Postproduktion von Sprachaufnahmen

Die Auswirkungen neuer Produktionswege auf die Restauration und das Sound Design

Masterarbeit

Universität für Musik und darstellende Kunst Graz
Institut für Elektronische Musik und Akustik

Fachhochschule Joanneum
Institut für Design und Communication

Betreuer:
Univ.Prof. Dipl.-Ing. Dr.techn. Alois Sontacchi

vorgelegt von:
Lukas Steinegger
01440931

Graz, Jänner 2024

FH JOANNEUM
University of Applied Sciences

kunst uni graz

LUKAS STEINEGGER **01440931**

....

Name in Blockbuchstaben Matrikelnummer

Ehrenwörtliche Erklärung

Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen nicht benutzt und die den Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen inländischen oder ausländischen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht. Die vorliegende Fassung entspricht der eingereichten elektronischen Version.

22.01.2024

Graz, am ..... Steinegger Lukas

.....

Unterschrift des Verfassers*der Verfasserin

I. Inhaltsverzeichnis

II. Abstract

Neue Wege in der Produktion von Sprachaufnahmen über das Internet und dessen Übertragungsmedien können bei falscher Handhabung Artefakte verursachen. Diese sollen in der Postproduktion restauriert werden. Zudem sollten die produzierten Sprachaufnahmen einem emotional ansprechenden Sound Design unterzogen werden. In dieser Arbeit werden zuerst die tontechnischen Grundlagen und Literatur zu Sprachaufnahmen miteinander verglichen, um daraus objektive Anforderungen an Sprachaufnahmen abzuleiten. Daraus werden dann in Kombination mit Beispielen aus der Praxis die Erreichbarkeit der Anforderungen mit einer konventionellen und einer daten- und KI-basierten Postproduktion getestet. Diese Ergebnisse werden dann mit einer vollständig automatisierten Postproduktion verglichen. Diese Arbeit ergab, dass die im Produktionsprozess entstandenen Artefakte keine vollständige Restauration zulassen und dass ein zufriedenstellendes Ergebnis nur in Kombination mit daten –und KI-basierten Tools zu erreichen ist.

New ways of producing voice recordings via the Internet and its transmission media can cause artifacts if handled incorrectly. These should be restored in post-production. In addition, the voice recordings produced should be subjected to an emotionally appealing sound design. In this work, the sound technology basics and literature on voice recordings are first compared in order to derive objective requirements for voice recordings. From this, in combination with practical examples, the achievability of the requirements is tested with conventional and data- and AI-based post-production. These results are then compared to fully automated post-production. This work revealed that the artifacts created during the production process do not allow for complete restoration and that a satisfactory result can only be achieved in combination with data and AI-based tools.

III. Abkürzungsverzeichnis

et. al. ..... und andere
S. ..... Seite
vgl. ..... vergleiche
bzw. ..... beziehungsweise
Hz..... Hertz
s..... Sekunden
µs..... Mikrosekunden
F..... Frequenz
t..... Zeit
dB..... Dezibel
ca. ..... cirka
dBFS..... Dezibel-relativ-zu-Full-Scale
K.I. .....Künstliche Intelligenz

1. Einleitung

1.1 Problemstellung

„Die Radiostimme geht direkt ins Ohr, sie wirkt anders. Anders, weil sie keinen Raum hat. Sie kann sich nicht ausbreiten, sondern wird über das Radiomikrofon direkt in die elektrischen Schaltkreise der Sendeanlage eingespeist. Es ist nicht nur die Stimme an sich, sondern gerade diese Stimme ohne Raum, die den Sound des Mediums Radio ausmacht“ (Patka 2015, 1).

Das einleitende Zitat stellt einen Idealfall bei der Produktion von Sprachaufnahmen dar. Aufnahme und Bearbeitung finden zur gleichen Zeit und am gleichen Ort statt. Durch die Digitalisierung ist es möglich geworden, die Produktionsschritte einer Sprachaufnahme zu trennen.
Verursacht wurde diese Bewegung durch die Entwicklung von integrierten Schaltkreisen, welche die Gestaltung von komplexeren Signalverarbeitungs-systemen zuließ. Die entstandenen Analog-Digital-Wandler ermöglichen in der Audioprozesstechnik die Umwandlung von kontinuierlichen Signalen in diskrete Signale. (vgl. Walden 1999, 539)

Diese Entwicklung ermöglichte somit, dass die Aufnahme und die Bearbeitung einer Sprachaufnahme auf verschiedene Speichermedien vorgenommen werden kann. Das Internet und dessen Möglichkeiten gewährleistete zusätzlich mit Übertragungsmedien die örtliche Trennung des Aufnahmeprozesses. In Summe kann die Aufteilung der Aufnahme, Bearbeitung und Speicherung eine Minderung der Qualität der Sprachaufnahme hervorrufen.

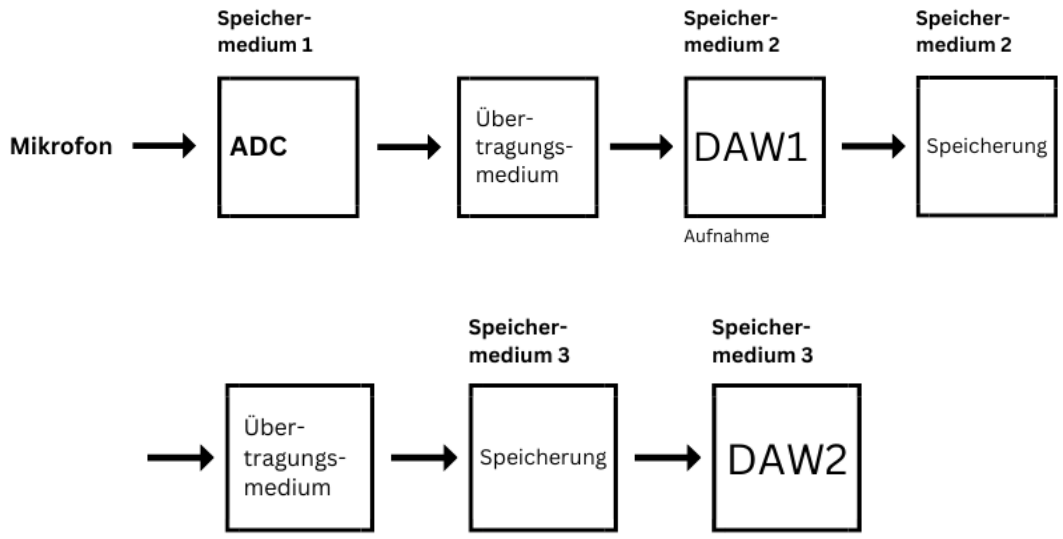

Abbildung 1: Der „neue Produktionsweg“ für die Produktion von Sprachaufnahmen: Örtlich getrennte Aufnahme durch ein Übertragungsmedium im Internet (trennt Speichermedium 1 vom Speichermedium 2) und ausgelagerter Postproduktion getrennt durch eine weiteres Übertragungsmedium (trennt Speichermedium 2 vom Speichermedium 3). [ADC - Analog-Digital Wandler, DAW - Digital Audio Workstation]

Besonders die Übertragungsmedien über das Internet und das Mikrofon können bei falscher Handhabung eine zusätzliche Fehlerquelle darstellen. Dieser neue Produktionsweg (siehe Abbildung 1 - vom Mikrofon bis zur DAW2) kann nun auch neue Formen von Artefakten in Sprachaufnahmen hervorrufen.

Diese Arbeit behandelt zwei Sprachaufnahmen, welche über diesen neuen Produktionsweg (siehe Abbildung 1) entstanden sind. Beide Sprachaufnahmen wurden über ein Übertragungsmedium mit einem im Speichermedium (1) integrierten Mikrofon auf einem örtlich getrennten Speichermedium (2) aufgenommen. Für die Postproduktion wurde diesen Sprachaufnahmen schlussendlich auf ein drittes Speichermedium übertragen. Ziel dieser Arbeit ist es, die entstandenen Sprachaufnahmen unter Berücksichtigung der theoretischen Grundlagen und der tontechnischen Theorie qualitativ zu beurteilen und nach den aktuellen Standards einer Postproduktion zu unterziehen.

Daraus sollen die Auswirkungen des neuen Produktionsweges auf die Audiodateien sichtbar und hörbar gemacht werden und die Umsetzbarkeit einer Postproduktion getestet werden. Unter Postproduktion wird in dieser Arbeit die Restauration und das Sound Design von Aufnahmen verstanden. Zudem sollen die Aufnahmen einer Postproduktion mit künstlicher Intelligenz (kurz K.I.) unterzogen werden und mit der zuvor vorangegangenen manuellen Ergebnissen mit konventioneller (ohne K.I.) Signalverarbeitung verglichen werden.

1.2 Methode und Forschungsfragen

Daraus ergeben sich folgende Forschungsfragen, welche im Zuge dieser Arbeit beantwortet werden:
FF1: Inwieweit verändern die neuen Produktionswege über das Internet die Qualitätsanforderungen an Sprachaufnahmen?
FF2: Inwieweit können Sprachaufnahmen auf Basis des vorangegangenen neuen Produktionsweges restauriert werden?
FF3: Inwiefern können daten- und KI-basierte Tools in die Postproduktion integriert werden?
FF4: Inwiefern können Qualitätsunterschiede zwischen manueller und einer allumfassenden KI-basierten Postproduktion festgestellt werden?

Für die Beantwortung der ersten Forschungsfrage wird eine qualitative Inhaltsanalyse durchgeführt. Dafür sollen zuerst die technischen Grundlagen gegenübergestellt und daraus im Kontext mit Sprachaufnahmen die Anforderungen für die Produktion diskutiert werden. Aufbauend auf diesen Erkenntnissen sollen die definierten Anforderungen mit den Auswirkungen der neuen Produktionswege verglichen und neue objektive Qualitätsanforderungen abgeleitet werden.

Für die Beantwortung der zweiten Forschungsfrage werden anhand einer Literaturdiskussion die Bestandteile einer Postproduktion und die Möglichkeiten der Restauration zur Herstellung einer finalen Sprachaufnahme diskutiert werden. Die gewonnen Erkenntnisse werden anschließend anhand der Sprachaufnahmen umgesetzt. Dafür werden nach allen wesentlichen Arbeitsschritten die Zwischenergebnisse hörbar gemacht und analysiert. In diesem Teil der Arbeit sollen typische Artefakte, welche über den neuen Produktionsweg entstehen mit manuellen und KI-basierten Signalverarbeitungsalgorithmen entfernt werden.

Für die Beantwortung der dritten Forschungsfrage soll zuerst der beste zu erreichende Bearbeitungsgrad mit einer manuell adaptierten und konventionellen Signalverarbeitung dargestellt werden. Ausgehend von den bestmöglichen Ergebnissen sollen zur weiteren Bearbeitung KI-basierte Signalverarbeitungsalgorithmen zusätzlich in den Prozess der Postproduktion integriert werden. Um die Forschungsfrage zu beantworten, soll in diesem Teil der Arbeit die sinnvolle Integration von KI-basierten Signalverarbeitungsalgorithmen im Kontext mit den Artefakten aus den neuen Produktionswegen diskutiert werden. Hier sollen die Grenzen bzw. die Fähigkeiten der parametrischen und der KI-basierten Signalverarbeitungsalgorithmen sichtbar werden.

Für die Beantwortung der letzten Forschungsfrage werden die Ergebnisse mit konventionelle Signalverarbeitungsalgorithmen und einer allumfassenden KI-basierten Postproduktion miteinander verglichen. Dafür sollen die finalen Ergebnisse der Postproduktion analysiert und die Unterschiede aufgezeigt werden. Abschließend soll anhand einer Umfrage im Vergleich die wahrgenommene Qualität der final bearbeiteten Sprachaufnahmen und der wesentlichen Zwischenschritte abgefragt werden. Die Ergebnisse der Umfrage sollen zudem als Hilfe bei der Beantwortung aller Forschungsfragen dienen.

2. Die Eigenschaften einer Sprachaufnahme

2. 1 Begriffserklärung Digital Audio

Unter Schall versteht man Wellen - Longitudinalwellen, welche durch das Medium Luft von einem Punkt zum anderen übertragen werden (vgl. Roederer 2000, 1). Für die Umwandlung dieser Wellen in ein digitales Format muss ein elektroakustisches Audiosignal gesamplet werden. (vgl. Ciesla 2022, 1)
Die Entwicklung von digitaler Technologie ermöglichte diese Umwandlung und erweiterte die Handlungsmöglichkeiten, welche mit analoger Technologie nicht möglich gewesen wären (vgl. Watkinson 2003, 1). Verantwortlich dafür war im Speziellen die Entwicklung des Analog-zu-Digital-Konverters, kurz auch AD-Wandler genannt (vgl. Walden 1999, 1). Analoge akustische Signale werden damit in den digitalen Raum transformiert und in ein datenbasiertes Objekt umgewandelt (vgl. Watkinson 2003, S. 1). Das bedeutet, dass aus einem zeitlichen Signal ein diskretes Signal hergestellt wird (vgl. Walden 1999, 1). Somit können analoge akustische Signale gespeichert, bearbeitet und übertragen werden (vgl. Watkinson 2013, 1).

2.2 Die Dimensionen einer Audiodatei

2.2.1 Sample-Rate und Bit-Tiefe

Für die Digitalisierung muss ein Audiosignal in ein zeit- und amplitudendiskretes Signal umgewandelt werden. Die Sampling-Rate gibt an, wie oft der Zustand eines Signals in einer Sekunde abgetastet wird (vgl. Ciesla 2022, 1). Die Sample-Rate von 44100 Hz bedeutet somit, dass in einer Sekunde der Signalamplitudenwert 44100 mal abgetastet wird.

F [Hz] = 1 / t [s]
44100 = 1 / t
44100 * t = 1
t = 1 / 44100
t = 0.00002268 [s] = 22,68 µs

Bei einer Sampling-Rate von 44100 Hz wird ein Signal also alle 22,68 µs abgetastet. Die zugrunde liegende Formel für diese Berechnung ist
F [Hz] = 1 / t [s].

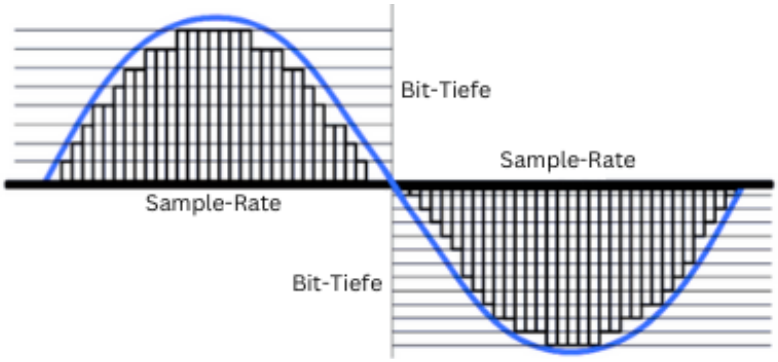

Abbildung 2: Sampling-Rate und Bit-Tiefe. Die Quantisierungsschritte und Bit-Tiefen sind in dieser Abbildung symbolisch (Abbildung für diese Arbeit modifiziert). [1]

Das analoge Signal ist in der Abbildung 2 als blaue Linie dargestellt. Die schwarze eckige Linie stellt die Auswirkung einer definierten Sampling-Rate und einer definierten Bit-Tiefe auf die Amplitudenquantisierung des gesampleten Signals dar. Die vertikalen Linien (im Bild schwarz) stellen dabei die Sample-Rate dar. Bei einer Frequenz von 44100 Hz würde dieser Abstand zwischen den schwarzen vertikalen Linien 22,68 µs betragen. Die horizontalen Linien (im Bild schwarz) stellen die Bit-Tiefe beim Sampling dar. Die Bit-Tiefe bestimmt somit die Anzahl der Bits, welche jedem Sample zugeordnet sind und definiert den möglichen Dynamikumfang des gesampleten Zustands (vgl. Ciesla 2022, 2-3).

Bit Depth	Typical Usage Scenario	Possible Amplitude Values	SNR
4	Sampled 1980s digital audio	$2^4 = 16$	≈ 24 dB
8	1990s video game audio	$2^8 = 256$	≈ 48 dB
16	CD and modern video game audio	$2^{16} = 65536$	≈ 96 dB
24	Recording studio	$2^{24} = 16777216$	≈ 144 dB
32	High-end recording studio	$2^{32} = 4294967296$	≈ 192 dB

Abbildung 3: Bit-Tiefe und Dynamikumfang (Ciesla 2022, S. 3)

[1] *https://www.masteringthemix.com/blogs/learn/113159685-sample-rates-and-bit-depth-in-a-nutshell*; abgerufen am 14.4.2023

Die Abbildung 3 erklärt den Zusammenhang zwischen Bit-Tiefe und dem möglichen Dynamikumfang eines Zustands. Ein mit 24 bit gesampleter Zustand hat einen möglichen Dynamikumfang von 144 dB.

2.2.2 Pulse-Code-Modulation und Pulse-Amplitude-Modulation

Der Vorgang des Digitalisierens wird in der Digitaltechnik als Pulse Code Modulation bezeichnet (vgl. Ciesla 2022, 4). Wie schon zuvor erwähnt wird das Signal auf Basis der Sampling-Rate periodisch gemessen. Diese Ebene wird in Abbildung 4 auf der y-Achse (Time) repräsentiert. Zudem wird in jeder Situation (T1, T2, T3...Tn) die Spannung gemessen und einer Zahl zugeordnet (vgl. Watkinson 2003, 4). Dieser Prozess wird als Quantisierung bezeichnet. Je höher also die Sampling-Rate und die Bit-Tiefe während des Vorgangs, desto genauer wird das digitale Abbild der analogen Signalform (vgl. Ciesla 2022, 3). Am Ende des Prozesses wird nun jedes Sample (x-Achse) mit einem bestimmten Wert (y-Achse) repräsentiert (vgl. Walden 1999, 540). In Abbildung 4 ist der Vorgang der Quantisierung visuell dargestellt.

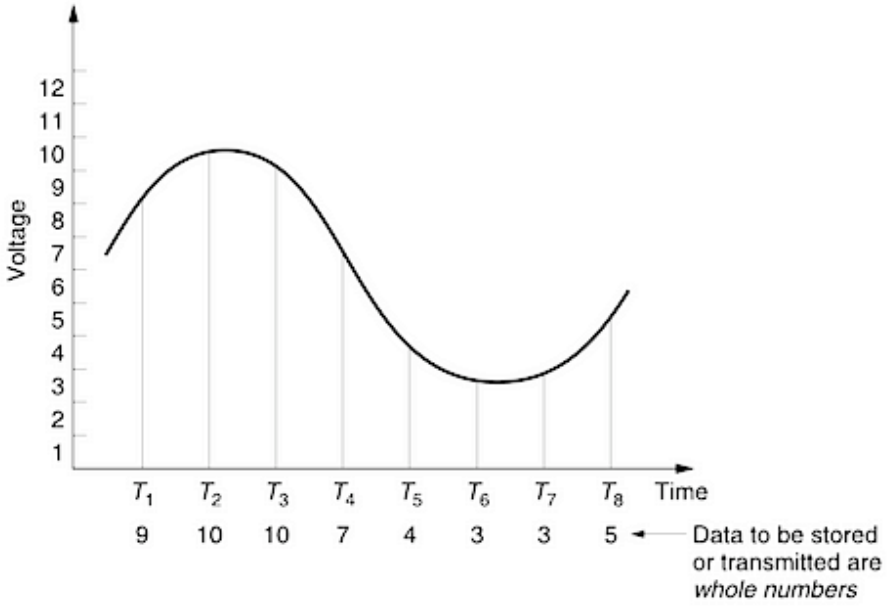

Abbildung 4: Quantisierung (Watkinson 2003, 4)

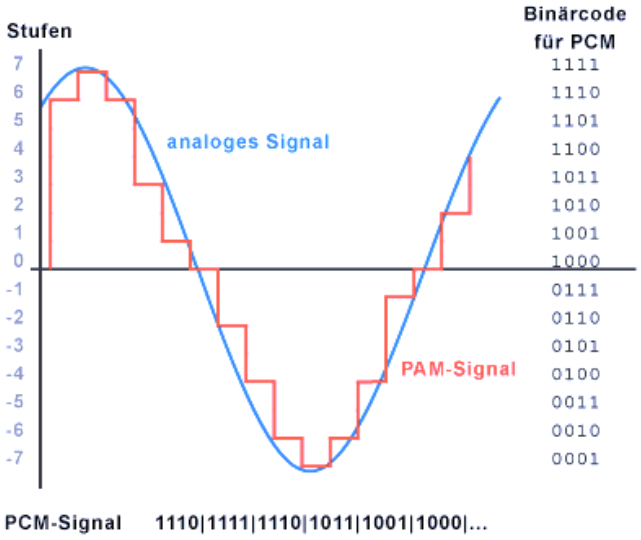

Abbildung 5: Pulse-Code-Modulation und Pulse-Amplituden-Modulation [2]

In Abbildung 5 ist die quantisierte Wellenform (eckig und rot) erkennbar. Diese Wellenform wird PAM (Pulse-Amplitude-Modulation) genannt. Hier wird wie zuvor erwähnt, Die Spannung jedes Samplezustands einem Zahlenwert zugeordnet. Anhand der Stufen wird dann jedes Sample einem Binär-Code zugeschrieben. Das binäre Signal wird danach PCM(Pulse-Code-Modulation) - Signal genannt. Während der Quantisierung kann es zu Fehler kommen. Dieser Fehler wird immer dann maximal, wenn bei der Quantisierung der analoge Wert zwischen zwei Quantisierungsstufen liegt. Zu Fehlern kommt es somit, wenn der wahre Wert irgendwo zwischen zwei definierte und quantisierte Amplitudenwerte liegt. (vgl. Ciesla 2022, 4)

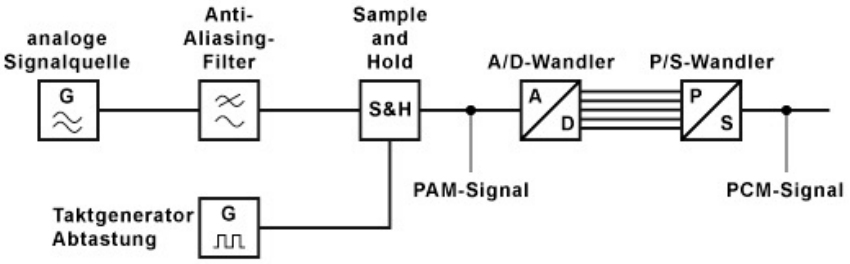

Abbildung 6: Pulse-Code-Modulation [2]

[2] https://www.elektronik-kompendium.de/sites/kom/0312281.htm; Elektronik-Kompendium.de Patrick Schnabel, Droste-Hülshoff-Straße 22/4, D-71642 Ludwigsburg, abgerufen am 7.5.2023.

