Fakultät I – Geistes- und Bildungswissenschaften
Institut für Sprache und Kommunikation
Studiengang Audiokommunikation und -technologie

Masterarbeit für die Prüfung zum Master of Science

Evaluation der akustischen Szenenklassifikation bei Cochlea-Implantaten

Clarissa Staudt
[redacted]

Betreuender Professor: Prof. Stefan Weinzierl
TU Berlin Fachgebiet Audiokommunikation
Betreuer Klinik: PD Dr. Dietmar Basta
HNO Klinik Zentrum für angewandte Medizintechnologie
Unfallkrankenhaus Berlin

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt gegenüber der Fakultät I der Technischen Universität Berlin, dass die vorliegende, dieser Erklärung angefügte Arbeit selbstständig und nur unter Zuhilfenahme der im Literaturverzeichnis genannten Quellen und Hilfsmittel angefertigt wurde. Alle Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind kenntlich gemacht. Ich reiche die Arbeit erstmals als Prüfungsleistung ein. Ich versichere, dass diese Arbeit oder wesentliche Teile dieser Arbeit nicht bereits dem Leistungserwerb in einer anderen Lehrveranstaltung zugrunde lagen.

Titel der schriftlichen Arbeit:

Evaluation der akustischen Szenenklassifikation bei Cochlea-Implantaten

Verfasserin:

Staudt, Clarissa Beate

[redacted]

Betreuende Dozenten:

Prof. Dr. Stefan Weinzierl

PD Dr. Dietmar Basta

Mit meiner Unterschrift bestätige ich, dass ich über fachübliche Zitierregeln unterrichtet worden bin und diese verstanden habe. Die im betroffenen Fachgebiet üblichen Zitiervorschriften sind eingehalten worden. Eine Überprüfung der Arbeit auf Plagiate mithilfe elektronischer Hilfsmittel darf vorgenommen werden.

Berlin, den 05.03.2019

Clarissa Beate Staudt

Zusammenfassung

Cochlea-Implantate dienen Menschen als Hörprothese, um akustische Reize verarbeiten zu können. Mit fortschreitender Entwicklung der Technik rückt die Musikerkennung zunehmend in den Fokus der Signalverarbeitungsstrategien. Diese Arbeit fokussiert sich daher darauf, wie akkurat Musik mit dem gegenwärtigen Stand der Technik detektiert werden kann und welche Faktoren Einfluss hierauf nehmen. Die SCAN-Automatik der Firma Cochlear detektiert mittels einer automatischen Situationserkennung unter anderem Musik. In verschiedenen Testszenarien wurden die Musikerkennungsrate und -zeit der SCAN-Automatik ermittelt und getestet, welche Faktoren signifikanten Einfluss darauf nehmen. Allgemein scheinen einzelne Genres von der Automatik besser erkannt zu werden als andere. Es zeigt sich, dass vor allem der Lautstärkepegel der Musik Einfluss auf die Musikerkennungsrate nimmt, nicht hingegen auf die Musikerkennungszeit. Störgeräusche haben zu einen gewissen Grad Einfluss auf Musikerkennungsrate und -zeit. Die Richtung, aus der die Störgeräusche abgespielt werden, zeigt hingegen allgemein keinen Zusammenhang zur Musikerkennungsrate oder -zeit auf. Zum weiteren Verständnis der Erkennungsmöglichkeiten von Musik wird mit einem an der TU Berlin entworfenen Audio Content Analyse Algorithmus zur Musikdiskrimination getestet, welche Trennschärfe zwischen zwei akustischen Gruppen erreicht werden kann. Die Ergebnisse dazu liegen zwischen 69-99%.

Abstract

Cochlear Implants serve humans as hearing prosthesis to process acoustic stimuli. As technology advances, music recognition is increasingly becoming the focus of signal processing strategies. The aim of this work is to examine systematically how accurate music can be detected with the technology available today and which factors influence the detection rate for music. The SCAN automatic by Cochlear unlimited detects music via automatic situation recognition. In different test scenarios the rate and time of music detection were tested and determined which factors take a significant influence on it.

The results show that some music genres are better detected than others. It is shown that the music sound level has an influence on the detection rate but not on the time to detect music. To a certain degree, there is a significant relation between noise and the music detection rate. The direction of the signal shows no significant impact on the music detection rate and -time. For better understanding of music detection, a specifically designed algorithm of audio content analysis from the TU Berlin is tested which accuracy between two groups of sounds lies between 69-99%.

Inhaltsverzeichnis

Eidesstattliche Erklärung ii

Zusammenfassung iii

Abstract iii

Inhaltsverzeichnis iv

Abbildungsverzeichnis vii

Tabellenverzeichnis ix

Abkürzungsverzeichnis xi

1. Einleitung und Forschungsfrage 1

2. Grundlagen 4

2.1. Schall 4

2.2. Digitale Signalverarbeitung 5

2.2.1. Fouriertransformation 5

2.2.2. Filterung und Cepstrum 7

2.3. Der Hörprozess 8

2.4. Psychoakustik 10

2.4.1. Lautheit - wahrgenommene Lautstärke 10

2.4.2. Tonheit - wahrgenommene Tonhöhe 12

2.5. Störungen im Hörprozess 13

2.6. Cochlea-Implantat (CI) 15

2.6.1. Aufbau 15

2.6.2. Funktionsweise 16

2.6.2.1. SCAN-Automatik 16

2.6.2.2. Datalogging 17

3. Methode zur Untersuchung der SCAN-Automatik 18

3.1. Auswahl der Musikbeispiele 18

3.1.1. Spektrogramme 19

3.1.2. Vorstellung der genrespezifischen Musikausschnitte 20

3.1.2.1. Blues ... 20
3.1.2.2. Disco ... 21
3.1.2.3. Electro ... 22
3.1.2.4. Hip Hop ... 23
3.1.2.5. Jazz ... 24
3.1.2.6. Klassik ... 25
3.1.2.7. Metal ... 26
3.1.2.8. Pop ... 28
3.1.2.9. Reggae ... 28
3.1.2.10. Rock ... 29
3.2. Messung ... 30
3.2.1. Versuchsaufbau ... 31
3.2.2. Messablauf ... 32
4. Untersuchungsergebnisse SCAN-Automatik ... 34
4.1. Einfluss des Nutzpegels auf die Musikerkennung in ruhiger Umgebung ... 34
4.2. Einfluss des Nutzpegels auf die Musikerkennung unter Störgeräuschen ... 39
4.3. Einfluss von Rauschen von vorne auf die Musikerkennung ... 41
4.4. Einfluss von Sprache von vorne auf die Musikerkennung ... 46
4.5. Einfluss der Richtung des Rauschens auf die Musikerkennung ... 49
4.6. Einfluss der Richtung von Sprache auf die Musikerkennung ... 53
5. Musikdiskrimination ... 57
5.1. Auswahl des Trainingsmaterials ... 59
5.2. Verwendete Features ... 59
5.2.1. 13 Mel Frequency Cepstral Coefficents ... 60
5.2.2. 13 Shifted Delta Cepstral Coefficients ... 61
5.2.3. Spectral Flux ... 62
5.2.4. Spectral Flatness ... 63
5.2.5. Spectral Centroid ... 64
5.2.6. Spectral Spread ... 64
5.2.7. Spectral Tonal Power Ratio ... 65

