. Abbildung 3), damit das Modell auch mit frequenzabhängigen Informationen arbeiten kann (Mohammed & Radhika, 2022, S. 8).

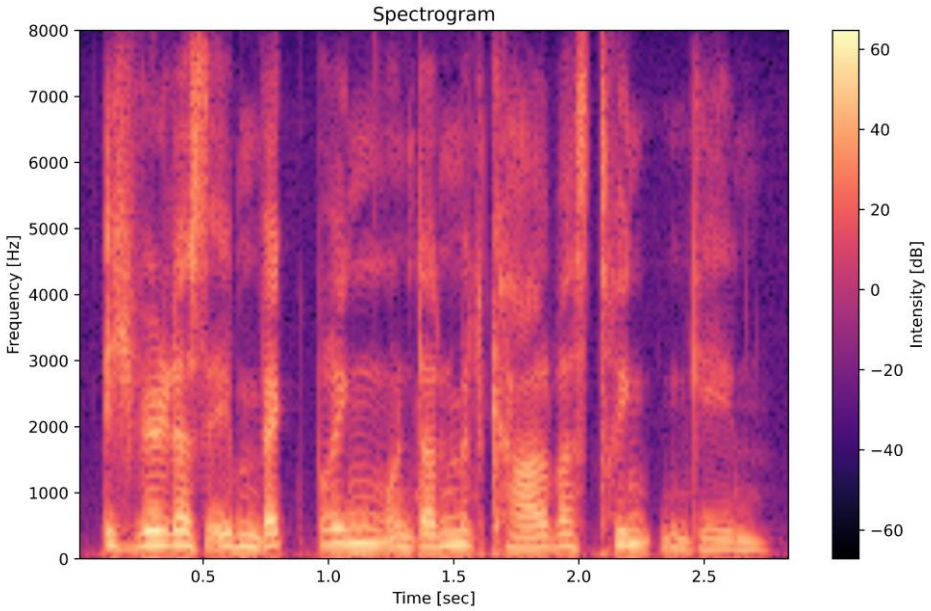

Abbildung 3: Beispiel eines Audio-Spektrogramms (eigene Darstellung)

2.2.2. Mastering

Eine weitere Anwendung von KI in der Tonpostproduktion erfolgt beim letzten Schritt, nämlich dem Mastering. Hierbei gibt es jedoch die unterschiedlichsten Varianten: Von Tools, die lediglich unterstützend auf KI setzen, wie z.B. „Ozone“ (iZotope Inc., 2023), bis hin zu KI-basierten Lösungen, die den kompletten Prozess des Masterings übernehmen, z.B. „LANDR“ (LANDR, o. D.). Beide Systeme funktionieren jedoch in ihren Grundzügen ähnlich.

Das Training der KI erfolgt wieder mit Audiodaten, die in ihre zugehörigen Spektrogramme gewandelt wurden. Anders als bei der Aufgabe des „Denoisings“ bestehen die Daten hierbei jedoch aus zahlreichen gemasterten Songs mit passenden Beschreibungen, wie die jeweiligen Toningenieur*innen dabei vorgegangen sind. Die KI lernt während des Trainings also, welche Schritte beim Mastering welche Veränderungen im Spektrogramm mit sich bringen (Birtchnell, 2018, S. 2). Da die Kunst des Masterings jedoch sehr

genreabhängig ist, bringt das Benutzen einer KI den großen Nachteil mit sich, dass nur jene Genres verlässlich verarbeitet werden können, die Teil der Trainingsdaten waren. Somit kann es bei Songs oder Filmen, die keinem spezifischen Genre zugeordnet werden können, oder deren Genre nicht von den Trainingsdaten abgedeckt wurde, zu unvorhersehbaren Ausgaben kommen (Birtchnell, 2018, S. 7).

Während das Training ähnlich abläuft, unterscheiden sich jedoch die KI-Systeme von LANDR und Ozone in ihrer Flexibilität. Während Nutzer*innen bei Ozone lediglich durch den sog. „Master Assistant“ Vorschläge für das Mastering bekommen, dabei aber in jeden Signalverarbeitungsschritt eingreifen können, ist dies bei LANDR nicht möglich.

Abbildung 4: Benutzeroberfläche des „Master Assistants“ in Ozone (iZotope Inc., 2023)

Ozone lässt zahlreiche Anpassungen bezüglich Frequenzgang, Dynamik und Stereobreite zu (vgl. Abbildung 4). Im Vergleich dazu haben Nutzer*innen bei LANDR nur die Auswahl zwischen verschiedenen Stilen und Lautheiten (vgl. Abbildung 5).

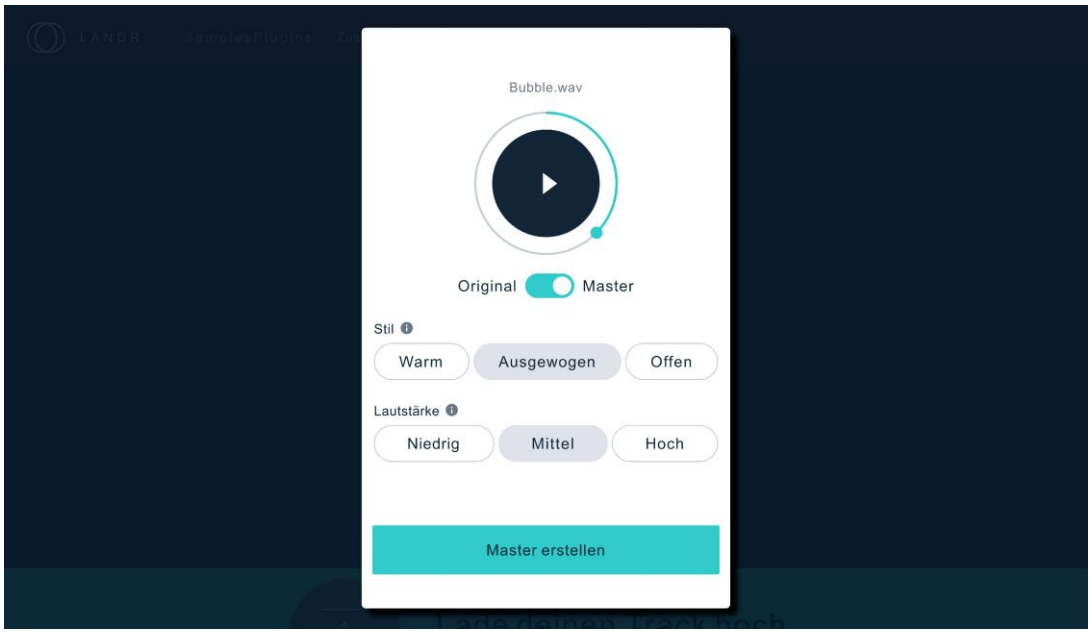

Abbildung 5: Benutzeroberfläche von LANDR (LANDR., o.D.)

Anders als bei der Aufgabe des Denoisings, bei der das KI-System nahezu alle Probleme selbstständig mit entsprechend hoher Qualität bewältigt, ist bei den Anforderungen des Masterings ein KI-System nach wie vor klassischen Toningenieur*innen mit Expertise unterlegen, wenn es um die Qualität des Masterings geht. KI kann den Prozess jedoch deutlich effizienter gestalten und auch für unerfahrene Künstler*innen zugänglicher machen (Birtchnell, 2018, S. 14).

2.2.3. Sounddesign

Die wohl aktuellsten Entwicklungen zum Thema KI in der Tonpostproduktion gibt es unter anderem beim Thema Sounddesign. Wie auch schon bei den anderen Anwendungsgebieten, gibt es auch hier verschiedene Arten, wie ein KI-System genutzt wird, um bei dem Prozess des Sounddesigns zu unterstützen. Es gibt Systeme, die lediglich bei der Suche und der Verarbeitung bestimmter Klänge unterstützen, wie z.B. der „AI Ambience Generator“ (Krotos, 2024), welcher jedoch nicht generativ arbeitet, also keine Klänge durch künstliche Synthese generiert. Auf der anderen Seite des Spektrums gibt es auch experimentelle Tools, wie z.B. „Sonic Alchemist“ (Sonomagic & Puronas, 2024), welches den kompletten Prozess eines Sounddesigns übernehmen kann. Die Funktionsweise dieses Tools soll in diesem Kapitel näher erläutert werden.

Sonic Alchemist besteht in der Basis aus drei Teilsystemen, die jeweils unterschiedliche Aufgaben erfüllen. Das erste Teilsystem nutzt eine KI, um das eingegebene Videomaterial zu analysieren und Zeitpunkte und Objekte zu identifizieren, die mit Sounds versehen werden sollen. Das zweite Teilsystem ist für die Generierung dieser Sounds verantwortlich (Puronas, 2023, S. 97). Das KI-System kann dabei sowohl auf Sounds einer Datenbank zurückgreifen als auch auf Basis von vorhandenen Sounds neue generieren. Wie auch schon bei den anderen KI-Anwendungen arbeitet das System dabei mit den Spektrogrammen der Audiodaten, um auch frequenzabhängige Informationen zu verarbeiten (Puronas, 2023, S. 101). Das dritte Teilsystem ist dafür zuständig, die in den ersten beiden Systemen generierten Daten zu vereinen. Die generierten Sounds werden also synchron zu den korrespondierenden Zeitpunkten im Videomaterial angelegt und noch weiter auditiv bearbeitet (Puronas, 2023, S. 97).

Sonic Alchemist ist jedoch nicht in der Lage konventionelle Ergebnisse zu liefern, die bisherigen Produktionsstandards entsprechen (Puronas, 2023, S. 106). Dies soll jedoch nach der Meinung des Entwicklers Vytis Puronas auch nicht das Ziel dieses Tools sein. KI soll hierbei viel mehr bei dem kreativen Prozess unterstützen und Sounddesigner*innen helfen, herkömmliche Methoden abzulegen und Innovationen zu entdecken und zu nutzen (Puronas, 2023, S. 106).

2.3. Sprachsynthese und KI

In diesem Kapitel soll in den Grundzügen erklärt werden, wie ein KI-basiertes TTS-Modell funktioniert, aus welchen Komponenten es besteht und wieso diese für die in dieser Arbeit untersuchten Anwendung nötig sind. Da der Großteil der für die Studie benötigten Ausschnitte mithilfe des KI-Systems „IMS-Toucan“ (Lux et al., 2024) generiert wurden, wird das Prinzip der Sprachsynthese beispielhaft an diesem System näher erklärt. IMS-Toucan besteht in den Grundzügen aus drei verschiedenen Systemen, die jeweils eine andere Aufgabe im Prozess der Sprachsynthese übernehmen. Es gibt ein System, dass den eingegebenen Text in Phoneme wandelt, anschließend wandelt ein weiteres System diese Phoneme in Spektrogramme, und als letzten Schritt werden diese Spektrogramme in Wellenformen gewandelt (Lux et al., 2023, S. 1). Um einen Stimmenausschnitt als Referenz für die Sprachsynthese nutzen zu können, ist noch ein weiteres System nötig. Im Fall von IMS-Toucan werden hierfür zwei verschiedene Modelle zur Sprecher*innen-

Einbettung genutzt, welche zusammengeführt werden. Der grundlegende Aufbau des Systems ist in Abbildung 6 dargestellt.

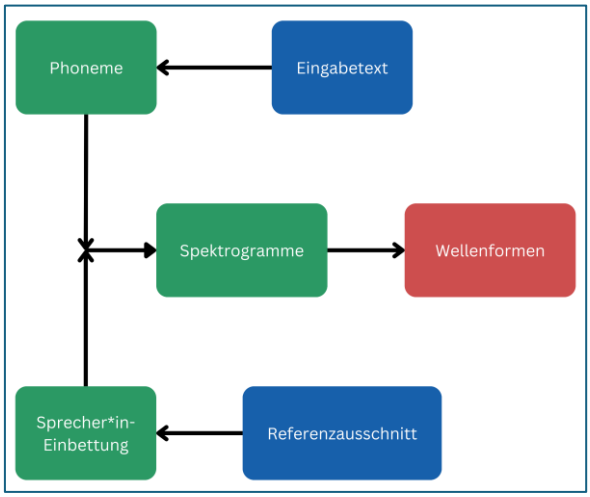

*Abbildung 6: Grundlegende Architektur von IMS Toucan. Die Farben stellen die Eingabe (blau), Ausgabe (Rot) sowie die für Nutzer*innen nicht sichtbaren Umwandlungsschritte (grün) dar. (eigene Darstellung)*

2.3.1. Text zu Phonem

Um den Eingabetext in Phoneme zu wandeln, nutzt IMS-Toucan ein System namens „Phonemizer“ (Bernard & Titeux, 2021, zitiert nach Lux et al., 2023, S. 1). Um diese Sequenz von Phonemen nun für die anderen Komponenten lesbar zu machen, werden sie zusätzlich noch mithilfe von den Systemen „PanPhon“ (Mortensen et al., 2016, zitiert nach Lux & Vu, 2022, S. 3) und „Papercup“ (Staib et al., 2020, zitiert nach Lux & Vu, 2022, S. 4) in eine Sequenz von Vektoren umgewandelt. Diese Vektoren sind sog. „one-hot-encodings“ und repräsentieren durch eine Folge der Ziffern 0 und 1 das Verhalten des menschlichen Stimmtrakts beim Produzieren der jeweiligen Phoneme. (Lux et al., 2023, S. 1)

2.3.2. Sprecher*innen-Einbettung

Um das Klonen einer Stimme zu ermöglichen, werden zwei verschiedene Modelle verwendet, mit denen die Einbettung der Stimmen der Sprecher*innen möglich ist. Der schlussendliche Vektor für die Weiterverarbeitung der Daten besteht dabei aus Werten, die von beiden Modellen berechnet wurden. Das erste Modell nutzt sog. „X-Vektoren“ (Snyder et al., 2018, zitiert nach Lux et al., 2022, S. 2). Dabei werden Features in Form von „Mel Frequency Cepstral Coefficients“ (MFCC) aus Audiosignalen extrahiert (Snyder et al., 2017, S. 2). MFCCs bestehen dabei aus in bestimmten Zeitintervallen berechneten Spektrogrammen, die im Anschluss durch Filter und Umwandlungen weiter bearbeitet

werden (Abdul & Al-Talabani, 2022, S. 2–3). Diese gewonnenen Informationen werden anschließend in Form von X-Vektoren gespeichert, um weiterverarbeitet zu werden.

Das zweite Modell nutzt die „ECAPA-TDNN“-Architektur (Desplanques et al., 2020, zitiert nach Lux et al., 2022, S. 2), welche auf den X-Vektoren basiert, jedoch einige Verbesserungen mit sich bringt. Der finale Vektor zur Repräsentation der Features der Sprecher*innen bei IMS-Toucan besteht aus 704 Dimensionen, wovon die ersten 192 Positionen durch die ECAPA-TDNN-Architektur berechnet werden, und die letzten 512 durch die klassischen X-Vektoren (Meyer et al., 2022, S. 2). Dieser Vektor wird dann schlussendlich mit den Vektoren aus Kapitel 2.3.1 kombiniert und dimensional angepasst, wodurch nun sowohl die textlichen Phoneme als auch die stimmlichen Eigenschaften in denselben Vektoren repräsentiert werden (Meyer et al., 2022, S. 2).

2.3.3. Phonem zu Spektrogramm

Um diese Vektoren nun in Spektrogramme umzuwandeln, werden sowohl „FastSpeech 2“ (Ren et al., 2020, zitiert nach Lux et al., 2023, S. 2), als auch „FastPitch“ (Łańcucki, 2020, zitiert nach Lux et al., 2023, S. 2) genutzt, und an die Anwendung in IMS-Toucan angepasst. Durch die Kombination dieser beiden Systeme ist während der Inferenz (also dem eigentlichen Generieren von Sprache) eine Kontrolle über diverse Parameter möglich (Lux et al., 2023, S. 2). So ist es z.B. möglich die Dauer (bzw. die Geschwindigkeit), die Varianz der Stimmhöhe und Energie oder auch die Prosodie zu steuern. Prosodie ist die „Gesamtheit spezifischer sprachlicher Eigenschaften wie Akzent, Intonation, Quantität, (Sprech-) -Pause.“ (Bussmann, 2008, S. 559). Um diesen Schritt simpler und effizienter zu gestalten, werden Spektrogramme lediglich mit einer Samplerate von 16kHz generiert (Lux et al., 2023, S. 2).

2.3.4. Spektrogramm zu Wellenform

Im letzten Schritt müssen die generierten Spektrogramme wieder in eine Wellenform transformiert werden. Hierfür wird ein neuronaler Vocoder genutzt, also ein Vocoder, welcher auf einem neuronalen Netz basiert. Im Fall von IMS-Toucan wurde der Vocoder „HiFi-GAN“ (Kong et al., 2020, zitiert nach Lux et al., 2024a, S. 2) genutzt. Im selben Schritt wird außerdem die Samplerate durch sog. „Upsampling“ von 16kHz auf 24kHz erhöht, um die generierte Sprache nochmals qualitativ aufzuwerten.