Die in Abbildung 6 dargestellte Grafik stellt alle Elemente bis zum Endresultat Pulse-Code-Modulation dar.

1) analoge Signalquelle: damit ist die elektroakustische Welle gemeint
2) Anti-Aliasing-Filter: Abhängig von der Sampling-Rate und der Bit-Tiefe wird gemäß dem Nyquist-Theorem die Bandbreite des Signals so begrenzt, sodass die Sampling Frequenz zumindest doppelt so groß ist wie die höchste Frequenz des zu digitalisierenden Signals (vgl. Ciesla 2022, 11)
3) Taktgenerator: Die Word-Clock ist dafür verantwortlich, dass der Abstand zwischen den einzelnen Samplezuständen immer gleich groß ist (vgl. Ciesla 2022, 15).
4) Sample and Hold
5) PAM: Quantisierung (siehe S. 6)
6) A/D-Wandler: Dieser Wandler schreibt jeder Stufe einen Binärcode zu.
7) P/S Wandler: Dieser Wandler schichtet den Binärcode um, um ihn auslesbar zu machen. Die Stufen (bzw. der Code) werden somit in die richtige Reihenfolge gebracht.
8) PCM: Diskretes Signal (siehe S. 6)

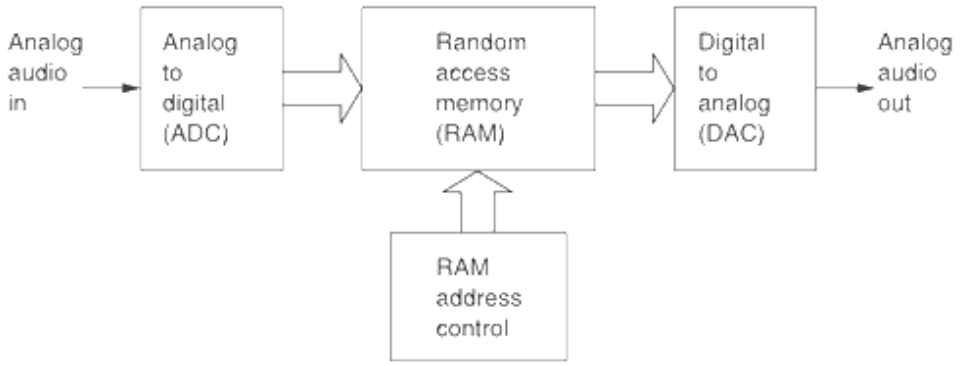

Abbildung 7: Random Access Memory (*Watkinson* 2003, 11)

In Abbildung 7 wird der Speicherplatz des Binärcodes bzw. des PCM Signal dargestellt. Für die Speicherung des Binär-Codes existiert das sogenannte Random Access Memory. Das RAM dient im digitalem Raum als ein begrenztes Speichermedium (vgl. Watkinson 2003, 11). Artefakte beim Sampling entstehen dann, wenn die Sample-Rate nicht doppelt so groß ist wie höchste abzutastende Frequenz.

2.2.3 Dateiformate und Kompression

Digital Audio ist zusammenfassend eine andere Möglichkeit Audioinformationen zu übertragen und zu speichern (vgl. Watkinson 2003, 3).

Grundsätzlich können bezüglich den Formate zwei Kategorien unterschieden werden. Die erste Kategorie nennt sich Lossless Audio. Die zweite ist das Gegenteil von Lossless Audio und wird Lossy Audio bezeichnet. Zur Kategorie Lossless Audio werden jene Formate gezählt, welche keine Verluste bei der Erzeugung vorweisen. Im Gegensatz zu Lossless Audio spricht man von Lossy Audio, wenn bei der Erzeugung der Audiodatei Verluste entstehen. (vgl. Ciesla 2022, 11)
Im digitalen Raum wird das Kürzel PCM auf manchen Geräten als Synonym für verlustfreie Audio verwendet (vgl. Ciesla 2022, 4). Dazu zählen Waveform Audio File Format (WAV) und Audio Interleaved File Format (AIFF).

Mp3 wird der Kategorie Lossy Audio zugeordnet (vgl. Ciesla 2022, 16). Bei dem Format mp3 handelt es sich um einen Codec, welcher auf Basis des Hörsinnes des Menschen erstellt wurde (vgl. Fuchs und Maxwell 2016, 1). Bei einem Codec handelt es sich um eine Software, welche Informationen zuerst encodiert und danach wieder decodiert (vgl. Ciesla 2022, 15). Der Encoder entfernt beim Encodieren jene Teil des Audiosignals, welche für die auditive Wahrnehmung des Menschen nicht nötig sind (vgl. Fuchs & Maxwell 2016, 1). Mp3 hat im Gegenteil zu WAVE und AIFF keine Bit-Tiefe jedoch aber eine Bit-Rate (vgl. Ciesla 2022, 15).

Hier wird dann von bits per second (kbps) gesprochen (vgl. Fuchs und Maxwell 2016, S.1). Je höher die Bit-Rate der mp3-Datei, desto besser ist schlussendlich die Qualität der Audio.

Eine sehr niedrige Bit-Rate kann den Effekt „double-speak“ verursachen. Wenn die Audiodatei dann auf einem Kopfhörer oder hochauflösenden Lautsprechern abgehört wird, kann ein Echo der Aufnahme festgestellt werden (vgl. Brandenburg 1999, S. 7).

File Format	Parameters	File Size for One Minute of Audio
WAV/AIFF	16 bits, 44.1 kHz, stereo	≈ 10.5 megabytes
	24 bits, 96 kHz, stereo	≈ 32 megabytes
	16 bits, 44.1 kHz, mono	≈ 5 megabytes
MPEG-1 Audio Layer 3 (MP3)	16 bits, 44.1 kHz, stereo, encoded at 128 kbps	≈ 0.9 megabytes
	16 bits, 44 kHz, stereo, encoded at 320 kbps	≈ 2.4 megabytes

Abbildung 8: Dateigröße im Vergleich zwischen WAVE (in der Abbildung WAVE /AIFF) und mp3 pro Minute (*Ciesla* 2022, 18)

In Abbildung 8 erkennt man die Dateigrößen der wichtigsten Dateiformate. In dieser Darstellung ist der Einfluss der Sample-Rate und Bit-Tiefe auf die Größe von Lossless Formaten und der Einfluss der kbps auf die Größe von mp3-Dateien klar ersichtlich. Trotz identer Sample-Rate und Bit-Tiefe (bei WAVE/AIFF und mp3) ist bei unterschiedlicher kpbs der Größenunterschied zwischen den oben dargestellten mp3-Dateien (zwischen 128 kbps und 320 kbps) evident.

Bei abnehmender Größe der kbps wird zunehmend die Bandbreite der Audiodatei eingeschränkt. In Anbetracht der kompletten Bandbreite gehen die höheren Frequenzen zuerst verloren. (vgl. Brandenburg 1999, 7). WAVE und AIFF-Formate sind somit während des Produktionsprozesses einer Audiodatei von Vorteil, da das ursprüngliche Audiosignal am detailreichsten abgebildet wird. Mp3 Formate eignen sich eher für die Bereitstellung eines finalen

Produktes (vgl. Ciesla 2022, 16). Während des Produktionsprozesses soll auch keine Veränderung des Formats von WAVE auf mp3 stattfinden, weil dadurch Informationen verloren gehen. Auch diese Fehler können in der Postproduktion zu nicht behebaren Artefakten führen.

2.3 Die Anforderungen an eine Aufnahme für die optimale Abbildung von Sprache

Grundsätzlich soll der Anspruch darin liegen die finale Aufnahme so bereitzustellen, sodass

- keine Ermüdung des Zuhörers durch störende Frequenzen ausgelöst wird und
- der Zuhörer emotional von dem finalen Produkt angesprochen wird. (vgl. Ciesla 2022, 58)

Bezugnehmend auf Sprachaufnahmen sprechen Beerends & Beerends (2015, 176) zusammenfassend davon, dass eine „warme Radiostimme“ als qualitativ höherwertiger von Konsumenten und Konsumentinnen wahrgenommen wird.

2.3.1 Die Qualität der Stimme

Für die Beurteilung der Qualität einer Sprachaufnahme müssen zwei Ebenen betrachtet werden. Die erste Ebene bezieht sich auf die Qualität der aufgenommenen Stimme. Die zweite bezieht sich auf die Qualität der Audiodatei aus der technischen Perspektive. Dazu zählen die Auflösung, der Grad der Komprimierung, die Verzerrung und der Rauschpegel in einer Audiodatei. Die Kombination beider Ebenen wird als Sprachqualität bezeichnet. (vgl. Beerends & Beerends 2015,174)

Der erste Parameter betrifft somit die Charakteristik der Stimme. Seifert (2020, 244) spricht davon, dass eine Stimme nach drei Kriterien beurteilt werden kann:

1) Jitter

2) Shimmer
3) Klang-Lärm Verhältnis

Unter Jitter versteht man die kurzfristige Stabilität des Grundtons F0. F0 (Fundamentalfrequenz) ist der Grundton einer Stimme im Frequenzspektrum.

Unter Shimmer versteht man die kurzfristigen Unterschiede aufeinanderfolgender Amplituden. (vgl. Ferreira & Fernandes 2017, 1)

Der Klang einer Stimme entsteht durch einen Grundton und der entstehenden harmonischen Obertonreihe. Das entstehende auditive Bild wird als Klangfarbe oder Timbre bezeichnet. Je nach Zusammensetzung und Lautstärke der einzelnen Obertöne, verändert sich auch die Klangfarbe der Stimme. (vgl. Stassen 1995, 34)
Das Klang-Lärm Verhältnis ist somit das Laustärkenverhältnis zwischen Fundamentalfrequenz inklusive harmonischer Obertonreihe und Lärm (vgl. Ferreira & Fernandes 2017, 1).

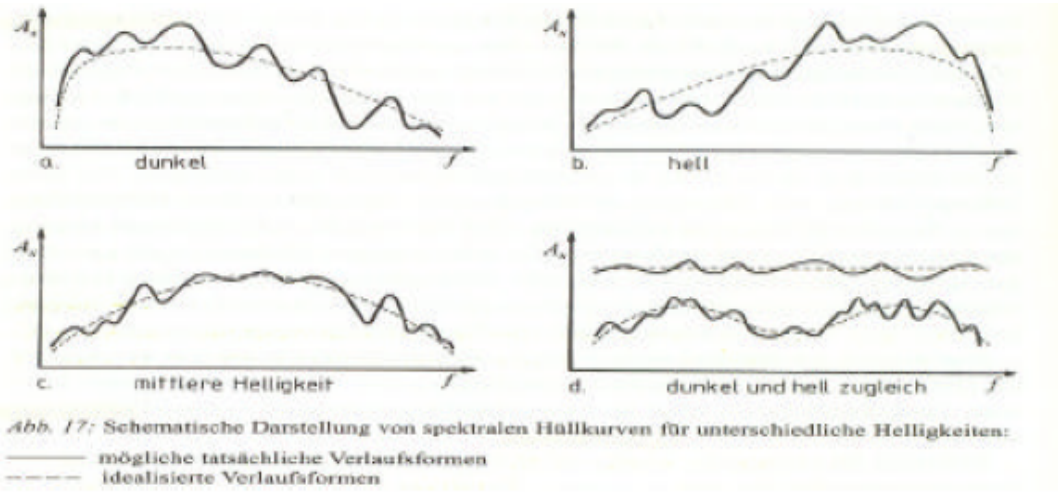

Abbildung 9: Klangfarben der Stimme (Neppert 1999, 66)

In der Abbildung 9 können verschiedene Verläufe der Amplituden der Teiltöne gesichtet werden. Je nach Verlauf der Amplituden verändert sich Klangfarbe der Stimme (oben: dunkel, hell, mittlere Helligkeit, dunkel und hell zugleich).

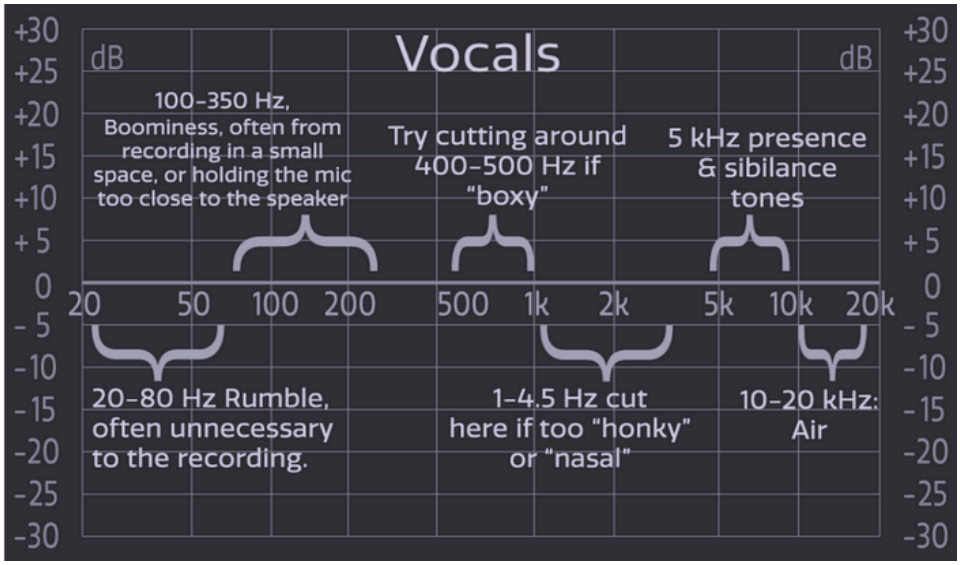

Abbildung 10: Charakteristik von Frequenzbereichen [3]

In Abbildung 10 kann eine detailliertere Analyse des Frequenzbandes für die Beurteilung von Sprachaufnahmen betrachtet werden. Die oben angegebenen Werte und Bereiche sind nicht bei jeder Sprachaufnahme gleich. Diese Frequenzen stellen Richtbereiche dar und können in der Postproduktion zur Orientierung dienen. Die Bereiche bieten eine Ausgangsbasis für die Bearbeitung von der aus durch aktives Hören die Frequenzen der vorliegenden Sprachaufnahme adaptiert werden können.

An dieser Stelle ist noch zu erwähnen, dass die Eigenschaften von Sprachkrankheiten in dieser Arbeit nicht behandelt werden. Es wird grundsätzlich bei den nachfolgenden Beispielen von einem gesunden Sprachapparat ausgegangen. Diese Arbeit bezieht sich auf jene Eigenschaften und Informationen einer Sprachaufnahme, welche für die Postproduktion, d.h. für die Restauration und das Sound Design von Bedeutung sind.

[3] *https://www.bhphotovideo.com/explora/pro-audio/tips-and-solutions/8-eq-tips-videographers*, B & H Foto & Electronics Corp. 420 9th Ave, New York, NY 10001; abgerufen am 24.5.2023.

2.3.1 Die Qualität der Audiodatei

Die zweite Ebene für die Beurteilung der Qualität einer Sprachaufnahme betrifft die technischen Parameter. Dazu zählen die Auflösung (Sample-Rate und Bit-Tiefe), der Grad der Komprimierung, die Verzerrung und der Rauschpegel in einer Aufnhame. (vgl. Beerends und Beerends 2015, 1)

a) Auflösung: (SNR: Sound-to-Noise Ratio): Darunter versteht man die Geräuschentwicklung bei nicht adäquater Sample-Rate und Bit-Tiefe (vgl. Ciesla 2022, 3).

b) Grad der Komprimierung: Je nach Dateiformat, kann das Objekt formatbedingte Artefakte vorweisen – Umformatierung (siehe S. 7).

c) Verzerrung: Um Verzerrungen zu vermeiden muss ein Audiosignal unter 0 dBFS aufgenommen werden (vgl. Ciesla 2022, 136-137).

d) Rauschpegel in der Aufnahme

Die gesamte Qualität der Aufnahme, welche sich aus der Summe der Fehler der Sprache und der Audiodatei ergeben, sollte in der Postproduktion so gut wie möglich behoben werden. Hierfür soll für die Verbesserung der Sprachaufnahme in der Postproduktion

- Veränderungen an der Klangfarbe
- und der Lautstärke

vorgenommen werden.

Zudem sollten Effekte wie

- De-Esser,
- Reverb-Entferner und

- Geräuschentferner

für die weitere Fehlerkorrektur miteinbezogen werden. (vgl. Beerends und Beerends 2015, 176)
Die Aufgabe in der Postproduktion ist somit nicht nur das Beheben von Fehlern, sondern auch das Herstellen eines emotional ansprechbaren Produktes. In diesem Punkt überschneiden sich die Aufgaben der Restaurierung und des Sound Designs.

3 Elemente für die Produktion und Postproduktion von Sprachaufnahmen

Die Fehler in einer Sprachaufnahme müssen so gut wie möglich korrigiert werden. In der vorangegangenen Diskussion wurde über diese möglichen Fehler gesprochen, jedoch nicht die über dazu nötigen Bearbeitungsalgorithmen. Für die Postproduktion einer Sprachaufnahme stehen grundsätzlich folgende Signalverarbeitungselemente zur Verfügung:

- Equalizer,
- Kompressoren,
- Limiter und
- Filter.

Zusätzlich zu diesen Elementen existieren Audioeffekte, welche eine „Lern“-Funktion besitzen. Dazu zählen zum Beispiel zum

- Reverb-Entferner und
- Geräuschentferner.

Ciesla (2022, 59) spricht zusätzlich davon, dass das Erreichen der richtigen Lautstärke einer Sprachaufnahme ebenfalls Teil der Postproduktion ist. Bei der Postproduktion einer einzigen Spur, in diesem Fall eine Sprachaufnahme muss es nicht unbedingt zu einer Trennung der Postproduktion und des Masterings

kommen. Diese zwei Vorgänge sind bei der Herstellung eines musikalischen Stückes jedoch üblicherweise voneinander getrennt. Bevor es jedoch zur Postproduktion einer Aufnahme kommt, muss diese zuerst hergestellt werden. Welchen Einfluss der Aufnahmeprozess auf die Abbildung der Sprache hat, wird im nächsten Kapitel erklärt.

3.1 Aufnahme

Der Fokus dieser Arbeit liegt auf die Restauration und das Sound Design (gemeinsam Postproduktion) von Sprachaufnahmen, jedoch bedarf der Einfluss der Aufnahmedurchführung auf die Qualität einer näheren Erklärung.
Für eine qualitativ hochwertige Aufnahme der Sprache sind laut Ciesla (2022, 128) unter optimalen Bedingungen vier Elemente zu berücksichtigen:

1) Mikrofon und Interface
2) Vorverstärker
3) Schallisolierung
4) Abhörmöglichkeit

Dieses Kapitel behandelt vorrangig den Punkt 1, da die Wahl des Mikrofons und die Durchführung der Aufnahme unter anderem als Teil des aktiven Sound Designs gesehen werden kann. Diese beiden Entscheidungen sind in der Gestaltung nicht zu vernachlässigen und haben einen wesentlichen Einfluss auf die Qualität der Sprachaufnahme.

Mikrofone wandeln Schall in Wechselspannung um (vgl. Dickreiter et.al. 2014, 139). Die Wahl des Mikrofons hat einen wesentlichen Einfluss auf die Qualität der aufgenommen Sprache. Bei der Wahl des Mikrofons ist darauf zu achten, dass der jeweilige Typ des Mikrofons zu Stimme passt (vgl. Ciesla 2022, 129).

