**Masterarbeit**

Im Studiengang Audiovisuelle Medien

# Kreatives Sound Design mit KI: Potenziale und Herausforderungen für immersive narrative Formate

vorgelegt von Manuel Matern

Matr. Nr. 44979

an der Hochschule der Medien Stuttgart am 21.07.25

zur Erlangung des akademischen Grades eines Master of Engineering

Erstprüfer: Prof. Oliver Curdt

Zweitprüfer: Prof. Dr. Andreas Koch

## Ehrenwörtliche Erklärung

Hiermit versichere ich, Manuel Matern, ehrenwörtlich, dass ich die vorliegende Masterarbeit mit dem Titel: “Kreatives Sound Design mit KI: Potenziale und Herausforderungen für immersive narrative Formate” 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.

Ort, Datum: Stuttgart, 21.07.25

**Hilfsmittel Hinweise:** Die KI-basierten Werkzeuge "Chat-GPT" und "DeepL Translator" wurden eingesetzt, um bei sprachlichen Formulierungen zu unterstützen und die Lesbarkeit des Textes zu verbessern. Der Hinweis gilt für die gesamte Arbeit. Nach der Nutzung der Tools wurden die entsprechenden Texte gründlich vom Autor geprüft und überarbeitet. Der Autor übernimmt die volle Verantwortung für die Inhalte der Masterarbeit.

## Kurzfassung

Diese Masterarbeit untersucht die Potenziale und Herausforderungen des kreativen Sound Designs mit Künstlicher Intelligenz (KI) in narrativen und immersiven Medienformaten. Ausgehend von einer theoretischen Auseinandersetzung mit der Wirkung und Funktion von Sound Design in Film, Journalismus, Videospielen und Audioformaten, wird dargelegt, wie Klang zur Narration, Emotionalität und Immersion beiträgt. Anschließend werden verschiedene KI-Tools im Bereich Sound Design analysiert und anhand einer praktischen Produktion, bestehend aus einer Filmszene und einer Audiostory, mit traditionellen manuellen Methoden verglichen.

Zentrale Leitfragen betreffen das kreative Potenzial aktueller KI-Tools, ihre Einbindung in bestehende Workflows sowie ethische und gestalterische Fragestellungen. Die Ergebnisse zeigen, dass KI zwar in der Lage ist, einfache Klangaufgaben zu übernehmen und Produktionsprozesse zu unterstützen, in komplexen dramaturgischen Kontexten jedoch qualitative und emotionale Schwächen aufweist. Die Arbeit kommt zu dem Schluss, dass KI-gestütztes Sound Design eine sinnvolle Ergänzung im kreativen Prozess darstellen kann, die Verantwortung für Wirkung, Gestaltung und Authentizität jedoch weiterhin beim Menschen liegt.

## Abstract

This work examines the potentials and challenges of creative Sound Sesign with artificial intelligence (AI) in narrative and immersive media formats. Based on a theoretical examination of the effect and function of sound design in film, journalism, video games and audio formats, it explains how sound contributes to narration, emotionality and immersion. Various AI tools in the field of sound design are then analysed and compared with traditional manual methods using a practical production consisting of a film scene and an audio story.

Key questions concern the creative potential of current AI tools, their integration into existing workflows and ethical and creative issues. The results show that while AI is capable of performing simple sound tasks and supporting production processes, it exhibits qualitative and emotional weaknesses in complex dramaturgical contexts. The study concludes that AI-supported sound design can be a useful addition to the creative process, but that the responsibility for effect, design and authenticity still lies with humans.

## Inhaltsverzeichnis

**Ehrenwörtliche Erklärung ..... I**

**Kurzfassung ..... II**

**Abstract ..... II**

**Inhaltsverzeichnis ..... III**

**1. Motivation und Zielsetzung ..... 1**

**2. Methodik ..... 2**

**3. Begriffsdefinition ..... 2**

**4. Grundlagen des Sound Designs in narrativen Formaten hinsichtlich Zweck und Wirkung ..... 3**

4.1. Ziel und Zweck der auditiven Ebene ..... 3

4.2. Entwicklung der Nachvertonung von Bildern ..... 6

4.3. Bedeutung von Klängen ..... 8

4.4. Erwecken von Emotionen ..... 9

4.5. Sound Design in Videospielen ..... 10

4.6. Sound Design im Journalismus (TV, Radio und Podcasts) ..... 11

4.7. Risiken und ethische Überlegungen zu technischen Fortschritten ..... 15

**5. Einführung in KI und KI-Tools ..... 18**

5.1. Was ist KI ..... 18

5.2. Künstliche Intelligenz im Bereich des Sounds ..... 19

5.3. SFX Engine ..... 20

5.4. ClipmoveAI ..... 21

5.5. MyEdit ..... 22

5.6. Filmora 14 ..... 22

5.7. ElevenLabs: Custom AI Sound Effects Generator ..... 23

5.8. MMAudio: Revolutionäre KI-gestützte Video-zu-Audio-Generierung ..... 24

5.9. Meta: AudioCraft ..... 25

5.10. Meta: Audiobox ..... 26

5.11. Google: SoundStorm ..... 27
**6. Manuelles vs. KI-gestütztes Sound Design ..... 28**
6.1. Auswahl und Beschreibung der KI-Tools ..... 28
6.2. Beschreibung der Beispielproduktionen ..... 29
6.3. Vorgehensweise bei der Evaluation ..... 30
**7. Konzeption und Umsetzung ..... 31**
7.1. Beispielproduktion 1: Vertonung von Filmszenen ..... 31
7.2. Beispielproduktion 2: Produktion und Vertonung einer Audiostory ..... 32
7.3. Vertonung der Filmszenen und der Audiostory ..... 32
7.4. Vertonung der Audiostory ..... 33
7.5. Ergebnisse ..... 33
7.6. Herausforderungen und Lösungen während der Umsetzung ..... 35
**8. Evaluation und Analyse ..... 39**
8.1. Wirkung des Sound Designs auf Narrative und Immersive ..... 39
8.2. Unterschiede in Kreativität, Effizienz und emotionaler Wirkung ..... 43
8.3. Kritische Reflexion: Vor- und Nachteile von KI-gestütztem Sound Design im Vergleich zum manuellen Ansatz ..... 45
**9. Diskussion ..... 46**
9.1. Kreatives Potenzial und Grenzen von KI im Sound Design ..... 46
9.2. KI-Tools als Co-Creator im kreativen Prozess ..... 49
9.3. Herausforderungen, Risiken und ethische Überlegungen beim Einsatz von KI im Sound Design ..... 50
9.4. Verstärkung von bestehenden Risiken ..... 50
9.5. Ethische und rechtliche Überlegungen ..... 52
9.6. Veränderte Verantwortung in der Produktion ..... 54
9.7. Risikominderung durch KI ..... 55
**10. Fazit und Ausblick ..... 55**
10.1. Empfehlungen für den Einsatz von KI im Sound Design ..... 57

10.2. Weiterführende Forschung und Entwicklung ....58

**Literaturverzeichnis ....61**

**Abbildungsverzeichnis ....76**

**Tabellenverzeichnis ....77**

**Anhang ....78**

1. Whispering Woods ....78
2. Skizzenhaftes Protokoll der Eigenproduktion ....80
3. MMAudio Protokoll ....81
4. Filmora 14 Protokoll ....86
5. Skizzenhaftes Protokoll: Audiobox ....91
6. Skizzenhaftes Protokoll: ElevenLabs ....93
7. Checkliste für den verantwortungsvollen Einsatz von KI-Tools ....96

# 1. Motivation und Zielsetzung

*„Das Medium Klang wird immer wieder unterschätzt und viel zu spärlich als Gestaltungsmittel eingesetzt. Dabei sind die Möglichkeiten bewusster Klanggestaltung immens und können eine beachtliche Wirkung entfalten.“ – Jan Paul Herzer/Max Kullmann 2018*

Die Gestaltung von Klang ist Kernaufgabe eines jeden Tonschaffenden. Egal ob Film, Rundfunk oder Musik. Der Bereich des Sound Designs wird dabei aber oft unterschätzt. Die Wirkung des Sounds ist meist eher subtil als voll im Fokus. In Kinofilmen, Hörspielen, YouTube-Videos, Dokumentationen, journalistischen Beiträgen in Radio und TV, oder jeglichem Shortform-Content auf Plattformen wie TikTok oder Instagram, gibt es zu jeder Zeit etwas zu hören. Im Radio und Hörspiel wird sogar komplett auf die visuelle Ebene verzichtet. So wie Filme zu Zeiten des Stummfilms in den 1890er bis 19020er Jahren komplett ohne Ton produziert wurden(vgl. Ganguly 2022). Doch nicht nur im Bereich der Medienproduktion hat sich seither einiges getan. Die künstliche Intelligenz, kurz KI, wird nicht ohne Grund neben andere „Zukunftstechnologien“ wie der Dampfmaschine, der Eisenbahn, der Elektrizität oder der Kernenergie gestellt (vgl. Künstliche Intelligenz. Die letzte Erfindung der Menschheit? - MIK o. D.). Auch diese großen Erfindungen haben in der Vergangenheit „große Begeisterung, hohe Erwartungen an ihre Veränderungskraft und weitreichende Visionen zu ihren Einsatzmöglichkeiten, aber auch Kritik, Ängste vor ihrer Unkontrollierbarkeit und große Sorgen um ihre Folgen“ in die Gesellschaft gebracht (Künstliche Intelligenz. Die letzte Erfindung der Menschheit? - MIK o. D.). So fand die KI in den letzten Jahren auch ihren Einzug in die Medienproduktion.
In dieser Masterarbeit soll deshalb zunächst dargelegt werden, wie Sound in narrativen Medienformaten eingesetzt wird, und anschließend erforscht werden, wie künstliche Intelligenz die Produktion von Medienformaten beeinflusst und wie diese für die Produktion kreativ eingesetzt werden kann. Dafür sollen folgende Gesichtspunkte bearbeitet werden. Was sind die Grundlagen des Sound Designs in narrativen und immersiven Formaten? Welches (kreative) Potential beinhalten aktuelle KI-Tools hinsichtlich Sound Design und wie können sie die Narrative und die Immersive unterstützen? Was sind die Risiken von Sound Design und wie erweitert sich das durch die Nutzung von KI? Oder kann eine KI sogar Probleme lösen?

Im zweiten Teil dieser Arbeit wird dann genauer auf den Vergleich zweier Produktionen eingegangen. Hinsichtlich der Qualität des Sound Designs wird untersucht, wie sich manuelles und KI-gestütztes Sound Design in ihrer (emotionalen) Wirkung unterscheiden. Außerdem wird anschließend das Potential der KI in Bezug auf den Workflow im Sound Design eingeschätzt und geklärt, wie KI-Tools in das Tool Set Einzug finden könnten. Hat KI Vorteile in der Effizienz?
Im dritten Teil werden die Ergebnisse abschließend zusammengefasst und Potential und Grenzen der KI im Sound Design diskutiert. Fragen der Ethik und Verantwortung werden vom Autor eingeordnet und in Bezug auf die Zukunft, die KI als Co-Creator im kreativen Prozess hat, eingeschätzt.

## 2. Methodik

Für die Erforschung der Fragestellung wird zunächst eine qualitative Literaturrecherche durchgeführt. Auf Basis der Darlegung des aktuellen Stands der Forschung und der Technik wird sich einer Antwort auf die Forschungsfrage angenähert. Zusätzlich werden mit Hilfe einer praktischen Arbeit, in Form einer beispielhaften Produktion und einem Vergleich von KI basiertem, und manuellem Sound Design, die zuvor theoretisch zusammengefassten Inhalte untermalt, beziehungsweise geprüft. Dies ermöglicht einen detaillierten Einblick in das Potential und gleichzeitig die Risiken und Probleme des Themas.

## 3. Begriffsdefinition

Wenn es um Sound Design geht, tauchen schnell andere Begriffe auf, die auf den ersten Blick dasselbe zu beschreiben scheinen. Klanggestaltung oder Tongestaltung und Sound Design sind zum Beispiel Begriffe, die immer wieder und teilweise fälschlicherweise synonym verwendet werden. Diese Begriffe bezeichnen aber unterschiedliche Aufgaben im Audiobereich. Um terminologische Verwirrung zu vermeiden, werden in diesem Abschnitt die in dieser Arbeit verwendeten Begriffe definiert.
Die Begriffe Klanggestaltung, Tongestaltung und Sound Design (seltener: Sounddesign) werden mehr oder weniger austauschbar eingesetzt. Von Sound Design und Klanggestaltung wird jedoch auch gesprochen, wenn mithilfe von Synthesizern Klänge erzeugt und verändert werden. Dabei geht es, um die technische Bearbeitung und Veränderung der verschiedenen Parameter des Klangerzeugers, sodass der gewünschte Klang oder Sound erzielt wird. Insbesondere klangliche und musikalische

Wahrnehmungsmerkmale wie Lautstärke, Klangfarbe, Lokalisation, Raumeindruck oder Tonhöhe, werden in Gewichtung zwischen Schallquellen oder Instrumenten, und im zeitlichen Verlauf beeinflusst (vgl. Dickreiter 2014: 370). In enger Verbindung damit liegt die gezielte Klangsynthese, bei der es um die heutzutage meist digitale Erzeugung von Klängen geht (vgl. Dickreiter 2014: 105). Es stehen also die Praxis der Klangerzeugung und der Klangveränderung eng beieinander.

Im Bereich der Beschallung kommt der Begriff Klanggestaltung und damit auch der aus dem Englischen kommende Begriff Sound Design ebenfalls vor. Die Kunst, unterschiedlichste Räume mit der gegebenen Technik (bspw. Eines vorhanden Soundsystems) so zu beschallen, dass der bestmögliche Klang erzeugt wird, hat auch mit der Gestaltung von Klang zu tun. Je nach Einstellung und Position der Lautsprecher, oder dem Einsatz von Equalizern, wird der Klang, welcher über Beschallungsanlage auf die Zuschauer trifft (Public Address, kurz PA) verändert (vgl. Dickreiter 2014: 565).

Die Begriffe Klanggestaltung und Sound Design beschreiben in dieser Arbeit die kreative und technische Arbeit mit Sounds, Klängen und Geräuschen, welche in Form einer Aufnahme oder einer künstlichen Klangerzeugung (z.B. durch KI-Generierung) produziert, und für Medienformate wie Film oder Hörspiel genutzt werden, Inhalte zu vertonen. Dabei spielen technische, sowie künstlerische und gestalterische Aspekte eine Rolle.

# 4. Grundlagen des Sound Designs in narrativen Formaten hinsichtlich Zweck und Wirkung

Bevor auf den Einfluss der KI auf das Sound Design und dessen Wirkung eingegangen werden kann, werden im folgenden Kapitel die Grundlagen des Sound Designs in narrativen Formaten dargelegt. Dabei wird insbesondere auf Ziel und Zweck der auditiven Ebene in Bezug auf Narrative und Immersive eingegangen.

## 4.1. Ziel und Zweck der auditiven Ebene

Um die Wirkung der auditiven Ebene untersuchen zu können, muss zunächst geklärt werden, welches Ziel beim Einsatz von Ton jeglicher Art überhaupt verfolgt wird. Dabei gibt es verschiedene tontheoretische Ansätze, die sich historisch betrachtet einerseits durch Faktoren wie zum Beispiel technischer Fortschritt verändert und weiterentwickelt haben, und andererseits immer noch den gleichen Zweck haben wie vor 100 Jahren,

als die ersten Filme mit Musik und Geräuschen untermalt wurden (vgl. Ganguly 2022). Die Rolle der auditiven Ebene wird im Folgenden am Beispiel von Filmen erklärt.

Das grundlegende Ziel der auditiven Ebene in narrativen Formaten ist es, die Erzählung zu unterstützen, zu erweitern und zu vertiefen (vgl. Dickreiter 2014: 338). Im Bereich des Films wird dazu der Begriff „Diegese“ verwendet. Der französische Philosoph und Ästhetiker Etienne Souriau beschreibt die Diegese als „alles, was man als vom Film dargestellt betrachtet und was zur Wirklichkeit, die er in seiner Bedeutung voraussetzt, gehört“ (Souriau 1997: 151; vgl. Wikipedia-Autor*innen 2018). In anderen Worten wird unter der Diegese eine imaginäre, erzählte Welt und „alles, [...] was sich laut der vom Film präsentierten Fiktion ereignet und was sie implizierte, wenn man sie als wahr ansähe“ verstanden (Souriau 1997: 156). Des Weiteren wird nach Souriaus Beschreibung klar, dass es um Fiktion und nicht um die Realität geht. Die Diegese wird als „mentaler Akt“ verstanden, welcher vom Rezipienten geleistet wird (vgl. ebd.: 153). Erstes Ziel bei der Produktion eines Films ist also, dass die Darstellung, und damit die Diegese, Narrative und Immersion, also das Eintauchen in die virtuelle Welt des Films, schlüssig ist, und mit ausreichend Information gefüllt ist, damit das Eintauchen des Zuschauers erreicht wird. Da nicht nur die visuelle Ebene zur Darstellung beiträgt, sondern auch die auditive Ebene, gelten diese Ziele auch für eben diese. Die auditive Ebene unterstützt also zunächst die audiovisuelle. Orte und Geschehnisse werden über die Ton Ebene durch Geräusche beschrieben. Ein Türschließen, ein Schuss oder ein bellender Hund in Verbindung mit zum Gesehenen passenden atmosphärischen Klängen komplettieren die Darstellung einer Szene. Der Forscher Michel Chion nennt solche Geräusche „territory sounds“ (vgl. Chion 1994: 75). Die Aktionen werden durch passende Klänge untermalt, ohne dass dabei die Frage entsteht, woher die Geräusche stammen. Die Kombination aus visuellen und akustischen Reizen erzeugt nach Chion zwei Phänomene: die Synchrese und den Mehrwert (vgl. für folgenden Abschnitt ebd.: 63; Flückiger 2011: 162). Der Begriff Synchrese ist eine Wortschöpfung, die sich aus den Elementen Synchronismus und Synthese zusammensetzt. Dieses Phänomen beschreibt die Verbindung eines akustischen und eines kurzen optischen Ereignisses, beim zeitgleichen Auftreten. Diese Verknüpfung tritt unabhängig von jeder rationalen Logik auf, was bedeutet, dass unser Gehirn akustische und visuelle Eindrücke automatisch als zusammengehörig interpretiert, sobald sie gleichzeitig wahrgenommen werden. Als Mehrwert wird der „Ausdrucks- und Informationswert, mit dem ein Ton ein gegebenes Bild auf

verschiedenen Ebenen anreichert“, beschrieben (Flückiger 2011: 162; vgl. Chion 1994: 63). Chion beschreibt des Weiteren, dass fälschlicherweise oft angenommen wird, dass der Eindruck, oder die Bedeutung, bereits vollständig in dem enthalten ist, was gesehen wird. Sozusagen eine Verdopplung des Eindrucks, was den Ton überflüssig machen würde. In seiner Theorie werden zwei zentrale Wirkungsprinzipien audiovisueller Wahrnehmung dargelegt: Das erste Prinzip betrifft die perzeptive Ebene, in der durch die zeitliche Koinzidenz von optischen und akustischen Reizen eine unmittelbare Sinnesverbindung entsteht. Das zweite Prinzip wirkt auf semantischer Ebene: Ton und Bild modifizieren sich wechselseitig in ihrer Bedeutung, sodass eine neue, gemeinsame Aussage entsteht.

Ein besonders aufschlussreiches Beispiel für diese Verflechtung liefert der McGurk-Effekt, ein in 1976 von Harry McGurk und John MacDonald beschriebenes Experiment (vgl. Mcgurk/Macdonald 1976; brainrulesbook 2008). Es demonstriert, wie visuelle Informationen (z. B. Lippenbewegungen) die akustische Wahrnehmung von Sprachlauten unwillkürlich verändern können. Dieses Phänomen macht deutlich, dass Hören und Sehen eng miteinander verbunden sind und auf komplexe Weise zusammenwirken. Genau diesen Mechanismus untersuchte Chion in seinen Theorien systematisch.

Bis hier hin bleibt festzuhalten, dass die klangliche Ebene Informationen vermittelt. Flückiger bezieht sich in ihrer Arbeit „Zum Mehrwert Ton und Bild“ beispielsweise auf den Dialog. Es werden „Informationen über den emotionalen Zustand des Sprechers oder der Sprecherin und über seinen oder ihren sozialen oder ethnischen Hintergrund“ vermittelt (Flückiger 2011: 166). Attribute wie körperliche Kondition, Alter oder gesundheitlicher Zustand der sprechenden Person können allein über die Sprache erkannt werden. Weiterführend beschreibt sie den Unterschied von sensorischen und semantischen Informationen von Klangobjekten. Als Beispiel zieht sie hier eine sich schließende Tür heran, welche einerseits bedeuten kann, dass jemand kommt oder geht, oder andererseits, dass jemand eingeschlossen wird. Zusätzliche Informationen, wie die Beschaffenheit der Türe werden ebenfalls übermittelt. Größe, Gewicht oder Zustand des Objekts können über den Ton erkannt werden, oder lassen zumindest darauf schließen (vgl. ebd.). Durch diese Informationsvermittlung kann die virtuelle Welt erweitert und detaillierter beschrieben werden, wodurch die Narrative unterstützt, und der Grad der Immersion erhöht werden.

Nachdem in diesem Kapitel die Grundlagen der Vertonung und dessen Ziel und Zweck erklärt wurden, wird im nächsten Kapitel die Entstehung der Nachvertonung dargelegt. Dabei wird konkret auf die frühe Entwicklung von Nachvertonungen eingegangen und anschließen auf die Entstehung des Berufs und der Aufgaben von Sounddesigner*innen.

### 4.2. Entwicklung der Nachvertonung von Bildern

Bereits im ersten Jahrhundert v. Chr. wurden mit Hilfe von hydraulischen Maschinen Windgeräusche, oder mit so genannten Donnerblechen Donner erzeugt (vgl. ebd.:168). Bis heute werden so genannte Windmaschinen für Aufführungen benutzt (vgl. Stuttgart o. D.). In den 1920er Jahren wurde mit der Einführung des Tonfilms viel auf Originalaufnahmen gesetzt. Alfred Hitchcock ließ beispielsweise für seinen Film *Erpressung (OG: Blackmail)* aus dem Jahr 1929 auf den Straßen vor dem Studiogelände Mikrofone aufstellen, um die dort herrschenden Straßen Geräusche aufzunehmen (vgl. Flückiger 2011: 168; IMDb o. D.a). Durch diesen technischen Fortschritt wurde es möglich, eine Szene sehr präzise wiederzugeben. Entweder wurde der Originalton exakt übernommen oder es kam ein ähnlicher Ton zum Einsatz, der zur jeweiligen Situation passte. Die damalige Begrenzung auf nur eine Mono-Tonspur führte jedoch zu Problemen, da sich verschiedene Geräusche überlagerten und gegenseitig verdeckten. Aus diesem Grund griff man, wie Flückiger erklärt, vor allem auf Klänge zurück, die kulturell stark verankert waren und einen markanten akustischen Charakter besaßen, wie zum Beispiel das Hupen eines Autos oder das Klingeln eines Telefons (vgl. Flückiger 2011: 168). Nach Wesley C. Miller sind die besten Ergebnisse erzielt worden, wenn Filmgeräusche durchdacht und künstlerisch eingebunden wurden (vgl. Miller 1931: 211; Flückiger 2011: 168). Es ist also die Rede von kreativer Klangsubstitution. Flückiger zieht hier eine entscheidende Verbindung zum „Continuity System“ aus Hollywood, „dessen Ziel es ist, die Erzählinstanz zu maskieren und die erzählerischen Elemente so anzuordnen, dass sie wie von selbst ein kohärentes Ganzes bilden und die fragmentierte und addierte Darstellungsweise des Films verschwinden lassen“ (Flückiger 2011: 168f.; Bordwell et al. 1985b 287ff.). Die Klänge und Geräusche, die eingesetzt werden, verfolgen demnach das Ziel, das Gesehene möglichst authentisch darzustellen, ohne dabei dem Rezipienten großartig aufzufallen, oder aus dem Raster zu fallen. Der oben beschriebene McGurk-Effekt veranschaulicht, wie in einer typischen Kinosituation die Wahrnehmung des Zuschauers beeinflusst wird. Da Ton und Bild im Film häufig getrennt voneinander

aufgenommen werden, stimmt die sichtbare Quelle oft nicht mit dem tatsächlich gehörten Ton überein. Dies führt dazu, dass die Zuschauer manchmal etwas anderes wahrnehmen, als tatsächlich dargestellt wird. Klangelemente wurden in diesem System auf ihre erzählerischen Ziele hin funktionalisiert und so eingesetzt, um die Aufmerksamkeit der Zuschauer auf bestimmte Dinge zu lenken. Wie oben beschrieben, können bestimmte Geräusche bereits eine Reaktion bei den Zuschauern hervorrufen. Hört man eine Tür quietschen, kommt oder geht jemand, hört man ein Hund bellen, wird im nächsten Schnitt ein Hund erwartet. Die Gestaltung der auditiven Ebene bezieht sich demnach nicht nur auf den Klang und die Auswahl der Geräusche, sondern auch auf das Timing.

In den 1970er Jahren wurden im Abspann von Filmen wie „Star Wars“ oder „Apocalypse Now“ zum ersten Mal Sounddesigner*innen erwähnt (vgl. für folgenden Abschnitt Flückiger 2023: 13-18). Heute ist oft auch die Rede von „Sound Editor“ oder einfach „Sound Effects“. Sounddesigner*innen arbeiten von Beginn der Planungsphase, über die eigentliche Zeit am Set und im Dreh, zur Postproduktion eng mit den Produzent*innen und Regisseur*innen zusammen. Es geht darum, sich ein Bild von dem gewünschten Sound des Films zu machen und bereits ein Konzept für die spätere Vertonung, den Klang der Sounds und deren Timing zu erstellen (vgl. Carlsson o. D.). Das Ziel ist die Gestaltung eines Ton-ästhetischen Gesamtkonzepts, welches in Einklang mit der visuellen Ästhetik des Films funktioniert. Walter Murch beschreibt in seinem Buch „The Conversations“ seine einzigartige Vorgehensweise, Szenen zu vertonen. Bereits im Rohschnitt wird das Material direkt stumm geschaltet, sodass Murch sich den Sound und die Stimmung im Kopf vorstellen kann. Dabei bezieht er sich ausschließlich auf zu sehende Aspekte wie Mimik oder Körpersprachen (vgl. Ondaatje/Murch 2002: 289f.). Erst anschließend wird der Ton hinzugeschalten, um zu sehen, wie gut die Nachvertonung funktioniert. Die Produktion eines Soundtracks für eine Medienproduktion umfasst aber nicht nur Sound Effekte, sondern auch Musik. Da für die Produktion des musikalischen Soundtracks, also der Filmmusik, Komponist*innen zuständig sind, müssen Sounddesigner*innen eng mit ihnen zusammenarbeiten, da Soundeffekte, welche die Atmosphäre einer Szene mitbestimmen, auf die Musik angepasst werden müssen. Die Sound Effekte, oder Geräusche, werden meist entweder aus Archiven gesammelt, oder eigens für den Film erzeugt. Unter dem Begriff „Foleyartist“ versteht man zum Beispiel eine Person, deren Hauptaufgabe es ist, mit Hilfe aller möglicher Gegenstände Geräusche synchron zum

Film zu erzeugen und aufzunehmen (vgl. filmsound.org o. D.). Das können beispielsweise Schritt- oder Jackengeräusche sein (vgl. Flückiger 2023: 93). Sounddesigner*innen können jedoch auch für die Erstellung einzelner, bestimmter Sounds angestellt werden, wenn es sich um spezielle Geräusche handelt. So wurde Ron Bochar nur für die Vertonung der Verwandlung von Jack Nicholson in einen Wolf, in Mike Nichols gleichnamigen Film „Wolf“, angeheuert (vgl. IMDb o. D.b; Carlsson o. D.).

Wie Flückiger weiter diskutiert, kann der Ton auch dazu beitragen, Objekte, welche nicht direkt identifizierbar sind, über die auditive Ebene vorstellen. Als Beispiel zieht sie hier das Lichtschwert aus der Star-Wars Saga heran (vgl. für folgenden Abschnitt Flückiger 2023: 18, 121, 144). Die von George Lucas entworfene Fantasie-Waffe konnte zum Zeitpunkt des ersten Auftritts in „A New Hope“ von den Zuschauer*innen nicht als Waffe identifiziert werden. Erst als durch den elektrischen Sound beim ersten „Zünden“ des Lichtschwerts, blitzartige Klänge und Geräusche, die an durch Hitze verschmortes Fleisch und Kleidungsstücke erinnerten, klar wurde, dass dieses Objekt sehr gefährlich war, wurde es in den Köpfen der Zuschauer*innen als Waffe verankert. Im Sound Design können also unmittelbar bevorstehende Ereignisse angekündigt, Geschehnisse im Umfeld einer Szene kommuniziert, und die Bedeutung von zu sehenden Objekten erklärt werden.

Nach dem das Ziel und der Zweck der auditiven Ebene dargelegt wurde, wird im folgenden Kapitel erklärt wie Geräusche und Klänge Bedeutungen kommunizieren.

### 4.3. Bedeutung von Klängen

Klänge und Geräusche werden wie oben dargelegt also nicht nur eingesetzt, damit Zuschauer*innen auch etwas hören, sondern beinhalten oft auch eine weitere Bedeutung. Barbara Flückiger spricht in ihrer Arbeit „Sound Design. Die virtuelle Klangwelt des Films“ von Klangobjekten höherer Ordnung *(vgl. für folgenden Abschnitt Flückiger 2023: 158-175).* Sie unterteilt sie in „Signale“, „Symbole“, „Key Sounds“ und „Leitmotive“. Signale und Symbole beinhalten eine Bedeutung, welche historisch oder kulturell verankert ist. Wie oben beschrieben, ergeben sich durch den immer wieder aufkommenden Einsatz bestimmter Sounds in bestimmten Szenen Bedeutungen einzelner Geräusche. Hierfür nutzt Flückiger den Begriff Key Sounds, also Klangobjekte, die durch ihre wiederholte Verwendung im Film eine symbolische

Bedeutung erhalten. Der Film selbst muss diese Bedeutung jedoch erst etablieren. Signalgeräusche beinhalten oft eine Handlungsaufforderung, während Symbole, wie beispielsweise das Ticken einer Uhr, durch die Interpretation des Zuschauers beispielsweise für die Vergänglichkeit stehen könnte. Stereotypen hingegen sind Klangobjekte, die erfahrene Zuschauer an bestimmten Stellen erwarten. Diese Erwartungshaltung ist so stark, dass Filmemacher sich oft gezwungen sehen, Stereotypen zu verwenden, um das Publikum nicht zu enttäuschen. Die auditive Ebene im Film ist also komplex und reicht von direkten Signalen über interpretationsbedürftige Symbole bis hin zu Stereotypen, die den Erwartungshorizont des Publikums prägen. Das oben bereits aufgeführte Beispiel der quietschenden Türe, welches einerseits Signal für das Öffnen oder Schließen einer Türe ist, wird im Genre des Horrorfilms plötzlich zu einem gruseligen Geräusch. Ähnliches passiert mit dem Lachen oder Weinen eines Kindes. Je nach Szene wird aus einem vermeintlich harmlosen Alltagsgeräusch etwas Unheilvolles (vgl. Clasen 2022). Diese Geräusche werden funktionalisiert und bewusst eingesetzt, um eine gewisse Wirkung bei den Rezipient*innen zu erzeugen und Emotionen zu erwecken.

