VERFAHREN ZUR ANALYSE VON AUDIOSIGNALEN UND -SYSTEMEN

Bachelorarbeit im Studiengang Audiovisuelle Medien an der Hochschule der Medien Stuttgart zur Erlangung des akademischen Grades „Bachelor of Engineering“

Vorgelegt von:	Sam Brauner
Matrikelnummer:	34982
Mail:	sam.jj@web.de
Betreuer:	Prof. Oliver Curdt (Erstprüfer)
	Prof. Dr. Michael Felten (Zweitprüfer)

Vorgelegt am: 02. August 2021

EHRENWÖRTLICHE ERKLÄRUNG

„Hiermit versichere ich, Sam Brauner, ehrenwörtlich, dass ich die vorliegende Bachelorarbeit mit dem Titel: „Verfahren zur Analyse von Audiosignalen und -systemen“ selbstständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist noch nicht veröffentlicht oder in anderer Form als Prüfungsleistung vorgelegt worden.

Ich habe die Bedeutung der ehrenwörtlichen Versicherung und die prüfungsrechtlichen Folgen (§26 Abs. 2 Bachelor-SPO (6 Semester), § 24 Abs. 2 Bachelor-SPO (7 Semester), § 23 Abs. 2 Master-SPO (3 Semester) bzw. § 19 Abs. 2 Master-SPO (4 Semester und berufsbegleitend) der HdM) einer unrichtigen oder unvollständigen ehrenwörtlichen Versicherung zur Kenntnis genommen.“

Hilzingen, 01.08.2021

Ort, Datum

Sam Brauner

KURZFASSUNG

Die vorliegende Arbeit beschäftigt sich mit der Frage, wie verschiedene Verfahren zur Analyse von Audiosignalen und -systemen bestmöglich angewendet werden können. Hierfür wird ein fundierter Überblick über verschiedene Methoden für sowohl technische als auch künstlerische Anwendungsbereiche der Audiotechnik gegeben. Dabei wird gezeigt, dass Analyseverfahren auf einzelne Parameter fokussiert werden sollten, um möglichst aussagekräftige Ergebnisse erzielen zu können. Auch in welcher Form künstliche Intelligenz bei der Analyse von Audiosignalen eingesetzt wird und welche Vorteile diese mit sich bringt, ist Teil dieser Arbeit. Um die theoretischen Inhalte anschaulich zu vermitteln, werden die gewonnenen Erkenntnisse gezielt in verschiedenen Experimenten umgesetzt und vertieft.

ABSTRACT

This thesis deals with the question how different methods for the analysis of audio signals and systems can be optimally applied. For this purpose, a well-founded overview of different techniques for both technical and artistical applications of sound engineering is given. It is shown that analysis procedures should be focused on individual parameters in order to achieve the most meaningful results. In what form artificial intelligence is used in the analysis of audio signals and what advantages it brings, is also part of this work. In order to convey the theoretical content clearly, the knowledge gained is specifically implemented and deepened in various experiments.

INHALTSVERZEICHNIS

EHRENWÖRTLICHE ERKLÄRUNG II
KURZFASSUNG III
ABSTRACT III
INHALTSVERZEICHNIS IV
ABBILDUNGSVERZEICHNIS VII
FORMELVERZEICHNIS IX
ABKÜRZUNGSVERZEICHNIS XI
GENDERHINWEIS XII
1 EINFÜHRUNG 1
1.1 ZIELSETZUNG 1
1.2 AUFBAU DER ARBEIT 1
1.3 ANMERKUNGEN 2
2 SCHALL 3
3 SCHALLVERARBEITUNG IM MENSCHLICHEN GEHÖR 7
3.1 ANATOMIE DES OHRES 7
3.2 INTERAURALE LAUFZEITDIFFERENZ 8
3.3 INTERAURALE PEGELUNTERSCHIEDE 9
3.4 AUßENOHRÜBERTRAGUNGSFUNKTION 10
3.5 GEHÖRGANG ALS RESONATOR 10
3.6 FREQUENZANALYSE DER COCHLEA 12
4 QUANTIFIZIERUNG SUBJEKTIVER SCHALLWAHRNEHMUNG 13
4.1 LAUTSTÄRKEPEGEL IN PHON 13
4.2 LAUTHEIT 14
4.3 FREQUENZGRUPPEN 15
4.4 TIMBRE 16
4.5 TONHÖHE 16
4.6 SCHÄRFE UND RAUHEIT 17
5 ANALYSEVERFAHREN FÜR AUDIOSIGNALE UND -SYSTEME 18
5.1 AUDIOSIGNALE 18
5.1.1 Kontinuierliche und diskrete Signale 18
5.1.2 Stochastische und determinierte Signale 19

5.1.3 Energie- und Leistungssignale 20
5.1.4 Stationäre und instationäre Signale 20
5.2 AUDIOSYSTEME 21
5.2.1 Dirac-Stoß 22
5.2.2 Faltung 23
5.2.3 Impulsantwort 24
5.2.4 Faltungshall 24
5.3 VERZERRUNGEN 26
5.3.1 Lineare Verzerrungen 27
5.3.2 Nichtlineare Verzerrungen 27
5.4 ZWEI- ODER MEHRKANALIGE AUDIOSYSTEME 31
5.5 LÄRMANALYSE 33
5.6 AKUSTISCHE ANTENNEN 35
5.7 VOM ZEIT- IN DEN FREQUENZBEREICH 36
5.7.1 Transformations-Analyse 36
5.7.2 Fourier-Transformation 38
5.7.3 Diskrete Fourier-Transformation 40
5.7.4 Schnelle Fourier-Transformation 42
5.7.5 Störspektrum 42
5.7.6 Kurzzeit-Fourier-Transformation 43
5.7.7 Rauschunterdrückung mittels FFT 44
5.7.8 Wavelet-Transformation 45
5.7.9 Oktavanalyse 46
5.8 AKUSTISCHER FINGERABDRUCK 46
5.8.1 Erstellung des akustischen Fingerabdrucks 47
5.8.2 Datenbank durchsuchen 49
5.8.3 Shazam 50
6 KÜNSTLICHE INTELLIGENZ FÜR AUDIOANALYSEN 55
6.1 KÜNSTLICHE INTELLIGENZ IN PLUGINS 55
6.2 KÜNSTLICHE NEURONALE NETZWERKE 57
6.3 AUTOMATISCHE KLASSIFIZIERUNG VON AUDIOSIGNALEN 58

7 FAZIT 61
DANKSAGUNG XIII
LITERATURVERZEICHNIS XIV
ANHANG XX

ABBILDUNGSVERZEICHNIS

Abb. 1: Teilchenschwingung im zeitlichen Schalldruckverlauf p(t). 4

Abb. 2: Schallmuster von Sinuston (oben), Klang (mittig)....5

Abb. 3: Schematische Darstellung der Anatomie des Außen-, Mittel- und Innenohrs. 7

Abb. 4: Im Winkel Θ (Theta) schräg einfallende, ebene Welle.....9

Abb. 5: Schematischer Querschnitt der Cochlea. 12

Abb. 6: Kurven gleichen Lautstärkepegels (Isophone) mit der Hörschwelle bei 3 Phon.....14

Abb. 7: Die Lautheitsfunktion für den 1 kHz Sinuston (Sone-Skala). 15

Abb. 8: Klassifizierung von Signalen. 19

Abb. 9: Systemantwort y(t) auf das Eingangssignal x(t).....21

Abb. 10: Rechtecksignal d(t) der Breite T_0, der Höhe $1/T_0$ und der Fläche 1. 22

Abb. 11: Sporthalle, deren Impulsantwort berechnet wurde.....25

Abb. 12: Lissajous-Figuren für Mono-Signal (a), nur auf dem linken Kanal ein Signal (b), ...33

Abb. 13: Schematische Darstellung der konventionellen Analyse und der Transformations-Analyse.....37

Abb. 14: Zerlegung eines Signals x(t) in seine sinusförmigen Komponenten.....39

Abb. 15: Beispiel für die Anwendung der DFT. Für eine bessere Anschaulichkeit sind die Signale nicht als diskrete, sondern als kontinuierliche Kurven dargestellt.....41

Abb. 16: Störspektrum der Signalkette D/A-Wandler-Gitarrenverstärker-A/D-Wandler.....43

Abb. 17: Signale f_noise und f_clean (oben), Leistungsdichtespektrum (mittig) 44

Abb. 18: Frequenz- und Zeitauflösung bei der STFT (links) und WT (rechts). 46

Abb. 19: Schematische Übersicht der Erstellung eines 32 Bit Hashwertes.....48

Abb. 20: Fingerabdruckblöcke der originalen Musik (a) und der komprimierten Version (b), ... 49

Abb. 21: Shazam Spektrogramm eines Audiofiles. 51

Abb. 22: Shazam Constellation Maps. 51

Abb. 23: Festlegung eines Ankerpunkts mit zugehöriger Target Zone. 52

Abb. 24: Bildung des Hash:Zeit Wertes eines Target Zonen Punkt mit dessen Ankerpunkt. 53

Abb. 25: Vereinfachte Darstellung eines künstlichen neuronalen Netzwerks. 58

Abb. 26: Python Code zur Rauschunterdrückung mit FFT. XX

Abb. 27: Python Code zur Vorbereitung des Datensatzes für das CNN. XXI

Abb. 28: Python Code für das CNN-Training - Teil 1. XXII

Abb. 29: Python Code für das CNN-Training - Teil 2. XXIII

FORMELVERZEICHNIS

(1) Schalldruckpegel L_p 5

(2) Schalldruck über Mittelungsdauer $\tilde{p}$ 6

(3) Maximale Laufzeitdifferenz τ_{max} 8

(4) Problematische Frequenz bei Schalleinfall aus 90° 9

(5) Eigenfrequenzen f_n des Gehörgangs 10

(6) Energie E_x eines zeitkontinuierlichen Signals 20

(7) Mittlere Leistung P_x eines zeitkontinuierlichen Signals 20

(8) Energie E_x eines zeitdiskreten Signals 20

(9) Mittlere Leistung P_x eines zeitdiskreten Signals 20

(10) Skalierungsfaktor H eines LTI-Systems 21

(11) Approximation einer Zeitfunktion durch Rechtecksignale 22

(12) Systemantwort auf die Approximation durch Rechtecksignale 23

(13) Approximation des Eingangssignals durch die Dirac-Funktion 23

(14) Approximation des Ausgangssignals durch die Dirac-Funktion 23

(15) Faltung in der Kurzschreibweise 23

(16) Total Harmonic Distortion THD 28

(17) Klirrfaktor k 29

(18) Teilklirrfaktor k_n 29

(19) Summentöne f_S und Differenztöne f_D 30

(20) Intermodulationsfaktor 2. Ordnung m_2 30

(21) Beziehung zwischen LTI-System und Eigenfunktion 38

(22) Aus der Impulsantwort $h(t)$ wird die Übertragungsfunktion $H(f)$ 38

(23) Aus der Übertragungsfunktion $H(f)$ wird die Impulsantwort $h(t)$ 38

(24) Fourier-Transformation $F\{x(t)\}$ 38

(25) Inverse Fourier-Transformation $F^{-1}\{X(f)\}$ 38

(26) Amplitudenspektrum $|X(f)|$ 40

(27) Phasenspektrum $\varphi(f)$ 40

(28) Diskrete Fourier-Transformation $F\{x(n)\}$ 40

(29) Inverse diskrete Fourier-Transformation $F^{-1}\{X(k)\}$ 40

(30) Substitution für die Matrixdarstellung der DFT ω 41

(31) DFT in der Matrixdarstellung 41

(32) Histogramm aus Zeitkoordinate des Datenbank Files und Zeitkoordinate der Handyaufnahme δt_k 54

ABKÜRZUNGSVERZEICHNIS

A/D-Wandler:	Analog-Digital-Wandler
AKF:	Autokorrelationsfunktion
BER:	Bitfehlerrate (von engl. bit error rate)
DFT:	Diskrete Fourier-Transformation
FFT:	Schnelle Fourier-Transformation (von engl. fast fourier transform)
iFFT:	Inverse schnelle Fourier-Transformation (von engl. inverse fast fourier transform)
ILD:	Interaurale Pegelunterschiede (von engl. interaural level difference)
ITD:	Interaurale Laufzeitdifferenz (von engl. interaural time difference)
KKF:	Kreuzkorrelationsfunktion
LDS:	Leistungsdichtespektrum
LTI-Systeme:	Linear zeitinvariante Systeme (von engl. linear time-invariant)
LUT:	Lookup-Tabelle
MFCC:	Mel-Frequenz-Cepstrum-Koeffizient (von engl. Mel Frequency Cepstral Coefficient)
SNR:	Signal-Rausch-Verhältins (von engl. signal-to-noise ratio)
SSG:	Stereosichtgerät
THD:	Total-Harmonic-Distortion
THD+N:	Total-Harmonic-Distortion-Plus-Noise
WT:	Wavelet-Transformation

GENDERHINWEIS

In dieser Arbeit wird aus Gründen der besseren Lesbarkeit das generische Maskulinum verwendet. Es sind dabei alle Geschlechteridentitäten ausdrücklich stets eingeschlossen.

1 EINFÜHRUNG

„Wenn man mit günstigem Winde einem ruft, ist darum der Ton nicht lauter […]. Man nützt nur die Verhältnisse aus“ (Wilhelm, 1928, S. 221).

Ob gesprochene Sprache, Morsecode, geblasene Hörner oder läutende Glocken, Audiosignale werden seit jeher von Menschen genutzt, um Informationen zu übertragen. Während früher meist das menschliche Ohr als Empfänger für Audiosignale diente, gibt es heute zahlreiche Geräte, welche Schall in elektrische Ströme umwandeln und in weiteren Schritten nutzbar und verarbeitbar machen. So sind bei Audiosignalen nicht nur akustische, sondern vor allem auch elektrische oder elektromagnetische Schwingungen gemeint.[1] Daraus wurden in der Audiotechnik zahlreiche Verfahren entwickelt, um die bestehenden „Verhältnisse“ zu analysieren und dadurch bestmögliche Ergebnisse bei der Darstellung, Übertragung oder Verarbeitung der Audiosignale erzielen zu können.

1.1 Zielsetzung

Diese Bachelorarbeit soll durch einen fundierten Überblick zeigen, wie sowohl etablierte als auch neuere, auf einzelne Parameter fokussierte Methoden zur Analyse verschiedener Problematiken im Audiobereich genutzt werden. Durch Anwendungsbezogene Erläuterungen, ergänzende Experimente und Quellcode-Beispiele sollen für den Leser komplexere Inhalte verständlich und anschaulich gemacht werden. Auch wie künstliche Intelligenz bei Audioanalysen Anwendung findet und inwieweit es diese ermöglicht, anspruchsvolle Verfahren für Laien zugängliche zu machen, soll Teil dieser Arbeit sein.

1.2 Aufbau der Arbeit

Zu Beginn der vorliegenden Arbeit wird sowohl auf Grundlagen und charakteristische Eigenschaften des Schalls, als auch auf dessen Verarbeitung und Analyse durch das menschliche Gehör eingegangen. Anschließend werden mögliche Verfahren dargestellt, um die subjektive auditive Wahrnehmung zu quantifizieren und Analysen

[1] Vgl. Görne (2014), S. 135 f.

dementsprechend anpassen zu können. Im Kern der Arbeit wird auf die verschiedenen Methoden zur Analyse und Betrachtung von Audiosignalen und -systemen eingegangen. Abschließend wird gezeigt, wie dabei künstliche Intelligenz Anwendung findet und wie Algorithmen wie akustische Fingerabdrücke und Neuronale Netze genutzt werden können, um bspw. Musik auf Eigenschaften zu analysieren und zu kategorisieren.

1.3 Anmerkungen

Durch ausgewählte Abbildungen, Experimente und praktische Beispiele, sollen dem Leser komplexe Thematiken verständlich aufgezeigt werden. Dennoch sind mathematische und technische Grundkenntnisse nötig, um die Erläuterungen in dieser Arbeit in vollem Umfang nachvollziehen zu können. Dazu gehört vor allem ein Verständnis der mathematischen Analysis mit Begriffen wie Integralrechnung, Folgen und Reihen, Grenzwerte oder komplexe Zahlen. Außerdem werden zum Nachvollziehen der programmierten Anwendungen Kenntnisse der Programmiersprache Python vorausgesetzt. Auch auf die Erläuterung nebensächlicher audiotechnischer Grundlagen wie die Funktionsweise von Mikrofonen, Lautsprechern, verschiedenen Filtern oder der A/D-Wandlung usw. wird bewusst verzichtet.

2 SCHALL

In der Audiotechnik gibt es verschiedene technische Geräte und Verfahren, die dazu dienen, akustische Ereignisse (Schall) messen und analysieren zu können. Als Grundlage der in dieser Arbeit betrachteten Thematiken, wird in diesem Abschnitt zunächst auf die Entstehung und die Eigenschaften des Schalls eingegangen.

Schall entsteht immer dann, wenn ein vibrierendes Objekt (Schallquelle) Schwingungen auf das umgebende Medium weitergibt. Im Medium kommt es daraufhin zu Druckunterschieden, welche sich entsprechend der Richtcharakteristik der Schallquelle ausbreiten.[2] Der Schalldruck p wird in der Einheit Pascal (Pa) angegeben. Bei einem Ton mit 1000 Hz befinden sich üblicherweise wahrgenommene Schalldruckwerte zwischen der Hörschwelle von 20 µPa und der Schmerzgrenze von etwa 150 Pa.[3] Für die Ausbreitung eines Schalls muss das Medium elastisch sein. Das bedeutet, dass durch Krafteinwirkung entstandene Deformationen automatisch durch Rückstellkräfte wieder rückgängig gemacht werden.[4] Werden die Teilchen im Medium also aus ihrer Ruhelage gebracht, beginnen diese zu schwingen. Eine räumlich-zeitliche Ausbreitung dieses Schwingungszustandes wird als Schallwelle bezeichnet.[5] Dabei stößt das eine Teilchen sein benachbartes an, dieses wiederum das Nächste und so weiter. Im Zusammenhang mit der menschlichen Wahrnehmung findet die Schallausbreitung meist über die Luft (Luftschall) statt.[6] Stattdessen kann Schall aber auch durch Flüssigkeiten (z. B. Wasser) oder Festkörper (Körperschall) übertragen werden.[7] Die Entstehung der Druckunterschiede ist in Abb. 1 vereinfacht dargestellt. Zum Zeitpunkt A sind alle Luftteilchen in ihrer Ruhelage, der Schalldruck ist gleich Null. Wird die Saite losgelassen, werden die umliegenden Teilchen aus ihrer Ruhelage ausgelenkt und stoßen gegen die benachbarten Teilchen. Im grau gekennzeichneten Bereich kommt es daher zum Zeitpunkt B zu einem Schalldruckmaximum. Da die Teilchen sich daraufhin auseinander bewegen, kommt es zum Zeitpunkt C zu einem Minimum. Wird in diesem Beispiel nun der Schalldruck p in der grauen Fläche über die Zeit t betrachtet, ergibt sich wie in Abb. 1 links dargestellt, eine Sinusschwingung.

[2] Vgl. Hellbrück (1993), S. 51.
[3] Vgl. Dickreiter/Hoeg (2014), S. 13.
[4] Vgl. Hellbrück (1993), S. 46.
[5] Vgl. Hellbrück (1993), S. 51.
[6] Vgl. Hellbrück (1993), S. 51.
[7] Vgl. Hellbrück (1993), S. 51.

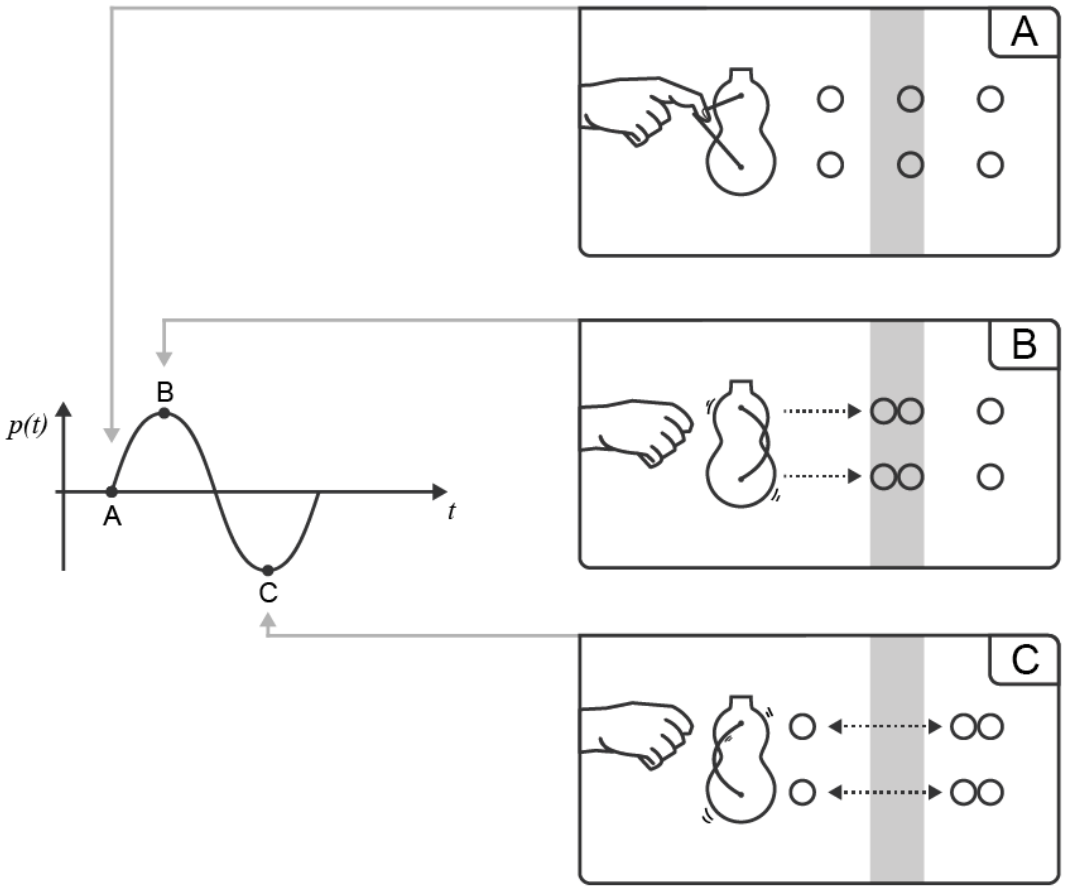

Abb. 1: Teilchenschwingung im zeitlichen Schalldruckverlauf p(t).[8]

Die sogenannte Schallschnelle gibt dabei die momentane Geschwindigkeit an, mit der ein Teilchen um seinen Ruhepunkt schwingt.[9] Da die Schallschnelle neben einem Zahlenwert auch eine Richtungsangabe besitzt, stellt sie eine vektorielle Größe dar.[10] Während in den meisten Anwendungsbereichen hauptsächlich der Schalldruck von Interesse ist, findet die Schallschnelle vor allem bei der Untersuchung der physikalischen Details von Schallfeldern ihren Einsatz.[11] Um die Schallschnelle direkt zu messen, können Hitzedrahtanemometer genutzt werden. Diese bestehen aus Platinwiderständen in Form von zwei sehr dünnen, nah beieinander liegenden Drähten, die ihre Wärmeenergie an die umgebende Luft abgeben (Betriebstemperatur von etwa 200 °C bis 400 °C).[12] Kommt es nun durch einen Schall zu einer lokalen Luftströmung, ändert sich die Temperaturverteilung asymmetrisch und es entstehen Widerstands- und Spannungsdifferenzen.[13] Aus diesen Differenzen kann schließlich die Schallschnelle

[8] In Anlehnung an Lindsay/Norman (1977), S. 127.
[9] Vgl. Dickreiter/Hoeg (2014), S. 13.
[10] Vgl. Dickreiter/Hoeg (2014), S. 13.
[11] Vgl. Müller u. a. (2017), S. 5.
[12] Vgl. Müller u. a. (2017), S. 5.
[13] Vgl. Müller u. a. (2017), S. 5.

abgeleitet werden.[14] Mit einem solchen Schnellesensor kann im Bereich von etwa 100 nm/s bis 0,1 m/s gemessen werden.[15]

Trifft ein Schall auf das menschliche Trommelfell und versetzt dieses mit einer hörbaren Frequenz (zwischen 20 Hz und 20 kHz) in Schwingung, wird eine Schallwahrnehmung ausgelöst.[16] Dabei kann sich das Schallmuster entweder periodisch oder nicht periodisch in der Zeit wiederholen. Die einfachste periodische Schwingung stellt der Sinuston dar (vgl. Abb. 2 (oben)). Ebenfalls periodisch sind Klänge, jedoch sind diese in ihrer Zusammensetzung komplexer (vgl. Abb. 2 (mittig)). Dazu zählen unter anderem Instrumentenklänge und Tierlaute. Als nicht periodisch gelten die übrigen Geräusche wie z. B. Bachplätschern oder das Rauschen von Blättern im Wind (vgl. Abb. 2 (unten)).

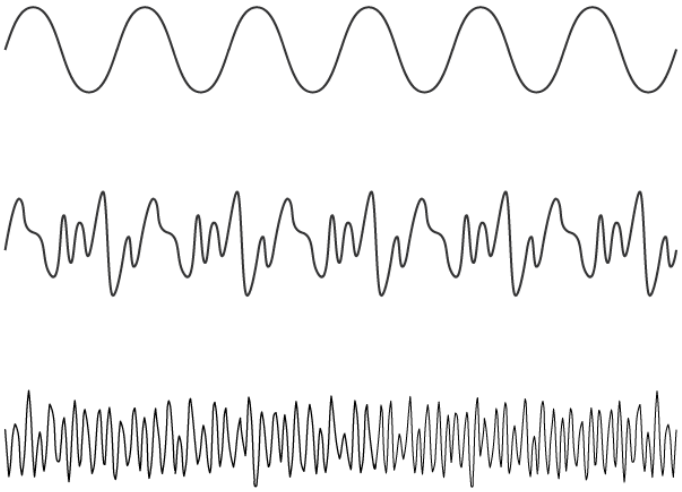

Abb. 2: Schallmuster von Sinuston (oben), Klang (mittig) und Geräusch (unten).