5.2.8. Zero Crossing Rate 65
5.2.9. Root Mean Square (RMS) Energy 65
5.2.10. Time Peak Envelope 66
5.3. Support Vector Machine 66
5.4. Dreifache Kreuzvalidierung 68
6. Ergebnisse des Musikdiskriminators 70
6.1. Musik-Genre und Noise 71
6.2. Musik-Genre und Sprache 71
6.3. Musik und Mix Musik-Noise im gleichen Verhältnis 72
6.4. Musik und Mix Musik-Noise im Verhältnis 60 dB zu 50 dB 73
6.5. Musik und Mix Musik-Noise im Verhältnis 80 dB zu 50 dB 73
6.6. Musik gegenüber Mix Musik-Sprache im gleichen Verhältnis 74
6.7. Musik und Mix Musik-Sprache 60 dB zu 50 dB 75
6.8. Genre und Mix Musik-Sprache 80 dB zu 50 dB 75
7. Zusammenfassung und Diskussion 77
8. Literaturverzeichnis 82
9. Anhang 86

Abbildungsverzeichnis

Abbildung 1: Das Ohr und seine Funktion (Möller, 2015)..... 9
Abbildung 2: Lautheitsfunktion nach Zwicker für 1 kHz nach Fastl und Zwicker (2007)..... 11
Abbildung 3: Kurven gleicher Lautstärke – Isophone nach Zwicker und Fastl (1999)..... 12
Abbildung 4: Mel-Skala nach Stevens et al (1937) aus Houtsma (1995)..... 12
Abbildung 5: Tonhöhenskalierung in Bezug auf die Basilarmembran nach Fastl und Zwicker (2007)..... 13
Abbildung 6: Das Hörorgan mit Cochlea-Implantat..... 15
Abbildung 7: Schematische Darstellung eines Cochlea-Implantats (Mühler, 2010) 16
Abbildung 8: Vergleich der Spektren von Metal (m07) oben mit weißem Rauschen unten..... 27
Abbildung 9: Schematische Darstellung Versuchsaufbau..... 31
Abbildung 10: Messaufbau – Beschallung von vorne..... 31
Abbildung 11: Messaufbau – Beschallung von vorne und hinten..... 33
Abbildung 12: Musikerkennungshäufigkeit und –zeit bei 60dB Nutzpegel in ruhiger Umgebung..... 35
Abbildung 13: Musikerkennungshäufigkeit und –zeit bei 80dB Nutzpegel in ruhiger Umgebung..... 35
Abbildung 14: Musikerkennungshäufigkeit und –zeit bei 60dB Nutzpegel unter 50dB Rauschen von vorne..... 42
Abbildung 15: Musikerkennungshäufigkeit und –zeit bei 80dB Nutzpegel unter 50dB Rauschen von vorne..... 43
Abbildung 16: Musikerkennungshäufigkeit und –zeit bei 60dB Nutzpegel unter 50dB Sprache von vorne..... 46
Abbildung 17: Musikerkennungshäufigkeit und –zeit bei 80dB Nutzpegel unter 50dB Sprache von vorne..... 47
Abbildung 18: Musikerkennungshäufigkeit und –zeit bei 60dB Nutzpegel unter 50dB Rauschen von hinten..... 50
Abbildung 19: Musikerkennungshäufigkeit und –zeit bei 80dB Nutzpegel unter 50dB Rauschen von hinten..... 51

Abbildung 20: Musikerkennungshäufigkeit und –zeit bei 60dB Nutzpegel unter 50dB Sprache von hinten 54
Abbildung 21: Musikerkennungshäufigkeit und –zeit bei 80dB Nutzpegel unter 50dB Sprache von hinten 54
Abbildung 22: Schematische Darstellung zur Musikerkennung 58
Abbildung 23: Filterbank mit Dreiecksfiltern zur mel frequency cepstrum coefficient (MFCC) Generierung nach Davis et al. (1980) 61
Abbildung 24: Berechnung der Shifted Delta Cepstral (Calvo et al., 2007) 62
Abbildung 25: Lineare Datentrennung zweidimensional links und dreidimensional rechts 67
Abbildung 26: Trennung der Klassen durch optimale Hyperplane 68
Abbildung 27: Spektrogramme für die Musikstücke im Genre Blues 86
Abbildung 28: Spektrogramme für die Musikstücke im Genre Disco 87
Abbildung 29: Spektrogramme für die Musikstücke im Genre Electo 88
Abbildung 30: Spektrogramme für die Musikstücke im Genre Hip Hop 89
Abbildung 31: Spektrogramme für die Musikstücke im Genre Jazz 90
Abbildung 32: Spektrogramme für die Musikstücke im Genre Klassik 91
Abbildung 33: Spektrogramme für die Musikstücke im Gerne Metal 92
Abbildung 34: Spektrogramme für die Musikstücke im Genre Pop 93
Abbildung 35: Spektrogramme für die Musikstücke im Genre Reggae 94
Abbildung 36: Spektrogramme für die Musikstücke im Genre Rock 95