Wie dieses Erwecken von Emotionen funktioniert und wie es in Medienformaten durch kreatives Sound Design erreicht werden kann wird im folgenden Kapitel behandelt.

### 4.4. Erwecken von Emotionen

Geräusche unterstützen nicht nur das grundlegende Level der Narrative, also das simple Erzählen einer Geschichte, sondern werden gezielt eingesetzt, um bei den Zuschauer*innen Emotionen zu erwecken (die folgende Argumentation basiert auf früheren Überlegungen des Autors; vgl. hierfür Matern 2022: 5ff.). Der Emotionsforscher Klaus Scherer unterscheidet zwischen überlebensdienlichen, utilitaristischen Emotionen und zweckfreien, ästhetischen Emotionen (vgl. für folgenden Abschnitt Scherer 2005: 706f.; Altenmüller/Bernatzky 2015: 224f.). Dem äußeren Ausdruck von Emotionen, also der physiologischen Ebene, geht stets das innere emotionale Erleben voraus, das als subjektive Komponente bezeichnet wird. Plötzlich auftretende Geräusche, die häufig mit einer abrupten Lautstärkezunahme einhergehen, können beispielsweise schreckhafte Reaktionen wie eine beschleunigte Atmung oder eine erhöhte Herzfrequenz hervorrufen. Auch Mimik und Gestik geben Aufschluss über emotionale Zustände. Wie Flückiger beschreibt, haben sogenannte Key Sounds und akustische Stereotype eine gelernte Bedeutung. Wenn

Rezipient*innen allein durch die Erwartung eines bestimmten Geräuschs körperlich reagieren, spricht man von einem „konditionierten Reiz-Reaktions-Schema“ (vgl. Altenmüller/Kopiez 2005: 225).
Im musikalischen Kontext treten häufig „ästhetische Emotionen“ auf, die sich sogar auf der Ebene des autonomen Nervensystems physiologisch nachweisen lassen. Besonders intensive Reaktionen, sogenannte SEMs (Starke Emotionen beim Musikhören), wurden unter anderem von Gabrielsson untersucht (vgl. Gabrielsson 2010). Ein bekanntes Beispiel hierfür sind „Chills“ oder Gänsehautmomente beim Musikhören. Sowohl akustische als auch visuelle Reize können solche starken emotionalen Reaktionen auslösen. Während melancholisch-nostalgische Melodien tiefgreifende Gefühle hervorrufen können, zeigen sich ähnliche Reaktionen auch bei unangenehmen Geräuschen wie dem Kratzen von Kreide auf einer Tafel oder dem Quietschen von Besteck auf einem Teller. Solche Reize führen zur Ausschüttung von Dopamin, was die emotionale Erregung und die Motivation steigert. Durch diese neurobiologische Reaktion werden entsprechende Erlebnisse besonders stark im Langzeitgedächtnis verankert. Dies macht deutlich, wie eng unsere sensorische Wahrnehmung mit emotionalen und kognitiven Prozessen verbunden ist und welche zentrale Rolle sie für unser Erinnerungsvermögen und subjektives Erleben spielt (vgl. Altenmüller/Bernatzky 2015: 227; Schramm 2008: 139). Für Sound Designer*innen sind diese Grundlagen wertvoller Begleiter beim Einsatz von Geräuschen. Es ist essenziell für die Qualität jeglicher Medienproduktion, unabhängig des Formats, zu jeder Zeit zu wissen und zu erkennen, welche Wirkung die auditive Ebene allein, und in Verbindung mit der visuellen Ebene hat.

Im folgenden Kapitel wird untersucht, wie die im vorherigen Kapitel dargelegten Ziele der auditiven Ebene in verschiedenen Medienformaten umgesetzt wird, und welche Rolle Sound Design darin spielt. Dafür werden einzelne Medienformate analysiert und auf die narrative und immersive Wirkung von Sound spezifisch eingegangen.

### 4.5. Sound Design in Videospielen

Sound Design in Videospielen ist wichtiger Bestandteil für die Erreichung einer möglichst hohen Immersion, also des Grads der „Eintauchung“ in die erschaffene Welt (vgl. Leenders 2012: 10). „Der Ton ermöglicht es dem Geschichtenerzähler, eine fesselnde Umgebung zu schaffen und zu ergänzen, die die emotionale Wirkung der Szene oder des Spiels verstärkt und die aktive Beteiligung des Spielers erhöht“

(Sinclair 2020: 12). Im Bereich der Videospiele ist die Interaktion des Spielers mit der fiktiven Welt entscheidender Unterschied zum Film. Der Soundtrack des Spiels muss nicht nur mit der visuellen Darstellung kohärent sein und entertainen, sondern den Spieler*innen zu jedem Zeitpunkt nötige Informationen über die virtuelle Umgebung liefern (vgl. Sinclair 2020: 8ff.): Wo befindet sich mein Charakter, wie weit sind eventuelle andere Spieler*innen oder KI-Gegner entfernt und wo befinden sich diese? Die „Präsenz der Umwelt" als Informationsfluss muss ununterbrochen laufen, egal ob Videospiel oder Film (vgl. Flückiger 2023: 299). Attribute wie Ortung und Distanz müssen vom Soundtrack möglichst präzise dargestellt werden. Technologien wie *Spatial Audio*[1]helfen dabei, die Immersion durch genauere Ortung von Geräuschen zu erhöhen. Der Soundtrack des Videospiels ist nicht linear, sondern muss den Handlungen des Spielers folgen (adaptiv). Dabei geht es nicht nur um Geräusche, sondern auch um Musik, egal, ob *Sidescroller*[2] oder *3D-Open-World*[3]. Des Weiteren ist die Stimmsynthese für Dialoge in Spielen ein zentrales Themengebiet. In einer großen virtuellen Welt begegnet man meist vielen verschiedenen Charakteren und *NPC*s[4], welche alle eine andere Stimme benötigen. Mit Hilfe von künstlicher Stimmsynthese ist es Spielentwickler*innen schnell und einfach möglich, diese Stimmen ohne aufwendige Aufnahme-Sessions zu generieren (vgl. Summers 2024).

Im folgenden Abschnitt wird das Sound Design in TV- und, Radioformaten, Podcasts und dem damit verbundenen Journalismus untersucht. Es wird analysiert, wie sorgfältig ausgewählte Tonelemente Elemente die Berichterstattung aufwerten, sie fesselnder machen und zu einer größeren Wirkung führen können.

### 4.6. Sound Design im Journalismus (TV, Radio und Podcasts)

Die Elemente Musik, Sound Effekte, Foley und Voice-Over zählen in der Produktion zu den Hauptaufgaben von Sounddesigner*innen (vgl. Score a Score 2025). Unter Musik versteht man in Medienformaten wie zum Beispiel dem Fernsehen, Rundfunk oder Podcasts, auch die Produktion von so genannten Jingles. Diese stellen kurze

[1] Der Begriff Spatial Audio (engl.), auch 3D-Audio, beschreibt, wenn Ton aus allen Richtungen (auch von oben) auf die zuhörende Person trifft (vgl. Schlüter 2024).
[2] Der Begriff Sidescroller (engl.) bezeichnet Videospiele, bei denen die spielende Person von der Seite auf das Spielgeschehen schaut (vgl. Davison 2005: 298).
[3] Der Begriff 3D-Open-World (engl. open world „offene Welt") bezeichnet Videospiele, welche eine offene, nicht-lineare Spielwelt, in der sich die Spieler*innen frei bewegen können, bieten (vgl. Thöing 2008).
[4] NPCs (enlg. Non-Playable-Charakter „Nichtspielcharakter) sind Spielfiguren die nicht von den Spieler*innen gesteuert werden können (vgl. Unterhaltungssoftware Selbstkontrolle (USK) 2019).

Ton- oder Musikausschnitte dar, welche in neue Rubriken einführen, Themenblöcke voneinander trennen, oder als Intro oder Outro verwendet werden (vgl. Kolbeck o. D.). Je nach Inhalt oder Thema wird dabei auf unterschiedliche Musik gesetzt. Die Jingles müssen teilweise unter dem Gesprochenen wie ein Bett liegen, ohne die Sprachverständlichkeit zu stören. Gleichzeitig müssen sie aber auch zum Inhalt passen, um Fehlkommunikation zu vermeiden und gleichermaßen einprägsam sein, um einen möglichst hohen Widererkennungswert zu haben. Während die Musik in Formaten der Unterhaltung die gleiche Rolle spielt wie im Film oder Videospiel, wird sie in informativen Bereichen wie dem Journalismus in Bezug auf die Immersion weniger mit Entertainment, sondern der Fokussierung auf den Inhalt verbunden. Journalistische Beiträge im TV, Radio oder dem Internet beinhalten meist eine Form von Dialog oder Sprache und gelegentlich atmosphärische Klänge in Form von O-Tönen[5] oder musikalischen Elementen. Wie in den vorherigen Kapiteln dargelegt, bereichern Ton und Musik seit vielen Jahren die Erzählung von Geschichten. Audioelemente schaffen emotionale Verbindungen, erregen Aufmerksamkeit und fügen Bedeutung hinzu. In der heutigen journalistischen Landschaft, insbesondere mit der wachsenden Popularität von Podcasts und Audio-Dokumentationen, setzen Journalist*innen Klanglandschaften und Musik gezielt ein, um Erzählungen zu gestalten, Stimmungen zu erzeugen und die Zuhörer durch Erzählbögen zu führen. Die bereits erwähnten Jingles spielen dabei eine wichtige Rolle, beispielsweise zur Einleitung eines *Newsflash*[6] oder neuer Themenblöcke.

Wie Elke Huwiler darlegt, ist die Musik im Hörspiel ebenfalls zweifelsohne fester Bestandteil der Hörspiel-Narratologie (vgl. Huwiler 2005: 293). Während die Rolle von Sprache und O-Tönen bei Interviews oder Straßenberichten selbsterklärend ist, ist der Einsatz von nachsynchronisiertem Ton nicht ganz so schnell ersichtlich. Der Ton im Journalismus schafft für das Publikum eine eindringliche Erfahrung, die es ihm ermöglicht, sich im Augenblick präsent zu fühlen. Wie K.B. Jensen (vgl. Jensen 2006: 8ff.) feststellt, gibt es drei Bereiche der Klangforschung: Speech (Sprache), Music (Musik) und Soundscapes (Klanglandschaften). Durch das Einbinden von *O-Tönen* oder das Hinzufügen von Umgebungsgeräuschen, Musik und Stimmen können Journalist*innen die Zuhörer*innen an den Schauplatz einer Geschichte versetzen und

[5] Ein O-Ton bezeichnet den Originalton, also den ursprünglichen, unbearbeiteten Ton einer Audioaufnahme (von zum Beispiel eines Films) (vgl. Duden 2025).

[6] Als Newsflash werden kurze Nachrichtenbeiträge (z.B. im TV oder Radio) bezeichnet, welche die neusten Informationen über wichtige oder unerwartete Ereignisse enthalten (vgl. Cambridge University Press 2025) [übersetzt durch Autor].

sie dadurch lebendiger und fesselnder machen. Im Vergleich zu früheren Jahren, in denen News nur durch Mund-zu-Mund Propaganda oder durch Schrift in Zeitungen verbreitet werden konnten, ist durch die Erfindung des Radios bereits eine Steigerung der Immersion für die Zuhörer*innen erlangt worden. Wie Segal et al. (2025) feststellen, ist die Kommunikation von Angesicht zu Angesicht sehr wichtig, um die Botschaft richtig zu verstehen und Missverständnisse zu vermeiden, da auch die nonverbale Kommunikation, wie Gestik und Mimik, starke Signale aussendet. Das Grundschema der Kommunikation von Laswell (1948) zeigt, dass Kommunikation immer aus fünf Komponenten besteht. Dem Sender, der Botschaft, dem Kanal, dem Empfänger und der Wirkung. Wenn der Sender nicht sichtbar ist, kann die Botschaft missverstanden werden. Dennoch hat die menschliche Stimme, vielleicht das grundlegendste Element des Audiojournalismus, eine einzigartige Kraft beim immersiven Geschichtenerzählen. Sie kann Nuancen, Emotionen und Persönlichkeit auf eine Weise vermitteln, die Text allein nicht leisten kann. Wie Crook (1999) feststellt, schafft die Intimität der menschlichen Stimme im Radio- und Podcast-Journalismus ein Gefühl persönlicher Verbundenheit zwischen Sprecher*in und Zuhörer*in. Dennoch müssen Radiojournalist*innen ruhig, direkt, eloquent und professionell bleiben. Auf diese Weise wird sichergestellt, dass die Bedeutung und die potenziellen Auswirkungen von Eilmeldungen richtig vermittelt werden. Die TV-Nachrichtenübertragungen hingegen bietet die Möglichkeit, die Bilder in Verbindung mit dem Gehörten persönlich zu interpretieren. Der zweite Schritt zur Erhöhung der Immersion und damit zur Verbesserung des Erlebens der Erzählung ist die Einbindung von Musik. Die Einbindung von Musik im Hintergrund kann die Botschaft und den Inhalt unterstreichen und die Sicherheit, dass die Botschaft richtig verstanden wird, erhöhen (die folgende Argumentation basiert auf früheren Überlegungen des Autors; vgl. hierfür Matern 2022: 13-15.). In einer Studie von Frances Rauscher et. al. im Jahre 1993 wurde beispielsweise eine Leistungsverbesserung von Studierenden festgestellt, welche vor der Bearbeitung bestimmter Aufgaben für 10 Minuten die „Sonate für zwei Klaviere in D-Dur KV 448“, von Wolfgang Amadeus Mozart gehört hatten (vgl. Rauscher et al. 1993). Auch wenn diese Verbesserung der Leistung nachträglich nicht immer bestätigt werden konnte, wurde trotzdem immer wieder ein grundsätzlich positiver Einfluss von Musik für kognitive Leistungen beobachtet. Das als „Mozart-Effekt“ bezeichnete Phänomen konnte auch bei anderer Musik festgestellt werden und bezog sich in folgenden Studien nicht nur auf die visuell-räumliche Verarbeitung von

Informationen, sondern auch auf Kreativität und Verarbeitungsgeschwindigkeit (vgl. Degé/Schwarzer 2015: 360; Reinberger/Altenmüller 2012). Die sogenannte „Erregungs- und Stimmungshypothese“ des kanadischen Forschers und Komponisten Glenn Schellenberg der Universität Toronto besagt, dass durch einen positiven Stimulus, also zum Beispiel Musik, die Erregung und Stimmung eines Rezipienten erhöht werden, wodurch sich die resultierende, verbesserte kognitive Leistung erklären ließe (vgl. E. Glenn Schellenberg 2024; Nantais/Schellenberg 1999). Weitere Studien haben gezeigt, dass das Erzählen von Geschichten mit Ton das Engagement des Publikums sogar erheblich steigern kann. Eine Studie von Rodero (2012) belegt, dass Audioerzählungen die Vorstellungskraft der Zuhörer*innen wirksamer anregen als textbasierte Geschichten, was die Beobachtungen von Rauscher aus dem Jahr 1993 erneut bestätigt. Diese Steigerung der Immersion führt zu einer stärkeren emotionalen Beteiligung und schließlich zu einer besseren Erinnerung an die Informationen. Das Konzept der akustischen Erzählung hat in den öffentlichen Medien dadurch an Zugkraft gewonnen. Nach der Definition von Mchugh/Lindgren (2013) wird bei akustischen Erzählungen eine Kombination aus Stimme, Umgebungsgeräuschen und Musik verwendet, um eine multidimensionale Erzählung zu schaffen. Diese Technik ermöglicht es Journalist*innen, Fakten mit Atmosphäre und Emotionen zu untermalen. Im Journalismus kann die Einbeziehung von Geräuschkulissen oder musikalischen Elementen zu einem tieferen emotionalen Engagement und einer persönlichen Bindung an die erzählte Geschichte führen und somit das Publikum zu aktiven Teilnehmer*innen an Medienerlebnissen und nicht zu passiven Konsument*innen machen (vgl. Schmidgen 2021: 121f.; McLuhan 1994: 22-32). Das Engagement zu erhöhen bedeutet, die Grundlage für eine Diskussion zu schaffen, was Teil der Regel sechs der zehn Regeln des Journalismus nach Kovach/Rosenstiel (2021: 27) ist. Jingles beispielsweise besitzen einen hohen Wiedererkennungswert, damit bei jedem erneuten Einsatz klar ist, was als nächstes passiert. Ein hoher Grad der Wiedererkennung und die Vertrautheit sind wiederum ebenfalls Attribute, welche die Motivation und Erregung der Rezipient*innen, und damit die Wirkung auf kognitive Prozesse positiv beeinflussen (vgl. Reinhardt/Rötter 2012: 134).

Zusammenfassend lässt sich sagen, dass das immersive Storytelling mit Sound Journalist*innen leistungsstarke Werkzeuge an die Hand gibt, um das Publikum zu fesseln und komplexe Geschichten zu vermitteln. Durch die geschickte Kombination

von Stimme, Umgebungsgeräuschen und Musik können reichhaltige, mehrdimensionale Geschichten geschaffen werden, die bei den Zuhörer*innen auf große Resonanz stoßen und sowohl die Wirkung als auch die Einprägsamkeit ihrer Berichterstattung verstärken. Im folgenden Kapitel werden Überlegungen zu Risiken und ethischen Fragen durch technische Fortschritte angestellt.

### 4.7. Risiken und ethische Überlegungen zu technischen Fortschritten

Durch technologische Fortschritte konnte wie bereits gezeigt die Immersion erhöht und damit die bessere Überlieferung der Narrative oder Information sichergestellt werden. Erfindungen wie die *Text-to-Speech-Technologie*[7], *Voice Assistants*[8] und die breitere Verfügbarkeit von Audioinhalten über das Internet ermöglichen auch sehbehinderten Menschen, oder Menschen mit Leseschwierigkeiten den Zugang zum Informationsfluss und verbessern die allgemeine Zugänglichkeit und Barrierefreiheit (die folgende Argumentation basiert auf früheren Überlegungen des Autors; vgl. hierfür Matern 2024: 4-9.). Neben der größeren Inklusion bieten Audio-Inhalte auch einen höheren Nutzungskomfort. Podcasts zum Beispiel führen nicht nur zu einer verringerten Bildschirmzeit, was für manch einen ein zu begrüßender Nebeneffekt sein mag, sondern sind auf Abruf verfügbar, meist unabhängig und haben die gleichen Vorteile wie das Radio. Diese Bequemlichkeit und Benutzerfreundlichkeit spiegelt sich auch in den Nutzungsstatistiken wider. Im Jahr 2021 gab es bereits über 2 Millionen Podcasts in mehr als hundert Sprachen, während es im Jahr 2018 noch rund 550 Tausend waren (vgl. Winn 2021). Im Jahr 2024 gibt es bereits 4,25 Millionen Podcasts weltweit (vgl. Singh 2024). Noch nie war es so einfach, etwas in hoher Qualität aufzunehmen. Durch den einfachen Zugang zu hochwertigen Mikrofonen, Audio-Interfaces und digitalen Audio-Workstations (DAWs) sind professionelle Produktionen von zu Hause produzierten Inhalten kaum noch zu unterscheiden. Gerade auf Social-Media wird dies deutlich. Da gerade junge Menschen viel Zeit auf Instagram, TikTok und Co. verbringen, werden auch über eben diese Kanäle News konsumiert (vgl. Prakash 2024). Durch den privaten Zugang und die schnelle Verbreitung von Inhalten sind diese aber sehr schwer zu kontrollieren. Spätestens seit 2016, während Donald Trumps Präsidentschaftswahlkampfes, ein Thema aktueller denn je, hat das

[7] Text-to-Speech (auf Deutsch „Text-zu-Sprache“) bezeichnet die Technologie, die es Computern ermöglicht, Text zu natürlich klingenden Audiosignalen zu wandeln. Sie wird auch als „Vorlesetechnologie“ bezeichnet (vgl. Hu/Downie 2025).

[8] Der Begriff Voice Assistant (auf Deutsch „Sprachassistent) bezeichnet die Technologie, die es Menschen ermöglicht, mittels Sprache unter anderem Informationen von Computern abzufragen (vgl. Sorth 2024).

allgemeine Bewusstsein für Fake News zugenommen (vgl. Lilleslåtten 2022). Wenn man dazu bedenkt, dass die Inhalte, die in den sozialen Medien geteilt werden, von jedem in hoher Qualität produziert werden können, gibt es viele Risiken, auch in Bezug auf die Verbreitung von Fake News. In Verbindung mit Sound Design spielt dabei weniger die grundlegende Recherche von Fakten als mehr die unmissverständliche Kommunikation und Überlieferung ebendieser eine entscheidende Rolle. Als Beispiel der gezielten Verwendung von Musik in Nachrichtenberichten kann die rechtsgerichtete Partei „Alternative für Deutschland“ (AfD) angeführt werden (vgl. für den folgenden Abschnitt Schipmann et al. 2022). Die AfD hat im Vergleich zu anderen deutschen Parteien mit Abstand die meisten Accounts auf TikTok. Die Verwendung von kurzen Videos, die nur eine sehr einseitige Perspektive eines Vorfalls zeigen, in Kombination mit Musik und Sounds, die die vermittelte Botschaft verstärken, ist Teil ihrer erfolgreichen Strategie. Dies zeigt, wie wirkungsvoll Musik und Ton bei der Vermittlung einer Botschaft sein können. Auch die Anhänger*innen der AfD nutzen diese Strategie, um ihre Botschaft unabhängig von den offiziellen Aktivitäten der Partei zu verbreiten. Im Mai 2024 feierten AfD-Anhänger*innen eine Party auf der Insel Sylt. Videos dieser Party zeigen junge Menschen, die zu *Gigi D'Agostinos* Hit „L'Amour Toujours“ tanzen, trinken und singen. Ihre Überzeugungen finden sich im Text des Lieds wieder, der neu interpretiert wurde, wodurch das Lied zum Propagandalied wurde. Daraufhin tauchte das Lied unter zahlreichen Videos von AfD-Anhänger*innen auf.

Hintergrundmusik kann jedoch, wenn sie mit Bedacht eingesetzt wird, den emotionalen Ton einer Geschichte auch in positiver Weise bestimmen und die Zuhörer*innen durch die verschiedenen Erzählabschnitte führen. Es ist dabei sehr wichtig, Musik zu verwenden, die die Handlung darstellt oder unterstützt, um den Zuschauer nicht zu manipulieren. Eine übermäßige Verwendung von Soundeffekten oder Musik und ein übermäßiger Einsatz von Produktionstechniken können den eigentlichen journalistischen Inhalt überschatten und die übermittelte Botschaft verzerren (vgl. Tompkins o. D.). Die Gesamtqualität des Nachrichtenberichts wird damit herabgesetzt und könnte, als nicht vertrauenswürdige Quelle betrachtet werden. Die Verwendung von Klängen, die nicht mit dem Originalereignis in Verbindung stehen, könnte zu einer Fehlinterpretation der Realität führen und das Publikum in die Irre führen. In der Nachrichtenberichterstattung würde diese Verzerrung bedeuten, dass die Öffentlichkeit die Botschaft der Nachricht falsch versteht. Wenn das Publikum feststellt,

dass die hinzugefügten Geräusche und Musikstücke gefälscht, irreführend oder manipulativ sind, könnte die Glaubwürdigkeit der Journalist*in oder des Medienunternehmens schwer beschädigt werden. Daher sollte bei journalistischer Tonbearbeitung das gleiche Maß an ethischer Professionalität und Genauigkeit an den Tag gelegt werden wie bei der Auswahl von Worten und Fakten (vgl. ebd.). Die zehn Regeln des Journalismus nach Kovach/Rosenstiel (2021: 27) können also wie oben erwähnt, auch auf den Umgang mit Ton angewendet werden und spielen zur Sicherstellung des richtigen Umgangs mit Ton in Medienformaten eine gleichermaßen wichtige Rolle. Wenn Töne beispielsweise von einer Szene in eine andere verschoben werden, werden Fakten zu Fiktion. Daher muss für das Publikum immer klar sein, welche Töne oder Effekte bearbeitet oder hinzugefügt wurden. Ganz anders als im Bereich des Entertainments. Dies ist zwar die Grundlage der Audiobearbeitung im Allgemeinen, doch sollten Journalist*innen berücksichtigen, ob hinzugefügte Musik einen redaktionellen Ton erzeugt, der sonst nicht vorhanden wäre. Ein Dokumentarfilm zielt zum Beispiel darauf ab, reale Ereignisse oder Situationen darzustellen und sich auf Fakten und historische Gegebenheiten zu konzentrieren (vgl. Nichols 2017: 5). Dafür ist es wichtig, die Atmosphäre des Gezeigten einzufangen. Ein Beispiel für eine Rundfunkproduktion, die zu einer Verwirrung des öffentlichen Verständnisses dessen führte, was real und was fiktiv ist, ist Orson Welles' Radiosendung "The War of the Worlds". Im Jahr 1938 sorgte Welles' Adaption des Romans "The War of the Worlds" von H. G. Wells aus dem Jahr 1898 für Verwirrung. Einige Zuhörer glaubten, es habe eine Invasion der Marsmenschen stattgefunden (vgl. Burke 2023). Die kühle Darstellung und der gekonnte Einsatz der menschlichen Stimme durch Welles lösten bei einem Teil des Publikums Verwirrung und Panik aus.

In Bezug auf die Darstellung von realen Ereignissen ist wie eben beschrieben das Feingefühl beim Einsatz von Klängen sehr wichtig. Der Einsatz von Ton und Musik kann die Nachvollziehbarkeit und Wirkung von Nachrichten verbessern, jedoch muss der kulturelle Kontext richtig verstanden sein und dargestellt werden. Kulturelle Sensibilität bei der Produktion ist vielleicht wichtiger denn je. Bestimmte Klänge oder Musik können in verschiedenen Kulturen unterschiedliche Bedeutungen haben (vgl. Chilvers et al. 2023: 286). Beispielsweise unterscheidet sich die allgemeine Wahrnehmung von Musik in westlichen und östlichen Teilen der Welt. Daher sollte darauf geachtet werden, dass Ton nicht in einer Weise verwendet wird, die als unsensibel oder beleidigend empfunden werden könnte. Insbesondere in einem

Dokumentarfilm kann es sehr heikel sein, Musik zu verwenden, die stark westlich oder östlich geprägt ist, da ein Dokumentarfilm verschiedene Themen wie soziale Fragen, historische Ereignisse und kulturelle Unterschiede untersucht und darauf abzielt, zu informieren und den Wunsch zu wecken, mehr zu erfahren. Andererseits versuchen Dokumentarfilme auch, das Publikum einzubeziehen, indem sie neue Perspektiven eröffnen, das Bewusstsein schärfen und ein tieferes Verständnis der realen Welt fördern (vgl. Nichols 2017: 5–7). Musik kann dabei helfen, die Botschaft zu unterstreichen, wenn sie, wie bereits erwähnt, mit Bedacht eingesetzt wird.

In diesem Kapitel wurde gezeigt, dass der Einsatz von KI im Sound Design nicht nur neue kreative Möglichkeiten eröffnet, sondern auch ethische Risiken, insbesondere in narrativen und journalistischen Formaten birgt. Klänge beeinflussen die Wahrnehmung stark, und wenn ihre Herkunft unklar ist, kann dies zu Manipulation oder Fehlinterpretation führen. Diese Art von Immersion ist kein Qualitätsmerkmal, da sie auf irreführenden Klanggestaltungen beruht. Außerdem ersetzt KI nicht das bewusste, verantwortungsvolle Gestalten von Klang durch Sounddesigner*innen. Die Wirkung auf Narrative und Immersion hängt maßgeblich davon ab, wie transparent und überlegt Klänge eingesetzt werden. Besonders im Journalismus ist der verantwortungsbewusste und ethische Umgang mit Sound essenziell, um Vertrauen und Glaubwürdigkeit zu wahren.

# 5. Einführung in KI und KI-Tools

Nachdem in den vorherigen Kapiteln die Grundlagen von Sound Design und dem generellen Einsatz von Ton in verschiedenen Medienformaten dargelegt wurden, wird in diesem Kapitel die Verbindung zur künstlichen Intelligenz (KI) hergestellt. Dabei wird im Besonderen geklärt, was KI im Audiobereich eigentlich ist. Außerdem wird ein Blick auf das kreative Potential von KI in Bezug auf die Tonproduktion geworfen und verschiedene KI-Tools als Beispiele herangezogen. Letztlich wird untersucht, inwiefern die bereits beschriebenen Risiken bei manuellem Sound Design auch auf das Sound Design mit KI zutreffen und ob sie dort von größerer, oder vielleicht sogar kleinerer Bedeutung sind.

## 5.1. Was ist KI

Um genauer auf den Einsatz von KI in der Medienproduktion eingehen zu können, muss zunächst definiert werden, was Künstliche Intelligenz, kurz KI, eigentlich ist. Das Fraunhofer Institut beschreibt KI wie folgt: „Künstliche Intelligenz (KI) ist ein Teilgebiet

der Informatik. Sie imitiert menschliche kognitive Fähigkeiten, indem sie Informationen aus Eingabedaten erkennt und sortiert. Diese Intelligenz kann auf programmierten Abläufen basieren oder durch maschinelles Lernen erzeugt werden“ (Fraunhofer-Institut IKS o. D.).