Um einzelne Schalldruckwerte besser differenzieren zu können, wird der sogenannte Schalldruckpegel L_p in dB verwendet. Dieser beträgt das 20-fache des logarithmierten Verhältnisses des Schalldrucks p zu einem vereinbarten Bezugsschalldruck p_0:[17]

$$L_p = 20 \cdot \log\left(\frac{p}{p_0}\right) \tag{1}$$

[14] Vgl. Müller u. a. (2017), S. 5 f.
[15] Vgl. Müller u. a. (2017), S. 6.
[16] Vgl. Hellbrück (1993), S. 46.
[17] Vgl. Dickreiter/Hoeg (2014), S. 15.

Dabei wird für den Bezugsschalldruck meist die Hörschwelle mit 20 µPa gewählt.[18] Da der Schalldruck häufig über eine bestimmte Mittelungsdauer T_m betrachtet wird, kann für p auch $\tilde{p}$ mit

$$\tilde{\mathrm{p}} = \sqrt{\frac{1}{T_m}\int_0^{T_m} p^2(t)dt} \tag{2}$$

eingesetzt werden, um den Effektivwert eines Schalldruck-Zeitverlaufs $p(t)$ zu erhalten.[19] Aufgrund der Berechnung des Effektivwertes wird dies im englischen auch häufig mit der Abkürzung RMS (Root Mean Square) bezeichnet.[20] Je nachdem, ob ein Zeitsignal $p(t)$ eher gleichbleibend oder eher impulsartig ist, wird für die Mittelwertbildung zwischen verschiedenen, in internationalen Normen vereinbarten Zeitkonstanten gewählt (bspw. FAST = 125 ms oder SLOW = 1 s).[21] Die Zeitkonstante SLOW hat den Vorteil, dass die Schallpegelanzeige langsamer schwankt und daher besser abgelesen werden kann.[22] Werden jedoch kurze Impulsspitzen betrachtet, ist FAST zu wählen, da diese mit SLOW möglicherweise nicht erfasst werden.[23]

[18] Vgl. Müller u. a. (2017), S. 13.
[19] Vgl. Müller u. a. (2017), S. 13.
[20] Vgl. Müller u. a. (2017), S. 13.
[21] Vgl. Müller u. a. (2017), S. 13.
[22] Vgl. Müller u. a. (2017), S. 13.
[23] Vgl. Müller u. a. (2017), S. 13.

3 SCHALLVERARBEITUNG IM MENSCHLICHEN GEHÖR

Ohne die Fähigkeit des Menschen hören zu können, wären die Audiotechnik und damit auch die Verfahren zur Analyse des Schalls zum größten Teil irrelevant. Viele Techniken orientieren sich am Prinzip des menschlichen Ohres. Außerdem ist es häufig von Interesse, Analysen bestimmter Parameter den Eigenschaften des Hörens anzupassen. Im folgenden Abschnitt wird auf die Anatomie und auf die Signalverarbeitung des Ohres eingegangen. Außerdem wird gezeigt, wie dem Menschen das Richtungshören und die Frequenzanalyse des Schalls möglich ist.

3.1 Anatomie des Ohres

Das Ohr wird in die drei anatomischen Bereiche Außenohr, Mittelohr und Innenohr unterteilt.[24] Auf dem Weg von außen nach innen trägt jeder Bereich dazu bei, die Schallwellen in für das Gehirn verwertbare neuronale Impulse zu wandeln. Das Außenohr umfasst hierbei die Ohrmuschel und den äußeren Gehörgang und endet mit dem Trommelfell am Mittelohr (vgl. Abb. 3).[25] Die Ohrmuschel dient zur Aufnahme des Schalls aus der Umwelt.

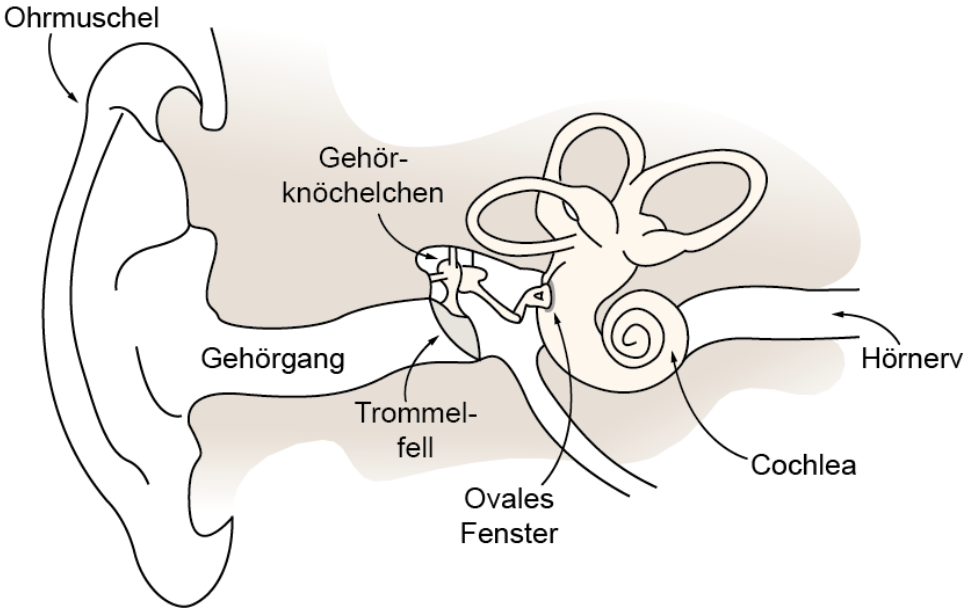

Abb. 3: Schematische Darstellung der Anatomie des Außen-, Mittel- und Innenohrs.[26]

[24] Vgl. Hellbrück (1993), S. 83.
[25] Vgl. Hellbrück (1993), S. 83.
[26] In Anlehnung an Lindsay/Norman (1977), S. 126.

Die Form des menschlichen Kopfes, des Oberkörpers, aber vor allem des komplexen Außenohrs, kodieren richtungsabhängige Information wie Phasen-, Amplituden- und Spektralverzerrungen auf den eintreffenden Schall.[27] Dadurch ist es dem Menschen möglich, die Richtung aus welcher der Schall eintrifft, zu bestimmen.[28] Für die Ortung auf der Horizontalebene sind vor allem die Laufzeit- und Pegelunterschiede zwischen den beiden Ohren ausschlaggebend.

3.2 Interaurale Laufzeitdifferenz

Wenn eine ebene Welle seitlich auf den Kopf trifft, kommt diese bei dem zugewandten Ohr früher an als bei dem abgewandten, wodurch Laufzeitdifferenzen entstehen (vgl. Abb. 4). Den maximalen Umweg um den Kopf hat der Schall, wenn er genau aus 90° eintrifft.[29] Dieser beträgt durchschnittlich rund 21 cm und führt geteilt durch die Schallgeschwindigkeit c (bei 20 °C = 343 m/s)

$$\tau_{max} = \frac{0{,}21\ m}{c} \tag{3}$$

zu einer maximalen Laufzeitdifferenz τ_{max} (Tau) der Ohrsignale von ungefähr 0,61 ms.[30] Fällt die ebene Welle stattdessen von 0° bzw. 180° ein, kommt es zu keiner interauralen Laufzeitdifferenz (ITD). Am genausten können Richtungsänderungen wahrgenommen werden, wenn der Schall von vorne eintrifft.[31] Hier kann der Mensch Winkeländerungen von 1° differenzieren, die einer ITD von etwa 0,01 ms entsprechen.[32]

[27] Vgl. Kuhn (1987), S. 3.
[28] Vgl. Kuhn (1987), S. 3.
[29] Vgl. Görne (2014), S. 126.
[30] Vgl. Görne (2014), S. 126.
[31] Vgl. Görne (2014), S. 126.
[32] Vgl. Görne (2014), S. 126.

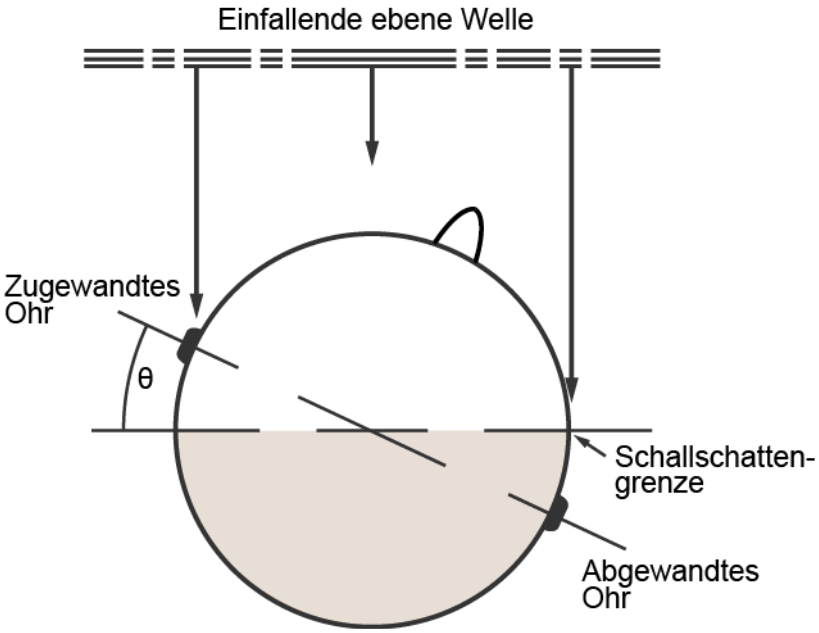

Abb. 4: Im Winkel Θ (Theta) schräg einfallende, ebene Welle.[33]

Bei interauralen Laufzeitdifferenzen kann es unter Umständen zu Fehlortungen kommen. Dies lässt sich gut mit dem Schalleinfall von der Seite (Θ = 90°) verdeutlichen. Aus der maximalen Laufzeitdifferenz (aus Formel (3)) lässt sich eine Frequenz *f* von

$$f = \frac{1}{\tau_{max}} = 1{,}63\ kHz \qquad (4)$$

bestimmen. Fällt nun eine Schallwelle mit dieser Frequenz aus 90° ein, sind die Signale beider Ohren phasengleich.[34] Da solch eine Phasengleichheit auch bei Schalleinfall von vorne oder von hinten auftritt, kann es zur mehrdeutigen Ortung kommen. Dies geschieht vor allem dann, wenn das Signal zusätzlich hochfrequent (mindestens eine Frequenz von 1,63 kHz) und im zeitlichen Verlauf näherungsweise gleichbleibend ist.[35]

3.3 Interaurale Pegelunterschiede

Trifft eine ebene Welle schräg auf den Kopf, entstehen neben den Laufzeitunterschieden auch Pegelunterschiede (ILD) zwischen den beiden Ohrsignalen. Dabei ist nicht nur die unterschiedliche Entfernung der Ohren zur Schallquelle ausschlaggebend. Ist die Wellenlänge kleiner als der Kopfdurchmesser, wird die Welle am Kopf reflektiert.[36]

[33] In Anlehnung an Kuhn (1987), S. 5.
[34] Vgl. Görne (2014), S. 126 f.
[35] Vgl. Görne (2014), S. 127.
[36] Vgl. Görne (2014), S. 127.

Dadurch kommt es am zugewandten Ohr zu einem Druckstau (höherer Schalldruckpegel). Das abgewandte Ohr hingegen liegt im Schallschatten (vgl. Abb. 4). So ist bspw. bei einem Kopfdurchmesser von ca. 16,5 cm davon auszugehen, dass Frequenzen $f \geq 2078$ Hz reflektiert werden. Bei niedrigeren Frequenzen kommt es stattdessen zur Beugung, wodurch diese bei ILD nicht weiter ins Gewicht fallen.[37]

3.4 Außenohrübertragungsfunktion

Die Fähigkeit Schallquellen in der Horizontalebene zu lokalisieren, sowie die Unterscheidung hinten/vorne, ist unter anderem durch die komplexe Form der Ohrmuschel gegeben.[38] Je nachdem aus welcher Richtung der Schall eintrifft, kommt es durch Überlagerung von den Reflektionen an der Ohrmuschel und dem Direktschall zu charakteristischen Interferenzmustern.[39] Diese spiegeln sich im Signal am Gehörgangeingang als hochfrequentes Kammfilter mit Pegelunterschieden von ca. 20 dB wider.[40] Das Hörorgan und das Gehirn sind später in der Lage, diese Interferenzmuster zu analysieren und Informationen über die Richtung, aus der ein Schall eintrifft, zu entnehmen.[41]

3.5 Gehörgang als Resonator

Der gekrümmte Gehörgang mit einer ungefähren Länge von 22,5 mm und einem Durchmesser von 7,5 mm dient als Verbindung zum Trommelfell.[42] Durch eine geöffnete Seite und eine durch das Trommelfell geschlossen Seite, verhält sich der Gehörgang wie ein λ/4-Resonator.[43] Seine Eigenfrequenzen f_n lassen sich durch

$$f_n = (2n-1)\frac{c}{4L}, \quad n = 1, 2, 3, \ldots \tag{5}$$

bestimmen, wobei c mit 343 m/s für die Schallgeschwindigkeit (bei 20 °C) und L mit 0,0225 m für die Länge des Gehörganges stehen.[44] Demzufolge kommt es im Bereich

[37] Vgl. Görne (2014), S. 127.
[38] Vgl. Kuhn (1987), S. 3.
[39] Vgl. Görne (2014), S. 111 ff.
[40] Vgl. Görne (2014), S. 111.
[41] Vgl. Görne (2014), S. 111 ff.
[42] Vgl. Zacharov (2018), S. 33.
[43] Vgl. Zacharov (2018), S. 33.
[44] Vgl. Görne (2014), S. 61.

der Grundfrequenz $f_1 \approx 3{,}8$ kHz zu Anhebungen um (wie Abb. 6 zu entnehmen) ca. 10 dB. Entsprechend bildet sich eine weitere Anhebung bei der Oberschwingung $f_2 \approx 11{,}4$ kHz. In diesen durch den Gehörgang verstärkten Frequenzbereichen befinden sich einige Laute, welche häufig in der menschlichen Sprache zu finden sind. Daher wirken sich diese Verzerrungen am ursprünglichen Signal positiv auf die Sprachverständlichkeit aus. Wird die Länge des Resonators L im Nenner der Formel (5) anstatt mit 4 mit 2 multipliziert, lassen sich Frequenzen bestimmen bei denen es zu einer Absenkung des Pegels kommt. Im hörbaren Bereich geschieht dies dementsprechend bei etwa 7,6 kHz (vgl. Abb. 6).

Nach dem Gehörgang trifft der Schall auf das Trommelfell.[45] Dieses wandelt einfallende Schallwellen in mechanische Vibrationen um und gibt diese im Mittelohr an die Gehörknöchelchen (Hammer, Amboss und Steigbügel) weiter.[46] Über den Steigbügel sind diese wiederum mit dem sogenannten ovalen Fenster der Gehörschnecke verbunden (vgl. Abb. 3).[47] Die Gehörschnecke, auch Cochlea genannt, ist eine spiralförmige Röhre, welche mit beinahe nicht komprimierbarer Lymphflüssigkeit gefüllt ist.[48] Im Mittelohr finden durch die Transformierung des Trommelfells auf das kleinere ovale Fenster und durch die Hebelwirkung der Gehörknöchelchen Impedanzanpassungen statt.[49] So wird vermieden, dass die meiste Schallenergie aufgrund des Impedanzunterschiedes zwischen den beiden Übertragungsmedien am ovalen Fenster reflektiert und daher ungenutzt bleibt. Im schematischen Querschnitt der Cochlea in Abb. 5 ist zu erkennen, wie die Lymphflüssigkeit durch die Basilarmembran und die Reissnersche Membran getrennt wird.[50] Auf der Basilarmembran sitzt das Cortische Organ, welches mit Haarzellen ausgestattet ist.[51]

[45] Vgl. Hellbrück (1993), S. 83.
[46] Vgl. Hellbrück (1993), S. 84.
[47] Vgl. Hellbrück (1993), S. 84.
[48] Vgl. Moore (2007), S. 9.
[49] Vgl. Moore (2007), S. 8.
[50] Vgl. Moore (2007), S. 9.
[51] Vgl. Harrison/Hunter-Duvar (1988), S. 146.

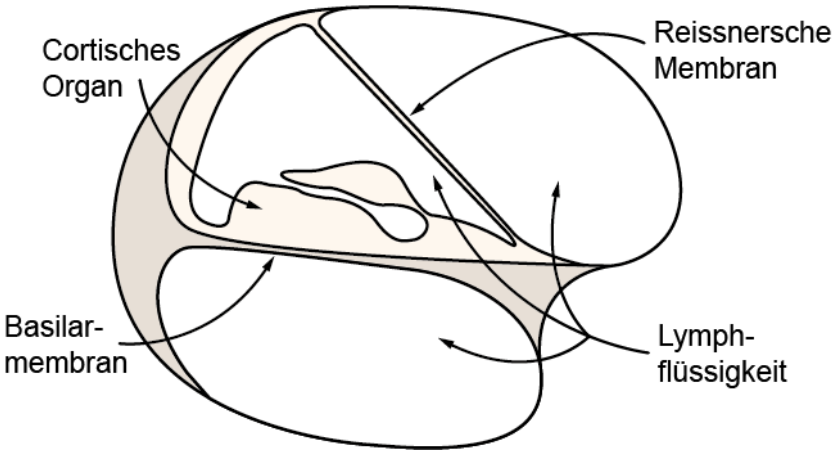

Abb. 5: Schematischer Querschnitt der Cochlea.[52]

Treffen nun mechanische Wellen über die Gehörknöchelchen auf das ovale Fenster, wird die Lymphflüssigkeit des Innenohrs und somit die Basilarmembran in Schwingung versetzt.[53] In Folge dessen werden die kleinen Härchen der Haarzellen ausgelenkt und es werden neuronale Impulse (Aktionspotentiale) in der angrenzenden Nervenzelle ausgelöst.[54]

3.6 Frequenzanalyse der Cochlea

Die Schwingung der Basilarmembran wird in Form einer Wanderwelle ausgelöst.[55] Treffen also durch Schall ausgelöste Schwingungen auf das ovale Fenster, wird die Basilarmembran erst dort und dann immer tiefer in der Cochlea ausgelenkt. Dabei ist die Frequenz des Schalls entscheidend, wie weit die Wanderwelle eindringt.[56] Während niedrigere Frequenzen tiefer in die Cochlea gelangen, bewirken höhere Frequenzen bereits nahe des ovalen Fensters die maximale Auslenkung der Basilarmembran.[57] Durch diese Frequenz-Orts-Transformation ist es dem menschliche Ohr möglich, komplexe Schalle auf darin enthaltene Frequenzen zu analysieren.[58] Werden bspw. zwei Sinustöne überlagert, lösen diese zwar eine komplexere Schwingung des Trommelfells aus, können aber dennoch als einzelne Töne differenziert werden.[59]

[52] In Anlehnung an Allen (1988), S. 247.
[53] Vgl. Moore (2007), S. 10.
[54] Vgl. Moore (2007), S. 11.
[55] Vgl. Görne (2014), S. 114.
[56] Vgl. Görne (2014), S. 114.
[57] Vgl. Görne (2014), S. 114.
[58] Vgl. Hellbrück (1993), S. 116 f.
[59] Vgl. Hellbrück (1993), S. 117.

4 QUANTIFIZIERUNG SUBJEKTIVER SCHALLWAHRNEHMUNG

Eine Schallwelle, die auf das Ohr trifft, ist ein durch Parameter wie Schalldruckpegel, enthaltene Frequenzen, Dauer und Einfallsrichtung ein eindeutig physikalisch beschreibbarer Reiz.[60] Dieser wird als Schallereignis bezeichnet.[61] Durch die Eigenschaften des Ohres und der anschließenden Nervenbahnen und Nervenzentren wird der Reiz zu einer Empfindung.[62] So wird aus einem Schallereignis ein sogenanntes Hörereignis.[63] Da Hörereignisse keine physikalische Größe sind, können diese auch nicht direkt durch physikalische Analysen quantifiziert werden.[64] Um dennoch Angaben zu erhalten, können Versuche mit Personen durchgeführt und die Aussagen über die betrachteten Parameter ausgewertet werden.[65] Im Folgenden werden verschiedene Verfahren dargestellt, die bestmöglich quantifizieren sollen, wie laut eine Person ein bestimmtes Schallereignis wahrnimmt. Außerdem werden die für die Klanganalyse wichtigen Eigenschaften Timbre, Tonhöhe, Schärfe und Rauheit erläutert.

4.1 Lautstärkepegel in Phon

Der Mensch nimmt verschiedene Frequenzen unterschiedlich laut war. So wirkt ein Ton mit z. B. 50 dB bei 100 Hz deutlich leiser als bei 4 kHz. Um dies bei Analysen und technischen Anwendungen mit einfließen zu lassen, wurden unter anderem die Kurven gleichen Lautstärkepegels entwickelt. Hierfür wurden Hörversuche durchgeführt, in denen Probanden ausgehend von einem 1 kHz Ton eines bestimmten Schallpegels (z. B. 40 dB) die restlichen Frequenzen so einpegeln sollten, dass sie von ihnen als gleich laut empfunden wurden.[66] Die daraus entstandenen Kurvenverläufe (Isophone) erhalten die Einheit Phon und sind in Abb. 6 zu sehen. Hier ist zu erkennen, wie bspw. tiefe Frequenzen weniger gut wahrgenommen werden als die zwischen 2 kHz und 5 kHz.

[60] Vgl. Dickreiter/Goeres-Petri (2014), S. 118.
[61] Vgl. Dickreiter/Goeres-Petri (2014), S. 118.
[62] Vgl. Dickreiter/Goeres-Petri (2014), S. 118.
[63] Vgl. Dickreiter/Goeres-Petri (2014), S. 118.
[64] Vgl. Dickreiter/Goeres-Petri (2014), S. 118.
[65] Vgl. Dickreiter/Goeres-Petri (2014), S. 118.
[66] Vgl. Hellbrück (1993), S. 64.

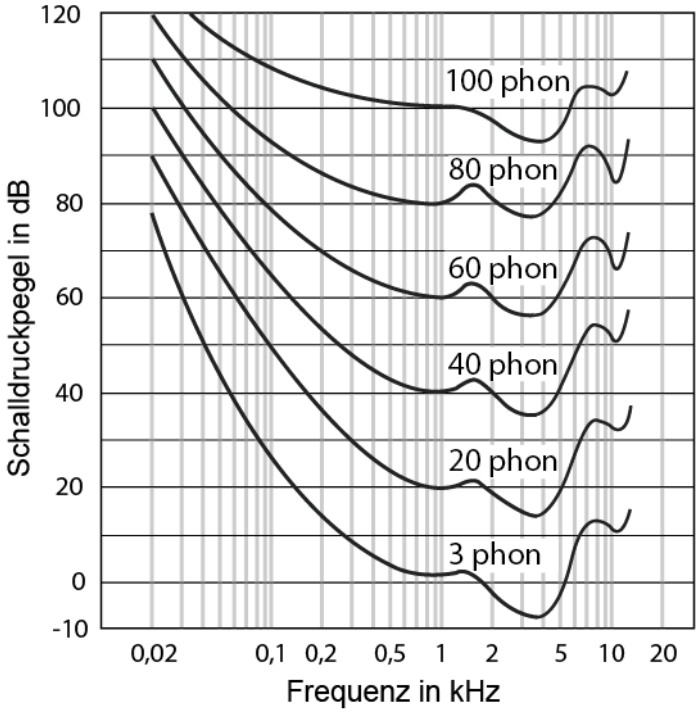

Abb. 6: Kurven gleichen Lautstärkepegels (Isophone) mit der Hörschwelle bei 3 Phon.[67]

4.2 Lautheit

Da Isophone allein noch keinen Aufschluss darüber geben, wie laut das menschliche Ohr bestimmte Schallereignisse tatsächlich wahrnimmt, wurde die Lautheit (engl. loudness) eingeführt. Mit der Maßeinheit Sone hat ein Sinuston mit einer Frequenz von 1 kHz und 40 dB die Lautheit 1 sone (vgl. Abb. 7).[68] Wird die Lautstärke um 10 dB erhöht, verdoppelt sich die Lautheit auf 2 sone. Bei weiteren 10 dB auf 4 sone usw. Dadurch erhält ein als doppelt so laut empfundenes Schallereignis auch einen doppelt so hohen Wert auf der Skala. Bei Werten unterhalb von 40 dB ist dieser potenzielle Verlauf nicht mehr vorhanden und die Lautheitsfunktion fällt deutlich schneller. So ist bspw. bei 20 dB im Vergleich zu 40 dB von einer Absenkung der Lautheit um den Faktor 6,6 anstatt dem Faktor von 4 auszugehen (vgl. Abb. 7).[69] Mit Hilfe der Isophone kann nun auch die Lautheit für die restlichen Frequenzen ermittelt werden.[70]

[67] In Anlehnung an Dickreiter/Goeres-Petri (2014), S. 119.
[68] Vgl. Fastl/Zwicker (2007), S. 206.
[69] Vgl. Fastl/Zwicker (2007), S. 207.
[70] Vgl. Hellbrück (1993), S. 76.