Tabellenverzeichnis

Tabelle 1: Erkennungszeiten bei 60dB und 80dB Nutzpegel in ruhiger Umgebung 36
Tabelle 2: Einfluss Lautstärkepegel in ruhiger Umgebung auf die Erkennungshäufigkeit 36
Tabelle 3: Tests der Zwischensubjekteffekte 37
Tabelle 4: Mehrfachvergleiche 38
Tabelle 5: Ergebnisse des t-Tests für die Musikerkennungszeit bei 60dB und 80dB in ruhiger Umgebung 38
Tabelle 6: Einfluss des Nutzpegels unter verschiedenen Störgeräuschen auf die Musikerkennungshäufigkeit 39
Tabelle 7: p-Werte für den Levene-Test und der Zwischensubjekteffekte 40
Tabelle 8: Einfluss des Lautstärkepegels auf die Musikerkennungszeit 41
Tabelle 9: Durchschnittliche Erkennungszeit je Genre bei 60dB und 80dB Nutzgeräusch in ruhiger Umgebung und unter 50dB Rauschen 42
Tabelle 10: Einfluss von Rauschen von vorne auf die Musikerkennungshäufigkeit 44
Tabelle 11: Tests der Zwischensubjekteffekte 44
Tabelle 12: Mehrfachvergleiche 45
Tabelle 13: Tests der Musikerkennungszeiten für den Einfluss von Rauschen . 45
Tabelle 14: Durchschnittliche Erkennungszeit je Genre in ruhiger Umgebung und unter 50dB Sprache bei 60dB und 80dB Nutzgeräusch 47
Tabelle 15: Einfluss von Sprache von vorne auf die Musikerkennungshäufigkeit 48
Tabelle 16: Tests der Zwischensubjekteffekte 48
Tabelle 17: Mehrfachvergleiche 49
Tabelle 18: Einfluss von Sprache von vorne auf die Musikerkennungszeit..... 49
Tabelle 19: Durchschnittliche Erkennungszeit je Genre bei Rauschen von vorne und hinten bei 60dB und 80dB Nutzpegel 50
Tabelle 20: Einfluss der Richtung von Rauschen auf die Musikerkennungshäufigkeit 52
Tabelle 21: Tests der Zwischensubjekteffekte 52
Tabelle 22: Mehrfachvergleiche 53

Tabelle 23: Einfluss der Richtung des Rauschens auf die Musikerkennungszeit53
Tabelle 24: Durchschnittliche Erkennungszeit je Genre bei Rauschen von vorne und hinten bei 60dB und 80dB Nutzgeräusch 55
Tabelle 25: Einfluss der Richtung von Sprache auf die Musikerkennungshäufigkeit 55
Tabelle 26: Tests der Zwischensubjekteffekte 56
Tabelle 27: Einfluss der Richtung von Sprache auf die Musikerkennungszeit ... 56
Tabelle 28: Separierungsgenauigkeit für jedes Genre gegenüber Rauschen 71
Tabelle 29: Separierungsgenauigkeit für jedes Genre gegenüber Rauschen 72
Tabelle 30: Separierungsgenauigkeit für Musik gegenüber der Mix-Klasse Musik und Noise gleich gewichtet 72
Tabelle 31: Separierungsgenauigkeit für Musik gegenüber der Mix-Klasse Musik 60dB und Noise 50dB 73
Tabelle 32: Separierungsgenauigkeit für Musik gegenüber der Mix-Klasse Musik 80dB und Noise 50dB 74
Tabelle 33: Separierungsgenauigkeit für Musik gegenüber der Mix-Klasse Musik und Sprache gleich gewichtet 74
Tabelle 34: Separierungsgenauigkeit für Musik gegenüber der Mix-Klasse Musik 60dB und Sprache 50dB 75
Tabelle 35: Separierungsgenauigkeit für Musik gegenüber der Mix-Klasse Musik 80dB und Sprache 50dB 76

Abkürzungsverzeichnis

ACE	Advanced Combination Encoder
AGC	Automatic Gain Control
AD	Analog Digital Wandler
bzw.	beziehungsweise
BPF	Bandpassfilter
CI	Cochlea-Implantat
CIS	Continuous Interleaved Sampling
dB	Einheit Dezibel
DIN	Deutsches Institut für Normung
DFT	Discrete Fourier Transformation
F1	Formant eins
FFT	Fast Fourier Transformation
GTZAN	Music Genre Dataset nach G. Tzanetakis
HF	High Frequency
Hz	Einheit Hertz
JPEG	Bildformat nach Norm der Joint Photographic Experts Group
kHz	Einheit Kilohertz
MFCC	Mel Frequencies Cepstral Coefficients
MPEAK	Multipeak
MPEG	Multimedia Content Description Interface
MUSAN	Music, Speech and Noise Corpus
Pa	Einheit Pascal
RMS	Root Mean Square
s	Sekunde
SCAN	Automatik zur Erkennung für sechs verschiedene Situationen
SPEAK	Spectral-Peak-Strategy
STFT	Short Time Fourier Transformation
SDCC	Shifted Delta Cepstral Coefficients
UKB	Unfallkrankenhaus Berlin
wav	Wave Audiodatei
z.B.	zum Beispiel

1. Einleitung und Forschungsfrage

Das Cochlea-Implantat (CI) ist die technische Verwirklichung zur Ersetzung eines Sinnesorgans. Mittels eines Cochlea-Implantats ist es möglich, vorher taube oder stark hörgeschädigte Menschen an ihrer Umgebung akustisch teilhaben zu lassen. Dabei konzentriert sich die Entwicklung der Technik erfolgreich auf die Verständlichkeit von Sprache (Adunka und Kiefer 2005, Hughes 2013, Büchner und Gärtner 2017). Mit der Einführung von signalprozessorischen Verarbeitungsstrategien, wie sie bei Hörgeräten bereits seit längerem üblich sind, verbessert sich auch zusehends das Verstehen von Sprache in geräuschvoller Umgebung (Büchner und Gärtner 2017, Wilson et al. 2003). Die Studien von Mauger (2014) und Gilden (2016) haben beispielweise für den Soundprozessor CP900 der Firma Cochlear gezeigt, dass eine Verbesserung der Verständlichkeit von Sprache allgemein und in geräuschvoller Umgebung erreicht wird.

Neben der Verständlichkeit von Sprache ist der Musikgenuss für CI-Patienten ein entscheidender Faktor zur Erhöhung der Lebensqualität (McDermott, 2004). Viele CI-Träger berichten jedoch von Einschränkungen bei der Musikwahrnehmung nach einer CI-Operation (Gfeller et al. 2000, Migirov et al. 2009, Leal et al. 2003). Dies könnte u.a. daran liegen, dass CI-Träger Musik eher über Gesang erkennen als durch bestimmte musikalische Eigenschaften, wie Melodie und Harmonie (Leal et al. 2003). Auch Kong et al. (2004) zeigen, dass CI-Träger gegenüber Normalhörenden signifikant schlechter in der Identifizierung von komplexen musikalischen Mustern abschneiden.