Eine KI erschafft also nicht komplett Neues, noch nie Dagewesenes, sondern nutzt bereits vorhandene Informationen, verarbeitet diese und gibt eine Kombination dieser wieder aus. Im Bereich der textbasierten KI werden mit Hilfe des „Prompts“ Texte generiert und gestaltet. Im Bereich der Bild-KI (Text-to-Picture) wird ein Bild generiert, welches sich bereits bestehende Bilder zunutze macht, um möglichst nahe an die Beschreibung des Prompts heranzukommen. Ähnliche KI-Tools gibt es auch im Audiobereich. Musik kann mit Hilfe eines einfachen Text Prompts generiert werden. Dabei können die Wahl der Instrumente und Stimmung, sowie Attribute wie BPM, Tonart und Länge bestimmt werden. Auch passende Lyrics können von der KI geschrieben und mit Hilfe von KI-Stimmen generiert werden. In der Musikproduktion wird KI auch benutzt, um Sound Archive zu organisieren. Die KI sortiert dafür Samples [9] in Attribute wie Instrument oder Sound Effekt, Stimmung, Tonart und BPM. Ein Beispiel hierfür ist das Plugin *COSMOS* der Firma *Waves Audio Ltd* (vgl. Waves Audio Ltd. o. D.).

Nachdem geklärt wurde, was Künstliche Intelligenz ist wird im Folgenden geklärt, welche Rolle KI im Audiobereich spielt. Damit wird die Grundlage für den späteren Vergleich gelegt.

### 5.2. Künstliche Intelligenz im Bereich des Sounds

Im Audiobereich ist man sich teilweise uneinig, wie gut KI-Tools im Vergleich zur Bildgenerierung funktionieren können. John Lehmkuhl, auch bekannt als *PlugInGuru*, Plug-in Enthusiast und langjähriger Synthesizer-Sounddesigner bei *Korg* ist skeptisch, da die Vorstellung von Bildern eine andere sei als die von Geräuschen: „Wenn man sich eine ‚Akustikgitarre mit Flügeln, die über den Ozean schwebt,‘ wünscht, hat man ein ziemliches klares Bild im Kopf, wie das Ganze aussehen könnte. Bei Klängen sieht das anders aus. Augen und Ohren funktionieren unterschiedlich“ (Sauer 2024). Um diesem Potential weiter nachgehen zu können, werden nachfolgend einige KI Sound

[9] Ein Sample (engl. für „Auswahl“ oder „Beispiel“) bezeichnet hier eine kurze Audiodatei, welche beispielsweise ein Geräusch, oder Ton eines Instruments, oder kurze Melodie beinhaltet (vgl. Musik: Was ist eigentlich Sampling? 2016).

Design Tools vorgestellt. Dabei geht es um KI-Tools, welche durch einen manuellen Input in Form eines Text Prompts Sounds generieren.

### 5.3. SFX Engine

Auf der Website wird versprochen, dass dieser Soundeffect Generator es einem ermöglicht „individuelle Audioerlebnisse zu schaffen“ welche immer ideal zum jeweiligen Projekt passen, egal ob Film, Videospiel oder Musikproduktion (SFX Engine o. D.a) [übersetzt von Autor]. Es handelt sich dabei um den bereits oben beschriebenen Fall. User*innen geben im Graphical User Interface (GUI) einen Text Prompt ein und bekommen passende Sounds zur Auswahl (vgl. SFX Engine o. D.a). Dabei wird hohe Qualität versprochen und es sei „als ob man einen professionellen Sounddesigner zur Hand hätte“ (SFX Engine o. D.b) [übersetzt von Autor]. Weiter wird versprochen, Sounds unterschiedlichster Art generieren zu können. Geräusche der Natur wie Regen oder Tiergeräusche, menschliche Geräusche wie Schritte, musikalische Instrumente, Atmosphären wie Wälder oder Städte, aber auch Spezial-Effekte wie Explosionen sollen maßgeschneidert für die individuellen Bedürfnisse generiert werden können. Durch die passgenaue Generierung der Sounds wird nicht nur eine effizientere Arbeitsweise versprochen, auch eine höhere Immersion soll dadurch erzielt werden.

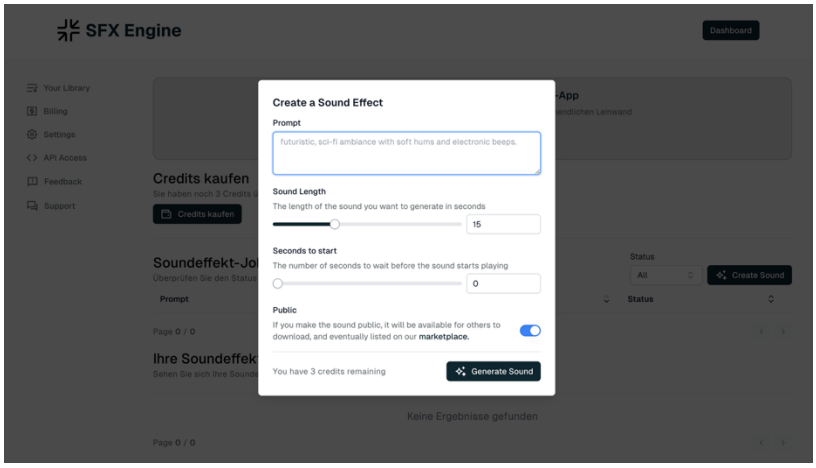


*Abbildung 1 GUI SFX Engine https://sfxengine.com/de/dashboard, abgerufen am 12.05.25.*

Das Tool bietet dazu die Möglichkeit, die Länge der Sounds frei anzupassen. Im Text Prompt Fenster sind dazu Längen zwischen Null und 47 Sekunden möglich. Außerdem kann laut Blog-Website der Stil, in Bezug auf angewandte Effekte wie Echo, Hall oder Verzerrung verändert werden, jedoch wird nach kurzem Test schnell klar, dass dies nur rein über den Text Prompt möglich ist (vgl. ebd.). Der gewünschte Sound soll außerdem in mehreren Versionen generiert werden können, aus denen gewählt werden kann. Diese Funktion ist jedoch nicht ersichtlich. Die einzige Möglichkeit besteht darin, einen neuen Sound mit angepasstem Text Prompt generieren zu lassen.

Negative Prompts zum Auslassen bestimmter Attribute sollen ebenfalls möglich sein. Ein Tag-System soll es zusätzlich möglich machen, schnell und effizient die Stimmung eines Geräuschs, beispielsweise von bedrohlich zu freundlich, zu verändern. Zuletzt wird angegeben, dass das Tool fähig ist, aus selbst ausgewählten und hochgeladenen Sounds etwas Neues zu generieren. Ein bereits existierendes Geräusch wird als Startpunkt genutzt, um noch genauere Ergebnisse zu erzielen. Jedoch sind beide Funktionen im Test nicht auffindbar gewesen.

Das Tool bietet außerdem die Möglichkeit, aus einer Sammlung an Geräuschen auszuwählen, die alle lizenzfrei genutzt werden können. Leider ist über die Arbeitsweise der KI keine Dokumentation vorhanden. Es sind keinerlei Informationen über das Unternehmen, welches hinter dem Tool steht, zu finden, was die Glaubwürdigkeit der lizenzfreien Nutzung der Sounds stark in Frage stellt. Die Angaben der Website stimmen außerdem nicht mit den tatsächlichen Funktionen des Tools überein, was die allgemeine Glaubwürdigkeit dieses Tools einschränkt.

### 5.4. ClipmoveAI

Dieses Unternehmen aus San Diego, CA, USA bietet gleich mehrere KI-Tools. Egal ob Text-to-Video, Image-to-Video, Videoenhancing (Automatische Videoverbesserung) oder ganze Musikvideo Generierung. Außerdem gibt es die Möglichkeit, sich mit Hilfe des Tools „Free AI Soundeffect Generator Online“ per Text Prompt passende Sounds generieren zu lassen. Ähnlich wie bei *SFX Engine* wird auch hier versprochen, hohe Qualität zu erzielen und eine große Bandbreite an Geräuschen, wie Atmosphären oder Spezial-Effekte, generieren zu können. Die KI soll außerdem „äußerst realistisch und von realen Aufnahmen nicht zu unterscheiden“ sein (ClipMove o. D.) [übersetzt von Autor]. Hierzu wird angegeben, dass das KI-Modell „anhand eines umfangreichen Datensatzes tatsächlicher Tonaufnahmen trainiert“ wurde und so „verschiedene Klangszenarien genau nachbilden“ kann [übersetzt].

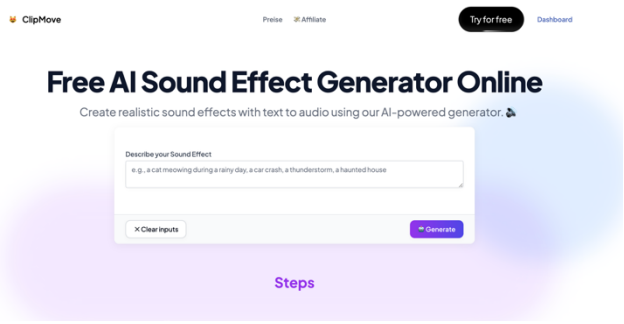


*Abbildung 4 GUI ClipmoveAI https://www.clipmove.com/de/ai-sound-effect-generator abgerufen am 06.07.25*

### 5.5. MyEdit

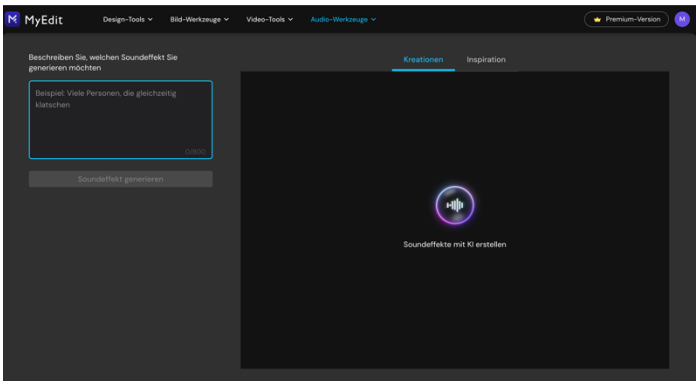


*Abbildung 7 GUI MyEdit https://myedit.online/de/audio-editor/ai-sound-effect-generator/edit abgerufen am 12.05.25*

Auch MyEdit stellt zahlreiche KI-Tools zur Erstellung von Musik, Videos, Bildern oder Sundeffekten bereit. Der „AI-Sound-Generator“ bietet auch hier wieder ähnliche Features wie die vorherigen Tools. Egal ob für Content-Creator*innen, Musiker*innen & Songwriter*innen oder Gamer*innen, das Tool verspricht eine schnelle und kostenlose Generierung von Geräuschen, ganz ohne Software (vgl. ClipMove o. D.) Die Generierung von Sounds erfolgt auch hier in drei Schritten: Beschreibung des Sounds; KI-basierte Sound Generierung; kostenloser Download. Es handelt sich also erneut um die Text-to-Sound Variante. Die KI stellt pro Generierung drei verschieden generierte Sounds zur Auswahl. Der Download der Dateien erfolgt schließlich in gängigen Formaten wie WAV, FLAC oder M4A.

### 5.6. Filmora 14

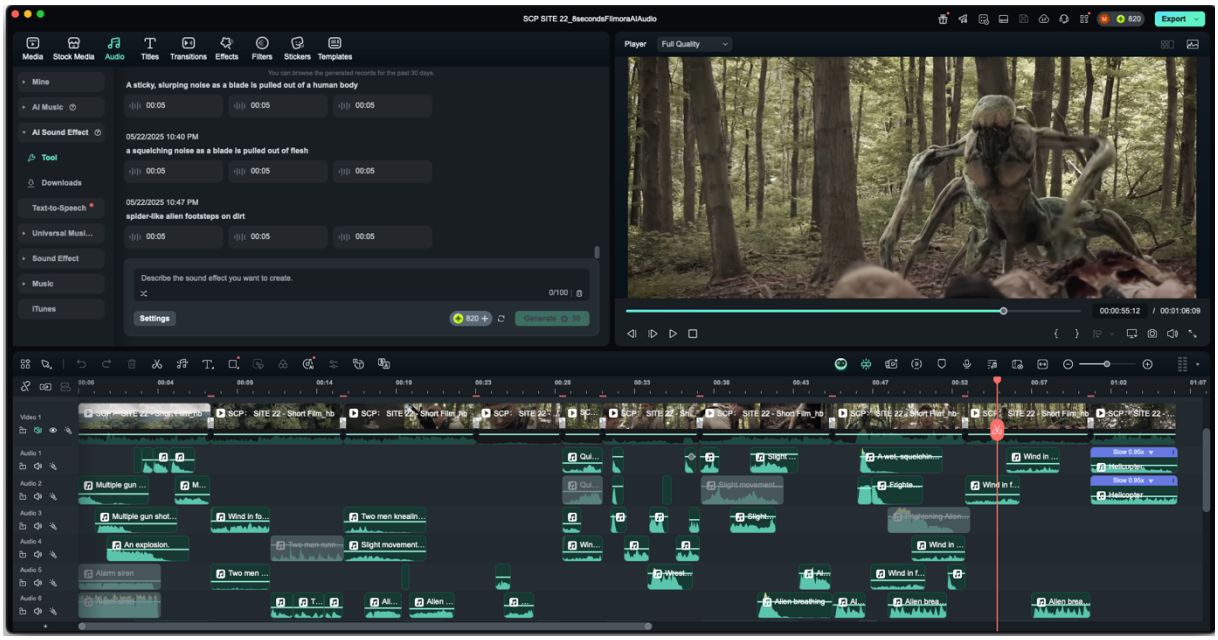


*Abbildung 10 Filmora 14 GUI*

Der Entwickler *Wondershare Technology* bietet mit Filmora 14 ein echtes „KI-Powertool“ zur Content Erstellung an. Neben normalem Video-Editing werden den Nutzer*innen KIs zur Verfügung gestellt, die automatisch Videos verbessern oder Kurzvideos mit der „Smart Short Clip“ Funktion erstellen (vgl. Wondershare o. D.b). Neben zahlreicher anderer KIs, wird auch hier ein „AI Sound Effect Generator“

angeboten. Das Tool nutzt erneut das „Text-to-Sound“ Modell und ist direkt in die Software integriert. Die KI verspricht die Generierung von Umwelt Geräuschen bis hin zu Sci-Fi Effekten und auch hier ist die Zeitersparnis in der Produktion fester Teil der Vermarktung. Die fertigen Sounds sind wieder in verschiedenen Formaten wie WAV, MP3, FLAC oder M4A zum Download verfügbar. Die drei Schritte zur Generierung sind im Fall von *Filmora* noch etwas ausgefeilter, da das Tool direkt in die Videobearbeitungssoftware eingebaut ist. So können beispielsweise die Länge und andere Attribute der generierten Geräusche direkt an das Video angepasst werden (vgl. Wondershare o. D.a).

### 5.7. ElevenLabs: Custom AI Sound Effects Generator

Auch bei *ElevenLabs* handelt es sich um einen Text-to-Sound Generator. Direkt auf der Website bietet ein Textfeld die Möglichkeit direkt mit der Beschreibung der zu generierenden Sounds zu beginnen. Das Tool generiert darauf hin vier Samples, welche laut Website „in Sekunden“ zur Verfügung stehen und „Auswahl und Workflow“ beschleunigen (vgl. für folgenden Abschnitt ElevenLabs 2024b) [übersetzt von Autor]. Des Weiteren wird garantiert, präzise Kontrolle über die generierten Sounds zu haben, da durch nuancierte Beschreibungen im Text-prompt die Ergebnisse nach Wunsch angepasst werden. *ElevenLabs* gibt außerdem an, dass alle generierten Sounds lizenzfrei sind und die Produktion mit „originalgetreuem Sound“ und „realistischem Klang“ bereichert wird. *ElevenLabs* bietet zusätzlich eine Funktion namens „Infinite Soundboard“ (vgl. ElevenLabs 2024a). Nutzer*innen können mit Hilfe eines virtuellen Soundboards Soundeffekte frei kombinieren und damit live neue Soundscapes kreieren. Das Interface bietet die Möglichkeit, frei neue Buttons mit Sounds hinzuzufügen, welche miteinander kombiniert werden können. Dabei kann aus vorgegebenen Presets gewählt oder eigene erstellt werden. Das Soundboard bietet einen Volume Regler und einen Loop Button, sowie einen „Edit mode“, in dem die Sounds mit der bereits bekannten Text Prompt Maske bearbeitet oder neu generiert werden können.

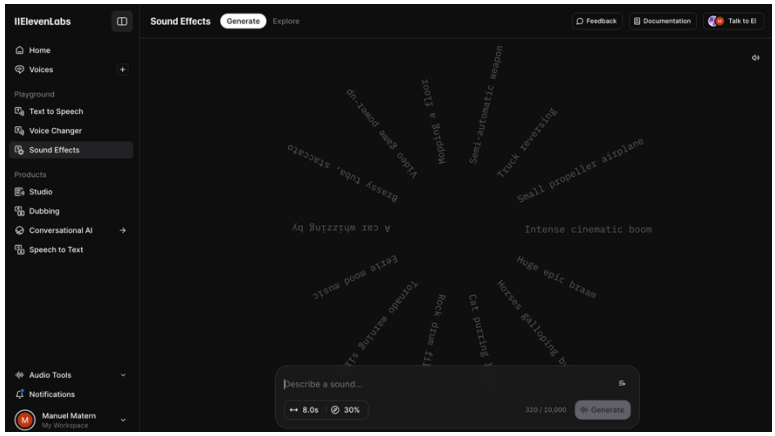


*Abbildung 13 ElvenLabs GUI*
*https://elevenlabs.io/app/sound-effects/generate*
*abgerufen am 12.05.25*

Besonders für die Erstellung von Audiobooks ist die eigens hierfür bereitgestellte Umgebung von *ElevenLabs* nützlich. In einem Textdokument kann die zu erzählende Geschichte eingefügt oder geschrieben werden und direkt mit passenden KI-Stimmen, welche aus einer großen Bibliothek ausgesucht werden können, versehen werden. Auch Soundeffekte können direkt, wie Satzzeichen im Text, eingefügt und per Text Prompt generiert werden.

### 5.8. MMAudio: Revolutionäre KI-gestützte Video-zu-Audio-Generierung

*MMAudio* bietet im Vergleich zu den anderen vorgestellten Tools, welche alle rein auf dem Text-to-Sound Modell basieren, die Möglichkeit, Videos hochzuladen. Die KI erstellt dann passende Geräusche, welche mit Hilfe eines Text Prompts und eines negativ Prompts beschrieben und angepasst werden können. Die Pro Version bietet zusätzlich die Möglichkeit, die Länge anzupassen und diverse Variablen zur Generierung zu verstellen. Zum Beispiel “Number of inference steps”, also die Anzahl der Schritte oder Iterationen, die die KI macht und durchläuft, um die Geräusche zu generieren; “CFG Strength”, kurz für “classifier-free guidance scale”, welche angibt, wie sehr die KI bei der Generierung dem Text prompt folgt; und den “Seed”, welcher auf -1 gestellt zufällig gewählt wird und den Startpunkt für die KI zur Generierung darstellt (vgl. Stablecog Inc. 2023; NightCafe Studio Pty Ltd o. D.; getimg.ai o. D.). Im Vergleich zu den anderen Tools bietet *MMAudio* nicht nur eine einzigartige Weise, Geräusche direkt in Verbindung mit dem Visuellen zu erstellen, sondern auch tiefgreifendere Möglichkeiten zur Manipulierung der generierten Geräusche. *MMAudio* gibt an, durch “Intelligente Umgebungsgeräuschsynthese” den Kontext der eingespeisten Szene zu analysieren und “geeignete Umgebungsgeräusche zu erzeugen”, welche “eine reichhaltige, immersive Klanglandschaft” schaffen (MMAudio.net 2025). Durch “Multi-Modale KI-Analyse”, bei der “gleichzeitig visuelle

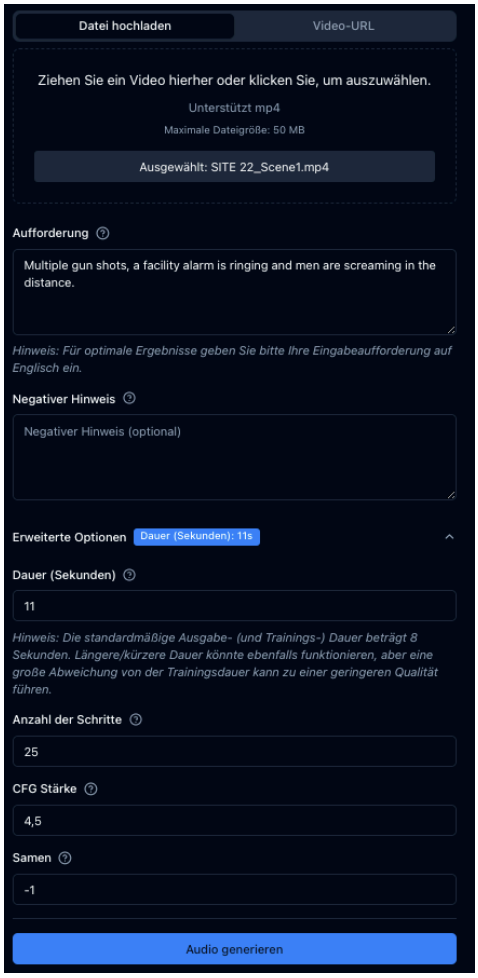


*Abbildung 16 MMAudio GUI https://mmaudio.net/de/text-to-audio abgerufen am 20.05.25*

Hinweise, Bewegung und Kontext” verarbeitet werden, wird außerdem versprochen, “perfekt synchronisierten Audio zu erstellen”(ebd.). Des Weiteren werden auch hier hohe “Studioqualität”, die Generierung innerhalb von Minuten versprochen und Audioanpassungen wie Lautstärke und Effekte geboten. Das Tool soll vielseitig einsetzbar sein. Egal ob Bildungsinhalte, Film- und Videoproduktionen, Spieleentwicklung, historische Filmverbesserung, Soziale Medien oder das Erzählen von Geschichten sollen damit möglich sein. Die Möglichkeit, die Sounds kommerziell nutzen zu können, wird auch hier garantiert; auch wenn wieder wenig Einblick in das KI machine-learning System geboten wird (ebd.).

### 5.9. Meta: AudioCraft

Auch der Großkonzern *Meta* lässt es sich nicht entgehen, im KI und Musik Business mitzumischen. Mit *AudioCraft* stellt *Meta* ein open source KI-Modell zur Verfügung, welches Audio in Form von Musik und Soundeffekten generieren kann. *AudioCraft* nutzt ebenfalls das Modell des Text Inputs zu Audio und verspricht “einen bedeutenden Fortschritt in der Entwicklung interaktiver KI-Systeme welcher es den Menschen ermöglicht, auf einfache und natürliche Weise mit KI-Modellen zu arbeiten” (Meta 2025a) [übersetzt durch Autor]. *Meta* nutzt hierfür die Formulierung “co-create with AI models”, was suggeriert, dass es nicht das Ziel ist, die Arbeit des Creators zu ersetzen, sondern zu unterschützen. Anders als die vorherigen Anbieter bietet *Meta* einen genauen Einblick in die KI (ebd.). *AudioCraft* wird beschrieben als Vereinfachung des Designs generativer Audio-Modelle im Vergleich zu früheren Ansätzen. Sowohl *MusicGen* als auch *AudioGen* basieren laut Website jeweils auf einem einzigen autoregressiven Sprachmodell, welches die Text Inputs verarbeitet (vgl. ebd.; Copet et al. 2023; Kreuk et al. 2022). Diese basieren wiederum auf so genannten Audio-Tokens. Diese Tokens stammen aus *Metas EnCodec-Codec*, der Audiosignale in mehrere parallele Token-Ströme umwandelt. (vgl. Défossez et al. 2022). Der Einfachheit wegen soll hier nicht weiter auf den technischen Hintergrund eingegangen werden, jedoch ist es wichtig, zu verstehen, was diese Herangehensweise bedeutet. *Meta* beschreibt die Methode als besonders effizient da diese Token-Ströme ineinander verflochten werden, und *AudioCraft* dadurch mit nur einem Modell effizient Audio, inklusive langer Abhängigkeiten im Klangverlauf, generieren kann. Der autoregressive Ansatz bedeutet vereinfacht gesagt, dass ein Stück nach dem anderen, zum Beispiel Sekunde für Sekunde oder Sample für Sample, generiert wird. Dabei schaut das Modell immer auf das, was es vorher selbst erzeugt hat, und macht dann

den nächsten Schritt (vgl. Noble 2025). Komponiert man beispielsweise selbst ein Lied Note für Note, hängt jede neue Note von der Note ab, die zuvor geschrieben wurde. Der Vorteil solcher Modelle ist die präzise Kontrolle über jedes einzelne Sample. Jedoch ist die Generierung eher langsam und anfällig für Fehler, da ein Fehler sich durch die Abhängigkeit generierter Samples durch das gesamte Ergebnis zieht. Für die Generierung von Audio in Form von Umweltgeräuschen bietet *Meta* ihr Modell *AudioGen: Textually-Guided Audio Genereation* (vgl. Meta 2025b). Auf einen Text Prompt wird einem ein 5-sekündiges Audiosignal ausgegeben, welches sich an die vorgegebene Textbeschreibung hält.

In Bezug auf die rechtliche Lage der Samples gibt *Meta* an, zumindest für *MusicGen* die KI mit eigener, lizensierter Musik trainiert zu haben (vgl. Weiß 2023). Da die KI zwar mit Einschränkungen öffentlich ist und auch so trainiert werden kann, ist jedoch nicht ganz klar, ob eine kommerzielle Nutzung bedenkenfrei gegeben ist.

### 5.10. Meta: Audiobox

*Meta* bietet mit *Audiobox* ein weiteres, vielseitiges KI-Modell zur Generierung von Audio [vlg. für folgenden Abschnitt Audiobox Team, Fundamental AI Research (FAIR) at Meta 2023; Meta 2024a; Meta 2023). Die Funktionen reichen von Generierung von Sprache und Geräuschen basierend auf Textbeschreibungen, zu Voice Cloning (Klonen der eigenen Stimme), Voice Restyling (Manipulation einer Stimme durch Text-Prompts) sowie Audioreparatur in Form von automatischer Rausch Entfernung. Der *Audiobox Maker* erweitert diese Möglichkeiten durch eine Oberfläche zur Kombination und Bearbeitung verschiedener Audioelemente wie Stimmen und Soundeffekte in einer Timeline (vgl. Meta 2024b). Dies ermöglicht die Erstellung komplexer Audiogeschichten, indem Nutzer*innen verschiedene Segmente arrangieren und bearbeiten können. Die Benutzeroberfläche erinnert dabei stark an klassische Audio-Editing-Software.

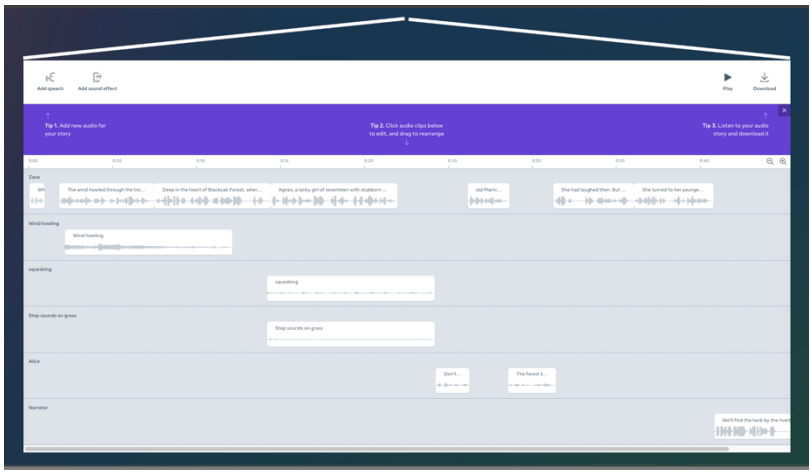

*Abbildung 19 Audiobox GUI https://audiobox.metademolab.com/maker abgerufen am 25.05.25*

Derzeit ist *Audiobox* hauptsächlich für nicht-kommerzielle Zwecke verfügbar, und die Nutzung unterliegt bestimmten Einschränkungen, insbesondere in Bezug auf die Replikation von Stimmen. *Meta* hat dazu Maßnahmen wie digitale Wasserzeichen und Stimmverifizierung implementiert, um Missbrauch zu verhindern (vgl. Kaminskyi 2023).

### 5.11. Google: SoundStorm

Mit *SoundStorm* liefert der Großkonzern *Google* ebenfalls ein vielversprechendes KI Modell zur Generierung von Geräuschen (vgl. für folgenden Abschnitt Google 2023). *Google* gibt in ihrem offiziell veröffentlichten Blog an, gleich wie *Metas Audiocraft,* rohes Audiomaterial zunächst in eine kompakte Repräsentation in Form von Audiotkens, mit Hilfe von Audio-Codes wie *SoundStream (ein Audio Codec von Google)* oder *EnCodec*, umzuwandeln (vgl. Google 2021). Die Tokens dienen hier erneut zur Erfassung von Klangmerkmalen und zeitlichen Strukturen und ermöglichen, mit Sprachmodellen Audio mit Text-to-Speech zu generieren. *Soundstrom* arbeitet jedoch, anders als *AudioCraft*, nicht autoregressiv. Es wird also nicht Token für Token, was wie zuvor erwähnt zu einer hohen Klangqualität, aber zu einer langen Inferenzzeit führt, sondern mehrere Teile gleichzeitig generiert. Die Inferenzzeit bezeichnet die Zeit, die die KI braucht, um ein Ergebnis aus dem bereits gelernten zu generieren (vgl. Google 2023; Hewlett Packard Enterprise Development LP 2025). *SoundStorm* umfasst eine speziell von Google angepasste Architektur für Audio-Tokens sowie ein paralleles Dekodierverfahren, das laut dem Blog von *MaskGIT* von einer Bild-KI inspiriert wurde (vgl. Chang et al. 2022). Die parallele Erzeugung von Tokens ermöglicht eine um das 100-Fache verkürzte Berechnungszeit, bei gleichbleibender, hoher Klangqualität.

*Googles Soundstorm* ist somit ein sehr effizientes Audio-Generationsmodell, was durch innovative Architektur hohe Qualität mit schnellem Output vereint. Ein weiteres Feature dieses Modells ist die Generierung von Dialogen durch die Integration von *SPEAR-TTS*, einem Text-to-Speech System, welches mehrere Stimmen zu einem Dialog generieren kann (vgl. Kharitonov et al. o. D.).