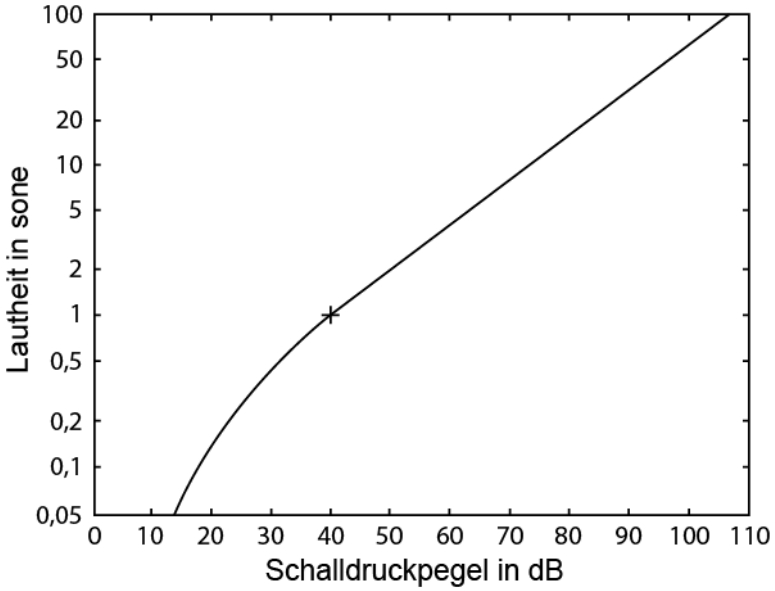

Abb. 7: Die Lautheitsfunktion für den 1 kHz Sinuston (Sone-Skala).[71]

4.3 Frequenzgruppen

Eine subjektiv wahrgenommene Lautheit kann noch immer stark von einem Lautstärkepegel in Phon oder einer Lautheit in Sone abweichen.[72] Das menschliche Ohr hat die Eigenschaft, Frequenzen, die nah beieinander liegen, in bestimmten Frequenzbändern, den so genannten Frequenzgruppen, zusammenzufassen.[73] Sind zwei subjektiv gleich laute Schalle nicht in einer Frequenzgruppe, dann addieren sich ihre Lautheiten zur Summenlautheit.[74] Zusammen werden sie also doppelt so laut wahrgenommen. Befinden sich die beiden Schalle hingegen in der gleichen Frequenzgruppe, so wirken sie weniger verstärkt.[75]

Im hörbaren Bereich ergeben sich insgesamt 24 Frequenzgruppen, die in einer Skala mit der Einheit Bark dargestellt werden.[76] Während die Frequenzgruppen unterhalb von 500 Hz einen konstanten Abstand ihrer Mittenfrequenzen von etwa 100 Hz besitzen, werden sie oberhalb zunehmend breiter.[77]

[71] In Anlehnung an Fastl/Zwicker (2007), S. 207.
[72] Vgl. Görne (2014), S. 123.
[73] Vgl. Hellbrück (1993), S. 79.
[74] Vgl. Hellbrück (1993), S. 79.
[75] Vgl. Hellbrück (1993), S. 79 f.
[76] Vgl. Hellbrück (1993), S. 80.
[77] Vgl. Hellbrück (1993), S. 80.

4.4 Timbre

Das Timbre, auch Klangfarbe genannt, beschreibt die Beschaffenheit eines Klanges. Es ermöglicht z. B., Instrumente, die den gleichen Ton mit der gleichen Lautstärke spielen, voneinander zu unterscheiden.[78] Dabei können nicht nur z. B. Streicher und Bläser, sondern auch bspw. die eine von der anderen Violine unterschieden werden. Das Timbre wird dabei von mehreren Parametern bestimmt. Dazu zählen u.a. die spektrale Zusammensetzung (z. B. Oberschwingungen), die Tonalität, der zeitliche Verlauf und die Atack- und Releasezeit eines Klanges.[79]

4.5 Tonhöhe

Eine wichtige Rolle bei der Musiktheorie und der Musikanalyse spielen die tonalen Eigenschaften der Klänge. Aus den Tonhöhen innerhalb eines Stückes entstehen unter anderem die Melodie, die Harmonien und die Tonart.[80] Die Tonhöhe kann mit der Grundfrequenz eines Klangs gleichgesetzt werden. Da bei gleichbleibender Grundfrequenz auch die Tonhöhe konstant ist, spielt es keine Rolle, ob sich die Wellenform eines Klangs ändert. Menschen können Tonhöhen normalerweise auch nur anhand der Struktur der Obertöne wahrnehmen, sollte die Grundfrequenz im Signal nicht oder nur leicht vorhanden sein.[81] So lassen sich bspw. beim Telefonieren männliche und weibliche Stimmen meist unterscheiden, obwohl die Grundfrequenzen durch starke Filterung oft nicht übertragen werden.

Aufgrund der Eigenschaften des menschlichen Innenohrs ist bei höheren Frequenzen ein höherer Grundfrequenzabstand als bei tiefen Frequenzen nötig, um den als gleich empfundenen Tonhöhenunterschied zu erzielen.[82] Unterhalb von 500 Hz entspricht eine Verdopplung der Frequenz auch der Verdopplung der wahrgenommenen Tonhöhe.[83] Oberhalb der 500 Hz ist diese Proportionalität jedoch nicht mehr gegeben.[84] So ist eine Verdopplung der empfundenen Tonhöhe eines Tones von z. B. 1,3 kHz erst

[78] Vgl. Zacharov (2018), S. 44.
[79] Vgl. Hoeg u. a. (2014), S. 1334.
[80] Vgl. Lerch (2012), S. 79.
[81] Vgl. Lerch (2012), S. 79.
[82] Vgl. Lerch (2012), S. 79.
[83] Vgl. Hellbrück (1993), S. 78.
[84] Vgl. Hellbrück (1993), S. 78.

bei einem Wert von 8 kHz erreicht.[85] Dieser nichtlineare Zusammenhang kann durch Skalen wie bspw. der sogenannten Mel-Skala angenähert werden.[86]

4.6 Schärfe und Rauheit

Wie scharf ein Audiosignal wahrgenommen wird, hängt hauptsächlich von der Umhüllenden des Spektrums und bei schmalbandigen Signalen von deren Mittenfrequenz ab.[87] Steigt die Anzahl an hohen Frequenzen im Signal bzw. steigt die Mittenfrequenz eines schmalbandigen Signals, so wirkt es schärfer.

Enthält ein Signal schnelle Pegelschwankungen, vor allem zwischen 20 Hz und 150 Hz, wird es als rau wahrgenommen.[88] Die Rauheit ist dabei bei ca. 70 Hz am größten und wird in der Einheit Asper angegeben.[89] Ein Sinuston von 1 kHz und 60 dB, der mit 70 Hz größtmöglich moduliert wird, hat dabei 1 asper.[90]

[85] Vgl. Hellbrück (1993), S. 78.
[86] Vgl. Hellbrück (1993), S. 79.
[87] Vgl. Dickreiter/Goeres-Petri (2014), S. 126.
[88] Vgl. Dickreiter/Goeres-Petri (2014), S. 125.
[89] Vgl. Dickreiter/Goeres-Petri (2014), S. 125.
[90] Vgl. Dickreiter/Goeres-Petri (2014), S. 125.

5 ANALYSEVERFAHREN FÜR AUDIOSIGNALE UND -SYSTEME

Um Audiosignale und deren Übertragungssysteme optimieren zu können, werden diese in den verschiedensten Weisen betrachtet und ausgewertet. So gibt es je nach Anwendungsfall Verfahren, mit denen sich Aussagen treffen lassen und passende Ergebnisse erzielt werden können. Aufgrund der Vielfalt an Audiosignalarten, wird im Folgenden zunächst auf Möglichkeiten der Klassifizierung eingegangen. Im Anschluss wird der Begriff Audiosystem geklärt und anhand eines Experiments gezeigt, wie ein solches modelliert werden kann. Zudem werden Verfahren aufgezeigt, die zur Qualitätsanalyse von Audiosignalen und -systemen genutzt werden können. Auch Möglichkeiten, Schallfelder zu betrachten werden erläutert. Der Schwerpunkt dieses Abschnittes liegt auf der Transformation von Audiosignalen aus dem Zeit- in den Frequenzbereich und deren Anwendung in der Praxis.

5.1 Audiosignale

Signale dienen auch in der Audiotechnik zur Übertragung von Informationen. Sie können dabei durch akustische, elektrische oder elektromagnetische Schwingungen übertragen werden.[91] Audiosignale werden häufig als Zeitfunktion betrachtet, wobei deren Informationsparameter (wie z. B. Amplitude, Phase oder Frequenz) den Wertverlauf einer physikalischen oder technischen Größe bilden.[92]

5.1.1 Kontinuierliche und diskrete Signale

Signale lassen sich in zeitkontinuierliche und zeitdiskrete Signale klassifizieren. Bei zeitkontinuierlichen Signalen ist für jeden Zeitpunkt ein Signalwert definiert. Dies ist bei analogen Signalen der Fall. Durch A/D-Wandler können analoge Signale in digitale (zeitdiskrete) Signale umgewandelt werden. Dazu wird mit einer gewissen Rate das analoge Eingangssignal abgetastet, und somit den Werten an gewissen Zeitpunkten Zahlen zugeordnet. Entsprechend können Signale zwischen wertkontinuierlich und wertdiskret unterschieden werden. Wie in Abb. 8 dargestellt, kann bei Signalen also zwischen werte- und zeitkontinuierlich (analoges Signal), wertekontinuierlich und

[91] Vgl. Görne (2014), S. 135 f.
[92] Vgl. Beucher (2015), S. 16.

zeitdiskret, wertediskret und zeitkontinuierlich (digitales Signal in der Schalttechnik) und werte- und zeitdiskret (digitales Signal) unterschieden werden. Digitale, also wert- und zeitdiskrete Signale werden in der Praxis nicht wie in Abb. 8 gezeigt, sondern in binärer Form gespeichert und verarbeitet.

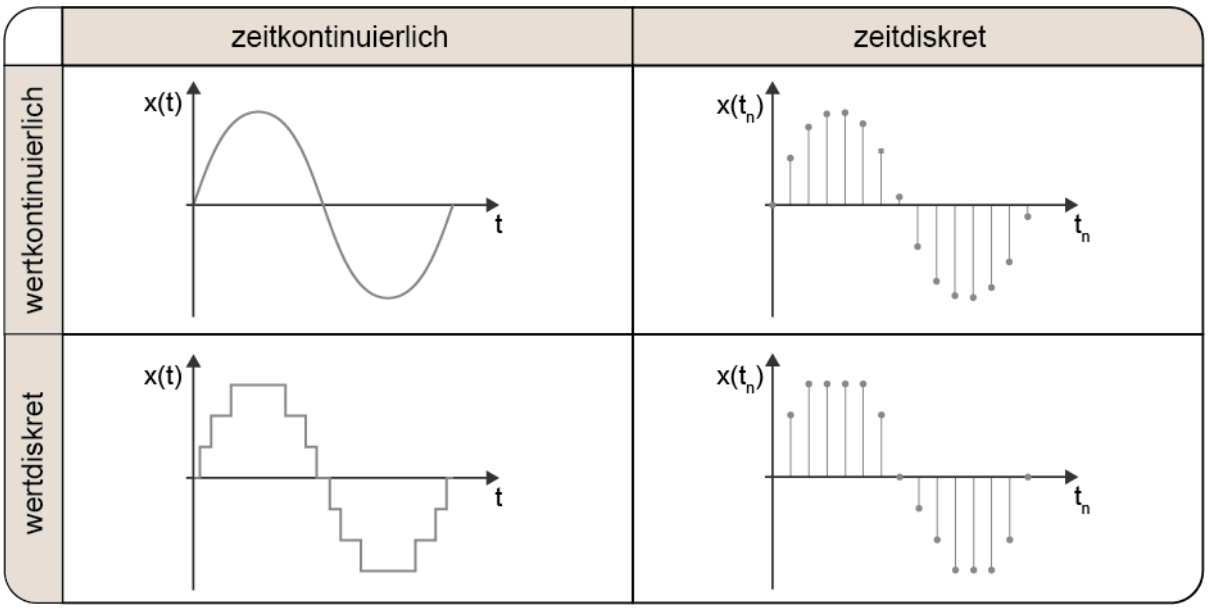

Abb. 8: Klassifizierung von Signalen.[93]

5.1.2 Stochastische und determinierte Signale

Bei einem stochastischen Audiosignal wird jedem Zeitpunkt eine Menge an möglichen Werten zugeordnet, aus der dann weitgehend zufällig ein bestimmter Wert festgelegt wird.[94] Diese Zufälligkeit kann mit Methoden der Wahrscheinlichkeitsrechnung dargestellt werden.[95] Bei determinierten Signalen hingegen ist jedem Zeitpunkt ein eindeutiger Zahlenwert zugeordnet, wodurch diese mit Funktionen beschrieben werden können.[96]

[93] In Anlehnung an Hoffmann/Wolff (2014), S. 5.
[94] Vgl. Meffert/Hochmuth (2018), S. 20.
[95] Vgl. Meffert/Hochmuth (2018), S. 20.
[96] Vgl. Meffert/Hochmuth (2018), S. 20.

5.1.3 Energie- und Leistungssignale

Signale können auch in Energie- und Leistungssignale eingeteilt werden. Dabei berechnet sich die Energie E_x eines deterministischen zeitkontinuierlichen Signals $x(t)$ durch

$$E_x = \int_{-\infty}^{\infty} |x(t)|^2 \, dt \,.^{97} \tag{6}$$

Von einem Energiesignal ist dann die Rede, wenn E_x endlich ist.[98] Stattdessen ist $x(t)$ ein Leistungssignal, wenn die Energie unendlich, aber die mittlere Leistung

$$P_x = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} |x(t)|^2 \, dt \tag{7}$$

endlich ist.[99] Entsprechend gilt für die Energie eines zeitdiskreten Signals $x(t_n)$

$$E_x = \sum_{t_n=-\infty}^{\infty} |x(t_n)|^2 \tag{8}$$

und für die mittlere Leistung

$$P_x = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{t_n=-N}^{N} |x(t_n)|^2 \,.^{100} \tag{9}$$

5.1.4 Stationäre und instationäre Signale

Ein Signal, dessen statistische Eigenschaften nicht von der Zeit abhängen, wird als stationär bezeichnet.[101] Die Betrachtung und Verarbeitung von stationären Signalen ist meist deutlich einfacher, wodurch einige Verfahren ausschließlich auf stationäre Signale anwendbar sind.[102] Sind die statistischen Eigenschaften eines Signals je nach Betrachtungszeitpunkt verschieden, gelten diese als instationär.

---

[97] Vgl. Mertins (2020), S. 7.
[98] Vgl. Mertins (2020), S. 7.
[99] Vgl. Mertins (2020), S. 7.
[100] Vgl. Mertins (2020), S. 7 f.
[101] Vgl. Meffert/Hochmuth (2018), S. 20.
[102] Vgl. Meffert/Hochmuth (2018), S. 20.

5.2 Audiosysteme

In der Audiotechnik können Signale mit den unterschiedlichsten Dingen übertragen und transformiert werden. Diese Dinge, bzw. Prozesse werden als Systeme bezeichnet.[103] So gelten auch alle Geräte der Audiotechnik, wie z. B. Mikrofon, Verstärker, Filter oder Lautsprecher, als System.[104] Wird ein beliebiges Eingangssignal, z. B. die Zeitfunktion $x(t)$, von einem System verarbeitet, wird das Ausgangssignal $y(t)$ als die Systemantwort bezeichnet (vgl. Abb. 9).[105]

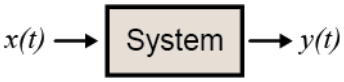

Abb. 9: Systemantwort y(t) auf das Eingangssignal x(t).[106]

Bewirken Systeme höchstens lineare Verzerrungen, also Pegeländerungen und Änderungen der Phasenlage des Signals, so werden sie als lineare zeitinvariante Systeme (LTI-Systeme) bezeichnet.[107] Dabei müssen die folgenden Anforderungen erfüllt sein:

1. Für die Linearität muss das Superpositionsgesetz erfüllt sein. Dafür muss die Systemantwort auf zwei addierte Signale der Summe der beiden einzelnen Systemantworten entsprechen.[108]
2. Für die Erfüllung der Zeitinvarianz muss bei einer zeitlichen Verschiebung des Eingangssignals entsprechend auch die Systemantwort verschoben sein.[109]

Eine Schwingung mit der komplexen Schreibweise $x(t) = \mathrm{e}^{\mathrm{j}\omega t}$ wird durch ein LTI-System höchstens in Amplitude und Phase verändert.[110] Dies wird mit dem Komplexen Skalierungsfaktor H, der auch eine mögliche Phasendrehung beinhaltet, wie in Formel (10) angegeben:[111]

$$x(t) = \mathrm{e}^{\mathrm{j}\omega t} \quad \Rightarrow \quad y(t) = H\mathrm{e}^{\mathrm{j}\omega t} \tag{10}$$

[103] Vgl. Görne (2014), S. 136.
[104] Vgl. Görne (2014), S. 136.
[105] Vgl. Görne (2014), S. 136.
[106] In Anlehnung an Görne (2014), S. 136.
[107] Vgl. Görne (2014), S. 136.
[108] Vgl. Görne (2014), S. 137.
[109] Vgl. Görne (2014), S. 137.
[110] Vgl. Görne (2014), S. 136 f.
[111] Vgl. Görne (2014), S. 137.

Daher sind die harmonischen Schwingungen die Eigenfunktionen der LTI-Systeme.[112] Wie gut Systemverhalten analysiert oder entworfen werden können, ist abhängig von der mathematischen und simulativen Modellierbarkeit des technischen Systems.[113] LTI-Systeme sind aufgrund ihrer Vorhersagbarkeit besonders geeignet. Obwohl Systeme in der Praxis niemals ganz linear und zeitinvariant sind, können Nichtlinearitäten bei geeigneter Betrachtung oftmals vernachlässigt und die Systeme dennoch als LTI-Systeme behandelt werden.[114] Dies ist oft die einzige Möglichkeit Systeme mit Nichtlinearitäten mathematisch zu betrachten.

5.2.1 Dirac-Stoß

Ein wichtiges theoretisches Hilfsmittel bei der Beschreibung von Systemen stellt der Dirac-Stoß dar. Für dessen Erläuterung wird zunächst ein Rechtecksignal mit der Höhe $1/T_0$ und der Breite T_0 betrachtet (vgl. Abb. 10).[115] Durch $T_0 \cdot 1/T_0 = 1$ ist die Fläche stets endlich.

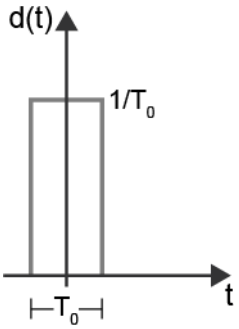

Abb. 10: Rechtecksignal d(t) der Breite T_0, der Höhe $1/T_0$ und der Fläche 1.[116]

Mit der Summe solcher Rechtecke $d(t)$ lässt sich eine beliebige Zeitfunktion $x(t)$ wie folgt approximieren:[117]

$$x(t) \approx x_A(t) = \sum_{n=-\infty}^{\infty} x(nT_0)\, d(t - nT_0)\, T_0 \qquad (11)$$

[112] Vgl. Görne (2014), S. 136.
[113] Vgl. Beucher (2015), S. 24 f.
[114] Vgl. Görne (2014), S. 137.
[115] Vgl. Görne (2014), S. 137.
[116] In Anlehnung an Görne (2014), S. 127.
[117] Vgl. Görne (2014), S. 137.

Dabei ist $d(t - nT_0)T_0$ ein um nT_0 nach rechts verschobenes Rechteck mit der Höhe 1. Die Amplitude des verschobenen Rechtecks ist $x(nT_0)$, also der Wert von $x(t)$ an der Stelle nT_0.

Durch das Superpositionsgesetz und der Zeitinvarianz lässt sich bei LTI-Systemen die Systemantwort $y_A(t)$ auf die Approximation $x_A(t)$ als Summe der einzelnen Rechteck-Antworten $y_d(t)$ darstellen:[118]

$$y(t) \approx y_A(t) = \sum_{n=-\infty}^{\infty} x(nT_0)\, y_d(t - nT_0)\, T_0 \tag{12}$$

Werden die Rechtecke jetzt unendlich schmal, wird die Approximation exakt. Um dies zu kennzeichnen, wird nT_0 durch τ und T_0 durch $d\tau$ ersetzt.[119] Solch ein Rechteck wird Dirac-Stoß (bzw. Dirac-Funktion) $\delta(t)$ genannt. Da weiterhin für die Fläche $T_0 \cdot 1/T_0$ gilt, hat der unendlich schmale Dirac-Stoß mit ebenfalls unendlicher Amplitude die Fläche 1. Wird nun die Summe durch ein Integral ersetzt und wird die Rechteck-Antwort $y_A(t)$ als Impulsantwort $h(t)$ bezeichnet, gilt für die Approximation des Eingangssignals durch die Dirac-Funktion

$$x(t) = \int_{-\infty}^{\infty} x(\tau)\, \delta(t - \tau)\, d\tau \tag{13}$$

und für das Ausgangssignal

$$y(t) = \int_{-\infty}^{\infty} x(\tau)\, h(t - \tau)\, d\tau.^{120} \tag{14}$$

5.2.2 Faltung

Das Integral aus Formel (14) wird auch Faltungsintegral genannt. Dessen Berechnung, bzw. die Operation $*$ in der Kurzschreibweise

$$y(t) = x(t) * h(t) \tag{15}$$

[118] Vgl. Görne (2014), S. 138.
[119] Vgl. Görne (2014), S. 138.
[120] Vgl. Görne (2014), S. 138.

ist die sogenannte Faltung der Signale.[121] Aus Formel (14) geht hervor, dass bei der Faltung das zweite Signal zunächst an der Zeitachse gespiegelt (t – τ), und beim anschließenden gegeneinander Verschieben alle Werte der Signale multipliziert werden.

5.2.3 Impulsantwort

Wie durch Formel (15) gezeigt, wird ein LTI-System vollständig durch dessen Impulsantwort $h(t)$ beschrieben. Ist also das Eingangssignal und die Impulsantwort eines LTI-Systems bekannt, kann durch die eben beschriebene Faltung das Ausgangssignal mathematisch bestimmt werden. Das Arbeiten mit Impulsantworten ist daher in der Praxis beim Analysieren von LTI-Systemen von großer Bedeutung. In der Audiotechnik lassen sich so unter anderem sogenannte Faltungshallgeräte realisieren.

5.2.4 Faltungshall

Wenn sich zwei Menschen in einem Raum (z. B. Kirche, Gewölbekeller oder Tunnel) unterhalten, kann dieser Raum auch als LTI-System betrachtet werden. Der Raum (bzw. das LTI-System) hat ganz bestimmte Eigenschaften, wie z. B. Form, Größe oder Beschaffenheiten des Materials, welche sich alle auf das Eingangssignal auswirken. Dabei stellt ein gesprochenes Wort der einen Person das Eingangssignal dar. Das Signal, das schließlich nach den Einwirkungen durch den Raum (z. B. Reflektionen und Echos) am Ohr der anderen Person ankommt, kann als Ausgangssignal betrachtet werden. Wie in Abschnitt 5.2.3 erläutert, können die akustischen Eigenschaften des Raumes analysiert werden, indem dessen Impulsantwort berechnet wird. Da in der Praxis die Herstellung, bzw. Wiedergabe eines Dirac-Stoßes (unendlich kurz mit unendlich hoher Amplitude) nicht umsetzbar ist, wird häufig ein lauter, breitbandiger Knall (z. B. Pistolenschuss oder Luftballonplatzen) genutzt. Wie folgender Versuch zeigt, kann auch ein Sinus-Sweep zur Erstellung der Impulsantwort eines Raumes dienen. Hierzu wurde zunächst mit der Funktion Zirpen von Audacity[122] ein 20 Sekunden langer Sinus-Sweep erstellt, welcher logarithmisch die Frequenzen von 20 Hz bis 20 kHz durchläuft. Dieser Sinus-Sweep wurde über einen Lautsprecher in einer kleinen Sporthalle (vgl. Abb. 11) abgespielt und mit einer XY-Anordnung von zwei Beyerdynamic M300N(C) aufgezeichnet. Der Aufbau befand sich zwischen 7-m-Linie und Mittellinie

[121] Vgl. Beucher (2015), S. 69.
[122] Mazzoni u. a.: Audacity. Version 3.0.0 (2021).

des Handballfeldes der Sporthalle, wobei der Abstand zwischen Lautsprecher und Mikrofonpaar etwa 8 m betrug.

Abb. 11: Sporthalle, deren Impulsantwort berechnet wurde.

Um eine Verbesserung des SNR zu erhalten, wurde die Halle in acht Durchläufen von dem Sinus-Sweep angeregt und jeweils das resultierende Signal aufgezeichnet. Durch Übereinanderlegen der acht Stereospuren konnte eine Anhebung des Nutzsignals um ca. 18 dB erzielt werden (je Verdopplung der Spurenanzahl ca. 6 dB). Da der Störanteil (z. B. durch Lüftung in der Halle oder Verkehr vor der Halle) in den einzelnen Aufzeichnungen verschieden ist, wurde dieser bei jeder Verdopplung der Spurenanzahl nur um etwa 3 dB angehoben. Somit konnte der SNR um knapp 9 dB vergrößert werden, was zu einer qualitativen Verbesserung der späteren Impulsantwort führte. Die anschließende Transformierung in eine Impulsantwort wurde mit dem in REAPER[123] enthaltenen Plugin ReaVerb durchgeführt. Hierzu musste nur der ursprüngliche Sweep und der Bounce der acht gelayerten Spuren selektiert werden, und das Plugin berechnete automatisch die resultierende Impulsantwort der Sporthalle. Diese konnte nun in einem weiteren Schritt in einem Faltungshall hinterlegt werden. So war es möglich, jedes beliebige Signal mit der Impulsantwort zu falten und somit mit den Eigenschaften des Raumes zu versehen. Um das Resultat zu testen, wurde hier ein möglichst trockenes Snare Sample gewählt. Als Referenz diente eine Aufnahme davon, wie dieses Sample über den Lautsprecher in der Sporthalle abgespielt, und ebenfalls mit der XY-

[123] Cockos Inc.: REAPER. Version 5.981 (2019).