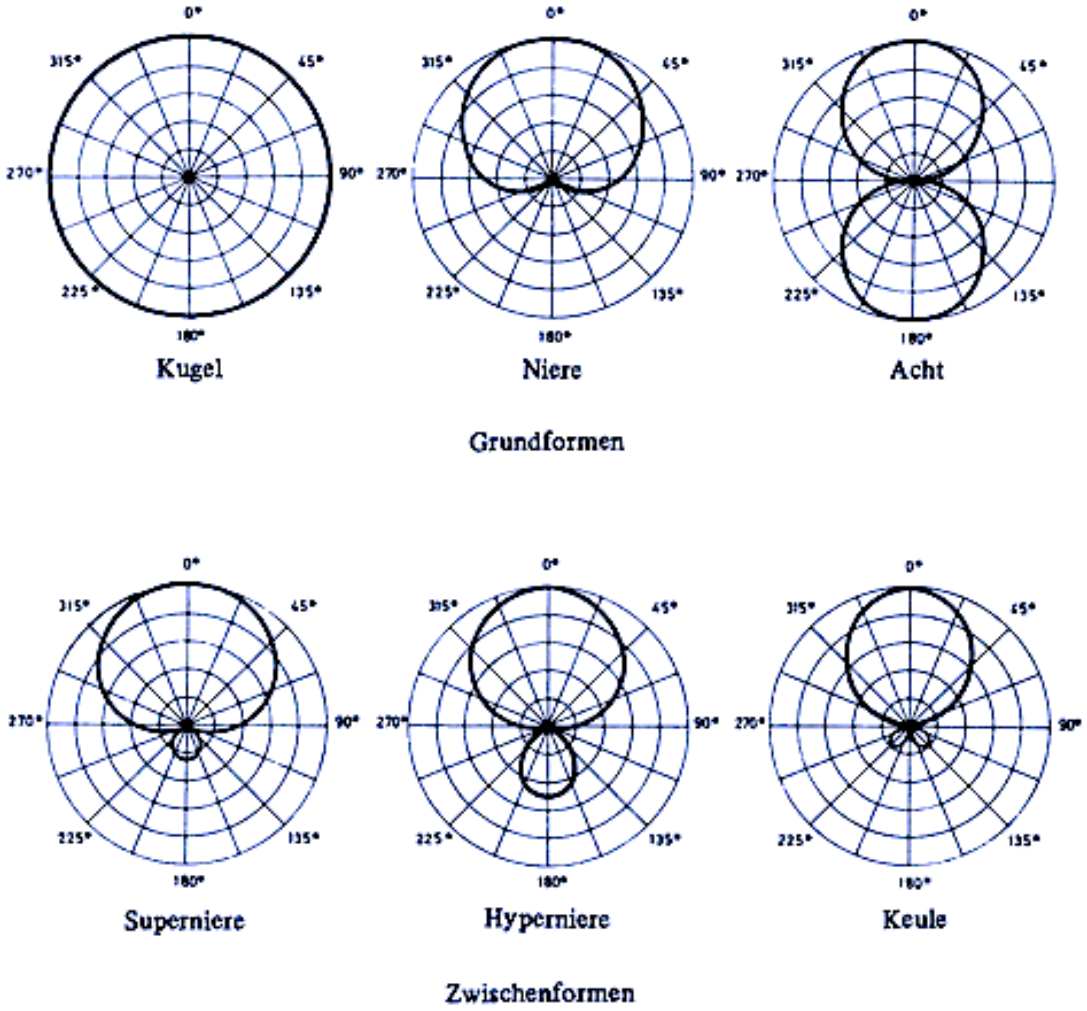

Abbildung 11: Richtcharakteristiken bei Mikrofone (Dickreiter et.al. 2020, 151)

Des Weiteren besitzen Mikrofone aufgrund ihrer Bauform auch eine bestimmte Richtcharakteristik. In Abbildung 11 sind die Grundformen (erste Reihe oben) und die Zwischenformen (zweite Reihe unten) dargestellt.
Idealerweise sollten Mikrofone in allen Frequenzbereichen dieselbe Richtcharakteristik beziehungsweise die exakt gleiche Form aufweisen. In der Praxis ist dieser Zustand aber nicht gegeben und es kommt daher zu Schwächen und Stärken in bestimmten Frequenzbereichen. (vgl. Dickreiter et.al. 2014, 151)

Grundsätzlich können zwei Mikrofontypen voneinander unterschieden werden, nämlich Dynamische Mikrofone und Kondensatormikrofone. Diese beiden Typen werden auch für den professionellen Einsatz für Sprachaufnahmen verwendet (vgl. Dickreiter et.al. 2014, 152).

3.1.1 Dynamische Mikrofone

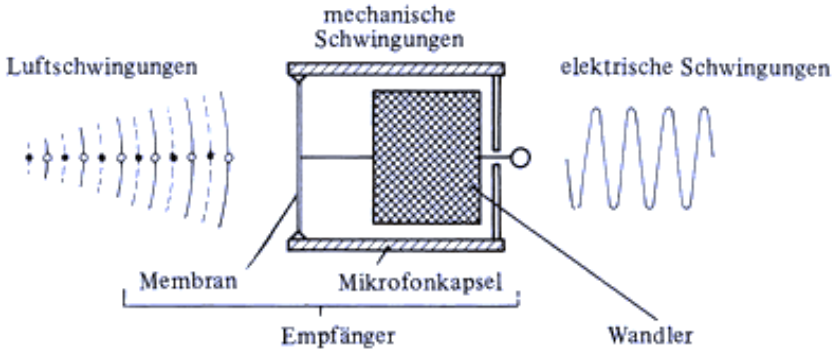

Abbildung 12: Prinzip eines Dynamischen Mikrofons - vereinfacht (Dickreiter et.al. 2020, S. 139)

Beim Dynamischen Mikrofon treffen die Schallwellen zuerst auf eine Membran. Diese Membran ist mit einem Wandler verbunden, welche sich in einer Mikrofonkapsel befindet. Die Membran bewegt sich aufgrund der auftreffenden Luftschwingungen und bewegt somit auch den Wandler in der Kapsel. Durch die Bewegung des Wandlers in der Kapsel wird nur Strom induziert. Durch diese Übertragung wird bei einem Dynamischen Mikrofon Schallschwingungen in elektrische Spannung umgewandelt. (siehe Abbildung 12)

Der Nachteil Dynamischer Mikrofone ist, dass sie im Vergleich zu Kondensatormikrofonen eine geringere Impulstreue und einen geringeren Ausgangspegel haben. Dynamische Mikrofone neigen dazu, nicht so schnell zu verzerren und eignen sich deshalb für die Bühnenumgebung. (vgl. Hansch und Rentschler 2013, 91-92)

3.1.2 Kondensatormikrofone

Kondensatormikrofone arbeiten mit einer zusätzlichen Phantomspeisung. Die Membran dient als Elektrode und der Kondensator als Gegenelektrode. Die Luft dazwischen ist das Dialektrikum. Wirkt Schall auf die Membran ein, ändert sich der Abstand zwischen Membran und Kondensator. (vgl. Bernstein 2019, 194-195)

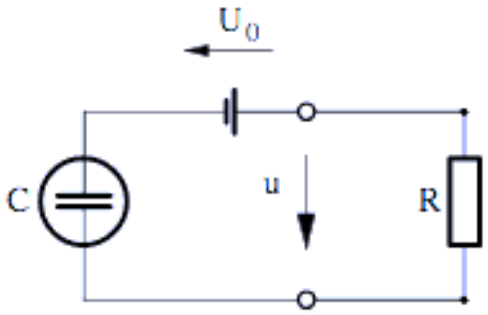

Abbildung 13: Funktionsprinzip eines Kondensatormikrofons (U_0 – Spannung, C – Kapazität, u – Spannungs, R-Widerstand)

In Abbildung 13 ist ein Schaltkreis eines Kondensatormikrofons dargestellt. U_o ist in diesem Fall die induzierte Spannung. Die Membran wird durch Schalleinwirkung bewegt und damit ergeben sich somit Kapazitätsunterschiede zwischen Membran (Elektrode) und Gegenelektrode. R ist der Widerstand und bewerkstelligt einen Spannungsabfall.

Für detailliertere Aufnahmen bieten sich eher die Kondensatormikrofone an, da sie aufgrund ihrer Wirkweise empfindlicher gegenüber Schalleinwirkungen sind. Der Nachteil ist jedoch die Neigung früher zu verzerren. Der Vorteil Dynamischer Mikrofone für die Aufnahme von Sprachaufnahmen liegt bei der höheren Toleranz bezüglich größerer Dynamikunterschiede. Die Wahl des Mikrofons ist somit vom gewünschten Endergebnis abhängig.

Da Dynamische Mikrofone und Kondensatormikrofone auch Frequenzbereiche unterschiedlich wiedergeben, kann das Abgleichen mit der Klangfarbe und die bewusste Entscheidung für einen Mikrofontyp der Aufnahme eine bestimmte Charakteristik verleihen. In der Praxis ist die Auswahl eher ein Luxusproblem, jedoch kann die Wahl des Mikrofons schon in der Aufnahmephase der Stimme als Teil des Sound Designs gesehen werden. Jedes Mikrofon hat Vor- und Nachteile.

3.1.3 Nahbesprechungseffekt

Der Nahbesprechungseffekt verursacht ein Anheben der Frequenzen im Bassbereich. Dieser Effekt entsteht jedoch nur bei Druckgradientenempfänger. Dazu zählen Mikrofone mit Nieren- sowie Achtrichtcharakteristik. Zudem noch alle dazugehörigen Zwischenformen. (vgl. Dickreiter et. al. 2014, 143-144)

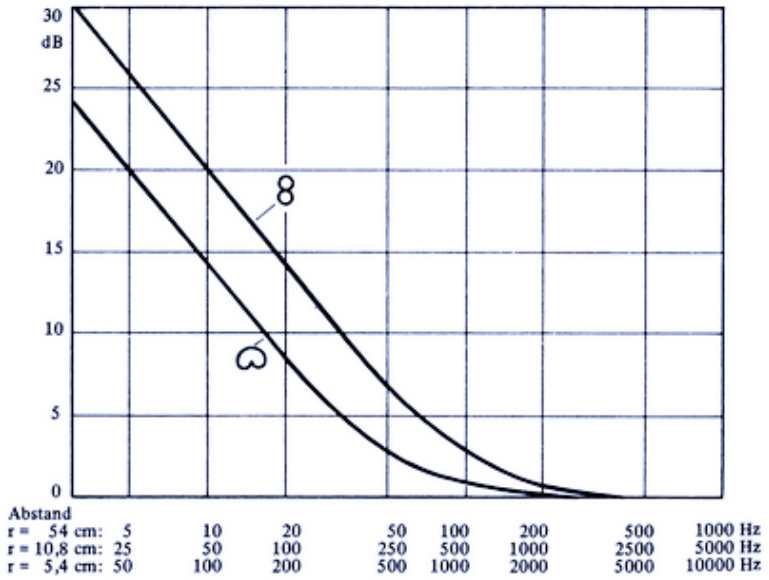

Abbildung 14: Der Einfluss des Wirkungsbereichs auf die Anhebung von (tiefen) Frequenzen. (Dickreiter et. al. 2014, S. 145)

Wichtig dabei ist zu erwähnen, dass die Sprachverständlichkeit durch den Nahbesprechungseffekt nicht negativ beeinflusst wird (vgl. Bernstein 2019, 194-195). In Abbildung 14 ist die Abhängigkeit der Lautstärke der tiefen Frequenzen abhängig vom Abstand der Schallquelle zum Mikrofon dargestellt. Je näher die Schallquelle – desto lauter die Frequenzen im Bassbereich.

Die Besonderheit bei Druckgradientenmikrofone ist, dass die Schallwellen auf die Membran sowohl auf die Vorderseite als auch auf die Rückseite einwirken. Dadurch kommt dieser Effekt zustande.

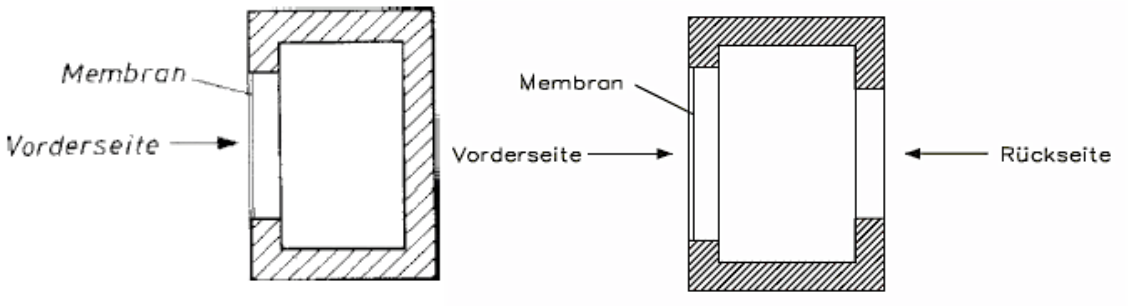

Abbildung 15: Mikrofonkapsel mit und ohne Öffnung (Bernstein 2019, 199-201)

Wie in der Abbildung 15 dargestellt, können bei der rechten Mikrofonkapsel Schallwellen auch auf der Rückseite der Membran auftreffen. Je nach Bedarf kann dieser Effekt bei der Aufnahme genutzt werden um der Stimme mehr Sättigung im Bassbereich zu verleihen. Die Wahl des Mikrofons und die Art der Aufnahme entscheiden schon sehr früh, welche Charakteristik die aufgenommene Stimme haben kann.

3.2 Equalizer

Der Ursprung des Equalizers liegt in der Telekommunikationstechnik. Dieser wurde dazu verwendet um abgeänderte Frequenzbereiche zu korrigieren. Zu leise Frequenzen wurden damit lauter gemacht und zu laute Frequenzen wurden damit leiser gemacht. Ziel war es, das ursprüngliche Signal wieder so gut wie möglich zu rekonstruieren. Ab diesem Zeitpunkt wird der Begriff „Equalization“ für die Veränderung der Lautstärken von Frequenzbereichen verwendet. (vgl. Välimäki und Reiss 2016, 1)

In der Produktion von Audioprodukten zählt der Equalizer zu den wichtigsten Elementen zur Manipulation von Frequenzen. Wie schon zuvor erwähnt, einerseits um Korrekturen vorzunehmen und andererseits um der Audioaufnahme bewusst eine bestimmte Klangfarbe zu verleihen. (vgl. Ciesla 2023, 57)
Die verschiedenen Arten von Equalizer werden in den nächsten Unterkapiteln erklärt.

Grundsätzlich gibt es drei Arten von Equalizern:

1) Parametric Equalizer
2) Graphic Equalizer
3) Linear Equalizer

3.2.1 Parametrische Equalizer

Der parametrische Equalizer ist im Vergleich zu den anderen Equalizern der flexibelste. Diese Flexibilität entsteht durch die individuelle Wahl

1) des zu bearbeitenden Frequenzbereiches Hz (Bandbreite),
2) der Filtergüte Q (breit oder schmal),
3) und der Lautstärkenveränderung. (vgl. Välimäki und Reiss 2016, 3)

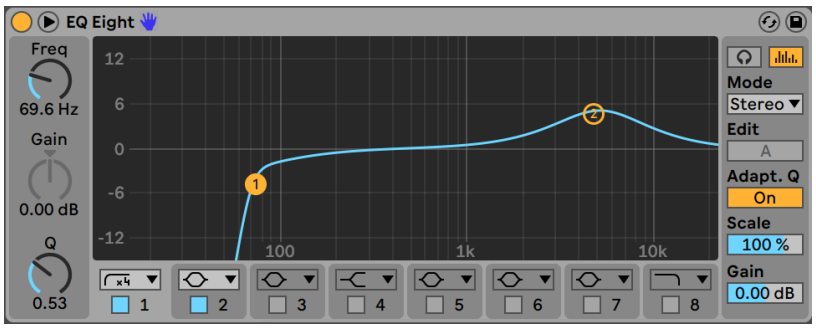

Abbildung 16: Screenshot eines Parametrischen Equalizers in Ableton Live 10.

Der in Abbildung 16 dargestellte Parametrische Equalizer ist ein digitaler Effekt der DAW Ableton Live 10[4]. Bei diesem Equalizer kann die zu bearbeitende Frequenz mit dem Regler *Freq*, die Filtergüte Q mit dem Regler *Q* und die Anpassung der Lautstärke mit *Gain* angepasst werden. Die Größe der Filtergüte bestimmt die Breite des zu bearbeitenden Frequenzbereiches (0: breit, 10 schmal). Für sehr genaue Eingriffe in der Sprachbearbeitung bietet sich dieser Equalizer aufgrund seiner genauen Parametrisierung sehr gut an.

[4] www.ableton.com

3.2.2 Grafischer Equalizer

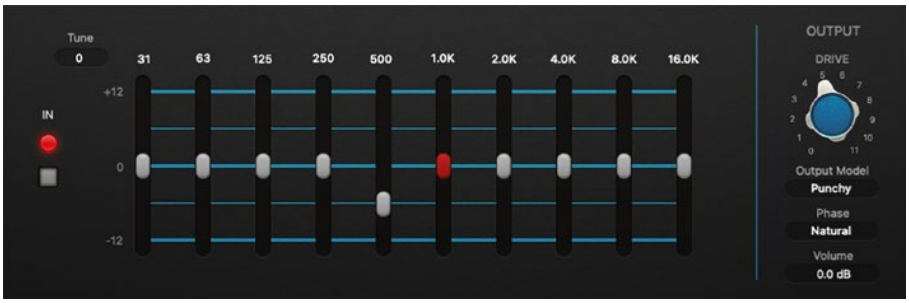

Abbildung 17: Graphic Equalizer der DAW Logic Pro (Ciesla 2023, 57)

Wie in Abbildung 17 dargestellt, ist die Bearbeitung bei einem Grafischen Equalizer auf bestimmte Frequenzbänder und einer fixierten Filtergüte limitiert (vgl. Välimäki und Reiss 2016, 11). Je nach Model können diese Frequenzbänder variieren. Mit der auf und ab Bewegung des Faders (im Bild grau und rot) können die Anpassungen der Lautstärke des Frequenzbandes vorgenommen werden.

3.1.3 Linearer Equalizer

Der lineare Equalizer arbeitet auf dem Prinzip der Phasenumkehr. Dieser Art des Equalizers ist neutral und bietet sich auch für genaue Korrekturen im Frequenzspektrum an (vgl. Ciesla 2023, 57). Auch dieser Typ eignet sich für die Bearbeitung von Sprachaufnahmen, jedoch sind diese Equalizer sehr kostenintensiv.

3.3 Kompressor

Kompressoren zählen zu den wichtigsten Signalverarbeitungsalgorithmen in der Nachbearbeitung von Audiospuren (vgl. Ciesla 2023, 67). In diesem Prozessschritt kann der Dynamikumfang einer Tonspur verringert werden (vgl. Giannoulis et. al., 399). Kompressoren verursachen somit eine Komprimierung der Audioaufnahme. Die leiseren Abschnitte werden verstärkt und die lauteren Abschnitte abgedämpft (vgl. Ciesla 2023, 67).

Der Kompressor gehört aufgrund seiner vielseitigen Einsetzbarkeit zu den komplexeren Audioeffektgeräten (vgl. Giannoulis et. al, 399). Eine extreme Form des Kompressors stellt der Limiter dar (vgl. Ciesla 2023, 67). Die Reduktion der Dynamik passiert nicht sofort sondern in Stufen (vgl. Giannoulis et. al, 399).

Für die Einstellung eines Kompressors benötigt man vier Parameter (vgl. Ciesla 2023, 67-68).

1) Treshold: Dieser Parameter bestimmt, ab welcher Laustärke der Kompressor einsetzt (vgl. Giannoulis et. al, 400).

2) Ratio: Verhältnis zwischen Input und Output Gain

3) Attack: Dieser Parameter bestimmt, wie schnell der Kompressor die Laustärke reduziert (vgl. Giannoulis et. atl, 400).

4) Release: Dieser Parameter bestimmt, wie lang der Kompressor braucht, um wie auf seine Ausgangsituation zurückzukommen (bei keinem erneuten Triggern des Kompressors). (vgl. Giannoulis et. al, 400).

5) Knee: unter Hard Knee versteht man das abrupte einsetzten eines Kompressor; unter soft Knee versteht man das kontinuirliche Einsetzten eines Kompressors;

6) Make-Up Gain: Dieser Regler wird für die Kompensation der verlorengegangenen Lautstärke durch die Kompression verwendet.

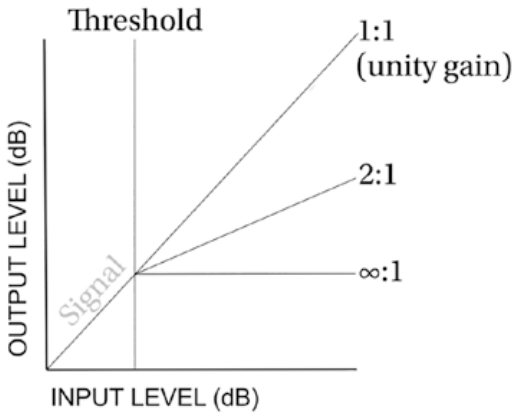

Abbildung 18: Darstellung eines Kompressors (Ciesla 2023, S. 67)

Das Einstellen der oben genannten Parameter hängt von dem gewünschten Ergebnis ab.

3.3.1 Hardware-Typen von Kompressoren

Grundsätzlich gibt es laut Ciesla (2020, 69-70) vier Kompressor Typen:

1) Tube Compressor: arbeitet mit Vakuumröhren. Diese verzerren bei etwas intensiverer Gain Reduction.

2) FET Kompressor: arbeitet mit Field Effect Transistoren. Diese zeichnen sich durch ihre schnellen Attack-Zeiten aus. FET Kompressoren werden oft für die Kompression von Stimmen verwendet.

3) Optical Kompressor: arbeitet mit einem von einer Lampe abhängigen Resistor.

4) VCA Kompressor: Voltage Controlled Amplifier regelt die Laustärke proportional zu einer Spannungskontrolle im Kompressor.

3.4 Sound Design in der Postproduktion

In der Postproduktion von Sprachaufnahmen überscheiden sich zwei Gebiete. Einerseits das Gebiet der Restauration der Audioaufnahme und anderseits das Gebiet des Sound Designs. Der Begriff der Restauration bezieht sich in dieser Arbeit auf jene Tätigkeiten, welche eine Aufnahme von Fehlern befreien sollen. Der Begriff Sound Design hingegen auf die aktive und bewusste Gestaltung einer Klangästhetik. Bei den zuvor erwähnten Effekten für die Audiobearbeitung existieren unter anderem Effekte, welche dem zu bearbeitenden Material eine bestimmte Klangfarbe verleihen. Bei der Kreation einer bestimmten Klangfarbe spielen Equalizer und vor allem die Kompressoren eine Rolle. Beerends & Beerends (2015, 176) postulieren die Wichtigkeit einer „warmen Stimme“, da diese eine höhere Qualität für die Konsumenten ausstrahlt.