# 6. Manuelles vs. KI-gestütztes Sound Design

Um den Einfluss von KI auf die Qualität von Produktionen medialer Formate in Bezug auf die Narrativität und Immersion untersuchen zu können, soll im folgenden Teil dieser Arbeit ein praktischer Vergleich durchgeführt werden. Hierzu wird vom Autor eigens eine Produktion traditionell manuell vertont und diese dann mit der Vertonung mit Hilfe verschiedener KI-Tools verglichen. Zunächst werden im folgenden Kapitel die genutzten KI-Tools kurz genauer vorgestellt und die Gründe für die Verwendung dargelegt. Anschließend wird der Aufbau des Vergleichs genauer erklärt, die Beispielproduktion beschrieben und schließlich die Vorgehensweise der Bewertung vorgestellt.

## 6.1. Auswahl und Beschreibung der KI-Tools

Aus den zuvor vorgestellten KI-Tools soll eine engere Auswahl für den Vergleich genutzt werden (siehe Tabelle 1). Die anfängliche Google Recherche ergab eine Vielzahl an Ergebnissen von KI-Tools, welche Text-to-Sound Generierung versprechen. Bei genauerem Hinsehen wurde jedoch klar, dass viele Tools nicht das hielten, was sie versprachen, sondern vielmehr wie eine Art Abzocke oder oberflächliche KI-Geschäfte wirkten. Die verschiedenen Modelle wurden deshalb auf Basis der gebotenen Features und Anpassungsmöglichkeiten getestet, bewertet und für die praktische Arbeit ausgewählt. Zunächst fiel die Wahl auf das Video-zu-Audio-Generationsmodell von *MMAudio*. Dieses Modell ist das einzige, welches die Möglichkeit bietet, Videos hochzuladen und diese von der KI analysieren zu lassen. Diese Funktionsweise bietet eine einzigartige Möglichkeit für die Produktion, welche wiederum für den Vergleich sehr interessant ist. Auch die Sound KI in *Filmora 14* bietet durch die direkte Integration in die Video Editing Software einen einzigartigen Workflow, welcher wie bei *MMAudio*, insbesondere auf das bildbezogene Sound Design abzielt. Ein weiteres Modell mit einer einzigartigen Arbeitsweise ist *METAs Audiobox* mit dem Tool *Audiostory Maker*. Dieses Tool zielt anders als die vorherigen auf die Produktion von Audiobooks ab, wodurch der Bezug zum Bild gänzlich entfällt. Als letztes Tool zum Vergleich fiel die Wahl auf *ElevenLabs*. Auch hier gibt es die Option, ein Audiobook zu erstellen. Wie oben beschrieben, wird eingegebener Text von KI-Stimmen gesprochen und es gibt die Möglichkeit mit Text-to-Sound direkt KI-Geräusche zu generieren und einzufügen. Dieses Tool soll einerseits als leistungsstarkes und umfangreiches KI-Tool, als auch als Vertreter jeglicher, simpler Sound KIs in den Vergleich mit einbezogen werden. Die restlichen KI-Tools, wurden

auf Grund verschiedener Faktoren nicht ausgewählt. *Audiocraft* und *Soundstorm* bieten derzeit kein eigenständiges UI, wodurch die Nutzung einer eigenen lokalen Installation bedarf. *SFX Engine, Clipmove AI und MyEdit* wurden entweder auf Grund von nicht funktionierender Generierung (*Clipmove*) oder unzufriedenstellender Qualität bei der Generierung von Sounds (*SFX Engine, MyEdit*) ausgeschlossen. Für die Evaluierung der generierten Sounds der verschiedenen KI-Tools wurden jeweils einfache, grundlegende Text Prompts eingegeben. Beispielsweise „Slow footseps of boots on wooden floor“ oder „An old door creaking while being opened”. Diese simplen Geräusche konnten von den ausgeschlossenen Tools jedoch nicht zufriedenstellend generiert werden. Es scheiterte entweder an der Audioqualität, an zu vielen Artefakten, oder an zu großer Abweichung vom Textprompt.

| Tool | Text-to-Sound | Video-analyse | Bearbeitung im Tool | Lizenzfreie Nutzung | Stimmen-Generierung | Fokus |
|---|---|---|---|---|---|---|
| **MMAudio** | Ja | Ja | Nein | Ja | Nein | Szenenanalyse |
| **Filmora 14** | Ja | Nein | Ja, zumindest im Videoeditor | Ja | Nein | visuelle Bearbeitung |
| **Audiobox (Meta)** | Ja | Nein | Nur zeitliches Arrangement | Nur Demo, eingeschränkt | Ja | Audiostories |
| **ElevenLabs Studio** | Ja | Nein | teilweise, sonst extern nötig | Ja | Ja | Storytelling und Dialog |

*Tabelle 1 Auswahlkriterien*

### 6.2. Beschreibung der Beispielproduktionen

Auf Basis der getroffenen Auswahl von KI-Tools fiel die Wahl der Beispielproduktion auf zwei verschiedene Formate. Einerseits wird anhand einer kurzen Filmszene das bildbezogene Sound Design mit und ohne KI verglichen und andererseits, in einer zweiten Produktion, das rein auf Audioebene basierende Sound Design untersucht. Hierzu wird eine kurze Audiostory produziert, welche einmal mit, und einmal ohne KI, also manuell vom Autor selbst, vertont wird.

### 6.3. Vorgehensweise bei der Evaluation

Für die Evaluation der Qualität der gesamten Produktionen und der generierten Geräusche wurde eigens ein Bewertungsbogen erstellt (siehe Tabelle 2).

| Kriterium | Beschreibung | Punktebereich |
|---|---|---|
| 1. **Audioqualität** | Artefaktfreiheit, Klarheit, Frequenzbalance | 1–5 |
| 2. **Prompt - Übereinstimmung** | Wie gut entspricht der generierte Sound der Beschreibung im Prompt? | 1–5 |
| 3. **Natürlichkeit / Realismus** | Klingt der Sound wie aus einer echten Umgebung bzw. natürlichen Quelle? | 1–5 |
| 4. **Details / Nuancen** | Sind Feinheiten wie Materialklang, Rauminformation oder Bewegungsdynamik erkennbar? | 1–5 |
| 5. **Anpassbarkeit** | Wie gut lässt sich der Sound durch Variieren des Prompts oder Anpassungsmöglichkeiten im UI kontrollieren? | 1–5 |

*Tabelle 2 Bewertungskriterien*

Die verschiedenen Kriterien wurden außerdem mit einer Gewichtung versehen, da diese für die Gesamtqualität unterschiedlich ausschlaggebend sind (siehe Tabelle 3). Nach reiflicher Auseinandersetzung mit den theoretischen Annahmen aus 4.1 und wissenschaftlichen Erkenntnissen aus Kapitel 6.1, wurden Audioqualität und Prompt-Übereinstimmung mit je 25% am höchsten gewichtet, da diese am Ehesten zu einer zufriedenstellenden Generierung von Sounds führt. Eine klare, Artefakt-freie Generierung mit sauberem Klang, idealerweise in einem verlustfreien Dateiformat (WAV vs. MP3 o.Ä.), inklusive hoher Samplerate (44,1kHz vs. 48kHz oder höher) und Bitrate (z.B. 24bit in WAV oder 320kbps in MP3) sind bei professioneller Produktion wünschenswert, da diese die Grundlage professioneller Produktionen sind (vgl. Dickreiter 2014: 657-659). Insbesondere bei der Weiterverarbeitung sind verlustfreie Formate wichtig, um Artefakte zu vermeiden. Ebenso wichtig ist die Genauigkeit, mit der die KI die Intention und den Inhalt des Prompts trifft. Sobald die KI nicht zumindest nahe an den gewünschten Sound herankommt, ist das KI-Tool nicht zu gebrauchen, wodurch auch die Prompt-Übereinstimmung als sehr wichtig angesehen wird. Fast genauso wichtig ist die Qualität des Klangs in Bezug auf die Glaubwürdigkeit, beziehungsweise des Realismus. Klingt ein Schritt Sound auf Holz tatsächlich glaubwürdig nach Schritten auf Holzboden? Falls nicht verringert die fehlende Glaubwürdigkeit die Immersion der Produktion und damit die allgemeine Qualität. Die

Natürlichkeit wurde daher mit 20% gewichtet. Die zwei letzten Kriterien, Details und Anpassbarkeit, werden mit je 15% gewichtet. Da diese Kriterien als finale Feinjustierung dienen, um hohe Qualität und zielgenaue Generierung zu ermöglichen, werden diese am wenigsten stark gewichtet. Details und Nuancen werden idealerweise bereits durch eine hohe Promptübereinstimmung erzielt und auch die Anpassbarkeit ist zwar im Zuge eines guten Uis gut für die einfache Generierung, jedoch sollte dies ebenfalls bereits durch eine gute Umsetzung der Intention des Prompts erreicht werden.

| **Kriterium** | **Gewichtung (%)** |
|---|---|
| Audioqualität | 25% |
| Prompt-Übereinstimmung | 25% |
| Natürlichkeit / Realismus | 20% |
| Details / Nuancen | 15% |
| Anpassbarkeit | 15% |

*Tabelle 3 Gewichtungstabelle*

# 7. Konzeption und Umsetzung

In diesem Kapitel wird die Konzeption der Beispielproduktion dargelegt. Überlegungen für die Produktion und die Auswahl narrativer Formate werden beschrieben und begründet. Anschließend wird der Einsatz der KI-Tools und die Umsetzung dargelegt. Der kreative Prozess und die Ergebnisse werden daraufhin präsentiert. Außerdem werden Herausforderungen und deren Lösungen in der Durchführung der Produktion erläutert. Für das Schreiben passender Prompts wurde die Anleitung und Empfehlungen der *Harvard University* berücksichtigt (vgl. HUIT 2025).

## 7.1. Beispielproduktion 1: Vertonung von Filmszenen

In der ersten Beispielproduktion wurden die KI-Tools *MMAudio* und *Filmora 14* eingesetzt. Für beide Tools wurde die Standardversion gekauft, da nur so genügend Generierungsdurchgänge zur Verfügung standen. Ziel war die nachträgliche Vertonung ausgewählter Filmszenen. Als Ausgangsmaterial diente der unter der Lizenz Creative Commons BY-SA 3.0 auf YouTube veröffentlichte Kurzfilm „SCP: SITE 22 – Short Film“ des Nutzers MrKlay (vgl. MrKlay 2018; Creative Commons o. D.). Die Lizenz erlaubte die Nutzung und Veränderung des Materials unter Namensnennung, sodass eine Reduktion auf zentrale Szenen vorgenommen werden konnte. Die Auswahl der Szenen erfolgte nach inhaltlichen und technischen Kriterien. Sie sollten vielfältige Möglichkeiten zur Soundgestaltung bieten und sowohl

grundlegende als auch spezifischere Geräusche enthalten. Die final gewählten Szenen umfassten Pistolenschüsse, Explosionen, Schritte, Rascheln, Waldatmosphäre, Vogelgezwitscher, Kampfgeräusche, Monster- und Alienlaute, eine Tötung durch Erstechen sowie Helikoptergeräusche. Die Szenen wurden lose aneinandergereiht, wobei die ursprüngliche Handlung entfiel. Diese Reduktion diente dazu, den Fokus auf das Sound Design zu legen und das Potential der KI-Tools isoliert zu analysieren.

### 7.2. Beispielproduktion 2: Produktion und Vertonung einer Audiostory

Für die zweite Beispielproduktion wurde eine eigene kurze Audiostory mit dem Titel „Whispering Woods“ erstellt (siehe Anhang 1). Die Geschichte wurde mithilfe von *ChatGPT* verfasst, wobei gezielt Stellen eingebaut wurden, die akustische Gestaltungsmöglichkeiten bieten. Auch hier sollten sowohl grundlegende Geräusche als auch spezifischere Soundeffekte berücksichtigt werden. Als Referenz wurde die Geschichte vom Autor selbst eingesprochen. Die Vertonung erfolgte mit der kostenlosen Version von *Soundly*, die eine ausreichende Auswahl an frei verfügbaren Basisgeräuschen bereitstellt (siehe Anhang 2) (vgl. Soundly o. D.). Auf ein professionelles Tonstudio und umfangreiche Soundbibliotheken wurde bewusst verzichtet, um die KI-Tools unter minimalen Produktionsbedingungen zu testen. Für den Vergleich kamen die Tools *Audiobox* und *ElevenLabs* Studio zum Einsatz. Auch hier musste für die Nutzung von *ElevenLabs* die Standardversion gekauft werden, um genug Download-Credits zur Verfügung zu haben. Bei *Audiobox* handelt es sich um freie Demoversion, bei der es keine Möglichkeit zum Kauf einer Vollversion gibt.

### 7.3. Vertonung der Filmszenen und der Audiostory

Für die Vertonung mit *MMAudio* wurden die Filmszenen einzeln hochgeladen und jeweils mit einem spezifischen Text-Prompt versehen, der die gewünschten Geräusche beschrieb (für ein detailliertes Produktionsprotokoll siehe Anhang 3). Im Vergleich dazu bot *Filmora 14* eine integrierte Lösung, bei der die Szenen nicht einzeln hochgeladen werden mussten (siehe Anhang 4). Stattdessen wurden die Geräusche durch die Integration der KI in die Software direkt in der Bearbeitungsumgebung hinzugefügt. Auf weiterführende Bearbeitungen wie Panning, Effekte wie Hall oder EQ wurde bewusst verzichtet, da ausschließlich das Potenzial der KI-Tools getestet werden sollte. Dabei lag ein besonderer Fokus auf der Frage, wie gut die reine Soundgenerierung funktioniert und ob sie ausreicht, um qualitativ überzeugende Ergebnisse zu erzielen.

### 7.4. Vertonung der Audiostory

Das KI-Tool *Audiobox* bietet eine Timeline direkt im Browser. Alle Dialoge und Geräusche werden dort generiert und zeitlich arrangiert. Nach Auswahl einer passenden Erzählerstimme werden die Dialoge in die Textbox eingefügt und mit der ausgewählten KI-Stimme gesprochen. Die Geräusche wurden über entsprechende Text Prompts generiert und anschließend ebenfalls manuell im Zeitverlauf arrangiert (siehe Anhang 5). *ElevenLabs Studio* ermöglicht einen deutlich effizienteren Workflow. Hier kann der gesamte Text als Dokument hochgeladen werden. Die Software erkennt automatisch den Sprechertext und ordnet den verschiedenen Charakteren wahlweise bereits Stimmen zu. Aus einer großen Bibliothek an Stimmen werden die Stimmen anschließend nach Wunsch ausgetauscht und passende Geräusche per Text Prompt generiert (siehe Anhang 6).

### 7.5. Ergebnisse

Die im Zuge der praktischen Arbeit produzierten Beispiele werden im Folgenden vorgestellt und evaluiert. Diese können über diesen Link angehört und heruntergeladen werden.

Das Ziel, das vorliegende Material mit KI-Tools zu vertonen wurde grundsätzlich erreicht. Für die Kurzfilm-Vertonung entstanden zwei finale Versionen, die ausschließlich mit KI-generiertem Sound untermalt sind. Bei der Produktion mit *MMAudio* ist es für die vollständige Evaluation der Qualität jedoch nötig gewesen, eine bearbeitete, und eine nicht bearbeitete Version fertigzustellen. Der Grund dafür ist, dass *MMAudio* eigentlich verspricht, die eingespeisten Szenen vollständig vertonen zu können. Die Ergebnisse waren jedoch nicht zufriedenstellend. In der zweiten Version wurden nachträglich manuell von *MMAudio* generierte Audiospuren kombiniert, um ein besseres Ergebnis zu erzielen. Da für die Produktion mit *Filmora* bereits die Arbeit in der Editing Software voraussetzt, fand bei dieser Version ebenfalls etwas Bearbeitung, in Form von Kürzungen und Lautstärkeanpassungen der generierten Geräusche statt. Der Umfang wurde allerdings auf die Anpassung von Lautstärke und Länge und die zeitliche Anordnung der Geräusche beschränkt, um das Sound Design mit KI nicht zu viel manuell zu beeinflussen.

**Dateiübersicht:**

- SCP_SITE_22
    - SCP_SITE_22_OGAudio.mp4 Zusammengestellte Szenen mit Originalton
    - SCP_SITE_22_MMAudioAIAudio_unedited.mp4 Szenen mit *MMAudio* vertont und unbearbeitet
    - SCP_SITE_22_MMAudioAIAudio_edited.mp4 Szenen mit *MMAudio* vertont und bearbeitet
    - SCP_SITE_22_FIlmoraAIAudio.mp4 Szenen mit *Filmora 14* vertont
- Whispering_Woods
    - Whispering_Woods_Eigenproduktion.wav Eigenproduktion mit Sprachaufnahme durch Autor und manuelle Vertonung mit *Soundly*
    - Whispering_Woods_Audiobox.wav Produktion mit *Audiobox* (verkürzt auf Grund der Demoversion)
    - Whispering_Woods_ElevenLabs.wav Produktion mit *ElvenLabs* (Nachproduziert in *FL Studio* auf Grund der nicht korrekt funktionierenden Sound Einbindung in *ElevenLabs*)

Auch die Produktion des Audiobooks war weitestgehend erfolgreich. Nach der manuellen Eigenproduktion der Audiostory, welche als Vergleich und Ausgangspunkt dient, wurden zwei KI-Versionen produziert. Da *Audiobox* nur eine Demo Version ist, konnte hier jedoch nicht die vollständige Geschichte ausproduziert werden. Die Bearbeitung beschränkte sich aber ausschließlich auf das Interface, welches *Audiobox* bereitstellt. Es erfolgte keine Postproduktion in einer anderen Software. Mit *ElevenLabs* konnte die Kurzgeschichte vollständig ausproduziert werden. Leider war auch hier wieder die Bearbeitung in einer anderen Software nötig, um die Produktion abzuschließen.

Im folgenden Kapitel werden im Verlauf der Produktion aufgekommene Probleme und deren Lösungen beschrieben und erklärt.

### 7.6. Herausforderungen und Lösungen während der Umsetzung

Jedes Tool erforderte die Überwindung bestimmter Probleme (siehe Tabelle 4). Bei der Produktion mit *MMAudio* sorgten insbesondere viele technische und funktionale Probleme für Hindernisse, welche die Produktionsdauer verlängerten (siehe Anhang 3).

| Tool | Vorteile | Probleme / Einschränkungen |
|---|---|---|
| **MMAudio** | • Automatisierte Analyse und Vertonung von Szenen | • Schlechte Audioqualität bei längeren Sounds<br>• Unzuverlässige Prompt-Verarbeitung<br>• Abbrüche, Artefakte<br>• Kein vollständiges Sounddesign möglich<br>• Browser neu laden nach jeder Aktion |
| **Filmora 14** | • Direkt in Videoschnitt integriert<br>• Benutzerfreundliches Interface | • Begrenzte Prompt-Länge<br>• Schlechte Genauigkeit<br>• Lange Generierungszeit<br>• Schlechte Tonqualität (Artefakte, Loops)<br>• Eingeschränkte Bearbeitung und fehlerhafte Effekte |
| **Audiobox (Meta)** | • Fokus auf Storytelling<br>• Integrierte Timeline<br>• Dialoge möglich<br>• Simpler Workflow | • Nur Demo-Version (Längenbegrenzung)<br>• Eingeschränkte Stimmvielfalt<br>• Eingeschränkte Navigation<br>• Ungenaue Soundgenerierung<br>• Technisch instabil (Fehlermeldungen) |
| **ElevenLabs Studio** | • Sehr gute Stimmenqualität<br>• Umfangreiche Bibliothek<br>• Effizienter Workflow mit Textdokumenten | • Soundeffekte teilweise fehlerhaft<br>• Länge-Eingaben ignoriert<br>• Nachbearbeitung in externer DAW notwendig |

*Tabelle 4 Vor- und Nachteile der KI-Tools*

Die Qualität der generierten Sounds variierte stark. Da die KI nur auf acht Sekunden spezialisiert ist, sind längere Szenen nicht möglich, die Soundqualität nimmt rapide ab und die höhere Länge wird einfach nicht beachtet. Die Geräusche werden meist abrupt abgeschnitten. So beispielsweise die Explosion in der ersten Szene, welche nicht ganz ausklingt. Auch die Textprompt Verarbeitung funktionierte nicht fehlerfrei. Prompts wurden häufig ignoriert oder nur teilweise umgesetzt. Dabei ist jedoch oft nicht zu erkennen, ob die Umsetzung durch den Prompt oder durch Analyse des Videos erfolgte. Teilweise gab es keinen Unterschied zwischen den generierten Sounds mit oder ohne Prompt. Die ursprüngliche Hoffnung, die KI mit den Szenen zu füttern und vollständiges Sound Design generiert zu bekommen, wurde ebenfalls nicht erfüllt. Die KI erforderte die Fokussierung auf einzelne Geräusche pro Szene, da komplexere Textprompts mit Beschreibungen verschiedener Geräusche nicht funktionierten. Die

Prompts führten eher zum Gegenteil und schienen die KI zu überfordern, da die Ergebnisse oft Geräusche enthielten, die weder etwas mit der Szene, noch mit dem Prompt zu tun hatten. Dies führte auch zu dem Problem, dass mit Hilfe des Prompts Geräusche, die von Quellen kommen, welche nicht zu sehen sind, nicht generiert werden konnten. Die KI erzeugt ausschließlich Geräusche für sichtbare Elemente im Bild. In Szene neun führt das dazu, dass die Schritte des Monsters nicht mehr zu hören sind, sobald es den Bildausschnitt verlässt; obwohl es sich akustisch noch im Raum befinden müsste. Da nach *Chion* (1994) und *Souriau* (1997), wie oben beschrieben, das Monster auch außerhalb des Zusehenden noch zur Diegese, also zur Wirklichkeit des Films gehört, sollte es auch dann noch gehört werden können.
Zusätzlich dazu ergaben die Einstellungsmöglichkeiten wie CFG-Stärke, Schrittzahl und Seed keine positiven Änderungen. Es schien, als würden diese Parameter keine Änderung hervorrufen oder sogar eher für mehr Artefakte und Abweichungen sorgen. Deshalb wurden diese Parameter auch größtenteils auf dem Standardwert gelassen. Auffällig war auch, dass alle gewaltbezogenen Geräusche, wie auch zum Beispiel der Tod durch Erstechen in Szene acht, von der KI nicht vertont wurden. Es ist denkbar, dass hier inhaltliche Filter oder ethische Einschränkungen der KI greifen. Die browserbasierte Anwendung führte ebenfalls zu Problemen. Nach jeder Generierung musste die Seite neu geladen und die Videodatei wieder neu hochgeladen werden, wodurch alle Einstellungen und eingegebene Textprompts verloren gingen. All die Hindernisse führten zu einem hohen Zeitaufwand und erschwerten den kreativen Prozess erheblich. Für Szene eins dauerte es mit der Kürzung über eine Stunde, bis überhaupt eine Art zufriedenstellendes Ergebnis vorlag.

Die Produktion mit *Filmora 14* erforderte ebenfalls das Überwinden vieler technischer und funktioneller Probleme, welche den kreativen Prozess als auch die Qualität der Ergebnisse massiv beeinträchtigen (siehe Anhang 4). Komplexe Textprompts, welche sämtliche Geräusche der Szene beschreiben, konnten von der KI nicht umgesetzt werden. Die erzeugten Geräusche wirkten oft zufällig und hatten meist wenig mit der eigentlichen Intention des Prompts zu tun. Zusätzlich ist die Textlänge auf 100 Zeichen beschränkt, was längere, präzisere Beschreibungen weiter erschwert. Die Generierung der Sounds ist zudem sehr zeitaufwändig. Pro Durchlauf, werden standartmäßig drei, und optional bis zu sechs Varianten generiert. Die voreingestellte Länge beträgt dabei fünf Sekunden, und kann auf bis zu 30 Sekunden erhöht werden.

Da die Generierung bei drei Sounds mit je fünf Sekunden, trotz guter Internetverbindung, bereits lange dauert, wurde darauf verzichtet, um den Workflow nicht weiter zu stören. Trotz des aufwändigen Prozesses sind die Ergebnisse überwiegend unzufriedenstellend und die Audioqualität häufig unzureichend. Viele Sounds enthalten stark hörbare Artefakte oder sind schlichtweg fehlerhaft. Zusätzlich zu dieser minderen Qualität lässt auch die Genauigkeit zu wünschen übrig. Untergründe, beispielweise bei simplen Schrittsounds, wurden falsch interpretiert und statt Gras oder Kies Geräusche eines harten Steinbodens erzeugt. Dies erforderte häufige Neugenerierung, um authentische Sounds zu erlangen. Besonders dynamische Szenen, wie Szene sechs, in der die zwei Charakter miteinander kämpfen und viele verschiedene Geräusche des Untergrunds und der Schläge, der Klamotten und so weiter zu hören sein müssten, überforderte die KI komplett. Hier kam es oft zu unnatürlichen Sound Loops, welche an eine hängende CD erinnerten. Diese Sounds können ohne Nachbearbeitung nicht eingesetzt werden. Darüber hinaus traten unerwünschte räumliche Effekte wie ungefragte Panning-Effekte auf, die ohne dramaturgische Begründung zwischen den Kanälen wechseln. Auffällig war wie bei der Arbeit mit *MMAudio*, dass Gewaltsounds oft nur sehr verzerrt oder abstrakt generiert wurden. Zu all dem ergaben sich bei der Nachbearbeitung mit *Filmora* zusätzliche Schwächen. Übergänge, wie etwa Crossfades[10], wirkten oft unnatürlich, enthielten hörbare Fehler oder entsprachen nicht den erwarteten Kurven (linear, konstante Leistung, exponentiell).
Interessanterweise funktionierten Prompts, die mithilfe von ChatGPT erstellt wurden, wesentlich besser und führten zu besseren Ergebnissen. Dies stellte jedoch den Nutzen eines integrierten KI-Tools infrage, da der Mehrwert einer direkt in das System eingebetteten Prompt-Logik dadurch verloren geht.

Das KI-Tool *Audiobox* barg ähnliche Probleme in Bezug auf die Nutzungsfreundlichkeit und Audioqualität (siehe Anhang 5). Die Demoversion erlaubt nur eine maximale Audiolänge von 1 Minute und 45 Sekunden, was bedeutete, dass die Kurzgesichte nicht vollständig produziert werden konnte. Anders als bei der Produktion des Kurzfilms, lag der Fokus neben den Geräuschen auch auf den generierten Stimmen, welche den Kern einer Audiostory darstellen. Die verfügbaren vier Stimmen passten

[10] Ein Crossfade (engl. für Überblendung) bezeichnet die kontinuierliche Überblendung von einer Szene in die nächste, oder von einer Audiospur zur nächsten (vgl. Wikipedia-Autor*innen 2004).

stilistisch und tonal leider nicht zu den Figuren der Geschichte. Die Möglichkeit, neue Stimmen über Textprompts zu generieren, brachte keine zufriedenstellenden Resultate. Aus Datenschutzgründen wurde die Funktion, eigene Stimmen per Sprachaufnahme zu erstellen, nicht genutzt, weshalb Stimmen mehrfach eingesetzt werden mussten. Die Texteingabe war außerdem auf 125 Zeichen pro Prompt limitiert, wodurch die Erzähltexte einzeln und satzweise eingegeben werden mussten und den Prozess stark verlangsamten. Auch die Generierung von Dialogen erwies sich als fehleranfällig. Direkt zu Beginn wurde der Titel „Whispering Woods“ zu „Whispering Wood“ verkürzt. Erst durch das Hinzufügen eines weiteren „s“ am Ende konnte die korrekte Aussprache erreicht werden. Auch später wurden immer wieder Silben oder Buchstaben verschluckt, was eine erneute Eingabe eines angepassten Prompts und eine erneute Generierung bedeutete. Das Interface bot zudem nur eingeschränkte Navigationsmöglichkeiten. So war es etwa nicht möglich, durch Klicken auf die Timeline gezielt zu bestimmten Stellen zu springen, was die Bearbeitung und Platzierung einzelner Elemente erschwerte. Darüber hinaus wurden wieder Textprompts falsch interpretiert. Der Prompt „wind howling through trees“ führte zum Beispiel zu Wasserrauschen und Vogelgezwitscher. Auch technisch kam es immer wieder zu Fehlermeldungen, vor allem bei hoher Serverauslastung („Too many requests“) oder wenn bestimmte Prompts nicht verarbeitet werden konnten (ohne Begründung). Auch die Bearbeitung der Sounds war eingeschränkt. Generierte Audiospuren ließen sich weder herunterladen noch innerhalb der Anwendung kürzen. Dadurch stimmten viele Geräusche zeitlich nicht mit der Erzählung überein und konnten nicht angepasst werden. Auch das Verschieben der Clips in der Timeline funktionierte nur, wenn man sehr vorsichtig und langsam vorging.

Das Audiostory Tool in *ElevenLabs Studio* funktionierte im Vergleich am besten (siehe Anhang 6). Mit der Stimmen Generierung gab es keinerlei Probleme, jedoch funktionierte die integrierte Soundeffekt Generierung nicht richtig. Die Eingabe der Länge wurde ignoriert und die Qualität der Generierung in Bezug auf Audioqualität und Beachtung des Prompts schien schlechter als bei ersten Tests zu Auswahl der KI-Tools. Auf Grund dessen wurden die Geräusche nachträglich in der Text-to-Sound Umgebung separat generiert. Dadurch war jedoch eine Postproduktion in einer geeigneten DAW nötig. Die generierten Sounds aus der Audiostory waren im Verlauf aufzufinden und es war zu erkennen, dass die Funktion der Länge dort einen Bug hat

und nicht richtig funktionierte. Die erneute Generierung der Geräusche mit den gleichen Text Prompts führte direkt zu besseren Ergebnissen. Auch die Länge konnte zuverlässig eingestellt werden und die Prompt Influence ermöglichte es, Geräusche zu erzeugen, die nahe an der Beschreibung des Prompts lagen. Wie die anderen Tools hatte jedoch auch *ElevenLabs* hin und wieder Probleme die Anforderungen beim ersten Versuch direkt zufriedenstellend umzusetzen, auch wenn die Sounds beständig näher an die Anforderungen herankamen als die der anderen KI-Tools. Es war jedoch auch abseits der notwendigen Postproduktion nötig, einige Sounds nachträglich zu layern, um ein möglichst optimales Ergebnis zu erzielen. Da die Soundeffekte nur in MP3 Format verfügbar sind, ist eine nachträgliche Bearbeitung jedoch nicht ideal, da dies zu Artefakten und Minderung der Audioqualität führen kann.

# 8. Evaluation und Analyse

In diesem Kapitel wird auf Basis der Ergebnisse und der Bewertungen evaluiert, wie gut das Sound Design funktioniert hat. Um die Rolle von KI im Sound Design abschließend einordnen und bewerten zu können, werden dabei die Wirkung sowie die narrative, immersive und allgemeine Qualität analysiert.