Zusammenfassend wird die Sprachsynthese bei einem KI-basierten TTS-System also grundsätzlich durch drei, bzw. im Falle von IMS-Toucan durch 4 unterschiedliche Komponenten ermöglicht. Die beiden Eingaben, also der Text und die Referenzstimme werden parallel verarbeitet und in Form eines gemeinsamen Vektors abgebildet. Anschließend werden die Phoneme mit den extrahierten Eigenschaften der Referenzstimme in Spektrogramme umgewandelt. Hierbei können bei der Inferenz einige Anpassungen bezüglich Geschwindigkeit, Prosodie sowie Stimmhöhe und Energie gemacht werden. Im letzten Schritt werden diese Spektrogramme durch einen Vocoder in ihrer Samplerate erhöht und wieder in eine Wellenform gebracht.

3. Stand der Forschung

In diesem Kapitel wird der aktuelle Forschungsstand im Bereich der KI-basierten Sprachsynthese und aktuelle Entwicklungen im Bereich des ADR-Verfahrens in der Filmtonpostproduktion dargestellt. Gerade im Bereich der künstlichen Intelligenz gab es in den letzten Jahren einige Fortschritte bezüglich der Qualitätsverbesserung und Kontrolle einiger Systeme. Diese Fortschritte sollen in den folgenden Abschnitten nun deshalb näher erläutert werden.

3.1. Aktuelle Entwicklungen in der Sprachsynthese

In der Sprachsynthese, bzw. auch bei der TTS-Sprachsynthese kam es in mehreren Bereichen zu einigen Fortschritten, welche in diesem Kapitel näher beleuchtet werden sollen. Aufgrund der Relevanz für die hier vorliegende Arbeit wird jedoch nur im speziellen auf die Synthese von emotionaler Sprache und das Klonen von Stimmen eingegangen.

3.1.1. Synthese von emotionaler Sprache

Im Jahr 2019 waren die fortschrittlichsten TTS-Sprachmodelle bereits in der Lage natürlich klingende Sprache zu generieren. Was jedoch gefehlt hat war präzise Kontrolle über die Ausgaben des Modells und über die Emotionen der generierten Sprache (Tits et al., 2019, S. 1-2). Der Grund hierfür lag hauptsächlich in der schweren Verfügbarkeit von geeigneten Datensätzen, die zum einen groß genug sind, und zum anderen emotionale Sprache enthalten (Tits et al., 2019, S. 2). Während man dieses Problem im Jahr 2019 durch sog. „Finetuning“ (das erneute Trainieren und Anpassen des Modells auf einem kleineren Datensatz) gelöst hat, gibt es mittlerweile einige andere vielversprechende Ansätze.

Erst dieses Jahr wurde „EmoSphere-TTS“ (Cho et al., 2024) vorgestellt, mit dem Ziel nicht nur vordefinierte Emotionen generieren zu können, sondern diese auch uneingeschränkt in Intensität und Variation kontrollieren zu können.

Hierzu haben die beteiligten Forscher*innen einen „spherical emotion vector space“ (Cho et al., 2024, S. 1), also einen sphärischen Vektorraum, der die Emotion abbilden soll, eingeführt. Dieser Raum wird durch drei Dimensionen definiert: „Dominance“ (Dominanz), „Arousal“ (Erregung)“ und „Valence“ (Valenz). Hierdurch soll es möglich

sein, jegliche Emotionen in diesem Vektorraum darzustellen. Vor der Darstellung in einem sphärischen Raum ist es allerdings notwendig, die Emotionen in einem dreidimensionalen kartesischen Koordinatensystem darzustellen und dieses dann in ein sphärisches Koordinatensystem zu transformieren (vgl. Abbildung 7).

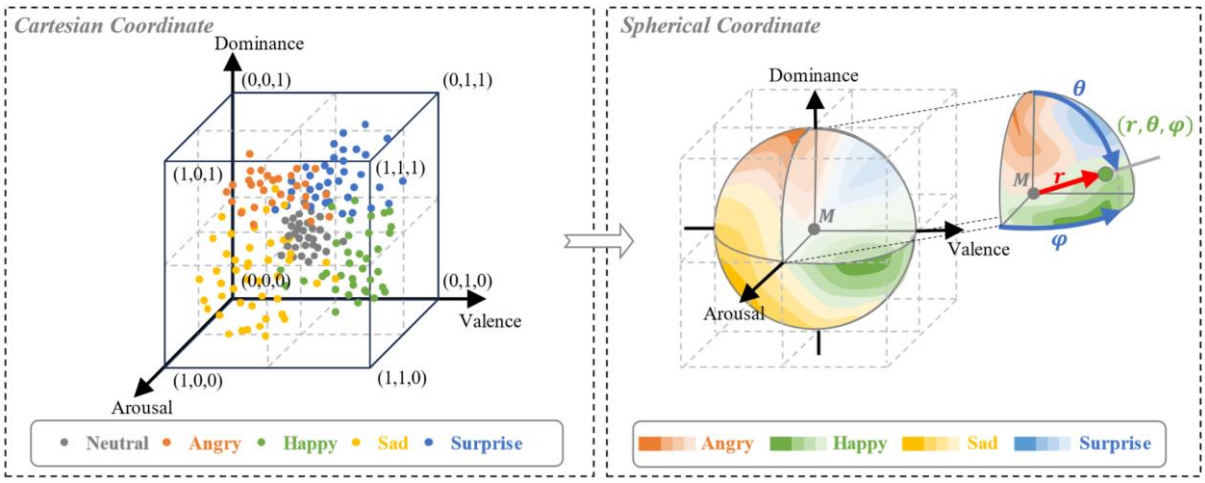

Abbildung 7: Transformation der abgebildeten Emotionen von einem kartesischen Koordinatensystem in ein sphärisches Koordinatensystem (Cho et al., 2024, S. 2)

Die Entfernung der einzelnen Punkte (welche die gewünschte Emotion repräsentieren) vom Ursprung des Koordinatensystems zeigt dabei an, wie stark die jeweilige Emotion ausgeprägt sein soll. Der Winkel, bzw. die Richtung des Vektors steht dabei für die Art der Emotion. Durch das Training des KI-Systems mit dieser Art der Emotionsrepräsentation ist es bei der Inferenz möglich jegliche Art von Emotion zu generieren (Cho et al., 2024, S. 3). Durch Experimente und Evaluationen konnte gezeigt werden, dass „EmoSphere-TTS“ schon jetzt effektiv komplexe Emotionen synthetisieren kann. Die beteiligten Forscher*innen wollen diesen Ansatz in Zukunft noch weiter verbessern, mit dem Ziel die Emotionen nicht nur auf Satz-Ebene, sondern auch auf Phonem-Ebene kontrollieren zu können (Cho et al., 2024, S. 4).

Einen weiteren Ansatz gezielt Emotionen zu synthetisieren, bietet „EmoKnob“ (Haozhe et al., 2024). Ähnlich wie „EmoSphere-TTS“ berechnet das Modell hierbei ebenfalls einen Vektor, der die Art der Emotion repräsentieren soll (Haozhe et al., 2024, S. 4). Grundlage der Berechnung dieses Vektors sind Audiopaare, die bestimmte Voraussetzungen erfüllen müssen. Zum einen müssen beide Audioausschnitte von derselben Person gesprochen sein, und zum anderen muss ein Ausschnitt mit der gewünschten Emotion vorliegen und

ein neutraler Ausschnitt. „EmoKnob“ kann anschließend auf Basis der analysierten Unterschiede zwischen beiden Ausschnitten einen Vektor berechnen, der die gewünschte Emotion repräsentiert (vgl. Abbildung 8). Dieser Vektor wird anschließend normalisiert, um Nutzer*innen die Möglichkeit zu bieten, durch Multiplikation mit einem „Emotion Control Strength“-Parameter die Intensität der Emotion in der generierten Sprache zu steuern (Haozhe et al., 2024, S. 4).

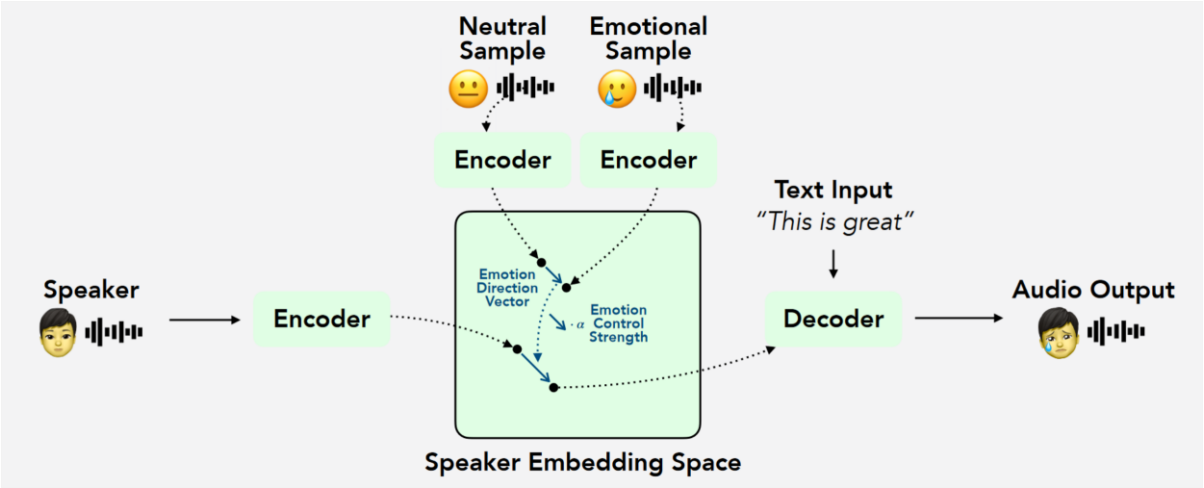

Abbildung 8: Grundlegende Architektur von „EmoKnob“ (Haozhe et al., 2024)

Dieser Vektor wird anschließend, wie bei anderen TTS-Modellen, zusammen mit dem Eingabetext weiterverarbeitet, um abschließend eine Audiodatei zu generieren (vgl. Abbildung 8).

Zusammenfassend hat das Forschungsgebiet der emotionalen Sprachsynthese in den letzten Jahren erhebliche Fortschritte gemacht. Mittlerweile ist es möglich Emotionen präzise in Form von Vektoren darzustellen und sie somit kontrolliert bei der Synthese von Sprache miteinfließen zu lassen. Aktuelle Systeme sind jedoch zurzeit noch nicht für den Gebrauch mit deutscher Sprache ausgelegt, weshalb sie im Zusammenhang mit dieser Arbeit noch keine Anwendung finden konnten. In dieser Arbeit soll deshalb untersucht werden, wie die bisherigen deutschsprachigen Modelle in Bezug auf emotionale Sprache und ADR abschneiden. Mit zusätzlichem Training auf deutschsprachigen Daten in Zukunft, könnten die hier beschriebenen Systeme jedoch auch für den Gebrauch beim ADR-Prozess in Frage kommen, um die Emotionen der generierten Sprache präzise zu kontrollieren und sie so möglichst verlässlich der visuellen Komponente anzupassen.

3.1.2. Voice-Cloning

Ein weiterer Forschungsbereich, der v.a. für den in dieser Arbeit beschriebenen Anwendungsfall wichtig ist, ist das sog. „Voice-Cloning“, also das Klonen, bzw. Imitieren einer Stimme durch eine KI. Die neusten KI-Systeme sind dabei bereits in der Lage, Stimmen verlässlich und natürlich klingend zu klonen und benötigen dabei lediglich einen Ausschnitt von wenigen Sekunden als Eingabe.

Das wohl fortschrittlichste Modell ist aktuell Microsofts „VALL-E“ (Wang et al., 2023), bzw. dessen Nachfolger „VALL-E 2“ (Chen et al., 2024), welches erst im Juni dieses Jahres an die Öffentlichkeit gebracht wurde und laut Angaben der Entwickler*innen Ausgaben synthetisieren kann, die mit der Qualität und Wertigkeit menschlicher Sprache zu vergleichen sind (Chen et al., 2024, S. 1).

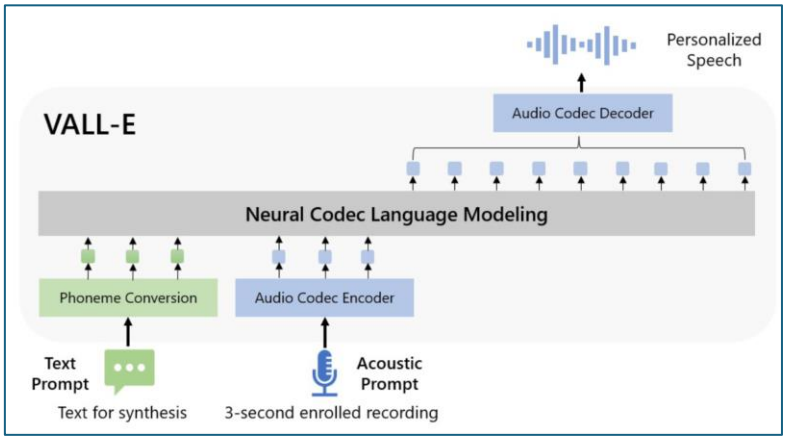

Abbildung 9: Grundlegende Architektur von „VALL-E“ (Wang et al., 2023)

Der Aufbau des Systems entspricht dabei in den Grundzügen aus den in Kapitel 2.3 erwähnten Komponenten, jedoch mit dem Unterschied, dass die aus dem Eingabetext stammenden Phoneme nicht in Spektrogramme, sondern in einen diskreten Audiocodierungs-Code umgewandelt werden (vgl. Abbildung 9) (Wang et al., 2023, S. 1). Der Unterschied zu bisherigen TTS-Modellen besteht also in der Repräsentation der Eingaben und deren Weiterverarbeitung, da VALL-E nicht auf Basis von Spektrogrammen sondern mit sog. „Acoustic Tokens“ (Wang et al., 2023, S. 2) arbeitet. Dies bietet verschiedenste Vorteile: Zum einen besitzt das Modell dadurch die Fähigkeit des sog. „In Context Learnings“, welches dem Modell ermöglicht aus dem Kontext der Eingaben zu lernen, wodurch zusätzliche Anpassungen, wie z.B. Finetuning, reduziert werden können

(Wang et al., 2023, S. 3). Andererseits wird dadurch die benötigte Rechenzeit während der Sprachsynthese erheblich verringert (Wang et al., 2023, S. 6). In Bezug auf diese Arbeit zeigt sich, dass die Nutzung von VALL-E und vergleichbare Modelle durch den gewählten Ansatz, die Aufgabe der Sprachsynthese in Form einer Sprachmodellierung mit diskreten Audiokodierungs-Codes zu realisieren, für die Sprachsynthese beim ADR-Prozess einige Vorteile (siehe oben) bieten würde. VALL-E, bzw. VALL-E 2 sind zum jetzigen Zeitpunkt jedoch nicht öffentlich nutzbar und werden lediglich zu wissenschaftlichen Zwecken weiterentwickelt (Chen et al., 2024, S. 3), weshalb im Rahmen dieser Arbeit alternative TTS-Systeme genutzt werden müssen.

Zusammenfassend gab es in den vergangenen Jahren also diverse Innovationen auf dem Gebiet der künstlichen Sprachsynthese. Mittlerweile ist es möglich gezielt die Emotionen der Ausgabe zu steuern und im Bereich des Voice-Clonings sind Sprachausgaben möglich, die mit der Qualität eines Menschen vergleichbar sind. Was jedoch auffällt ist, dass die meisten dieser innovativen KI-Systeme lediglich dazu in der Lage sind englische Ausgaben zu generieren. Für den in dieser Arbeit betrachteten Anwendungsfall sind jedoch Systeme erforderlich, die in der Lage sind, auch deutsche Ausgaben zu generieren. Dies könnte potenziell zu Herausforderungen für die KI-Systeme führen, den Anforderungen für einen Ersatz des ADR-Verfahrens gerecht zu werden.

3.2. Aktuelle Entwicklungen beim ADR-Verfahren

Auch wenn das ADR-Verfahren an sich schon mehrere Jahrzehnte alt ist und in den Grundzügen immer gleich aufgebaut ist, gibt es trotzdem Weiterentwicklungen, v.a. in Form von neuer oder überarbeiteter Software, die den ADR-Prozess einfacher und weniger zeit- und kostenintensiv gestalten soll. Eine weitere Innovation ist außerdem das sog. „Remote ADR“, welches sich v.a. durch die Corona Pandemie und die damit einhergehende Kontaktbeschränkung etablieren konnte. In diesem Kapitel soll beispielhaft näher erläutert werden, welche Entwicklungen es im Bereich des ADR in den letzten Jahren gab, und wie diese sich auf den Gesamtprozess ausgewirkt haben.

3.2.1. Softwaretechnische Entwicklungen

Der Einsatz von diversen Tools ist mittlerweile im Prozess des ADRs fest verankert, sowohl in Form von Plugins, die lediglich bei kleineren Teilaufgaben unterstützen, als auch in Form von kompletten Softwarepaketen, die nahezu bei der kompletten ADR-Produktion unterstützen.