Anordnung aufgenommen klang. Das Snare Sample wurde nun mit Hilfe des Convolution Reverbs von Ableton Live Suite[124] mit der zuvor gewonnenen Impulsantwort der Sporthalle gefaltet. Durch die Regler innerhalb des Plugins konnten nachträglich Anpassungen in der Räumlichkeit des Klangs durchgeführt werden. Beim Vergleich zwischen der Aufnahme in der Halle und der Variante mit Faltungshall fiel auf, dass der Klang und die Länge der Hallfahnen tatsächlich sehr ähnlich sind.
Bei der Wiedergabe des Sinus-Sweeps, wurde die Vorrichtung, auf welcher der Lautsprecher positioniert war, bei einer Frequenz von etwa 220 Hz angeregt und zum Vibrieren gebracht. Es wurde vermutet, dass darauf der in der Hallfahne des Faltungshalls enthaltene tonale Anteil (mit der Grundfrequenz in diesem Frequenzbereich) zurückzuführen ist. Die in diesem Experiment entstandenen Audiodateien sind auf der beigelegten CD abgespeichert und können gerne verglichen werden.
Das Besondere bei einem Faltungshall ist der sehr natürliche Klang, der, wie im vorangegangenen Beispiel gezeigt, die einzigartigen Eigenschaften bestimmter Orte wiedergeben kann. Wurde bspw. die Impulsantwort einer Kirche aufgezeichnet, kann eine Szene einer Sounddesignmischung mit der Raumakustik dieser Kirche versehen werden, ohne dass die Sprecher tatsächlich vor Ort aufgenommen werden müssen. Bei vielen Plugins besteht außerdem die Möglichkeit, den Effekt mit bestimmten Parametern wie z. B. Dry/Wet, Stereobreite oder Entfernung anzupassen, wodurch noch mehr Freiheiten in der Gestaltung entstehen. Mit dieser Vorgehensweise können auch analoge Effektgeräte analysiert und anschließend digital simuliert werden, wobei jedoch schnell die Grenzen der Möglichkeiten erreicht werden. Aufgrund der Eigenschaften der Faltung ist diese Vorgehensweise ausschließliche für LTI-Systeme anwendbar und kann bei nichtlinearen Manipulationen, wie Verzerrungen und Kompression oder zeitvarianten Systemen wie Chorus oder Flanger, nicht eingesetzt werden.

5.3 Verzerrungen

Veränderungen, die durch einen Verarbeitungsschritt bei Audiosignalen entstehen, werden Verzerrungen genannt.[125] Dabei gelten die Zeitverschiebung oder proportionale Amplitudenänderungen, wie Verstärkung oder Dämpfung des Signals, nicht als Verzerrung.[126] Um ungewünschte negative Auswirkungen von analogen

[124] Ableton: Ableton Live Suite. Version 11.0.2 (2021).
[125] Vgl. Dickreiter (2014), S. 617.
[126] Vgl. Dickreiter (2014), S. 617.

Übertragungswegen auf die Klangtreue zu vermeiden, wird versucht, die Verzerrung des Signals so gering wie möglich zu halten. Wie stark ein Signal durch einen Übertragungsweg verändert wird, kann durch unterschiedliche Messungen der einzelnen Übertragungskomponenten untersucht werden. Dabei ist zwischen linearen und nichtlinearen Verzerrungen zu unterscheiden.

5.3.1 Lineare Verzerrungen

Bei linearen Verzerrungen kommt es zu frequenzabhängigen Amplitudenänderungen.[127] Diese können sich unter Umständen negativ auf den Klang des Audiosignals auswirken.[128] Kommt es bspw. in der Aufnahmekette einer Gitarrenaufnahme zu linearen Verzerrungen in Form von Dämpfung hoher Frequenzen, wirkt der Klang des Instrumentes dumpfer. Um lineare Verzerrungen, die durch Systeme verursacht werden, zu messen, werden nacheinander am Eingang verschiedene Frequenzen angelegt und am Ausgang gemessen. Wird dies für alle Frequenzen im hörbaren Bereich durchgeführt, ist der sogenannte Frequenzgang des Systems das Ergebnis. Mit diesem kann dann z. B. die lineare Verzerrung eines Mikrofons oder eines Lautsprechers veranschaulicht werden.

5.3.2 Nichtlineare Verzerrungen

Während das Ausgangssignal bei einer linearen Verzerrung die gleichen Frequenzen wie beim Eingangssignal besitzt, kommt es bei nichtlinearen Verzerrungen im Ausgangssignal zu Frequenzen, die im Eingangssignal nicht enthalten sind. Diese neu entstandenen Frequenzen sind jeweils ganzzahlige Vielfache einer im Eingangssignal enthaltenen Frequenz und werden Harmonische oder Obertöne genannt.[129] Um die Anteile von nichtlinearen Verzerrungen im Ausgangssignal eines Systems bzw. eines Audiogerätes zu quantifizieren, können verschiedene Messverfahren angewendet werden.

[127] Vgl. Dickreiter (2014), S. 617 ff.
[128] Vgl. Dickreiter (2014), S. 619.
[129] Vgl. Dickreiter (2014), S. 620.

Total Harmonic Distortion

Betrachtet man die Frequenzen, die in einem Rechtecksignal der Grundschwingung 220 Hz enthalten sind, ist im Spektrum eine sinusförmige Grundschwingung mit einer Frequenz von 220 Hz zu erkennen. Außerdem sind in dem Rechtecksignal Obertöne mit dem 3-, 5-, 7-, 9-, 11-fachen usw. der Grundfrequenz enthalten. Die Total Harmonic Distortion (THD) in der Audiotechnik gibt dabei das Verhältnis (in Prozent) zwischen der Summe der einzelnen Effektivspannungen aller Obertöne und der Effektivspannung der Grundschwingung an, und kann mit

$$THD = \frac{\sqrt{{U_2}^2 + {U_3}^2 + {U_4}^2 + \cdots + {U_n}^2}}{U_1} \cdot 100\ \% \tag{16}$$

bestimmt werden.[130] Dabei steht U_n für die Effektivspannung der n-ten Oberschwingung und U_1 für die Effektivspannung der Grundschwingung. Audiosysteme sind in der Praxis nicht exakt linear, sondern führen zu nichtlinearen Verzerrungen in Form von harmonischen Anteilen im Ausgangssignal. Werden diese Verzerrungen zu stark, kann es zu einer ungewollten Änderung des Timbres führen. Um die Qualität eines Audiosystems beurteilen zu können, kann es daher von Interesse sein, die Stärke dieser Verzerrungen mit Hilfe der THD zu analysieren. Dazu wird bspw. ein Sinuston (da dieser keine signifikanten harmonischen Anteile besitzt) in das zu analysierende System gegeben.[131] Das Ausgangssignal ist schließlich der Sinuston mit den durch das System verursachten Verzerrungen. Wird nun das Eingangssignal vom Ausgangssignal subtrahiert, bleiben nur die vom System hinzugefügten harmonischen Anteile übrig.[132] Deren einzelne Effektivspannungen werden wie in Formel (16) aufsummiert und mit der Effektivspannung der Grundschwingung, also des Eingangssignals, ins Verhältnis gesetzt.[133] Um die Beeinflussung des Eingangssignals durch das Audiogerät noch genauer zu bestimmen, wird häufig auch die Total Harmonic Distortion Plus Noise (THD+N) angegeben, wobei neben den nichtlinearen Verzerrungen auch das Grundrauschen des Geräts miteinbezogen wird.[134]

[130] Vgl. Slone (2002), Abschn. Harmonic Distortion.
[131] Vgl. Slone (2002), Abschn. Harmonic Distortion.
[132] Vgl. Slone (2002), Abschn. Harmonic Distortion.
[133] Vgl. Slone (2002), Abschn. Harmonic Distortion.
[134] Vgl. Slone (2002), Abschn. Harmonic Distortion.

Klirrfaktor

Ähnlich wie die THD gibt auch der Klirrfaktor k (in Prozent) an, wieviel an ungewollter, nichtlinearer Verzerrung zum Eingangssignal dazugekommen ist. Im Unterschied dazu werden die Effektivspannungen der entstehenden Harmonischen jedoch mit der Effektivspannung des verzerrten Gesamtsignals U_{ges} in das Verhältnis

$$k = \frac{\sqrt{{U_2}^2 + {U_3}^2 + {U_4}^2 + \cdots + {U_n}^2}}{U_{ges}} \cdot 100\ \% \tag{17}$$

gesetzt.[135] Klirrfaktor und THD sind für kleine Verzerrungen (heute üblich) sehr ähnlich und werden deshalb oft fälschlicherweise als Synonyme verwendet. Werden die Verzerrungen jedoch stärker, weichen auch die beiden Messgrößen deutlicher voneinander ab. Bei Geräten in der Audiotechnik, wie z. B. Vorverstärkern, Mikrofonen, Audio Interfaces, Mischpulten oder Lautsprechern, werden möglichst geringe Klirrfaktoren gewünscht (z. B. 0,002%), um eine möglichst hohe Klangtreue zu erzielen.

Der sogenannte Teilklirrfaktor k_{n} gibt das Verhältnis der Effektivspannung einer einzelnen Harmonischen zur Effektivspannung des verzerrten Gesamtsignals U_{ges} an:[136]

$$k_n = \frac{U_n}{U_{ges}} \tag{18}$$

Dabei ist U_{n} die Effektivspannung der n-ten Harmonischen. Da höhere Harmonische das Messergebnis meist kaum noch beeinflussen, werden oft nur die Teilklirrfaktoren k_2 und k_3 berücksichtigt.[137]

Intermodulationsfaktor

Da die nichtlinearen Audiogeräte meist nicht nur einen einzelnen Sinuston, sondern ein Gemisch vieler Schwingungen übertragen, entstehen nicht nur die Harmonischen, sondern auch zusätzliche Kombinationstöne.[138]

[135] Vgl. Dickreiter (2014), S. 624.
[136] Vgl. Dickreiter (2014), S. 624.
[137] Vgl. Dickreiter (2014), S. 624.
[138] Vgl. Dickreiter (2014), S. 625.

Aus den Grundschwingungen oder deren ganzzahligen Vielfachen entstehen Summentöne f_S oder Differenztöne f_D:[139]

$$f_S = f_H + f_T \qquad \text{und} \qquad f_D = f_H - f_T \tag{19}$$

mit f_H für die Frequenz des höheren Tons und f_T für die Frequenz des tieferen Tons. Zur Messung des Intermodulationsfaktors m werden mit einem Tongenerator zwei sinusförmige Signale mit den Frequenzen f_H und f_T mit einem Spannungsverhältnis von 1:4 in das zu analysierende Audiogerät gegeben.[140] Die Frequenz f_H wird auf 4 kHz oder 5kHz eingestellt, und für f_T wird meist frei eine tiefere Frequenz gewählt.[141] Nun werden die Effektivspannungen der am Ausgang neu entstehenden Summen- und Differenztöne aufsummiert und mit dem Effektivwert der Spannung des höheren Messtons f_H ins Verhältnis gesetzt:[142]

$$m_2 = \frac{U_{(f_H - f_T)} + U_{(f_H + f_T)}}{U_{f_H}} \cdot 100\ \% \tag{20}$$

Hier wird der Intermodulationsfaktor als m_2 bezeichnet, da es sich um die 2. Ordnung handelt. Aufgrund der Tatsache, dass auch Vielfachen der Frequenz f_T in Kombinationstönen mit f_H resultieren, können auch Intermodulationsfaktoren höherer Ordnung bestimmt werden.[143] So wird sich bspw. für den Intermodulationsfaktor der 3. Ordnung m_3 auf die Frequenzpaare $f_H \pm 2\,f_T$ beschränkt.[144]

Der Intermodulationsfaktor ist aufgrund der Berücksichtigung dieser zusätzlichen Schwingungen genauer als der zuvor beschriebene Klirrfaktor, wird aber aufgrund der aufwendigen Messung nur selten angewendet.[145]

Differenztonfaktor

Während der Klirrfaktor vor allem für tiefe und mittlere Frequenzen und der Intermodulationsfaktor nur für mittlere Frequenzen aussagekräftig ist, gibt es mit der Bestimmung des Differenztonfaktors ein Messverfahren für die entstehende nichtlineare

[139] Vgl. Dickreiter (2014), S. 625.
[140] Vgl. Dickreiter (2014), S. 625.
[141] Vgl. Dickreiter (2014), S. 625.
[142] Vgl. Dickreiter (2014), S. 625.
[143] Vgl. Dickreiter (2014), S. 625.
[144] Vgl. Dickreiter (2014), S. 625.
[145] Vgl. Dickreiter (2014), S. 625.

Verzerrung bei mittleren und hohen Frequenzen.[146] Wie der Intermodulationsfaktor ist auch der Differenztonfaktor ein Messverfahren, bei dem zwei sinusförmige Signale in das betrachtete Audiogerät eingespeist werden. Deren Frequenzabstand soll dabei 70 Hz betragen.[147] Der Differenzfaktor d gibt dann das Verhältnis zwischen der Effektivspannung derer Differenztöne und dem Effektivwert der Spannung des gesamten Ausgangssignals an.[148]

5.4 Zwei- oder mehrkanalige Audiosysteme

Eine Methode, um die Ähnlichkeit von zwei Signalen zu untersuchen, ist die der Korrelationsfunktionen. Dabei ist die sogenannte Kreuzkorrelationsfunktion (KKF) zweier Signale ähnlich der Faltung, nur wird das zweite Signal nicht an der vertikalen Achse gespiegelt.[149] Wird ein Signal mit sich selbst korreliert, ist von der Autokorrelationsfunktion (AKF) die Rede.[150] Die Fourier-Transformierte[151] der AKF führt zu dem sogenannten Leistungsdichtespektrum (LDS), welches die Verteilung der Leistung in Abhängigkeit der Frequenz beschreibt.[152] Genutzt werden können solche Korrelationsanzeigen unter anderem, um die Beziehung zwischen zwei Audiosignalen (z. B. rechtes und linkes Signal bei Stereo) messtechnisch darzustellen.[153] Solch eine Korrelationsanzeige (auch Korrelationsgradmesser genannt) wird auf einer Skala angezeigt und dient vor allem der Beurteilung der Mono-Kompatibilität einer Mischung.[154] Die Skala ist dabei auf Werte von –1 bis +1 normiert. Ist der Korrelationsgrad genau +1 gibt es keine Phasenverschiebung und die beiden Signale sind identisch. Bei einem Wert von 0 sind beide Signale komplett unterschiedlich oder es liegt nur auf einem Kanal ein Signal an.[155] Liegt der Korrelationsgrad dazwischen, herrscht eine Phasenverschiebung zwischen 90°und 0°, wodurch das Stereosignal mehr oder weniger monokompatibel ist.[156] Ist dieser jedoch kleiner als 0, kommt es zu Phasenverschiebungen zwischen 90° und 180° und bei einer Wiedergabe in mono zu teilweisen oder

[146] Vgl. Dickreiter (2014), S. 623.
[147] Vgl. Dickreiter (2014), S. 625.
[148] Vgl. Dickreiter (2014), S. 625.
[149] Vgl. Görne (2014), S. 139.
[150] Vgl. Mertins (2020), S. 77 f.
[151] Nähere Ausführungen dazu finden sich in Abschn. 5.7.2.
[152] Vgl. Mertins (2020), S. 198.
[153] Vgl. Hoeg u. a. (2014), S. 1306.
[154] Vgl. Hoeg u. a. (2014), S. 1306.
[155] Vgl. Hoeg u. a. (2014), S. 1307.
[156] Vgl. Hoeg u. a. (2014), S. 1307.

ganzen Auslöschungen im Gesamtsignal, was bei der Musikproduktion nicht wünschenswert ist.[157] Ein Korrelationsgradmesser ermöglicht zwar das Erkennen einer Verpolung der Stereosignale, ist aber dennoch kein eindeutiges Kriterium für die Kompatibilität einer Mischung.[158] Daher wird eine Korrelationsanzeige häufig in Kombination mit einem sogenannten Stereosichtgerät (SSG) verwendet. Mit einem SSG wird die Beziehung der beiden Kanäle in grafischer Form, auch als Lissajous-Figur bezeichnet, dargestellt.[159] Diese ändert sich dynamisch in Abhängigkeit des Signalverlaufes und ermöglicht es, Informationen zu Seitenzuordnung, Richtungsverteilung, Stereobalance, Phasenlage oder die Kompatibilität einer Stereomischung abzuleiten.[160] Beträgt der Wert im Korrelationsgradmesser +1, also liegen auf beiden Kanälen identische Signale, bildet die Lissajous-Figur eine mittige, vertikale Linie (vgl. Abb. 12 (a)). Liegt nur auf dem linken Kanal ein Signal an, wird wie in Abb. 12 (b) eine um 45° nach links gekippte Linie angezeigt (bzw. bei einem nur rechts anliegendem Signal entsprechend nach rechts gekippt). Beträgt der Wert des Korrelationsgradmessers hingegen –1, zeigt das SSG eine horizontale Linie. Auch ob ein Stereosignal zu breit oder zu monoartig ist, kann abgelesen werden. Abb. 12 (c) zeigt bspw. eine Lissajous-Figur, die weder zu breit noch zu schmal ist und sich damit gut zur Wiedergabe in Mono eignet. Wird bei einem der Kanäle die Phase um 180° gedreht, ist der Korrelationsgradmesser nahe der –1 und im SSG ergibt sich ein mit Abb. 12 (d) vergleichbares Bild. Solch eine Mischung fällt in mono meist völlig in sich zusammen und sollte bei der Musikproduktion vermieden werden.

[157] Vgl. Hoeg u. a. (2014), S. 1307.
[158] Vgl. Hoeg u. a. (2014), S. 1307.
[159] Vgl. Hoeg u. a. (2014), S. 1307.
[160] Vgl. Hoeg u. a. (2014), S. 1308 f.

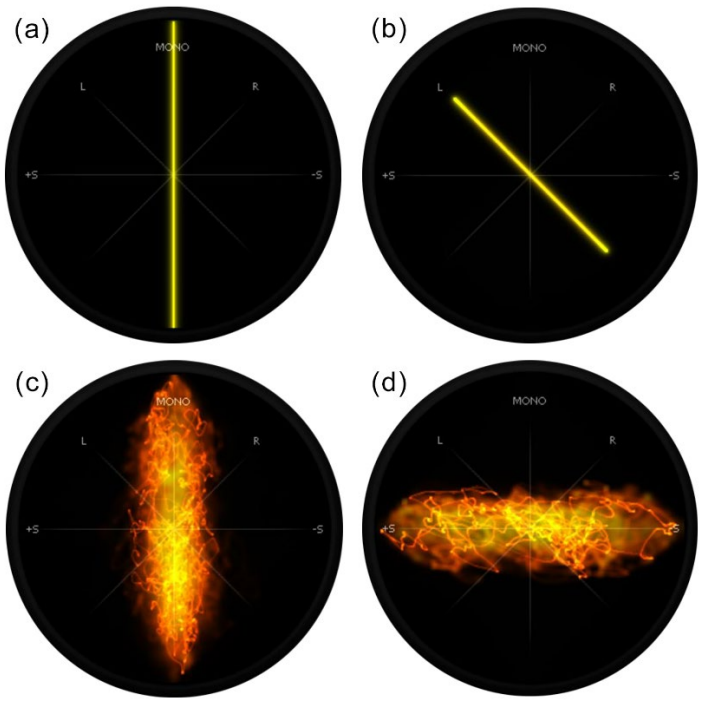

Abb. 12: Lissajous-Figuren für Mono-Signal (a), nur auf dem linken Kanal ein Signal (b), unproblematische Stereobreite (c) und rechter Kanal mit 180° Phasendrehung (d). Erstellt mit Stereo Tool V3.[161]

Auch mehrkanalige Audiosignale oder Surroundsysteme lassen sich durch ein sogenanntes Mehrkanalsichtgerät auf ihre Kompatibilität analysieren.[162] Dabei sind zur visuellen Klangkontrolle zusätzlich Parameter wie Richtungsgleichgewicht innerhalb des Front-Klangbildes, Balance zwischen Front- und Surroundsignalen, Kanalzuordnung und die Überprüfung der Stereo-Kompatibilität von Interesse.[163]

5.5 Lärmanalyse

Ab wann ein Schallereignis als störend eingestuft wird, ist vor allem eine subjektive Bewertung. Während für den einen bspw. ein beschleunigender Sportwagen ein Genuss ist, wird er von einem anderen bereits als belästigend wahrgenommen. Um Lärm dennoch einigermaßen angemessen quantifizieren zu können, müssen gehörrichtige Schallanalysen und Messverfahren angewendet werden. Dabei werden grundsätzliche Eigenschaften betrachtet, die dazu beitragen, dass ein Schallereignis allgemein als Lärm empfunden wird. Vor allem der Schalldruckpegel eines Schallereignisses ist entscheidend. Je höher dieser ansteigt, desto einheitlicher wird er als Lärm

[161] FLUX SOFTWARE ENGINEERING: Stereo Tool V3. Version 20.12.0.49880 (2020).
[162] Vgl. Hoeg u. a. (2014), S. 1309 ff.
[163] Vgl. Hoeg u. a. (2014), S. 1309.

empfunden.[164] Auch der Anteil an hohen Frequenzen ist ausschlaggebend dafür, wie störend ein Schallereignis wahrgenommen wird.[165] So wird bspw. quietschendes Styropor oder das Summen einer Mücke nahe des Ohres meist schneller als lästig eingestuft als niedrig frequentere Geräusche. Außerdem können auch die Rauigkeit oder vermehrte Schwankungen im zeitlichen Verlauf (z. B. in der Tonhöhe oder Frequenz) zur störenden Wirkung eines Schallereignisses beitragen.[166] Selbst die Tageszeit kann entscheidend sein.[167] So wirkt bspw. das Ticken einer Uhr in der Nacht u. U. belästigender als das Holzhacken des Nachbars tagsüber (obwohl Letzteres deutlich lauter ist). Welche Schallanalyse zur Bewertung des Lärms nun letztendlich angewendet wird, unterscheidet sich oft stark, je nach Quelle oder Umfeld (z. B. Fluglärm, Baustellenlärm oder öffentliche Veranstaltungen), um für jede Situation die bestmögliche Bewertung zu erstellen. Ein Messverfahren, welches in der Lärmbewertung häufig Anwendung findet, ist der A-bewertete Schalldruckpegel mit der Einheit dB(A). Diese Messung ist zwar daran angepasst, dass das menschliche Ohr für sehr hohe und sehr tiefe Frequenzen relativ unempfindlich ist, vernachlässigt dabei aber, dass diese Frequenzabhängigkeit gerade bei hohen Pegeln abnimmt.[168] Dadurch kann es u. U. zu gesundheitlichen Auswirkungen durch den Lärm kommen, welche aus der fehlerbehafteten Messung nicht abzuleiten waren. Zeitlich veränderliche Lärmsituationen lassen sich mittels energieäquivalentem Dauerschallpegel L_{eq} bewerten.[169] Dieser gibt den Pegel eines konstanten Lärms an, welcher in einem vorgegebenen Zeitraum die gleiche Wirkung wie der zeitlich veränderliche Lärm hervorruft.[170] Da jedoch bspw. ein intermittierendes Geräusch als störender empfunden werden kann als ein konstantes Geräusch, kann auch diese Messung zu Fehlern führen.[171] Bei der Analyse und Bewertung von Lärm ist es also sehr wichtig die Messungen bestmöglich an die Situation anzupassen, um Ungenauigkeiten zu vermeiden und um ein möglichst objektives Ergebnis zu erhalten.

[164] Vgl. Sommer u. a. (1998), S. 3.
[165] Vgl. Puls (2007), S. 11.
[166] Vgl. Puls (2007), S. 11.
[167] Vgl. Thießen/Schnorr (2005), S. 8.
[168] Vgl. Henn u. a. (2008), S. 194.
[169] Vgl. Henn u. a. (2008), S. 195.
[170] Vgl. Henn u. a. (2008), S. 195.
[171] Vgl. Henn u. a. (2008), S. 204 f.

5.6 Akustische Antennen

Soll die Zusammensetzung einer Lärmsituation aus mehreren Schallquellen genauer analysiert werden, können Messungen mit Mikrofonarrays durchgeführt werden. Diese Methode mit sogenannten akustischen Antennen wird auch als Beamforming bezeichnet und dient unter anderem zur Lokalisation von Schallquellen. Je nach Entfernung, die ein Schall von seiner Quelle bis zu dem Mikrofonarray zurücklegt, kommt es zu einer Verzögerung des Signals. So kommt es auch je nach Einfallswinkel zu unterschiedlichen Entfernungen der einzelnen Mikrofone zur Quelle und damit jeweils zu unterschiedlichen Verzögerungen innerhalb des Arrays. Durch Zeitverschiebung des vom jeweiligen Mikrofon aufgenommenen Schallsignals, kann die Mikrofonanordnung auf einzelne Messpunkte fokussiert werden.[172] Werden diese zeitlich verschobenen Schallsignale summiert, ergibt sich ein Zeitsignal für jeden Messpunkt.[173] Aus diesem können wiederum zeitlich gemittelte Größen wie der Schalldruckpegel oder das Leistungsdichtespektrum gewonnen werden.[174] Resultierende Heatmaps aus Farb- oder Grauwerten können anschließend mit einem Video oder einem 3D-Modell des Messobjekts kombiniert werden, um identifizierte Schallquellen und deren Eigenschaften anzuzeigen.[175] Dies wird dann auch als akustische Kamera bezeichnet. Dadurch lassen sich konkretere Rückschlüsse, wie genauer Ort oder enthaltene Frequenzen der Schallquelle, ermitteln. Beim Beamforming gibt es je nach Anforderungen verschiedene Möglichkeiten das Mikrofonarray anzuordnen und anschließend auszuwerten. So finden in der Praxis u.a. Kreuz- und X-Arrays, Ring-Arrays, Spiral-Arrays oder Kugel-Arrays Anwendung.[176] Beamforming kann bspw. effizient bei der Lärmbewertung von Schallereignissen[177], der Optimierung von Fahrzeug-[178] oder Raumakustik[179] oder zur Analyse bewegter Schallquellen[180] genutzt werden.