3.4.1 Die Rolle der Kompressoren 1176 und LA2A für das Sound Design

In diesem Kapitel wird die Relevanz des FET-Kompressors 1176 und des Optical Kompressors LA-2A als klangfärbende Effekte für das Sound Design von Sprachaufnahmen diskutiert.
Zuerst ist zu erwähnen, dass viele Produzenten und Konsumenten von der Färbung der Kompressoren positiv emotional angesprochen werden. Die emotionale Reaktion besteht aufgrund der Vertrautheit der Artefakte mit den Konsumenten [5].

[5] https://www.musicexpo.co/post/la-2a-vs-1176-and-compression-explained; OFF THE ONE LLC. abgerufen am 10.10.2023

Abbildung 19: 1176 Kompressor Plugin (Hersteller: Waves)

In Abbildung 19 ist der Kompressor 1176 als Plugin dargestellt. Die Möglichkeiten zur Parametrisierung des 1176 sind sehr beschränkt. Wie oben dargestellt kann bei diesem Kompressor nur Ratio, Attack und Release verändert werden. Wie stark der Kompressor das Audiosignal komprimiert, hängt vom Pegel des Eingangssignals ab. Um wie viel Dezibel das Signal reduziert wird, ist aus der Anzeige abzulesen.
Derr 1176 wird in der Sprachbearbeitung unter anderem oft als Kompressor für die Färbung von Sprachaufnahmen benutzt. Bei ausreichender Kompression tritt die typische Verzerrung des 1176 ein und verursacht diese typische Färbung. (vgl. Moore und Wakefield 2017, 1)

A LA-2A is a really popular compressor from recording studios all the way to radio. It was really popular in radio. A lot of radio personalities would typically go through an LA-2A type compressor when they speak [6].

[6] https://www.musicexpo.co/post/la-2a-vs-1176-and-compression-explained; OFF THE ONE LLC.; abgerufen am 10.10.202

Abbildung 20: LA2A Kompressor Plugin (Hersteller Waves)

Der in Abbildung 20 dargestellte Kompressor CLA-2A ist ein Opto-Kompressor und somit langsamer als der 1176 FET-Kompressor. Der LA2A hat noch weniger Möglichkeiten zur Parametrisierung. Die Kompression des Audiosignals wird mit dem Pegel des Eingangssignals gesteuert. Die Kompression wird in der Anzeige in Dezibel abgelesen. Je lauter das Eingangssignal, desto stärker die Kompression.

Die Kompressoren 1176 und LA2A werden von Postproduzenten oft in Serie verwendet. Für diese Serienschaltung wird zuerst der schnelle Kompressor 1176 und danach der langsamere Kompressor LA2A verwendet. Postproduzenten sprechen nach dem Einsatz der Kompressoren von einem polierten und emotional ansprechenden Endergebnis. (sinngemäß wiedergegeben und übersetzt [7]).

Hörbeispiel 1: *unkomprimiert.wav*

Hörbeispiel 2: *komprimiert.wav* (1176: Ratio 4:1; Attack 7; Release 3; Gain Reduction bei beiden Kompressoren maximal ca. 5 dB)

[7] https://mixanalog.com/products/vocals762a; Distopik d.o.o., Tobačna ulica 5, 1000 Ljubljana, Slovenia; abgerufen am 15.10.2023.

3.4.3 Die Rolle des Equalizers für das Sound Design

In diesem Kapitel wird die Relevanz von Equalizer als klangfärbender Effekt für das Sound Design von Sprachaufnahmen diskutiert. Die Schlussfolgerung aus dem Kapitel 3.1 ist, dass Equalizer grundsätzlich ein klangneutrales Ergebnis herstellen und von sich selbst aus das Audiomaterial nicht färben. Das Sound Design kann somit hauptsächlich mit der Manipulation der Lautstärke verschiedene Frequenzbereiche vorgenommen werden. Das Erreichen einer „warmen Radiostimme“ muss somit manuell mit der Adjustierung der zur Verfügung stehenden Parameter des Equalizers erfolgen.

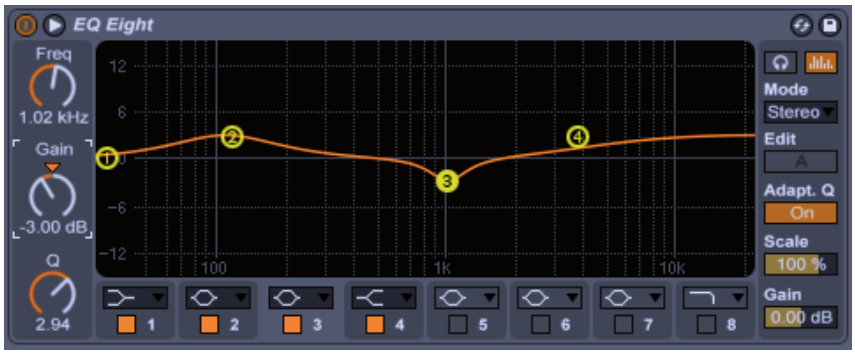

Abbildung 21: Standard-Einstellung eines parametrischen Equalizers für das Sound Design einer „warmen Radiostimme“ (100 Hz +3 dB Q1, 1000 Hz -3 dB Q3, 4000 Hz +3 dB).

Die Abbildung 21 dargestellte Einstellung kann als Ausgangspunkt für die Herstellung einer „warmen Radiostimme“ herangezogen werden. Jede Stimme besitzt einen individuellen Charakter und somit müssen auch die Einstellungen des Equalizers entsprechend an die Stimme angepasst werden. In Abhängigkeit der Klangfarbe der Stimme und der Qualität der Aufnahme kann der gewünschte „warme“ Effekt in manchen Fällen nur mehr schwer erreicht werden. Lineare Verzerrungen, welche durch den neuen Produktionsweg entstanden sind können mit einem Equalizer gut bearbeitet werden – dynamisch abhängige Verzerrungen hingegen nicht.

3.5 Postproduktion von Sprachaufnahmen mit künstlicher Intelligenz

Machines that do 'deep learning' — like LANDR — are a form of advanced artificial intelligence (A.I.). They deal with large and complex sets of data. They're capable of high levels of abstract understanding. They adapt. They learn how to learn (Thorau 2022)

Dieses Zitat stammt vom Gründer und Produktdesigner auf machine learning basierten Tools LANDR[8]. Dieses Tool ist für die Postproduktion von Audioaufnahmen entwickelt worden. LANDR spezialisiert sich auf die Postproduktion bzw. das Mastering von musikalischen Werken. Das KI-basierte Bearbeitungstool Auphonic[9] hingegen spezialisiert sich auf die Postproduktion und Restauration von Sprachaufnahmen.

3.5.1 Funktionsweise von künstlicher Intelligenz in der Postproduktion

Das Prinzip beim machine learning in der Audiopostproduktion funktioniert mit einem Vergleich. Künstliche Intelligenz vergleicht ungemasterte und gemasterte Versionen von Musikstücken von Mastering Ingenieure und lernt daraus, welche Veränderungen vorgenommen wurden. Die Veränderungen in diesem Prozessschritt sind jedoch nicht gleich und unterscheiden sich von Musikstück zu Musikstück. (vgl. *Birtchnell und Elliot* 2018, 77)
Diese Daten werden einer machine zum Lernen bereitgestellt. Diese wendet die Erkenntnisse an und erweitert diese mit neuen Audiostücken ständig.

[8] www.landr.com
[9] www.auphonic.com

3.5.2 Postproduktion von Sprachaufnahmen mit Auphonic

Das Tool Auphonic behandelt alle Schritte, welche Teil einer Postproduktion für Sprachaufnahmen sind. Alle Veränderungen werden somit automatisiert von einer künstlichen Intelligenz vorgenommen. Die zu bearbeitende Spur wird über die Website in das System hochgeladen und nachfolgend bearbeitet wieder bereitgestellt.

Die für die Postproduktion von Sprachaufnahmen relevanten Funktionen sind:

1) Noise & Reverb Reduction: Geräuschreduktion
2) Filtering & Auto EQ: Korrekturen bei den Frequenzen
3) Intelligent Leveler: Lautstärkenanpassung
4) Cut Filler Words and Silence: Füllwörter und Stille werden entfernt
5) Multitrack Alogrithms: Anpassen der Pegel zwischen mehreren Spuren

Auphonic ist eine all-in-one Lösung für Sprachaufnahmen. Mit dieser kann eine komplett unbearbeitete Sprachaufnahme einer kompletten Postproduktion unterzogen werden. Für die Bearbeitung einer schon geschnittenen einkanaligen Sprachaufnahme werden für die Postproduktion nur Nachhall und Störgeräusche (1) reduziert und Anpassungen am Frequenzband (2) vorgenommen. Diese zwei wesentlichen Funktionen werden nachfolgenden kurz näher beschrieben.

1) Entfernung der Störgeräusche:
Diese Funktion trennt den Anteil der Sprache von Hintergrundgeräuschen und Raumhall.

2) Entzerrung:
Diese Funktion nimmt Korrekturen am Frequenzband vor.

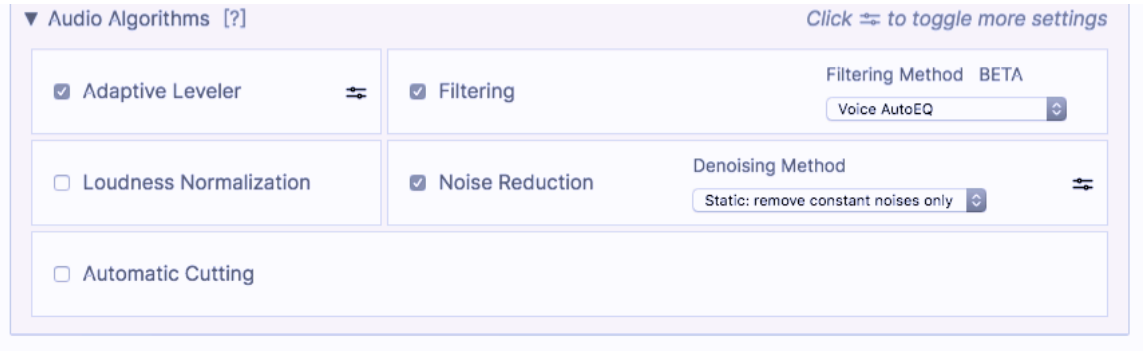

Abbildung 22: Parametrisierung von Auphonic für die Postproduktion in Kapitel 4

Wie in Abbildung 22 dargestellt, wird jede Sprachaufnahme mit diesen Einstellungen einer automatischen Postproduktion unterzogen. Die restlichen Funktionen von Auphonic werden in dieser Arbeit nicht beachtet, da die für die Postproduktion bereitgestellten Aufnahmen diesen Korrekturen nicht unterzogen werden müssen.

Die Digitalisierung und die daraus resultierende zeitliche und örtliche Trennung der einzelnen Schritte in der Produktion können bei falscher Durchführung Artefakte verursachen, welche korrigiert werden müssen. Im nächsten Kapitel werden typische Aufnahmen, welche über diesen „neuen Produktionsweg“ entstanden sind restauriert und einem Sound Design unterzogen. Zuerst soll versucht werden, nur mit konventionellen Signalverarbeitungsalgorithmen die gesamte Postproduktion durchzuführen. Dazu zählen alle konventionellen Signalverarbeitungsalgorithmen - ausgenommen sind KI-basierte Methoden. Danach sollen KI- und datenbasierte Tools in die Restauration und das Sound Design integriert werden. Im nachfolgenden Kapitel werden die Ergebnisse der manuellen Postproduktion mit dem Ergebnis von Auphonic verglichen und Unterschiede herausgearbeitet.

4 Postproduktion von Sprachaufnahmen

In diesem Kapitel werden 2 Sprachaufnahmen einer Postproduktion unterzogen. Alle Sprachaufnahmen weisen aufgrund ihres Produktionsweges (siehe Abbildung 1) verschiedenste Probleme bezüglich ihrer Qualität auf. Diese Probleme werden in diesem Kapitel dargestellt, erklärt und versucht gelöst zu werden. Für die Visualisierung der Frequenzgänge wird das Tool Matlab [10] benutzt.

Der Ablauf aller Experimente ist wie folgt aufgebaut:

1) Vergleich einer guten Aufnahme mit der schlechten Aufnahme; auch Ciesla (2023, 137) spricht davon, gute Referenzen für die Bearbeitung des Materials zu verwenden, um die Mängel der Sprachaufnahme überhaupt feststellen zu können.

    a) Audioaufnahmen werden beide auf -16 LUFS gesetzt
    b) beide Aufnahmen werden mittels einer Spektralanalyse dargestellt
    c) Beschreibung der Probleme (Ergebnis aus einer Höranalyse und einer technischen Analysen)
2) Vergleich zwischen einer automatisierten Postproduktion von Auphonic und dem unbearbeiteten Material
3) Vergleich der Postproduktion mit Auphonic mit einer manuellen Post-Produktion
4) Einbindung von daten- und KI-basierten Tools für die Postproduktion

Im ersten Punkt werden die offensichtlichen Mängel der zu bearbeitenden Aufnahme sichtbar gemacht und beschrieben. Im zweiten Punkt werden die Ergebnisse der automatisierten Postproduktion im Vergleich mit dem Ausgansmaterial visuell dargestellt. Im dritten Punkt werden die Ergebnisse der vollautomatisierten Postproduktion mit einer manuell durchgeführten Postproduktion miteinander verglichen.

[10] www.matlab.com (Matlab-Skript vom IEM–Universität für Musik und Kunst Graz, 2024)

Abschließend werden im Punkt 4 noch bestehende Mängel aus Punkt 3 dargestellt und dafür Lösungen mit daten- und KI-basierte Tools gesucht und mögliche Vorteile resultierend aus dieser Integration gefunden werden.

4.1 Postproduktion Nr. 1

Die Aufnahme *Sprachaufnahme_zwei.wav* wurde im Zuge einer Produktion für die Podcast-Serie Blaulichthelden [11] aufgenommen. Die Aufnahme wurde nicht im Studio sondern via Zoom und mit einem Headset-Mikrofon hergestellt – also über den zuvor beschriebenen neuen Produktionsweg.

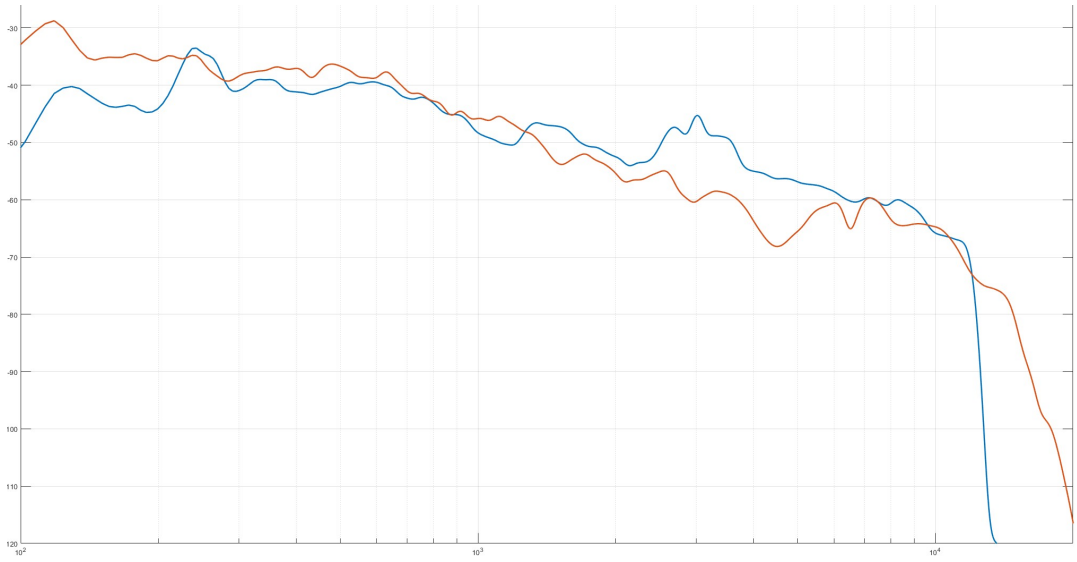

Abbildung 23: Gemittelter Frequenzgang Vergleich 1 - über die gesamte Referenzaufnahme *Sprachaufnahme_eins.wav* (orange) und der zu bearbeitenden Aufnahme *Sprachaufnahme_zwei.wav* (blau) mit logarithmierter Amplituden – und Frequenzachse (relative Amplitudenwerte in dB und Frequenz in Hertz). Darstellung: halbtonbreite (1/12 Oktave) Frequenzglättung. (Betrachtungsbereich: 100 Hz bis 20 kHz, -120 dB bis -30 dB)

Die *Sprachaufnahme_eins.wav* weist nach Basis einer Höranalyse keine hörbaren und auffälligen Artefakte auf. Auch die Darstellung des Frequenzgangs beschreibt in dieser Form keine Auffälligkeiten.

Hörbeispiel: *Sprachaufnahme_eins.wav*

[11] www.blaulichthelden.at

zu bearbeitende Aufnahme: *Sprachaufnahme_zwei.wav*

4.1.1 Analyse der Aufnahme *Sprachaufnahme_zwei*

in den Höhen keine Sprachinformation mehr hörbar
die Frequenzen im Bassbereich sind unterrepräsentiert
Bei 00:05:000 s und 00:14:500 s ist eine Resonanzfrequenz im Zischlaut „sch“ hörbar
Bei 00:19:975 s ist ein Plosiv hörbar
Artefakte zwischen 2000 Hz und 10000 Hz hörbar: dauerhaft, klingt verzerrt - hier liegt eventuell ein Enkodierungs-, Dekodierungs– bzw. Übertragungsfehler vor und wird im Hörtest als „Bröseln“ wahrgenommen
Die Frequenzen im Mittenbereich sind überrepräsentiert
„Knacksen“ bei 00:24:450 s und 00:28:190 s hörbar

a) Qualität der Aufnahme aus technischer Perspektive

Die technischen Limitationen, welche für den Low-Pass-Filter ab circa 12 kHz verantwortlich, sind können an dieser Stelle nicht mehr mit Sicherheit festgestellt werden. Vermutlich ist für die Bandbegrenzung die schlechte Qualität des Mikrofons oder das Übertragungsmedium Zoom verantwortlich. Das Übertragungsmedium ist mit hoher Wahrscheinlichkeit auch für die Artefakte in den oberen Mitten und in den Höhen verantwortlich. Neben dem Übertragungsmedium und dem Mikrofon kann für diese Degradierungen („Bröseln“) in der Signalkette die Qualität des Analog-Digital Converters verantwortlich sein (siehe Abbildung 1).

b) Qualität der Sprache

Durch die technischen Limitationen wirkt die Sprache in mittleren Frequenzen überrepräsentiert. Auditiv können Resonanzfrequenzen bei der Aussprache des Frikativs „sch“ und einmal ein Plosiv festgestellt werden.

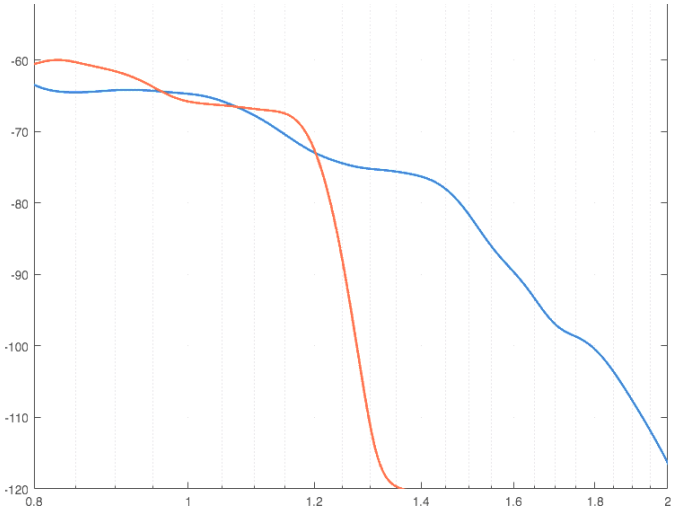

Abbildung 24: Frequenzgang über die gesamte Referenzaufnahme *Sprachaufnahme_eins.wav* und der *Sprachaufnahme_zwei.wav* - mit logarithmierter Amplituden – und Frequenzachse (relative Amplitudenwerte in dB und Frequenz in Hertz). halbtonbreite (1/12 Oktave) Frequenzglättung. (Betrachtungsbereich: 8000 Hz bis 20000 Hz, -120 dB bis -50 dB)

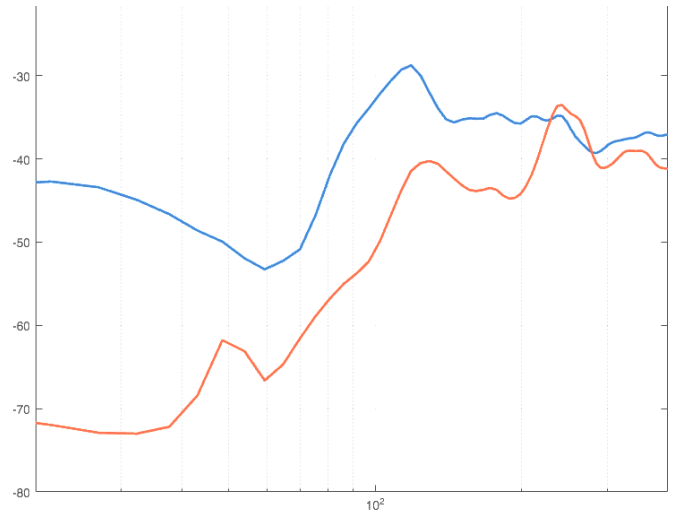

Abbildung 25: Gemittelter Frequenzgang über die gesamte Referenzaufnahme *Sprachaufnahme_eins.wav* (orange) und der *Sprachaufnahme_zwei.wav* (blau) -mit logarithmierter Amplituden – und Frequenzachse (relative Amplitudenwerte in dB und Frequenz in Hertz). halbtonbreite (1/12 Oktave) Frequenzglättung. (Betrachtungsbereich: 100 Hz bis 400 Hz, -80 dB bis -30 dB)

Wie auf der Seite zuvor im Hörversuch beschrieben, sind die Frequenzen im Bassbereich unterrepräsentiert und in den Höhen ab 12 kHz nicht mehr vorhanden. Diese Annahme wird auch mit einer Analyse in Matlab im Vergleich mit der Referenzaufnahme bestätigt.