Ein Plugin, welches sowohl bei der Aufnahme als auch bei der anschließenden Nachbearbeitung der Dialoge hilft, ist „Dialogue Match“ (iZotope Inc., 2019). Aufgabe dieses Plugins ist es, den neu eingesprochenen Dialog klanglich an den ursprünglich am Set aufgenommenen Originalton anzupassen. Diese Angleichung findet dabei über verschiedene Faktoren, wie z.B. das Frequenzspektrum, den Hall oder auch das Hintergrundrauschen statt. Da das Plugin diese Angleichung innerhalb kurzer Zeit und mit nur wenigen Klicks bewerkstelligen kann, ist es außerdem möglich schon direkt während einer Aufnahme die von Dialogue Match verarbeiteten Dialoge in den Film einzubetten, wodurch eine verlässlichere Bewertung des aufgenommenen Dialogs durchgeführt werden kann. Dadurch kann v.a. Zeit eingespart und eine höhere Qualität sichergestellt werden.

Neben Plugins, die innerhalb der „Digital Audio Workstation“ (DAW), also der Software, die für die Nachbearbeitung der Dialoge genutzt wird, verwendet werden, gibt es mittlerweile auch komplette Softwarepakete, die bei jedem Produktionsschritt eingesetzt werden können.

Das wahrscheinlich gängigste Tool ist hierbei „VoiceQ“ (Kiwa Digital Ltd., 2024), welches sowohl beim Cueing und Programmieren als auch bei der Aufnahme zum Einsatz kommen kann. Durch den fortschreitenden Anstieg der Nutzung von KI-Technologie gibt es mittlerweile auch in VoiceQ die Möglichkeit, diverse veröffentlichte KI-Tools, die den Arbeitsprozess erheblich beschleunigen sollen, innerhalb der Software zu verwenden. Ein Beispiel hierfür wäre das Tool „MediaCat“ (XL8 Inc., 2022), welches durch ein KI-System Transkripte von Dialogausschnitten inklusive der benötigten Zeitstempel erstellen kann. Durch die Integration von MediaCat in VoiceQ kann der*die Toningenieur*in somit mehrere Schritte, wie z.B. das Exportieren und Importieren benötigter Daten, umgehen, wodurch viel Zeit eingespart werden kann. Die Arbeit mit

einer einzigen Software erlaubt es außerdem während der Aufnahme Änderungen in kürzester Zeit vorzunehmen, um den Ablauf möglichst reibungslos zu gestalten.

Neben Plugins und dedizierter Software gibt es mittlerweile auch eigens von den Herstellern der DAWs entwickelte Features, die speziell für die Anwendung während des ADR-Prozesses gedacht sind. Eine DAW, welche besonders viel Wert auf solche Features legt, ist „Nuendo“ (Steinberg, 2023). Nuendo bietet mittlerweile alle Funktionen, die in der Vorproduktion, der eigentlichen Produktion und in der Postproduktion des ADR-Prozesses benötigt werden. Dies ist ein großer Vorteil, da somit der komplette Filmtonpostproduktionsprozess nur innerhalb einer einzigen DAW stattfinden kann, wodurch die Postproduktion zeiteffizienter und flexibler gestaltet werden kann. Der Datenaustausch zwischen verschiedenen Toningenieur*innen ist durch die Nutzung von lediglich einer einzigen Software ebenfalls um ein Vielfaches einfacher, da nicht auf softwareunabhängige Dateiformate gesetzt werden muss.

3.2.2. Remote-ADR

Eine sehr aktuelle Entwicklung, die v.a. im Zusammenhang mit der 2020 aufgekommenen Corona-Pandemie in Verbindung zu bringen ist, ist das sog. „Remote-ADR“. Anders als beim herkömmlichen ADR-Prozess befinden sich hierbei die Schauspieler*innen und Toningenieur*innen nicht am selben Ort, bzw. im selben Tonstudio, sondern können flexibel von überall aus arbeiten, sofern das nötige Equipment zur Verfügung steht. Um die Kommunikation und den Datenaustausch hierbei so einfach wie möglich zu halten, gibt es mittlerweile speziell für diesen Anwendungszweck entwickelte Software.

Am weitesten verbreitet ist hierbei „Source Connect“ (Source Elements LLC., 2024). Source Connect bietet zahlreiche Funktionen, um latenzfreie und qualitativ hochwertige Dialogaufnahmen nur über eine Internetverbindung zu erstellen. Da die Software jedoch nicht speziell für die ADR-Anwendung, sondern vielmehr für Remote-Aufnahmen im Allgemeinen entwickelt wurde, bietet Source Connect keine Funktionen an, die das Cueing und Programmieren innerhalb der Software ermöglichen. Das Anfertigen der Cues sowie die unterstützende Nutzung von Balken oder Pieptönen muss also über eine zweite separate Software stattfinden, was natürlich die Fehleranfälligkeit erhöht.

Zusammenfassend lässt sich also sagen, dass das ADR-Verfahren in seinen Grundbestandteilen zwar seit Jahrzehnten gleichgeblieben ist, es jedoch in den letzten Jahren einige Innovationen gab, sowohl was die Plugins, bzw. Software, als auch die Integration von KI-Systemen angeht. Motivation dieser Neuerungen ist dabei immer, den ADR-Prozess noch effizienter und kostengünstiger zu gestalten, und mögliche Probleme, die mit der zeitlichen und örtlichen Verfügbarkeit der Schauspieler*innen zusammenhängen, zu umgehen. Insgesamt kam es jedoch im Bereich des ADR im Vergleich zum Bereich der KI-Sprachsynthese zu deutlich weniger Innovationen, was jedoch mit großer Wahrscheinlichkeit an dem steigenden generellen Interesse an KI-Systemen liegt.

4. Methodik

In den folgenden Kapiteln soll näher darauf eingegangen werden, wie die in der Wahrnehmungsstudie verwendeten Sprachausschnitte künstlich generiert wurden. Dabei wird zunächst die Wahl des verwendeten TTS-Systems, sowie dessen grundlegende Benutzung erläutert. Anschließend werden die für die Studie relevanten Evaluationskriterien vorgestellt, gefolgt von einer detaillierten Beschreibung des Arbeitsprozesses während der eigentlichen Sprachsynthese, sowie der durchgeführten Schritte in der weiteren Nachbearbeitung.

4.1. Sprachsynthese-Modell

In diesem Kapitel soll zunächst erklärt werden, anhand welcher Kriterien die finalen KI-TTS-Systeme ausgewählt wurden, und anschaulich beschrieben werden, wie sie während der Inferenz, also dem Generieren der Ausschnitte, genutzt wurden und welche Möglichkeiten sie boten, die Ausgabe zu manipulieren.

4.1.1. Wahl des TTS-Systems

Um ein für die Wahrnehmungsstudie optimal geeignetes TTS-System zu finden, wurden bestimmte Anforderungen aufgestellt, die zur Beantwortung der Forschungsfrage nötig waren, und nach Möglichkeit von dem System hätten erfüllt sein sollen.

Diese lauteten wie folgt:

1) Das System soll dazu in der Lage sein, deutschsprachige Ausschnitte zu generieren.

   Da sich die Untersuchung in dieser Arbeit auf den deutschsprachigen Raum bezieht und im Rahmen der Studie nur Beispiele aus deutschen Filmproduktionen genutzt werden, ist es unabdingbar, dass das TTS-System deutschsprachige Ausgaben generieren kann.

2) Das System soll als Eingabe sowohl einen zu generierenden Text als auch einen zu klonenden Stimmausschnitt entgegennehmen.

   Für den Einsatz eines TTS-Systems im Rahmen einer Filmtonpostproduktion und als Ersatz für das ADR-Verfahren ist es nötig, dass Nutzer*innen den Eingabetext frei

wählen können, um die benötigten Dialogausschnitte generieren zu können. Da außerdem Sätze bestimmter Schauspieler*innen ersetzt werden sollen, muss das KI-System dazu in der Lage sein, Stimmen anhand eines kurzen Ausschnitts zu klonen.

3) Das System soll dazu in der Lage sein, Sprache mit einer vorgegebenen Emotion zu generieren.

Da die im ADR-Prozess zu vertonenden Dialogausschnitte selten nur monotone Sprache beinhalten, ist es unabdingbar, dass das KI-System emotionale Sprache generieren kann und diese Emotionen auch möglichst kontrollierbar sind.

Zum jetzigen Zeitpunkt existiert noch kein öffentlich zugängliches KI-TTS-System, das alle drei Anforderungen erfüllt, weshalb die Nutzung zweier verschiedener Systeme für die Wahrnehmungsstudie unausweichlich ist. Die beiden final ausgewählten Systeme waren „IMS-Toucan“ (Lux et al., 2024b), welches die Anforderungen 1) und 2) erfüllt sowie „Bark“ (Suno Inc., 2023), welches Anforderung 1) und von allen getesteten Systemen am ehesten Anforderung 3) erfüllt. In welchem Ausmaß die beiden Systeme die Anforderungen erfüllen konnten wird in den folgenden Abschnitten genauer erläutert.

4.1.2. IMS-Toucan

IMS-Toucan (Lux et al., 2024b) kam v.a. in Frage, da es von einem Entwicklerteam der Universität Stuttgart entwickelt wurde und somit einerseits gut erreichbare Ansprechpartner*innen zur Verfügung standen und andererseits vom System deutschsprachige Ein- und Ausgaben unterstützt wurden.

Wie schon im letzten Kapitel, sowie in Kapitel 2.3, erwähnt ist IMS-Toucan dazu in der Lage einen Eingabetext und einen Referenzausschnitt entgegenzunehmen, und auf deren Basis einen geklonten Sprachausschnitt zu generieren. In der Theorie sollte das System also Anforderungen Nummer 1) und 2) erfüllen. Inwiefern das Klonen der Stimmen auf deutsch in der Praxis funktioniert, wird im Rahmen der noch folgenden Wahrnehmungsstudie genauer bewertet.

Die Inferenz, also die tatsächliche Synthese, funktioniert bei IMS-Toucan über das Initialisieren eines Objektes „ToucanTTSInterface“ und dem anschließenden Ausführen einer Methode mit den gewünschten Parametern.

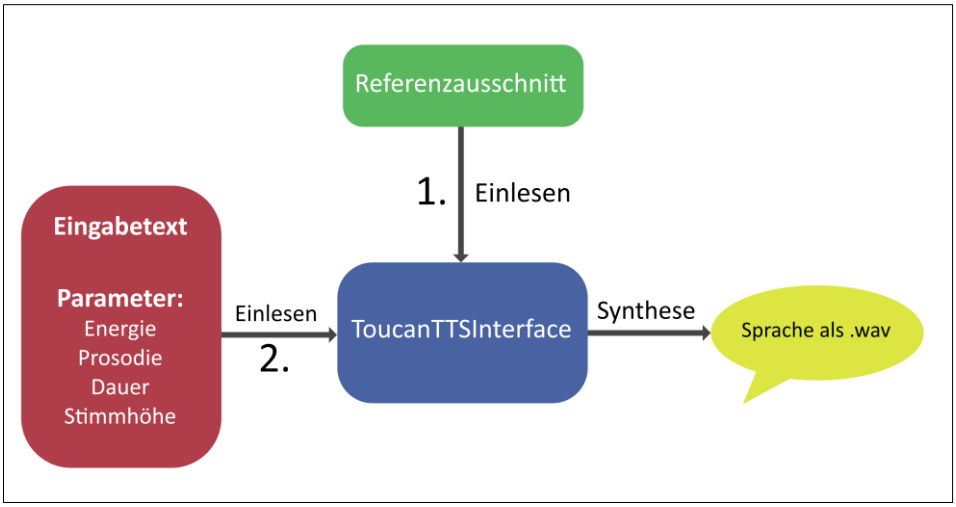

Abbildung 10: Inferenzprozess bei IMS-Toucan (eigene Darstellung)

Der genaue Ablauf während dem Prozess der Inferenz ist in Abbildung 10 genauer dargestellt. Dem ToucanTTSInterface-Objekt muss vor der eigentlichen Inferenz über eine separate Methode der zu klonende Referenzausschnitt übergeben werden. Im nächsten Schritt können ebenfalls der erforderliche Eingabetext sowie einige Parameter für die Kontrolle der Ausgabe übergeben werden. Nutzer*innen können dabei die Varianz der Energie, der Prosodie, der Dauer und der Stimmhöhe beeinflussen, wodurch einerseits sehr monotone Sprachsynthese möglich ist, aber andererseits auch sehr dynamische und ausdrucksstarke Sprachvariationen, was v.a. für die Anwendung im ADR-Bereich sehr von Vorteil ist.

Während des Testens von IMS-Toucan brachte das System leider nicht die gewünschten Ergebnisse im Hinblick auf das Klonen von Stimmen. Das Modell bietet jedoch die Möglichkeit es auf weiteren Daten zu finetunen. Da für die Wahrnehmungsstudie lediglich das Klonen von zwei verschiedenen Stimmen vorgesehen war, wurde der Ansatz des Finetunings getestet und konnte letztendlich auch zu besseren Ergebnissen führen. Das System wurde dabei nochmals mit jeweils 10-15 Audioausschnitten mit einer Länge von bis zu 15 Sekunden trainiert und im Anschluss nochmals evaluiert.

4.1.3. Bark

Da es während der Untersuchung in Form dieser Arbeit noch kein öffentlich zugängliches TTS-System gab, was dazu in der Lage war, menschliche Emotionen kontrolliert in

deutscher Sprache zu generieren, musste hierfür auf ein alternatives System umgestiegen werden.

Der Grund warum Bark (Suno Inc., 2023) für die Bewältigung dieser Aufgabe gewählt wurde war, weil dieses TTS-System ähnlich wie das in Kapitel 3.1.2 erwähnte VALL-E-System (Wang et al., 2023) ebenfalls auf Basis diskreter Audiocodes arbeitet und auch Ausgaben auf deutsch generieren kann. Die Eingabe wird zudem ohne weitere Umwandlung direkt in einen Audioausschnitt gewandelt, was somit auch Klänge ermöglicht, die über einfache Sprache hinausgehen, wie z.B. Lachen, Seufzen oder Räuspern. Zusätzlich besteht auch die Möglichkeit durch Großschreibung bestimmter Wörter die Betonung des Eingabetextes anzupassen. Durch die Repräsentation der Eingabe als diskreten Code gibt es außerdem keine Beschränkungen, was den Eingabeprompt angeht. Im Fall von Bark können solche Prompts durch eckige Klammern ([...]) in den Eingabetext miteingebunden werden, wodurch sich auch die Option ergibt, durch diesen Prompt die gewünschte Emotion zu kontrollieren. Eine beispielhafte Eingabe würde dadurch wie folgt aussehen:

“[HAPPY] Ich liebe die Hochschule der Medien!“

Beim Testen des Modells wurde jedoch schnell klar, dass die gezielte Kontrolle der Emotionen durch den Eingabeprompt sehr unzuverlässig und inkonsistent ist, weshalb das System die dritte in Kapitel 4.1.1 genannte Anforderung nur in Teilen erfüllen kann. Mit welcher Genauigkeit die von Bark generierten Emotionen erkannt werden wird ebenfalls im Rahmen der Wahrnehmungsstudie untersucht.

4.2. Evaluierungskriterien

Im folgenden Kapitel soll näher auf die Evaluierungskriterien, die sowohl für das Erstellen der Wahrnehmungsstudie als auch das Synthetisieren der Stimmausschnitte relevant sind, eingegangen werden. Diese Kriterien bilden die Grundlage für den Rahmen der Umfrage und die Formulierung der an die Teilnehmer*innen gestellten Fragen. Die Evaluierungskriterien gewährleisten eine objektive und vergleichbare Beurteilung der Stimmausschnitte.

Die Wahrnehmungsstudie soll drei zentrale Fähigkeiten der KI-Modelle evaluieren:

Im ersten Teil der Studie soll die Fähigkeit bewertet werden, menschliche Originalstimmen anhand eines kurzen Ausschnitts zu klonen. Zur Bewertung wurde der sog. „similarity mean opinion score (SMOS)“ verwendet, durch den die Testhörer*innen die Ähnlichkeit auf einer Skala von 1 (überhaupt nicht ähnlich) bis 5 (sehr ähnlich) bewerten konnten. Dieser Score wurde bereits bei der Bewertung einiger TTS-Modelle, wie z.B. VALL-E (Wang et al., 2023, S. 8) oder auch dessen Nachfolger VALL-E 2 (Chen et al., 2024, S. 10) verwendet und ist somit wissenschaftlich anerkannt und bringt eine gewisse Aussagekraft mit sich.