Die Reduzierung der Komplexität von Musikstücken kann eine Möglichkeit sein, den Musikgenuss von CI-Trägern zu steigern (Kohlberg et al. 2015, Buyens 2017). Ferner zeigt sich, dass Musiktraining die Musikwahrnehmung bei CI-Patienten verbessern kann (Looi 2008, Looi et al. 2016, Petersen et al 2012). Neue, verbesserte Musikverarbeitungsstrategien könnten eine weitere Option sein, den Musikgenuss und die Musikwahrnehmung von CI-Trägern zu steigern (Kong et al. 2004 und Leal et al. 2003).

Wesentliche Voraussetzung für eine bessere Musikverarbeitungsstrategie ist jedoch eine zuverlässige Musikerkennung durch den CI Prozessor. Nur wenn

der Prozessor Musik erkennt, kann er anschließend entsprechend die Verarbeitungsstrategie anpassen. Mit der SmartSound iQ® mit SCAN, im folgenden SCAN-Automatik genannt, bietet die Firma Cochlear als einziger Hersteller eine offen erkennbare Musikerkennungsstrategie an. Die automatisierte Szenenklassifizierung unterscheidet zwischen sechs verschiedenen akustischen Situationen (Büchner und Gärtner, 2017).

Musik als akustisches Ereignis ist sehr viel komplexer als Sprache, da sie aus einer Vielzahl von Elementen besteht, die sich in unzähligen Kombinationen und Stilen zusammensetzen lassen (Leal et al. 2003). Entsprechend schwierig gestaltet sich daher auch die Musikerkennung in CI Geräten. Diese Arbeit soll sich daher der Frage widmen, wie akkurat Musik mit dem gegenwärtigen Stand der Technik detektiert werden kann und welche Faktoren Einfluss hierauf nehmen. Dafür werden beispielhaft zwei Musikerkennungssysteme betrachtet. Hierfür wird zunächst das bereits auf dem Markt erhältliche CI System, der Prozessor CP910 des Nucleus 6 CI Systems des Unternehmens Cochlear, hinsichtlich der Musikerkennung in verschiedenen Szenarien untersucht. Es wird vermutet, dass

- die Musikerkennungshäufigkeit der CI SCAN-Automatik genrespezifisch variiert
- sich mit zunehmendem Musiklautstärkepegel die Musikerkennungshäufigkeit und -zeit der CI SCAN-Automatik verbessert
- Störgeräusche die Musikerkennungshäufigkeit und -zeit der CI SCAN-Automatik beeinflussen
- die Richtung des Störgeräuschs die Musikerkennungshäufigkeit und -zeit der CI SCAN-Automatik beeinflusst.

Welcher Algorithmus und welche technische Funktionsweise sich hinter der SCAN-Automatik verbergen ist jedoch Betriebsgeheimnis. Ein an der TU Berlin entwickelten Musikdiskriminators, dessen Funktionsweise bekannt ist, wird daher zu Hilfe genommen, um zu überprüfen, welche technischen Möglichkeiten zur Trennschärfe zwischen zwei akustischen Ereignissen möglich sind.

In dieser Arbeit vereinen sich verschiedene thematische Felder, die für das weitergehende Verständnis dieser Arbeit notwendig sind. In Kapitel 2 werden

daher zunächst die Grundlagen zur Akustik und Psychoakustik erklärt, um zu verstehen wie sich Schall ausbreitet und vom Menschen wahrgenommen wird. Daneben wird auf die digitale Signalverarbeitung eingegangen, welche grundlegend für die Erstellung des Musikdiskriminators ist. Ferner wird der Hörprozess beschrieben, welche Störungen dabei auftreten können und wie das CI das Hörorgan bei einer Schädigung desselben ersetzen kann. Kapitel 3 befasst sich anschließend mit der Untersuchungsmethode der Musikerkennung des CP900 Prozessors. Musik zeichnet sich durch eine große Bandbreite an Genres und Stilen aus. Für die Untersuchung der Musikerkennungsfähigkeiten des CI Systems wurden daher verschiedene Musikrichtungen und Lieder ausgewählt, welche mithilfe von Spektrogrammen und einer kurzen musikalischen Interpretation des Musikstückes vorgestellt werden. Im 4. Kapitel werden schließlich die Hypothesen zur Musikerkennung der SCAN-Automatik untersucht. Das 5. Kapitel beschäftigt sich mit den Features des Musikdiskriminators. Die Ergebnisse zur Trennschärfe des Diskriminators werden anschließend im Kapitel 6 denen der Musikerkennung der SCAN-Automatik gegenübergestellt. Zusammenfassend wird die eingangs formulierte Fragestellung vor dem Hintergrund der Ergebnisse diskutiert.

2. Grundlagen

Nachfolgend werden einige Grundbegriffe erläutert, die zum grundlegenden Verständnis der verschiedenen, in die Arbeit involvierten Themengebiete beitragen sollen.

2.1. Schall

Schall wird verursacht durch ein schwingendes Element, die Schallquelle, welche das sie umgebende Medium anregt zu schwingen. Die Schallweiterleitung erfolgt im angeregten Medium Luft durch Druck- und Dichteänderungen, die sich wellenförmig ausbreiten und von der Schallsenke, beispielsweise dem menschlichen Ohr, als Schallereignis verarbeitet werden können (Möser, 2009). Durch das Hin- und Herschwingen der Luftteilchen entsteht ein Wechselschalldruck. Die Stärke der Luftdruckänderung wird durch den Effektivwert des Wechselschalldruckes ausgedrückt, als Schalldruck *p* mit der Einheit Pascal [Pa] bzw. $\frac{N}{m^2}$ (Möser, 2009).