4.1.2 Vergleich zwischen der Postproduktion mit Auphonic und der unbearbeiteten *Sprachaufnahme_zwei*

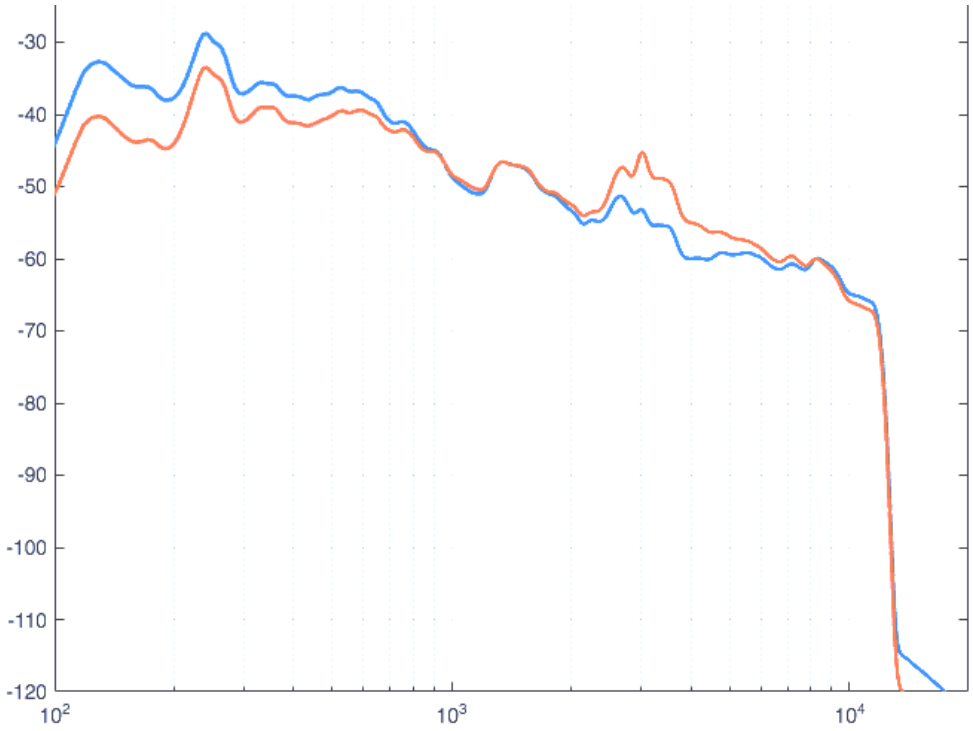

Abbildung 26: Frequenzgang über die gesamte unbearbeitete *Sprachaufnahme_zwei.wav* (orange) und der *Sprachaufnahme_zwei_Auphonic.wav* (blau) - mit logarithmierter Amplituden – und Frequenzachse (relative Amplitudenwerte in dB und Frequenz in Hertz). halbtonbreite (1/12 Oktave) Frequenzglättung. (Betrachtungsbereich: 100 Hz bis 20 kHz,-120 dB bis-30 dB)

Die Veränderungen der automatisierten Postproduktion von Auphonic reduzieren sich lediglich auf chirurgische Anpassungen einzelner Frequenzbereiche. Das Programm unternimmt nur leichte Eingriffe in die Ästhetik der Sprachaufnahme. Die Zischlaute und die Plosive werden nicht korrigiert und auch die Artefakte zwischen 6000Hz und 10000Hz bleiben bestehen. Das Defizit im unter 200Hz wird ebenfalls nur leicht angepasst.

Hörbeispiel: *Sprachaufnahme_zwei_Auphonic.wav*

4.1.3 Manuelle Postproduktion der Sprachaufnahme.zwei

Die manuelle Postproduktion wird in Ableton durchgeführt. Zum Einsatz kommen in diesem Schritt nur digitale Effekte von Ableton.

a) Equalizer

Zuerst werden auf Basis der Erkenntnisse von 4.1.1 die nachfolgenden Korrekturen anhand zweier Achtband-Equalizer durchgeführt. Diese Korrekturen resultieren aus den Defiziten, welche in Abbildung 23 dargestellt werden.

1) 11714 Hz: -1dB, Q: 8
2) 8355 Hz: -3 dB, Q: 8
3) 6589 Hz: -8 dB, Q: 19
4) 4479 Hz: -10 dB, Q: 4
5) 3015 Hz: -14 dB, Q: 6
6) 1206Hz: +4 dB, Q: 10
7) 431 Hz: +4 dB, Q: 4
8) 150 Hz: +10 dB, Q: 0.6
9) 100Hz: +3db, Q:0.71

Auf Basis der Höranalyse werden zudem die Resonanzfrequenzen des Frikativs „sch“ vermindert.

1) 3010 Hz:-15dB, Q:10
2) 2930 Hz: -6 dB: Q:14
3) 2700 Hz: -15 dB, Q:10
4) 2600 Hz: -6 dB, Q: 14
5) 2500Hz: --12dB, Q:16

Hörbeispiel: *Sprachaufnahme_zwei_manuell_Equalizer.wav*

An dieser Stelle werden noch zusätzliche ästhetische Eingriffe gemacht, welche auf einer subjektiven Beurteilung beruhen.

1) 700 Hz, 4 dB, Q:1
2) 3740 Hz, -3dB, Q: 3

b) Korrekturen durch Schnitte

Die Lautstärke des Plosivs wurde von mir manuell mittels einer Automation verringert. Auch das „Knacksen“ musste entfernt werden, da die vorhandenen Tools diese nicht entfernen konnten. Das Filtern dieses Artefaktes hätte zu einer zu starken Beeinträchtigung der restlichen Sprachaufnahme geführt.

c) Kompression:

Wie am Anfang in Punkt 4 besprochen, gilt es eine möglichst „warme“ Radiostimme zu erzeugen. Dies erfordert unter anderem eine adäquate Kompression der Aufnahme, um diesen Zustand zu erreichen. Die Gestaltung der Kompression für diese Sprachaufnahme ist wie folgt aufgebaut:

1) Kompressor mit schneller Attack und schneller Release, um die Transienten der Aufnahme zu kontrollieren.

Ratio: 6:1
Attack: 6ms
Release: 40 ms
Soft-Knee: 12ms
Lookahead = 1ms
maximal 3 dB Gain Reduction

2) Kompressor mit langsamer Attack und langsamer Release für die Lautheit der Aufnahme:

Ratio: 8:1
Attack: 40ms
Release: 200 ms

Soft-Knee: 12ms

Lookahead = 1ms

Maximal 3 dB Gain Reduction

Hörbeispiel: *Sprachaufnahme_zwei_manuell.wav*

4.1.4 Einbindung von datenbasierte Tools für die Bearbeitung der *Sprachaufnahme.zwei*

Das Problem der hörbaren Artefakte zwischen 2000Hz und 10000Hz konnte nicht gelöst werden. Diese Fehler sind mit einem Equalizer nicht zu beseitigen, da die störenden Frequenzen in deren Gesamtheit nicht mehr feststellbar sind. Für die Lösung dieses Problems werden in diesem Kapitel zwei datenbasierte adaptive Tools von Waves [12] für die Restauration dieses Fehlers verwendet:

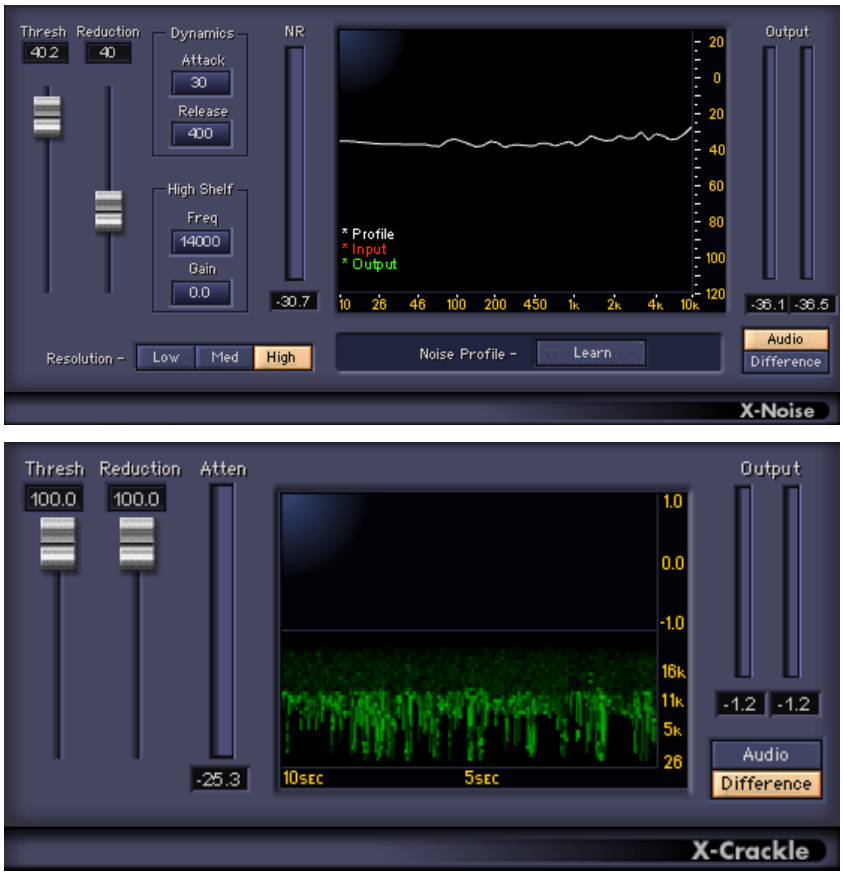

Abbildung 27: Screenshot von X-Noise und X-Crackle (Hersteller Waves)

[12] www.waves.com

a) X-Crackle

Dieses Plugin arbeitet auf Basis eines „psycho-acoustic noise reduction algorithms“. Laut Hersteller können damit Störsignale wie „crackles“ und „pops“ in einem Audiosignal gefunden werden. Dazu zählen zum Beispiel Artefakte, welche durch beschädigte Schallplatten ausgelöst werden können. Mit diesem Plugin können diese störenden Artefakte (in diesem Fall zwischen 2 kHz – 10 kHz) festgestellt und hörbar gemacht werden.

- Der Fader „Treshhold“ bestimmt die Amplitude des Störsignals, welches entfernt werden sollen. *Getroffene Einstellung: 100*

- Der Fader „Reduction“ bestimmt die Verringerung der Lautstärke der Störsignale. *Getroffene Einstellung: 100*

Mit der Funktion „Difference“ kann das zu entfernende Störsignal isoliert von dem restlichen harmonischen Signal wiedergegeben werden.

Hörbeispiel: *Sprachaufnahme_zwei_Störsignal.wav*

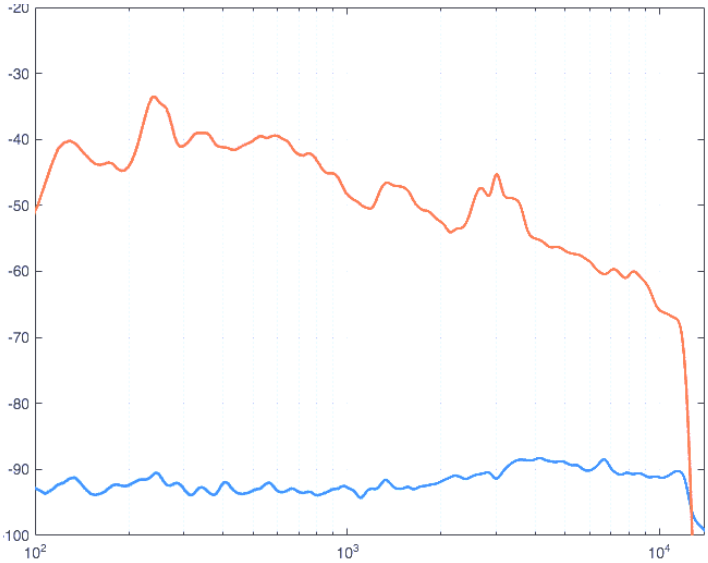

Abbildung 28: Frequenzgang über die gesamte unbearbeitete *Sprachaufnahme_zwei.wav* (orange) und der *Sprachaufnahme_zwei_Störsignal* (blau) - mit logarithmierter Amplituden – und Frequenzachse (relative Amplitudenwerte in dB und Frequenz in Hertz). halbtonbreite (1/12 Oktave) Frequenzglättung. (Betrachtungsbereich: 100Hz bis 20 kHz, -100 dB bis -20 dB)

Mit dem Plugin X-Crackle kann nun das Störsignal identifiziert und isoliert dargestellt werden. Wie in Abbildung 28 in blau dargestellt, ist das isolierte Störsignal zwischen 2 kHz bis 10 kHz am lautesten. In Abbildung 28 kann zudem das Laustärkenverhältnis zwischen dem isolierten Störsignal und der unbearbeiteten *Sprachaufnahme_zwei.wav* abgelesen werden. Der große Dynamikunterschied erklärt die hochgradigen Einstellungen der verwendeten Audioeffekte.

b) X-Noise

Dieses Plugin arbeitet auf Basis eines „Fingerprint"-Systems. Mit der Funktion „learn" erlernt das Plugin das Profil eines wiedergegebenen Signals. Nachdem das Plugin das Störsignal erfasst hat, kann das Störsignal mehr oder weniger aus dem Signal entfernt werden.

- Der Fader „Treshold" bestimmt die Amplitude des Störsignals, welches entfernt werden sollen. *Getroffene Einstellungen: 40.2*

- Der Fader „Reduction" bestimmt die Verringerung der Lautstärke des erfassten Störsignals. *Getroffene Einstellung: 40*

Das isolierte Störsignal aus dem Plugin X-Crackle kann nun mit dem Plugin X-Noise erlernt werden. Dieses Signal kann in Abbildung 27 im Plugin X-Noise abgelesen werden und bestätigt den dargestellten Amplitudenverlauf in der Abbildung 28.

Finales Hörbeispiel: *Sprachaufnahme_zwei_final.wav* (alle Schritte kombiniert)

4.1.5 Fazit aus der Postproduktion Nr. 1

Die fehlenden Frequenzen des Sprachinhalts ab ca. 10 kHz konnten in der Restauration nicht angehoben werden. Diese sind in der Ausgangsdatei nicht vorhanden und können deshalb auch nicht restauriert werden. Die Frequenzen im Bassbereich sind zwar vorhanden jedoch unterrepräsentiert. Mit einem Equalizer konnten die Frequenzen in diesem Bereich angehoben werden. Die Resonanzfrequenzen des Frikativs „sch“ konnten mit einem Equalizer eingedämmt werden. Um die störenden Frequenzbereiche besser identifizieren zu können, wurde punktuell im oberen Bereich des Frequenzbandes Frequenzbereiche lauter gemacht. Auch die Frequenzen im mittleren Bereich wurden mit diesem Prinzip identifiziert und danach entsprechend verringert. Die Artefakte, welche vermutlich durch das Übertragungsmedium Zoom entstanden sind konnten nur mit den datenbasierten Plugins X-Crackle und X-Noise eingedämmt werden. Das Plosiv bei 00:19:975 s und das Knacksen bei bei 00:24:450 s und 00:28:190 s konnte in letzte Instanz nur durch einen Schnitt entfernt werden.

Bei der Bearbeitung der Störlaute wurde insbesondere klar, dass sich ab einem gewissen Grad der Bearbeitung eines Artefaktes der Prozessschritt selbst negativ auf den Rest der Sprachaufnahme auswirken kann. Rein theoretisch wäre das Entfernen des Störgeräusches wie z.B. das „Knacksen“ ohne Schnitt bis zu einem Grad schon möglich, jedoch nur dann, wenn die restlichen Audioinformationen nicht zu sehr in Mitleidenschaft gezogen werden. Im Zuge der Postproduktion war es möglich, alle erkennbare Artefakte in irgendeiner Form zu bearbeiten und die gesamte Sprachaufnahme im Vergleich zu Ausgangssituation zu verbessern. Keines der erkannten Artefakte wurde unbearbeitet belassen.

4.2 Postproduktion Nr. 2

Die Aufnahme *Sprachaufnahme_vier.wav* wurde ebenfalls im Zuge einer Produktion für die Podcast-Serie „Blaulichthelden“ aufgenommen. Die Aufnahme wurde nicht im Studio sondern via Zoom mit einem im Laptop (keine Informationen vorhanden) eingebauten Mikrofon hergestellt.

Referenzaufnahme: *Sprachaufnahme_drei.wav*

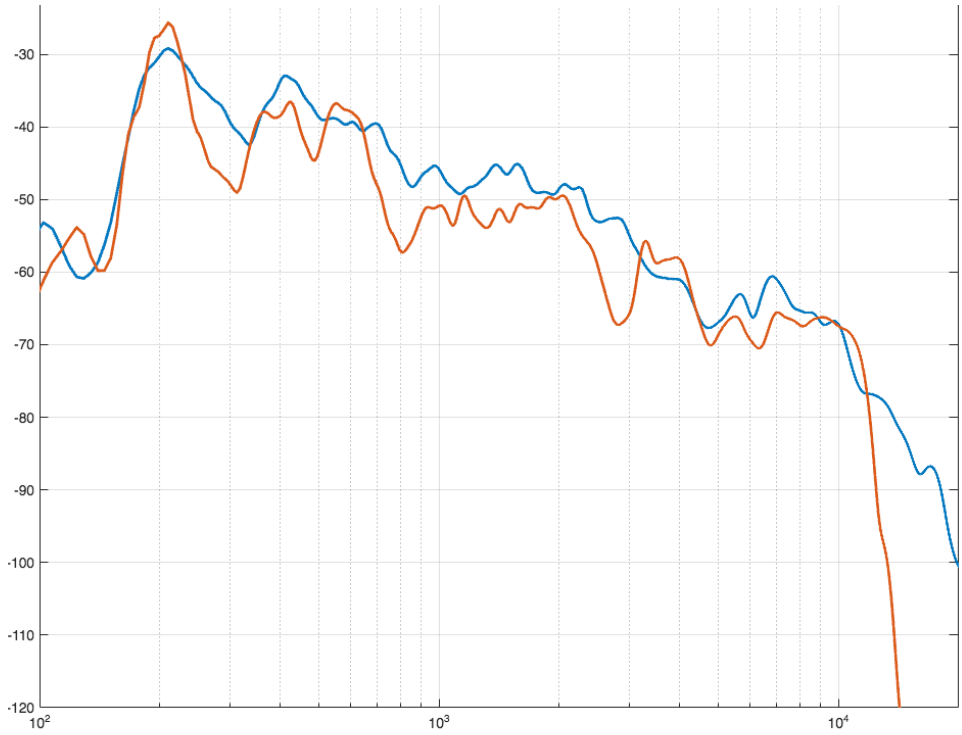

Abbildung 29: Gemittelter Frequenzgang über die gesamte Referenzaufnahme *Sprachaufnahme_drei.wav* (blau) und der zu bearbeitenden Aufnahme *Sprachaufnahme_vier.wav* (orange) - mit logarithmierter Amplituden – und Frequenzachse (relative Amplitudenwerte in dB und Frequenz in Hertz). Darstellung: halbtonbreite (1/12 Oktave) Frequenzglättung. (Betrachtungsbereich: 100Hz bis 20 kHz, -120 dB bis -30 dB)

zu bearbeitende Aufnahme: *Sprachaufnahme_vier.wav*

4.2.1 Analyse der Aufnahme *Sprachaufnahme_vier*

Raumresonanz in den unteren Mitten hörbar
Bei 00:03:750 s und bei 00:12:892 s ist eine Resonanzfrequenz im Frikativ „sch“ hörbar
Zwischen 3000 Hz und 5000 Hz ist eine Resonanzfrequenz hörbar (feststellbar bei 00:08:145)
Bei 00:01:260 s ist eine Verzerrung hörbar
Artefakte („Knistern“) in den höheren Frequenzen hörbar
Low-Pass-Filter hörbar (aber nur sehr leicht)

a) Qualität der Aufnahme aus technischer Perspektive

Der Aufnahme zur Folge wurde das Interview in einem größeren nicht schallbehandelten Raum durchgeführt. Dies führt zu einer durch den Raum ausgelösten Resonanzfrequenz in den unteren Mitten im Frequenzspektrum. Zudem sind Artefakte wie das „Knistern“ und eine kurze von der Sprachinformation unabhängige Verzerrung hörbar. Diese Artefakte haben mit großer Wahrscheinlichkeit wieder einen Zusammenhang mit der Aufnahmeart über das Übertragungsmedium Zoom.

b) Qualität der Sprache

Die weibliche Person hat auf Basis des Hörtests eine hellere Stimme. Zudem können Resonanzfrequenzen bei der Aussprache des Frikativs „sch“ festgestellt werden. Weitere Auffälligkeiten, welche die Sprache betreffen wurde nicht festgestellt.