[172] Vgl. Möser (2010), S. 378.
[173] Vgl. Möser (2010), S. 378.
[174] Vgl. Möser (2010), S. 378 f.
[175] Vgl. Döbler/Jensen (2019), Abschn. Einleitung.
[176] Vgl. Möser (2010), Abschn. 6.4.
[177] Vgl. Alloza/Vonrhein (2019).
[178] Vgl. Meyer/Döbler (2006).
[179] Vgl. Kümmritz/Kerscher (2019).
[180] Vgl. Meyer/Döbler (2014).

5.7 Vom Zeit- in den Frequenzbereich

Betrachtet man ein Audiosignal ausschließlich im Zeitbereich kann man noch keine Aussagen über die enthaltenen Frequenzen treffen. Daher wurden verschiedene Möglichkeiten entwickelt, den Zeitbereich in den Frequenzbereich umzuwandeln.

5.7.1 Transformations-Analyse

In der Signalanalyse kann es hilfreich sein, eine Problemstellung zunächst zu transformieren, d. h. in eine andere Betrachtungsweise zu wandeln. Als Veranschaulichung dieser Methode dient das Beispiel in Abb. 13. Dabei soll die Problemstellung (Multiplikation von 123 und 234) ohne Rechner gelöst werden. Auf der linken Seite ist die konventionelle Herangehensweise schematisch dargestellt. Auf der rechten Seite wurde zunächst eine Transformation in die Form der sogenannten Linien-Multiplikation durchgeführt, wodurch im folgenden Schritt eine einfache Addition der entsprechenden Anzahl von Linien-Schnittpunkten durchgeführt werden kann. Die anschließende inverse Transformation führt letztendlich zur Lösung der Problemstellung. Die Berechnung von $123 \cdot 234$ kann somit durch eine einfache Addition, anstelle einer händischen Multiplikation, gelöst werden.

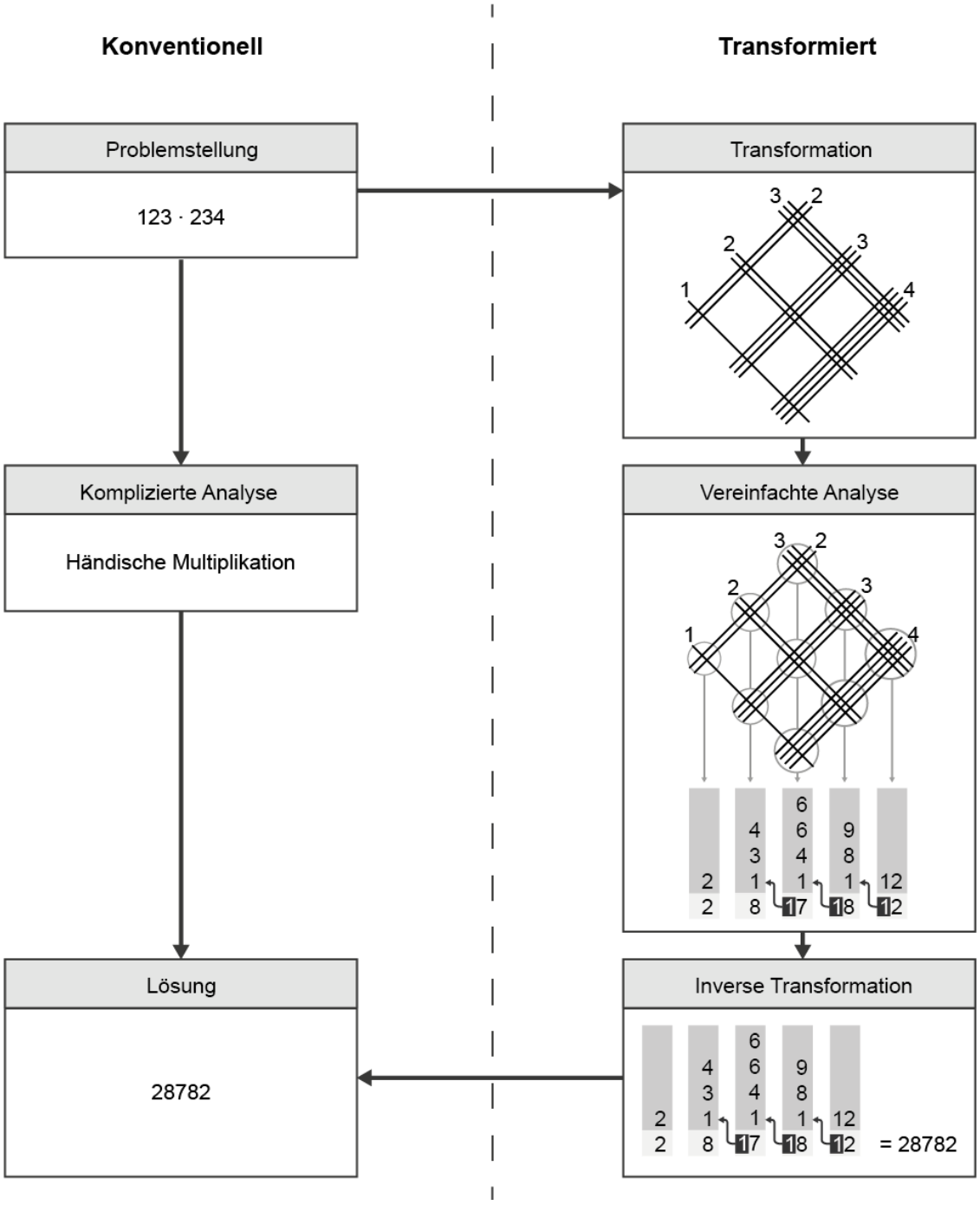

Abb. 13: Schematische Darstellung der konventionellen Analyse und der Transformations-Analyse.[181]

Transformationen können also zur Vereinfachung von Problemlösung genutzt werden. Eine dieser Transformationen ist die Fourier-Transformation, die im folgenden Abschnitt näher erläutert wird.

[181] In Anlehnung an Brigham (1987), S. 14.

5.7.2 Fourier-Transformation

Wie in Abschnitt 5.2.2 gezeigt, beschreibt die Faltung des Eingangssignals mit der Impulsantwort des Systems die Systemantwort eines LTI-Systems: $y(t) = x(t) * h(t)$. Die Beziehung zwischen LTI-System und Eigenfunktion (vgl. Formel (10)) lässt sich für eine Schwingung der Kreisfrequenz $\omega_0 = 2\pi f_0$ also darstellen als

$$e^{j\omega_0 t} * h(t) = H e^{j\omega_0 t}. \tag{21}$$

Durch Ausschreiben des Faltungsintegrals und Umformen der Gleichung, kann H bestimmt und die Funktion der Frequenz gebildet werden:[182]

$$H(f) = \int_{-\infty}^{\infty} h(t)\, e^{-j2\pi f t}\, dt \tag{22}$$

Dies wird als Fourier-Transformation $F\{h(t)\}$ bezeichnet. Aus der Funktion der Zeit (Impulsantwort $h(t)$) wird die Funktion der Frequenz (Übertragungsfunktion $H(f)$).[183] Dabei wird $H(f)$ auch als Fourier-Transformierte eines Signals bezeichnet. Durch die inverse Fourier-Transformation $F^{-1}\{H(f)\}$

$$h(t) = \int_{-\infty}^{\infty} H(f)\, e^{-j2\pi f t}\, df \tag{23}$$

lässt sich die Impulsantwort aus der Übertragungsfunktion zurückgewinnen.[184] Entsprechend gilt für jedes beliebige Signal x(t) die Fourier-Transformation

$$X(f) = \int_{-\infty}^{\infty} x(t)\, e^{-j2\pi f t}\, dt \quad \text{bzw.} \quad X(f) = F\{x(t)\} \tag{24}$$

und die inverse Fourier-Transformation

$$x(t) = \int_{-\infty}^{\infty} X(f)\, e^{-j2\pi f t}\, dt \quad \text{bzw.} \quad x(t) = F^{-1}\{X(f)\}.^{185} \tag{25}$$

[182] Vgl. Görne (2014), S. 142.
[183] Vgl. Görne (2014), S. 142.
[184] Vgl. Görne (2014), S. 143.
[185] Vgl. Görne (2014), S. 143.

Wie in Abb. 14 dargestellt, wird eine von −∞ bis ∞ definierte Funktion x(t) durch die Fourier-Transformation also in ihre sinusförmigen Anteile zerlegt. Die resultierende Fourier-Transformierte X(f) wird als Funktion der Frequenz mit entsprechenden Amplituden angegeben (vgl. Abb. 14 unten). Da negative Frequenzen bei realen Systemen keine Bedeutung haben, werden diese häufig weggelassen.

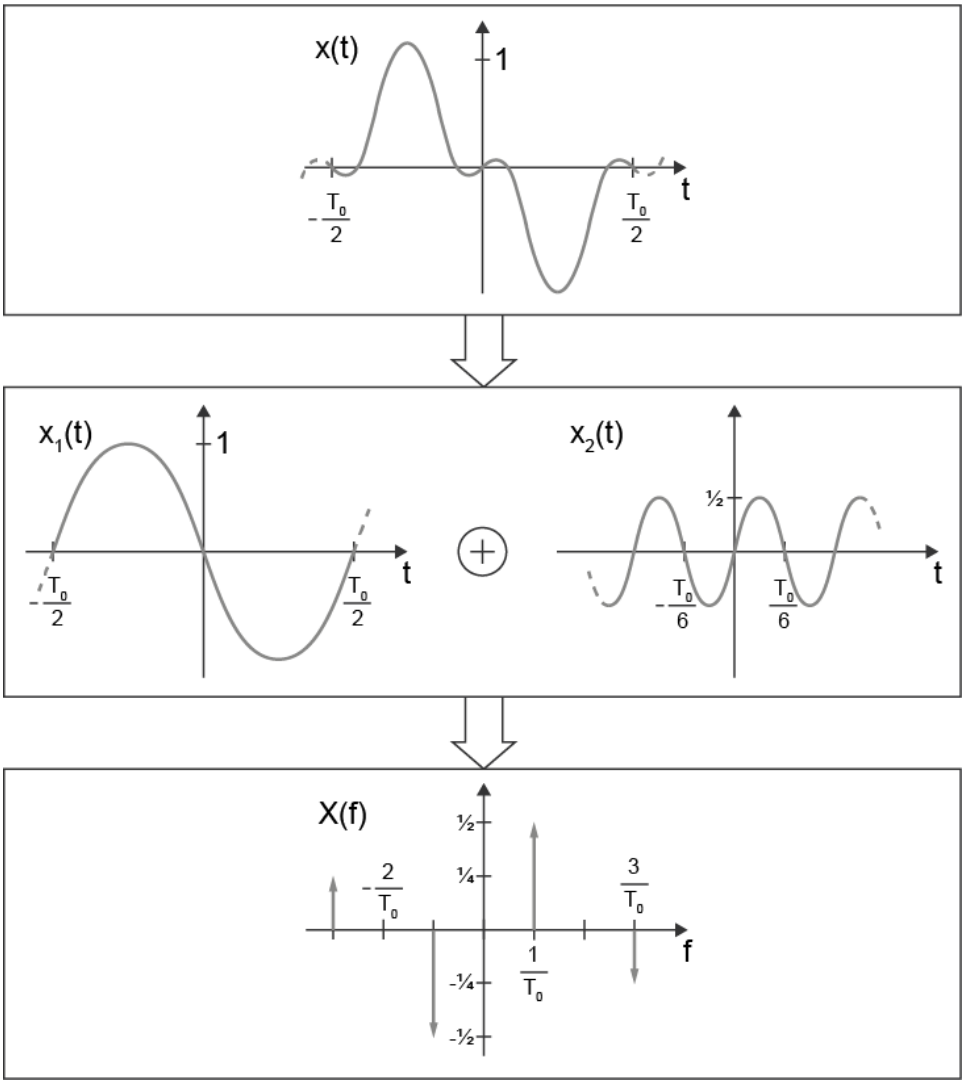

Abb. 14: Zerlegung eines Signals x(t) in seine sinusförmigen Komponenten.[186]

Während ein Signal im Zeitbereich durch Dirac-Stöße approximiert werden kann (vgl. Abschnitt 5.2.1), kann es im Frequenzbereich also durch eine Folge von harmonischen Schwingungen approximiert werden.[187] Das heißt, dass jedes beliebige Audiosignal aus einer Überlagerung von unendlich vielen sinusförmigen Funktionen generiert, bzw. in diese Bestandteile zerlegt werden kann.

[186] In Anlehnung an Brigham (1987), S. 16.
[187] Vgl. Görne (2014), S. 143.

Durch die Zerlegung der auch als Frequenzspektrum bezeichneten komplexen Funktion $F\{x(t)\} = X(f)$ in Betrag

$$|X(f)| = \sqrt{[Re\{X(f)\}]^2 + [Im\{X(f)\}]^2} \tag{26}$$

und Phase

$$\varphi(f) = \arctan\left(\frac{Im\{H(f)\}}{Re\{H(f)\}}\right) \tag{27}$$

entstehen das Amplituden- bzw. das Phasenspektrum.[188] So können auch Informationen über diese Signalparameter gewonnen werden.
$X(f)$ ist dabei zunächst ein Spektrum, welches die Leistung der enthaltenen Frequenzen wiedergibt. Wird der Logarithmus dessen gebildet, entsteht ein Spektrum, in dem die Amplituden (in dB) der einzelnen Frequenzen zu finden sind.

5.7.3 Diskrete Fourier-Transformation

Da Signale in der Digitaltechnik zeitdiskrete Folgen x(n) sind, wird aus dem Fourier-Integral die Summe der diskreten Fourier-Transformation (DFT)

$$F\{x(n)\} = X(k) = \sum_{n=0}^{M-1} x(n)\, e^{-\frac{\mathrm{j}2\pi nk}{M}} \qquad k = 0, \dots, M-1 \tag{28}$$

bzw. die inverse diskrete Fourier-Transformation

$$F^{-1}\{X(k)\} = x(n) = \frac{1}{M}\sum_{n=0}^{M-1} X(k)\, e^{\frac{\mathrm{j}2\pi nk}{M}} \qquad n = 0, \dots, M-1.^{189} \tag{29}$$

Da die DFT bevorzugt für periodische Signale eingesetzt wird, wird bei einem nicht periodischen Signal nur ein Signalausschnitt verwendet.[190] M entspricht dabei der Länge des Signalausschnittes bzw. der Anzahl, wieviele Abtastwerte betrachtet werden sollen.[191] Wird die DFT z. B. für den Ausschnitt M des Signals in Abb. 15 (oben)

[188] Vgl. Hoffmann/Wolff (2014), S. 107.
[189] Vgl. Görne (2014), S. 146.
[190] Vgl. Hoffmann/Wolff (2014), S. 118.
[191] Vgl. Hoffmann/Wolff (2014), S. 118.

durchgeführt, resultiert dies in dem Spektrum der periodischen Fortsetzung dieses Ausschnittes (vgl. Abb. 15 (unten)).[192]

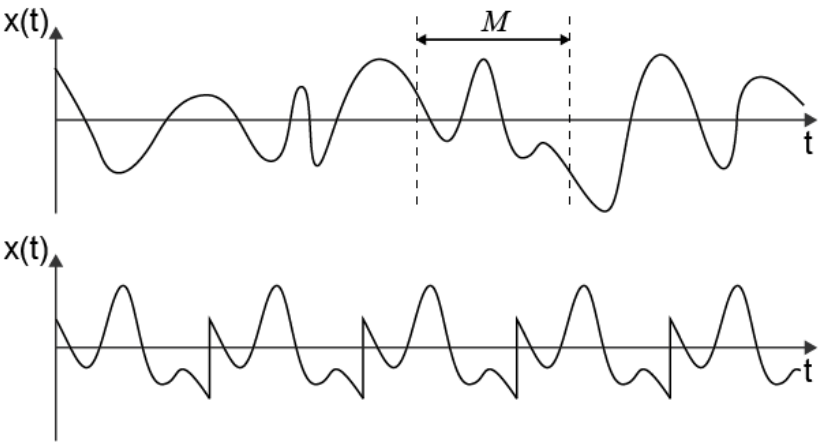

Abb. 15: Beispiel für die Anwendung der DFT. Für eine bessere Anschaulichkeit sind die Signale nicht als diskrete, sondern als kontinuierliche Kurven dargestellt.[193]

Aufgrund dieser Einteilung in Signalausschnitte und der damit verbundenen Einschränkung auf eine bestimmte Anzahl von betrachteten Werten, ist die DFT grundsätzlich fehlerbehaftet.[194] Der bei dieser Einteilung resultierende Fehler kann durch geeignete Fensterfunktionen[195] minimiert werden.[196] Wird nun die Formel aus (28) für M Werte mit der Substitution

$$\omega = e^{-\frac{j2\pi}{M}} \tag{30}$$

berechnet, kann die DFT auch als Matrix dargestellt werden:[197]

$$\begin{bmatrix} X(0) \\ X(1) \\ X(2) \\ \vdots \\ X(M-1) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{M-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(M-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{M-1} & \omega^{2(M-1)} & \cdots & \omega^{(M-1)(M-1)} \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ \vdots \\ x(M-1) \end{bmatrix} \tag{31}$$

Dabei werden die Koeffizienten des Eingangssignals als Vektor mit der DFT als Matrix multipliziert, um schließlich die Koeffizienten der Fourier-Transformierten (ebenfalls als Vektor) zu erhalten.[198] Da die Matrix durch die Substitution aus Formel (30) komplex

[192] Vgl. Hoffmann/Wolff (2014), S. 118.
[193] In Anlehnung an Hoffmann/Wolff (2014), S. 118.
[194] Vgl. Hoffmann/Wolff (2014), S. 134.
[195] Genaueres zu geeigneten Fensterfunktionen z. B. in Butz (2015), Abschn. 3.
[196] Vgl. Hoffmann/Wolff (2014), S. 134.
[197] Vgl. Brunton/Kutz (2017), S. 66.
[198] Vgl. Brunton/Kutz (2017), S. 66.

ist, ist auch der resultierende Vektor komplex.[199] Dabei entspricht jeder Koeffizient einer im Eingangssignal vorkommenden sinusförmigen Funktion, wobei sowohl die Frequenz als auch die Amplitude und Phase enthalten sind.[200] So können auch hier Frequenz-, Amplituden- und Phasenspektrum abgeleitet und für Analysen verwendet werden.

5.7.4 Schnelle Fourier-Transformation

Die Berechnung des resultierenden Vektors der DFT in Formel (31) wird mit steigendem M schnell aufwändiger. In der O-Notation lässt sich das Verhalten mit $O(M^2)$ beschreiben.[201] Hat bspw. ein Audiosignal von 100 Sekunden eine Samplerate von 44,1kHz, ergibt sich ein M von $4{,}41 \cdot 10^6$.[202] Somit müssen bei der DFT mit M^2 ungefähr 10^{13} Matrix-Multiplikationen berechnet werden, wodurch schnell viel Rechenleistung benötigt wird. Wie der Name vermuten lässt, wird dies durch den Algorithmus der schnellen Fourier-Transformation (FFT) stark vereinfacht. Bei der Berechnung der FFT gilt dabei $O(M \log_2(M))$, also ein sogenannter linear-logarithmischer Anstieg.[203] Daraus ergeben sich im Vergleich nur etwa 10^8 Matrix-Multiplikationen. Die Berechnung der DFT ist in diesem Beispiel um das ca. 100.000-fache aufwändiger als die der FFT. Diese enorme Vereinfachung wird vor allem durch eine geschickte Umstrukturierung der komplexen Multiplikation in Formel (31) gewonnen.[204]

5.7.5 Störspektrum

Durch eine FFT lässt sich ein Audiogerät auf mögliche Störsignale analysieren, die sich negativ auf das Audiosignal auswirken können. Für Abb. 16 wurde in einen Gitarrenverstärker ein Sinuston mit 997 Hz eingespeist, und nach anschließender A/D-Wandlung und einer FFT mit dem SPAN Plugin von Voxengo[205] das Spektrum dargestellt. Bei 997 Hz ist deutlich der Sinuston des Eingangssignals zu erkennen. Die in Abschnitt 5.3.2 beschriebenen nichtlinearen Verzerrungen des Tons durch die

[199] Vgl. Brunton/Kutz (2017), S. 66.
[200] Vgl. Brunton/Kutz (2017), S. 66.
[201] Vgl. Brunton/Kutz (2017), S. 65.
[202] Vgl. Brunton/Kutz (2017), S. 68.
[203] Vgl. Ohm/Lüke (2014), S. 133.
[204] Implementierungen der FFT Algorithmen bspw. in Brigham (1987), S. 181 ff.
[205] Voxengo: SPAN. Version 3.11 (2021).

Signalkette gehen im ebenfalls entstandenen Rauschen unter. Im tiefen Bereich sind hingegen deutlich die Störungen durch Netzbrummen zu sehen.

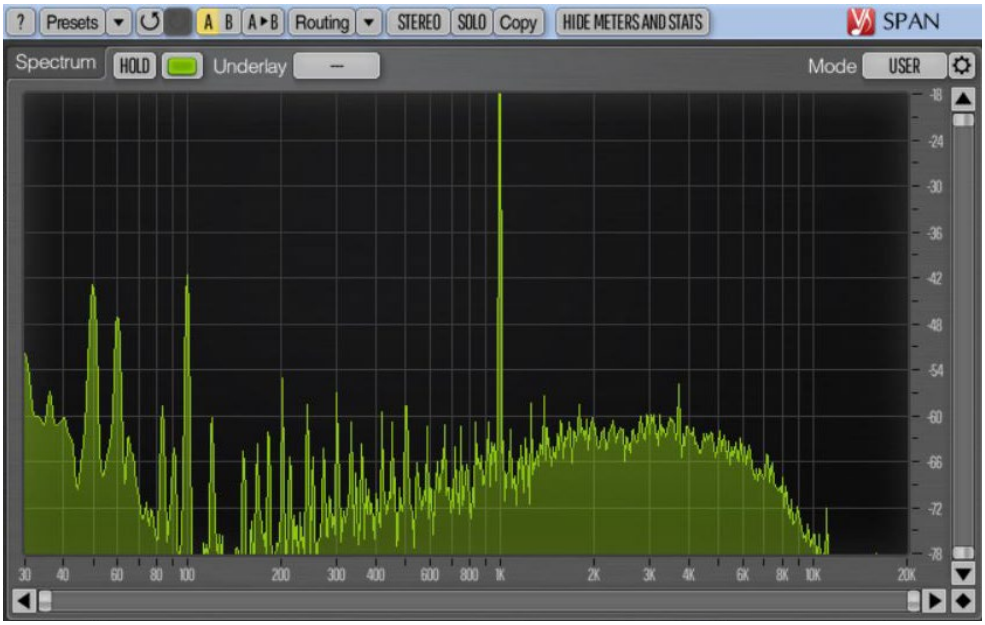

Abb. 16: Störspektrum der Signalkette D/A-Wandler-Gitarrenverstärker-A/D-Wandler.[206]

5.7.6 Kurzzeit-Fourier-Transformation

Da die meisten Signale, die in der Audiotechnik verarbeitet werden, nicht periodisch sind, ist es von großer Bedeutung, die spektrale Zusammensetzung auch bei einem variierenden Verlauf über die Zeit anschaulich wiedergeben zu können. Dies kann durch das Analyseverfahren der Kurzzeit-Fourier-Transformation (STFT) umgesetzt werden. Dabei wird das Signal zunächst in kürzere gleichgroße Abschnitte eingeteilt, um anschließend jeweils eine Fourier-Transformation (z. B. FFT) durchzuführen.[207] Dieser zeitliche Verlauf des Frequenzspektrums ist dann in einem sogenannten Spektrogramm zu finden.[208] Mit der Zeit als x-Achse, der Frequenz als y-Achse und Farben (oder Grauwerten) für die jeweiligen Amplituden, wird das Spektrogramm als Heatmap dargestellt. Auch hier ist die Wahl der richtigen Fensterfunktion entscheidend, um den durch das Einteilen resultierenden Fehler zu minimieren.

[206] Voxengo: SPAN. Version 3.11 (2021).
[207] Vgl. Brunton/Kutz (2017), S. 80.
[208] Vgl. Brunton/Kutz (2017), S. 80.

5.7.7 Rauschunterdrückung mittels FFT

Die FFT findet in vielen Bereichen Anwendung. So z. B. auch bei der Rauschunterdrückung in Signalen. Hierzu soll ein einfaches Python-Programm (vgl. Anhang 1, Abb. 26) als Veranschaulichung dienen. Zunächst wird eine Nutzsignalfunktion f_clean erstellt, welche sich aus der Summierung der drei Sinusfunktionen mit den Frequenzen 60 Hz, 100 Hz und 120 Hz ergibt. In der folgenden Zeile wird f_noise erstellt, indem f_clean mit einem vergleichsweise starken Rauschen überlagert wird. In Abb. 17 (oben) sind beide Funktionen grafisch dargestellt.