Die physikalische Spannweite von Schalldrücken, die das menschliche Hörorgan verarbeiten kann, umfasst sieben Zehnerpotenzen von $2 \cdot 10^{-5}$ Pa bis 200 Pa (Möser, 2009). Das technische Verhältnismaß skaliert diese Spannweite anhand des Logarithmus auf eine handlichere Dynamik. Pegel werden in der Verhältniseinheit Dezibel [dB] angegeben und der Bezugswert ist die Hörschwelle, welche mit 0 dB angesetzt ist und bis zur Schmerzgrenze eine Dynamik von etwa 140 dB aufweist (Weinzierl, 2008). Der physikalische Bezugswert der Hörschwelle ist über den Schalldruck definiert als $p_0 = 2 \cdot 10^{-5} \frac{N}{m^2}$ bzw. über die Schallintensität als $I_0 = 10^{-12} \frac{W}{m^2}$ (Möser, 2009). Somit ergibt sich der Schallpegel nach Möser (2009) wie folgt

$$L_p = 10\, log\left(\frac{p^2}{p_0^2}\right). \qquad (1)$$

Äquivalent dazu lässt sich dies auch mit der Schallenergie ausdrücken, welche pro Zeit und Fläche als Schallintensität I übertragen wird und den Schallintensitätspegel liefert (Weinzierl, 2008):

$$L_I = 10\ log\left(\frac{I^2}{I_0^2}\right). \qquad (2)$$

Innerhalb der Hördynamik des Menschen wird von einer linearen Schallwellenausbreitung ausgegangen (Möser, 2009). Das hat den Vorteil, dass die Intensität zweier Schallquellen addiert werden bzw. umgekehrt jedes Schallereignis in seine einzelnen Teilschallwellen zerlegt werden kann. So kann mittels Fouriertransformation eine Schallwelle mathematisch durch die Summe aller Sinusschwingungen ausgedrückt werden. Informationen über Frequenz (Tonhöhe), Amplitude (Lautstärke) sowie Phasenlage der Sinusschwingungen zueinander beschreiben in der Summe letztlich die Schallwelle. Aus einem sich zeitlich verändernden Schallereignis kann somit zu jedem Zeitpunkt das Frequenzspektrum bestimmt werden, welches die Frequenzverteilung aller vorkommenden Frequenzen aufzeigt (Weinzierl, 2008).
Mittels des Frequenzspektrums lässt sich die Charakteristik eines Schallereignisses ablesen. Tonale Geräusche setzen sich aus einer Grundfrequenz und ihren ganzzahligen Vielfachen zusammen (Obertöne), Rauschen dagegen zeigt eine stochastische Verteilung über eine Vielzahl angeregter Frequenzen (Weinzierl, 2008).

2.2. Digitale Signalverarbeitung

Für die Betrachtung von akustischen Signalen in Bezug auf ihre spektralen Eigenschaften ist die digitale Signalverarbeitung ein nützliches Werkzeug, um beispielsweise mittels Fouriertransformation Signale zu analysieren.

2.2.1. Fouriertransformation

Zur digitalen Signalverarbeitung wird das kontinuierliche, analoge Audiosignal in eine diskrete Darstellung des Schalldrucks in Abhängigkeit von der Zeit überführt. Der endliche Speicher des Computers benötigt eine endliche Anzahl an Werten des Audiosignals, sowohl im Zeitbereich als auch im Wertebereich

(Amplitude). Der Zeitbereich des Schallsignals wird mittels der Abtastfrequenz (sampling frequency) von üblicherweise 44100 Samples pro Sekunde für Audiosignale auf CD in diskreten Abtastwerten (Samples) erfasst. Dabei ist das Nyquist-Theorem zu beachten, nach dem die höchste abbildbare Frequenz der halben Abtastrate entspricht, in diesem Fall 22050 Hz (Weinzierl, 2008). Das entspricht in etwa dem Frequenzspektrum, welches der Mensch mit seinem Ohr maximal auflösen kann (siehe 2.3).

Die Diskretisierung des Wertebereichs erfolgt über Quantisierung. Bei 16 Bit, werden 2^{16} = 65536 Amplitudenstufen ermöglicht.In der Audiosignalverarbeitung sind die im Audiosignal auftretenden Frequenzen zur Signalbestimmung relevant. Dafür bietet sich das Frequenzspektrum an (siehe 2.1). Mittels der Fouriertransformation ist die Wandlung vom Zeit- in den Frequenzbereich mathematisch durch die Analysefunktion (3) und die Rücktransformation mittels der Synthesefunktion (4) verlustfrei möglich (Weinzierl, 2008):

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi f t} dt \qquad (3)$$

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(f) e^{j2\pi f t} df \qquad (4)$$

Dabei steht *x(t)* für ein zeitkontinuierliches komplexes Signal. Die Rücktransformation vom Frequenz- zum Zeitbereich erfolgt über die inverse Fouriertransformation. Betrachtet man ein endliches und wertediskretes Signal, so wendet man die diskrete Fouriertransformation (discrete fourier transformation (DFT)) an. Eine effektive Rechenoperation bietet die schnelle Fouriertransformation (fast fourier transformation (FFT)). Voraussetzung dafür ist, dass die Länge des Signals im Optimalfall einer Multiplikation von Zweier-Potenzen entspricht (Möller, 2016; Weinzierl, 2008).

Eine weitere Rechenoptimierung erfolgt durch die Kurzzeit-Fouriertransformation (short time fourier transformation (STFT)). Anstelle der Frequenzdarstellung des gesamten Signals, zeigt sie die Veränderung der spektralen Zusammensetzung von einem Zeitpunkt zum nächsten an. Das

Signal wird dazu in zeitlich gleichgroße Blöcke (frames) unterteilt, die DFT für jeden einzelnen Block vollzogen und als Resultat ein Spektrogramm erzeugt. Die Blöcke werden jeweils gefenstert und somit die äußeren Samples des einzelnen Blocks weniger gewichtet, z.B. mittels Hann- oder Hammingfenster. Damit wird dem Problem der Verschmierung (Leck-Effekt) im Frequenzbereich entgegengewirkt, welches auftritt, wenn ungewünschte Frequenzanteile im zu analysierenden Block vorkommen (Oppenheim et al., 2004). Die Fensterform des Hann- bzw. Hamming-Fensters entspricht in etwa einer Gaußverteilung. Um keine zeitliche Information bei der abschwächenden Fensterung an den Rändern zu verlieren, sollten sich die Blöcke überlappen. Letztlich müssen sich Frequenz- und Zeitauflösung bei der STFT die Waage halten, da die höhere Auflösung des einen eine geringere Auflösung des anderen bewirkt (Oppenheim et al., 2004). Die STFT zeichnet sich schließlich durch die Parameter Blocklänge (blocksize = Anzahl der Samples), Schrittweite (Ausmaß der Überlappung beim Verschieben der Fensterfunktion) und Art der Fensterung aus.

2.2.2. Filterung und Cepstrum

Zur weiteren Verarbeitung wird das Audiosignal üblicherweise gefiltert, um somit gewisse Frequenzbereiche anzuheben, abzuschwächen oder auszulöschen. Die Filterung ist dabei sowohl im Zeit- als auch im Frequenzbereich möglich. Nachfolgend wird sich auf die Filterung im Frequenzbereich konzentriert.