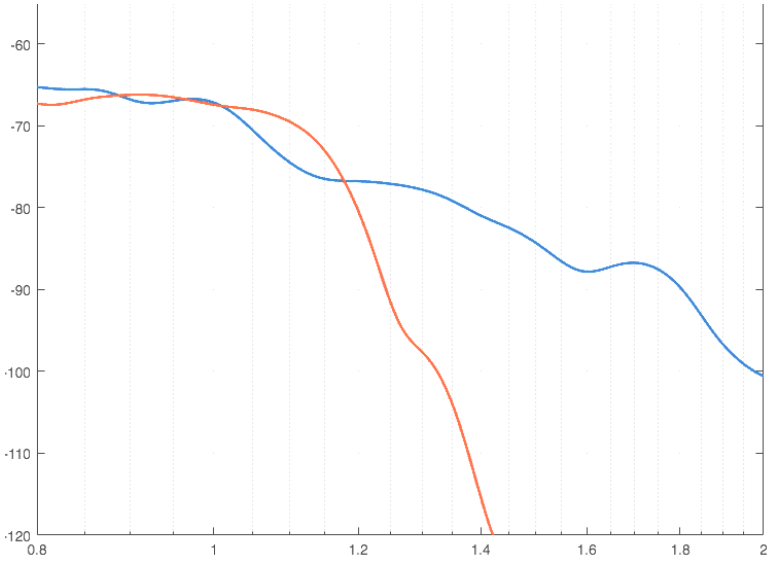

Abbildung 30: Gemittelter Frequenzgang über die gesamte Referenzaufnahme *Sprachaufnahme_drei.wav* (blau) und der *Sprachaufnahme_vier.wav* (orange) -mit logarithmierter Amplituden – und Frequenzachse (relative Amplitudenwerte in dB und Frequenz in Hertz). halbtonbreite (1/12 Oktave) Frequenzglättung. (Betrachtungsbereich: 8000 Hz bis 20000 Hz, -120 dB bis -50 dB)

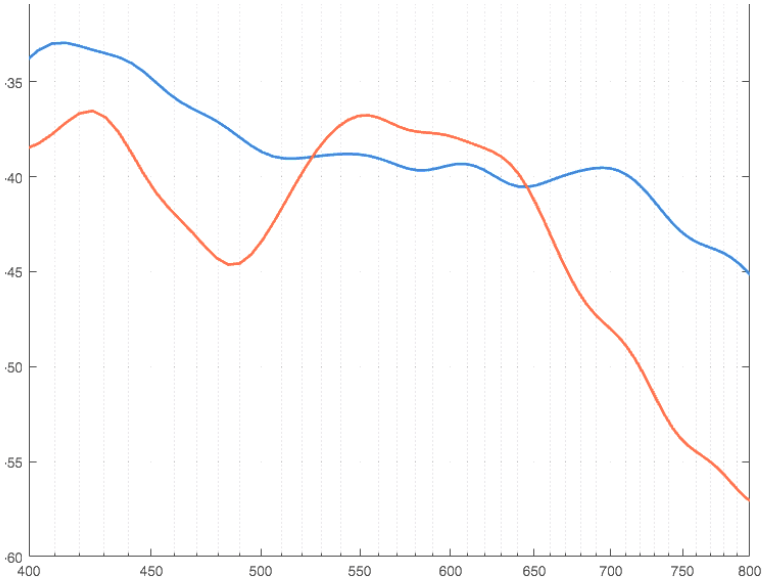

Abbildung 31: Gemittelter Frequenzgang über die gesamte Referenzaufnahme *Sprachaufnahme_drei.wav* (blau) und der *Sprachaufnahme_vier.wav* (orange) -mit logarithmierter Amplituden – und Frequenzachse (relative Amplitudenwerte in dB und Frequenz in Hertz). halbtonbreite (1/12 Oktave) Frequenzglättung. (Betrachtungsbereich: 400 Hz bis 800 Hz, -60 dB bis -20 dB)

Der gewählte Ausschnitt in Abbildung 30 hat den zuvor vermuteten Low-Pass Filter in diesem Bereich bestätigt. In dem in Abbildung 31 ausgewählten Bereich wird zusätzlich die Resonanz des Raumes vermutet.

4.2.2 Vergleich zwischen der Postproduktion mit Auphonic und der unbearbeiteten *Sprachaufnahme_vier*

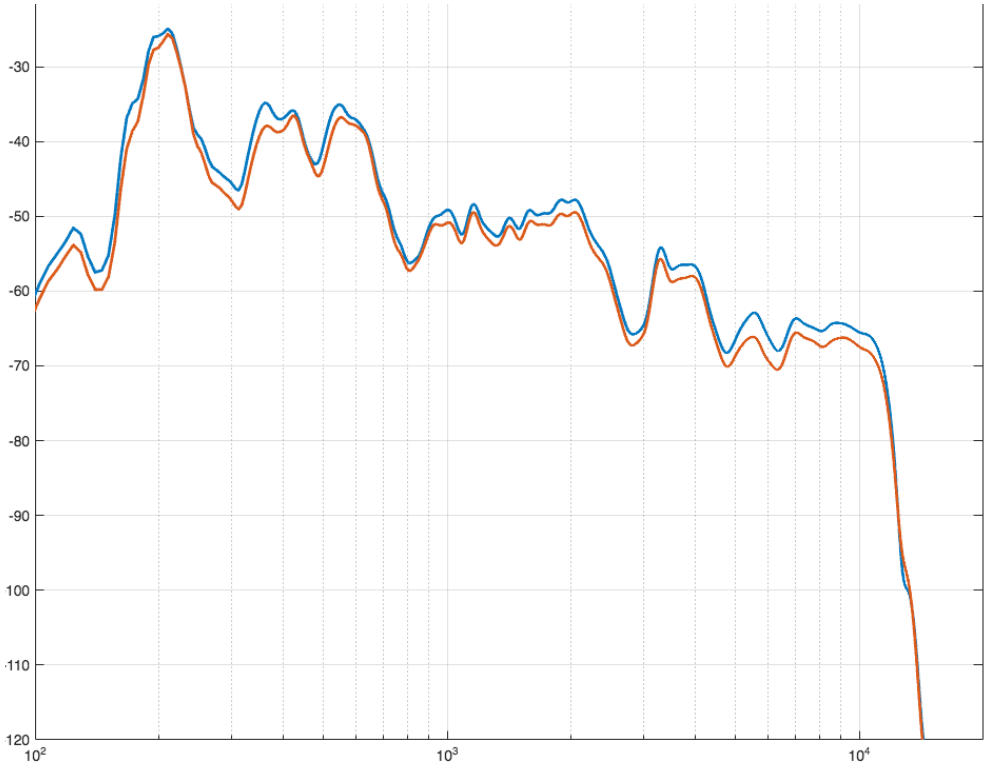

Abbildung 32: Frequenzgang über die gesamte unbearbeitete *Sprachaufnahme_vier.wav* (blau) und der *Sprachaufnahme_vier_Auphonic.wav* (orange) - mit logarithmierter Amplituden – und Frequenzachse (relative Amplitudenwerte in dB und Frequenz in Hertz). Halbtonbreite (1/12 Oktave) Frequenzglättung, Betrachtungsbereich: 100Hz bis 20 kHz, -120 dB bis -30 dB)

Die Veränderungen der automatisierten Postproduktion von Auphonic reduzieren sich wieder auf leichte Anpassungen über den gesamten Frequenzverlauf hinweg. Auf Basis des Hörtests bleiben die Artefakte („Knistern“) in den höheren Frequenzen hörbar. Auch die Resonanzfrequenzen des Frikativs „sch“ bleiben für das persönliche Hörempfinden unangenehm und zu laut. Im Vergleich zwischen unbearbeiteter Sprachaufnahme und bearbeiteter Aufnahme hat die bearbeitete von Auphonic im Hörtest einen geringeren Anteil des Raumhalls. Hier hat das Programm das Problem im mittleren Frequenzbereich erkannt und diese zum Teil hörbar bereinigt. Für die nachfolgende manuelle Postproduktion besteht vor allem das Problem den Raumhall und das „Knistern“ aus der Aufnahme zu entfernen.

Hörbeispiel: *Sprachaufnahme_vier_Auphonic.wav*

4.2.3 Manuelle Postproduktion der *Sprachaufnahme_vier*

Die manuelle Postproduktion wird in Ableton durchgeführt. Zum Einsatz kommen in diesem Schritt nur digitale Effekte von Ableton.

1) Equalizer:

Auf Basis der Analyse wird zuerst mit Hilfe eines Achtband-Equalizers die Resonanzfrequenz in den unteren Mitten beseitigt. Diese Einstellungen dienen dazu, die entstehenden Resonanzen verursacht durch den Raum zu entfernen.

1) 180 Hz: 6dB, Q: 2
2) 400 Hz: 6 dB, Q: 4
3) 683 Hz: 6 dB, Q: 4
4) 600 Hz: 6 dB, Q: 1
5) 194 Hz: 3dB, Q: 0.71

Anschließend werden mit einem nachfolgenden Achtband-Equalizer die beobachteten Resonanzfrequenzen der Frikative „sch“ und „s“ entfernt.

1) 850 Hz: 3dB, Q: 3
2) 3570 Hz: 3 dB, Q: 3
3) 3340 Hz: 10 dB, Q: 10

Die Resonanzfrequenzen im Frikativ „s“ konnten mittels eines Equalizers nicht zur Gänze entfernt werden. Anschließend soll dafür eine Lösung mit einem De-Esser gefunden werden. **Hörbeispiel:** *Sprachaufnahme_vier_manuell.wav*

2) De-Esser:

Für die zusätzliche Eindämmung des Frikativs „s“ wird De-Esser verwendet. Der De-Esser wurde auf den Frequenzbereich von 6000 Hz bis 20000 Hz eingestellt. Danach wird der Treshold (in Abbildung 33 - Processing) so lange adjustiert, bis die übrig gebliebenen Frequenzen für das Hörempfinden angenehm und natürlich empfunden werden.

Abbildung 33: De—Esser. Hersteller: Techivation

Hörbeispiel: *Sprachaufnahme_vier_Equalizer_De_Esser.wav*

4.2.4 Einbindung von intelligenten Tools für die Bearbeitung der *Sprachaufnahme_vier*

Trotz Eindämmung der Frequenzen, welche für die Raumresonanz verantwortlich sind ist immer noch eine gewisser Anteil an Hall hörbar. Die Eingriffe mit dem Equalizer sind nicht ausreichend. Deshalb wird für die weitere Bearbeitung das Plugin Clarity-Vx [14] des Hersteller Waves verwendet. Das Plugin trennt die Stimme (Direktschall) vom Raumhall (Diffusschall). Das Plugin beschreibt sich selbst als Plugin, welches auf Basis eines „Neural Networks®“ arbeitet[15]. Dieses Netzwerk analysiert Millionen von Beispielen, vergleicht diese mit der zu bearbeitenden Aufnahme, um dann Annahmen für diese zu treffen. Diese Annahmen können mit den Parametern adaptiert werden.

[13] www.techivation.com

[14] www.waves.com

[15] https://assets.wavescdn.com/pdf/plugins/clarity-vx-pro-v2.pdf

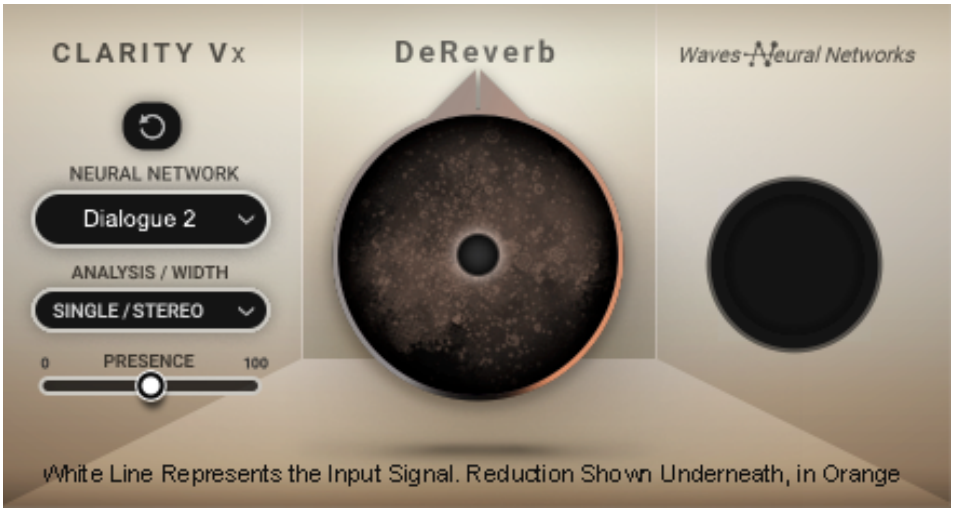

Abbildung 34: Clarity Vx für die Entfernung von Raumhall. (Hersteller Waves)

In Abbildung 34 sind die getroffenen Einstellungen für die *Sprachaufnahme_vier* dargestellt. Der Regler in der Mitte wurde solange adjustiert, bis der Raumhall hörbar verschwunden ist. Bei der Einstellung der Dämpfung wurde synchron die Qualität der Sprache berücksichtigt, sodass eine zu hochgradige Einstellung der Dämpfung nicht zu einer Verschlechterung der Sprachqualität führt. Hier wurde bei 49 % ein Kompromiss für diese Aufnahme gefunden.

Hörbeispiel: *Sprachaufnahme_vier_ohne_Reverb.wav*

Nach der Eindämmung des Raumhalls bleiben hörbare Artefakte, welche als „Knistern“ wahrgenommen werden in der Aufnahme übrig. Für die Restauration dieser, wird in diesem Schritt wieder auf zwei datenbasierte Plugins des Herstellers Waves zurückgegriffen. Dazu gehören die Plugins X-Crackle und X-Noise.

Abbildung 35: oben: X-Crackle von Waves. unten: X-Noise von Waves

Das Entfernen des Störsignals erfolgt bei dieser Sprachaufnahme mit dem gleichen Prinzip wie bei *Sprachaufnhame_zwei.wav*. Zuerst wird mit dem Plugin X-Crackle das Störgeräusch isoliert vom Sprachinhalt dargestellt und mit der Learn-Funktion mit dem X-Noise Plugin gelernt und danach entfernt. Die Einstellungen können aus der Abbildung 35 abgelesen werden. Zudem kann in beiden Plugins das Profil des Störsignals abgelesen werden. In der nachfolgenden Abbildung wird das unbearbeitete Signal in Kombination mit dem Störsignal noch einmal dargestellt.

Hörbeispiel: *Sprachaufnahme_vier_Störsignal.wav*

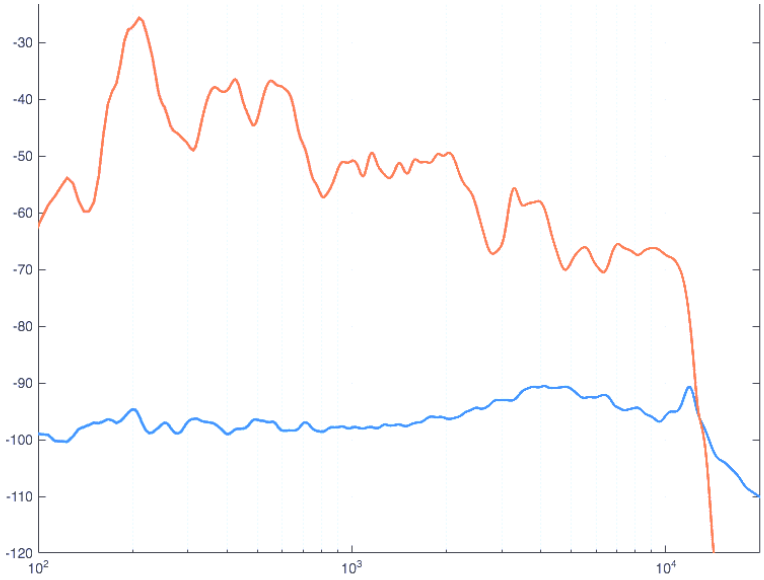

Abbildung 36: Frequenzgang über die gesamte unbearbeitete *Sprachaufnahme_vier.wav* (orange) und der *Sprachaufnahme_vier_Störsignal* (blau) - mit logarithmierter Amplituden – und Frequenzachse (relative Amplitudenwerte in dB und Frequenz in Hertz). halbtonbreite (1/12 Oktave) Frequenzglättung, Betrachtungsbereich: 100Hz bis 20 kHz, -120 dB bis -30 dB

In Abbildung 36 ist der Betragsfrequenzgang der unbearbeiteten Sprachaufnahme und des Störsignals dargestellt. Die lautesten Stellen des Störsignals befinden sich zwischen 3 kHz bis 5 kHz und um 12 kHz. Diese konnten mit dem zuvor beschriebenen Prozess nahezu unhörbar gemacht werden.

Hörbeispiel: *Sprachaufnhame_vier_AI.wav* (alle Schritte kombiniert)

Trotz dieser Berücksichtigung entstand aus den vorherigen Bearbeitungsschritten eine störende Resonanzfrequenz, welche mit einem Achtband-Equalizer mit den Einstellungen

1) 240 Hz, -6dB, Q4

korrigiert wurde. Diese Resonanzfrequenz konnte zum Schluss nur mehr begrenzt eingedämmt - aber nicht völlig entfernt werden.

4.2.5 Fazit aus der Postproduktion Nr. 2

Die im Hörtest wahrgenommene Resonanzfrequenz im Frikativ „sch“ konnte mit Equalizer auch diesmal gut restauriert werden. Trotzdem wurde zusätzlich für die Eindämmung der Frequenzen im Frikativ „s“ in diesem Prozessschritt ein De-Esser verwendet. Auch die Raumresonanz konnte mit Hilfe eines Equalizers verringert werden. Jedoch blieb der Raumhall trotz dieses Prozessschrittes hörbar, weshalb zusätzlich das KI-basierte Plugin Claritx-Vx für die weitere Verringerung herangezogen werden musste. Auch das im Hörtest beschriebene „Knistern“ konnte mit einem herkömmlichen Equalizer nicht restauriert werden. Dazu waren zwei datenbasierte Plugins, X-Crackle und X-Noise nötig. Die kurze Verzerrung, welche unabhängig von Sprachinformationen entstanden ist, konnte nicht entfernt werden. Die Versuche dieses Artefakt mit Equalizer und einem KI-basierten Filter zu entfernen, sind aufgrund zu starker Auswirkungen auf die gesamte Qualität der Aufnahme gescheitert. Dieser Filter hätte relevante Frequenzen für die Abbildung der Sprache zu stark reduziert. Im Gegensatz zur Postproduktion Nr. 1 befindet sich für Demonstrationszwecke dieses Artefakt noch in der Aufnahme.

Wie auch bei der Postproduktion Nr. 1 muss bei dieser Postproduktion wieder darauf geachtet werden, dass durch den Vorgang andere Aspekte der Aufnahme nicht verschlechtert werden. Jeder Prozessschritt, welcher ein explizites Problem behandelt, darf auch hier wieder nicht isoliert von allen anderen Eigenschaften der Sprachaufnahme betrachtet werden. Jeder Schritt kann durch dessen Parametrisierung der Effekte eine negative Auswirkung auf die anderen Eigenschaften haben. Wie auch zuvor sind durch die Postproduktion Artefakte nur verringert beziehungsweise nahezu unhörbar gemacht – jedoch nicht vollständig entfernt worden.

5. Befragung

In den vorangegangenen Kapiteln wurde das Ziel eine ansprechende Sprachaufnahme zu erzeugen diskutiert und mit Grafiken unterstützend erklärt. Die gewonnenen theoretischen Erkenntnisse wurden an zwei Sprachaufnahmen praktisch umgesetzt. Die übergeordnete Aufgabe war es, die Umsetzbarkeit einer emotional ansprechenden und technisch möglichst sauberen Sprachaufnahme trotz Artefakte, verursacht durch neue Produktionswege zu prüfen. Wie schon zuvor erklärt versteht man in dieser Arbeit unter dem neuen Produktionsweg die Produktion von Sprachaufnahmen über das Internet und dessen Übertragungsmedien. In diesem Kapitel wird wahrgenommene Qualität der Sprachaufnahmen geprüft. Der erste Versuch betrifft die Sprachaufnahme einer männlichen Person und die zweite Aufnahme einer weiblichen Person.

5.1 Aufbau der Umfrage

Der Test selbst besteht aus zwei Teilen. Im ersten Teil wird die wahrgenommene Qualität der bestehenden Aufnahme einer männlichen Stimme ohne Nachbearbeitung und drei alternativer Nachbearbeitungen abgefragt. Im zweiten Schritt die wahrgenommene Qualität einer weiblichen Stimme ohne Nachbearbeitung und drei alternativer Nachbearbeitungen. Für den gesamten Versuchsverlauf ist ein Kopfhörer zu verwenden. Alle Sprachaufnahmen besitzen eine Lautstärke von -16 LUFS. Den Probanden werden keine Qualitätsstandards erklärt. Die Probanden entscheiden somit nach ihren eigenen Kriterien was als qualitativ besser oder minderwertiger im Vergleich eingeschätzt wird. Die gesamte Darstellung der Befragung befindet sich im Appendix.

5.2 Befragungsart und Auswertung

Die Befragung findet mittels einer kontinuierlichen Beurteilungsskala statt. Die Probanden haben die Möglichkeit die Sprachaufnahmen von 0 (minderwertig) bis 100 (hochwertig) zu bewerten. Im Fragebogen wird die zentrale Frage:

Wie bewerten Sie die wahrgenommene Qualität der Sprachaufnahmen?

gestellt.