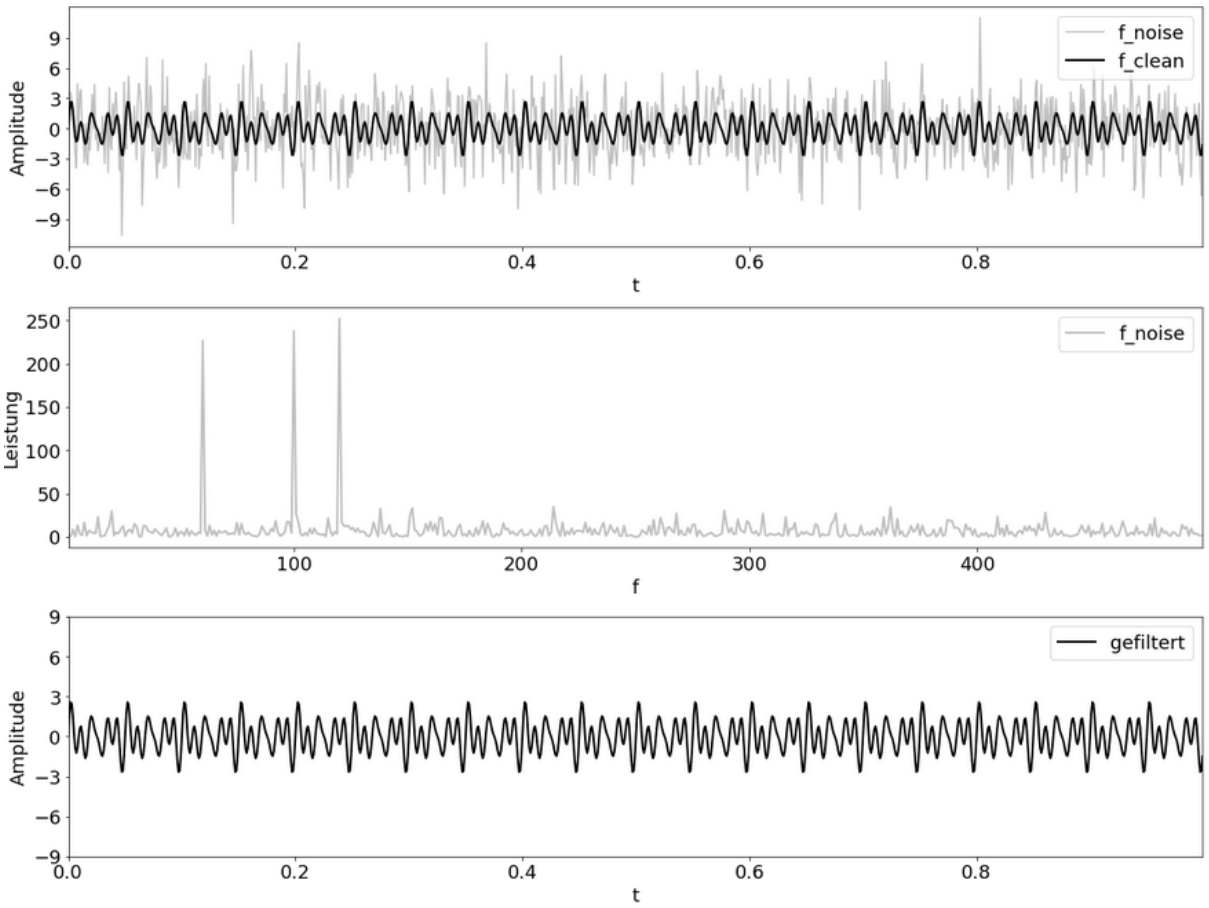

Abb. 17: Signale f_noise und f_clean (oben), Leistungsdichtespektrum (mittig) und gefiltertes Signal nach iFFT (unten).

Mit dem Einzeiler $np.fft.fft(f_noise, M)$ wird der Vektor der Fourier-Transformierten des Signals f_noise mit M Abtastwerten gebildet. Daraufhin erfolgt die Berechnung des LDS[209], welches in Abb. 17 (mittig) dargestellt ist. Hier ist nun deutlich zu sehen, dass die Frequenzen 60 Hz, 100 Hz und 120 Hz eine besonders hohe Leistung besitzen.

[209] Nähere Ausführungen dazu finden sich in Abschn. 5.2.2

Da diese genau der Frequenzen der drei sinusförmigen Komponenten entsprechen, liegt die Idee, alles unter einem gewissen Threshold wegzufiltern, nahe. Dazu wird dieser beispielhaft bei dem Leistungswert 100 gesetzt und somit alles unterhalb in 0 gewandelt. Anschließend kann durch eine inverse FFT (iFFT) wieder eine Zeitfunktion und somit ein gesäubertes Nutzsignal erstellt werden (vgl. Abb. 17 (unten)). Dabei ist zu beachten, dass dieses gefilterte Signal nicht exakt f_clean entspricht, da durch f_noise auch bei diesen Frequenzen Rauschen addiert wurde. Dies ist an den unterschiedlichen Ausschlägen der drei Frequenzen im LDS abzulesen.

5.7.8 Wavelet-Transformation

Bei der reinen Betrachtung der Wellenform eines Audiosignals herrscht eine hohe zeitliche Auflösung. Aussagen über die enthaltenen Frequenzen können jedoch nicht getätigt werden. Nach einer Fourier-Transformation hingegen ist die Frequenzauflösung gut, jedoch fehlt die Information der Zeit (vgl. Abschnitt 5.7.2). Einen Kompromiss zwischen angemessener Zeit- und Frequenzauflösung bietet das Spektrogramm (vgl. Abschnitt 5.7.6). Die Wavelet-Transformation (WT) ermöglicht dagegen verschiedene zeitliche Auflösungen für verschiedene Frequenzbereiche des Audiosignals. In Abb. 18 ist die konstante Frequenz- und Zeitauflösung der STFT (links) im Vergleich mit der variablen Frequenz- und Zeitauflösung der WT (rechts) schematisch dargestellt. Dabei steht die Breite der einzelnen Rechtecke für die Zeitauflösung und die Höhe für die Frequenzauflösung. Eine schlechtere Auflösung in der Zeit resultiert in einer besseren Auflösung in den Frequenzen (und umgekehrt). Da das menschliche Gehör hohe Frequenzen nicht so gut differenziert, können diese mit einer geringeren Frequenz- aber höheren Zeitauflösung dargestellt werden. Außerdem kann bei tiefen Frequenzen durch eine niedrigere Zeitauflösung eine deutlich bessere Frequenzauflösung erreicht werden. Dieser Vorgang wird auch Multiskalenanalyse genannt und kann damit zur verbesserten Anpassung an das Gehör genutzt werden.[210] Im Vergleich zur Fourier-Transformation wird ein Audiosignal bei der Wavelet-Transformation nicht in sinusförmige Schwingungen, sondern in transformationseigene Basisfunktionen, sogenannte Wavelets, zerlegt.[211] Dabei können je nach Anwendungen unterschiedliche Wavelets als Basisfunktion genutzt werden. Diese haben jeweils spezielle

[210] Vgl. Görne (2014), S. 153.
[211] Vgl. Görne (2014), S. 153.

Eigenschaften wie bspw. zeitkontinuierlich oder zeitdiskret, endliche oder unendliche zeitliche Ausdehnung, reell oder komplex, oder orthogonal oder nicht orthogonal.[212]

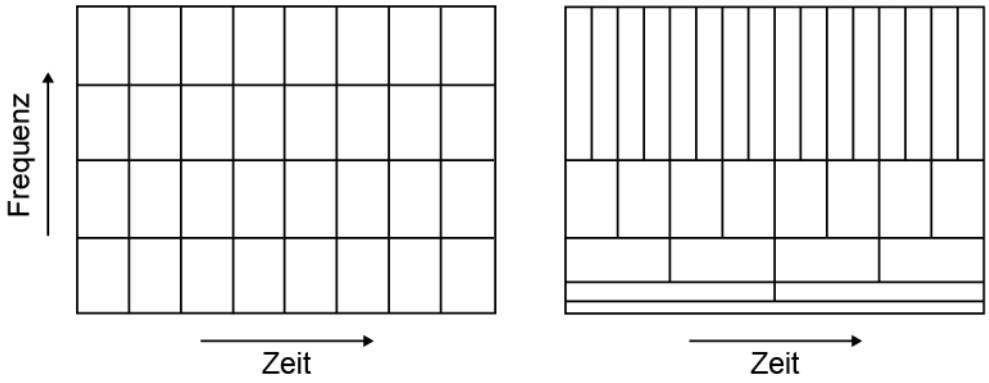

Abb. 18: Frequenz- und Zeitauflösung bei der STFT (links) und WT (rechts).[213]

Ähnlich dem Spektrogramm bei der STFT kann die WT als sogenanntes Skalogramm mit Farb- oder Grauwertdarstellung wiedergegeben werden.[214]

5.7.9 Oktavanalyse

Um bei einer Messung des Schalldruckpegels Informationen über die Pegelwerte einzelner Frequenzbereiche zu erhalten, lässt sich eine sogenannte Oktavanalyse anwenden.[215] Dabei wird das Audiosignal im Frequenzbereich durch eine Filterbank aufgeteilt. Eine Filterbank besteht aus mehreren Bandpassfiltern in Form von Oktavfiltern.[216] Deren Breite beträgt eine Oktave, d. h. die obere Grenzfrequenz entspricht dem Doppelten der unteren Grenzfrequenz. Um genauere Einteilungen zu erhalten, kann die Breite der Bandpassfilter verkleinert werden. Werden die Oktavfilter jeweils nochmals in drei Teile unterteilt, entstehen die sogenannten Terzfilter.

5.8 Akustischer Fingerabdruck

Ein akustischer Fingerabdruck gibt in einzigartiger und kompakter Form die für die Wahrnehmung relevanten Aspekte einer Audioaufnahme wieder.[217] Hierzu wird eine Audioaufnahme mit einer großen Anzahl an Bits, durch eine Funktion F einem

[212] Vgl. Hoffmann/Wolff (2014), S. 311.
[213] In Anlehnung an Hoffmann/Wolff (2014), S. 301.
[214] Vgl. Hoffmann/Wolff (2014), S. 312.
[215] Vgl. Müller u. a. (2017), S. 15.
[216] Vgl. Müller u. a. (2017), S. 15.
[217] Vgl. Lerch (2012), S. 163.

Fingerabdruck mit einer begrenzten Anzahl an Bits zugeordnet.[218] So kann beim Anlegen der Datenbank eine große Datenmenge eingespart und die Dauer der Suche verkürzt werden. Dies kann mit dem so genannten Perceptual Hashing umgesetzt werden. Im Vergleich zu den bspw. aus der Kyptographie bekannten Hash Funktionen, führt beim Perceptual Hashing eine kleine Änderung im Eingangswert nicht zu einer starken Änderung des Ausgangswertes. Genau wie der Großteil der Menschen bspw. einen Song in der WAV und der MP3 Version als gleich wahrnimmt, kann bei der Perceptual Hash Funktion trotz teils unterschiedlicher Wellenformen der gleiche Fingerabdruck zugeordnet werden.[219] Nach Haitsma (2002, S. 4) sind für ein System, welches auf akustischen Fingerabdrücken basiert, unter anderem folgende Parameter wichtig:

- **Robustheit**: legt fest, ob der richtige Fingerabdruck der Aufnahme auch nach starker Abwandlung (durch z. B. Hintergrundrauschen, Filterung oder Raumhall) noch identifiziert werden kann. Die Robustheit wird durch die falsch negativ Rate definiert, d. h. wie oft eigentlich gleich wahrgenommene Signale aufgrund einer zu großen Abweichung als nicht gleich festgelegt werden.
- **Verlässlichkeit**: gibt an, wie oft ein Titel fälschlicherweise einem anderen zugeordnet wird, also entsprechend der falsch positiv Rate.
- **Größe eines akustischen Fingerabdrucks**: ist entscheidend für die Geschwindigkeit der Suche und die Reduzierung der benötigten Serverkapazitäten der Datenbank.
- **Mindestdauer**: an Audiomaterial, welche für die Erkennung benötigt wird.
- **Geschwindigkeit und Skalierbarkeit der Suche**: sind entscheidend dafür, wie schnell ein akustischer Fingerabdruck in der Datenbank gefunden werden kann und wie diese sich bei steigender Größe verhalten.

5.8.1 Erstellung des akustischen Fingerabdrucks

Um einen akustischen Fingerabdruck einer unbekannten Audioaufnahme zu bestimmen, wird diese zunächst in 11,8 Millisekunden lange Intervalle eingeteilt.[220] Da nicht bekannt ist, welche jeweiligen Intervallgrenzen für die originalen Hashwerte in der

[218] Vgl. Haitsma (2002), S. 3.
[219] Vgl. Haitsma (2002), S. 3.
[220] Vgl. Haitsma (2002), S. 10.

Datenbank gewählt wurden, wird bei der Einteilung der Audioaufnahme mit hoher Überlappung gearbeitet.[221] Um anschließend die spektrale Darstellung der einzelnen Intervalle zu erhalten, wird jeweils eine Fourier-Transformation durchgeführt.[222] Aufgrund der Eigenschaft des menschlichen Gehörs, nicht besonders phasenempfindlich zu sein, wird im weiteren nur der Betrag des Spektrums verwendet.[223] In Anlehnung an die Bark-Skala werden nun 33 sich nicht überlappende Frequenzbänder zwischen 300 Hz und 2000 Hz mit logarithmischen Abständen gewählt.[224] Wird nun das Vorzeichen der Energieunterschiede der einzelnen Intervallbänder (sehr robust gegenüber vielen Signalveränderungen) betrachtet, lässt sich je Intervall ein 32 Bit Hashwert bilden.[225] Die Schritte, um diesen 32 Bit Hashwert zu erstellen, sind in Abb. 19 schematisch dargestellt.

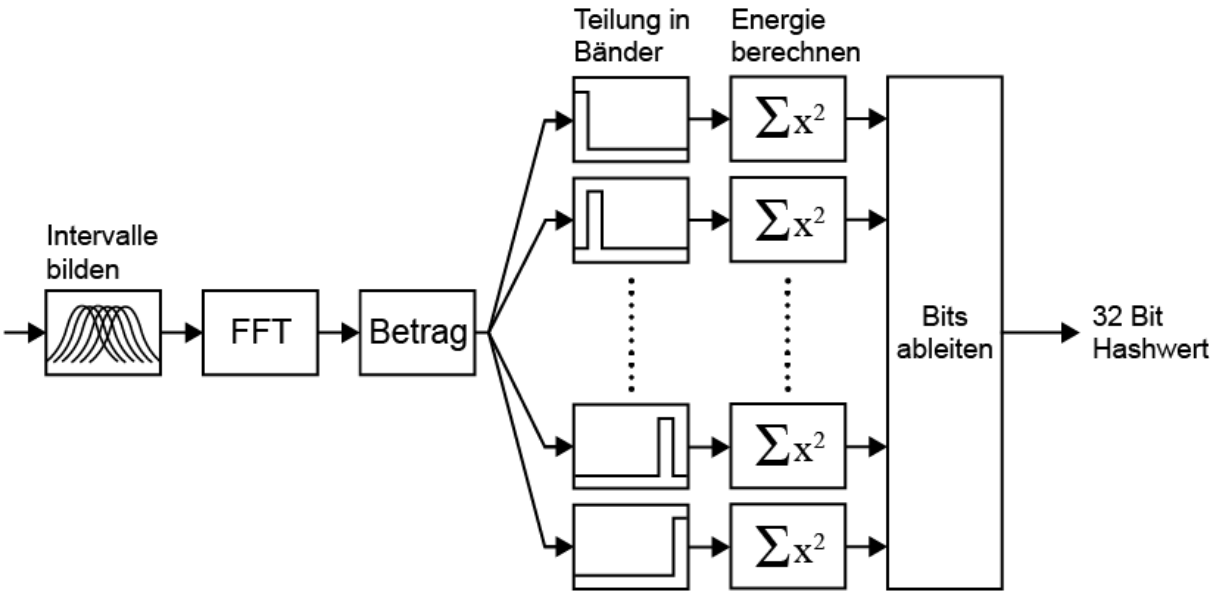

Abb. 19: Schematische Übersicht der Erstellung eines 32 Bit Hashwertes.[226]

In Abb. 20 sind 256 aufeinanderfolgende 32 Bit Hashwerte als Block dargestellt. Dieser entspricht einer ungefähren Länge von 3 Sekunden. Dabei entsprechen Bits mit einer 1 weißen Pixel, und Bits mit einer 0 schwarzen Pixel. Um aufzuzeigen, wie sich die Signalverarbeitung auf einen akustischen Fingerabdruck auswirken kann, wurde hier beispielhaft eine MP3 Kompression auf 32 kBit durchgeführt. Die Abweichung vom originalen Audiosignal (a), die durch die Kompression in (b) entstehen, sind in (c) als

[221] Vgl. Haitsma u. a. (o.J.), Abschn. 3.
[222] Vgl. Haitsma u. a. (o.J.), Abschn. 3.
[223] Vgl. Haitsma u. a. (o.J.), Abschn. 3.
[224] Vgl. Haitsma (2002), S. 10 f.
[225] Vgl. Haitsma (2002), S. 11.
[226] In Anlehnung an Haitsma u. a. (o.J.), Abschn. 2.

Bitfehler (schwarze Pixel) sichtbar. Die starke Kompression (vgl. Abb. 20) führt zu einer Bitfehlerrate (BER) von 0,115, d. h. von den 8192, sich innerhalb des Blocks befindenden Bits, sind mehr als 940 Bits fehlerhaft.

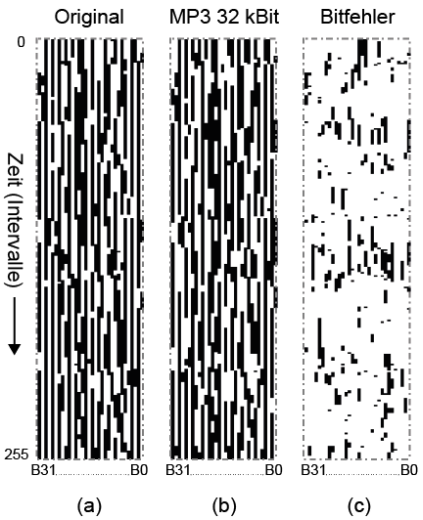

Abb. 20: Fingerabdruckblöcke der originalen Musik (a) und der komprimierten Version (b), sowie der Unterschied zwischen beiden anhand der Bitfehlerrate (BER = 0,115).[227]

Da für die robusten akustischen Fingerabdrücke nur Frequenzen unterhalb der 2000 Hz einbezogen werden, kann zunächst ein Down-Sampling (auf 5 kHz) und eine Konvertierung in mono durchgeführt werden, um entstehende ungenutzte Daten zu reduzieren.[228]

5.8.2 Datenbank durchsuchen

Um den Titel der unbekannten Audioaufnahme zu finden, muss der erstellte akustische Fingerabdruck nun mit einer Fingerabdruck-Datenbank von mehreren Millionen Songs abgeglichen werden. Das Anspruchsvolle daran ist, dass nicht der Gleiche, sondern der Fingerabdruck mit der meisten Ähnlichkeit gefunden werden muss. Würden hier alle Möglichkeiten ausprobiert werden (Brute-Force-Methode), könnte die Suche einige Zeit in Anspruch nehmen. Um dies zu beschleunigen, wird von einer hohen Wahrscheinlichkeit ausgegangen, dass in einem Fingerabdruckblock einer gesuchten Audioaufnahme mindestens einer der 32 Bit Hashwerte im Vergleich zum passenden Original in der Datenbank keinen Bitfehler hat.[229] Außerdem wird eine Lookup-Tabelle

[227] Haitsma u. a. (o.J.), Abschn. 3.
[228] Vgl. Haitsma/Kalker (2002), Abschn. 4.2.
[229] Vgl. Haitsma u. a. (o.J.), Abschn. 5.

(LUT) erstellt, welche alle möglichen 32 Bit Hashwerte der Datenbank enthält.[230] Die Einträge der LUT zeigen dabei auf den Song (bzw. die Songs) und die Positionen innerhalb dessen, wo der jeweilige 32 Bit Hashwert zu finden ist.[231] Werden nun die 256 Hashwerte in der LUT gesucht, ergibt sich eine Liste möglicher Kandidaten für den unbekannten Titel.[232] Zwischen dem Fingerabdruckblock der Audioaufnahme und der Kandidaten an den jeweiligen Stellen wird nun die BER berechnet.[233] So kann einfach bestimmt werden, welcher Titel aus der Datenbank am ehesten mit der Audioaufnahme übereinstimmt (der mit der niedrigsten BER).[234] Falls die BER jedoch über einem festgelegtem Threshold ist, wird kein passender Titel angegeben.[235] Dies trägt zur Verlässlichkeit des Systems bei, d. h. die falsch positiv Rate wird geringer.

5.8.3 Shazam

Shazam ist ein Dienst, um Musik, TV-Sendungen, Filme und Werbung automatisch zu analysieren und erkennen zu lassen.[236] Während nach der Markteinführung im Jahr 2002 noch eine Kurzwahlnummer mit dem Telefon angerufen werden musste, um anschließend den gesuchten Titel per SMS zu erhalten, sind heute alle Funktionen in der kostenlosen mobilen App nutzbar.[237] Dabei hat Shazam mittlerweile um die 200 Millionen monatliche Nutzer und insgesamt bereits über 50 Milliarden Tags erreicht.[238] Wie auch bei der zuvor beschriebenen Analyse werden bei Shazam akustische Fingerabdrücke verwendet, um Handyaufnahmen den zugehörigen Informationen wie Titel, Künstler, Möglichkeiten zum Streamen oder Kaufen usw. zuzuordnen. Im Folgenden wird beispielhaft das Vorgehen beim Analysieren eines Songs aufgezeigt. Dieser wird zunächst für eine Zeit von wenigen Sekunden vom Nutzer per Handy-App aufgezeichnet. Anschließend wird aus der Aufnahme ein Spektrogramm generiert (vgl. Abb. 21 und Abschnitt 5.7.6).

[230] Vgl. Haitsma u. a. (o.J.), Abschn. 5.
[231] Vgl. Haitsma u. a. (o.J.), Abschn. 5.
[232] Vgl. Haitsma u. a. (o.J.), Abschn. 5.
[233] Vgl. Haitsma u. a. (o.J.), Abschn. 5.
[234] Vgl. Haitsma u. a. (o.J.), Abschn. 5.
[235] Vgl. Haitsma u. a. (o.J.), Abschn. 5.
[236] Vgl. Shazam (o.J.).
[237] Vgl. Shazam (o.J.).
[238] Vgl. Shah (2021).

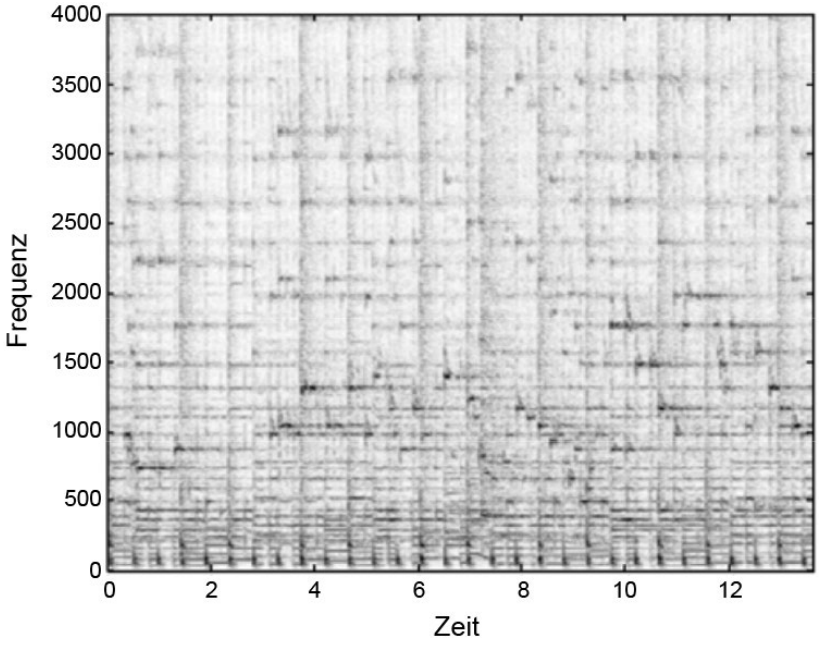

Abb. 21: Shazam Spektrogramm eines Audiofiles.[239]

In diesem werden dann einzelne Zeit-Frequenz-Punkte bestimmt, die im Vergleich zu den umliegenden Punkten eine höhere Energie, bzw. eine höhere Amplitude besitzen.[240] So wird aus dem Spektrogramm eine stark vereinfachte Form, in welcher nur noch die Koordinaten dieser Zeit-Frequenz-Punkte enthalten sind (vgl. Abb. 22).[241]

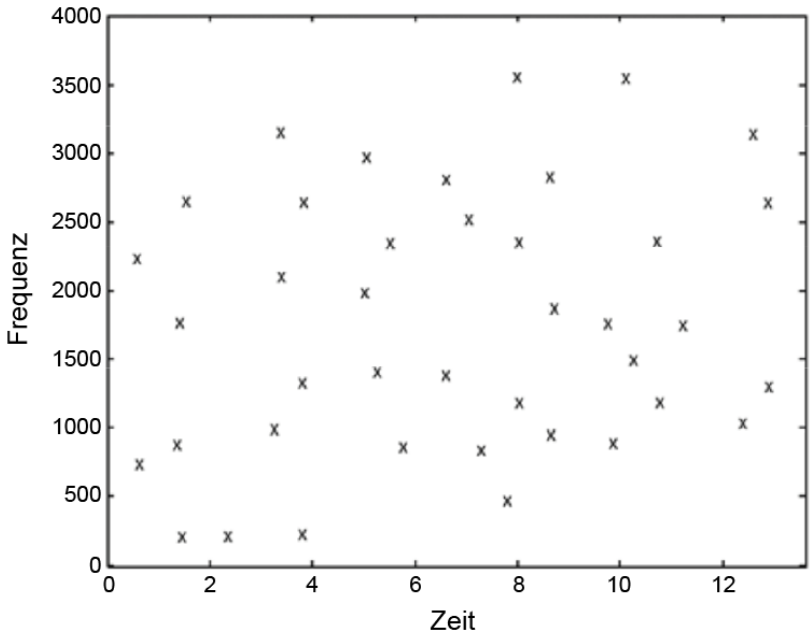

Abb. 22: Shazam Constellation Maps.[242]

[239] Wang (2003), Abschn. 2.
[240] Vgl. Wang (2003), Abschn. 2.1.
[241] Vgl. Wang (2003), Abschn. 2.1.
[242] Wang (2003), Abschn. 2.