Im zweiten Teil der Studie soll in Verbindung mit dem dazugehörigen Filmausschnitt die Fähigkeit bewertet werden, natürlich klingende Dialoge zu generieren. Hierfür wurden mehrere verschiedene Parameter hinzugezogen, anhand dessen die Testhörer*innen die generierte Sprache evaluieren sollten. Bisher existente KI-Systeme wurden dabei sowohl in den Kategorien „Prosodie“ (z.B. Huang et al. (2023, S. 8023)), Audioqualität (ebenfalls Huang et al. (2023, S. 8023)), als auch in der Kategorie „Natürlichkeit“ selbst (z.B. Lux et al. (2022b, S. 5)) bewertet. Diese Parameter wurden in Form eines „Mean Opinion Scores“ (MOS) von 1 bis 5 gemessen. Für die Bewertung im Rahmen dieser Wahrnehmungsstudie wurde der Prosodie-MOS in zwei Unterkategorien (Satzmelodie und Rhythmus) aufgeteilt, um die Bewertung der Prosodie für die Testhörer*innen so simpel wie möglich zu gestalten. Der Natürlichkeit-MOS wurde durch einen Gesamteindruck-MOS ersetzt, um Dopplungen und Verwirrungen bei den Testhörer*innen zu verhindern. Somit wurden im zweiten Studienteil insgesamt vier verschiedene MOSs jeweils in den Kategorien Satzmelodie, Rhythmus, Audioqualität und Gesamteindruck gemessen.

Diese Kategorien sind außerdem für die anschließende Sprachsynthese von Relevanz, da Satzmelodie, Rhythmus und Audioqualität entweder bei der Inferenz selbst oder in der darauffolgenden Bearbeitung gezielt angepasst werden können.

Im dritten Studienteil soll die Fähigkeit der KI-Systeme bewertet werden, Sprache mit bestimmten Emotionen zu generieren. Die Testhörer*innen bekommen dabei eine vorgegebene Liste von Emotionen und sollten jene wählen, welche nach ihrer Wahrnehmung am ehesten zu dem gehörten Audioausschnitt passt. Für die

anschließende Auswertung wird die Genauigkeit berechnet, mit der die jeweilige Emotion korrekt erkannt wurde.

Wie die Wahrnehmungsstudie konkret strukturiert und konzipiert wurde wird im noch folgenden Kapitel 5.2 genauer erläutert.

4.3. Sprachsynthese und weitere Bearbeitung

In diesem Kapitel soll erklärt werden, wie mithilfe von IMS-Toucan und Bark die für die Wahrnehmungsstudie gewählten Stimm- und Dialogausschnitte generiert wurden, und wie diese im Anschluss für die bestmöglichen Ergebnisse weiter bearbeitet wurden.

4.3.1. Vorbereitung

Vor der eigentlichen Sprachsynthese mussten zunächst Schauspieler*innen gesucht werden, die damit einverstanden waren, ihre Stimme im Rahmen dieser wissenschaftlichen Arbeit durch ein KI-System klonen zu lassen. Für eine möglichst hohe Generalisierbarkeit der Ergebnisse der Wahrnehmungsstudie wurden sowohl ein männlicher Schauspieler als auch eine weibliche Schauspielerin gesucht. Hierfür bereit erklärt haben sich die beiden deutschen Schauspieler*innen Franziska van der Heide und Arne Löber. Von beiden Schauspielern wurden im Anschluss diverse Ausschnitte gesammelt, die sowohl als Beispiele für die Wahrnehmungsstudie dienten, als auch für das benötigte Finetuning des IMS-Toucan Systems gebraucht wurden.

4.3.2. Synthese

Für die Synthese wurde sich an der Reihenfolge der in der Wahrnehmungsstudie zu untersuchten Fähigkeiten orientiert: Zuerst wurden die Stimmen der beiden Schauspieler*innen mit dem Fokus auf möglichst hohe Ähnlichkeit geklont. Danach wurden beispielhafte Dialogausschnitte in Verbindung mit den dazugehörigen Filmausschnitten geklont. Hierbei waren außer der Ähnlichkeit auch Aspekte wie Lippensynchronität und klangliche Einbettung in den Film wichtig. Zuletzt wurden mit Bark die emotionalen Stimmausschnitte synthetisiert. Da das Bark-System nativ nicht dazu in der Lage ist Stimmen zu klonen, wurde dabei auf eine jeweils schon vortrainierte männliche, bzw. weibliche Stimme zurückgegriffen.

Für die eigentliche Synthese der Stimmausschnitte wurde ein eigenes Python-Skript angelegt (im angehängten GitLab-Repository unter dem Namen „Inferenz.ipynb“ zu finden), welches jeweils für die Inferenz mit IMS-Toucan und Bark genutzt wurde.

Der allgemeine Sprachsynthese-Prozess bestand aus vielen „Trial-and-Error“-Versuchen, da sowohl IMS-Toucan als auch Bark viele inkonsistente Ergebnisse lieferten. Während bei IMS-Toucan unterschiedliche Referenzausschnitte **derselben** Person zu sehr anders klingenden Ausgaben führten, musste bei der Synthese mit Bark v.a. was den Eingabeprompt anging sehr viel experimentiert werden. So kam es beispielsweise bei einem Eingabeprompt von *„[HAPPY] Heute Abend könnte ich es ihm sagen!“* zu deutlich anderen Ausgaben als mit dem Prompt *„[HAPPY MAN] Heute Abend könnte ich es ihm sagen!“*. Die Sprachsynthese war deshalb ein sehr zeitintensiver und aufwändiger Prozess.

4.3.3. Klangliche Nachbearbeitung

Nach der Synthese wurden alle generierten Ausschnitte klanglich bearbeitet. Ziel der Nachbearbeitung war es die Stimmausschnitte klanglich zu verbessern und sie im Rahmen der Möglichkeiten an die jeweilige Anforderung (Ähnlichkeit, Natürlichkeit, Emotionalität) anzupassen.

Grundsätzlich wurden alle generierten Ausschnitte mithilfe der „Adobe Podcast-KI“ (Adobe, 2024) von bei der Synthese entstandenen Störgeräuschen bereinigt. Da für den ersten Studienteil die Fähigkeit Stimmen zu klonen untersucht wurde, wurden die von IMS-Toucan generierten Ausschnitte nach dem Entfernen von Störgeräuschen nicht weiter bearbeitet, um so eine wahrheitsgetreue Repräsentation der Fähigkeiten von IMS-Toucan zu erhalten.

Die Stimmausschnitte des zweiten Studienteils wurden im Gegensatz dazu klanglich mehr bearbeitet. Nach der Entfernung der Störgeräusche war es v.a. wichtig die KI-generierten Ausschnitte klanglich an die im Filmausschnitt zu erkennende Umgebung anzupassen. Hierbei kam v.a. das Tool „Dialogue Match“ (iZotope Inc., 2019) zum Einsatz welches das Klangprofil (Frequenzspektrum, Hall, Umgebungsrauschen) des originalen menschlichen Dialogausschnitts auf den von IMS-Toucan generierten Ausschnitt übertragen konnte. Bei unzureichender Zufriedenheit wurden die Ausschnitte hierbei

zusätzlich noch frequenzmäßig durch einen Equalizer und das Hinzufügen eines Halls bearbeitet. Für die Gewährleistung der Lippensynchronität wurden die generierten Ausschnitte außerdem auch zeitlich editiert. Unnötige Pausen, die bei der Sprachsynthese entstanden sind, wurden entfernt und die Geschwindigkeit wurde an den entsprechenden Filmausschnitt angepasst. Aufgrund der Entstehung von Artefakten bei der Geschwindigkeitsänderung konnten diese Anpassungen jedoch nur sehr klein gehalten werden. Größere Anpassungen der Geschwindigkeit mussten über die Parametereingabe bei der Sprachsynthese selbst erreicht werden (siehe Kapitel 4.1.2).

Die Ausgaben des Bark-Systems für den dritten Studienteil wurden nach dem Entfernen von Störgeräuschen noch mit dem Tool „Dialogue Countour“, welches Teil von „RX Advanced“ (iZotope Inc., 2024) ist, bearbeitet.

Abbildung 11: Benutzeroberfläche von Dialogue Contour (eigene Darstellung)

Mit Dialogue Contour ist es möglich die Satzmelodie, bzw. den Stimmhöhenverlauf von Sprache anzupassen (vgl. Abbildung 11). Diese Bearbeitung wurde unternommen, da die Stimmhöhe einer der wichtigsten Merkmale ist, wenn es um das Erkennen von Emotionen bei Sprache geht (Zhu et al., 2017, S. 3). Als Orientierung für den Stimmhöhenverlauf diente dabei immer ein entsprechender menschlicher Sprachausschnitt, welcher aus einer öffentlichen Datenbank stammte (Burkhardt et al., 2005).

Mit allen generierten, bzw. geklonten Sprachausschnitten konnte im Anschluss die Wahrnehmungsstudie in Form einer Umfrage erstellt und durchgeführt werden.

5. Wahrnehmungsstudie

Das folgende Kapitel soll sich der Wahrnehmungsstudie widmen. Dabei soll zuerst ausführlich auf die aus den Forschungsfragen abgeleiteten Hypothesen eingegangen werden und warum diese benötigt werden. Anschließend wird die eigentliche Durchführung detailliert beschrieben, indem näher auf den konkreten Aufbau der Umfrage und dessen Zusammenhang mit den jeweiligen zu testenden Hypothesen eingegangen wird. Am Ende dieses Kapitels werden schließlich die Ergebnisse der Umfrage präsentiert und auf Basis wissenschaftlicher Analysen ausgewertet und visuell dargestellt.

5.1. Hypothesenbestimmung

Die Grundlage für statistische Tests und deren Auswertung bilden Hypothesen. Diese sind v.a. wichtig, wenn es darum geht, zu bewerten ob wissenschaftliche Untersuchungen statistisch signifikante Ergebnisse liefern und sich somit verallgemeinern lassen oder nicht (Döring & Bortz, 2023, S. 646). Da sich diese Arbeit mit dem Einsatz von KI-Systemen und deren Eignung als Alternative für das ADR-Verfahren beschäftigt, ist es von Bedeutung die durchgeführte Wahrnehmungsstudie mithilfe von Hypothesentests auszuwerten, um so beweisen zu können, ob die künstliche Sprachsynthese durch die in Kapitel 4 beschriebenen KI-Systeme tatsächlich dazu in der Lage ist, das ADR-Verfahren bei der Filmtonpostproduktion zu ersetzen.

Da die Hypothesen auf Basis der Forschungsfragen formuliert werden, sind die aus der zentralen Forschungsfrage abgeleiteten Fragen im Folgenden nochmals aufgelistet:

1) Wie hoch wird die Ähnlichkeit einer KI-generierten geklonten Stimme im Vergleich zu der menschlichen Originalstimme wahrgenommen?
2) In welchem Ausmaß kann eine KI-generierte geklonte Stimme in Kombination mit einer visuellen Komponente die audiotechnischen und sprachlichen Anforderungen hinsichtlich der wahrgenommenen Natürlichkeit erfüllen?
3) Wie verlässlich werden Emotionen eines KI-generierten Dialogausschnitts im Vergleich zu der menschlichen Originalstimme erkannt?

Klassische Hypothesentests bestehen immer aus Hypothesenpaaren. Die zu prüfende Forschungshypothese ist dabei die sog. „Alternativhypothese“ (H_1). Sie wird der sog. „Nullhypothese“ (H_0) gegenübergestellt, wobei diese immer das Gegenteil der Alternativhypothese darstellt (Döring & Bortz, 2023, S. 648). Für die Formulierung der Alternativhypothesen bezogen auf den Anwendungsfall dieser Arbeit wird zunächst davon ausgegangen, dass ein KI-TTS-System das klassische ADR-Verfahren **nicht** ersetzen kann. Gründe für diese negative Annahme sind die extrem hohen Anforderungen, die ein KI-System erfüllen müsste, um den ADR-Prozess effektiv ersetzen zu können, v.a. was die Produktion von Emotionen in deutscher Sprache angeht. Wie schon in Kapitel 3.1 und 4.1 erläutert, gibt es zwar innovative KI-Systeme, die möglicherweise den Anforderungen entsprechen würden, jedoch nicht dazu in der Lage sind deutsche Ausgaben zu synthetisieren, was für die Untersuchung in dieser Arbeit jedoch einen ausschlaggebenden Punkt darstellt.

Für diese Annahme, dass ein KI-TTS-System das klassische ADR-Verfahren **nicht** ersetzen kann, lauten die Alternativhypothesen abgeleitet aus den jeweiligen Forschungsfragen wie folgt:

1) H_1: Die mittlere wahrgenommene Ähnlichkeit (μ^1) der durch Sprachsynthese generierten Sprachausschnitte ist **geringer** als die der menschlichen Ausschnitte.
2) H_1: Die mittlere wahrgenommene Natürlichkeit (μ^2) der KI-generierten Dialogausschnitte in Kombination mit der visuellen Komponente ist **geringer** als die der menschlichen Dialogausschnitte.
3) H_1: Die mittlere Genauigkeit (μ^3), mit der die Emotionen der KI-generierten Sprache erkannt wurden ist **geringer** als die der menschlichen Sprachausschnitte.

Auf Basis der formulierten Alternativhypothesen können nun die dazugehörigen Nullhypothesen formuliert werden:

1) H_0: Die mittlere wahrgenommene Ähnlichkeit (μ^1) der durch Sprachsynthese generierten Sprachausschnitte ist **nicht geringer** als die der menschlichen Ausschnitte.
2) H_0: Die mittlere wahrgenommene Natürlichkeit (μ^2) der KI-generierten Dialogausschnitte in Kombination mit der visuellen Komponente ist **nicht geringer** als die der menschlichen Dialogausschnitte.

3) H_0: Die mittlere Genauigkeit (μ^3), mit der die Emotionen der KI-generierten Sprache erkannt wurden ist **nicht geringer** als die der menschlichen Sprachausschnitte.

Für die Auswertung der Daten mithilfe verschiedener Signifikanztests ist es nun zusätzlich nötig die vorerst inhaltlich formulierten Hypothesen in statistische Hypothesen überzuführen (Döring & Bortz, 2023, S. 648).

Diese lauten im Falle dieser Arbeit wir folgt:

1) H_0: ($\mu^1_{KI} \geq \mu^1_{Mensch}$) und H_1: ($\mu^1_{KI} < \mu^1_{Mensch}$)

μ^1 gibt hierbei die mittlere wahrgenommene Ähnlichkeit der beiden Stimmausschnitte an.

2) H_0: ($\mu^2_{KI} \geq \mu^2_{Mensch}$) und H_1: ($\mu^2_{KI} < \mu^2_{Mensch}$)

μ^2 gibt hierbei die mittlere wahrgenommene Natürlichkeit der Stimmausschnitte an.

3) H_0: ($\mu^3_{KI} \geq \mu^3_{Mensch}$) und H_1: ($\mu^3_{KI} < \mu^3_{Mensch}$)

μ^3 gibt hierbei die mittlere Genauigkeit an, mit der die Emotionen erkannt wurden.

Da die wissenschaftliche Studie überprüfen soll, ob die KI-generierten Stimmausschnitte im Allgemeinen schlechter abschneiden als die menschlichen Originalstimmen, handelt es sich hierbei um einen sog. „einseitigen“ Hypothesentest. Dieser ermöglicht es gerichtete Alternativhypothesen zu formulieren, welche in der Regel eine höhere Aussagekraft haben als ungerichtete Hypothesen (Döring & Bortz, 2023, S. 649). Mit der Fertigstellung der Hypothesen kann nun die genauere Durchführung näher beschrieben werden.

5.2. Durchführung

Wie bereits in Kapitel 4.2 näher beschrieben, soll die Wahrnehmungsstudie in drei Teile geteilt werden, wobei jeder der Teile zur Bewertung einer anderen Fähigkeit dienen soll. Die Umfrage wurde online über „HdMSurvey“ (survey.hdm-stuttgart.de) durchgeführt und über mehrere E-Mail-Verteiler an die Studierenden der HdM Stuttgart weitergeleitet. Die Umfragedateien und die jeweiligen Ergebnisse befinden sich im externen Anhang.

Um einen möglichst objektiven Vergleich zwischen den originalen menschlichen Stimmausschnitten und den KI-generierten zu ermöglichen, wurden die Teilnehmer*innen softwareseitig automatisch in eine Experimental- und Kontrollgruppe

aufgeteilt. Teilnehmer*innen der Experimentalgruppe hörten den KI-generierten Inhalt, während die der Kontrollgruppe die menschlichen Ausschnitte hörten. Der Grund, warum zwei Umfragen durchgeführt wurden, ist, weil der ursprüngliche dritte Teil der Umfrage zur besseren Auswertung in Form einer zweiten Umfrage wiederholt wurde. Die erste Umfrage enthielt zur Auswertung also die Daten der ersten beiden Studienteile, während die zweite Umfrage die Daten für die Auswertung des dritten Studienteils enthielt.