## 8.1. Wirkung des Sound Designs auf Narrative und Immersive

Die vier KI-Tools, welche zum Einsatz kamen, haben ihre Versprechen größtenteils nicht einhalten können. Wie die Tabelle 5 zeigt, haben drei von vier Tools so schlecht abgeschnitten, dass die leichten Unterschiede in der Gesamtbewertung keine große Aussagekraft haben.

Lediglich das Tool von *ElevenLabs* konnte auf ganzer Linie überzeugen und die

| KI-Tool | Audio-qualität | Prompt-Übereinstimmung | Natürlichkeit / Realismus | Details / Nuancen | Anpass-barkeit | Gesamt-bewertung |
|---|---|---|---|---|---|---|
| MMAudio (unbearbeitet) | 2,00 | 1,00 | 1,00 | 1,00 | 0,00 | 1,10 |
| Filmora 14 | 1,00 | 2,00 | 0,00 | 1,00 | 1,00 | 1,05 |
| Audiobox | 1,00 | 1,00 | 1,00 | 1,00 | 0,00 | 0,85 |
| ElevenLabs | 4,50 | 4,00 | 4,00 | 4,00 | 4,00 | 4,13 |

*Tabelle 5 Ergebnisbewertung*

Anforderung fast komplett erfüllen. Die restlichen Tools konnten in den Kategorien Audioqualität und Prompt-Übereinstimmung, welche am schwersten gewichtet und damit am wichtigsten sind, nur eine Höchstwertung von zwei Punkten erreichen. Die Bewertung der Ergebnisse von *MMAudio* bezieht sich auf die unbearbeitete Version, da dies das originale Ergebnis der KI ist. Jedoch würde auch die bearbeitete Version

unwesentlich anders bewertet werden, da auch diese kein zufriedenstellendes Ergebnis darstellt. Die Bereitstellung der Sounds nur innerhalb des MPEG4 Format (*MMAudio*) spielte dabei jedoch die geringste Rolle, da auch *Filmora* trotz WAV-Format (44,1kHz, 16Bit) nicht überzeugen konnte. Bei *Audiobox* fehlt die Funktion, die Sounds herunterzuladen. Die Geräusche, die es nach mehrfacherer Generierung schließlich in den Film geschafft haben, enthalten immer noch entweder Artefakte oder andere Arten von Verzerrungen. Da hilft auch kein unkomprimiertes Format. Dabei muss jedoch auch erwähnt werden, dass für eine professionelle Produktion höhere Samplerates und Bitraten wünschenswert sind (z.B. mindestens 48kHz / 24Bit). Die Geräusche klingen außerdem meist sehr dünn (tiefe Frequenzen fehlen meist komplett) oder haben ein sehr störendes Highend (harsche hohe Frequenzen). Im Falle von *Audiobox* gilt diese Bewertung auch für die sehr dünn und roboterhaft klingenden Stimmen. Wie in Abbildung 9 zu sehen ist, sind ab 11kHz keine Informationen mehr enthalten, und das *Lowend*[11] fällt ebenfalls nach einem kleinen Peak[12] bei 100Hz schnell ab, wodurch die Stimmen zusätzlich sehr dünn klingen. Alle restlichen Sounds hatten so extreme Artefakte, oder waren so weit von der Prompt Intention entfernt, dass sie nicht weiterverwendet wurden. In den Kategorien Natürlichkeit / Realismus und Details / Nuancen beträgt die Höchstwertung sogar nur einen Punkt. Die Sounds der Tools klingen meist nur nach einer schwachen Annäherung an das, was im Prompt beschrieben wurde, aber sind weit von einem realistischen Sound entfernt. Details konnten nie wirklich integriert werden, was auch mit der niedrigen Wertung bei der Prompt-Übereinstimmung zusammenhängt. Lediglich einfache Klangkulissen wie Waldatmosphären mit Vogelzwitschern konnten erzeugt werden; mit einigen wenigen Details wie unterschiedlich lauten Vogelrufen und einem Geräuschbett, das in seiner Gesamtheit einen Wald auf klanglicher Ebene glaubwürdig darstellt. Generell klingen die Geräusche jedoch eher plump und wenig authentisch. In der letzten Kategorie der Anpassbarkeit konnten *MMAudio* und *Audiobox* auf Grund von nicht funktionierenden Reglern und Parametern, oder fehlender Funktionen wie Einstellung der Dauer, nicht Punkten. *Filmora* konnte auf Grund der Integration in eine Bearbeitungssoftware immerhin einen Punkt erzielen. Jedoch gehen diese Punkte damit eher auf das Konto der Software an sich, und nicht

[11] Das Lowend (engl. für tiefes Ende) bezeichnet im Audiobereich die tiefen Frequenzen eines Audiosignals, welche in der Musik meist durch Bassinstrumente erzeugt werden.
[12] Ein Peak (engl. Für Spitze) bezeichnet im Audiobereich die höchste Pegelspitze in einem Audiosignal (gemessen in Dezibel) (vgl. Tischmeyer 2025).

auf das des KI-Tools, welches ähnlich wenig Anpassung bot und lediglich die Möglichkeit einer Auswahl der Anzahl an Sounds bereitstellte. Diese Auswahl führte jedoch bei Vergrößerung der Anzahl der generierten Sounds zu einer erhöhen Wartezeit, was dieses Argument wieder ausgleicht.

Das Tool *ElevenLabs* hingegen konnte mit guter Audioqualität punkten, obwohl die Audiodateien nur in MP3 Format mit 44,1kHz und 128 kb/s zur Verfügung gestellt werden. Sowohl die generierten Geräusche als auch die generierten Stimmen überzeugen mit Artefakt freier Generierung und einem ausgeglichenen Frequenzspektrum, welches weder zu wenig oder zu viel Höhen bzw. Tiefen hat (siehe Abbildung 8).

*Abbildung 8 ElevenLabs*

*Abbildung 9 Audiobox*

Die Dialoge konnten im WAV-Format (44.1kHz, 16Bit, Mono) gespeichert werden, was ideale Voraussetzungen für eine weitere Bearbeitung bietet. Es ist aufgrund der nicht funktionierenden direkten Einbindung von Soundeffekten jedoch nicht gegeben, dass die generierten Sounds, welche teilweise Stereoinformationen beinhalten, auch im direkten Download, kombiniert mit den Dialogen, in Stereo verfügbar wären. Aufgrund der technischen Probleme wurde dies aber nicht in die Bewertung der Audioqualität miteinbezogen. In den restlichen Kategorien wurden jeweils vier Punkte vergeben. Die hohe Prompt Übereinstimmung ermöglicht eine gute Anpassbarkeit der Geräusche, wodurch auch Details und Nuancen einfach eingebunden werden können. Dies gilt auch für Stimmen, bei denen mit Hilfe von verschiedenen Parametern feine Abstimmungen der Stimmlage und Phonetik angepasst werden können. Dadurch können sogar Informationen über die Sprechende Person, wie es auch *Flückiger* bei echten Sprecher*innen beschreibt, bei den KI-Stimmen übermittelt werden (vgl. Flückiger 2011: 166). Auch die von *Crook* angesprochene Intimität und damit die Verbundenheit mit den Zuhörenden kann von KI erzielt werden (vgl. Crook 1999).

In Verbindung mit der hohen Audioqualität ergibt sich dadurch eine sehr hohe Authentizität und damit die erneut hohe Bewertung in der Kategorie Natürlichkeit/Realismus.
Auf Grund dieser Benotung kann der Einfluss auf die Narrative und Immersion der KI im Sound Design für Filme und Audiobooks wie folgt bewertet werden. Bei allen Tools, ausschließlich *ElvenLabs* kann kein positiver Einfluss festgestellt werden. Trotz erheblicher technologischer Fortschritte weisen die Tools deutliche Defizite in entscheidenden Bewertungskriterien auf. Die Gesamtbewertungen der Tools sind mit einer Höchstwertung von 1,10 sehr niedrig, was ihren kreativen Nutzen stark einschränkt. Die erzeugten Klänge wirken häufig generisch und wenig passend, da es an Klangqualität und Genauigkeit bei der Umsetzung der Eingabeanweisungen mangelt. Ebenso fehlt es den Sounds an natürlicher Wirkung und feinen Nuancen, wodurch zentrale Elemente für ein fesselndes, erzählerisches Erlebnis wie emotionale Tiefe und Atmosphäre, verloren gehen. Die begrenzten Anpassungsmöglichkeiten erschweren zudem die kreative Einflussnahme von Sounddesigner*innen, was eine gezielte dramaturgische Gestaltung nahezu unmöglich macht. Insgesamt tragen diese Tools dadurch kaum zur narrativen Qualität oder Immersion audiovisueller Produktionen bei.
Das Tool von *ElevenLabs* sticht mit einer Gesamtbewertung von 4,13 von 5 deutlich aus dem Feld der KI-basierten Tools hervor und demonstriert eindrucksvoll das Potenzial von KI im narrativen und immersiven Bereich. Besonders in den Kategorien Natürlichkeit, Audioqualität und Prompt Übereinstimmung liefert *ElevenLabs* präzise und stimmige Resultate, die sich nahtlos in erzählerische Kontexte einfügen. Die erzeugten Stimmen und Sounds wirken lebendig, glaubwürdig und emotional differenziert, was eine wesentliche Voraussetzung für authentische Immersion in Filmen und Audiobooks ist. Darüber hinaus bietet *ElevenLabs* umfangreiche Anpassungsmöglichkeiten, sodass Sounddesigner*innen gezielt dramaturgische Akzente setzen können. Dank dieser Stärken trägt das Tool maßgeblich zur atmosphärischen Verdichtung und zur Steigerung der narrativen Wirkung bei und markiert damit einen klaren Fortschritt gegenüber bisherigen Lösungen.

Nachdem die Ergebnisse der Tools auf ihre Wirkung auf Narrative und Immersive geprüft und evaluiert wurden, wird im folgenden Kapitel analysiert, welche

Unterschiede es bei den Tools in Bezug auf Kreativität, Effizienz und emotionaler Wirkung gibt.

### 8.2. Unterschiede in Kreativität, Effizienz und emotionaler Wirkung

Im Vergleich unterscheiden sich die analysierten KI-Tools nicht nur in der Qualität der Ergebnisse, sondern auch in technischer Ausführung, ihrer Auswirkung auf kreative Prozesse, Effizienz im Workflow und die emotionale Wirkung in den finalen Produktionen. Dabei zeigt sich deutlich, dass die jeweilige Herangehensweise an den Arbeitsablauf maßgeblich darüber entscheidet, wie sinnvoll das Tool in narrativen und immersiven Kontexten eingesetzt werden kann.

Der innovative Ansatz von *MMAudio* einer direkten Szenenanalyse zur automatisierten Soundgenerierung scheitert jedoch an der praktischen Umsetzung. Der kreative Prozess beschränkte sich auf die Überlegungen, wie die KI mit geeigneten Text-Prompts gefüttert werden kann, um optimale Ergebnisse zu erzielen und die anschließende Auswahl der Versionen. Jedoch wurde die Kreativität massiv von technischen Problemen beeinflusst. Die Einstellungsparameter, die *MMAudio* bietet, verschlimmerten wie bereits beschrieben die Ergebnisse eher. Außerdem sind die generierten Geräusche meist unbefriedigend und benötigen häufig manuelle Nachbearbeitung, was den kreativen Prozess weiter stark hemmt. Hinzu kommt eine fehleranfällige Website, auf der Szenen regelmäßig neu hochgeladen werden müssen. Dies wirkt sich negativ auf die Effizienz aus und macht *MMAudio* zu einem unzuverlässigen Werkzeug, das weder kreative Freiheit noch einen flüssigen Arbeitsablauf ermöglicht. Auch die emotionale Wirkung der Sounds bleibt aufgrund der niedrigen Qualität hinter den Erwartungen zurück.

Durch *Filmora*s direkte Einbindung der KI-Soundgeneration in eine bestehende Bearbeitungssoftware entfällt das Wechseln zwischen Anwendungen, was grundsätzlich die Effizienz und den Workflow verbessert. Dies setzt jedoch voraus, dass ohnehin eine Präferenz für die Arbeit mit *Filmora* besteht. Dagegen wird dieser Vorteil durch eine lange Generierungsdauer wieder relativiert. Das führt zu Wartezeiten, die den kreativen Fluss unterbrechen, was sowohl Effizienz als auch kreative Spontanität einschränkt. Die KI-Funktion war im Vergleich zu *MMaudio* dadurch außerdem weniger automatisiert, sodass pro Geräusch ein eigener Prompt nötig war und nicht Szene für Szene vorgegangen werden konnte. Dies erforderte etwas mehr manuellen Aufwand, erleichterte jedoch durch die direkte Integration den Überblick und die Koordination der Tonspuren. Die emotionale Wirkung der Sounds

ist solide, profitiert aber weniger von der KI als von der manuellen Gestaltung durch die Nutzer*innen.

Die webbasierte Timeline von *Audiobox* hingegen ermöglicht in der Theorie eine visuelle Organisation von Soundelementen, welche den Workflow und schließlich die Kreativität unterstützen. In der Praxis leidet jedoch sowohl die Usability als auch die Systemstabilität: Die Seite hängt häufig, was den Arbeitsfluss massiv stört. Da das Tool keine Möglichkeit zur Anpassung der Sounds bietet, ist die fehlende Option, generierte Dateien nicht einzeln herunterladen zu können, besonders problematisch und schränkt die kreative Bearbeitung stark ein. Nach Auswahl einer Erzählerstimme aus den ohnehin limitierten Optionen mussten die Texte satzweise eingefügt werden, da die Textbox nur begrenzte Zeichenmengen akzeptiert. Der kreative Prozess bestand vor allem in der Auswahl passender Prompts, der Wahl einigermaßen geeigneter Stimmen und der Abstimmung der Soundeffekte. Diese strukturellen und technischen Defizite behindern sowohl den kreativen Ausdruck als auch die Effizienz im Workflow. Die emotionale Wirkung bleibt ebenfalls blass, da Anpassungsmöglichkeiten für stimmige Dramaturgie fehlen und Stimmen roboterhaft wirken.

Der webbasierte Ansatz von *ElevenLabs* mit der klaren „All-in-One-Studio“-Strategie hingegen erzielt überzeugende Ergebnisse. Die intuitive Bearbeitung von Stimmen und Sounds im Kontext eines narrativen Textes ermöglicht ein hohes Maß an kreativer Kontrolle. Die umfangreichen Einstellmöglichkeiten für Parameter wie Tempo und Stil boten zudem hohen kreativen Spielraum. Die generierten Sounds lassen sich jederzeit einsehen und bei Bedarf neu laden, was einen sehr flexiblen Umgang mit allen generierten Sounds ermöglicht. Obwohl derzeit die direkte Einfügung von Sounds in das Dokument noch nicht vollständig funktioniert, bietet das Tool bereits jetzt ein sehr effizientes und nutzerfreundliches Arbeitsumfeld. Die gute Funktionalität relativiert sogar den Wunsch nach einer DAW Software Lösung, wie sie *Filmora* bietet. Durch die hohe Anpassbarkeit und natürlich wirkenden Geräusche gelingt es *ElevenLabs*, emotionale Tiefe und Atmosphäre in Geschichten effektiv zu transportieren.

Zusammenfassend lässt sich feststellen, dass Tools wie *MMAudio* und *Audiobox* aktuell mehr Hürden als Hilfen darstellen, während *Filmora* ein durchwachsenes Bild zeigt. Vergleicht man die in den Protokollen aufgeführten Bearbeitungszeiten der Eigenproduktion mit der mit KI, ist keine Zeitersparnis festzustellen (siehe Anhänge 2-

6). *ElevenLabs* hingegen demonstriert, wie durchdachtes Tooldesign und funktionale Integration kreative, effiziente und emotional wirkungsvolle Audioproduktion auf KI-Basis durchaus ermöglichen können. Um abschließend einordnen zu können, wie gut oder schlecht KI-Tools im Sound Design eingesetzt werden können, werden im folgenden Kapitel die Vor- und Nachteile von KI-gestütztem Sound Design, im Vergleich zum manuellen Ansatz dargelegt.

### 8.3. Kritische Reflexion: Vor- und Nachteile von KI-gestütztem Sound Design im Vergleich zum manuellen Ansatz

Die Einbindung künstlicher Intelligenz in das Sound Design eröffnet neue Möglichkeiten für die Audioproduktion, insbesondere, wenn es um Geschwindigkeit, einfache Bedienbarkeit und Automatisierung geht. Im direkten Vergleich zu traditionellen, manuellen Methoden treten jedoch grundlegende Fragen bezüglich Qualität, Kontrolle und künstlerischer Ausdrucksfähigkeit zutage.

Ein wesentlicher Vorteil KI-basierter Tools liegt in ihrer Effizienz. Sie ermöglichen das schnelle Erzeugen von einzelnen Sounds, ohne dass aufwändige Aufnahmen oder langwierige Recherchen in Sounddatenbanken nötig sind. Gerade für kleinere Projekte ohne Budget, oder die Entwicklung erster Konzepte kann dies, unter Auswahl der richtigen Tools und Beachtung einiger Punkte (siehe Anhang 7), eine Zeitersparnis bedeuten. Wichtig hierbei zu beachten ist jedoch, dass eine Person, welche einen routinierten Umgang mit Soundbibliotheken hat, wahrscheinlich unwesentlich länger braucht als bei der Arbeit mit einer generativen KI. Andererseits kommt hinzu, dass viele dieser Anwendungen auch ohne tiefgehende Fachkenntnisse im audio- oder produktionstechnischen Bereich genutzt werden können. Das senkt die Einstiegshürden und macht Sound Design für nicht-professionelle Anwender*innen oder Teams mit begrenzten Ressourcen zugänglicher.

Demgegenüber steht jedoch ein spürbarer Verlust an künstlerischer Kontrolle und Detailgenauigkeit. KI-generierte Sounds basieren auf vorgegebenen Algorithmen, deren Kreativität und Kontextverständnis begrenzt sind. Besonders in emotional anspruchsvollen Szenen fehlt es oft an der feinen Nuancierung, die erfahrene Sounddesigner*innen gezielt einsetzen. Komplexes Layering, dramaturgisch motivierte Klanggestaltung oder gezielt gesetzte Pausen und Spannungsbögen machen den entscheidenden Unterschied. Auch die Integration in bestehende narrative Strukturen gelingt der KI meist nur eingeschränkt, da ihr das tiefere Verständnis für den emotionalen Gehalt einer Szene fehlt. *MMAudio* beispielsweise

beinhaltet eine sehr interessante und innovative Idee, jedoch sind die Ergebnisse weit von nutzbaren Geräuschen und noch viel weiter von einer professionellen Produktion entfernt.

Ein weiteres Manko vieler KI-Tools ist die eingeschränkte Anpassungsfähigkeit. Während manuelles Sound Design nahezu unbegrenzte Bearbeitungsmöglichkeiten bietet, stoßen Nutzer*innen bei KI-generierten Inhalten oft auf technische oder gestalterische Grenzen. Individuelle Anpassungen sind häufig nur eingeschränkt oder mit Umwegen möglich, was die kreative Flexibilität merklich einschränkt und Workflows unterbricht. Andererseits zeigt der Erfolg einzelner, ausgereifter Tools, wie etwa *ElevenLabs*, dass eine gelungene Verbindung von Automatisierung und kreativer Kontrolle mit qualitativ hohem Output durchaus realisierbar ist.

# 9. Diskussion

Nach der Evaluation und Analyse der Wirkung des Sound Designs und Bewertung der allgemeinen Qualität werden in diesem Kapitel das kreative Potential, Grenzen und ethische Fragen beim Einsatz von KI im Sound Design diskutiert.

## 9.1. Kreatives Potenzial und Grenzen von KI im Sound Design

Künstliche Intelligenz erschließt im Sound Design ein kreatives Potenzial, das noch vor wenigen Jahren kaum vorstellbar gewesen wäre. Sie erlaubt es, innerhalb von Sekunden die verschiedensten Geräusche zu erschaffen, für die früher aufwendige Studioarbeit oder langwierige Field-Recordings nötig gewesen wären. Besonders in der Phase der Ideenfindung, beim Entwickeln von Prototypen oder beim Anfertigen erster Soundskizzen erweist sich KI als wertvolles Werkzeug. Sie inspiriert mit unerwarteten Klangvorschlägen und erweitert so die kreativen Möglichkeiten, insbesondere für Menschen ohne Vorerfahrung im Sound Design. Sie fungiert sozusagen als Impulsgeber. Durch die Eingabe eines noch vage formulierten Textprompts, der in irgendeiner Weise die Vision und Idee, die sich zu diesem Zeitpunkt noch nur im Kopf befindet, widerspiegelt, kann schnell eine erste Version eines Klangbilds erzeugt werden. Dabei werden vielleicht auch Klänge genutzt, die man selbst so nicht eingesetzt hätte, die wiederum zu neuen Ideen führen können. Die schnelle Produktion von Prototypen oder Demoversionen einer Vertonung ermöglicht es, Klangideen direkt hören zu können, ohne gleich eine ganze Session mit Field Recording oder Synthese durchführen zu müssen. Die limitierte Kontrolle über die eigentliche Erzeugung der Klänge und somit die Abgabe der Produktion an eine künstliche Intelligenz führt zu immer neuen Ergebnissen. In etwa vergleichbar mit zwei

Personen, die ein und dasselbe Video nachvertonen und trotz gleicher Anweisungen unterschiedliche Ergebnisse erzielen. Gerade in Bezug auf kreatives Sound Design können manche Modelle Klänge erzeugen, die jenseits traditioneller Synthese oder Sample-basierter Methoden liegen. Das kann stilistisch sehr reizvoll sein, zum Beispiel für Genres im Bereich des Sci-Fi oder Horror, in denen oft stark verzerrte, seltsam, oder nicht-von-dieser-Welt klingende Geräusche eingesetzt werden. In diesen ersten Schritten der Produktion können KI-Tools also als kreativer Helfer eingesetzt werden und die Ausarbeitung und Konkretisierung von Ideen vereinfachen, beschleunigen und vielleicht sogar durch Vorschläge neuer Klangideen in der Qualität verbessern.
Besonders interessant wird KI, wenn sie in erzählerische Kontexte eingebunden wird. Tools wie *ElevenLabs* demonstrieren, wie sich durch intelligente Kontextanalyse dynamische Klangverläufe von generierten Stimmen erzeugen lassen, die einen harmonischen Gesamteindruck machen, und gleichzeitig die Möglichkeit zur Einbindung qualitativ hochwertiger, generierter Geräusche bieten. In solchen Anwendungen agiert KI nicht nur als Generator, sondern als kreativer Partner im Produktionsprozess. Für grundlegende emotionale Atmosphären wie Spannung oder Ruhe, oder Geräusche zur Setzung von Aktzenten kann KI überzeugende Klangbilder liefern, was besonders für Hörspiele oder Podcasts ein großer Vorteil ist.
Allerdings stößt dieses kreative Potenzial an klare Grenzen. KI handelt nicht aus künstlerischer Motivation, sondern verarbeitet Datenmuster aus ihrem Trainingsmaterial. Echte Kontextreflexion, bewusste dramaturgische Entscheidungen oder gezielte Brüche mit Erwartungen bleiben bis auf ein paar Ausnahmen der KI größtenteils verwehrt und sind menschlichen Sounddesigner*innen vorbehalten. Während ein erfahrener Gestalter zum Beispiel gezielt Stilbrüche einsetzt, um Irritation oder Tiefe zu schaffen, produziert KI meist das Erwartbare und Wahrscheinliche. Dadurch entstehen häufig glatte, vorhersehbare Klangergebnisse, die die kreative Ausdruckskraft einschränken. Dadurch spielt in späteren Phasen der Produktion die Qualitätssicherung eine große Rolle. Gerade bei kommerziellen Produktionen müssen gewisse Qualitätsstandards eingehalten und eine genaue Umsetzung der Wünsche erzielt werden. Die Qualität von KI-generierten Sounds hängt jedoch maßgeblich vom eingesetzten Modell und der Gestaltung des Prompts ab. Die Tools und deren erzeugte Sounds erreichen in Bezug auf Präzision, Auflösung und Detailtiefe noch nicht durchgehend das Niveau professioneller Produktionen. Ein zentrales Problem bei der Nutzung von Text-to-Sound Tools besteht außerdem in der Vielfalt der

Klangtexturen und der Steuerbarkeit. KI-Modelle erzeugen oft interessante und überraschende Sounds, bieten jedoch bislang nur eingeschränkte Kontrolle über Details. Wer beispielsweise ein exakt zwei Sekunden langes, rhythmisch präzises, synchronisiertes Soundlogo benötigt, stößt mit diesen Systemen schnell an Grenzen. Für solche spezifischen Anforderungen ist der klassische Weg der manuellen Produktion mit DAWs, Synthesizern, Samplern und eigenen Recordings der KI weiterhin überlegen. Während bei KI-generierten Sounds keine Probleme bezüglich ungewollten Rauschens auftreten, gibt es bei anderen technischen Aspekten wie zum Beispiel der Frequenzbalance noch Einschränkungen. Viele KI-generierte Sounds weisen ein unausgewogenes Frequenzspektrum auf, Tiefen fehlen und hohe Frequenzen klingen harsch, was sie im professionellen Mix schwer kontrollierbar macht. Ein weiteres Problem ist die begrenzte Anpassbarkeit und Interaktivität vieler KI-Tools. Zwar lassen sich personalisierte Klänge erzeugen, doch sind diese oft nur eingeschränkt editierbar oder lassen sich nicht nahtlos in bestehende Workflows integrieren. Das erschwert die individuelle Feinarbeit, die für emotionale und dramaturgische Präzision in audiovisuellen Projekten unerlässlich ist.

Abschließend lässt sich festhalten, dass Künstliche Intelligenz vor allem als experimentelles und unterstützendes Werkzeug im Sound Design das kreative Spektrum bereichern kann. Ihre Stärke liegt in der schnellen Generierung von Ideen und Inspiration, weniger in der tiefgreifenden, dramaturgisch gesteuerten Ausarbeitung von auditiven Klanglandschaften. Um das volle Potenzial auszuschöpfen, sind hybride Arbeitsprozesse nötig, bei denen die KI als Impulsgeber dient, die finale kreative Kontrolle jedoch beim Menschen verbleibt. Die KI-Tools ersetzen zu diesem Zeitpunkt nicht das menschliche Ohr und den künstlerischen Feinsinn. Sie eignen sich jedoch hervorragend als kreatives Ausgangsmaterial oder als zusätzliche Layer in komplexeren Soundkompositionen. Nur so lassen sich hohe Audioqualität, narrative Tiefe, emotionale Wirkung und künstlerische Originalität auch beim Einsatz von KI-Tools sicherstellen. Mit zunehmender Modellreife, höherer Qualität und mehr Manipulationsmöglichkeiten dürfte der praktische, kreative Nutzen solcher KI-Tools in der Zukunft allerdings massiv zunehmen. Auf Basis dieser Einschätzung des kreativen Potentials von KI-Tools wird im folgenden Kapitel eingeschätzt, inwiefern KI-Tools auf dem heutigen Stand nur als Co-Creator in den kreativen Prozess einbezogen werden können.

### 9.2. KI-Tools als Co-Creator im kreativen Prozess

Der aktuelle Stand der Entwicklung macht deutlich: Künstliche Intelligenz ersetzt den kreativen Beruf des Sound Designers nicht, kann aber, wie es *Meta* bereits beschreibt, als wirkungsvoller Co-Creator einen bedeutenden Beitrag leisten (vgl. Meta 2025a). Gerade im Sound Design, wo Intuition, Erfahrung und ein feines Gespür für narrative Zusammenhänge gefragt sind, bleiben, wie oben gefolgert, menschliche Fähigkeiten unverzichtbar. Ein zentraler Vorteil von KI liegt darin, wiederkehrende Aufgaben und zeitintensive Recherchen zu übernehmen. Sie kann auf Basis kurzer Textvorgaben innerhalb von Sekunden erste Soundentwürfe liefern, Klanglandschaften vorschlagen oder einfache Atmosphären generieren. Für Sound Designer*innen bedeutet das Entlastung in frühen Produktionsphasen oder bei funktionalen Elementen, sodass mehr Raum für konzeptionelle und kreative Arbeit entsteht. Als Co-Creator eignet sich KI besonders für explorative Prozesse. Sie kann inspirieren, Varianten anbieten oder durch algorithmisch erzeugte Überraschungen neue kreative Wege eröffnen. Besonders für Sc-Fi Geräusche kann die KI dadurch interessante Klangtexturen generieren, die als Basis für weitere Bearbeitung dienen kann. Voraussetzung ist jedoch, dass KI nicht als fertige Lösung betrachtet wird, sondern als Rohmaterial, das durch menschliche Gestaltung veredelt und in einen dramaturgischen Kontext eingebettet werden muss. Der Ansatz der Co-Kreation verdeutlicht zudem die Verantwortung der menschlichen Gestalter*innen. Sounddesigner*innen müssen beurteilen, ob ein KI-generierter Sound dramaturgisch passt, emotional überzeugt oder dem Narrativ widerspricht. Dieses kritische Auswählen, Überarbeiten und Integrieren in das Gesamtwerk ist ein kreativer Prozess, der weit über die bloße Generierung hinausgeht und ausschließlich vom Menschen geleistet werden kann. Ähnlich zum Prozess bei der Arbeit mit Sound Archiven und Bibliotheken. Geräusche müssen gesucht und bearbeitet werden, um für das spezifische Projekt ideal zu passen.

Zusammengefasst liegt der größte Nutzen von KI nicht in der vollständigen Automatisierung, sondern in der intelligenten Kooperation mit dem Menschen. Besonders bei künstlerisch ambitionierten und emotional vielschichtigen Produktionen bleibt der Mensch mit seinem kreativen Gespür unverzichtbar. Als Co-Creator unterstützt die KI Sound Designer*innen dabei, effizienter, vielseitiger und experimentierfreudiger zu arbeiten, ohne deren kreative Rolle zu schmälern. Um

herauszufinden in welchem Ausmaß beim Einsatz von KI eventuelle Risiken oder Herausforderungen aufkommen, wird dies im folgenden Kapitel auf Basis der dargelegten Forschung und Ergebnisse auch in Bezug auf ethische Überlegungen untersucht.