Dafür wird das Spektrum mit einer Filterbank multipliziert, die einzelnen Frequenzbins werden somit gewichtet. Die Filterbank besteht dabei aus einer Anzahl sich überlappender Fensterfunktionen, die einen bestimmten Frequenzbereich betonen. Das Hervorheben eines gewissen Frequenzbereichs ist als Bandpass-Filterung bekannt. Die Filterbank wird auch zur Berechnung des Mel-skalierten[1] *Cepstrums* bzw. der mel frequency cepstral coefficients (MFCC)[2] benötigt. Das Cepstrum wurde erstmals von Bogert et al. erwähnt und

[1] Mel-Skala siehe Abschnitt 2.4 Psychoakustik
[2] Siehe Abschnitt 5.2.1

diente der Erkennung von Echos. Es ist eine Wortschöpfung aus dem Wort *Spectrum*, mit der ersten Silbe rückwärts gelesen (Möller, 2016).
Man erhält das Cepstrum durch Logarithmieren des Spektrums und die Anwendung der inversen Fouriertransformation darauf (Oppenheim und Schäfer, 2004).

$$\begin{aligned} C(x) &= F\{ln|G(j\omega)|\} \\ &= F\{ln|S(j\omega)| + ln|H(j\omega)|\} \\ &= C_1(x) + C_2(x) \end{aligned} \tag{5}$$

G(jω) ist die Ausgangsfunktion des Vokalfilters, *S(jω)* die Anregefunktion und *H(jω)* die Übertragungsfunktion des Vokaltraktfilters. Mathematisch bedeutend ist beim Cepstrum, dass durch den Logarithmus als nichtlineare Operation aus dem Produkt zwischen Eingangsspektrum und Übertragungsfunktion eine Addition als lineare Operation entsteht. Das wird vor allem bei der Sprachsignalanalyse genutzt, da „durch Filterung im Quefrency-Bereich Anregungssignal und Vokaltraktfilter getrennt werden können“ (Möller, 2016).
Hierbei entstehen weitere Wortschöpfungen, die von einem gewissen Humor der Schöpfer zeugt: Die Variable *x* steht für den quefrency Bereich (Silbendrehen aus frequency) und statt *filtering* heißt es nun *liftering* (Möller, 2016).

2.3. Der Hörprozess

Das Ohr fungiert für den Menschen als Empfänger für Schall. Es gliedert sich in drei Teile: Außen-, Mittel- und Innenohr. Jedem Part wird eine eigene Aufgabe zuteil. Das Außenohr sorgt durch seine spezielle Form und Platzierung am Kopf dafür, dass man sich anhand Pegel- und Zeitdifferenz orientieren kann, woher der Schall kommt. Außerdem wird der Schall durch die Resonanz des Ohrkanals und des äußeren Ohres um die sprachrelevante Frequenz (2,5 kHz) zusätzlich um 20 dB verstärkt (Ulrich et al., 2007). Eine weitere Verstärkung und Impedanzanpassung von Luft zu Flüssigkeit erfährt der Schall im Mittelohr durch die Flächentransformation von Trommelfell zu Steigbügelplatte und die

Hebelwirkung der Gehörknöchelchenkette. Im Innenohr wird schließlich der Schall mittels der Hörschnecke (Cochlea) in sein Frequenzspektrum aufgelöst und über den Hörnerv an das Gehirn weitergeleitet (Mrowinski et al., 2006). Dabei spielt nach der Wanderwellentheorie von Békesy vor allem die in Abbildung 1 gezeigte Basilarmembran eine entscheidende Rolle (Békésy, 1944). Sie trennt in der Hörschnecke die zwei Schneckengänge Scala Tympani und Scala Media voneinander. Durch ihren Aufbau (an der Schneckenbasis schmal, an der Schneckenspitze breit) weist sie ein tonotopisches Resonanzverhalten auf, nach dem hohe Frequenzen an der Schneckenbasis und tiefe Frequenzen an der Schneckenspitze über gereizte Haarsinneszellen verarbeitet werden (Ortsprinzip). Dieses Ortsprinzip gilt nicht nur für die Haarsinneszellen, sondern auch für die weiterverarbeitenden auditorischen Nervenzellen. So kann von einem normalhörenden Erwachsenen im Durchschnitt ein Frequenzbereich von 20 bis 20.000 Hz verarbeitet werden (Ulrich et al., 2007).

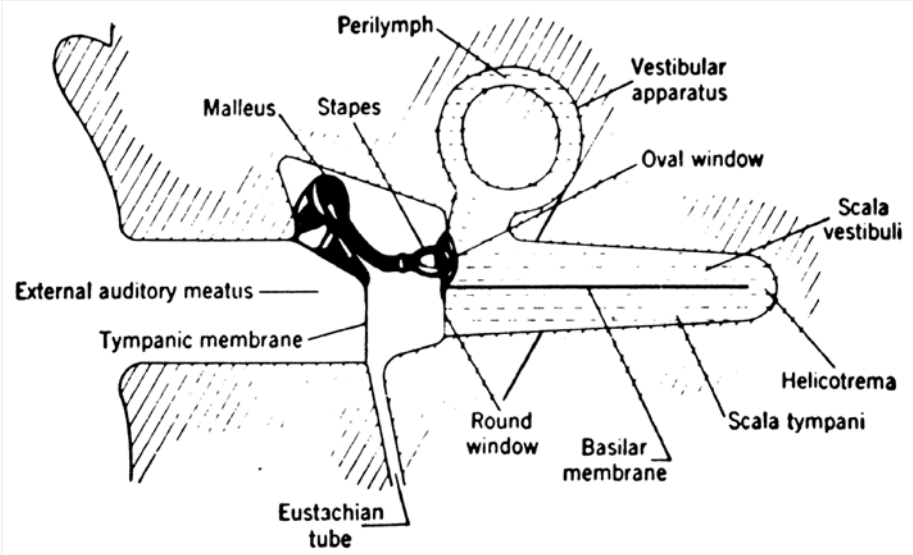

Abbildung 1: Das Ohr und seine Funktion (Möller, 2015)

Ferner befinden sich auf der Basilarmembran Haarsinneszellen, die an der Lautstärkewahrnehmung beteiligt sind. Je lauter das Signal, desto höher die Aktionspotenzialrate und desto mehr Haarsinneszellen werden angeregt (Janssen, 2014). Für das Periodizitätsprinzip gilt, dass tiefe Frequenzen bis etwa 1.000 Hz als zeitliches, neuronales Muster umgesetzt werden können

(Euler, 2004). Die genaue Art der Signalkodierung ist allerdings noch nicht vollständig erforscht (Battmer, 2009).