5.2.1. Störvariablen

Vor der Erstellung der Umfrage müssen zunächst mögliche Störvariablen identifiziert und eliminiert werden. Diese können die interne Validität stören und somit die Aussagekraft der Untersuchung verringern (Döring & Bortz, 2023, S. 198). Dabei wird zwischen **personenbezogenen** und **untersuchungsbedingten** Störvariablen unterschieden. Mit personenbezogene Störvariablen sind u.a. persönliche Merkmale oder Voraussetzungen gemeint. Untersuchungsbedingte Störvariablen beziehen sich auf Unterschiede bei den Untersuchungsgruppen während der Durchführung der Untersuchung. Folgende Störvariablen wurden im Vorfeld der Studie identifiziert:

- Vorerfahrung der Teilnehmer*innen (personenbezogen)
- Abhörsysteme der Teilnehmer*innen (untersuchungsbedingt)
- Andere personenbezogene Störvariablen (z.B. Alter, Geschlecht)

Um diese Störvariablen möglichst zu eliminieren, wurden gewisse Anpassungen gemacht. Grundsätzlich wurde, um die personenbezogenen Störvariablen zu eliminieren, mit der Technik der Randomisierung gearbeitet. Die Teilnehmer*innen wurden zufällig einer der beiden Untersuchungsgruppen zugeteilt, mit dem Ziel, dass sich diese hinsichtlich aller psychologischen und sozialen Merkmale nicht systematisch unterscheiden (Döring & Bortz, 2023, S. 198). Zusätzlich wurden die demographischen Merkmale aller Teilnehmer*innen sowie deren bisherige Vorerfahrung in den Bereichen Audiotechnik und KI am Ende der Umfrage erfasst. Um die untersuchungsbedingte Störvariable zu eliminieren, wurden die Teilnehmer*innen einerseits am Anfang der Umfrage gebeten, für das Bewerten der Audioausschnitte Kopfhörer zu verwenden, und andererseits am Ende der Umfrage dazu aufgefordert ihr genutztes Abhörsystem anzugeben.

Da die Studie aus verschiedenen Teilen besteht, kann es zu sog. „Positionseffekten“ kommen. Diese Effekte treten z.B. auf, wenn Teilnehmer*innen ermüden, sodass auf Fragen am Anfang der Umfrage unabhängig von der Art der Frage anders reagiert wird als am Ende der Umfrage. Um diesen Effekten entgegenzuwirken, wurden sowohl die Reihenfolge der Studienteile als auch die Reihenfolge der Fragen innerhalb der verschiedenen Teile softwareseitig automatisch randomisiert.

5.2.2. Teil 1: Ähnlichkeit

Im ersten Studienteil sollten die Teilnehmer*innen bewerten, wie ähnlich sie zwei verschiedene Stimmausschnitte wahrgenommen haben. Hierzu gab es insgesamt vier verschiedene Fragen, bei denen beide Untersuchungsgruppen jeweils zwei Stimmausschnitte hörten. Die Experimentalgruppe hörte dabei sowohl einen durch IMS-Toucan geklonten Ausschnitt als auch einen anderen menschlichen Ausschnitt derselben Person, während die Kontrollgruppe zwei verschiedene menschliche Ausschnitte derselben Person hörte. Wichtig zu erwähnen ist hierbei, dass alle Stimmausschnitte der beiden Untersuchungsgruppen sowohl inhaltlich als auch auf das Geschlecht des*der Schauspieler*in bezogen identisch waren. Für die Generalisierbarkeit enthielten zwei der Fragen männliche Stimmausschnitte von Arne Löber, und im Gegensatz dazu die zwei anderen Fragen weibliche Stimmausschnitte von Franziska van der Heide. Der einzige Unterschied bestand also darin, dass die Experimentalgruppe die Ähnlichkeit zwischen einem KI-generierten und einem menschlichen Ausschnitt bewerten sollte, während die Kontrollgruppe die Ähnlichkeit zwischen zwei verschiedenen menschlichen Ausschnitten bewerten sollte. Zur Bewertung wurde hierbei der in Kapitel 4.2 erwähnte „Similarity Mean Opinion Score“ (SMOS) genutzt.

5.2.3. Teil 2: Natürlichkeit

Im zweiten Studienteil sollten die Teilnehmer*innen bewerten, wie natürlich sie einen Dialogausschnitt in Verbindung mit dem dazugehörigen Filmausschnitt wahrgenommen haben. Beide Untersuchungsgruppen sollten dabei insgesamt vier Ausschnitte bewerten, wobei die Dialogausschnitte der Experimentalgruppe KI-generiert waren, während die der Kontrollgruppe den originalen unveränderten Dialogausschnitten entsprachen. Auch hier wurden für die Generalisierbarkeit jeweils zwei Ausschnitte beider Schauspieler*innen genutzt, um sowohl die Natürlichkeit einer männlichen Stimme bewerten zu können als

auch die einer weiblichen. Die Bewertung folgte auch hier durch MOSs, wobei die Teilnehmer*innen die Natürlichkeit in den Kategorien „Satzmelodie“, „Rhythmus“, „Audioqualität“ und „Gesamteindruck“ auf einer Skala von 1 bis 5 bewerten konnten.

5.2.4. Teil 3: Emotionserkennung

Im dritten Studienteil sollten die Teilnehmer*innen auswählen, welche Emotion sie bei den gehörten Stimmausschnitten meinen erkannt zu haben. Auch hier haben die Teilnehmer*innen der Experimentalgruppe von dem KI-System Bark generierte Stimmausschnitte gehört, während die der Kontrollgruppe menschliche Ausschnitte hörten. Als Quelle für die menschlichen Ausschnitte diente „EmoDB“ (Burkhardt et al., 2005), welches ein Datensatz ist, der aus ungefähr 500 verschiedenen Sprachaufnahmen von zehn verschiedenen Sprecher*innen in jeweils sieben verschiedenen Emotionen besteht. Enthalten sind dabei die folgenden Emotionen: Freude, Wut, Angst, Langeweile, Ekel, Trauer und Neutral. Die im Datensatz enthaltenen Emotionen dienten auch als Vorlage für die Antwortmöglichkeiten dieses Umfrageteils. Wie auch schon bei den anderen Studienteilen waren die gehörten Ausschnitte bei beiden Untersuchungsgruppen sowohl inhaltlich als auch auf das Geschlecht des*der Sprecher*in bezogen identisch. Für die anschließende Auswertung wurde die Anzahl der richtig erkannten Emotionen ermittelt, woraus im Anschluss eine relative Genauigkeit berechnet werden konnte.

5.3. Datenaufbereitung und Auswertung

Vor der eigentlichen Analyse der Ergebnisse ist es nötig die Daten aufzubereiten. Dies hat verschiedene Hintergründe. Zum einen müssen die Daten, damit sie bei der anschließenden Auswertung nicht zu unzuverlässigen Ergebnissen führen, qualitativ aufgewertet werden, indem mögliche fehlerhafte Ergebnisse ausgeschlossen werden. Andererseits trägt die Datenaufbereitung zu einer reibungslosen Datenanalyse und zu weniger Verzögerungen bei (Döring & Bortz, 2023, S. 573). Für die Datenaufbereitung der beiden durchgeführten Wahrnehmungsstudien wurden die Ergebnisse exportiert und in Microsoft Excel weiterverarbeitet. Die detaillierte Beschreibung der Datenbereinigung für die einzelnen Teile der Wahrnehmungsstudie erfolgt in den jeweiligen Unterkapiteln (5.3.2ff.), um eine bessere Übersichtlichkeit zu gewährleisten.

6.3.1. Stichprobenbeschreibung

Vor der inferenzstatistischen Analyse durch klassische Signifikanztests ist es notwendig die untersuchte Personenstichprobe deskriptiv zu analysieren (Döring & Bortz, 2023, S. 605). Die Teilnehmenden wurden nach Abschluss der beiden Studien gebeten folgende Angaben zu machen:

1) Geschlecht
2) Alter
3) Verwendetes Audiosystem
4) Vorkenntnisse im Bereich der Audiotechnik (nur bei erster durchgeführter Studie)
5) Vorkenntnisse im Bereich der KI

Da die Vorkenntnisse im Bereich der Audiotechnik lediglich für die Bewertung der Natürlichkeit relevant waren, wurde auf diese Angabe bewusst bei der Durchführung der zweiten Studie verzichtet.

Die erste durchgeführte Studie hatte 90 Teilnehmer*innen, während die zweite Studie sogar 126 Teilnehmer*innen hatte. Bei beiden durchgeführten Studien war der Großteil der Teilnehmer*innen weiblich (71, bzw. 70%), während lediglich 22, bzw. 27% männlich waren (vgl. Abbildung 12). Nur ein sehr kleiner Teil der Teilnehmenden gab „Divers“ an (4, bzw. 3%) oder wollte dazu keine Angabe machen (3%).

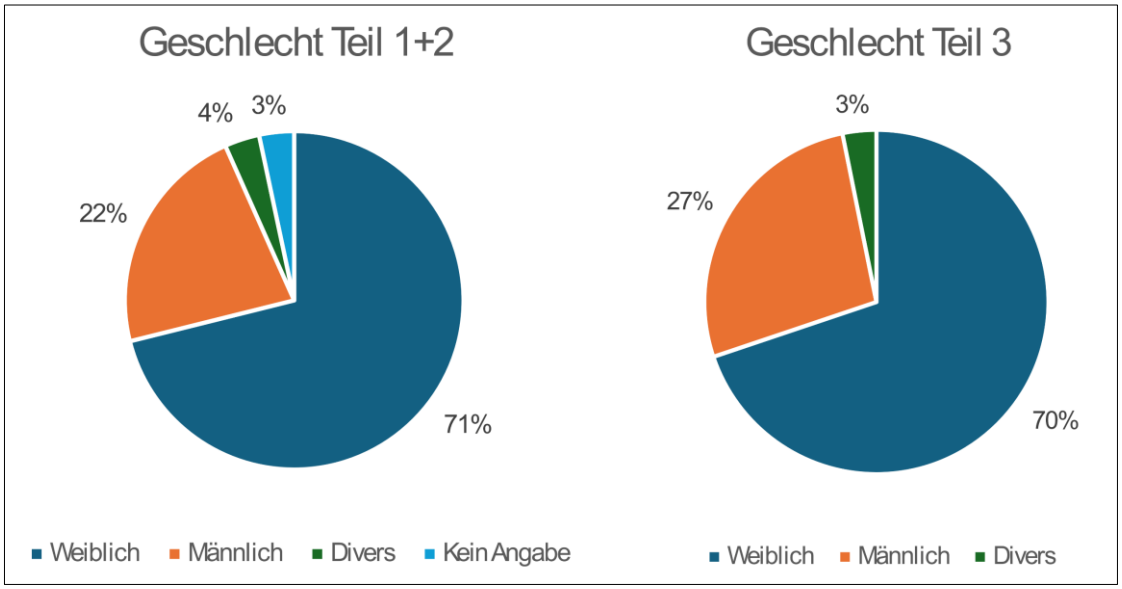

Abbildung 12: Geschlechterverteilung

Ebenfalls kaum Unterschiede gab es bei dem Anteil der verschiedenen Altersgruppen. Der Anteil der beiden Altersgruppen „18-24 Jahre“ und „25-34 Jahre“ ist nahezu identisch (vgl. Abbildung 13). Allerdings hatte die erste durchgeführte Studie zusätzlich auch Teilnehmer*innen, die der Altersgruppe „35-44 Jahre“ angehörten (1%). Am auffälligsten ist jedoch der gewogene Mittelwert, welcher bei der ersten Studie bei 23,2 Jahren liegt und bei der zweiten Studie bei 22,9 Jahren liegt. Grund für das niedrige mittlere Alter der Teilnehmer*innen könnte sein, dass die Umfrage an einer Hochschule durchgeführt wurde, wodurch v.a. jüngere Studierende als Zielgruppe angesprochen wurden.

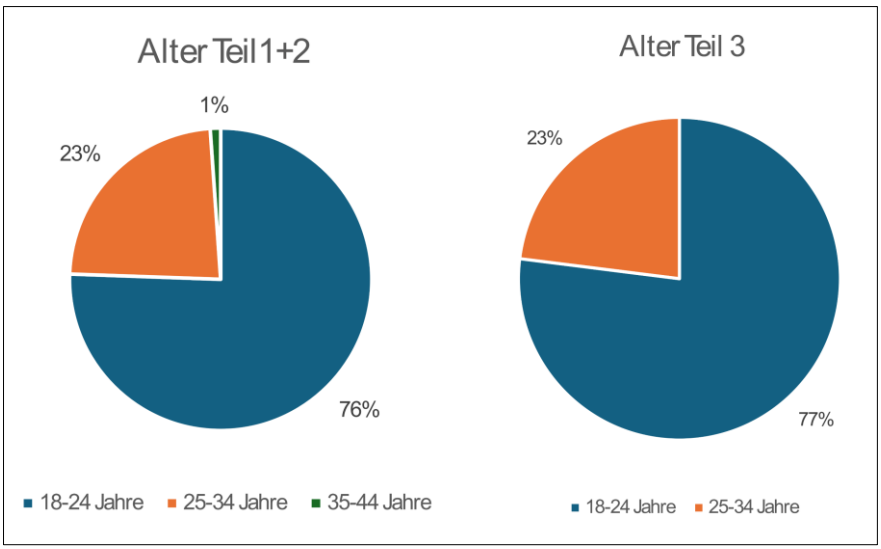

Abbildung 13: Altersanteil

Bei den genutzten Abhörsystemen gab es ebenfalls merkbare Unterschiede (vgl. Abbildung 14). Während bei der ersten Studie der größte Teil der Teilnehmenden die Audiobeispiele über einen Smartphone-/Tablet-Lautsprecher anhörte, ist bei der zweiten Studie der Anteil von Smartphone-/Tablet-Lautsprechern gegenüber Laptop-/Computer-Lautsprecher nahezu gleich. Auffällig ist ebenfalls, dass der Anteil an Teilnehmer*innen, die Kopfhörer genutzt haben, bei beiden Studien weniger als 25% beträgt, obwohl in der Umfrage ausdrücklich darum gebeten wurde Kopfhörer zu verwenden. Da das Abhörsystem eine Störvariable darstellt, ist es möglich, dass die Ergebnisse dahingehend zu einem gewissen Grad verfälscht sein könnten.

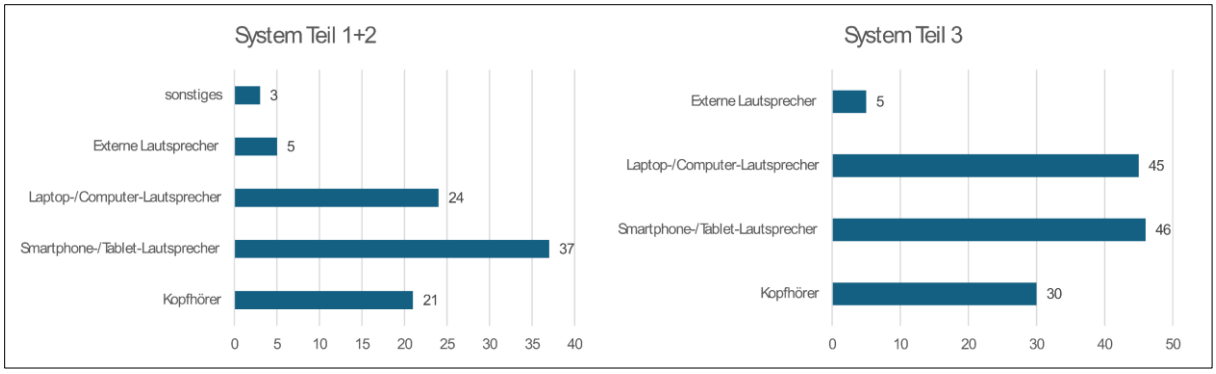

Abbildung 14: Genutzte Abhörsysteme

Die letzte Angabe, die die Teilnehmenden machen sollten, waren ihre bisherigen Vorkenntnisse in den Bereichen Audiotechnik und KI. Auffällig dabei ist, dass nur ca. ein Drittel der Teilnehmenden der ersten Studie überhaupt Kenntnisse im Bereich der Audiotechnik hatten, während im Bereich der KI bereits über der Hälfte der Teilnehmenden, sowohl bei der ersten als auch bei der zweiten Studie, Kenntnisse vorweisen konnten (vgl. Abbildung 15). Zudem ist der Anteil der Teilnehmenden, die angaben, fortgeschrittene oder sogar Expertenkenntnisse zu haben, bei beiden durchgeführten Studien eher gering. Wenn man jedoch in Betracht zieht, dass die Zielgruppen der meisten Filmproduktionen, bei denen in der Theorie KI-Systeme als Ersatz für ADR angewendet werden könnten, ebenfalls zu großen Teilen keine oder nur wenige Kenntnisse in den Bereichen Audiotechnik und KI vorweisen können, ist der Mangel an fortgeschrittenen und Expertenkenntnissen nicht als negativ zu bewerten.