## 9.3. Herausforderungen, Risiken und ethische Überlegungen beim Einsatz von KI im Sound Design

Nach der praktischen Analyse und dem Vergleich verschiedener KI-Tools im Sound Design wird deutlich, dass der Einsatz künstlicher Intelligenz im audiovisuellen Bereich nicht nur kreative Möglichkeiten eröffnet, sondern auch zahlreiche ethische und rechtliche Herausforderungen mit sich bringt. Besonders in sensiblen Bereichen wie Journalismus oder Dokumentarfilm bleibt der verantwortungsvolle Umgang mit KI-generiertem Audiomaterial eine zentrale Aufgabe. Risiken, die sich beim Einsatz von Ton in Medienformaten ergeben, bleiben zunächst auch beim Einsatz von KI erhalten und ethische Überlegungen oder Copyright Bedenken spielen auch bei KI-generiertem Inhalt eine Rolle. Jedoch stellt sich die Frage, ob durch den Einsatz von KI diese ausgeweitet werden, weitere Risiken hinzukommen oder ob es sogar zu einer Risikominderung kommen kann. Im folgenden Kapitel sollen Überlegungen diesbezüglich dargelegt und mögliche Antworten gegeben werden.

## 9.4. Verstärkung von bestehenden Risiken

Wie im Kapitel 4.7 bereits dargelegt wurde, ist der Unterschied zwischen Fiktion und Realität insbesondere im Journalismus oder dokumentarischen Formaten wichtig zu erkennen. Das Risiko von Desinformation oder gar bewusster Manipulation kann durch KI erhöht werden. Beispielsweise ist die Generierung von Stimmen durch KI-Tools wie *ElevenLabs* teilweise so gut, dass sie im ersten Moment täuschend echt erscheint. Die Fähigkeit, Stimmen auf Knopfdruck zu erzeugen, zu modulieren oder gezielt an bestimmte Sprecherprofile anzupassen, eröffnet neue kreative Wege, etwa für Hörbücher, Voiceovers oder digitale Assistenzen. Gleichzeitig birgt diese Technik erhebliche Risiken. Stimmen können täuschend echt nachgebildet und kontextuell manipuliert werden, was die Glaubwürdigkeit und Authentizität medialer Inhalte gefährden kann. Werden synthetische Stimmen eingesetzt, ohne dies offenzulegen, oder gezielt menschliche Stimmen imitiert, um Falschinformationen zu verbreiten, entsteht ein erhebliches ethisches Problem. Vergleicht man das mit dem in Kapitel 4.7 beschriebenen Vorfall mit Orson Welles Hörspiel im Rundfunk, so ist die Tragweite einer Täuschung wie sie in Kombination mit KI möglich wäre, nicht vorzustellen.

Transparenz und Kennzeichnungspflicht sind hier entscheidend, um Vertrauen zu erhalten und Missbrauch vorzubeugen. Insbesondere in Kombination mit anderen Sounds können so bewusst qualitative Makel im Ton kaschiert werden, wodurch die Grenzen zwischen real und künstlich erzeugt weiter verschwimmen. Die glaubwürdige Fälschung von Stimmen realer Personen kann in Bezug auf die Verbreitung von Fake News ein ernsthaftes Problem darstellen. Während die Veränderung von Bildern, oder die alleinige Anschuldigung der Nutzung von KI bereits häufiger zur Verbreitung von Fake News in der Politik, beispielsweise im Wahlkampf zwischen Kamala Harris und Donald Trump, führte, können auch künstlich generierte Interviews oder Sprachnotizen aktiv zur Verbreitung von Fake News genutzt werden (vgl. Judson 2024; Bundeszentrale Für Politische Bildung 2024). Bei der Nationalratswahl der Slowakei im Jahre 2023 wurde eine Sprachaufnahme über *Facebook* geteilt, auf welcher zwei Stimmen zu hören waren. Angeblich sollen Michal Šimečka, Vorsitzender der liberalen Partei Progressive Slowakei, und Monika Tódová von der Tageszeitung Denník N. darüber diskutieren haben, wie man die Wahlen manipulieren könnte. In der Aufnahme ist zu hören, wie über den Kauf von Stimmen der marginalisierten Roma-Minderheit des Landes diskutiert wird (vgl. Meaker 2023). Sogenanntes Deepfake Audio oder Voice Cloning ist nicht nur billiger als die Generierung gesamter Videoinhalte, sondern auch einfacher zu faken, da es sich allein um Informationen über die Audioebene handelt und visuelle Makel, wie schlecht synchronisierte Lippenbewegungen, nicht zur Identifizierung des Fakes beitragen können (vgl. Lowy Institute 2024). Durch bessere Fakes wird der Verlust der Nachvollziehbarkeit immer größer und gravierender wodurch KI-generierte Inhalte immer schwerer zu verifizieren sind, wenn keine klare Kennzeichnung erfolgt. Der bedachte Einsatz von KI, und die kritische Hinterfragung von Inhalten, welche von KI produziert sind, bleibt wichtiger Bestandteil beim Umgang mit KI-Tools zur Produktion von Inhalten. Gerade bei politischen Fragen sollte besonders Acht gegeben werden. Ein einschlägiges Beispiel hierfür ist das chinesische KI-Tool *DeepSeek.* Dieser Chatbot funktioniert gleich wie jeder andere KI Chatbot, jedoch fällt bei bestimmten Fragen auf, dass die Personen, die hinter *DeepSeek* stecken, nicht die gleichen Ansichten haben, wie man es aus der europäischen Perspektive vielleicht gewohnt ist. Wie eine Recherche von *The Guardian* im Januar 2025 zeigte, werden bei manchen Fragen die politischen Ansichten der Entwickler*innen und deren Regierung als Antworten ausgespielt oder bestimmte Themen gar nicht kommentiert. Auf Fragen zu politischen Protesten oder

Revolutionen gegen die chinesische Regierung findet *DeepSeek* keine Antwort. Auch zu Fragen nach Taiwans Unabhängigkeit antwortet die KI eindeutig aus chinesischer Sicht und gibt entgegen anderer Chatbots aus anderen Ländern an, dass Taiwan eindeutig zu China gehört und nie unabhängig war (vgl. Lu 2025).

Zusammenfassend bleibt festzuhalten, dass insbesondere politische Themen in Bezug auf KI ein heikles Thema sind. Je nach Herkunft und politischer Lage werden KIs darauf getrimmt, bestimmte Themen nicht, oder nur mit Zensur oder bestimmten Vorgaben zu kommentieren. Auf Basis dessen werden im folgenden Kapitel weitere ethische und rechtliche Überlegungen beim Einsatz von KI-Tools im Sound Design angestellt.

### 9.5. Ethische und rechtliche Überlegungen

Weitere zentrale ethische und rechtliche Überlegungen betreffen das Urheberrecht und das Eigentum. Datenschutz und rechtliche Grundlagen stellen eine Herausforderung dar, da sie bei vielen KI-basierten Soundtools oft unklar sind. Wer besitzt das geistige Eigentum an KI-generierten Sounds oder Stimmen? Mit welchem Datenmaterial wurde die KI trainiert? Wessen Stimme wurde, gegebenenfalls unrechtsmäßig, synthetisiert und kommerziell weiterverwendet? Ähnlich wie bei der Generierung von Bildern lässt sich bei Geräuschen und Stimmen häufig nicht nachvollziehen, ob urheberrechtlich geschützte Inhalte, sensible Sprachaufnahmen oder Persönlichkeitsrechte betroffen sind. Da künstliche Intelligenzen mit Hilfe von bestehenden Daten trainieren, ist oft unklar, ob das Ausgangsmaterial rechtlich geschützt ist. Die vorgestellten Tools geben meist keine genauen Angaben bezüglich der Art und Weise wie und womit die KIs trainiert wurden, wodurch auch Angaben bezüglich der urheberrechtsfreien Nutzung der generierten Sounds fraglich sind. Auch hier lässt sich das anhand von generierten Stimmen gut erklären. KI-Stimmen müssen trainiert, oder direkt geklont werden. Wenn dies auf Basis realer Stimmen passiert, können Persönlichkeitsrechte verletzt werden. Im Bereich der Musik kam dies bereits häufiger vor und führte zu rechtlichen Verfahren, welche die Debatte des Datenschutzes weiter anheizten. Der Rapper *Drake* wurde beispielsweise vom Nachlass des Rappers *2Pac* verklagt, weil er in einem Song einen Teil von einer Stimme rappen lies, welche der von *2Pac* stark ähnelte (vgl. Reinelt 2024). Andererseits wurde auch *Drake* Opfer einer künstlichen Generierung seiner Stimme (vgl. im Folgenden Snapes 2023). Im Jahr 2023 veröffentlichte ein *TikTok* User

namens *Ghostwriter977* einen Song des Rappers *Drake* und dem Sänger *TheWeeknd* zunächst auf *TikTok*, später aber auf auch Streaminganbietern wie *Spotify*. Dieser Track war jedoch kein offizielles Feature der beiden Künstler, sondern ein rein durch KI erstellter Song, welcher die Stimmen täuschend echt generierte. Der Song konnte in kurzer Zeit viele Streams sammeln und wurde nach wiederholter Anfrage der *Universal Music Group (UMG)* letztlich von allen Plattformen entfernt. Auch nachträgliche „Fake Features" von mehreren Rappern führen zu immer häufiger aufkommenden Remixen bereits existierender Songs durch nachträgliche Einfügung von Parts anderer Artists. *UMG* forderte daraufhin bessere KI-Erkennung und schnellere Entfernung solcher Inhalte, um Urheberrechte und Persönlichkeitsrechte zu schützen. Das Thema rund um Urheberrechte und Datenschutz ist nicht nur ein ethisches, sondern auch ein juristisches Problem. Ohne klare Angaben zur Herkunft des Trainingsmaterials ist die Nutzung solcher Sounds, insbesondere in kommerziellen Produktionen, rechtlich riskant, auch wenn von den Anbietern angegeben wird, dass alle Geräusche frei von Copyright sind. Medienproduzierende müssen sich daher aktiv mit den Nutzungsbedingungen der Tools auseinandersetzen und im Zweifel auf rechtlich einwandfreies Material zurückgreifen. Unklare Lizenzlagen können zu Abmahnungen oder Imageschäden führen.
Im Bereich der Audiobooks und Podcasts ist außerdem eine authentische Darstellung von Akzenten und Stimmcharakteristiken wichtig. *Google* äußert sich hierzu beispielsweise in der Dokumentation zur KI *Soundstorm* wie folgt (vgl. hierzu Google 2023): Die Verzerrungen, die in den Trainingsdaten enthalten sind, können die Hörbeispiele beeinflussen. Jedoch soll es durch gezieltes Prompting steuerbar sein, wie bestimmte Sprachmerkmale verantwortungsvoll generiert werden. *Google* bezieht sich hierbei auf die Bedeutung umfassender Analysen von Trainingsdaten und ihre grundlegenden Prinzipien in Bezug auf die Arbeit mit KI, damit beispielsweise rassistische Elemente keinen Einzug in die KI finden und in der Entwicklung der KI damit verantwortungsvoll umgegangen wird (vgl. Google o. D.).

Es lässt sich festhalten, dass die KI eine Realität abbildet, wie sie ihr von Menschen vermittelt wird. Dabei können politische Haltungen einfließen, was bei fehlender kritischer Reflexion leicht zu einer verzerrten Interpretation der Wirklichkeit führen kann. Fake News und Meinungsbildung auf gefälschten Fakten werden dadurch zur großen Gefahr, insbesondere wenn diese durch die oben genannten Fälschungen von

Interviews, Stimmen und allgemein journalistischen Inhalten gestützt werden. Aus diesem Grund wird im folgenden Kapitel detaillierter auf die veränderte Verantwortung der Produzent*innen beim Einsatz von KI in der Produktion eingegangen.

### 9.6. Veränderte Verantwortung in der Produktion

Die Verantwortung gegenüber dem Publikum rückt durch die oben genannten Punkte stärker in den Fokus. KI macht es zunehmend schwer, zwischen realen und künstlich erzeugten Klangquellen zu unterscheiden. Durch diese neuen Möglichkeiten, Inhalte zu produzieren, gilt auch eine veränderte Verantwortung in der Produktion. Einerseits verschiebt sich die redaktionelle Verantwortung, wenn KI Klangentscheidungen trifft. Etwa durch Auto-Mixing oder automatische Stimmenerkennung *(Filmora 14)* müssen Produzierende besonders darauf achten, diese Entscheidungen kritisch zu prüfen. Stimmen müssen weiterhin verständlich sein, fälschlich erkannte Stimmen müssen geprüft und eine gegebenenfalls bereits erfolgte Bearbeitung rückgängig gemacht werden, um eine konstant hohe Qualität und Authentizität zu gewährleisten. Auch die Kennzeichnung von KI-generierten Inhalten ist ein wichtiges Element bei der Inklusion von KI-Tools in der Produktion medialer Inhalte. Durch KI-Kennzeichnung auf Social Media ist beispielsweise erkennbar, wenn ein Inhalt künstlich erzeugt worden ist. In Audioproduktionen sollte, ähnlich wie bei visuellen Effekten, klar deklariert werden, welche Sounds KI-generiert wurden. Dies ist wie oben erwähnt insbesondere bei der Verwendung von künstlichen Stimmen sehr wichtig.

Zusammenfassend bleibt festzuhalten, dass, ähnlich wie beim Journalismus generell, ethische Leitlinien für den Einsatz von KI-generierten Sounds etabliert werden müssen. Dazu zählen zum Beispiel:

- Offenlegung: Zuhörer*innen müssen erkennen können, ob Klänge oder Stimmen synthetisch erzeugt wurden.
- Kontextsensibilität: Besonders bei kulturellen oder historischen Themen sollte der Klang mit Respekt gegenüber dem dargestellten Inhalt gewählt werden.
- Narrative Integrität: Sounds dürfen die Bedeutung oder Wahrnehmung einer Geschichte nicht manipulativ verzerren.

Diese Regeln lassen sich jedoch auf manuelles Sound Design, sowie Sound Design mit KI anwenden. In der Summe bleibt festzuhalten, dass KI narrative Welten erweitern und verändern kann, sich damit aber auch die Anforderungen an mediale Verantwortung verändern. Sounddesigner*innen, Journalist*innen und

Produzent*innen stehen vor der Aufgabe, nicht nur kreativ, sondern auch ethisch reflektiert zu arbeiten. Die Tools müssen transparent eingesetzt, klar gekennzeichnet und rechtlich abgesichert sein. Nur so kann KI ein Werkzeug für immersive, inklusive und glaubwürdige Erzählformen werden und nicht zu einem Instrument der Täuschung oder Manipulation. Um nicht nur die negativen Aspekte der KI zu beleuchten, wird im folgenden Kapitel untersucht, wie KI dazu beitragen kann, Risiken zu mindern.

### 9.7. Risikominderung durch KI

Jedoch ergeben sich bei der Nutzung von KI in der Produktion auch Möglichkeiten zur Risikominderung. Einerseits können automatisierte Prüfungen und Transparenztools dafür sorgen, gefakte oder KI-generierte Inhalte zu erkennen, ganz nach dem Motto: KI erkennt KI. Gleichermaßen kann KI trainiert werden um zum Beispiel extremistische oder irreführende Inhalte automatisiert zu blockieren. Allerdings sind diese Mechanismen keineswegs unfehlbar. Auch hier wäre eine erneute manuelle Prüfung nötig. Jedoch weitet die KI die zuvor erwähnte Erhöhung der Inklusion durch sprachlich synthetisierte Inhalte weiter aus. Barrierefreie Inhalte können schneller erstellt und einfacher zugänglich gemacht werden.

Zusammenfassend lässt sich festhalten, dass sich die Risiken und ethischen Überlegungen mit dem Einsatz von KI verschieben. Manche Probleme (wie Fake Audio) werden verschärft, andere (wie Barrierefreiheit) können verbessert werden. Zentraler Teil der Produktion mit KI bleibt, dass ein transparenter, reflektierter Einsatz von KI-Technologie, besonders im Journalismus und der Dokumentation sichergestellt werden muss, sodass der hohe Anspruch an Authentizität gewahrt wird.

## 10. Fazit und Ausblick

Die vorliegende Arbeit hat deutlich gemacht, dass Sound Design in narrativen und immersiven Medien weit über einen rein technischen Aspekt hinausgeht. Sound Design ist fester Bestandteil audiovisueller Kommunikation und trägt maßgeblich dazu bei, glaubwürdige, emotionale und eindrucksvolle Welten zu erschaffen. Sei es im Film, im Journalismus, in Podcasts oder in interaktiven Medien wie Videospielen. Die auditive Ebene vermittelt Informationen, erzeugt Bedeutung und löst Emotionen aus. Sie ergänzt das Bild, erweitert die erzählerische Dimension und intensiviert das Gesamterlebnis. Dabei zeigt sich, dass Sound Design ein vielschichtiger, kreativer Prozess ist, der sowohl technisches Know-how als auch gestalterische Sensibilität verlangt.

Mit dem Einzug künstlicher Intelligenz in die Medienproduktion eröffnen sich neue Chancen, aber auch Herausforderungen. Die Arbeit hat die kreativen Möglichkeiten aktueller KI-Tools im Sound Design untersucht und gezeigt, dass KI grundsätzlich in der Lage ist, realistische und kontextgerechte Geräusche zu erzeugen. Theoretisch verspricht der Einsatz von KI eine höhere Effizienz, kreative Unterstützung und die Möglichkeit, durch die schnelle Erzeugung schwer auffindbarer Klänge oder die Erweiterung bestehender Soundbibliotheken bisherige Grenzen zu überwinden.
Die praktischen Experimente dieser Arbeit haben jedoch auch die aktuellen Grenzen der KI-Tools aufgezeigt. Schwierigkeiten bei der Umsetzung von Prompts, eingeschränkte Audioqualität, mangelnde Synchronität und begrenzte Anpassungsoptionen machen deutlich, dass KI derzeit kein vollwertiger Ersatz für manuelles Sound Design ist. Besonders bei komplexen Szenen, in denen es auf Timing, emotionale Tiefe oder feine erzählerische Nuancen ankommt, stößt KI schnell an ihre Grenzen. Die viel versprochene Zeitersparnis wird dadurch nur bedingt erreicht, oder gar nicht erreicht. Auch ethische Fragen in Bezug auf kulturelle Bedeutungen von Sounds, oder der gezielte Einsatz zur Manipulation, bleiben weiterhin relevant und gewinnen durch künstlich erzeugte Klänge und Stimmen an Bedeutung.
Die Leitfrage dieser Arbeit lautete, welches kreative Potenzial KI im Sound Design bietet und wie sie Narrative und Immersion beeinflusst. Die Ergebnisse zeigen ein zweigeteiltes Bild. Einerseits kann KI, richtig eingesetzt, die Immersion steigern. Etwa durch individuell generierte Klanglandschaften, die durch die Kombination vieler Klänge interessante Sounds und Stimmungen erzeugen können, oder durch direkte Erzeugung von Klängen auf Basis visueller Inhalte. Andererseits bleibt die emotionale Wirkung, insbesondere in sensiblen oder dramaturgisch anspruchsvollen Szenen, oft hinter der Qualität von handgemachtem Sound Design zurück.

Für die Zukunft lässt sich festhalten, dass KI voraussichtlich eine unterstützende Rolle im kreativen Prozess einnehmen wird. Sound Designer*innen können diese Technologien nutzen, um Routineaufgaben zu erleichtern, Inspiration zu finden oder schneller Prototypen zu erstellen. Die zentrale Erkenntnis dieser Arbeit lautet daher, dass KI zwar als unterstützendes Werkzeug sinnvoll in kreative Prozesse eingebunden werden kann, aber nicht das menschliche Urteilsvermögen, die kreative Intuition und das narrative Feingefühl ersetzt. Sie eignet sich besonders für einfache,

wiederkehrende Aufgaben und kann bei der richtigen Wahl des Tools den Arbeitsablauf beschleunigen und neue Impulse liefern. Die Verantwortung für die Wirkung, Gestaltung sowie die kreative und ethische Ausgestaltung des Sounds bleibt jedoch weiterhin beim Menschen. Abschließend lässt sich außerdem festhalten, dass die Bedeutung des Sound Designs oft unterschätzt wird. Die Auseinandersetzung mit KI-Tools hat gezeigt, dass Klang ein mächtiges Medium bleibt, das umsichtig genutzt werden sollte. Die Zukunft des Sound Designs wird geprägt sein vom Zusammenspiel zwischen Menschen und KI, von Intuition und Algorithmus, von kreativer Freiheit und technischer Unterstützung. Auf Grund diesem prognostizierten Wachstum der KI im Audiobereich werden im folgenden Kapitel Empfehlungen gegeben, welche beim Einsatz von KI im Sound Design eine Sichere und Effiziente Arbeit ermöglichen.

### 10.1. Empfehlungen für den Einsatz von KI im Sound Design

Nach den Tests und Analysen verschiedener KI-Tools für Sound Design gibt es einige Tipps, die helfen können, diese Technik besser zu nutzen. Wichtig ist zuerst, genau zu überlegen, was man mit dem Tool machen will. Braucht man Inspiration für neue Klänge? Oder sucht man fertige Sounds auf Anforderung, die direkt genutzt werden können? Geht es um Sprachaufnahmen, Hintergrundgeräusche oder das Nachvertonen von Szenen? Wer diese Fragen beantwortet, weiß besser, welches Tool geeignet ist. Je nachdem, wie das Tool in den eigenen Workflow eingebunden werden soll, sollte man prüfen, ob es die richtigen Dateiformate, Exportmöglichkeiten und die passende Nutzeroberfläche bietet. Besonders bei der Wahl zwischen Online-Tools und Tools, die direkt in Bearbeitungssoftware eingebaut sind, lohnt es sich, auf die Kompatibilität zu achten. Wenn das Tool nicht zum eigenen Workflow passt oder die Formate nicht stimmen, kann das schnell zu Problemen führen. Es lohnt sich, die Tools nach Möglichkeit zu testen, um auch die verschiedenen Längen der Generierungszeit mit einplanen zu können.

Ein weiterer wichtiger Punkt ist die rechtliche Seite. Viele Tools sagen nicht genau, mit welchen Daten sie trainiert wurden oder wie die Lizenzen für die Sounds aussehen. Wer die Sounds für professionelle oder kommerzielle Projekte nutzen will, sollte sich unbedingt mit den Nutzungsbedingungen beschäftigen. Es ist auch sinnvoll, die verschiedenen Abo-Modelle und Funktionen der Tools zu vergleichen. Oft sind wichtige Funktionen wie bestimmte Exportformate oder Bearbeitungsmöglichkeiten nur in teureren Versionen verfügbar. Wer sich vorher informiert, kann späteren Ärger

vermeiden (siehe Anhang 7 für eine Checkliste für den verantwortungsvollen Einsatz von KI-Tools).

Es zeigt sich, dass viele KI-Tools noch keine Sounds liefern, die man ohne Nachbearbeitung direkt verwenden kann. Es ist daher besser, keine zu hohen Erwartungen zu haben. KI ist aktuell keine vollautomatische Lösung. Sie hilft am meisten, wenn sie als Unterstützung genutzt wird, zum Beispiel um Ideen zu bekommen oder verschiedene Varianten auszuprobieren. Wie bereits festgestellt: Am besten funktioniert KI als Co-Creator. Sie macht Vorschläge und eröffnet neue Möglichkeiten, aber die Feinarbeit und das Endergebnis kommen vom Menschen. Um zu sehen, welche Möglichkeiten und Potential in der Zukunft der KI möglich sein können, wird im folgenden Kapitel auf die zukünftige Forschung und Entwicklung der KI eingegangen.

### 10.2. Weiterführende Forschung und Entwicklung

KI-Technologien für Audioproduktion entwickeln sich ähnlich wie KIs in anderen Bereichen sehr schnell. In Zukunft werden sie noch mehr Möglichkeiten zur Generierung und Bearbeitung von Audio bieten. KI-Tools zur Sound Organisierung wie mit *COSMOS*, Filterung einzelner Instrumente wie in *FL Studio*, intelligente Rauschentfernung mit *lalal.ai*, Synthesizer Plug-Ins mit KI gestützter Re-Synthese von Klängen mit *Synplant 2* von *Soniccharge*, oder Text-to-Instrument Technologie mit *SynthGPT* von *fadr* sind im Audiobereich bereits Gang und gäbe (vgl. OmniSale GmbH o. D.; Lidström 2025; Fadr o. D.). Besonders spannend ist jedoch die Aussicht, dass in Zukunft komplette Audiogeschichten automatisch erstellt werden könnten, die authentisch und real klingen. Dazu gehören Stimmen, Geräusche und Musik. Das kann für kleine Firmen oder Projekte hilfreich sein, jedoch stellt dies ein großes Problem für Sprecher*innen dar. Ihre Arbeit könnte durch KI ersetzt werden und Aufträge immer schwerer zu finden sein, da der Einsatz von KI billiger und effizienter ist (vgl. KI-Statement | Verband deutscher Sprecher:innnen 2024).

Ein weiteres Thema ist die automatische Videoanalyse zur Sounderzeugung. Hier gibt es schon erste Versuche, wie zum Beispiel bei *MMAudio*. Bisher funktioniert das noch nicht gut. Wenn die KI aber besser versteht, was im Bild Geräusche macht, wie lange diese Geräusche zu hören sind, in welchem Raum man sich befindet, und auch die Prompt Genauigkeit besser wird, kann daraus ein starkes Werkzeug werden. Zwar

sind arbeitsmarktbezogene Bedenken derzeit noch von geringer Bedeutung, mit Blick auf die Zukunft jedoch durchaus relevant.

Auch die Nachbildung von Stimmen, das sogenannte Voice-Cloning ist interessant. Nutzer können ihre eigene Stimme klonen und im kreativen Prozess nutzen. Das macht die Audioproduktion flexibler und persönlicher. Es gibt aber Risiken beim Datenschutz. Die Kontrolle über die eigene Stimme ist wichtig. Man muss genau wissen, welche Datenschutzregeln gelten, vor allem wenn die Anbieter aus Ländern kommen,, in denen andere Regeln gelten als in Deutschland. Diese könnten eventuell weniger strenge Beschränkungen für das Training von künstlichen Intelligenzen beinhalten, als man es in Deutschland vielleicht gewohnt ist. In Verbindung mit diesen Themengebieten ist außerdem zu erwähnen, dass in dieser Arbeit fokussiert auf die Perspektive der Produzent*innen eingegangen wurde. Weitere Schritte in der Forschung dieses Themas ist Perspektive der Rezipient*innen zu beleuchten: Wie klingt KI in narrativen Medienformaten? Wie gut funktioniert die Immersion wirklich?

Ein weiteres spannendes Feld ist die Verbindung von KI mit Such- und Archivierungstechnologien. Tools wie *COSMOS* von *Waves Audio Ltd.* gehen hier erste Schritte zur Sortierung und Kategorisierung von Sounds. Sound Design Tools wie *Krotos Studio* nutzen innovative Ansätze zur Erstellung passender Geräusche. Die Kombination aus dem real-time Recording UI von *Krotos Studio* und KI Soundgenerierung könnte ein sehr starkes Sound Design Tool werden. Wenn KI bestehende Sounds analysiert und sortiert und in Zukunft sogar neu kombiniert oder daraus neue Klänge baut, entstehen starke Werkzeuge für die Organisation und Produktion von Audio.

Wahrscheinlich wird KI auch immer mehr direkt in Musik- und Videoprogramme eingebaut. Erste Beispiele wie das Stem-Splitting und AI-Mastering in *FL Studio*, oder die Soundgenerierung in *Filmora* zeigen, wie das aussehen kann. So muss man nicht mehr zwischen vielen Programmen wechseln, sondern kann direkt in einer Software mit KI arbeiten (vgl. Image Line Software 2023). Auch im Jahr 2025 zeigt sich bei *FL Studio* Updates, dass immer mehr KI-Tools Einzug in die Software nehmen (vgl. Image Line Software 2025).

Bisher werden Technologien wie binaurales Audio, die besonders realistische Klangwelten schaffen, kaum mit KI verbunden. Wenn KI in Zukunft auch räumliche Klänge automatisch und echt erzeugen kann, gibt es viele neue Möglichkeiten. Das ist zum Beispiel für Virtual Reality, Gaming oder Dolby Atmos spannend. Gerade dort, wo

Atmosphäre und Immersion wichtig sind, könnte KI-gesteuertes 3D-Audio viel verändern.

In Bezug auf die Barrierefreiheit von Medienformaten sind ebenfalls weitere Interessante Entwicklungen zu erwarten. Der *DareFightingICE*-Wettbewerb ist beispielsweise ein wegweisendes Forschungsprojekt aus dem Jahr 2022, das Sounddesign und KI in einem barrierefreien Kampfspiel zusammenführt. Das zentrale Ziel besteht darin, das Spiel für sehbehinderte Menschen zugänglich zu machen. Dazu gibt es zwei parallellaufende Wettbewerbe. Im *Sound Design Track Wettbewerb* entwickeln die Teilnehmenden akustische Signale, die es ermöglichen, das Spielgeschehen vollständig über den Hörsinn zu erleben. Im *Blind AI Track Wettbewerb* programmieren Teams KIs, die das Spiel ausschließlich anhand von Toninformationen steuern. Beide Bereiche ergänzen sich, fördern die Inklusion im Gaming und eröffnen neue Perspektiven für die KI-Forschung mit Schwerpunkt Audio (vgl. Khan et al. 2022).

Insgesamt ist klar: KI hat im Sound Design noch viel Potenzial. Es bleibt spannend, wie sich Technik, Kreativität und Ethik in den nächsten Jahren weiterentwickeln und welche neuen Werkzeuge für die Audioproduktion entstehen, und sich die Qualität der Ergebnisse verbessern wird. Es ist zu erwarten, dass KI-Systeme leistungsfähiger, präziser und besser steuerbar werden. Gerade deshalb ist es wichtig, das Bewusstsein für die Rolle des Klangs in der Mediengestaltung zu schärfen. Die eigentliche Herausforderung besteht dabei nicht nur in der technischen Umsetzung, sondern vor allem darin, KI und Klang mit Sorgfalt, Kreativität und Verantwortung einzusetzen.

## Literaturverzeichnis


Altenmüller, Eckart/Günther Bernatzky (2015): Musik als Auslöser starker Emotionen, in: *Musik und Medizin: Chancen für Therapie, Prävention und Bildung*, Springer, S. 221–236, [online] doi:10.1007/978-3-7091-1599-2_15.

Altenmüller, Eckart/Reinhard Kopiez (2005): Schauer und Tränen: zur Neurobiologie der durch Musik ausgelösten Emotionen., in: Claudia Bullerjahn/Heiner Gembris/Andreas C. Lehmann (Hrsg.), *Musik: gehört, gesehen und erlebt. Festschrift Klaus-Ernst Behne zum 65. Geburtstag.*, Hannover, Deutschland: Institut für Musikpädagogische Forschung der Hochschule für Musik und Theater, S. 159–181.