Da die Amplitudenkomponente des Spektrogramms entfällt, ist der Algorithmus weniger anfällig auf Absenkung oder Anhebung von Frequenzbereichen in der Handyaufnahme oder bei der restlichen Signalverarbeitung.[243] Diese Zeit-Frequenz-Diagramme werden in einem Paper von Shazam Mitgründer Avery Li-Chun Wang als Constellation Maps bezeichnet.[244] Um aus diesen nun Hashwerte zu gewinnen, werden Ankerpunkte mit zugehörigen Target Zonen festgelegt (vgl. Abb. 23).

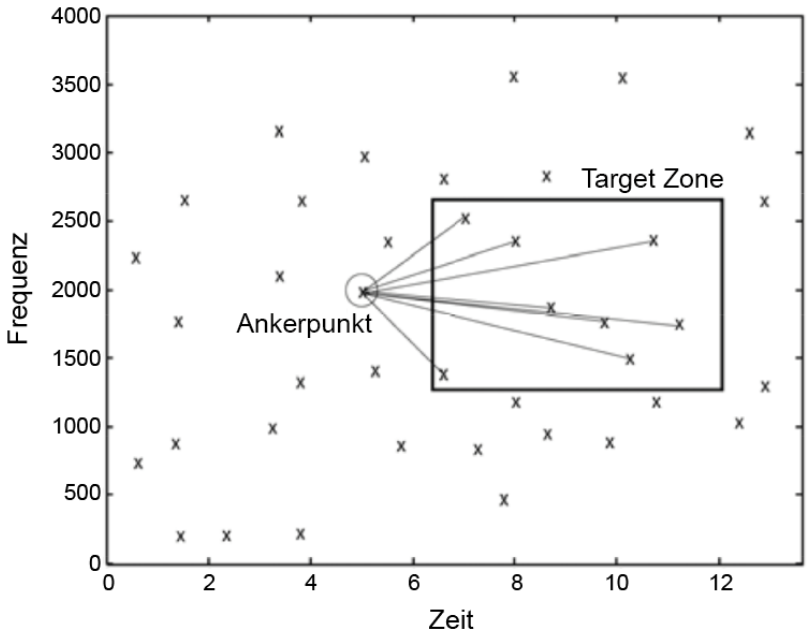

Abb. 23: Festlegung eines Ankerpunkts mit zugehöriger Target Zone.[245]

Ein Ankerpunkt bildet nacheinander mit den Punkten innerhalb seiner Target Zone Paare, wobei sich jedes Paar aus zwei Frequenzkomponenten und der Zeitdifferenz zwischen beiden Punkten ergibt. Diese Werte werden nun in einen 32-Bit Hashwert umgeformt und mit dem Zeitpunkt des Ankerpunktes innerhalb der Aufnahme zusammengebracht (vgl. Abb. 24).[246]

[243] Vgl. Wang (2003), Abschn. 2.1.
[244] Vgl. Wang (2003), Abschn. 2.1.
[245] Wang (2003), Abschn. 2.
[246] Vgl. Wang (2003), Abschn. 2.2.

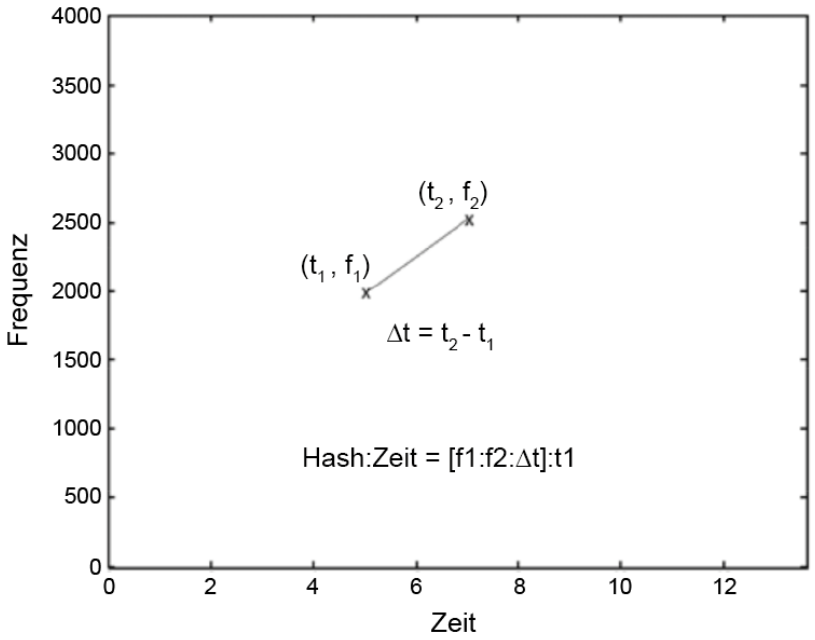

Abb. 24: Bildung des Hash:Zeit Wertes eines Target Zonen Punkt mit dessen Ankerpunkt.[247]

Dieser Vorgang wird auch für alle Files in der Datenbank durchgeführt. Hier wird jedoch zusätzlich noch die Song ID, mit angehängt. So ergibt sich ein 64-Bit Paket, bestehend aus einem 32-Bit Hashwert und 32 Bit für den Zeitpunkt innerhalb der Aufnahme und der Song ID.[248] Um nun die Handyaufnahme einer ID in der Datenbank zuzuordnen, wird für jedes Paket aus Hashwert und Zeitpunkt in der Aufnahme eine Suche in der Datenbank nach passenden Hashes durchgeführt.[249] Für jeden Treffer der Suche wird ein Zeitpaar aus dem Zeitpunkt innerhalb der Handyaufnahme und dem Zeitpunkt im passenden Song der Datenbank erstellt und in einem Bin[250] der jeweiligen Song ID abgelegt.[251] Sind nun alle Hashes der Aufnahme mit der Datenbank abgeglichen, werden die Bins mit Zeitpaaren der gefundenen Song IDs analysiert.[252] Dies kann in sogenannten Scatterplots veranschaulicht werden.[253] Wenn eine Song ID zu der Handyaufnahme passt, dann ist eine gleiche Hash-Sequenz sowohl in der Handyaufnahme, als auch in dem passenden File der Datenbank.[254] Im zugehörigen Scatterplot bildet sich ein auffälliges Cluster in Form einer Geraden.[255]

[247] Wang (2003), Abschn. 2.
[248] Vgl. Wang (2003), Abschn. 2.2.
[249] Vgl. Wang (2003), Abschn. 2.3.
[250] Eine Art virtuelles Behältnis für die Zeitpaare
[251] Vgl. Wang (2003), Abschn. 2.3.
[252] Vgl. Wang (2003), Abschn. 2.3.
[253] Vgl. Wang (2003), Abschn. 2.3.
[254] Vgl. Wang (2003), Abschn. 2.3.
[255] Vgl. Wang (2003), Abschn. 2.3.

Mit der Gleichung

$$\delta t_k = t'_k - t_k \tag{32}$$

kann nun ein Histogramm berechnet werden, wobei t_k‘ die Zeitkoordinate des Datenbank Files und t_k zugehörige Zeitkoordinate der Handyaufnahme ist.[256] Nachdem dies für alle Bins durchgeführt wurde, kann ausgewertet werden, welcher Song den größten Peak im Histogramm hat. Für diesen können dann die zugehörigen Informationen an den Nutzer übertragen werden.

[256] Vgl. Wang (2003), Abschn. 2.3.

6 KÜNSTLICHE INTELLIGENZ FÜR AUDIOANALYSEN

Nachdem in den vorangegangenen Abschnitten hauptsächlich auf Verfahren eingegangen wurde, die seit mehreren Jahren wichtiger Bestandteil der Audioanalyse darstellen, soll das Folgende auch als Ausblick auf die nahe Zukunft dienen.
Die Nutzung von künstlicher Intelligenz (KI) ist wie in den meisten technischen Bereichen auch in der Signalverarbeitung von Audio allgegenwärtig. Der Arbeitsaufwand in jeglichen Verfahren soll vermindert und damit vor allem beschleunigt werden. Dabei ist hauptsächlich die Reduzierung von administrativen Aufgaben von Interesse. In der Praxis wird künstliche Intelligenz jedoch auch immer häufiger bei kreativen Entwicklungsprozessen, wie bspw. beim Komponieren, beim Erweitern oder beim Mastering von Musik eingesetzt. Im Folgenden werden Beispiele für KI-Implementierungen in Plugins aufgezeigt, welche zur Analyse von verschiedenen Audioeigenschaften genutzt werden. Dabei steht zunächst die Vielfalt der Anwendungsmöglichkeiten und nicht die jeweils verwendeten KI-Algorithmen im Vordergrund. Eine beispielhafte Implementierung wird stattdessen in Abschnitt 6.3 dargestellt.

6.1 Künstliche Intelligenz in Plugins

Heutzutage ist es schon für ein vergleichsweise geringes Budget möglich, in die Audioproduktion einzusteigen. Durch ein großes Angebot an Informationen im Internet können auch komplexe Zusammenhänge der Audiotechnik selbstständig erlernt werden. Durch Plugins mit künstlicher Intelligenz können Arbeitsabläufe, die langwierig sind oder über das eigene Knowhow hinausgehen, automatisch erledigt werden.

Ein Beispiel wie KI in Plugins Anwendung finden kann, ist das Plugin Jamahook. Es ermöglicht auf Basis des aktuellen Mix, automatisch passende Samples und Sound-Loops zu finden.[257] Zur Umsetzung des Plugins ist das Schweizer Unternehmen Jamahook eine Kooperation mit dem Fraunhofer-Institut für Digitale Medientechnologie (IDMT) eingegangen.[258] Durch ein KI-basiertes Klassifizierungssystem (basierend auf der SoundsLike[259] Technologie) wird ein automatisches Kategorisieren und Versehen

[257] Vgl. Jamahook Corp. (o.J.).
[258] Vgl. Fraunhofer-Institut für Digitale Medientechnologie IDMT (o.J.a).
[259] Nähere Ausführungen dazu finden sich in Fraunhofer-Institut für Digitale Medientechnologie IDMT (o.J.b).

mit Metadaten der ungelabelten Audiodaten der Jamahook Datenbank ermöglicht.[260] Nachdem der Algorithmus einen Ausschnitt des aktuellen Mix hinsichtlich Stimmung, Harmonie, Rhythmus, Tonart, Tempo oder Genre analysiert hat, werden übereinstimmende Loops in der Datenbank gesucht und dem Nutzer vorgeschlagen.[261] Dies kann das zeitintensive manuelle Durchsuchen ablösen und den Workflow beschleunigen.

Einer ähnlichen Problemstellung widmet sich auch Algonaut Atlas 2. Hierbei können die auf dem eigenen Rechner installierten Drum Samples durch eine KI klassifiziert und in eine übersichtlichere Form sortiert und organisiert werden.[262] Außerdem können auch ganze Drum-Sets automatisch von der KI erstellt werden.[263]

Der Softwarehersteller iZotope Inc. hat bereits mehrere Plugins mit integrierter KI veröffentlicht. Dazu gehört u.a. auch das für das Mixing entwickelte Plugin Neutron 2. Der darin enthaltene Track-Assistent ermöglicht die automatische Erkennung des Instruments und passt die Parameter der enthaltenen Module wie Filter, Kompressor oder Exciter entsprechend der Signaleigenschaften an.[264] Auch der Workflow beim Mastering kann durch die KI des Master-Assistenten im iZotope Plugin Ozone 8 profitieren. Je nach gewünschtem Zielformat, legt der Master-Assistent die entsprechenden Parameter der Module fest.[265] So wird bspw. automatisch die gewünschte Lautheit erzielt oder der Effekt des Limiters auf den Gesamtsound festgelegt.[266]
Mit dem Vocal-Assistenten werden neben dem Vocal-Timbre auch der zeitliche Pitch-Verlauf betrachtet, um die Signalbearbeitungskette optimal anzupassen.[267] Auch problematische Zischlaute werden erkannt und mit einem integrierten De-Esser ausgeglichen.[268]

Durch das Plugin Adaptiverb von Zynaptiq ist es möglich, das Eingangssignal zu analysieren und dank künstlicher Intelligenz einen Reverb zu generieren, der sich optimal in die Mischung einfügt.[269] Auch hier werden automatisch die Werte der einzelnen Einstellmöglichkeiten angepasst und festgelegt.[270]

---

[260] Vgl. Fraunhofer-Institut für Digitale Medientechnologie IDMT (o.J.a), Abschn. Realisierung.
[261] Vgl. Fraunhofer-Institut für Digitale Medientechnologie IDMT (o.J.a), Abschn. Realisierung.
[262] Vgl. Algonaut (o.J.), Abschn. Sample Map.
[263] Vgl. Algonaut (o.J.), Abschn. Drum Kit.
[264] Vgl. iZotope (o.J.), Abschn. Track Assistant.
[265] Vgl. iZotope (o.J.), Abschn. Master Assistant.
[266] Vgl. iZotope (o.J.), Abschn. Master Assistant.
[267] Vgl. iZotope (o.J.), Abschn. Vocal Assistant.
[268] Vgl. iZotope (o.J.), Abschn. Vocal Assistant.
[269] Vgl. Zynaptiq (o.J.).
[270] Vgl. Zynaptiq (o.J.).

Vor allem Nutzer ohne besondere audiotechnische Ausbildung profitieren von KI-Implementierungen wie sie in den vorangegangenen Beispielen gezeigt wurden. So können komplexere Arbeitsschritte auch ohne Kenntnisse über die zugrundeliegenden Zusammenhänge durchgeführt werden. Dieser Markt ist auch für die Softwareentwickler von großem Interesse. Hersteller wie bspw. iZotope Inc. haben bereits angekündigt, auch in Zukunft weiter Einsatzmöglichkeiten von KI zu entwickeln und zu vertreiben.[271]

6.2 Künstliche neuronale Netzwerke

Die als ein Teilgebiet der künstlichen Intelligenz geltenden künstlichen neuronalen Netzwerke (KNN) finden auch in Bereichen der Audiotechnik immer häufiger Anwendung. Bei KNNs handelt es sich um eine mathematische Modellierung der parallelen Signalverarbeitungsstruktur des menschlichen Nervensystems.[272] Hierzu werden künstliche Neuronen, auch Nodes genannt, aufgebaut (in Abb. 25 vereinfacht als Kreise dargestellt). Ein künstliches Neuron gewichtet und summiert mehrere Inputs und transformiert diese entsprechend einer Übertragungsfunktion zu einem Output.[273] Dieser wird dann an die nächsten Neuronen weitergegeben und nach dem gleichen Prinzip weiterverarbeitet. KNNs bestehen üblicherweise aus einzelnen Layern (Schichten). Wie in Abb. 25 dargestellt, gibt es dabei eine Input-Layer, eine oder mehrere Hidden-Layers und eine Output-Layer. Um ein KNN zu erstellen werden grundsätzlich die drei Phasen, Trainieren, Speichern und Bereitstellen durchgeführt. Für das Training ist ein möglichst großer Datensatz nötig, der zunächst in eine einheitliche, vom Modell vorgegebene Form zu bringen ist. Bei Audiodaten könnte es sich dabei bspw. um eine benötigte Abtastrate, Gesamtlänge oder Bittiefe handeln. Anschließend kann die gewünschte Netzwerkstruktur implementiert und in mehreren Durchläufen (Epochen) trainiert werden. Bei großen Datensätzen, komplexer Netzwerkstruktur oder hoher Epochenanzahl kann das Training einige Zeit in Anspruch nehmen.[274] Je nach Aufwand der Berechnung und der Stärke der vorhandene Rechenleistung, kann das Trainieren von wenigen Minuten bis zu mehreren Tagen dauern.[275]

[271] Vgl. iZotope (o.J.).
[272] Vgl. Abraham (2005), S. 901.
[273] Vgl. Abraham (2005), S. 901.
[274] Vgl. Sarang (2021), S. 4 f.
[275] Vgl. Sarang (2021), S. 8.

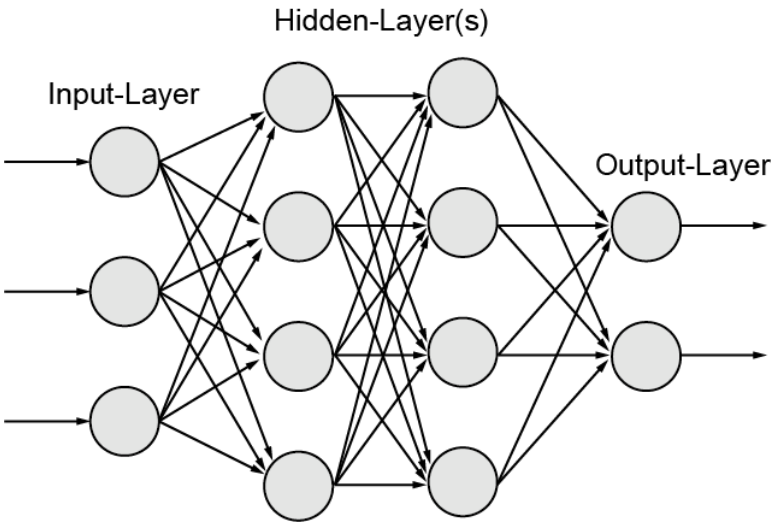

Abb. 25: Vereinfachte Darstellung eines künstlichen neuronalen Netzwerks.[276]

Um mit dem KNN am Ende bestmögliche Resultate zu erzielen, sollten Zwischenergebnisse analysiert und die Netzwerkstruktur dementsprechend angepasst werden. Auf diese Weise kann die sogenannte Accuracy (Treffergenauigkeit) des Netzwerks optimiert werden. Diese gibt in Prozent das Verhältnis der Anzahl korrekter Klassifizierungen zu der Anzahl aller Klassifizierungen an und stellt somit eine Bewertung dar, welche die Zuverlässigkeit des KNN wiedergibt. Sind die Resultate zufriedenstellend, kann das Netzwerk gespeichert und bspw. in einer Anwendung nutzbar gemacht werden.

6.3 Automatische Klassifizierung von Audiosignalen

Durch das Internet ist es um ein Vielfaches einfacher geworden, Audiomaterial wie Musik, Samples usw. für die Allgemeinheit verfügbar zu machen. Da die Datenbanken weiterwachsen, ist es sowohl im Interesse des Anbieters als auch des Nutzers, diese in Struktur und Ordnung zu halten. Immer wichtiger werden daher automatische Algorithmen, welche Audioinhalte auf bestimmte Eigenschaften untersuchen, vergleichen und auswerten können. Hierfür kann ein sogenanntes Convolutional Neural Network (CNN) eingesetzt werden, welches eine spezielle Form des künstlichen neuronalen Netzwerks darstellt. Dabei sind Anwendungsfälle, wie bspw. aus einem unbekannten Datensatz mit Audiodaten alle Snaredrum-Samples oder alle Songs des Genres Rock zusammenzuführen, denkbar. Auch könnte ein Musiktitel durch ein CNN auf

[276] In Anlehnung an Abraham (2005), S. 902.

bestimmte Eigenschaften analysiert werden und ähnliche, in der Datenbank enthaltene Titel vorschlagen. So können Metadaten vollständig automatisch erstellt und zeitintensive manuelle Eingaben eingespart werden.

In einem einfachen Beispiel, welches in Jupyter Notebook[277] in Python[278] implementiert wurde, nimmt ein CNN einen unbekannten Song (ohne Beschriftung, Metadaten, usw.) entgegen und ordnet diesen dem passenden Genre zu. Für das Training konnte dank der freundlichen Genehmigung des Autors der Datensatz GTZAN Genre Collection[279] verwendet werden. Dieser besteht aus 1000 Audiodateien, wobei jede 30 Sekunden lang ist. Jeweils 100 davon repräsentieren eines der 10 Genres: Blues, Klassik, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae und Rock. Die Audiodateien haben alle eine Abtastrate von 22050 Hz und eine Länge von 30 Sekunden, sind in Mono 16-bit und im WAV-Dateiformat. Um die Anzahl der Trainingsdaten zu erhöhen und diese in die von der Netzwerkstruktur vorgegebene Form zu bringen, mussten zunächst Anpassungen durchgeführt werden (vgl. Anhang 2, Abb. 26). Hierfür wurden die Audiodateien jeweils in 10 Segmente eingeteilt, wodurch sich die gesamte Anzahl auf 10.000 erhöhte. Anhand der Ordnerstruktur des Datensatzes konnten die Segmente mit Werten von 0 bis 9 gelabelt werden, um später das Rückschließen auf das entsprechende Genre zu ermöglichen (0 ≙ Blues, 1 ≙ Klassik, 2 ≙ Country, usw.). Da CNNs sich hervorragend für die Klassifizierung von Bildmaterial eignen, wurden aus den Segmenten sogenannte Mel-Frequenz-Cepstrum-Koeffizienten (MFCCs) gebildet. Dadurch wurden die Audiosegmente aus dem WAV-Format (also eine Funktion über der Zeit) in eine Art Matrizendarstellung umgeformt, welche jeweils die wichtigen Eigenschaften in kompakter Form enthält. Auf Basis dieser MFCCs lässt sich das Netzwerk sowohl einfacher aufbauen, als auch schneller trainieren. Die Genrebezeichnungen, MFCCs und Labels wurden dann unter *data_in_10_Segmente.json* abgespeichert. Auf dieses Jason-File kann nun im nächsten Schritt zugegriffen werden (vgl. Anhang 3, Abb. 28). Hier wurde der Datensatz zunächst aufgeteilt. Dabei wurden 20 % als Validierungsdaten und 25 % als Testdaten zurückgehalten. So konnten später die Validierungsdaten zur ständigen Kontrolle der Netzwerk Performance und die Testdaten zur Beurteilung des fertigen CNN genutzt werden. In der Funktion *build_model(input_shape)* in Anhang 3, Abb. 28, ist das Implementieren des Modells inklusive dessen

[277] Project Jupyter: Jupyter Notebook. Version 6.2.0 (2021).
[278] Python Software Foundation: Python. Version 3.9.0 (2020).
[279] Tzanetakis: GTZAN Genre Collection. Version 1 (o.J.).

Convolution-Layern zu sehen. In Anhang 4, Abb. 29, wurden die entsprechenden Funktionen aufgerufen um das CNN zu trainieren und letztendlich abzuspeichern. Zudem ist die Evaluierung der ersten Epochen abgebildet, wobei von Durchlauf zu Durchlauf ein starker Anstieg der Accuracy zu erkennen ist. Nach den gesamten 30 Epochen wird ein Wert von etwa 74 % erreicht, d. h., dass im Durchschnitt drei von vier unbekannten Songs durch das CNN richtig klassifiziert werden. In einer einfachen Anwendung kann das CNN nun aufgerufen und zur Genreklassifizierung unbekannter Songs genutzt werden.

Da Genres nicht genau definiert sind, haben Menschen unter Umständen Schwierigkeiten richtige Zuordnungen zu tätigen. Im Vergleich zu der automatischen Klassifizierung eines CNN, ist die des Menschen wesentlich subjektiver.[280] Durch die Berechnung der Eigenschaften durch MFCCs können mit einem CNN daher nicht nur höhere Accuracys, sondern auch konstantere Ergebnisse erreicht werden.

[280] Vgl. Lippens u. a. (2004), Abschn. 4.

7 FAZIT

In der jeweiligen Situation die optimalsten Verfahren anzuwenden, ist entscheidend um bestmögliche Ergebnisse zu erzielen. Dass dies bei der Analyse von Audiosignalen und -systemen voll und ganz zutrifft, hat sich durch die vorliegende Arbeit gezeigt. Es wurde erarbeitet, dass sich die verschiedenen Analyseverfahren der Audiotechnik meist auf einzelne Parameter fokussieren, um diese dann möglichst genau wiedergeben zu können. Sollen dagegen Audiosignale oder -systeme im Gesamten betrachtet werden, können verschiedene Verfahren kombiniert werden, um dadurch bessere Ergebnisse zu erzielen. So wurde gezeigt, dass bspw. bei der Bewertung von Lärm neben physikalisch messbaren Größen wie Schalldruckpegel oder enthaltene Frequenzen auch Phänomene wie empfundener Schalldruckpegel in Abhängigkeit der Frequenz, Rauigkeit, Schärfe oder Schwankungen der Tonhöhe berücksichtigt werden müssen. Durch verschiedene Methoden sollte sich daher nacheinander auf einzelne Aspekte (z. B. akustische Kamera zur Lokalisierung der Schallemission, STFT zur Bestimmung der enthaltenen Frequenzen, Messung des frequenzbewerteten Schalldruckpegels usw.) konzentriert werden, um eine bestmögliche Gesamtbewertung zu erzielen. Bei der Recherche ist aufgefallen, dass es aufgrund der vielen Methoden die eben nicht exakt, sondern annähernd sind, oftmals nur Vorschläge anstatt genormter Lösungswege für bestimmte Audioanalysen gibt. Dies resultiert in einer geringeren Vergleichbarkeit der Ergebnisse und kann, gerade bei der Bewertung von Lärm, zu Fehleinschätzungen führen.

In dieser Arbeit konnte durch das Erläutern der komplexen Funktionsweise des menschlichen Gehörs und das Aufzeigen verschiedener Quantifizierungsmöglichkeiten verdeutlicht werden, dass es für wahrnehmungsbezogene Betrachtungen essentiell ist, die Verfahren wie bspw. Schalldruckpegel-, Frequenz- oder Tonhöhenanalysen an die Eigenschaften des Ohres anzupassen. Auch hier ist zu sehen, dass sich oftmals noch auf keine standardisierte Analyse geeinigt wurde. So gibt es bspw. verschiedene Skalen für die Lautheit, und je nachdem welche Linie gleicher Lautstärke für die Messung herangezogen wird, kommt es zu einer anderen Frequenzabhängigkeit des Pegels.