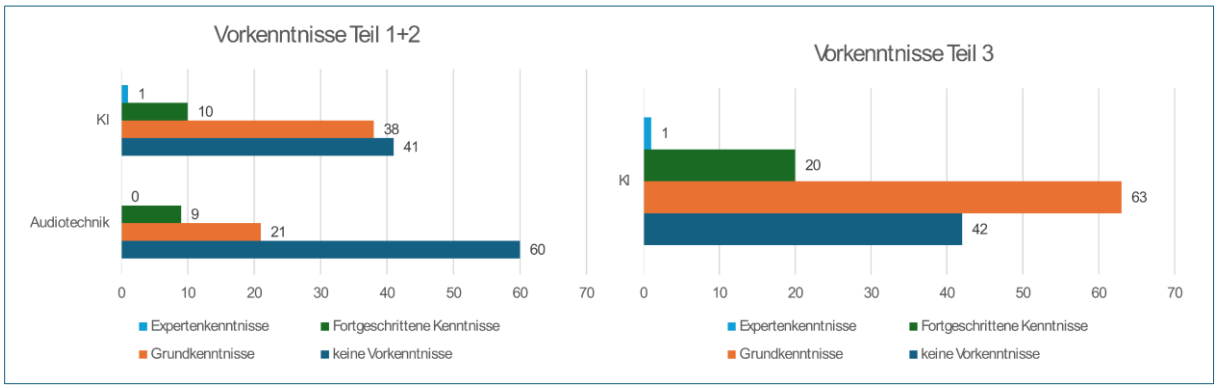

Abbildung 15: Vorkenntnisse in den Bereichen Audiotechnik und KI

Nachdem nun die Personenstichprobe näher beschrieben wurde, kann nun mit der Datenbereinigung und der eigentlichen inferenzstatistischen Analyse mithilfe von

verschiedenen Signifikanztests begonnen werden. Zur besseren Übersichtlichkeit werden diese für jeden thematischen Studienteil in separaten Kapiteln dargestellt.

6.3.2. Teil 1: Ähnlichkeit

Im ersten Teil der Studie sollten die Teilnehmer*innen subjektiv bewerten, wie ähnlich sie zwei verschiedene Stimmausschnitte wahrgenommen haben. Teilnehmer*innen der Experimentalgruppe hörten dabei jeweils einen durch die KI-geklonten Ausschnitt und einen Originalausschnitt der dazugehörigen menschlichen Stimme. Im Vergleich dazu hörten die Teilnehmer*innen der Kontrollgruppe jeweils zwei verschiedene Stimmausschnitte der jeweils gleichen Person. Ziel dieses Studienteils war es also zu testen, ob das gewählte KI-System dazu in der Lage ist menschliche Stimmen effektiv zu klonen. Die Ähnlichkeit der beiden zu vergleichenden Stimmausschnitte sollte im Anschluss von den Teilnehmer*innen auf einer Skala von 1 (überhaupt nicht ähnlich) bis 5 (sehr ähnlich) bewertet werden, um so eine statistische Repräsentation dieser wahrgenommenen Ähnlichkeit zu erhalten.

Für die Datenaufbereitung wurden zunächst alle erhaltenen Werte in jeweils einen Durchschnittswert pro Frage für Experimental- und Kontrollgruppe zusammengefasst. Dies ist nötig, um mit den Daten weitere Tests durchführen zu können. Da der Studienteil jeweils vier verschiedene Fragen in beiden Gruppen enthielt, beläuft sich die Gesamtanzahl der Durchschnittswerte also insgesamt auf acht.

Frage	Experimentalgruppe	Kontrollgruppe
1 (M)	1,8372093	2,46808511
2 (F)	2,44186047	2,34042553
3 (M)	2,25581395	3,63829787
4 (F)	2,1627907	2,55319149
Mittelwert + Standardabweichung	2,174 (±0,253)	2,75 (±0,599)

Tabelle 1: Durchschnittliche wahrgenommene Ähnlichkeit, „M“, bzw. „F“ stehen dabei für das Geschlecht der zu bewertenden Stimmausschnitte

Die durchschnittlichen Bewertungen der Ähnlichkeiten sind in Tabelle 1 dargestellt. Schon hier ist zu erkennen, dass die durchschnittliche Ähnlichkeit der Experimentalgruppe eher geringer ist als die der Kontrollgruppe. Um eine konkrete Aussage über die Annahme oder die Ablehnung der aufgestellten Hypothese treffen zu können müssen die Daten noch durch einen Signifikanztest weiter analysiert werden (Döring & Bortz, 2023, S. 606). Je nach vorhandener Datenlage kommt jedoch nicht jeder

Signifikanztest hierzu in Frage. Für die Wahl des passenden Signifikanztests (die sog. Indikation) müssen bestimmte statistische Voraussetzungen geprüft werden (Döring & Bortz, 2023, S. 606). Zunächst muss dabei die Anzahl der Ausprägungen der unabhängigen Variable beachtet werden (Schuff et al., 2023, S. 1213). Im Falle der im Rahmen dieser Arbeit durchgeführten Studie beträgt die Anzahl der Ausprägungen zwei, da zwischen einer Experimental- und einer Kontrollgruppe unterschieden wird. Im Anschluss muss evaluiert werden, ob ein gepaarter oder ungepaarter Signifikanztest in Frage kommt (Schuff et al., 2023, S. 1213). Im Falle der hier vorliegenden Studie sind die zu bewertenden Stimmausschnitte der beiden Gruppen miteinander verknüpft, da die Stimmausschnitte - unabhängig von der Verwendung von KI - sowohl inhaltlich als auch in Bezug auf die sprechende Person identisch waren. Somit würde nur ein gepaarter Signifikanztest in Frage kommen. Im nächsten Schritt muss überprüft werden, ob die Ergebnisse der Studie einer Normalverteilung entsprechen oder nicht. Zur Überprüfung der Verteilung eignet sich der „Shapiro-Wilk Test“ (Shapiro & Wilk, 1965).

Shapiro-Wilk Test	Experimentalgruppe	Kontrollgruppe
p-Wert	0.9893	0.09585
normalverteilt?	Ja	Ja

Tabelle 2: Ergebnisse des Shapiro-Wilk Tests

Beim Shapiro-Wilk Test wird durch die Berechnung eines p-Wertes bewertet, ob die vorliegenden Daten normalverteilt sind oder nicht. Der p-Wert gibt dabei an, wie wahrscheinlich es ist, eine solche vorliegende Stichprobe zu ziehen, wenn sie aus einer Normalverteilung stammt. Je höher dieser Wert also ist, desto eher sind die Daten normalverteilt. Der Test geht davon aus, dass wenn der Wert ein Signifikanzniveau von 0,05 überschreitet, von einer Normalverteilung der Daten ausgegangen werden kann. Da dieser Wert sowohl bei der Experimentalgruppe als auch bei der Kontrollgruppe oberhalb des Signifikanzniveaus liegt, kann bei beiden Gruppen von einer Normalverteilung ausgegangen werden. In einem solchen Fall ist es angebracht einen gepaarten t-Test als Signifikanztest durchzuführen, um zu überprüfen, ob die Nullhypothese dieses Studienteils abgelehnt wird oder nicht (Schuff et al., 2023, S. 1213).

gepaarter t-Test	Ergebnisse
Freiheitsgrade (df)	3
t-Statistik	-1,861894293
p(T<=t)	0,079764966
Kritischer t-Wert	2,353363435

Tabelle 3: Ergebnisse des gepaarten t-Tests

Die Nullhypothese und Alternativhypothese für diesen Studienteil waren folgende:

$$\mathbf{H_0: (\mu^1_{KI} \geq \mu^1_{Mensch})\ und\ H_1: (\mu^1_{KI} < \mu^1_{Mensch}),}$$

μ^1 entspricht dabei der mittleren Ähnlichkeit zwischen den beiden Stimmausschnitten.

In Tabelle 3 sind die Ergebnisse des gepaarten t-Tests aufgezeigt. Die berechneten Werte werden im Folgenden nochmals in Kürze erläutert.

Die Freiheitsgrade ergeben sich aus *Stichprobengröße-1*. Da die Stichprobengröße in diesem Fall bei 4 lag, ergibt sich ein Freiheitsgrad von 3. Aus dem Freiheitsgrad, dem Signifikanzniveau (welches in diesem Fall 0,05 betrug) und der Art des Tests (in diesem Fall aufgrund der gerichteten Hypothese einseitig) lässt sich durch Verwendung der t-Verteilungstabelle der kritische t-Wert ermitteln (Bobbitt, 2020). Mithilfe des Betrags der berechneten t-Statistik und dem p-Wert lässt sich nun eine Aussage über die Ablehnung der Nullhypothese treffen. Da der p-Wert **größer** als das Signifikanzniveau ist und der Betrag der t-Statistik **kleiner** als der kritische t-Wert ist, kann die Nullhypothese **nicht** abgelehnt werden und das Ergebnis ist **nicht** statistisch signifikant.

Zusammenfassend kann die Nullhypothese also **nicht** abgelehnt werden und es lassen sich somit auch keine signifikanten Aussagen über die wahrgenommene Ähnlichkeit der KI-generierten und der originalen Stimmausschnitte tätigen.

6.3.3. Teil 2: Natürlichkeit

Im zweiten Teil der Studie sollten die Teilnehmer*innen die wahrgenommene Natürlichkeit eines Dialogausschnitts in Verbindung mit dem dazugehörigen Filmausschnitt bewerten. Dabei sollten folgende Kategorien im Hinblick auf Natürlichkeit auf einer Skala von 1 (überhaupt nicht natürlich) bis 5 (sehr natürlich) bewertet werden:

1) Satzmelodie
2) Rhythmus
3) Audioqualität
4) Gesamteindruck

Teilnehmer*innen der Experimentalgruppe hörten Dialogausschnitte, die KI-generiert, bzw. geklont waren, während die Kontrollgruppe die originalen Ausschnitte zur Bewertung hörte. Ziel dieses Studienteils war es also zu testen, ob das KI-System dazu in der Lage ist, natürlich klingende Dialoge in Verbindung mit dem dazugehörigen visuellen Ausschnitt zu generieren. Dieser Studienteil kombiniert also auditive und visuelle Komponenten und kommt so dem klassischen ADR-Verfahren sehr nahe, weshalb die Ergebnisse der Datenanalyse eine zentrale Rolle bei der Beantwortung der Forschungsfrage spielen.

Zur Datenaufbereitung wurden auch hier Durchschnittswerte berechnet, um die Daten für die Signifikanztests vorzubereiten. Zunächst wurde für jede bewertete Kategorie (Satzmelodie, Rhythmus, Audioqualität, Gesamteindruck) ein durchschnittlicher Wert pro Frage ermittelt. Anschließend wurden diese Werte nochmals zu einem Durchschnittswert kombiniert, um pro Frage einen Gesamtdurchschnitt zu erhalten, der alle Kategorien zusammenfasst. Auch hier beläuft sich die Gesamtanzahl der Durchschnittswerte analog zum ersten Studienteil auf insgesamt acht.

Frage	Experimentalgruppe	Kontrollgruppe
1 (F)	2,20930233	4,10638298
2 (M)	3,25	3,93085106
3 (F)	2,83139535	3,36702128
4 (M)	2,95348837	3,53191489
Mittelwert + Standardabweichung	2,811 (±0,438)	3,734 (±0,343)

Tabelle 4: Durchschnittliche wahrgenommene Natürlichkeit, „M“, bzw. „F“ stehen dabei für das Geschlecht der zu bewertenden Stimmausschnitte

Die Ergebnisse des zweiten Studienteils sind in Tabelle 4 zu entnehmen. Auch hier ist schon eine Tendenz zu erkennen: Die wahrgenommene Natürlichkeit der Experimentalgruppe ist bei jeder gegebenen Frage geringer als bei der Kontrollgruppe. Um jedoch beurteilen zu können, ob diese Beobachtung statistisch signifikant ist, muss ebenfalls ein Signifikanztest angewendet werden. Hierfür muss zunächst wieder eine

Indikation durchgeführt werden, um zu überprüfen, welcher Signifikanztest für die hier gezeigte Datenlage in Frage kommt.

Hierbei muss zunächst wieder überprüft werden, wie viele Ausprägungen die unabhängige Variable hat, bevor überprüft werden muss, ob ein gepaarter oder ungepaarter Signifikanztest in Frage kommt und ob die Daten normalverteilt sind (Schuff et al., 2023, S. 1213). Da der zweite Studienteil dieselbe Struktur wie der erste Teil aufweist, besitzt die unabhängige Variable auch hier zwei Ausprägungen und es kommt aufgrund der inhaltlichen Verbindung der Beispiele aus Experimental- und Kontrollgruppe auch ein gepaarter Signifikanztest in Frage. Im Anschluss muss nun auf Normalverteilung getestet werden, was auch in diesem Studienteil mit dem Shapiro-Wilk Test bewerkstelligt wird.

Shapiro-Wilk Test	Experimentalgruppe	Kontrollgruppe
p-Wert	0.8801	0.8576
normalverteilt?	Ja	Ja

Tabelle 5: Ergebnisse des Shapiro-Wilk Tests

Wie in Tabelle 5 zu entnehmen ist, liegt der p-Wert beider getesteten Gruppen deutlich oberhalb des Signifikanzniveaus von 0,05. Deshalb kann davon ausgegangen werden, dass die Daten beider Gruppen normalverteilt sind, wodurch sich auch in diesem zweiten Studienteil eine Datenanalyse mit einem gepaarten t-Test anbieten würde, um die mögliche Ablehnung der Nullhypothese zu ermitteln.

Die Nullhypothese und Alternativhypothese für diesen Studienteil waren folgende:

$$\mathbf{H_0: (\mu^2_{KI} \geq \mu^2_{Mensch})\ und\ H_1: (\mu^2_{KI} < \mu^2_{Mensch}),}$$

μ^2 entspricht dabei der mittleren wahrgenommenen Natürlichkeit der Dialog- bzw. Filmausschnitte.

gepaarter t-Test	Ergebnisse
Freiheitsgrade (df)	3
t-Statistik	-2,830225829
p(T<=t)	0,033088673
Kritischer t-Wert	2,353363435

Tabelle 6: Ergebnisse des gepaarten t-Tests

In Tabelle 6 sind die Ergebnisse des gepaarten t-Tests aufgezeigt. Aufgrund der Stichprobengröße von vier, beträgt auch in diesem Teil der Studie der Freiheitsgrad drei. Auch hier lässt sich durch den Betrag der t-Statistik und den berechneten p-Wert eine Aussage über die Ablehnung der Nullhypothese tätigen. Da der p-Wert **kleiner** als das Signifikanzniveau (0,05) ist und der Betrag der t-Statistik **größer** als der kritische t-Wert ist, kann die Nullhypothese **abgelehnt** werden und das Ergebnis ist somit **statistisch signifikant.** Dies würde im Kontext der Hypothese also bedeuten, dass die wahrgenommene Natürlichkeit der KI-generierten Dialogausschnitte schlechter als die der echten Dialoge ist.

Um die Bedeutsamkeit des Signifikanztests zu bewerten, muss jedoch bei einem statistisch signifikanten Ergebnis immer zusätzlich auch die sog. Effektgröße bestimmt werden (Döring & Bortz, 2023, S. 655). Eines der bekanntesten Maße zur Berechnung der Effektgröße ist das sog. „*d*-Maß“ von Cohen (Cohen, 1988). Im Falle eines gepaarten t-Tests kann Cohen's *d* über folgende Formel berechnet werden (Lakens, 2013, S. 4, zitiert nach Rosenthal, 1991):

$$Cohen's\ d = \frac{t}{\sqrt{n}}$$

Das „t“ steht dabei für den Wert der t-Statistik, bzw. dessen Betrag, während das „n“ für die Größe der Stichprobe steht. Setzt man nun die durch den t-Test berechneten Werte ein, erhält man folgende Effektgröße:

$$Cohen's\ d = \frac{2{,}830225829}{\sqrt{4}} = 1{,}41511291$$

Ab einem Wert von 0,8 kann bei der Effektgröße von einem großen Effekt gesprochen werden (Döring & Bortz, 2023, S. 804), wodurch sich im Falle der hier durchgeführten Studie ein großer Effekt ergeben würde.

6.3.4. Teil 3: Emotionserkennung

Im dritten Teil der Studie hörten die Teilnehmer*innen verschiedene Stimmausschnitte und sollten die von ihnen wahrgenommene Emotion angeben. Teilnehmer*innen der Kontrollgruppe hörten Ausschnitte aus dem Datensatz „EmoDB“ (Burkhardt et al., 2005), welcher Sprachaufnahmen für diverse Emotionen enthält (Freude, Trauer, Langeweile,

Ekel, Wut, Angst, Neutral). Jede Emotion wurde in der Umfrage durch eine Sprachaufnahme repräsentiert. Die Experimentalgruppe hörte analog dazu KI-generierte Ausschnitte, welche hinsichtlich der Emotion, des Inhalts und des Geschlechts des*der Sprecher*in identisch mit dem jeweiligen Gegenstück der Kontrollgruppe waren. Ziel dieses Studienteils war es also zu testen, ob das KI-System dazu in der Lage ist Sprachaufnahmen mit menschlichen Emotionen zu erzeugen und wie genau diese von den Hörer*innen erkannt werden können.

Zur Datenaufbereitung wurde die Anzahl der korrekt identifizierten Emotionen beider untersuchten Gruppen gezählt und daraus im Anschluss eine Genauigkeit ermittelt. Da der verfügbare Datensatz sieben Emotionen enthielt und jede Emotion in der Umfrage durch ein Beispiel repräsentiert wurde, ergeben sich also insgesamt sieben Werte für die Genauigkeit. Da es innerhalb der Umfrage ebenfalls die Möglichkeit gab, keine der sieben Emotionen auszuwählen und „unsicher“ anzugeben wurden alle Antworten, die auf diese Art beantwortet wurden ignoriert, um Verfälschungen zu vermeiden.