Audiobox Team, Fundamental AI Research (FAIR) at Meta (2023): Audiobox: Unified Audio Generation with Natural Language Prompts, Meta (Hrsg.), ai.meta, [online] https://ai.meta.com/research/publications/audiobox-unified-audio-generation-with-natural-language-prompts/ [abgerufen am 02.05.2025].

Bordwell, David/Janet Staiger/Kristin Thompson (1985): *The Classical Hollywood Cinema*, London, Vereinigtes Königreich.

brainrulesbook (2008): McGurk Effect, [YouTube] https://www.youtube.com/watch?v=I1XWDOwH47Y [abgerufen am 23.03.2025].

Bundeszentrale Für Politische Bildung (2024): Wenn der Schein trügt – Deepfakes und die politische Realität, bpb.de, [online] https://www.bpb.de/lernen/bewegtbild-und-politische-bildung/551578/wenn-der-schein-truegt-deepfakes-und-die-politische-realitaet/ [abgerufen am 09.05.2025].

Burke, Myles (2023): „I had no idea I'd become a national event": Orson Welles on the mass hysteria of The War of the Worlds, bbc.com, [online] https://www.bbc.com/culture/article/20231027-behind-the-broadcast-orson-welles-on-the-mass-hysteria-of-the-war-of-the-worlds [abgerufen am 12.07.2025].

Cambridge University Press (2025): newsflash, in: *Dictionary.Cambridge*, [Wörterbucheintrag] https://dictionary.cambridge.org/de/worterbuch/englisch/newsflash [abgerufen am 11.07.2025].

Carlsson, Sven E (o. D.): Sound designer, Filmsound, [online] https://filmsound.org/terminology/designer.htm [abgerufen am 06.03.2025].

Chang, Huiwen/Han Zhang/Lu Jiang/Ce Liu/William T. Freeman (2022): MaskGIT: Masked Generative Image Transformer, in: *The IEEE Conference On Computer Vision And Pattern Recognition (CVPR)*, [online] https://masked-generative-image-transformer.github.io/.

[zugegriffen am 30.04.25]

Chilvers, Alex/Yixue Quan/Kirk N Olsen/William Forde Thompson (2023): The effects of cultural source sensitivity on music appreciation, in: *Psychology Of Music*, Bd. 52, Nr. 3, S. 284–304, [online] doi:10.1177/03057356231201846.

Chion, Michel (1994): *Audio-Vision: Sound on Screen*, Choice Reviews Online, New York, Vereinigte Staaten: Columbia University Press.

Clasen, Mathias, PhD (2022): Horror movies use sounds that we naturally associate with danger, in: *Psychology Today*, 17.09.2022, [online] https://www.psychologytoday.com/us/blog/why-we-love-horror/202209/why-sound-is-so-important-in-horror-movies.

[abgerufen am 25.03.25]

ClipMove (o. D.): AI Sound Effect Generator - ClipMove, ClipMove, [online] https://www.clipmove.com/de/ai-sound-effect-generator [abgerufen am 28.03.2025].

Copet, Jade/Felix Kreuk/Itai Gat/Tal Remez/David Kant/Gabriel Synnaeve/Yossi Adi/Alexandre Défossez (2023): Simple and Controllable Music Generation, in: *arXiv (Cornell University)*, [online] doi:10.48550/arxiv.2306.05284.

Creative Commons (o. D.): Deed - Namensnennung - Weitergabe unter gleichen Bedingungen 3.0 Deutschland - Creative Commons, Creativecommons, [online] https://creativecommons.org/licenses/by-sa/3.0/de/deed.de.

Crook, Tim (1999): *Radio Drama: Theory and practice*, Psychology Press.

Davison, Andrew (2005): *Killer game programming in Java*, [online] http://ci.nii.ac.jp/ncid/BA73282585.

Défossez, Alexandre/Jade Copet/Gabriel Synnaeve/Yossi Adi (2022): High fidelity neural audio compression, in: *arXiv (Cornell University)*, [online] doi:10.48550/arxiv.2210.13438.

Degé, Franziska/Gudrun Schwarzer (2015): Musik und kognitive Entwicklung, in: Günther Bernatzky/Gunter Kreutz (Hrsg.), *Musik und Medizin: Chancen für Therapie, Prävention und Bildung*, Wien, Österreich: Springer, S. 359–373, [online] doi:10.1007/978-3-7091-1599-2_23.

Dickreiter, Michael (2014): *Handbuch der Tonstudiotechnik*, 8. Aufl., De Gruyter Saur, [E-Book].

Duden (2025): Originalton, Duden, [online] https://www.duden.de/node/106548/revision/1243183 [abgerufen am 11.07.2025].

E. Glenn Schellenberg (2024): Department Of Psychology, [online] https://www.psych.utoronto.ca/people/directories/all-faculty/e-glenn-schellenberg [abgerufen am 11.07.2025].

ElevenLabs (2024a): ElevenLabs SB1, ElevenLabs, [online] https://elevenlabs.io/sound-effects/soundboard [abgerufen am 12.05.2025].

ElevenLabs (2024b): Free AI Sound Effect Generator | Text to Sound Effects | ElevenLabs, ElevenLabs, [online] https://elevenlabs.io/sound-effects [abgerufen am 13.04.2025].

Fadr (o. D.): AI Instruments - SynthGPT 2 - Fadr, Fadr, [online] https://fadr.com/synthgpt [abgerufen am 08.07.2025].

filmsound.org (o. D.): FOLEY, Filmsound, [online] https://www.filmsound.org/terminology/foley.htm.

Flückiger, Barbara (2023): *Sound design*, Zürcher Filmstudien, 7. Aufl., [eBook] doi:10.23799/9783741002472.

Flückiger, Barbara (2011): Zum Mehrwert in der Ton/Bild-Beziehung, in: *Brill | Fink eBooks*, S. 161–177, [online] doi:10.30965/9783846751718_010.

Frauenhofer-Institut IKS (o. D.): Künstliche Intelligenz (KI) und maschinelles Lernen, Fraunhofer-Institut IKS, [online] https://www.iks.fraunhofer.de/de/themen/kuenstliche-intelligenz.html [abgerufen am 11.03.2025].

Gabrielsson, Alf (2010): Strong experiences with music, in: Patrik N. Juslin/John N. Sloboda (Hrsg.), *Handbook of Music and Emotion Theory, Research, Applications.*, Oxford, Vereinigtes Königreich: Oxford University Press, S. 547–574.

Ganguly, Martin (2022): Der Stummfilm, bpb.de, [online] https://www.bpb.de/lernen/filmbildung/230667/der-stummfilm/ [abgerufen am 03.07.2025].

getimg.ai (o. D.): Guide to Stable Diffusion CFG Scale Parameter, getimg.ai, [online] https://getimg.ai/guides/interactive-guide-to-stable-diffusion-guidance-scale-parameter [abgerufen am 13.04.2025].

Google (o. D.): Google AI - AI Principles, Google Ai, [online] https://ai.google/responsibility/principles/ [abgerufen am 09.05.2025].

Google (2023): SoundStorm: Efficient parallel audio generation, Google Research, [online] https://research.google/blog/soundstorm-efficient-parallel-audio-generation/ [abgerufen am 02.05.2025].

Google (2021): SoundStream: an End-to-End neural audio Codec, Google Research, [online] https://research.google/blog/soundstream-an-end-to-end-neural-audio-codec/ [abgerufen am 02.05.2025].

Herzer, Jan P./Max Kulmann (2018): Akustische Szenografie - Anregungen und praktische Empfehlungen für die Arbeit mit Klang im Raum, in: A. Bienert/D. Haffner (Hrsg.), *Konferenzband EVA 2013 Berlin Elektronische Medien & Kunst, Kultur, Historie*, 6. Aufl., Heidelberg, Deutschland: arthistoricum.net, S. 13–16, [online] doi:10.11588/arthistoricum.159.c744.

Hewlett Packard Enterprise Development LP (2025): Was ist KI-Inferenz? | Glossar, HPE Deutschland, [online] https://www.hpe.com/de/de/what-is/ai-inference.html#:~:text=Die%20KI%2DInferenz%20beim%20maschinellen,auf%20zuvor%20ungesehene%20Daten%20anwendet [abgerufen am 23.04.2025].

Hu, Charlotte/Amanda Downie (2025): Text to Speech, Think, [online] https://www.ibm.com/de-de/think/topics/text-to-speech [abgerufen am 09.07.2025].

HUIT (2025): Getting started with prompts for text-based Generative AI tools, Harvard University Information Technology, [online] https://www.huit.harvard.edu/news/ai-prompts [abgerufen am 08.07.2025].

Huwiler, Elke (2005): Sound erzählt. Ansätze einer Narratologie der akustischen Kunst, in: Harro Segeberg/Frank Schätzlein (Hrsg.), *Sound. Zur Technologie und Ästhetik des Akustischen in den Medien*, Marburg, Deutschland: Schüren

(Schriftenreihe der Gesellschaft für Medienwissenschaft (GfM) 12), S. 285–305, [online] doi:10.25969/mediarep/14249.

Image Line Software (2023): FL STUDIO 21.2 | What's New?, Image-line, [online] https://www.image-line.com/fl-studio-news/fl-studio-21-2-whats-new [abgerufen am 12.07.2025].

Image Line Software (2025): FL Studio 2025 Start faster, finish stronger, Image-line, [online] https://www.image-line.com/release [abgerufen am 12.07.2025].

IMDb (o. D.a): Erpressung (1929), IMDb, [online] https://www.imdb.com/de/title/tt0019702/ [abgerufen am 03.07.2025].

IMDb (o. D.b): Ron Bochar, IMDb, [online] https://www.imdb.com/de/name/nm0090580/ [abgerufen am 11.07.2025].

Jensen, Klaus Bruhn (2006): Sounding the Media, in: *Nordicom Review/NORDICOM Review*, Bd. 27, Nr. 2, S. 7–33, [online] doi:10.1515/nor-2017-0228.

Judson, Ellen (2024): „AI-generated" is the new „fake news", Global Witness, [online] https://globalwitness.org/en/campaigns/digital-threats/ai-generated-is-the-new-fake-news/ [abgerufen am 09.05.2025].

Kaminskyi, Bohdan (2023): Meta hat Audiobox eingeführt - KI für Spracherzeugung und Soundeffekte, in: *Gagadget*, 12.12.2023, [online] https://gagadget.de/368718-meta-hat-audiobox-eingefuhrt-ki-fur-spracherzeugung-und-soundeffekte/?utm_source=chatgpt.com. [abgerufen am 02.05.25]

Khan, Ibrahim/Van Nguyen Thai/Xincheng Dai/Ruck Thawonmas (2022): DareFightingICE Competition: A Fighting Game Sound Design and AI Competition, in: *arXiv (Cornell University)*, [online] doi:10.48550/arxiv.2203.01556.

Kharitonov, Eugene/Damien Vincent/Zalán Borsos/Raphaël Marinier/Sertan Girgin/Olivier Pietquin/Matt Sharifi/Marco Tagliasacchi/Neil Zeghidour/Google

Research (o. D.): SPEAR-TTS, Google Research, [online] http://google-research.github.io/seanet/speartts/examples/ [abgerufen am 30.04.2025].

KI-Statement | Verband deutscher Sprecher:innnen (2024): Verband Deutscher Sprecher:Innnen, [online] https://www.sprecherverband.de/ki-statement/ [abgerufen am 12.07.2025].

Kolbeck, Philipp (o. D.): Jingle Definition - Lingner Marketing GmbH, Lingner Digital GmbH, [online] https://www.lingner.de/de/marketing-glossar/jingle [abgerufen am 11.07.2025].

Kovach, Bill/Tom Rosenstiel (2021): *The Elements of Journalism, Revised and Updated 4th Edition: What Newspeople Should Know and the Public Should Expect*, Crown.

Kreuk, Felix/Gabriel Synnaeve/Adam Polyak/Uriel Singer/Alexandre Défossez/Jade Copet/Devi Parikh/Yaniv Taigman/Yossi Adi (2022): AudioGen: Textually Guided Audio Generation, in: *arXiv (Cornell University)*, [online] doi:10.48550/arxiv.2209.15352.

Künstliche Intelligenz. Die letzte Erfindung der Menschheit? - MIK (o. D.): MIK, [online] https://mik-osnabrueck.de/presse/kuenstliche-intelligenz-die-letzte-erfindung-der-menschheit/ [abgerufen am 12.03.2025].

Laswell, Harold D. (1948): The structure and function of communication in society, in: Lyman Bryson (Hrsg.), *The communication of ideas*, New York, Vereinigte Staaten: Harper and Row, S. 37–51.

Leenders, Matts Johan (2012): *Sound für Videospiele: Besondere Kriterien und Techniken bei der Ton- und Musikproduktion für Computer- und Videospiele*, [online] doi:10.5771/9783741000126.

Lidström, Magnus & Fredrik (2025): Synplant 2, Sonic Charge, [online] https://soniccharge.com/synplant [abgerufen am 08.07.2025].

Lilleslåtten, Mari (2022): Linguistic cues could be key to exposing fake news, [online] https://www.hf.uio.no/ilos/english/research/news-and-events/news/2022/linguistic-cues-could-be-key.html [abgerufen am 12.07.2025].

Lowy Institute (2024): Don't play it by ear: Audio deepfakes in a year of global elections, The Interpreter, [online] https://www.lowyinstitute.org/the-interpreter/don-t-play-it-ear-audio-deepfakes-year-global-elections [abgerufen am 09.07.2025].

Lu, Donna (2025): We tried out DeepSeek. It worked well, until we asked it about Tiananmen Square and Taiwan, in: *The Guardian*, 31.01.2025, [online] https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan. [abgerufen am 06.07.25]

Matern, Manuel (2024): *The role of sound in journalism and how it influences the storytelling experience.* Essay, Volda University College, [online] https://drive.google.com/file/d/1aEdvLIrmTxGxL4APDzdBQ9rtGbTyhz1n/view [abgerufen am 19.07.25]

Matern, Manuel (2022): *Zwischen Gänsehaut und Halluzination - Welchen Einfluss hat Musik? Die Wirkung von Musik auf Emotion und Psyche*. Bachelorarbeit, Hochschule Furtwangen University, [online] https://drive.google.com/file/d/1nvx08G9xO_I8G8zZDBsqY0gzodtwC2cT/view [abgerufen am 19.07.25]

Mcgurk, Harry/John Macdonald (1976): Hearing lips and seeing voices, in: *Nature*, Bd. 264, Nr. 5588, S. 746–748, [online] doi:10.1038/264746a0.

Mchugh, Siobhan/Mia Lindgren (2013): Not Dead Yet: emerging trends in radio documentary forms in Australia and the US, in: *The Australian Journalism Review*, 2. Aufl., S. 101–113.

McLuhan, Marshall (1994): *Understanding Media. The Extensions of Man*, Cambridge, Massachusetts, Vereinigte Staaten: MIT Press.

Meaker, Morgan (2023): Slovakia's Election Deepfakes Show AI Is a Danger to Democracy, in: *WIRED*, 03.10.2023, [online] https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/#:~:text=Fact-checkers%20scrambled%20to%20deal%20with%20faked%20audio%20recordings%20released%20days. [abgerufen am 09.05.25]

Meta (2024a): Audiobox | Meta FAIR, Audiobox.Metademolab, [online] https://audiobox.metademolab.com/ [abgerufen am 02.05.2025].

Meta (2024b): Audiobox | Meta FAIR, Audiobox.Metademolab, [online] https://audiobox.metademolab.com/maker [abgerufen am 02.05.2025].

Meta (2023): Audiobox: Generating audio from voice and natural language prompts, Meta Blog, [online] https://ai.meta.com/blog/audiobox-generating-audio-voice-natural-language-prompts/ [abgerufen am 02.05.2025].

Meta (2025a): AudioCraft: generating high-quality audio and music from text, Meta, [online] https://ai.meta.com/resources/models-and-libraries/audiocraft/ [abgerufen am 25.04.2025].

Meta (2025b): AudioGen: Textually-Guided Audio Generation, Audiocraft Metademolab, [online] https://audiocraft.metademolab.com/audiogen.html [abgerufen am 22.04.2025].

Miller, Wesley C. (1931): The Illusion of Reality in Sound Pictures, in: Lester Cowan (Hrsg.), *Recording Sound for Motion Pictures*, S. 210–217.

MMAudio.net (2025): MMAudio - AI Voice Generation Tool, MMAudio, [online] https://mmaudio.net/de [abgerufen am 13.04.2025].

MrKlay (2018): SCP: SITE 22 - Short film, [YouTube] https://www.youtube.com/watch?v=JISSJs6HNFw [abgerufen am 14.06.2025].

Musik: Was ist eigentlich Sampling? (2016): in: *Augsburger-allgemeine*, 31.05.2016, [online] https://www.augsburger-allgemeine.de/panorama/Musik-Was-ist-eigentlich-Sampling-id37957082.html. [abgerufen am 12.07.25]

Nantais, Kristin/Glenn Schellenberg (1999): The Mozart Effect: An Artifact of Preference, in: *Psychological Science*.

Nichols, Bill (2017): *Introduction to documentary*, 3rd ed, Indiana University Press.

NightCafe Studio Pty Ltd (o. D.): What are seeds in AI and how they work?, NightCafe Studio, [online] https://help.nightcafe.studio/portal/en/kb/articles/what-are-seeds-in-ai-and-how-they-work [abgerufen am 13.04.2025].

Noble, Joshua (2025): Autoregressives Modell, Think, [online] https://www.ibm.com/de-de/think/topics/autoregressive-model [abgerufen am 22.04.2025].

OmniSale GmbH (o. D.): AI Voice Cleaner & Background Noise Remover | LALAL.AI, lalal.ai, [online] https://www.lalal.ai/voice-cleaner [abgerufen am 08.07.2025].

Ondaatje, Michael/Walter Murch (2002): *The conversations: Walter Murch and the Art of Editing Film*, Alfred A. Knopf (Hrsg.).

Prakash, Prarthana (2024): Gen Z's new TikTok news habit widens the generational gap with TV-loving boomers, Fortune Europe, [online] https://fortune.com/europe/2024/09/12/gen-z-tiktok-instagram-news-generational-divide-ofcom/ [abgerufen am 12.07.2025].

Rauscher, Frances/Gordon L. Shaw/Catherine N. Ky (1993): Music and spatial task performance, in: *Nature 365*.

Reinberger, Stefanie/Eckart Altenmüller (2012): Hören von Musik aktiviert limbisches System im Gehirn, dasGehirn.info, [online]

https://www.dasgehirn.info/denken/musik/das-musikalische-gehirn [abgerufen am 11.07.2025].

Reinelt, Paul (2024): Wegen KI-Fake: Tupac Shakurs Erben klagen gegen Drake, in: *Musikexpress*, 25.04.2024, [online] https://www.musikexpress.de/wegen-ki-fake-tupac-shakurs-erben-klagen-gegen-drake-2660469/. [abgerufen am 09.05.25]

Reinhardt, Jan/Günther Rötter (2012): Musikpsychologischer Zugang zur Jugend-Musik- Sozialisation, in: Robert Heyer/Sebastian Wachs/Christian Palentien (Hrsg.), *Handbuch Jugend - Musik – Sozialisation*, Springer VS, S. 127–155.

Rodero, Emma (2012): Stimulating the Imagination in a Radio Story: The Role of Presentation Structure and the Degree of Involvement of the Listener, in: *Journal Of Radio & Audio Media*, Bd. 19, Nr. 1, S. 45–60, [online] doi:10.1080/19376529.2012.667022.

Sauer, Matthias (2024): KI für Sound- und Musikdesign - Bonedo, Bonedo, [online] https://www.bonedo.de/artikel/ki-fuer-sound-und-musikdesign/ [abgerufen am 08.07.2025].

Scherer, Klaus R. (2005): What are emotions? And how can they be measured?, in: *Social Science Information*, Bd. 44, Nr. 4, S. 695–729, [online] doi:10.1177/0539018405058216.

Schipmann, Jan/Linda Friese/Constantin Hofsäß/Lennart Glaser (2022): So funktioniert die TikTok-Strategie der AfD, in: tagesschau.de, [online] https://www.tagesschau.de/investigativ/funk/afd-tiktok-101.html [abgerufen am 17.07.2025].

Schlüter, Rebekka (2024): Spatial Audio: Was ist das und wie kannst du es ausprobieren?, HIFI.DE, [online] http://hifi.de/news/spatial-audio-was-ist-das-und-wie-kannst-du-es-ausprobieren-181737 [abgerufen am 09.07.2025].

Schmidgen, Henning (2021): Surface Media: McLuhan, the Bauhaus and the Tactile Values of TV, in: *Body & Society*, Bd. 28, Nr. 1–2, S. 121–153, [online] doi:10.1177/1357034x211008960.

Schramm, Holger (2008): Rezeption und Wirkung von Musik in den Medien, in: Stefan Weinacht/Helmut Scherer (Hrsg.), *Wissenschaftliche Perspektiven auf Musik und Medien.*, Weinheim, Deutschland: Beltz Verlag, S. 135–153, [online] doi:10.1007/978-3-531-90859-5_9.

Score a Score (2025): What is Sound Design?, Score A Score, [online] https://scoreascore.com/blog/what-is-sound-design/ [abgerufen am 07.03.2025].

Segal, Jeanne, PhD/Melinda Smith MA/Lawrence Robinson/Greg Boose (2025): Body Language and Nonverbal Communication, HelpGuide.org, [online] https://www.helpguide.org/relationships/communication/nonverbal-communication [abgerufen am 11.07.2025].

SFX Engine (o. D.a): AI Sound Effect Generator: The Complete Guide, SFX Engine Blog, [online] https://sfxengine.com/blog/ai-sound-effect-generator-guide [abgerufen am 28.03.2025a].

SFX Engine (o. D.b): Free AI Sound Effect Generator, SFX Engine, [online] https://sfxengine.com/ [abgerufen am 28.03.2025].

Sinclair, Jean-Luc (2020): *Principles of Game Audio and Sound Design : Sound Design and Audio Implementation for Interactive and Immersive Media*, [online] https://openlibrary.org/books/OL28044399M/Principles_of_Game_Audio_and_Sound_Design.

Singh, Shubham (2024): Podcast Statistics (2024) → Active Listeners & Trends, DemandSage, [online] https://www.demandsage.com/podcast-statistics/ [abgerufen am 12.07.2025].

Snapes, Laura (2023): AI song featuring fake Drake and Weeknd vocals pulled from streaming services, in: *The Guardian*, 18.04.2023, [online] https://www.theguardian.com/music/2023/apr/18/ai-song-featuring-fake-drake-and-weeknd-vocals-pulled-from-streaming-services. [abgerufen am 09.05.25]

Sorth, Nina (2024): Sprachassistenten – IT-P GmbH, IT-P GmbH – Ihr Partner für Erfolgreiche Digitale Transformation, [online] https://www.it-p.de/lexikon/sprachassistenten/.

Soundly (o. D.): Soundly - the complete sound effects platform, Soundly, [online] https://getsoundly.com/ [abgerufen am 09.07.2025].

Souriau, Étienne (1997): Die Struktur des filmischen Universums und das Vokabular der Filmologie, in: *Montage AV. Zeitschrift für Theorie und Geschichte Audiovisueller Kommunikation*, Bd. 6, Nr. 2, S. 140–157, [online] doi:10.25969/mediarep/24.

Stablecog Inc. (2023): What is Inference Steps? | Guide, Stablecog, [online] https://stablecog.com/guide/generation-settings/inference-steps [abgerufen am 13.04.2025].

Stuttgart, Staatsoper (o. D.): Zwei Windmaschinen in Spezialanfertigung, Staatsoper-stuttgart, [online] https://www.staatsoper-stuttgart.de/magazin/windmaschine/ [abgerufen am 03.07.2025].

Summers, Rob (2024): Exploring the Boundaries of Artificial Intelligence in Sound Design, Musical Theatre, [online] https://robsummers.co.uk/exploring-the-boundaries-of-artificial-intelligence-in-sound-design/ [abgerufen am 08.07.2025].

Thöing, Sebastian (2008): Die Welt ist nicht genug - Das Open-World Prinzip in Spielen, PC GAMES, [online] https://www.pcgames.de/Spielemarkt-Thema-117280/Specials/Die-Welt-ist-nicht-genug-Das-Open-World-Prinzip-in-Spielen-636177/ [abgerufen am 11.07.2025].

Tischmeyer, Friedemann (2025): Peak und Lautheit beim Audio Mastering am Computer, delamar.de, [online] https://www.delamar.de/mastering/audio-mastering-am-computer-peak-und-lautheit-3960/#:~:text=Die%20h%C3%B6chste%20Pegelspitze%20einer%20digitalen,ausgedr%C3%BCckt%20die%20h%C3%B6chste%20Aussteuerung%20dar. [abgerufen am 17.07.2025].

Tompkins, Al (o. D.): Video & Audio Editing - Radio Television Digital News Association, rtdna.org, [online] https://www.rtdna.org/video-and-audio-editing [abgerufen am 12.07.2025].

Unterhaltungssoftware Selbstkontrolle (USK) (2019): Non-Player character - USK lexikon, Unterhaltungssoftware Selbstkontrolle, [online] https://usk.de/alle-lexikonbegriffe/non-player-character/#:~:text=NPC%20oder%20Non%2DPlayer%2DCharacter,hin%20zu%20tragende%20Rollen%20spielen [abgerufen am 09.07.2025].

Waves Audio Ltd. (o. D.): COSMOS AI-Powered Sample Finder, Waves, [online] https://www.waves.com/plugins/cosmos-sample-finder [abgerufen am 12.07.2025].

Weiß, Eva-Maria (2023): Meta veröffentlicht AudioCraft: Eine Art ChatGPT für Musiker, in: *Heise Online*, 03.08.2023, [online] https://www.heise.de/news/Meta-veroeffentlicht-AudioCraft-Eine-Art-ChatGPT-fuer-Musiker-9233473.html. [abgerufen am 22.04.25]

Wikipedia-Autor*innen (2018): Étienne Souriau, [online] https://de.wikipedia.org/wiki/%C3%89tienne_Souriau [abgerufen am 23.03.2025].

Wikipedia-Autor*innen (2004): Überblendung (Film), Wikipedia, [online] https://de.wikipedia.org/wiki/%C3%9Cberblendung_(Film) [abgerufen am 09.07.2025].

Winn, Ross (2021): 2021 Podcast Stats & Facts (New Research From Apr 2021), Podcast Insights®, [online] https://www.podcastinsights.com/podcast-statistics/ [abgerufen am 12.07.2025].

Wondershare (o. D.a): Generate high quality AI sound effects with text prompts, Wondershare, [online] https://filmora.wondershare.com/ai-sound-effect-generator.html [abgerufen am 06.04.2025].

Wondershare (o. D.b): WonderShare Filmora – AI Video Editor & App, Wondershare, [online] https://filmora.wondershare.com/ [abgerufen am 06.04.2025].

## Abbildungsverzeichnis

Abbildung 1 GUI SFX Engine https://sfxengine.com/de/dashboard, abgerufen am 12.05.25. .....20
Abbildung 2 GUI ClipmoveAI https://www.clipmove.com/de/ai-sound-effect-generator abgerufen am 06.07.25 .....21
Abbildung 3 GUI MyEdit https://myedit.online/de/audio-editor/ai-sound-effect-generator/edit abgerufen am 12.05.25 .....22
Abbildung 4 Filmora 14 GUI .....22
Abbildung 5 ElvenLabs GUI https://elevenlabs.io/app/sound-effects/generate abgerufen am 12.05.25 .....23
Abbildung 6 MMAudio GUI https://mmaudio.net/de/text-to-audio abgerufen am 20.05.25 .....24
Abbildung 7 Audiobox GUI https://audiobox.metademolab.com/maker abgerufen am 25.05.25 .....26
Abbildung 8 Audiobox Frequenzspektrum .....41
Abbildung 9 ElevenLabs Frequenzspektrum .....41

## Tabellenverzeichnis

Tabelle 1 Auswahlkriterien .....29
Tabelle 2 Bewertungskriterien.....30
Tabelle 3 Gewichtungstabelle .....31
Tabelle 4 Vor- und Nachteile der KI-Tools .....35
Tabelle 5 Ergebnisbewertung .....39

# Anhang

## 1. Whispering Woods

Anmerkung: Kursiv geschriebene Texte stellen Markierungen zur Einfügung von Geräuschen dar.

The wind *howled* through the trees like a mysterious ghost as the sun disappeared behind the dense canopy. Deep in the heart of Blackoak Forest, where few dared to go, an old story began to come alive - once again.

Agnes, a lanky girl of seventeen with stubborn eyes and a *rusty lantern, walked past the gnarled roots of an oak tree.*
"Don't go in after dark," old Marnie had warned her back in the village. "The forest listens. And it remembers."
She had laughed then. But now, even her breathing seemed too loud.
She turned to her younger brother, Lucas, who clutched his teddy like it was a shield.

"We'll find the herb by the riverbend and be out before the moon's high. I promise."
Lucas didn't answer. He only nodded.

Hours passed. The deeper they went, the less the forest looked like anything natural. *Trees twisted in impossible shapes*. Moss covered even the air itself, heavy and wet.
And then they heard it. A whisper, slithering through the dark.
*"Agnesss… why are you here?"*
Agnes froze.
"You heard that too?" she asked.
Lucas voice was a breath: "It knows your name."
Suddenly, the *lantern flickered*.
From behind a knotted oak, *a shape appeared* — neither man nor beast. It had no face, only the suggestion of one, as if a shadow tried to remember what it once looked like.
"You should not have come" it said.

Agnes stood still, *heart drumming*. "What do you want?"
The faceless thing tilted its head. "A trade. The boy for your safe return. That is the old law."
Agnes *put herself between the creature and Lucas*. "No deal!"
*Laughter echoed through the trees — not from one voice, but hundreds, layered and shrill*. "Then run," the whisper said. "Run before the forest wakes."

*They fled. Branches clawed at them like fingers. Roots rose to trip them*. The forest itself *bent and shifted*, changing paths, looping them back to the same clearing again and again. "There's no way out," Lucas sobbed. "Agnes, we're stuck!"

*Agnes dropped to her knees*. Her promise echoed in her mind. We'll be out before the moon's high. That was the key.

"Lucass" she said, "close your eyes and remember the way we came. Not with your eyes. With your heart." He obeyed. In the stillness, *a faint breeze stirred*. It smelled of rosemary and river water. "That way," he said.