Ein weiteres Ergebnis dieser Arbeit ist, dass geeignete Analysen auch hervorragende Hilfsmittel zur Optimierung von Übertragungs- und Wiedergabesystemen sein können.

So wurde gezeigt, dass durch Bestimmung von THD, THD+N, Klirrfaktor, Intermodulationsfaktor oder Differenztonfaktor Verzerrungen angegeben werden können. Durch diese lässt sich bspw. die Klangtreue von Audioequipment vergleichen oder die Verzerrung eines einzelnen Übertragungsglieds ausfindig machen. Außerdem wurde gezeigt, dass es Stereo- oder Mehrkanalsichtgeräte durch Korrelationsberechnungen und Betrachtung der Signale auf den einzelnen Kanälen ermöglichen, Aussagen zu den resultierenden Audiosignalen zu machen, um so die Kompatibilität einer Mischung zu gewährleisten.

Durch theoretische Erläuterungen, aber auch praktische Teile wie das Erstellen eines Störspektrums oder der Programmierung einer Anwendung zur Rauschunterdrückung wurde verdeutlicht, dass die Transformationen aus dem Zeit- in den Frequenzbereich in der Audiotechnik sehr bedeutend sind, um weitere Informationen über die Signalzusammensetzung zu erhalten. Dadurch wurde ebenfalls veranschaulicht, dass die schnelle Fourier-Transformation hierfür besonders effektiv eingesetzt werden kann. In einem weiteren Experiment wurde gezeigt, wie ein Raum durch Impulsantworten analysiert und modelliert werden kann, um dessen akustische Eigenschaften mit einem Faltungshall wiederzugeben. Hier wurden mit einem vergleichsweise geringen Arbeitsaufwand klanglich sehr gute Ergebnisse erzielt. Außerdem können durch Einstellungsmöglichkeiten innerhalb der Plugins zusätzlich kreative Anpassungen vorgenommen werden, die bei einer direkten Aufzeichnung innerhalb des Raums nicht möglich sind.

Im letzten Teil der Arbeit wurden Anwendungsmöglichkeiten künstlicher Intelligenz aufgezeigt und verdeutlicht, dass diese es Nutzern mit geringer Vorerfahrung ermöglichen, auch Arbeitsschritte, welche deren Kenntnisse normalerweise überschreiten würden, durchzuführen. Außerdem wurde durch die erläuterte Funktionsweise des akustischen Fingerabdrucks und der Klassifizierung von Musik gezeigt, dass diese durch Algorithmen auf bestimmte Eigenschaften untersucht und so schnell und automatisch erkannt, klassifiziert oder mit Metadaten versehen werden kann. Angesichts der stetig wachsenden Audiodatenbanken werden solche Algorithmen immer wichtiger, um diese strukturiert und durchsuchbar zu halten.

Durch den fundierten Überblick, den diese Arbeit liefert lässt sich abschließend sagen, dass in der Audiotechnik nur durch an die Zielsetzung angepasste Analysen, bestehende Verhältnisse ausreichend betrachtet und daraus bestmögliche Schlussfolgerungen gezogen werden können.

DANKSAGUNG

Für die Betreuung meiner Bachelorarbeit bedanke ich mich recht herzlich bei Herrn Prof. Oliver Curdt und Herrn Prof. Dr. Michael Felten. Vielen Dank auch für Ihre eindrucksvollen Vorlesungen, angebotenen Workshops und Exkursionen im Laufe meines Studiums. Für die hervorragende Assistenz bei der Umsetzung der Experimente bedanke ich mich bei Anna-Maria Straub und Hannah Brauner. Danke auch an Herrn Prof. George Tzanetakis für die Bereitstellung des Datensatzes.
Besonderer Dank für die seelische und moralische Unterstützung während meines Studiums und für das Korrekturlesen gilt meinen Eltern und Anna-Maria. Sowohl für das Korrekturlesen, als auch für die ambitionierten und immer von guter Laune getragenen gemeinsamen Projekte bedanke ich mich bei Christian Tobias. Danke auch an den Rest meiner Familie und Freunde, die hier zwar nicht aufgeführt sind, mich aber dennoch sehr in dieser besonderen Zeit unterstützt haben.

LITERATURVERZEICHNIS

Ableton (2021): Ableton Live Suite. Version 11.0.2, URL: https://www.ableton.com/de/live/, Stand: 5. Juli 2021.

Abraham, Ajith (2005): Artificial Neural Networks, in: Sydenham, Peter H./Thorn, Richard (Hrsg.): Handbook of measuring system design, Chichester, S. 901–908.

Algonaut (o.J.): Algonaut - Turn samples into music, URL: https://algonaut.audio/, Stand: 22. Juli 2021.

Allen, John B. (1988): Cochlear Signal Processing, in: Jahn, Anthony F. (Hrsg.): Physiology of the ear, New York/NY, S. 243–270.

Alloza, Pablo/Vonrhein, Benjamin (2019): Noise source localization in industrial facilities, URL: https://www.acoustic-camera.com/fileadmin/acoustic-camera/paper/2019_Internoise_Alloza_Vonrhein_NoiseSourceLocalization-Industrial-Facilities.pdf, Stand: 15. Juli 2021.

Beucher, Ottmar (2015): Signale und Systeme: Theorie, Simulation, Anwendung. Eine beispielorientierte Einführung mit MATLAB, 2. Aufl., Berlin/Heidelberg.

Brigham, Elbert Oran (1987): FFT. Schnelle Fourier-Transformation, 3. Aufl., München.

Brunton, Steven L./Kutz, Nathan J. (2017): Data Driven Science & Engineering. Machine Learning, Dynamical Systems, and Control, URL: http://databookuw.com/databook.pdf, Stand: 29. Juli 2021.

Brunton, Steven L./Kutz, Nathan J. (2019): Data-driven science and engineering. Machine learning, dynamical systems and control, Cambridge.

Butz, Tilman (2015): Fourier Transformation for Pedestrians, 2. Aufl., Cham.

Cockos Inc. (2019): REAPER. Version 5.981, URL: https://www.reaper.fm/download.php, Stand: 5. Juli 2021.

Dickreiter, Michael (2014): Analoge Tonmesstechnik, in: Dickreiter, Michael u. a. (Hrsg.): Handbuch der Tonstudiotechnik. 8., überarbeitete und erweiterte Auflage, Berlin/Boston, S. 613–656.

Dickreiter, Michael/Goeres-Petri, Jürgen (2014): Schallwahrnehmung, in: Dickreiter, Michael u. a. (Hrsg.): Handbuch der Tonstudiotechnik. 8., überarbeitete und erweiterte Auflage, Berlin/Boston, S. 115–135.

Dickreiter, Michael/Hoeg, Wolfgang (2014): Grundlagen der Akustik, in: Dickreiter, Michael u. a. (Hrsg.): Handbuch der Tonstudiotechnik. 8., überarbeitete und erweiterte Auflage, Berlin/Boston, S. 1–65.

Döbler, Dirk/Jensen, Sascha (2019): Berechnung und Darstellung von Schallquellen im dreidimensionalen Raum mit Mikrofonarrays und Beamforming, URL: https://www.acoustic-camera.com/fileadmin/acoustic-camera/paper/2019_DAGA_Doebler_Jensen_BerechnungDarstellungSchallquellen3D-Raum.pdf, Stand: 15. Juli 2021.

Fastl, Hugo/Zwicker, Eberhard (2007): Psychoacoustics. Facts and models, 3. Aufl., Berlin.

FLUX SOFTWARE ENGINEERING (2020): Stereo Tool V3. Version 20.12.0.49880, Frankreich, URL: https://www.flux.audio/project/stereo-tool-v3/, Stand: 30. Juni 2021.

Fraunhofer-Institut für Digitale Medientechnologie IDMT (o.J.a): Jamahook - Fraunhofer IDMT, URL: https://www.idmt.fraunhofer.de/de/business_units/m2d/references/jamahook.html, Stand: 22. Juli 2021.

Fraunhofer-Institut für Digitale Medientechnologie IDMT (o.J.b): SoundsLike - Fraunhofer IDMT, URL: https://www.idmt.fraunhofer.de/de/institute/projects-products/soundslike.html#2, Stand: 22. Juli 2021.

Görne, Thomas (2014): Tontechnik. Hören, Schallwandler, Impulsantwort und Faltung, digitale Signale, Mehrkanaltechnik, tontechnische Praxis, 4. Aufl., München.

Haitsma, J. A. (2002): Audio fingerprinting. a new technology to identify music, URL: https://pure.tue.nl/ws/portalfiles/portal/46792916/560633-1.pdf, Stand: 14. Juni 2021.

Haitsma, Jaap/Kalker, Ton (2002): A highly robust audio fingerprinting system, URL: http://ismir2002.ismir.net/proceedings/02-fp04-2.pdf, Stand: 18. Juni 2021.

Haitsma, Jaap/Kalker, Ton/Oostveen, Job (o.J.): Robust Audio Hashing for Content Identification, URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.16.2893&rep=rep1&type=pdf, Stand: 17. Juni 2021.

Harrison, Robert V./Hunter-Duvar, Ivan M. (1988): An Anatomical Tour of the Cochlea, in: Jahn, Anthony F. (Hrsg.): Physiology of the ear, New York/NY, S. 159–171.

Hellbrück, Jürgen (1993): Hören. Physiologie, Psychologie und Pathologie, Göttingen.

Henn, Hermann/Sinambari, Gholam Reza/Fallen, Manfred (2008): Ingenieurakustik. Physikalische Grundlagen und Anwendungsbeispiele, 4. Aufl., Wiesbaden.

Hoeg, Wolfgang u. a. (2014): Qualitätssicherung, in: Dickreiter, Michael u. a. (Hrsg.): Handbuch der Tonstudiotechnik. 8., überarbeitete und erweiterte Auflage, Berlin/Boston, S. 1267–1352.

Hoffmann, Rüdiger/Wolff, Matthias (2014): Intelligente Signalverarbeitung 1. Signalanalyse, 2. Aufl., Berlin.

iZotope (o.J.): iZotope and Assistive Audio Technology, URL: https://www.izotope.com/en/learn/izotope-and-assistive-audio-technology.html, Stand: 22. Juli 2021.

Jamahook Corp. (o.J.): Jamahook. The Sound Playground, URL: https://jamahook.com/, Stand: 22. Juli 2021.

Kuhn, George F. (1987): Physical Acoustics and Measurements Pertaining to Directional Hearing, in: Yost, William A. (Hrsg.): Directional Hearing, New York, S. 3–25.

Kümmritz, Sebastian/Kerscher, Michael (2019): The Acoustic Camera as a Tool for Room Acoustic Optimisation, URL: https://www.acoustic-camera.com/fileadmin/acoustic-camera/paper/2019_Internoise_Kuemmritz_Kerscher_Acoustic-Camera-as-tool-RoomAcousticOptimisation.pdf, Stand: 15. Juli 2021.

Lerch, Alexander (2012): An introduction to audio content analysis. Applications in signal processing and music informatics, Hoboken, NJ/Piscataway, NJ.

Lindsay, Peter H./Norman, Donald A. (1977): Human information processing. An introduction to psychology, 2. Aufl., New York.

Lippens, Stefaan u. a. (2004): A Comparison of Human and Automatic Musical Genre Classification, URL: https://www.researchgate.net/profile/George-Tzanetakis/publication/200806218_A_Comparison_of_Human_and_Automatic_Musical_Genre_Classification/links/0f31753ac84f3059ba000000/A-Comparison-of-Human-and-Automatic-Musical-Genre-Classification.pdf, Stand: 22. Juli 2021.

Mazzoni, Dominic u. a. (2021): Audacity. Version 3.0.0, URL: https://www.audacity.de/downloads/, Stand: 5. Juli 2021.

Meffert, Beate/Hochmuth, Olaf (2018): Werkzeuge der Signalverarbeitung.

Mertins, Alfred (2020): Signaltheorie. Grundlagen der Signalbeschreibung, Filterbänke, Wavelets, Zeit-Frequenz-Analyse, Parameter- und Signalschätzung, 4. Aufl., Wiesbaden.

Meyer, Andy/Döbler, Dirk (2006): Noise source localization within a car interior using 3D-Microphone Arrays, URL: https://www.acoustic-camera.com/fileadmin/acoustic-camera/paper/2006_BeBeC_Doebler_Meyer_Source_Localization_3D_micarrays.pdf, Stand: 15. Juli 2021.

Meyer, Andy/Döbler, Dirk (2014): Effiziente Methoden zur Analyse bewegter Schallquellen mit hochkanaligen Mikrofonarrays, URL: https://www.acoustic-camera.com/fileadmin/acoustic-camera/paper/2014_Daga_Doebler_Meyer_Analyse_MovingSoundSources.pdf, Stand: 15. Juli 2021.

Moore, Brian C. J. (2007): Cochlear hearing loss. Physiological, psychological and technical issues, 2. Aufl., Chichester/Hoboken, N.J.

Möser, Michael (2010): Messtechnik der Akustik, Berlin.

Müller, Gerhard/Vorländer, Michael/Möser, Michael (Hrsg.) (2017): Akustische Messtechnik, Berlin/Heidelberg.

Ohm, Jens-Rainer/Lüke, Hans Dieter (2014): Signalübertragung. Grundlagen der digitalen und analogen Nachrichtenübertragungssysteme, 12. Aufl., Berlin/Heidelberg.

Project Jupyter (2021): Jupyter Notebook. Version 6.2.0, URL: https://jupyter.org/install, Stand: 18. Juli 2021.

Puls, Thomas (2007): Verkehrslärm in der Diskussion. Möglichkeiten und Grenzen des Lärmschutzes in Deutschland, Köln.

Python Software Foundation (2020): Python. Version 3.9.0, URL: https://www.python.org/downloads/, Stand: 18. Juli 2021.

Sarang, Poornachandra (2021): Artificial Neural Networks with TensorFlow 2. ANN Architecture Machine Learning Projects, Berkeley CA.

Shah, Saqib (2021): Apple says Shazam identifies over a billion songs per month, URL: https://www.engadget.com/apple-shazam-billion-songs-140058264.html, Stand: 24. Juni 2021.

Shazam (o.J.): Unternehmen - Shazam, URL: https://www.shazam.com/de/company, Stand: 24. Juni 2021.

Slone, G. Randy (2002): The audiophile's project sourcebook, New York.

Sommer, Heini/Suter, Stefan/Höin, Reto (1998): Externe Lärmkosten des Verkehrs. Schlussbericht Vorstudie I, URL: https://www.ecoplan.ch/download/laer_sb1_de.pdf, Stand: 2. Juli 2021.

Thießen, Friedrich/Schnorr, Stephan (2005): Immobilien und Fluglärm, URL: http://www.windland.ch/doku_immo/509studieimmobienfluglaerm.pdf, Stand: 2. Juli 2021.

Tzanetakis, George (o.J.): GTZAN Genre Collection. Version 1, URL: http://marsyas.info/downloads/datasets.html, Stand: 20. Juli 2021.

Velardo, Valerio (2020a): How to implement a CNN for music genre classification, URL: https://github.com/musikalkemist/DeepLearningForAudioWithPython/blob/master/16-%20How%20to%20implement%20a%20CNN%20for%20music%20genre%20classification/code/cnn_genre_classifier.py, Stand: 20. Juli 2021.

Velardo, Valerio (2020b): Preprocessing audio data for deep learning, URL: https://github.com/musikalkemist/DeepLearningForAudioWithPython/blob/master/11-%20Preprocessing%20audio%20data%20for%20deep%20learning/code/audio_prep.py, Stand: 21. Juli 2021.

Voxengo (2021): SPAN. Version 3.11, URL: https://www.voxengo.com/product/span/, Stand: 12. Juli 2021.

Wang, Avery Li-Chun (2003): An Industrial-Strength Audio Search Algorithm, URL: https://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf, Stand: 25. Juni 2021.

Wilhelm, Richard (Hrsg.) (1928): Frühling und Herbst des Lü Bu We, Jena.

Zacharov, Nick (2018): Sensory Evaluation of Sound, Milton.

Zynaptiq (o.J.): ADAPTIVERB. Harmonic Tracking Resynthesis Reverb, URL: https://www.zynaptiq.com/adaptiverb/, Stand: 22. Juli 2021.

ANHANG

Anhang 1 Rauschunterdrückung mit FFT in Python[281]

```
import numpy as np
import matplotlib.pyplot as plt
plt.rcParams['figure.figsize']=[16,12]
plt.rcParams.update({'font.size': 18})

#Signal mit 3 Frequenzen erstellen
dt = 0.001
t = np.arange(0,1,dt)
f_clean = np.sin(2*np.pi*60*t) + np.sin(2*np.pi*100*t) + np.sin(2*np.pi*120*t)
f_noise = f_clean + 2.5*np.random.randn(len(t))
```

```
#FFT durchführen und LDS erstellen
M = len(t)
f_transformiert = np.fft.fft(f_noise,M)
LDS = f_transformiert * np.conj(f_transformiert) / M
freq = (1/(dt*M)) * np.arange(M)
L = np.arange(1,np.floor(M/2),dtype='int')
```

```
#Rauschanteile bzw. Leistungswerte unter Threshold auf 0 setzen und inverse FFT
indizes = LDS > 100
f_transformiert = indizes * f_transformiert
ifftfilter = np.fft.ifft(f_transformiert)

#Graphen plotten
print('\n')

fig,axs = plt.subplots(3,1, constrained_layout=True)

plt.sca(axs[0])
plt.plot(t,f_noise,color='silver', LineWidth=1.5,label='f_noise')
plt.plot(t,f_clean,color='k', LineWidth=2,label='f_clean')
plt.yticks(np.arange(-9,12,3))
plt.xlabel('t')
plt.ylabel('Amplitude')
plt.xlim(t[0],t[-1])
plt.legend()

plt.sca(axs[1])
plt.plot(freq[L],LDS[L],color='silver',LineWidth=2,label='f_noise')
plt.xlabel('f')
plt.ylabel('Leistung')
plt.xlim(freq[L[0]],freq[L[-1]])
plt.legend()

plt.sca(axs[2])
plt.plot(t,ifftfilter,color='k',LineWidth=2,label='gefiltert')
plt.yticks(np.arange(-9,12,3))
plt.xlabel('t')
plt.ylabel('Amplitude')
plt.xlim(t[0],t[-1])
plt.legend()

plt.show()
```

Abb. 26: Python Code zur Rauschunterdrückung mit FFT.

[281] Vgl. Brunton/Kutz (2019), S. 60 f.

Anhang 2 Datensatz für CNN vorbereiten[282]

```
import os
import librosa
import math
import json

DATASET_PATH = "Genre_Dataset"
JSON_PATH = "data_in_10_Segmente.json"

SAMPLE_RATE = 22050
DURATION = 30 #in seconds
SAMPLES_PER_TRACK = SAMPLE_RATE * DURATION
```

```
def save_mfcc(dataset_path, json_path, n_mfcc=13, n_fft=2048, hop_length=512, num_segments=5):

    #Daten schreiben in:
    data = {
        "mapping": [],
        "mfcc": [],
        "labels": []
    }

    num_samples_per_segment = int(SAMPLES_PER_TRACK / num_segments)
    expected_num_mfcc_vectors_per_segment = math.ceil(num_samples_per_segment / hop_length) # 1,2 -> 2
                                                                                            # aufs höhere gerun

    #Durch Genres Loopen
    for i, (dirpath, dirnames, filenames) in enumerate(os.walk(dataset_path)):

        #Sicher gehen dass man nicht im Root Level ist
        if dirpath is not dataset_path:

            #semantic Label speichern
            dirpath_components = dirpath.split("\\") #genre_dataset/blues => ["genre_dataset", "blues"]
            semantic_label = dirpath_components[-1]
            data["mapping"].append(semantic_label)
            print("\nProcessing {}".format(semantic_label))

            #File processen um auf Genres aufzuteilen
            for f in filenames:

                #Audiofile laden
                file_path = os.path.join(dirpath, f)
                signal, sr = librosa.load(file_path, sr = SAMPLE_RATE)

                #Segmente Processen Mfcc erstellen und abspeichern
                for s in range(num_segments):
                    start_sample = num_samples_per_segment * s #s=0 -> 0
                    finish_sample = start_sample + num_samples_per_segment #s=0 -> num_samples_per_segment

                    mfcc = librosa.feature.mfcc(signal[start_sample:finish_sample],
                                                sr=sr,
                                                n_fft=n_fft,
                                                n_mfcc=n_mfcc,
                                                hop_length=hop_length)
                    mfcc = mfcc.T

                    #Mfcc für Segment speichern wenn es die erwartete Länge hat
                    if len(mfcc) == expected_num_mfcc_vectors_per_segment:
                        data["mfcc"].append(mfcc.tolist())
                        data["labels"].append(i-1)
                        print("{}, segment:{}".format(file_path, s+1))

    with open(JSON_PATH, "w") as fp:
        json.dump(data, fp, indent=4)

if __name__ == "__main__":
    save_mfcc(DATASET_PATH, JSON_PATH, num_segments=10)
```

Abb. 27: Python Code zur Vorbereitung des Datensatzes für das CNN.

[282] Vgl. Velardo (2020b).

Anhang 3 CNN trainieren - Teil 1[283]

```
import json
import numpy as np
from sklearn.model_selection import train_test_split
import tensorflow.keras as keras
```

```
DATA_PATH = "data_in_10_Segmente.json"
SAVED_MODEL_PATH = "Genreklassifizierer.h5"
```

```
def load_data(data_path):
    #return X = Inputs und y = Targets
    with open(data_path, "r") as fp:
        data = json.load(fp)

    X = np.array(data["mfcc"])
    y = np.array(data["labels"])
    return X, y

def prepare_datasets(test_size, validation_size):

    #Daten laden
    X, y = load_data(DATA_PATH)

    #train/test split erstellen
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=test_size)

    #train/validation split erstellen
    X_train, X_validation, y_train, y_validation = train_test_split(X_train, y_train, test_size=validation_size

    #3d array -> (130, 13, 1)
    X_train = X_train[..., np.newaxis] # 4d array -> (num_samples, 130, 13, 1)
    X_validation = X_validation[..., np.newaxis]
    X_test = X_test[..., np.newaxis]

    return X_train, X_validation, X_test, y_train, y_validation, y_test

def build_model(input_shape):

    #model erstellen
    model = keras.Sequential()

    #1st conv layer
    model.add(keras.layers.Conv2D(32, (3, 3), activation='relu', input_shape=input_shape))
    model.add(keras.layers.MaxPool2D((3, 3), strides=(2, 2), padding='same'))
    model.add(keras.layers.BatchNormalization()) #models get faster

    #2nd conv layer
    model.add(keras.layers.Conv2D(32, (3, 3), activation='relu', input_shape=input_shape))
    model.add(keras.layers.MaxPool2D((3, 3), strides=(2, 2), padding='same'))
    model.add(keras.layers.BatchNormalization()) #models get faster

    #3rd conv layer
    model.add(keras.layers.Conv2D(32, (2, 2), activation='relu', input_shape=input_shape))
    model.add(keras.layers.MaxPool2D((2, 3), strides=(2, 2), padding='same'))
    model.add(keras.layers.BatchNormalization())

    #Output flatten und an Dense layer übergeben
    model.add(keras.layers.Flatten())
    model.add(keras.layers.Dense(64, activation='relu'))
    model.add(keras.layers.Dropout(0.3))

    #output layer
    model.add(keras.layers.Dense(10, activation='softmax'))

    return model
```

Abb. 28: Python Code für das CNN-Training - Teil 1.

[283] Vgl. Velardo (2020a).

Anhang 4 CNN trainieren - Teil 2[284]

```python
if __name__ == "__main__":
    #train, validation und test sets erstellen
    X_train, X_validation, X_test, y_train, y_validation, y_test = prepare_datasets(0.25, 0.2)

    #CNN bauen
    input_shape = (X_train.shape[1], X_train.shape[2], X_train.shape[3])
    model = build_model(input_shape)

    #Des Netz Compilen
    optimizer = keras.optimizers.Adam(learning_rate=0.001)
    model.compile(optimizer=optimizer,
                  loss="sparse_categorical_crossentropy",
                  metrics=['accuracy'])

    #CNN trainieren
    model.fit(X_train, y_train, validation_data=(X_validation, y_validation), batch_size=32, epochs=30)

    #CNN auf das test set evaluieren
    test_error, test_accuracy = model.evaluate(X_test, y_test, verbose=1)
    print("\nAccuracy on test set is: {}".format(test_accuracy))

    #Model Speichern
    model.save(SAVED_MODEL_PATH)
```

```
Epoch 1/30
188/188 [==============================] - 6s 29ms/step - loss: 1.6968 - accuracy: 0.4084 - val_loss: 1.5267 -
val_accuracy: 0.4193
Epoch 2/30
188/188 [==============================] - 5s 28ms/step - loss: 1.2139 - accuracy: 0.5733 - val_loss: 1.1939 -
val_accuracy: 0.5687
Epoch 3/30
188/188 [==============================] - 5s 29ms/step - loss: 1.0411 - accuracy: 0.6323 - val_loss: 1.0551 -
val_accuracy: 0.6453
Epoch 4/30
188/188 [==============================] - 5s 29ms/step - loss: 0.9273 - accuracy: 0.6767 - val_loss: 1.1111 -
val_accuracy: 0.6233
Epoch 5/30
188/188 [==============================] - 5s 29ms/step - loss: 0.8612 - accuracy: 0.6997 - val_loss: 1.0676 -
val_accuracy: 0.6393
Epoch 6/30
188/188 [==============================] - 5s 29ms/step - loss: 0.7835 - accuracy: 0.7324 - val_loss: 0.9108 -
```

Abb. 29: Python Code für das CNN-Training - Teil 2.

[284] Vgl. Velardo (2020a).