Genauigkeit	absolut		relativ	
Emotion	Experimentalgruppe	Kontrollgruppe.	Experimentalgruppe	Kontrollgruppe
Wut	28	66	0,56	1
Langeweile	5	64	0,087719298	0,96969697
Angst	3	60	0,054545455	0,983606557
Trauer	4	54	0,075471698	0,857142857
Freude	35	59	0,636363636	0,9
Neutral	35	54	0,636363636	0,870967742
Ekel	0	63	0	1
Mittelwert + Standardabweichung	15,7 (±16,1)	60 (±4,726)	0,293 (±0,3)	0,940 (±0,062)

Tabelle 7: Genauigkeit der erkannten Emotionen, sowohl absolut als auch relativ zur Gruppengröße

In Tabelle 7 sind die Anzahl der korrekt erkannten Emotionen, bzw. die Genauigkeit, mit der die Emotion identifiziert wurden zu entnehmen. Die absolute Genauigkeit gibt an, wie viele Teilnehmer*innen aus der jeweiligen Gruppe die richtige Emotion erkannt haben. Die relative Genauigkeit setzt diese absolute Zahl ins Verhältnis zur Gesamtzahl der jeweiligen Gruppenteilnehmer*innen. Dies ist notwendig, da die Gruppengröße unterschiedlich war (Experimentalgruppe 60 Teilnehmer*innen, Kontrollgruppe 66 Teilnehmer*innen).

Wie auch schon bei den anderen beiden Studienteilen lässt sich hier ebenfalls schon eine Tendenz feststellen: Die Genauigkeit, mit der die Emotionen innerhalb der Experimentalgruppe identifiziert wurden, ist bei allen Emotionen deutlich geringer als in der Kontrollgruppe. Um die statistische Signifikanz dieses Studienteils zu ermitteln, muss jedoch auch hier ein Signifikanztest zum Einsatz kommen, weshalb zunächst mit der Indikation begonnen wird.

Wie auch bei den vorherigen Studienteilen besitzt die unabhängige Variable aufgrund der beiden verschiedenen Untersuchungsgruppen zwei Ausprägungen. Es muss ebenfalls ein gepaarter Signifikanztest zum Einsatz kommen, da die Ausschnitte der beiden Gruppen sowohl was den Inhalt als auch die Emotion und das Geschlecht des*der Sprecher*in angeht, identisch sind. Anschließend müssen die Daten der Experimental- und Kontrollgruppe auf ihre Normalverteilung überprüft werden. Da die absolute Genauigkeit die Gruppengröße nicht berücksichtigt, wird für die weiteren Tests ausschließlich die relative Genauigkeit verwendet.

Shapiro-Wilk Test	Experimentalgruppe	Kontrollgruppe
p-Wert	0.02487	0.1186
normalverteilt?	Nein	Ja

Tabelle 8: Ergebnisse des Shapiro-Wilk Tests

Nach Durchführen des Shapiro-Wilk Tests (Ergebnisse in Tabelle 8 zu entnehmen) konnte festgestellt werden, dass die Kontrollgruppe mit einem p-Wert oberhalb des Signifikanzniveaus (0,05) normalverteilt ist, der p-Wert der Experimentalgruppe jedoch deutlich unterhalb des Signifikanzniveaus liegt, weshalb nicht von einer Normalverteilung ausgegangen werden kann. Anders als bei den beiden vorherigen Studienteilen kann in diesem Fall also kein gepaarter t-Test zur Auswertung herangezogen werden. In diesem Fall muss der sog. „Wilcoxon-Vorzeichen-Rang-Test“ (Wilcoxon, 1945) angewendet werden, da dieser als Alternative für den t-Test gilt, falls keine Normalverteilung vorliegen sollte (Schuff et al., 2023, S. 1214).

Die Nullhypothese und Alternativhypothese für diesen Studienteil waren folgende:

$$\mathbf{H_0: (\mu^3_{KI} \geq \mu^3_{Mensch})\ und\ H_1: (\mu^3_{KI} < \mu^3_{Mensch}),}$$

μ^3 entspricht dabei der Genauigkeit, mit der die korrekten Emotionen identifiziert wurden.

Wilcoxon-Vorzeichen-Rang-Test					
Emotion	Differenz	Absolute Differenz	Rangfolge der abs. Differenz	Positive Ränge	Negative Ränge
Freude	-0,263636364	0,263636364	7		7
Neutral	-0,287634409	0,287634409	6		6
Wut	-0,44	0,44	5		5
Trauer	-0,781671159	0,781671159	4		4
Langeweile	-0,881977671	0,881977671	3		3
Angst	-0,929061103	0,929061103	2		2
Ekel	-0,954545455	0,954545455	1		1

Tabelle 9: Ergebnisse des Wilcoxon-Vorzeichen-Rang-Tests

In Tabelle 9 sind Teile der Ergebnisse des Wilcoxon-Vorzeichen-Rang-Tests zu entnehmen. Die Ergebnisse sind dabei nach dem Wert der absoluten Differenz sortiert, angefangen mit dem geringsten. Der Test basiert auf der Erstellung einer Rangfolge der absoluten Differenzen zwischen den beiden untersuchten Gruppen. Im ersten Schritt wird die Differenz zwischen Experimental- und Kontrollgruppe berechnet, bevor diese in die absolute Differenz, also den Betrag, umgewandelt wird. Auf Basis dieser absoluten Differenzen wird im Anschluss eine Rangfolge erstellt, welche im nächsten Schritt in „positive“ und „negative“ Ränge eingeteilt werden. Ob die Ränge als positiv oder negativ kategorisiert werden hängt vom Vorzeichen der im ersten Schritt berechneten Differenz zwischen den beiden Untersuchungsgruppen ab (Bobbitt, 2021). Da die Differenz im Falle der hier durchgeführten Studie jedoch bei jedem Datenpunkt negativ ist, wurden alle Ränge als negativ klassifiziert. Um aus diesen Rängen nun den Wert der t-Statistik zu berechnen, werden alle positiven Ränge aufsummiert und im Anschluss alle negativen.

$$W^+ = 0$$

$$W^- = 1 + 2 + 3 + 4 + 5 + 6 + 7 = 28$$

Der kleinere dieser beiden Werten (in diesem Fall W^+) wird nun zuletzt noch mit dem kritischen t-Wert verglichen, welcher in einer eigens dafür angelegten Tabelle zu finden ist (Bobbitt, 2021). Aus der Tabelle lässt sich entnehmen, dass W^{krit} bei einer Stichprobengröße von 7 und einem Signifikanzniveau von 0,05 den Wert 2 erhält.

Somit ergibt sich folgendes:

$$W^t < W^{krit}$$

Da der Wert der t-Statistik **kleiner** als der kritische Wert ist, kann die Nullhypothese **abgelehnt** werden und das Ergebnis ist somit **statistisch signifikant** (Bobbitt, 2021).

Wie auch schon im zweiten Studienteil muss jedoch bei einem statistisch signifikanten Ergebnis zusätzlich noch die Effektgröße berechnet werden (Döring & Bortz, 2023, S. 655). Da die Daten jedoch nicht normalverteilt waren, ist die Berechnung der Effektgröße mit Cohen's *d* eher ungeeignet und eine Berechnung mit „Cliff's Delta (δ)" (Cliff, 1993) verlässlicher (Meissel & Yao, 2024, S. 2).

Für die Berechnung des Cliff's Delta Werts wurde „CliffsDelta" (Ernst, 2021) in Python verwendet und kam zu folgendem Ergebnis:

$$\delta = -1$$

Der Cliff's Delta Wert vergleicht die beiden Untersuchungsgruppen und analysiert, wie sehr sich beide Verteilungen überlappen. Er kann Werte zwischen -1 und 1 annehmen, wobei die beiden Extremwerte eine vollständige Nicht-Überlappung anzeigen, während der Wert 0 auf eine vollständige Überlappung hinweist (Meissel & Yao, 2024, S. 2-3). Der für diesen Studienteil berechnete Wert von -1 deutet also auf keinerlei Überlappung hin, weshalb von einer hohen Effektgröße ausgegangen werden kann.

Nach der Auswertung aller drei Studienteile können also zwei der drei Nullhypothesen mit Sicherheit abgelehnt werden, während die Nullhypothese des ersten Studienteils nicht abgelehnt werden kann. Welche Bedeutung diese Ergebnisse in Bezug auf die inhaltlichen Hypothesen und die Forschungsfragen haben wird im nächsten Kapitel ausführlich erläutert.

6. Diskussion

In den folgenden Kapiteln sollen die ausgewerteten Ergebnisse der Wahrnehmungsstudie nochmals zusammengefasst werden und im Kontext der Forschungsfrage interpretiert werden. Im Anschluss sollen daraus mögliche Stärken und Schwächen der KI-Sprachsynthese, v.a. im Bezug auf die Anwendung als möglicher Ersatz für das ADR-Verfahren, gezogen werden. Mithilfe dieser Stärken und Schwächen und den Ergebnissen der Wahrnehmungsstudie soll am Ende dieses Kapitels schließlich die Forschungsfrage beantwortet werden.

6.1. Interpretation der Ergebnisse

In Kapitel 1.1 wurden drei zentrale Anforderungen an die Fähigkeiten eines KI-TTS-Modells gestellt, auf deren Grundlage die ausgewählten KI-Systeme im Rahmen einer Wahrnehmungsstudie getestet wurden. Im Rahmen dieser Studie wurde ausgewertet, wie gut die gewählten KI-Systeme Stimmen klonen, natürliche Ausgaben erzeugen und kontrolliert emotionale Sprache generieren können.

Im ersten Studienteil konnten durch die Auswertung keine statistisch signifikanten Ergebnisse erzielt werden. Zwar konnte die Nullhypothese damit nicht abgelehnt werden, allerdings zeigt sich durch den Vergleich der beiden Untersuchungsgruppen, dass die bewertete Ähnlichkeit in der Experimentalgruppe im Durchschnitt bei allen Teilfragen schlechter bewertet wurde. Dies würde also bedeuten, dass das genutzte KI-System die Anforderung in Bezug auf das Klonen der Stimmen nicht erfüllen konnte. Allerdings ist die bewertete Ähnlichkeit in der Kontrollgruppe mit einem durchschnittlichen Wert von 2,75 ebenfalls gering ausgefallen. Mögliche Gründe für diesen geringen Wert könnten die Auswahl der zu vergleichenden Ausschnitte oder die unterschiedliche subjektive Wahrnehmung der Ähnlichkeit durch die Teilnehmer*innen sein. Da die statistische Auswertung zu einem nicht signifikanten Ergebnis kam, lassen sich für diese Untersuchung jedoch keine genaueren Aussagen tätigen. Um zu präziseren Ergebnissen zu kommen wären weitere Untersuchungen notwendig.

Im zweiten Studienteil konnte die Nullhypothese abgelehnt werden. Durch die Auswertung und die Ermittlung einer sehr hohen Effektgröße lässt sich demnach schlussfolgern, dass das genutzte KI-System nicht dazu in der Lage ist, natürlich

klingende Sprache zu erzeugen und dadurch im Rahmen der Studie deutlich schlechter abschnitt als die menschlichen Originalausschnitte.

Auch im dritten Teil der Studie konnte die Nullhypothese abgelehnt werden und eine hohe Effektgröße ermittelt werden. Das genutzte KI-System konnte hier also den Anforderungen nicht gerecht werden und keine Emotionen kontrolliert generieren. Was jedoch stark auffällt sind die Unterschiede in der Genauigkeit, mit der die Emotionen von den Teilnehmer*innen erkannt wurden. Während Emotionen wie „Ekel“ oder „Angst“ nur von sehr wenigen Teilnehmer*innen der Experimentalgruppe erkannt wurde, weisen die Emotionen „Freude“ und „Wut“ eine deutlich höhere Genauigkeit von 0,64, bzw. 0,56 auf. Grund dafür könnte einerseits sein, dass das KI-System bestimmte Emotionen möglicherweise verlässlicher erzeugen kann, andererseits könnte die Begründung auch darin liegen, dass nicht alle Emotionen gleich einfach zu erkennen sind (Mozziconacci, 1998, S. 168). Mit Blick auf die Kontrollgruppe lässt sich die zweite mögliche Begründung jedoch größtenteils ausschließen, da hier wiederum andere Emotionen eine höhere Genauigkeit aufzeigen. Um hier also zu aussagekräftigeren Ergebnissen zu kommen, wären auch für diesen Studienteil weitere Untersuchungen notwendig.

6.2. Stärken und Schwächen der KI-Sprachsynthese

Im Folgenden soll auf die aus der Wahrnehmungsstudie resultierenden Stärken und Schwächen der KI-TTS-Systeme eingegangen werden. Diese sollen zum einen identifiziert und zum anderen in den Kontext des in dieser Arbeit untersuchten Anwendungsfalls, nämlich den Ersatz des ADR-Verfahrens, eingeordnet werden.

6.3.5. Stärken

Stärken der KI-Sprachsynthese als Ersatz für das ADR-Verfahren sind auf jeden Fall die hohe Flexibilität und die damit verbundene Kosten- und Zeitersparnis. Ein TTS-System ermöglicht es, Texte schnell und unkompliziert zu generieren oder zu ändern und das Wechseln zwischen verschiedenem zu klonendem Stimmen ist sehr unkompliziert, da für das Klonen lediglich kurze Sprachaufnahmen von wenigen Sekunden erforderlich sind. Somit wäre der Einsatz eine KI-TT-Systems optimal für Projekte mit kleinem Budget oder knappen Zeitplänen.

Zusätzlich würde auch eine hohe Wiederverwendbarkeit für den Einsatz von KI-Systemen sprechen. Sobald Modelle trainiert und gefinetuned sind, können diese ohne großen Mehraufwand für verschiedenste Projekte wieder genutzt werden.

6.3.6. Schwächen

Die Schwächen von KI-TTS-Systemen haben sich v.a. bei der Wahrnehmungsstudie gezeigt. So können die verwendeten Systeme trotz vielen Fortschritten im Bereich der Sprachsynthese noch keine Sprache mit der erwünschten Natürlichkeit und den gewünschten Emotionen generieren. Diese beiden Aspekte wären für ein KI-System, welches den klassischen ADR-Prozess ersetzen soll, jedoch unerlässlich, da durch mangelnde Natürlichkeit und Emotionen das Zuschauererlebnis erheblich beeinträchtigt werden würde.

Außerdem würde die letztendliche Qualität der Sprachsynthese sehr von den vorhandenen Daten für das Finetuning abhängen. So könnte es unter Umständen vorkommen, dass die Stimmen gewisser Schauspieler*innen besser und natürlicher geklont werden könnten als die Stimmen anderer Schauspieler*innen.

Die ethischen Aspekte sind bei der Verwendung von KI ebenfalls zu bedenken. Für den gesamten ADR-Prozess, der den Einsatz von KI umfasst, wäre nicht nur die ausdrückliche Zustimmung der Schauspieler*innen erforderlich, sondern es gäbe auch ein großes Risiko für den Missbrauch eines solchen Systems. Somit wäre einerseits der komplette ADR-Prozess, sofern dieser durch eine KI durchgeführt wird, allein von der Zustimmung der jeweiligen zu klonenden Person abhängig, während andererseits die Ausgaben solcher KI-Systeme mit speziellen Wasserzeichen ausgestattet werden müssten, um potenziellen Missbrauch zu verhindern.

Insgesamt zeigt sich also, dass ein KI-TTS-System zwar Stärken mit sich bringt, aber im Moment jedoch deutlich mehr Schwächen aufzeigt, v.a. was die erwünschte Qualität der Ausgaben betrifft.

6.3. Beantwortung der Forschungsfrage

Im Folgenden soll auf Basis der in dieser Arbeit gewonnenen Erkenntnissen die zentrale Forschungsfrage beantwortet werden.

Diese lautete wie folgt:

Inwiefern kann künstliche Sprachsynthese durch ein KI-TTS-Modell die menschliche Stimme während des ADR-Prozesses einer Filmtonpostproduktion ersetzen?

Diese wurde für eine präzisere Beantwortung durch drei spezifischere Fragen ergänzt:

1) Wie hoch wird die Ähnlichkeit einer KI-generierten geklonten Stimme im Vergleich zu der menschlichen Originalstimme wahrgenommen?
2) In welchem Ausmaß kann eine KI-generierte geklonte Stimme in Kombination mit einer visuellen Komponente die audiotechnischen und sprachlichen Anforderungen hinsichtlich der wahrgenommenen Natürlichkeit erfüllen?
3) Wie verlässlich werden Emotionen eines KI-generierten Dialogausschnitts im Vergleich zu der menschlichen Originalstimme von Testhörer*innen erkannt?