Für die Auswertung werden zuerst die Bewertungsdifferenzen pro Versuchsperson ermittelt. Da die unbearbeitete Aufnahme die Referenzaufnahme ist, wird die Bewertung von allen anderen Beurteilungen einer Versuchsperson subtrahiert. Daraus resultierend werden die Mittelwerte und die Streuwerte berechnet (siehe Abbildung 37 und 38). Basierend auf diesen Werten werden anschließend die paarweisen Unterschiede mittels t-Test mit Irrtumswahrscheinlichkeit von 5% für die jeweiligen Nachbearbeitungsmethoden berechnet. Für die Korrektur der p-Werte wird die Bonferroni-Dunn Methode angewandt. Die Bewertungstrends der einzelnen Versuchspersonen bezüglich der verschiedenen Nachbearbeitungsmethoden sind im Appendix dargestellt.

Frage 1:

Zu bewertende Aufnahmen
1. *Sprachaufnahme_zwei_Auphonic.wav*
2. *Sprachaufnahme_zwei_manuell.wav*
3. *Sprachaufnahme_zwei_final.wav*
4. *Sprachaufnhame_zwei.wav*

Frage 2:

Zu bewertende Aufnahmen*:
1. *Sprachaufnahme_vier_Auphonic.wav*
2. *Sprachaufnahme_vier_Equalizer_ De_esser.wav*
3. *Sprachaufnahme_vier_AI.wav*
4. *Sprachaufnhame_vier_unbearbeitet.wav*

5.3 Auswertung der Befragung

Die statistische Auswertung samt nachfolgender Darstellung wird mit dem Programm Prism [16] durchgeführt.

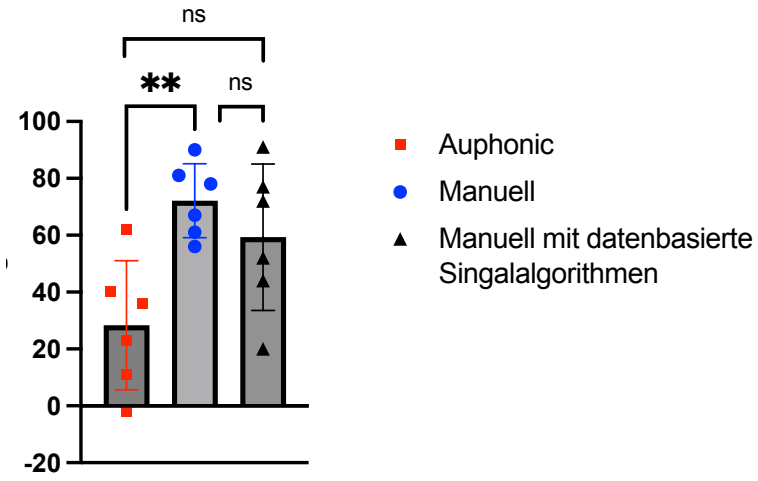

Abbildung 37: Balkendiagram der Ergebnisse der *Sprachaufnahme_zwei* (rot - *Sprachaufnahme_zwei_Auphonic.wav, blau - Sprachaufnahme_zwei_manuell.wav,* schwarz - *Sprachaufnhame_zwei_final.wav,* Balken - Mittelwert, Whisker (farbig) – Standardabweichung, *** - signifikant, ns- nicht signifikant)*

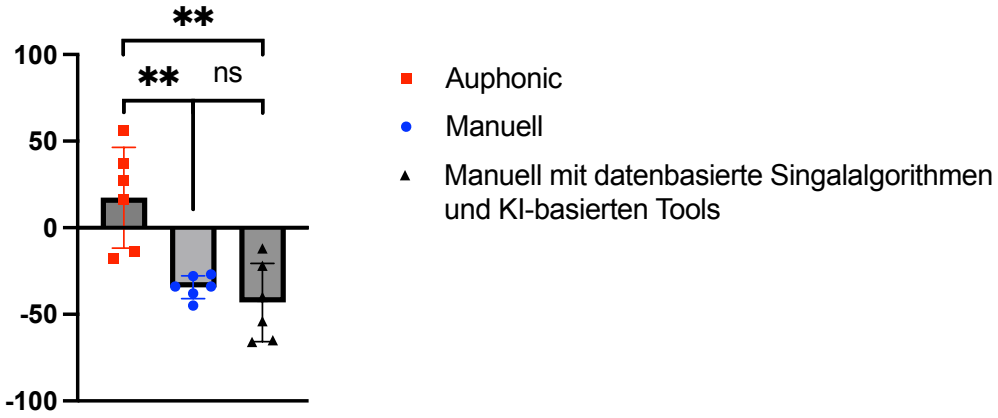

Abbildung 38: Balkendiagram der Ergebnisse der *Sprachaufnahme_vier* (rot - *Sprachaufnahme_vier_Auphonic.wav, blau - Sprachaufnahme_vier_Equalizer_De_esser.wav,* schwarz - *Sprachaufnahme_vier_AI.wav,* Balken - Mittelwert, Whisker (farbig) – Standardabweichung, *** - signifikant, ns- nicht signifikant)*

[16] www.graphpad.com

Information zu den Befragungen:

Anzahl der befragten Personen: 6 (2 – männlich, 4 – weiblich)

Altersregime: 24 - 64 Jahre

Wie in Abbildung 37 dargestellt, wurde die manuell bearbeitete Sprachaufnahme ohne datenbasierte und adaptive Signalverarbeitungsalgorithmen im Mittel mit der besten Qualität bewertet. Die Sprachaufnahme mit datenbasierten und adaptiven Signalverarbeitungsalgorithmen wurde von den Probanden nicht signifikant schlechter beurteilt. Die bearbeitete Aufnahme durch Auphonic (KI-basiert) wird als signifikant schlechter wahrgenommen. Wie in Abbildung 38 dargestellt, wurden die manuell bearbeiteten Sprachaufnahmen signifikant schlechter beurteilt als die bearbeitete Aufnahme durch Auphonic.

5.4 Interpretation der Ergebnisse

Bei der ersten Postproduktion wurden die bearbeiteten Aufnahmen als qualitativ höherwertiger beurteilt. Hierbei ist zu erwähnen, dass dem Bassbereich besonderes Augenmerk gewidmet wurde. Im Vergleich zur Aufnahme von Auphonic ist dieser in den bearbeiteten Beispielen viel präsenter. Der Unterschied zwischen den biesen beiden Aufnahmen ist auch signifikant. Die Probanden konnten jedoch keinen signifikanten Unterschied mehr zwischen dem manuellen Ergebnis und dem Ergebnis mit den datenbasierten Verarbeitungsalgorithmen feststellen.

Im Gegensatz zur ersten Postproduktion beurteilten die Probanden die bearbeiteten Sprachaufnahmen im Vergleich zu den unbearbeiteten schlechter. Die unbearbeitete Aufnahme wies zuerst eine Resonanzfrequenz in den unteren Mitten auf. Diese wurde im Zuge der Postproduktion entfernt. Dadurch wirken die bearbeiteten Aufnahmen „dünner“. Hier könnte unter anderem auch ein Bias durch die Interviewführung vorliegen. Die Annahme, dass Sprachaufnahmen mit einem präsenteren Bassbereich qualitativ höherwertiger ist, könnte nach der ersten auch für die zweite Versuchsreihe von den Probanden angenommen worden sein.

6. Conclusio

6.1 Zusammenfassung und Beantwortung der Forschungsfragen

Aufgabe dieser Arbeit war es typische Artefakte bei einer Produktion von Sprachaufnahmen über neue Produktionswege darzustellen, diese zu beschreiben und mit herkömmlichen Signalverarbeitungsalgorithmen und KI-basierter Tools zu restaurieren. Ziel war es, in der Postproduktion die Erreichbarkeit einer qualitativ hochwertigen Aufnahme zu prüfen. Das Wesentliche in dieser Arbeit war, zuerst anhand von Literatur die Relevanz der technischen Grundlagen im Kontext der neuen Produktionswege darzustellen, um daraus objektive Qualitätsstandards für Sprachaufnahmen abzuleiten. Dies bildete das Fundament für die anschließende Postproduktion. Im letzten Kapitel wurden diese Sprachaufnahmen mit Referenzaufnahmen verglichen und wahrgenommene Qualität von Probanden erfragt. Die im Kapitel 1.2 gestellten Forschungsfragen werden in diesem Kapitel unter Berücksichtigung der vorangegangenen Literaturdiskussion, der Postproduktion und der Ergebnisse der Befragung beantwortet.

Aufnahmen sind grundsätzlich etwas sehr dynamisches. Abhängig von der Stimmfarbe kann es in der Postproduktion zu erheblichen Einschränkungen der noch zu erreichenden Qualität der Sprachaufnahme führen. Zuerst ist zu erwähnen, dass Sprachkrankheiten keine Erwähnung finden, weil in dieser Arbeit grundsätzlich davon ausgegangen wird, dass diese bei der Aufnahme nicht vorliegen. Die Beurteilung der Qualität lässt sich in zwei Bereiche einteilen. Zuerst wird die Qualität der Sprache beurteilt. Im Fokus steht hier die Klangfarbe der Sprache. Je nach Zusammensetzung kann die aufgenommene Sprache in manchen Frequenzbereichen über- beziehungsweise unterrepräsentiert sein. Diese Auffälligkeiten werden im Kontext mit guten Referenzaufnahmen schnell hörbar und auch sichtbar. Für die Beantwortung der ersten Forschungsfrage sind jedoch die Auswirkungen der neuen Produktionswege wichtig.

Suboptimale Aufnahmesituationen und minderwertige Übertragungsmedien können ein Ungleichgewicht zwischen verschiedenen Frequenzbereichen hervorrufen. Je nach Schweregrad können diese nur bedingt restauriert werden. Die zweite Ebene betrifft die technischen Eigenschaften einer Sprachaufnahme. Die neuen Produktionswege können Artefakte und Verluste von Informationen hervorrufen, welche überhaupt nicht mehr zu beheben sind. Dazu zählen zum Beispiel Low Pass-Filter und High Pass-Filter und andere Artefakte ausgelöst durch die Komprimierung der Aufnahme. Die Beantwortung der ersten Forschungsfrage ist im Kontext dieser neuen Produktionswege keine eindeutige. Eine restaurierte Aufnahme frei von Artefakten kann in den meisten Fällen nicht als erreichbarer Zustand festgelegt werden. Ziel kann somit nur mehr das Erreichen einer Aufnahme sein, in welcher die Artefakte mehr oder minder kein großes Hindernis für den Konsum der Aufnahme darstellen. Somit kann man die möglichst angenehme Konsumierbarkeit des Resultats als wichtigstes Anforderungskriterium für diese Sprachaufnahmen festlegen.

Die Sprachaufnahmen im praktischen Teil machen das zuvor gesagte hörbar. Aufgrund der neuen Produktionswege sind auch neue Artefakte entstanden, welche zuvor nicht unbedingt vorgekommen sind. Das einleitente Zitat beschreibt den Idealfall einer Sprachproduktion. In diesem Fall waren die Studios so ausgelegt, dass es grundsätzlich zu einer ausbalancierten und artefaktfreien Aufnahme gekommen ist. Durch die Produktion über verschiedenste Übertragungsmedien ist diese Situation nicht mehr gegeben. Dadurch entstehen Aufnahmen welche mit den bestehenden Effekten nicht mehr zu restaurieren sind. Fehler wie zu laute oder zu leise Frequenzbereiche können mit herkömmlichen Mitteln sehr gut adaptiert werden. Auch große Dynamikunterschiede stellen kein Problem in der Postproduktion dar. Kritischer ist es jedoch, wenn Fehler wie zum Beispiel Verzerrungen, Raumhall und andere Kodierungsfehler entstehen. Bei der Postproduktion und Restauration wurde ganz klar ersichtlich das mit herkömmlichen Audioeffekten wie Equalizer, Kompressoren und Filter sehr gut Dynamikunterschiede und Anpassungen im Frequenzbereich vorgenommen werden konnten. All jene Fehler, welche jedoch destruktiv für die Audioinformation selbst sind konnten nur mit Hilfe von artifizieller Intelligenz eingedämmt werden.

Somit kann für die Beantwortung der zweiten Forschungsfrage gesagt werden, dass Fehler nie ganz behoben werden können. Bei der Postproduktion und Restauration muss mit bedacht vorgegangen werden, da zu viele oder unüberlegte Eingriffe der Sprachaufnahme zusätzlich schaden könnten. Keine Postproduktion und Restauration kann eine gute Aufnahme von Anfang an ersetzen.

Technische Fehler wie Verzerrungen, Raumhall und andere Kodierungsfehler konnten mit künstlicher Intelligenz eingedämmt werden. Mit einem Equalizer kann ich einzelne Frequenzen oder Bereiche im Spektrum anpassen. Im Gegensatz konnten all jene Effekte, welche auf künstlicher Intelligenz basieren die harmonische Informationen von breitbandigen Störsignalen trennen. Diese verwendeten Effekte können diese finden und isoliert von der restlichen Audioinformation darstellen. Das sinnvolle Integrieren sollte Schritt für Schritt passieren. Zuerst sollte eine vollständige Analyse der Audiodatei gemacht werden. Dort sollen alle Fehler benannt, beschrieben und kategorisiert werden. Danach sollte diese Schritte der Reihe nach abgearbeitet werden. Eine Integration von künstlicher Intelligenz sollte dann passieren, wenn nach dem Einsatz aller herkömmlichen Effekte (ohne künstlicher Intelligenz) noch offensichtliche Artefakte bestehen und den zuvor definierten angenehmen Konsum verhindern.

Um in dieser Arbeit einen objektiven Vergleich durchführen zu können, wurden die Endergebnisse aller Postproduktionen mit einer vollkommen auf künstlicher Intelligenz basierten Postproduktion verglichen. Hier hat sich herausgestellt, dass die intelligente Postproduktion nur leichte lineare Eingriffe auf das Frequenzband durchgeführt hat. Die manuelle Postproduktion ist detaillierter auf die einzelnen Artefakte eingegangen. Zum Teil hat die intelligente Postproduktion Artefakte übersehen oder nur sehr wage eingedämmt. Der größte Unterschied abseits der Korrektur von groben Fehlern wurde jedoch an der aktiven Klanggestaltung der Sprachaufnahme bemerkt.
Zusammenfassend kann gesagt werden, dass sich die intelligente Postproduktion sich eher auf die Korrektur von Fehlern konzentriert und weniger um die Herstellung eines emotional ansprechenden Produktes.

6.2 Diskussion

Für die Beantwortung der ersten Forschungsfrage wurde das Erreichen einer „konsumierbaren“ Aufnahme als Ziel festgelegt. An dieser Stelle muss jedoch noch das Ziel, eine emotional ansprechende Aufnahme zu erreichen erwähnt werden. Je nachdem welche Artefakte die Aufnahme stören beziehungsweise wie stark die Aufnahme komprimiert ist, können durchaus sehr gute Endresultate erreicht werden. Die unbearbeiteten Aufnahmen weisen sehr oft verschiedene Qualitätsunterschiede auf. Die verschiedenen Formen von Artefakten und Fehler können in den verschiedensten Kombinationen auftreten. Für das Erreichen einer warmen Radiostimme sind die Frequenzen im unteren Frequenzbereich notwendig. Bei stark komprimierten Aufnahmen kann dies deshalb sehr oft nicht mehr erreicht werden. Der generelle Anspruch, eine emotionale Aufnahme bereitzustellen wäre mit den derzeitigen Mitteln deshalb noch ein Schritt zu weit.

An dieser Stelle muss auch noch die Bezeichnung einer Sprachaufnahme als „emotional ansprechend“ diskutiert werden. Ein präsenter Bassbereich zählt aufgrund der Ergebnisse der Umfrage mit Sicherheit zu den Faktoren, welche eine Sprachaufnahme zumindest als qualitativ hochwertig erscheinen lässt. In Summe geht es in der Postproduktion darum, eine warme Sprachaufnahme herzustellen. Durch die synchron laufende Postproduktion in einem Radiostudio wird in das Frequenzband und in den Dynamikumfang einer Stimmaufnahme eingegriffen. Die Einstellungen sind so ausgelegt, sodass die Stimme auf allen möglichen Medien gut hörbar. Dadurch könnte eine gewisse Klangfarbe für die Konsumenten zur Gewohnheit geworden sein. Diese Sprachaufnahmen sind unter anderem üblicherweise sehr stark mit Kompressoren bearbeitet. Dadurch wirken sie sehr bass-lastig. Dieses Phänomen wird meiner Meinung nach daher als allgemein gültiger Qualitätsstandard und als „warm“ empfunden.

6.2 Ausblick

Die Qualität der KI-basierten Anwendungen ist nur eine Momentaufnahme. Diese Technologien entwickeln sich ständig weiter und werden in der Zukunft wahrscheinlich immer raffinierter werden. Da diese Systeme vielfältiger und besser im Einsatz werden, wird es in Zukunft wahrscheinlich auch möglich sein, stark beschädigte Sprachaufnahmen mit einer allumfassenden KI-basierten Anwendung zu restaurieren. Diese wird Fehler komplett beheben und verloren gegangene Informationen vollständig ersetzen können.
Resultierend aus den bisherigen Ergebnissen der manuellen Postproduktionen kann abschließend gesagt werden, dass die Weiterentwicklung der KI-basierten Verfahren noch die aktive Klanggestaltung zur Herstellung eines emotional ansprechenden Produktes vorrangig mitberücksichtigt werden d.h. gelernt werden muss.

7. Quellenverzeichnis

Patka, Kiron. 2015: „Radio als Sound. Von der enträumlichten Stimme zum Radio-Sounddesign.“ *Navigationen - Zeitschrift für Medien- und Kulturwissenschaften* 15, Nr. 2, S. 113–125. DOI: https://doi.org/10.25969/mediarep/1547.

Roederer, Juan G. 2000. *Physikalische und psychoakustische Grundlagen der Musik*. Berlin Heidelberg: Springer. https://doi.org/10.1007/978-3-642-57138-1_1.

Ciesla, Robert. 2022. *Sound and Music for Games: The Basics of Digital Audio for Video Games*. 1.Verlagsort: Apress Berkeley, CA. https://doi.org/10.1007/978-1-4842-8661-6.

Watkinson, John. 2003. *An Introduction to Digital Audio*. 2. Verlagsort: Routledge.

Walden, Robert. 1999. „Analog-to-digital converter survey and analysis.“ I*EEE Journal on Selected Areas in Communications* Vol. 17, Nr. 4 (April), 539-550, doi: 10.1109/49.761034.

Brandenburg, Karlheinz. 1999. „MP3 and AAC explained.“ Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding. Audio Engineering Society, Nr.009 (September), http://www.aes.org/e-lib/browse.cfm?elib=8079.

Fuchs, Robert und Maxwell, Olga. 2016. „The effects of mp3 compression on acoustic measurements of fundamental frequency and pitch range." *Speech prosody* Vol. 2016: 523-527.

Beerends, John, und Beerends, Imre. 2015. "On the assessment of high-quality voice recordings including voice postprocessing." *Journal of the Audio Engineering Society* Vol. 63, Nr. 3: 174-183. https://doi.org/10.17743/jaes.2015.0013.

Seifert, Eberhard. 2020. "Wie funktioniert die Stimme?." *Journal für Gynäkologische Endokrinologie/Schweiz* Vol. 3: 121-224. https://doi.org/10.1007/s41975-020-00164-x.

Ferreira, Aníbal, und Fernandes, Vânia. 2017. "Consistency of the F0, Jitter, Shimmer and HNR voice parameters in GSM and VOIP communication." *2017 22nd International Conference on Digital Signal Processing (DSP)*: 1-5. doi: 10.1109/ICDSP.2017.8096128.

Stassen, Hans H. 1995 . *Affekt und Sprache: Stimm- und Sprachanalysen bei Gesunden, depressiven und schizophrenen Patienten.* Berlin: Springer. https://doi.org/10.1007/978-3-642-79726-2

Dickreiter, Michael., Dittel, Volker., Hoeg, Wolfgang., Wöhr, Martin. 2014. *Handbuch der Tonstudiotechnik*. Germany: De Gruyter Saur.

Hansch, Pierre, und Rentschler, Christian. 2013. *Emotion@Web: Emotionale Websites durch Bewegtbild und Sound-Design*. Germany: Springer Berlin Heidelberg.

Bernstein, Herbert. 2019. *Elektroakustik*. Wiesbaden: Springer Fachmedien Wiesbaden.

Välimäki, Vesa, und Reiss, Joshua D.. 2016. "All About Audio Equalization: Solutions and Frontiers" *Applied Sciences* 6, Nr. 5: 129. https://doi.org/10.3390/app6050129

Moore, Austin, und Jonathan Wakefield. 2017. "An Investigation into the Relationship between the Subjective Descriptor Aggressive and the Universal Audio of the 1176 FET Compressor." *Audio Engineering Society Convention* 142, Nr. 9749 (Mai): 1-10. http://www.aes.org/e-lib/browse.cfm?elib=18625.

Birtchnell, Thomas, und Elliott, Anthony. 2018. "Automating the black art: Creative places for artificial intelligence in audio mastering." *Geoforum* 96 (2018): 77-86.

Thoreau, Guillaume. A.I. Music As Life-Form: Machine Learning Systems and Sound. LANDR BLOG, 17.12.2023. https://blog.landr.com/ai-as-life-form.

Neppert, Joachim. 1999. *Elemente einer Akustischen Phonetik.* Berlin: Buske Helmut Verlag GmbH.