2.4. Psychoakustik

Wie in den Grundlagen zum Schall bereits erwähnt, umfasst der Hörbereich des Menschen in der Lautstärke von der Hörschwelle bis zur Schmerzgrenze einen Bereich von $2 \cdot 10^{-5}$ Pa bis 200 Pa. Zur Verarbeitung dieses ausgedehnten Wertebereiches bedarf es einer nichtlinearen Verarbeitung. Die folgenden psychoakustischen Größen der Lautheit und der Tonheit sind Empfindungsgrößen. Das heißt sie beschreiben keinen physikalischen Zustand, sondern die Stärke der menschlichen Empfindung bezüglich einer physikalischen Größe wie Frequenz oder Schallintensität. Daher werden sie auch Reizgrößen genannt, weil sie den Reiz, der auf das menschliche Ohr trifft, skalieren und die logarithmische Empfindlichkeit des Sinnesorganes verdeutlichen (Ulrich et al., 2007).

2.4.1. Lautheit - wahrgenommene Lautstärke

Der Lautstärkepegel definiert sich, wie in den Grundlagen zum Schall beschrieben, über das Verhältnis des Schalldrucks. Die Lautheit beschreibt die Lautstärke in Bezug auf die menschliche Wahrnehmung. Die Lautheitsfunktion nach Zwicker zeigt, wie sich die empfundene Lautstärke für einen 1 kHz-Ton verhält. Ab 40 dB Lautstärkepegel verläuft die Funktion linear: Pro 10 dB Erhöhung verdoppelt sich die Lautheit mit der Einheit Sone (Fastl und Zwicker, 2007). In Richtung der Hörschwelle (10 dB) nimmt die Lautheit schneller zu (siehe Abbildung 2).

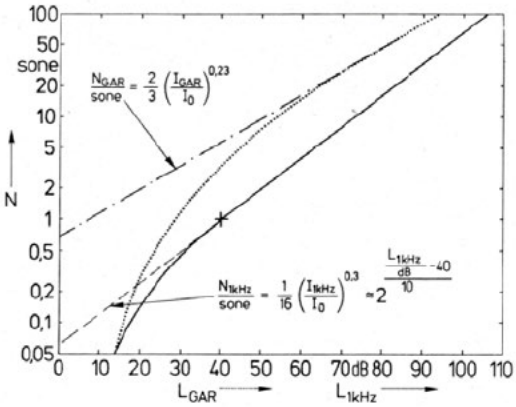

Abbildung 2: Lautheitsfunktion nach Zwicker für 1 kHz nach Fastl und Zwicker (2007)

Zu beachten ist, dass die Näherung nur für Töne über 30 dB gilt. Sie ist in der Norm ISO 1959 festgeschrieben und wird als Lautheitsskala eingesetzt. Das Lautheitsempfinden verhält sich für tiefe Frequenzen anders als für hohe (Ulrich et al., 2007).

Empirische Überprüfungen nach Fletcher und Munson zeigen, wie sich die subjektiv empfundene Lautstärke frequenzspezifisch verhält (Ulrich et al., 2007). Dazu wurden Versuchspersonen abwechselnd das zu bewertende Geräusch im Vergleich zu einem 1 kHz Standardton vorgespielt. Der Standardton ist anschließend auf einen gleich laut empfundenen Lautstärkepegel zum dargebotenen Geräusch einzuregeln. Die Frequenz des Standardtons wird beim Vergleichsverfahren auf 1kHz mit der DIN 1318 genormt. Der neu eingeregelte Schalldruckpegel des Standardtons ist durch die Maßeinheit der „Lautstärke“ und der Einheit Phon definiert.

$$L_n = 20 \lg\left(\frac{p_{1kHz}}{p_0}\right);\ p_0 = 2 \cdot 10^{-5} \qquad (6)$$

Bei 1 kHz gilt: Der Schalldruckpegel verhält sich äquivalent zum Lautstärkepegel.

Die Kurven gleicher Lautstärke (Isophone) verdeutlichen, dass die Empfindlichkeit des Menschen für Lautstärke bei etwa 3 kHz am höchsten ist (siehe Abbildung 3). Dies deckt sich auch mit dem Aufbau des äußeren Ohres, welches eine Resonanzfrequenz bei etwa 3 kHz erzeugt.

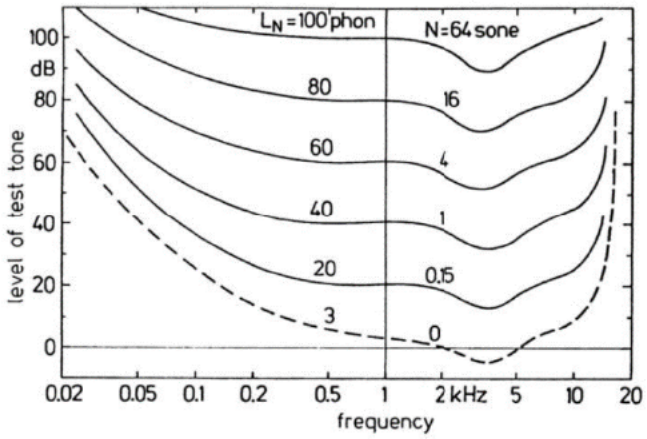

Abbildung 3: Kurven gleicher Lautstärke – Isophone nach Zwicker und Fastl (1999)

2.4.2. Tonheit - wahrgenommene Tonhöhe

Äquivalent zur Lautheit wird die Empfindungsgröße der empfundenen Tonhöhe bestimmt. Für die Tonhöhe gibt es die Möglichkeit der Verwendung einer Hörereignisskala, die zur Abbildung der Tonhöhenskala die Frequenzauflösung der Basilarmembran (Ortsprinzip) mit einbezieht (Möller, 2015). Den Zusammenhang von wahrgenommener Tonhöhe (Tonheit in mel) zur Frequenz (in Hz) gibt die Mel-Skala an, wie sie in Abbildung 4 zu sehen ist.