They followed his instinct. The forest protested — *groaned and howled and tried to trap them. But step by step*, it began to lose its grip.
Until at last — a break in the trees. *The stars, real stars, stared down at them like old friends.*

Old Marnie was waiting at the edge of the road, a lantern in his hand.
"You're lucky," he said, eyes hard. "Few make it out once it notices them."
Agnes didn't speak. She looked down at Lucas, still holding his teddy.
Marnie took *a long drag on his pipe and muttered*, "But it'll remember you now, girl. And it never forgets a name."
And behind them, deep in the dark, the forest whispered again.
*"Agnesss…"*

## 2. Skizzenhaftes Protokoll der Eigenproduktion

Die Aufnahme erfolgte zuhause in meinem Zimmer.
Mit Soundly free wurden fast alle Sounds gefunden, die aber alle noch etwas Nachbearbeitung im Sinne von Kürzung, oder sogar Pitch o.Ä. brauchten.
Der Candle flickering Sound fehlt.
Star dust war nicht vorhanden.

Die Sounds zu suchen, dauerte eine halbe Stunde.
Den Stardust Sound erstelle ich selbst mit dem BBC Orchestra PlugIn. Den Candle Sound habe ich selbst aufgenommen.
Hat grob 90 Minuten gedauert, weil die Sounds meist sehr lange sind in Soundly und man durch gehen muss um die perfekte Stelle zu finden. Mehr Auswahl = mehr Zeit aber auch besseres Ergebnis.

Recording 1 Stunde
Sounds suchen 30min
Post Produktion 1h 30min
**3 Stunden gesamt**

## 3. MMAudio Protokoll

| Szene | Textprompts | Negativer Prompt | Dauer (Sek) | Anzahl der Schritte | CFG | Samen |
|---|---|---|---|---|---|---|
| **1** | It's the first scene of a war short film. One hears multiple gun shots in the distance and at timecode 00:06min an explosion. An facility alarm is ringing and men are screaming. | | 8 | 25 | 4,5 | -1 |
| | It's the first scene of a war short film. One hears multiple gun shots in the distance and at timecode 00:06min an explosion. An facility alarm is ringing and men are screaming. | | 10 | 25 | 4,5 | -1 |
| | Multiple gun shots, a facility alarm is ringing and men are screaming in the distance. | | 11 | 25 | 4,5 | -1 |
| | Sound design from video, exterior, gun shots in the distance, men screaming in the distance. | | 11 | 25 | 4,5 | -1 |
| | Sound design from video, exterior, alarm siren, gun shots in the distance, men screaming in the distance. | | 11 | 25 | 4,5 | -1 |
| | Sound design from video, exterior, alarm siren, gun shots in the distance, men screaming in the distance. | | 8 | 25 | 4,5 | -1 |
| | Sound design from video, exterior, alarm siren, gun shots in the distance, men screaming in the distance. | | 8 | 25 | 9 | -1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Sound design from video, exterior, alarm siren, gun shots in the distance, men screaming in the distance. | | 8 | 50 | 4,5 | -1 |
| | Sound design from video, exterior, alarm siren, gun shots in the distance, men screaming in the distance. | | 8 | 25 | 4,5 | 1 |
| | Sound design from video, exterior, alarm siren, gun shots in the distance, men screaming in the distance. | | 8 | 50 | 9 | -1 |
| | Sound design from video, exterior, alarm siren, gun shots in the distance, men screaming in the distance. | | 8 | 25 | 4,5 | -1 |
| **2** | Sound design from video, exterior, two men running in the forest. | | 8 | 25 | 4,5 | -1 |
| | Sound design from video, exterior, two men running in the forest. | no music or atmospheric sounds | 8 | 25 | 4,5 | -1 |
| | Sound design from video, exterior, two men running in the forest. Gun shots in the background. | no music or atmospheric sounds | 8 | 50 | 10 | -1 |
| **3** | Sound design from video, exterior, two men kneeling in the forest on dry leaves, looking around. | | 8 | 25 | 4,5 | -1 |
| | Organic Sound design from video, exterior, two men kneeling in the forest on dry leaves, looking around. | No dialogue, no talking. | 8 | 25 | 4,5 | -1 |
| | Organic Sound design from video, exterior, two men kneeling in the forest on dry leaves, soft forest floor, looking around. A | No dialogue, no talking. | 8 | 25 | 4,5 | -1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Monster screams in the far distance. | | | | | |
| | Organic Sound Design from video. | No dialogue, no talking. | 8 | 25 | 4,5 | -1 |
| **4** | Organic Sound design from video, exterior, two men kneeling in the forest on dry leaves, one changing stance. | No dialogue, no talking. | 8 | 25 | 4,5 | -1 |
| | Organic Sound design from video, exterior, two men kneeling in the forest on dry leaves, one changing stance slightly. | No dialogue, no talking. | 8 | 25 | 4,5 | -1 |
| | Organic Sound design from video, exterior, two men kneeling in the forest on dry leaves, one changing stance slightly. Only soft, quiet noises. | No dialogue, no talking. | 8 | 25 | 4,5 | -1 |
| **5** | Organic Sound design from video, exterior. Quiet forest, birds chirping. | | 8 | 25 | 4,5 | -1 |
| | Organic Sound design from video, exterior. Quiet forest, peaceful. | | 8 | 25 | 4,5 | -1 |
| | Organic, quiet, peaceful forest | | 8 | 25 | 4,5 | -1 |
| **6** | Organic Sounddesign from video. Two men fighting, punching, grunting and falling on forest floor covered in leaves. | | 8 | 25 | 4,5 | -1 |
| | Organic Sounddesign from video. | | 8 | 25 | 4,5 | -1 |
| | Organic Sounddesign from video. | No music. | 8 | 25 | 4,5 | -1 |
| **7** | Organic Sounddesign from video. Leaves on the floor rustling, monster breathing and growling deeply. | | 8 | 25 | 4,5 | -1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Organic Sounddesign from video. Leaves on the floor rustling, monster breathing and growling threatening. | | 8 | 25 | 4,5 | -1 |
| **8** | Organic Sounddesign from video. Monster breathing and growling, stabbing man to death with thorn. Man falls to floor. | | 8 | 25 | 4,5 | -1 |
| | Organic Sounddesign from video. Monster stabbing man to death with thorn. Man falls to floor. | | 8 | 25 | 4,5 | -1 |
| | Monster stabbing man brutally to death with thorn. Man falls to floor. | | 8 | 25 | 4,5 | -1 |
| **9** | Organic Sounddesign from video. Monster breathing and growling threatening and walking away to the left. | | 8 | 25 | 4,5 | -1 |
| | Monster breathing and growling threatening and walking away to the left. | | 8 | 25 | 10 | -1 |
| **10** | Organic Sounddesign from video. Helicopter overhead from left to right. | | 8 | 25 | 4,5 | -1 |
| | Organic Sounddesign from video. Helicopter flying overhead from left to right. Radio message: "Drop your weapon and lay down on the ground". | | 8 | 25 | 10 | -1 |
| | Organic Sounddesign from video. Helicopter flying overhead from left to right. Radio message: "Drop your weapon and lay down on the ground". | | 8 | 25 | 4,5 | -1 |
| | Organic Sounddesign from video. Add radio message: | | 8 | 25 | 4,5 | -1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | "Drop your weapon and lay down on the ground". | | | | | |

**Gesamtdauer knapp über 4 Stunden.**

**Uneditierte Version:** Beinhaltet pro Szene lediglich eine generierte KI Audio Spur, ohne Bearbeitung (außer Fade in und out)

**Editierte Version:**

| Szene | Bearbeitungen |
|---|---|
| **1** | Verschiedene Teile gekürzt und kombiniert für authentisches Klangbild; Schüsse synchronisiert |
| **2** | Fehlerhafte Atmo geloopt und als Bed genutzt; Stepsounds aus Versuch 3 eingebettet; Distanz erzeugt; Tiefe Sounds als entfernte Explosionen interpretiert |
| **3** | Bewegungsgeräusche leiser gemacht und aus verschiedenen Versuchen kombiniert; Vogelzwitschern geloopt als ein Art „Bed“ (Kombi aus V1 und V2) |
| **4** | Nur Bewegungsgeräusche leiser gemacht (nur Versuch 1) |
| **5** | Versuch 1 nur leiser gemacht |
| **6** | Kombination aus V1 und V3; Timing korrigiert; Lautstärken angepasst; Geräusche rearrangiert, um Szene abzudecken |
| **7** | Kombination aus V1 und V2; Timing und Lautstärken angepasst; rearrangiert für durchgehende Atmo und angsteinflößenderes Monster |
| **8** | Kombination aus V1, V2, V3; Timings, Lautstärken und Anordnung angepasst; Erstechen besser erfasst und synchronisiert; Schreie für Übergänge platziert |
| **9** | Kombination aus V1 und V2; Nicht passende Monster-Sounds entfernt; Schritte fehlen weiterhin |
| **10** | Tonaussetzer am Anfang durch Duplizieren ausgebessert; unverständlicher Funkspruch am Ende eingefügt, näher am Original |

## 4. Filmora 14 Protokoll

| Szene | Uhrzeit | Prompt/Beschreibung | Ergebnis/Bewertung | Anmerkungen |
|---|---|---|---|---|
| **1** | 16:37 | It's the first scene of a war short film. Multiple gun shots, at timecode 00:01min an explosion, an alarm siren, and men are screaming in the distance. 8 Seconds long. | ist nicht möglich, da nur 100 Zeichen verfügbar sind | |
| | | Multiple gun shots, at 00:01min an explosion, alarm siren, men screaming | Funktioniert gar nicht. Sounds schwer zu unterscheiden, Sirene eher Polizei, Schrei sehr unklar, Sounds müssen wohl einzeln generiert werden | Immer drei Sounds (max. 6) zur Auswahl, bis zu 30 Sek., aber Generierung schon bei 3 mit 5 Sek sehr langsam |
| | | Multiple gun shots in the distance | Nicht zufriedenstellend, aber besser wird es nicht | Alle drei geladen |
| | | An explosion in the distance (verschiedene Varianten: an explosion, explosion) | Keine glaubwürdigen Sounds, Qualität schlecht | |
| | | Alarm siren | Schlechte Qualität, Artefakte, Verzerrungen | |
| | | Men screaming with fear | Schlechte Ergebnisse | alle drei geladen, weil auch wiederholte Generierung nicht besser wurde |
| | | War atmosphere | Meist Schuss-Soundeffekte, einer passend | |
| | 17:27 | | | *Arrangement Szene 1* |
| **2** | 17:34 | Two men running on gravel | Klingt wie Pferdehufe, teilwise Pflasterstein, einer genutzt | |
| | | Two men running in forest | Ergebnisse langsam, Untergrund wenig passend, zwei genutzt | Zu sehender Boden passt nicht mit Geräuschen überein. |

| | | | | |
|---|---|---|---|---|
| | | Two men breathing exhausted | Klingt wie verzerrtes Stöhnen, viele Artefakte, einer genutzt | |
| | | Two men running on gritty floor | Zwei Menschen hörbar, aber eher Steinboden | |
| | | Two men running on grit. | Etwas passender, haben aber irgendwie einen Pan von rechts nach links und umgekehrt. | Lade alle drei. |
| | | Wind in forest | Einer könnte als Atmo funktionieren | Da „War Atmo“ nicht funktioniert hat versuche ich damit etwas zu bekommen, was ich drunter legen kann. |
| | 17:51 | | | *Arrangement Szene 2* |
| **3** | 18:01 | Two men kneeling down on forest floor | Schlechte Qualität, Highend fehlt teilweise komplett, Boden nicht authentisch | zwei geladen |
| | | Monster screaming in distance | Artefakte, klingt wie Zombie-Knurren | |
| | | Alien screaming | Teilweise gut, einer geladen | |
| | 18:07 | | | *Arrangement Szene 3* |
| **4** | 18:23 | Slight movement on dry leaves | Sounds an sich nicht schlecht, Qualität aber schlecht | alle drei geladen, versuche das Beste draus zu machen |
| | 18:36 | | | *Arrangement Szene 4, nutze Wald Atmo erneut* |
| **5** | 18:40 | Quiet forest. Birds chirping | Sehr gute Sounds | |
| | 18:45 | | | *Arrangement Szene 5, Samples leiser gemacht, Wind leise daruntergelegt* |
| **6** | 18:49 | Two men fighting and punching each other | Sounds alle fast identisch, klingt komödiantisch, glitchy, schlechte Ergebnisse, | Punch sound effects nicht überzeugend, Wrestling nicht |

| | | | | |
|---|---|---|---|---|
| | | | manchmal nur rechter Lautpsrecher | passend. Mehrfach generiert. |
| | | Punch sound effects | Ich bekomme alles außer Schläge. | |
| | | Punch in the face sound effects | Klingt teilweise wieder nur wie Backpfeifen, aber weniger komödiantisch. | Nutze plural in der Hoffnung innerhalb eines Samples mehrere nutzen zu können |
| | | Two men wrestling | Nicht mal annähernd was mit Wrestling Sounds zu tun. | Auf Grund limitierter KI-Credits die ich kaufen musste, gehe ich zur nächsten Szene und komme falls noch welche übrig sind später nochmal zurück. |
| **7** | 19:12 | Slight movements on dry leaves | Gute Ergebnisse, alle drei geladen | Versuche den hier nochmal für die Bewegungen der Person am Anfang. |
| | | Alien breathing and growling deeply | Einer ok, Artefakte | |
| | | Alien breathing sounds | Voller Artefakte | Zwei geladen |
| | 19:25 | | | *Arrangement Szene 7* |
| **8** | 19:30 | Stabbing man to death | Komische gurgelartige Sounds, einer geht | |
| | | Sword stabs man to death | Ähnliche Sounds, Klingen nach Klangschale | einer genutzt |
| | | Man falling to ground | Eher Schritt-Sounds | einer genutzt |
| | | Frightening Alien scream | Sehr quitschig | Zwei genutzt |
| | *15min Pause* | | | |
| | 20:05 | | | *Arrangement Szene 8, Nutze Alien Sounds aus Szene davor Nutze Wind erneut* |

| | | | | |
|---|---|---|---|---|
| **9** | 20:13 | Huge spider with six legs walking / Big creature walking | Nicht die gewünschten Ergebnisse, einer genutzt | Erneute Generierung bringt auch nichts |
| | | Enormous spider walking | Nicht nutzbar | Hat nichts mit dem zu tun was zu sehen ist. |
| | 20:20 | | | *Arrangement Szene 9: Erneut nutze ich Wind und Alien Sounds*<br>*Steps passen wirklich absolut gar nicht auch erneute Versuche bringen nichts.*<br>*Hier auch dann der Moment, an dem ich ihn mir einfach selber bauen würde oder recorden würde, statt zu überlegen welchen prompt ich brauche.* |
| **10** | 20:27 | Helicopter Sounds | Gut | |
| | | Radio message: "Drop your weapon and lay down on the ground" | Klingt nach Funkspruch, aber unverständlich | War ein Versuch wert. |
| | 20:33 | | | *Arrangement Szene 10*<br>*Zwei Sounds gelayert und leicht Speed angepasst damit Länge passt.* |
| **Noch mal Szene 6** | 20:36 | Wrestling on dry leaves | Klingt wenigstens nach Laub | Punches und Geschrei gelayert, schwer abzustimmen |
| | 20:40 | | | *Arrangement Szene 6*<br>Punches und Geschrei gelayert, schwer abzustimmen auf Bild. *Besonders die Leaves und Wrestling Sounds* |

| | | | | |
|---|---|---|---|---|
| | | | | *haben meist eine Art Rhythmus, wahrscheinlich durch eine Art Loop, was Synchro schwer macht* |
| | | | | *BIS HIER UNGEFÄHR VIER STUNDEN ARBEIT* |
| | 22:30 | A wet, squelching sound of a sword piercing through flesh | Gutes Ergebnis. | Versuche nochmal das Erstechen besser hinzubekommen. Habe Prompt von ChatGPT genutzt. |
| | | A sticky, slurping noise as a blade is pulled out of a human body | Klingt eher dünn und nach sehr komischem Eislecken. | |
| | | A squelching noise as a blade is pulled out of flesh | Zwei davon könnten gehen aber klingen auch echt nicht gut | Habe die drei gelayered. Klingt jetzt ok ist aber eine komische Mischung aus Schlürfen und Keks Crunch Sounds. |
| | | spider-like alien footsteps on dirt | Weiter entfernt kann man nicht sein<br>Die restlichen Textprompts der KI passen ebenfalls nicht. | Gleiche Vorgehensweise erneut mit ChatGPT. |

## 5. Skizzenhaftes Protokoll: Audiobox

**Start der Bearbeitung: 15:30 Uhr**

Stimmen passen nicht richtig. Beispielsweise ist die Narrator Stimme nicht so wie erwartet. Stattdessen entschied ich mich für die Stimme Zane, weil sie am ehesten einer männlich klingenden Stimme ähnelt. Man kann immer 125 Zeichen in das Textfeld eingeben, ich muss also immer schauen, dass ich den Text in Stücke die klein genug sind einteile.
Die erste Generierung des Titels “Whispering Woods” wird abgeschnitten, sodass es nur nach “Whispering Wood” klingt. Lösung: Ich habe noch ein S hinzugefügt.

Es ist außerdem nervig, weil man nicht in die Timeline klicken kann, wodurch man immer von Beginn an abspielen muss.

Für den ersten Sound nutze ich diesen Prompt “wind howling through trees. mysterious” Ich bekomme Wasser rauschen und Vogel Gezwitscher. Es scheint statt meines Prompts eines der Standardbeispiele generiert zu haben. Beim erneuten Versuch bekomme ich Aufnahmen von Wind, als würde man einfach ein Mikrofon in den Wind halten. Die dritte Generierung klingt wenigstens etwas besser, jedoch ist kein Windpfeifen zu hören. “Wind howling” ergibt zwar gruselig klingende Sounds, jedoch eher nach Unterwasser statt nach Wind. Bei erneuter Generierung gab es ein Ergebnis, welches einigermaßen passt.

*Es ist nicht möglich die Sounds einzeln herunterzuladen. Außerdem gibt es keine Möglichkeit zu speichern. Es kommt auch vor, dass eine Fehlermeldung kommt, dass Audiobox gerade zu viel verwendet wird und man deshalb gerade nichts generieren kann.*

“Metal squeaking” ergab kein passendes Ergebnis. “Squeaking” ergab ein Ergebnis, was man nutzen kann. Jedoch ist auch dieses eher letzte Wahl und außerdem zu lang. Audiobox giebt einem jedoch keine Möglichkeit die Sounds zu kürzen. Eine Dauer in den Textprompt einzufügen funktioniert auch nicht.

“Steps on grass” ergab passende Ergebnisse, jedoch sehr leise und nicht sehr deutlich. Auch wieder zu lang.

Für die Stimme des Charakters Old Marnie gab es keine passende Stimme. Aus den 5 die zu wählen sind, sind außerdem alle weiblich klingend, was auch für die Stimme des Jungen Lucas oder für den Geist ein Problem darstellt.

*Problem: Wenn man Clips in der Timeline verschieben will muss man es sehr langsam machen, da man den Clip sonst “verliert” und er sich nur ein kleines Stück verschiebt.*

Die Generierung der Stimmen lief weitestgehend gut. Es gibt immer zwei Versionen und meist ist eine davon bei der Betonung am Ende des Satzes so, dass die Stimme etwas runter geht, was gut zur Audiostory passt.

“Wood creaking” funktionierte nicht. Die Ergebnisse enthielten entweder Geräusche, die nichts damit zu tun haben oder nur leises Knacken von Holzboden. “Wood creaking heavily” ebenfalls nicht. “Wood creaking deeply” auch nicht. “Wood murmuring deeply” ergab nichts Sinnvolles. Im Kopf habe ich Sounds die auch in Fluch der Karibik eingesetzt werden, wenn die großen Segelschiffe im Wind knarren. Wenn ich auf YouTube diese key words eingebe “Wooden sailboat creaking and groaning” mit “sound effect” bekomme ich direkt die richtigen Sounds. In Audiobox nicht mal annähernd. “deep wood creaking sound effect” ergab dann schließlich etwas, was einigermaßen funktioniert.

Eine der Stimmen (Emily) flüstert. Ich nutze diese für den Geist. Leider ist die Ausgabe aber nicht dazu zu bringen das “Agnesss…” länger zu ziehen. Auch die Art und Weise der Betonung ist sehr schnell und ähnelt eher einem Flüstern eines Mädchens im Klassenraum als dem eines Geists.

Für Lucas blieb dann keine Stimme mehr übrig. Der Versuch eine geeignete Stimme zu generieren, blieb erfolglos. Ich nutze die gleiche Stimme wie für Marnie, weil diese nur am Anfang und Ende kurz vorkommt, damit Agnes und Lucas zu unterscheiden sind. Fragezeichen, Ausrufezeichen und Punkte ändern die Betonung ein kleines bisschen. Jedoch ist keine passende zu finden, die sein Entsetzen authentisch reflektiert.

“A lantern flickering” funktioniert nicht
“A flame flickering” auch nicht.
“Fire flickering” auch nicht
“Fire flickering sound effect” schon eher, aber keinerlei saubere Sounds und oft mehrere in einer Datei, was durch die limitierten Bearbeitungsmöglichkeiten den Einsatz unmöglich macht.
“A sound for something scary suddenly appearing” ergibt Errors, das dieser Input nicht bearbeitet werden kann
“Something scary suddenly appearing” ergab komplett andere Sounds als erwartet oder erwünscht. Einer klang wie ein Alien kreischen, welches wenigstens etwas unheimlich klingt.
“Heart beating fast and loud” ergab zwar Herzschlagen aber mit viel Hintergrundgeräuschen. Ich versuche mal einen negativen Prompt. Hat nicht funktioniert.
“Heart beating” hat funktioniert, jedoch wieder sehr lange und nicht ganz sauber.

Leider ist da dann Schluss. Die KI lässt mich auch nach mehreren Versuchen keine Stimmen mehr generieren. Das Limit ist erreicht.
**Ende um18 Uhr**

## 6. Skizzenhaftes Protokoll: ElevenLabs

**Start um 10:45Uhr**

Ich lade mein Text Dokument hoch und wähle eine Default Stimme aus. Außerdem nutze ich die automatische Stimmen Erkennung. Für die Default Voice nutze ich George, da er die Tags Mature, und Narratve & Story hat.

Die Erkennung der Verschiedenen Stimmen und Charaktere hat super funktioniert. George (Narrator) habe ich direkt ausgetauscht, da mir die Stimme nicht gefallen hat. ElvenLabs bietet eine große Bibliothek an Stimmen an, die in einer Suchmaschine mit vielen verschiedenen Tags super einfach gesucht werden können. Nach dem ich alle Voices zu passenden ausgetauscht habe, und die verschiedenen Parameter abgestimmt habe, schaue ich mir jetzt die Sound Effekte an.

Ich kann an beliebigen Stellen Sound Effekte mit text-prompts generieren und habe die Möglichkeit die Länge einzustellen. Außerdem kann ich entscheiden, ob der Narrator unterbrochen werden soll, oder der Sound im Hintergrund läuft. Die Lautstärke kann auch eingestellt werden. Ich kann aus vier generierten Sounds auswählen. Leider funktioniert die Einstellung der Länge nicht richtig und ändert sich pro Generierung wieder. Ich nutze die gleichen Prompts wie bei Audiobox und bekomme ähnlich schlechte Resultate. Dass die Dauer nicht angepasst werden kann, ist ein großer Minuspunkt. Außerdem sind die Ergebnisse meist sehr weit vom Gewünschten entfernt. An diesem Punkt überlege ich keine Sound Effekte direkt einzufügen, sondern das Audiobook so zu lassen und nachträglich Sounds die von ElevenLabs generiert wurden, einzufügen.

**12:30Uhr**

**Pause**

**14:00 Uhr**

Ich versuche passende Sounds im ElevenLabs Studio Text-to-Sound zu generieren. Die generierten Sounds aus dem Audiobook sind hier im Verlauf auch aufzufinden und es ist zu erkennen, dass die Funktion der Länge dort einen bug hat und nicht richtig funktioniert. Die erneute Generierung von “Wind howling mysteriously” mi 8 Sekunden funktioniert einwandfrei und ich bekomme gute Ergebnisse. Es ist außerdem möglich die Prompt Influence einzustellen (Standard ist 30%). Auf 100% bekomme ich noch bessere Ergebnisse in dem Fall. Ich spiele für die weiteren Sounds damit herum, um die besten Ergebnisse zu erzielen.

“Metal squeaking” gibt mir quitschendes Metall
“metal lantern squeaking while swinging slightly” gibt mir genau was ich brauch.
“Steps on grass” funktioniert auch gut
“wood creaking heavily” funkioniert auch auf anhieb
“Wood murmuring deeply” funkionierte nicht. Es ergab eher tiefe Stimmen und Schläge
“deep wood creaking sound effect” ergab Geräusche, die eher nach Fußboden klingen
“A lantern flickering” gab Gitarren Sounds
“A flame flickering” gab stickflammen Sounds
“A canlde flickering” ergab ganz gute Sounds

“A sound for something scary suddenly appearing” ergab Alien geschreie, jedoch könnte einer gut funktionieren.
“Heart beating fast and loud” funkionierte direkt super
“A sound for when someone jumps forward on a forest floor” gibt Sounds, die für einen Sprung auf Steinboden gehen könnten
“A sound for when someone steps forward quickly on a forest floor covered with dry leaves” Klappt gut!
“Laughter from hundreds of voices echoing through a forest“ funktioniert ganz gut auch wenn es sehr verzerrt ist und es nicht so viele sind
“Laughter from hundreds of voices echoing through a forest. Lots of Reverb.” Die Ergebnisse sind alle ganz gut, aber ich werde alle Sounds layern müssen. Keines der Ergebnisse passt perfekt.
“Two persons running fast through a dense forest.” war zu langsam
“Two persons sprinting fast through a dense forest.” war zu schnell
“Two persons sprinting through a dense forest.” Geht einigermaßen, jedoch ist immer ein fade in und out mit drin, als würde man vorbeirennen. Außerdem hat nur ein Sound zwei Personen. Ich werde wieder mehrere layern müssen und vielleicht die Geschwindigkeit.
“Bushes rustling” gab sehr lautes unangenhemes rauschen teilweise, einen kann ich aber nehmen
“Leaves rustling” hat super funktioniert
“Roots rising like tentacles and trying to catch something” gab Dialog?! und White Noise riser sound effekte
“Big trees bending and creaking heavily” funktioniert
“A person dropping to its knees on forest floor” funktioniert gut, jedoch muss ich vielleicht etwas wegschneiden
“A faint breeze stirring “ ergab leises Wasser plätschern
“A faint breeze of wind stirring “ auch
“A faint breeze of wind” funktioniert
“A fairy forest groaning and howling” ergab sehr komische Geräusche
“An enchanted forest groaning and howling mysteriously “ ergab paar coole Geräusche
“Stars” war ein Test, ergab Sounds, die man gut in einem UI verwenden könnte
“Fairy star dust sound effect” funktionierte
“A long drag on a pipe “ hat nichts mit rauchen zu tun
“A man taking long drag on a pipe and breathing out smoke” ergab kreischende, verzerrte Sounds
“A man smoking and breathing out smoke” funktionierte
“A ghost whispering "Agneeesssss” funktionierte sogar, aber war weibliche Stimme
“A male ghost whispering "Agneeesssss”” hat funktioniert.

**15Uhr**
Fertig mit Generierung. Die Generierung geht immer nur ein paar Sekunden. Prompt Genauigkeit war immer auf 100%, ich habe auch mal damit rumgespielt aber nie ein Ergebnis genutzt bei weniger Prozent.
Für den Export gibt es verschiedene Parameter und Einstellungen (siehe Screenshot) Normalisierung für Audiobook Standards (-3db Peak stimmt) und wav / mp3 Formate. Quelle zum abgleich

https://www.linkedin.com/advice/3/how-do-you-master-audio-podcasts-audiobooks-loudness?lang=en also stimmen die Angaben und auch nach Testmessung passts bei -18LUFS.

Die Stimme von Marnie ist leider trotz Boost in ElevenLabs sehr leise. Ich füge jetzt die Soundeffekte ein und passe die Lautstärke von Marnie nachträglich an.

Die Sound Effekte von Eleven Labs sind MP3 was nicht gut für die Postproduction ist.

**Bis hier hin 2h30min**

Post-Produktion dauerte eine Stunde. **Insgesamt ungefähr 3h30min**.

## 7. Checkliste für den verantwortungsvollen Einsatz von KI-Tools

| Phase | Kategorie | Frage | Erfüllt? |
|---|---|---|---|
| **Vor dem Einsatz** | **Lizenz & Rechtliches** | Sind die generierten Inhalte rechtlich einwandfrei verwendbar (z. B. lizenzfrei)? | [ ] |
| | | Gibt es Informationen zu Trainingsdaten und Nutzungsrechten des KI-Tools? | [ ] |
| | | Werden rechtliche Grenzen wie Datenschutz (z.B. beim Voice Cloning) eingehalten? | [ ] |
| | **Kreativer Einsatz** | Gibt es ein gestalterisches Konzept hinter der Nutzung der KI? | [ ] |
| | | Kann die KI sinnvoll in den kreativen Workflow eingebettet werden? | [ ] |
| | | Ist der Einsatz dramaturgisch und inhaltlich begründet? | [ ] |
| | | Inwiefern dient der KI-Einsatz der kreativen Erweiterung und nicht nur der Effizienzsteigerung? | [ ] |
| | | Ist klar, wer die Verantwortung für die finale Soundgestaltung trägt? | [ ] |
| **Währendessen und danach** | **Qualität & Kontrolle** | Wurden die Sounds auf technische Qualität (z. B. Artefakte, Klarheit) geprüft? | [ ] |
| | | Wurde gegebenenfalls nachbearbeitet, um Qualität und Timing sicherzustellen? | [ ] |
| | | Unterstützen die Sounds die beabsichtigte narrative und immersive Wirkung? | [ ] |
| | **Transparenz & Nachvollziehbarkeit** | Sind KI-generierte Sounds eindeutig als solche erkennbar? | [ ] |
| | | Ist die Unterscheidung zwischen Original- und KI-Material möglich, wenn erforderlich? | [ ] |
| | **Ethische Reflexion** | Wird emotionale Wirkung gezielt, aber nicht manipulativ eingesetzt? | [ ] |
| | | Wird die Realität korrekt wiedergegeben, z. B. im journalistischen Kontext? | [ ] |
| | | Wurde kulturelle Sensibilität berücksichtigt? | [ ] |