8. Abbildungsverzeichnis

Appendix:

9 Tonträgerverzeichnis

(Dateien auf dem beigefügten Datenträger)

10 Appendix

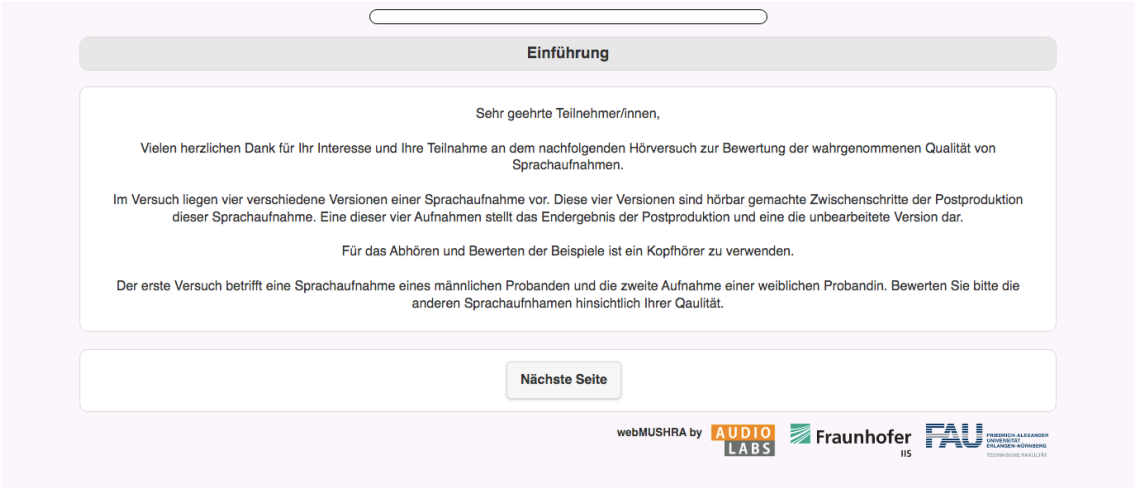

Abbildung 39: Seite 1 der Befragung

Abbildung 40: Seite 2 der Befragung

Abbildung 41: Seite 3 der Befragung

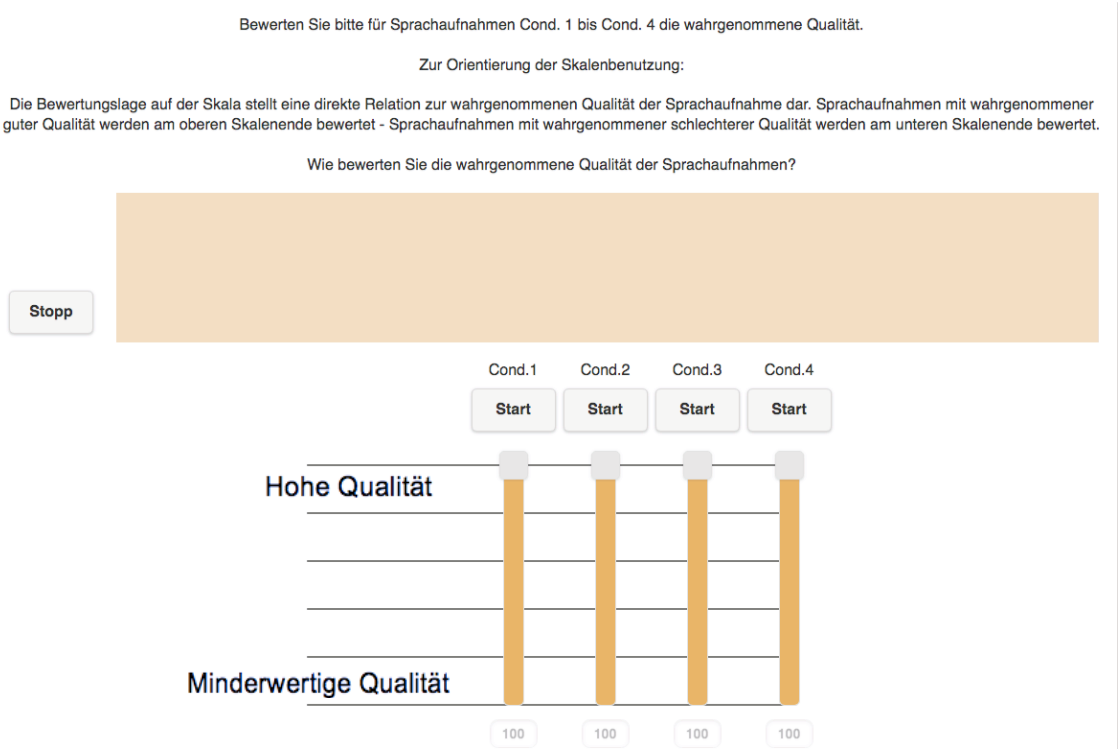

Abbildung 42: Seite 4 der Befragung

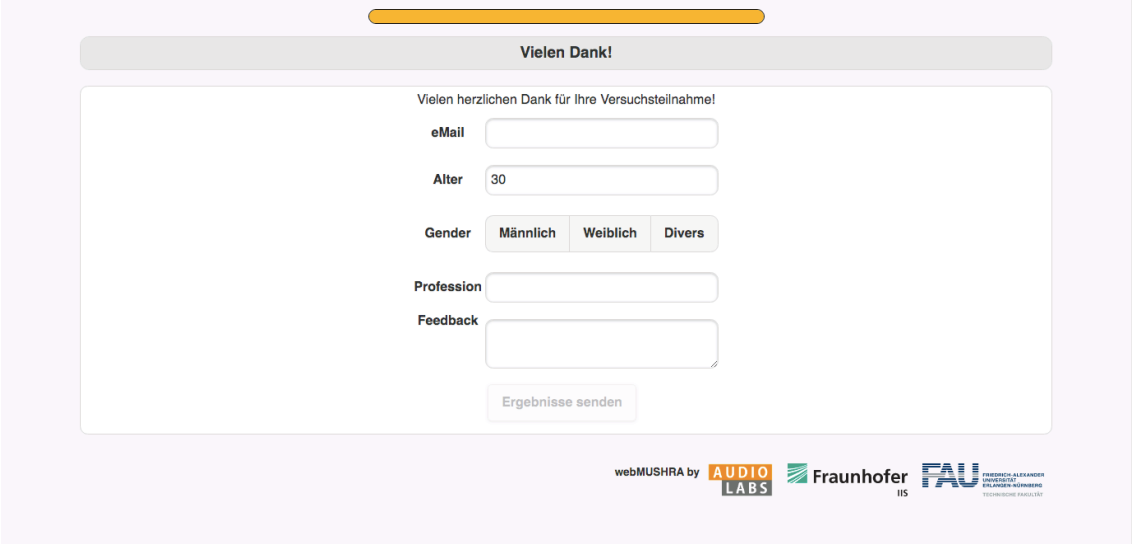

Abbildung 43: Seite 5 der Befragung

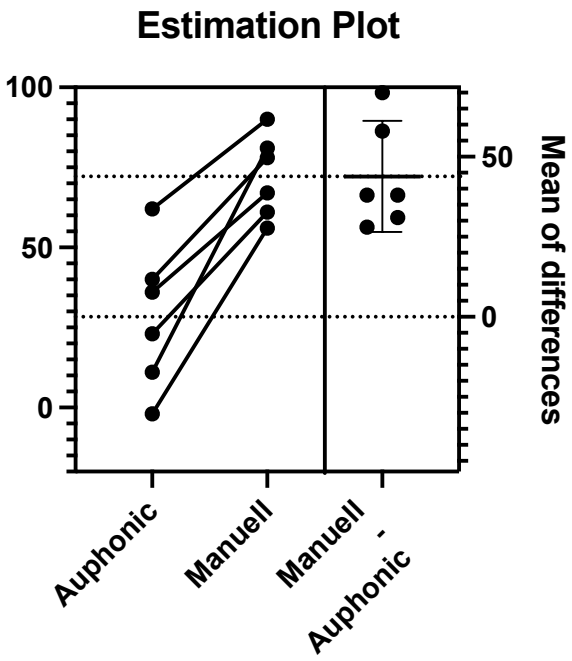

Abbildung 44: Bewertungstrend Auphonic – Manuell (Sprachaufnahme 2)

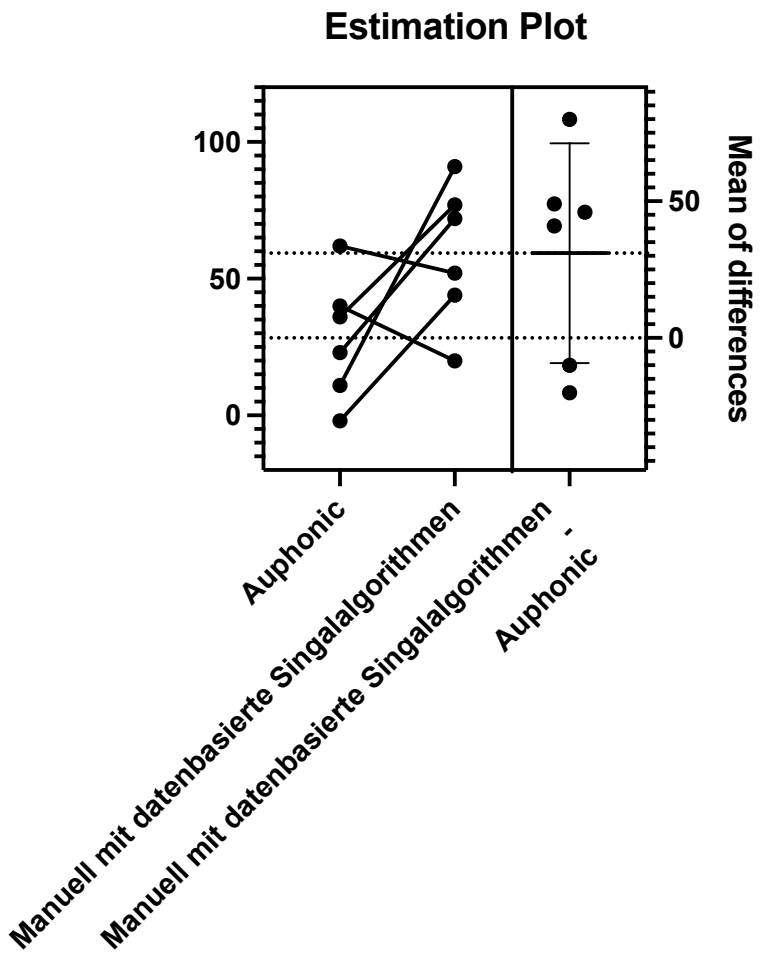

Abbildung 45: Bewertungstrend Auphonic – manuell mit datenbasierte Signalalgorithmen (Sprachaufnahme 2)

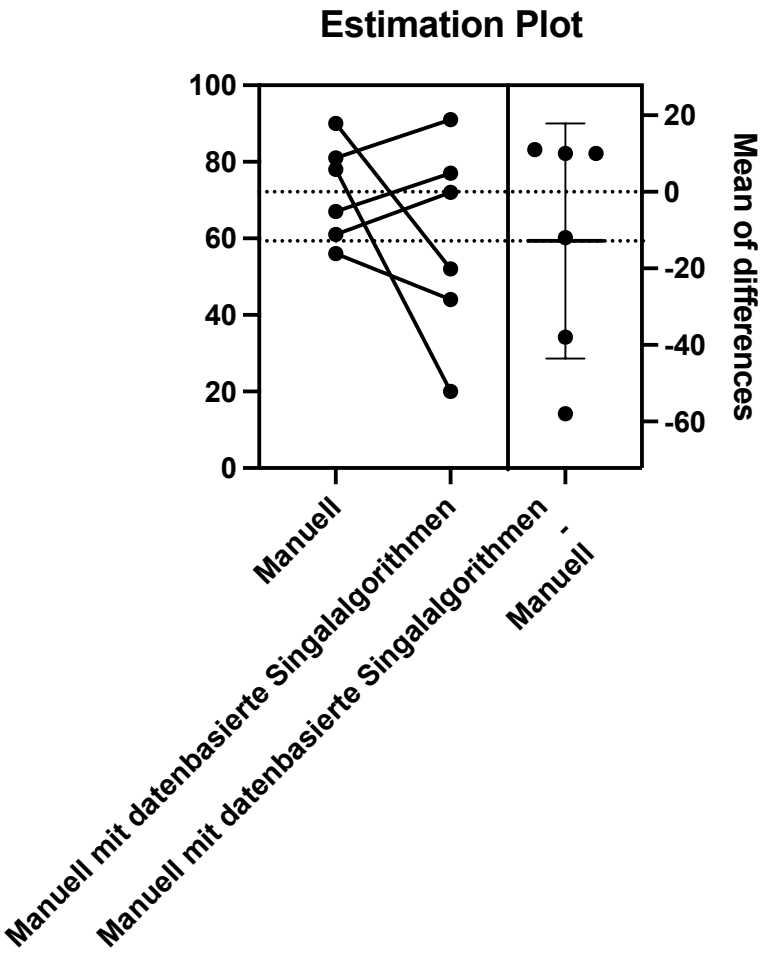

Abbildung 46: Bewertungstrend manuell – manuell mit datenbasierte Signalalgorithmen (Sprachaufnahme 2)

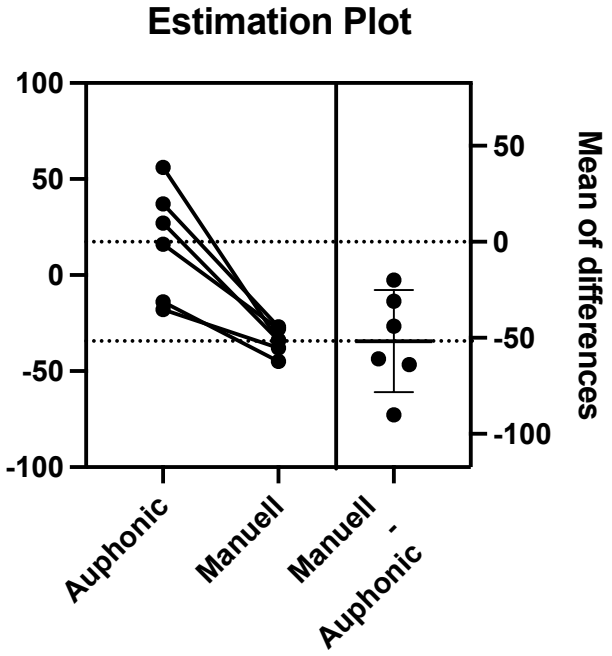

Abbildung 47: Bewertungstrend Auphonic – manuell (Sprachaufnahme 4)

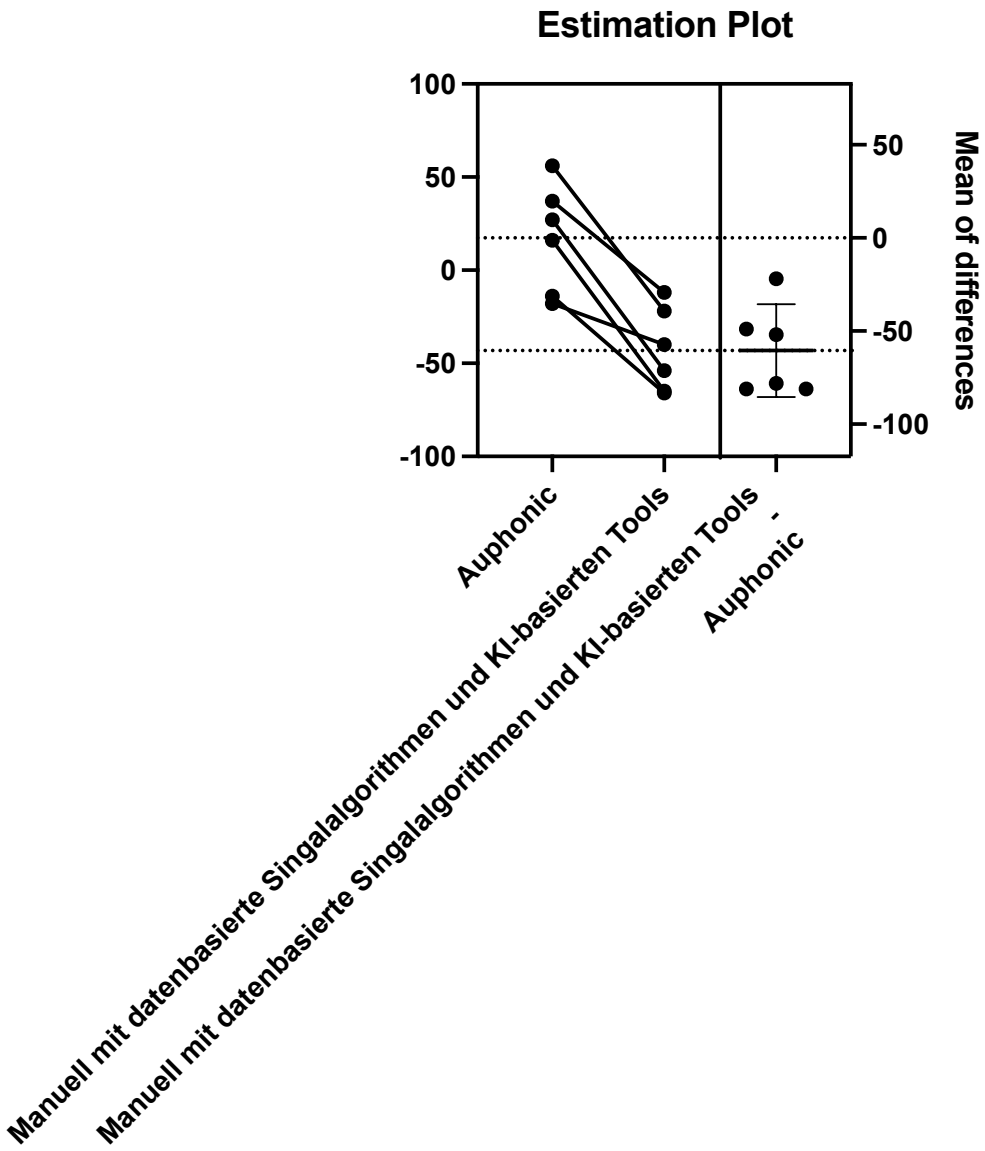

Abbildung 48: Bewertungstrend Auphonic – manuell mit datenbasierte Signalalgorithmen und KI-basierten Tools (Sprachaufnahme 4)

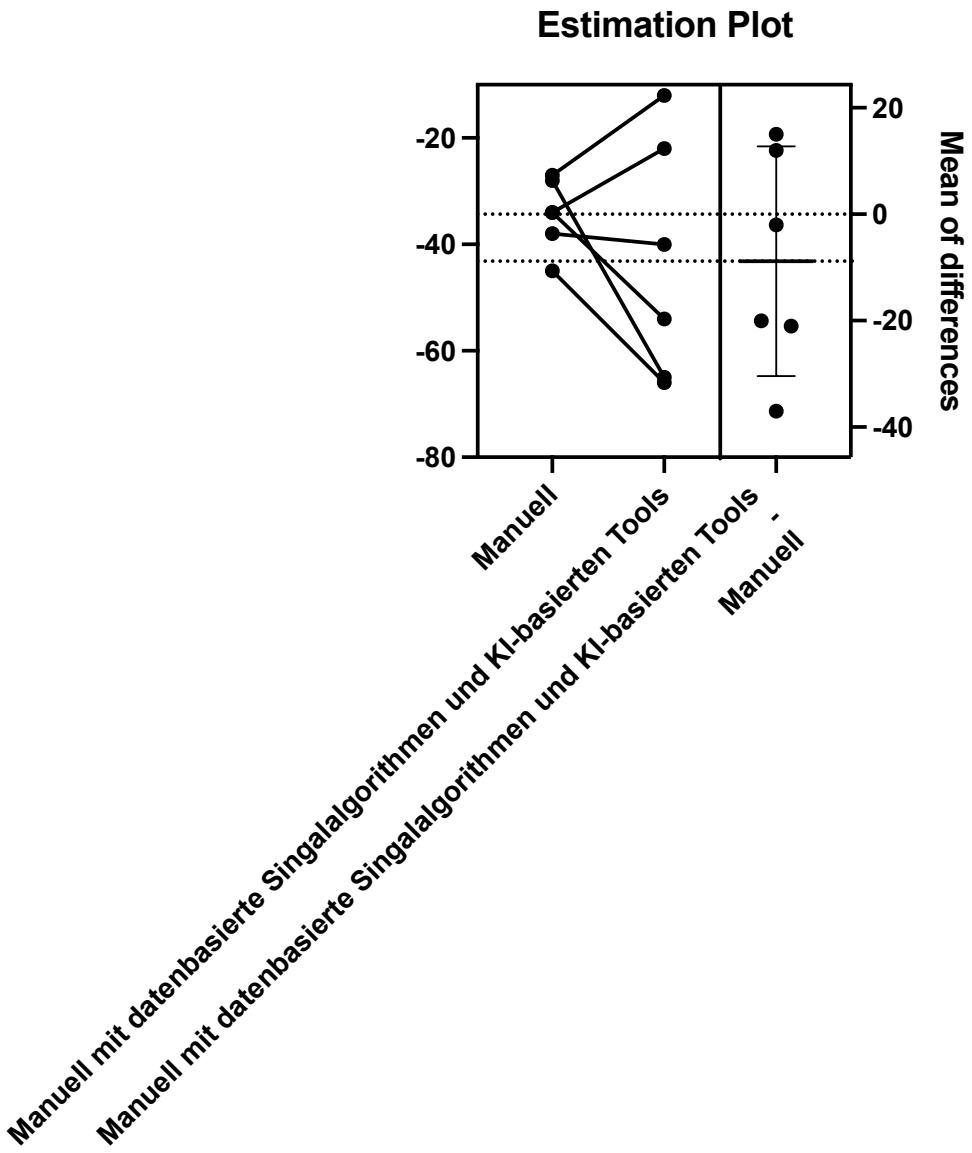

Abbildung 49: Bewertungstrend Manuell – manuell mit datenbasierte Signalalgorithmen und KI-basierten Tools (Sprachaufnahme 4)