Diese Fragen werden zunächst im Einzelnen beantwortet, bevor eine abschließende Bewertung vorgenommen wird, um die zentrale Forschungsfrage zu beantworten:

1) Die Ähnlichkeit einer KI-generierten geklonten Stimme wurde im Vergleich zu der menschlichen Originalstimme als eher gering wahrgenommen. Dennoch lassen sich aufgrund der statistisch nicht signifikanten Ergebnisse der Wahrnehmungsstudie und der ebenfalls unerwartet niedrigen Ähnlichkeit in der Kontrollgruppe nur begrenzte Schlussfolgerungen zu diesem Aspekt der Sprachsynthese ziehen.
2) Eine KI-generierte geklonte Stimme konnte in Kombination mit einer visuellen Komponente die audiotechnischen und sprachlichen Anforderungen hinsichtlich der wahrgenommenen Natürlichkeit wenig bis gar nicht erfüllen. Die Ergebnisse der Wahrnehmungsstudie deuten darauf hin, dass trotz Fortschritten in der KI-Sprachsynthese noch weitere Verbesserungen erforderlich sind, um den Anforderungen gerecht zu werden.
3) Emotionen eines KI-generierten Dialogausschnitts wurden im Vergleich zu der menschlichen Originalstimme von Testhörer*innen nur sehr unzuverlässig erkannt. Allerdings kam es hier zu deutlichen Unterschieden hinsichtlich der Genauigkeit, mit der die jeweiligen Emotionen erkannt wurden.

Nach der Beantwortung dieser Forschungsfragen folgt nun die zentrale Forschungsfrage.

Auf Basis der untersuchten Teilaspekte lässt sich zusammenfassend feststellen, dass die verwendeten KI-TTS-Modelle zwar vielversprechende Ansätze bieten, jedoch aktuell noch nicht in der Lage sind, die menschliche Stimme während des ADR-Prozesses einer Filmtonpostproduktion zu ersetzen. Die Ergebnisse zeigen, dass sowohl die Ähnlichkeit zur menschlichen Originalstimme als auch die Natürlichkeit der generierten Sprache Schwächen aufweisen. Das kann einen deutlichen Einfluss auf die Immersion und Authentizität von Filmdialogen haben und somit direkte Auswirkungen auf das Erlebnis der Zuschauer*innen. Diese Mängel machen klar, dass KI-basierte Sprachsynthese derzeit noch nicht die audiotechnischen und emotionalen Anforderungen des klassischen ADR-Prozesses erfüllen kann. Es muss jedoch auch berücksichtigt werden, dass die Ergebnisse dieser Arbeit durch bestimmte Einschränkungen, wie etwa die spezifische Auswahl der KI-Systeme, beeinflusst sein könnten.

Trotzdem zeigt der Ansatz großes Potenzial für zukünftige Anwendungen, v.a. in Verbindung mit Produktionen, die nur über begrenztes Budget und einen strengen Zeitplan verfügen. Um jedoch das ADR-Verfahren als ganzes zu ersetzen, bedarf es noch weitere Entwicklungen und Innovationen, insbesondere in Bezug auf die Arbeit mit deutschsprachigen Inhalten.

7. Fazit und Ausblick

In diesem abschließenden Kapitel soll die Arbeit in ihrer Gesamtheit nochmals knapp zusammengefasst werden, um anschließend einen Ausblick auf mögliche zukünftige Forschungsansätze zu bieten.

7.1. Zusammenfassung der Arbeit

Zur Beantwortung der Forschungsfrage wurden zunächst die Grundlagen der Filmtonpostproduktion inklusive des ADR-Verfahrens sowie der künstlichen Sprachsynthese erläutert. Die Grundlagen waren erforderlich, um die Anforderungen an ein KI-System, das den ADR-Prozess ersetzen soll, zu verdeutlichen und potenzielle Stärken und Schwächen im Rahmen der Untersuchung zu identifizieren. Anschließend wurde ein Überblick über aktuelle Innovationen sowohl im Bereich der Sprachsynthese als auch beim ADR-Verfahren gegeben, um die momentanen Möglichkeiten in beiden Feldern aufzuzeigen. Danach wurde detailliert auf die für die Wahrnehmungsstudie gewählten KI-Systeme und die mit ihnen durchgeführte Sprachsynthese eingegangen, um die grundlegende Funktionsweise und die dabei entstandenen Herausforderungen und darzustellen. Im Anschluss wurde die Wahrnehmungsstudie näher betrachtet. Es wurden zunächst wissenschaftliche Hypothesen aufgestellt und darauf basierend der Aufbau der Studie in Form einer Umfrage vorgestellt. Die Ergebnisse wurden durch verschiedene Signifikanztests analysiert und auf ihre Aussagekraft hin überprüft. Die erhaltenen Ergebnisse wurden abschließend im Kontext der Forschungshypothesen interpretiert, bevor die zentrale Forschungsfrage beantwortet wurde. Die Arbeit kam schlussendlich zu dem Ergebnis, dass die verwendeten KI-TTS-Modelle im Moment noch nicht dazu in der Lage sind das ADR-Verfahren der Filmtonpostproduktion zu ersetzen.

7.2. Ausblick auf zukünftige Forschung

Im Hinblick auf zukünftige Forschungsthemen steht die Entwicklung, bzw. das Training deutschsprachiger Modelle an oberster Stelle. Nur wenige für die Anwendung im ADR-Bereich geeigneten Systeme sind für die Arbeit mit deutschen Daten und Inhalten geeignet, weshalb hierbei v.a. die Weiterentwicklung und das Training der momentanen „State-of-the-Art“- Modelle auf deutschsprachigen Daten vorangetrieben werden sollte.

Darüber hinaus sollte auch im Bereich des „EmotionalTTS“ weiter geforscht werden, um sowohl die Synthese von emotionaler Sprache als auch die Kontrollierbarkeit der enthaltenen Emotionen zu verbessern und zu erweitern. Weitere Entwicklungen in diesem Bereich würden sich auch positiv auf die Anwendung von KI-Systemen im ADR-Prozess auswirken.

Der wohl wichtigste Forschungsbereich ist jedoch die Weiterentwicklung der KI-Systeme mit dem Ziel, das Risiko von Missbrauch zu minimieren. Nicht nur könnten damit die Schauspieler*innen, deren Stimme geklont werden würde, geschützt werden, sondern dies könnte auch dazu führen, dass mehr Schauspieler*innen ihre Zustimmung zur Nutzung von KI-Modellen erteilen. Ein solcher Fortschritt wäre entscheidend, wenn das ADR-Verfahren in Zukunft durch KI-TTS-Systeme ersetzt werden soll.

8. Literaturverzeichnis

Abdul, Z. K. & Al-Talabani, A. K. (2022). Mel Frequency Cepstral Coefficient and its Applications: A Review. *IEEE Access*, *10*, 122136–122158. https://doi.org/10.1109/access.2022.3223444

Adobe. (2024). *Adobe Podcast-KI*. https://podcast.adobe.com/enhance

Bernard, M. & Titeux, H. (2021). Phonemizer: Text to Phones Transcription for Multiple Languages in Python. *The Journal Of Open Source Software*, *6*(68), 3958. https://doi.org/10.21105/joss.03958

Birtchnell, T. (2018). Listening without ears: Artificial intelligence in audio mastering. *Big Data & Society*, *5*(2), 205395171880855. https://doi.org/10.1177/2053951718808553

Blackmagic Design. (2024). *Davinci Reolve* (Version 19) [Computer Software]. https://www.blackmagicdesign.com/products/davinciresolve/whatsnew

Bobbitt, T. (2020, Februar). *How to Conduct a Paired Samples t-Test in Excel*. https://www.statology.org/paired-samples-t-test-excel/

Bobbitt, Z. (2021). *How to Perform a Wilcoxon Signed Rank Test in Excel*. https://www.statology.org/wilcoxon-signed-rank-test-excel/

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F. & Weiss, B. (2005, 4. September). *A database of German emotional speech*. Interspeech 2005. https://doi.org/10.21437/interspeech.2005-446

Bussmann, H. (2008). *Lexikon der Sprachwissenschaft* (4. Aufl.). Alfred Kröner Verlag.

Chen, S., Liu, S., Zhou, L., Liu, Y., Tan, X., Li, J., Zhao, S., Qian, Y. & Wei, F. (2024). VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.05370

Cho, D., Oh, H., Kim, S., Lee, S. & Lee, S. (2024). EmoSphere-TTS: Emotional Style and Intensity Modeling via Spherical Emotion Vector for Controllable Emotional Text-to-Speech. *Interspeech 2024*. https://doi.org/10.21437/interspeech.2024-398

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494–509. https://doi.org/10.1037/0033-2909.114.3.494

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2. Aufl.). Lawrence Erlbaum Associates. https://www.utstat.toronto.edu/brunner/oldclass/378f16/readings/CohenPower.pdf

Desplanques, B., Thienpondt, J. & Demuynck, K. (2020, 25. Oktober). *ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification*. Interspeech 2020. https://doi.org/10.21437/interspeech.2020-2650

Döring, N. & Bortz, J. (2023). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (6. Aufl.). Springer-Verlag.

Ernst, N. (2021). *CliffsDelta* [Computer Software]. https://github.com/neilernst/cliffsDelta

Faul, F., Erdfelder, E., Lang, A. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Haozhe, C., Run, C. & Hirschberg, J. (2024). *EmoKnob: Enhance Voice Cloning with Fine-Grained Emotion Control*. https://doi.org/10.48550/arXiv.2410.00316

Huang, R., Zhang, C., Ren, Y., Zhao, Z. & Yu, D. (2023). Prosody-TTS: Improving Prosody with Masked Autoencoder and Conditional Diffusion Model For Expressive Text-to-Speech. *Findings Of The Association For Computational Linguistics: ACL 2023*. https://doi.org/10.18653/v1/2023.findings-acl.508

iZotope Inc. (2019). *Dialogue Match* [Computer Software]. https://www.izotope.com/en/products/dialogue-match.html

iZotope Inc. (2023). *Ozone* (Version 11) [Computer Software]. https://www.izotope.com/en/products/ozone.html

iZotope Inc. (2024). *RX Advanced* (Version 11) [Computer Software]. https://www.izotope.com/en/products/rx.html

Kiwa Digital Ltd. (2024). *VoiceQ* (8.2.0) [Computer Software]. https://www.voiceq.com/

Kizer, R. (2024). *ADR and Post-Sync Dialogue*. Focal Press. https://doi.org/10.4324/9781032414096

Kong, J., Kim, J. & Bae, J. (2020). HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2010.05646

Krotos. (2024). *Krotos Studio* (2.1.0) [Computer Software]. https://krotos.studio/

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Łańcucki, A. (2020). FastPitch: Parallel Text-to-speech with Pitch Prediction. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2006.06873

LANDR. (o. D.). *LANDR*. Abgerufen am 17. Oktober 2024, von https://www.landr.com/de/audio-online-mastern/

Lensing, J. U. (2009). *Sound-Design, Sound-Montage, Soundtrack-Komposition : über die Gestaltung von Filmton*.

Lux, F., Koch, J., Meyer, S., Bott, T., Schauffler, N., Denisov, P., Schweitzer, A. & Vu, N. T. (2023). The IMS Toucan System for the Blizzard Challenge 2023. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.17499

Lux, F., Koch, J. & Vu, N. T. (2022a). Exact Prosody Cloning in Zero-Shot Multispeaker Text-to-Speech. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2206.12229

Lux, F., Koch, J. & Vu, N. T. (2022b). Low-Resource Multilingual and Zero-Shot Multispeaker TTS. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2210.12223

Lux, F., Meyer, S., Behringer, L., Zalkow, F., Do, P., Coler, M., Habets, E. A. P. & Vu, N. T. (2024a). Meta Learning Text-to-Speech Synthesis in over 7000 Languages. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.06403

Lux, F., Meyer, S., Behringer, L., Zalkow, F., Do, P., Coler, M., Habets, E. A. P. & Vu, N. T. (2024b, September 22). *IMS-Toucan*. https://github.com/DigitalPhonetics/IMS-Toucan

Lux, F. & Vu, N. T. (2022). Language-Agnostic Meta-Learning for Low-Resource Text-to-Speech with Articulatory Features. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2203.03191

Meissel, K. & Yao, E. S. (2024). Using Cliff's Delta as a Non-Parametric Effect Size Measure: An Accessible Web App and R Tutorial. *Practical Assessment, Research, And Evaluation*, *29*(1). https://doi.org/10.7275/pare.1977

Meyer, S., Lux, F., Denisov, P., Koch, J., Tilli, P. & Vu, N. T. (2022). Speaker Anonymization with Phonetic Intermediate Representations. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2207.04834

Mohammed, S. J. & Radhika, N. (2022). Audio denoising using deep neural networks. In D. J. Hemanth, D. Pelusi & C. Vuppalapati (Hrsg.), *Intelligent Data Communication Technologies and Internet of Things. Lecture Notes on Data Engineering and Communications Technologies* (Bd. 101, S. 33–47). Springer. https://doi.org/10.1007/978-981-16-7610-9_3

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C. & Levin, L. S. (2016). PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. *International Conference On Computational Linguistics*, 3475–3484. https://www.aclweb.org/anthology/C16-1328.pdf

Mozziconacci, S. (1998). *Speech Variability and Emotion : Production and perception*. https://doi.org/10.6100/ir516785

Next Move Strategy Consulting. (2023). *Markvolumen für künstliche Intelligenz weltweit im Jahr 2021 und 2022 mit einer Prognose bis 2030 (in Millionen US-Dollar)* [Datensatz]. zitiert nach de.statista.com. https://de.statista.com/statistik/daten/studie/1405265/umfrage/kuenstliche-intelligenz-marktvolumen/

Ng, A. (2017, 2. Februar). *Artificial Intelligence is the New Electricity* (Stanford Graduate School of Business). Stanford MSx Future Forum, Stanford, Vereinigte Staaten. https://www.youtube.com/watch?v=21EiKfQYZXc

Purcell, J. (2007). *Dialogue Editing for Motion Pictures: A Guide to the Invisible Art*.

Puronas, V. (2023). Sonic Alchemist: AI's role in Amplifying creativity in Film sound. *INTERNATIONAL JOURNAL OF FILM AND MEDIA ARTS*, *8*(3), 89–107. https://doi.org/10.24140/ijfma.v8.n3.06

Qin, Z., Zhao, W., Yu, X. & Sun, X. (2023). OpenVoice: Versatile instant voice cloning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2312.01479

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z. & Liu, T. (2020). FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2006.04558

Roberts, E. & Backstage. (2023, August). *How Long Does It Take to Film a Movie?* https://www.backstage.com/magazine/article/how-long-does-it-take-to-film-a-movie-76171/

Rosenthal, R. (1991). *Meta-Analytic Procedures for Social Research*. SAGE Publications. https://doi.org/10.4135/9781412984997

Schuff, H., Vanderlyn, L., Adel, H. & Vu, N. T. (2023). How to do human evaluation: A brief introduction to user studies in NLP. *Natural Language Engineering*, *29*(5), 1199–1222. https://doi.org/10.1017/s1351324922000535

Shade Inc. (2024). *Shade* (2.0) [Computer Software]. https://shade.inc/

Shapiro, S. S. & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*(3/4), 591. https://doi.org/10.2307/2333709

Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D. & Khudanpur, S. (2018, 6. Juni). *Spoken Language Recognition using X-vectors*. Odyssey 2018. https://doi.org/10.21437/odyssey.2018-15

Snyder, D., Garcia-Romero, D., Povey, D. & Khudanpur, S. (2017, 16. August). *Deep Neural Network Embeddings for Text-Independent Speaker Verification*. Interspeech 2017. https://doi.org/10.21437/interspeech.2017-620

Sonomagic & Puronas, V. (2024). *Sonic Alchemist* (1.1.3) [Computer Software]. https://sonomagic.com/

Source Elements LLC. (2024). *Source Connect* (Version 4) [Computer Software]. https://www.source-elements.com/introducing-source-connect-4/

Staib, M., Teh, T. H., Torresquintero, A., Mohan, D. S. R., Foglianti, L., Lenain, R. & Gao, J. (2020, 25. Oktober). *Phonological Features for 0-Shot Multilingual Speech Synthesis*. Interspeech 2020. https://doi.org/10.21437/interspeech.2020-1821

Steinberg. (2023). *Nuendo* (Version 13) [Computer Software]. https://www.steinberg.net/de/nuendo/

Suno Inc. (2023). *Bark*. https://github.com/suno-ai/bark

Tits, N., Haddad, K. E. & Dutoit, T. (2019). Exploring Transfer Learning for Low Resource Emotional TTS. In *Advances in intelligent systems and computing* (S. 52–60). https://doi.org/10.1007/978-3-030-29516-5_5

Waddell, G. (2013). *Complete Audio Mastering*. McGraw Hill TAB.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S. & Wei, F. (2023). Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2301.02111

Waves Audio Ltd. (2022). *Clarity Vx* [Computer Software]. https://www.waves.com/plugins/clarity-vx

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80. https://doi.org/10.2307/3001968

XL8 Inc. (2022). *MediaCat* [Computer Software]. https://www.xl8.ai/products/mediacat

Zhu, L., Chen, L., Zhao, D., Zhou, J. & Zhang, W. (2017). Emotion Recognition from Chinese Speech for Smart Affective Services Using a Combination of SVM and DBN. *Sensors*, *17*(7), 1694. https://doi.org/10.3390/s17071694