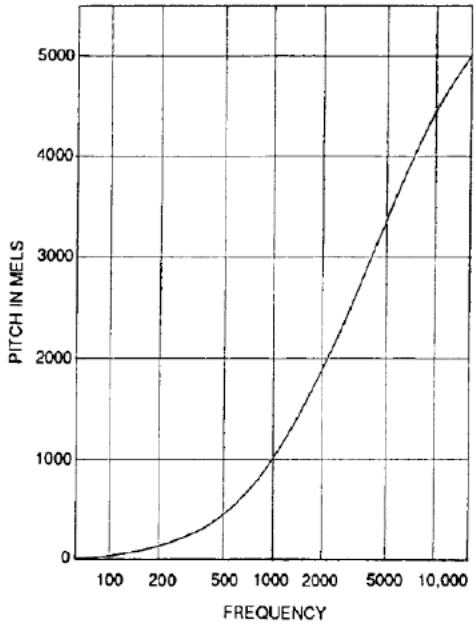

Abbildung 4: Mel-Skala nach Stevens et al (1937) aus Houtsma (1995)

Von Stevens et al. wurde bereits 1937 empirisch getestet, ab wann ein Ton doppelt so hoch wahrgenommen wird (Houtsma 1995). Dabei lässt sich

erkennen, wie die Tonheit bis 1.000 Hz linear empfunden wird und zu höheren Frequenzen annähernd logarithmisch verläuft (man beachte die logarithmische Aufteilung der X-Achse in Abbildung 4). Im Zusammenhang damit steht die Bark-Skala nach Fastl und Zwicker (2007). Hier wird der Begriff der *Frequenzgruppen* (critical bands) eingeführt (Fastl und Zwicker, 2007). So können für die menschliche Frequenzwahrnehmung 24 Frequenzgruppen ausgemacht werden. Jede einzelne umfasst die Wahrnehmungsbreite von 1 Bark = 100 mel. Während die Frequenzgruppen jeweils einen gleichlangen Teilbereich der Basilarmembran einnehmen (siehe Abbildung 5), steigt die Breite der Frequenzgruppe zur jeweiligen Frequenz proportional an. Das bedeutet, dass die Frequenzgruppenbreite für niedrige Frequenzen konstant 100 Hz beträgt, ab 500 Hz jedoch mit 20 % der Bandmittenfrequenz ansteigt (Möller, 2015).

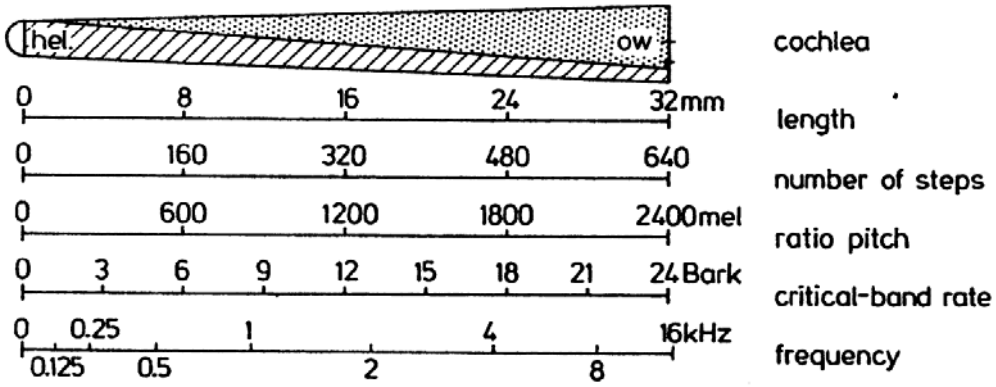

Abbildung 5: Tonhöhenskalierung in Bezug auf die Basilarmembran nach Fastl und Zwicker (2007)

2.5. Störungen im Hörprozess

Es gibt unzählige Möglichkeiten des Hörverlustes, die hier nicht im Einzelnen erläutert, sondern als grober Überblick umrissen werden.

Es wird unterteilt in die Beeinträchtigung der Schallleitung oder der Schallempfindung, je nachdem, welcher Teil des Ohres betroffen ist. Die Schallleitung findet über äußeres Ohr und Mittelohr statt, während die Schallempfindung im Innenohr vollzogen wird (Ulrich et al., 2007).

Ist das äußere Ohr blockiert, beispielsweise durch Cerumen (Ohrenschmalz), so kann der Schall schlecht an das Trommelfell gelangen und es kommt zu

einem dumpfen Klang der Schallereignisse. Durch professionelle Reinigung des Gehörgangs kann diese Art der Hörbeeinträchtigung behoben werden.

Im Mittelohr kann das Feder-Masse-System von Trommelfell und Gehörknöchelchen empfindlich gestört werden, zum Beispiel durch Trommelfellnarben oder -löcher sowie Schädigung der Gehörknöchelchenkette, wodurch die Verstärkung der Schallintensität beeinträchtigt ist. Dadurch kommt zu wenig Schallintensität am Innenohr an, um die Haarsinneszellen adäquat zu reizen (Mrowinski et al., 2006). Abhilfe ist hier unter anderem über ein Hörsystem oder ein Knochenleitungssystem BAHA (Bone anchored hearing aid) möglich. Während das Hörsystem am äußeren Ohr getragen wird und den Luftschall verstärkt, umgeht das BAHA System das äußere Ohr und das Mittelohr, indem es das Innenohr direkt über Vibrationen am Knochen anregt.

Liegt eine Innenohrschädigung bzw. kombinierte Schädigung von Mittel- und Innenohr vor, so kann der Hörverlust bis zu einem gewissen Grad durch Hörsysteme ausgeglichen werden. Im Innenohr werden zumeist die Stereozilien an den Haarsinneszellen durch zu hohe Schallintensitäten geschädigt. Dabei werden die Haarsinneszellen an der Basis der Hörschnecke meist zuerst funktionsunfähig, wodurch die Verarbeitung von hohen Frequenzen ungenügend erfolgt. Das Sprachverstehen wird erschwert, da hochfrequente Konsonantenteile fehlen und Sprache somit als undeutlich wahrgenommen wird (Ulrich et al., 2007).

Unterbleibt die Reizweiterleitung im Innenohr durch die Haarsinneszellen gänzlich, so werden Schallereignisse nicht verarbeitet. Ein Cochlea-Implantat kann die Hörwahrnehmung wieder ermöglichen. Voraussetzung dafür ist in jedem Fall ein intakter Hörnerv. Nach der aktuellen Leitlinie der Deutschen Gesellschaft für Hals-Nasen-Ohren-Heilkunde (Lenarz, 2012) ist das Cochlea-Implantat eine Alternative für folgende Patienten:

- Postlingual ertaubte und resthörige Kinder, Jugendliche und Erwachsene
- Prälingual taube (gehörlose) Erwachsene
- Prälingual taube (gehörlose) sowie perilingual ertaubte und resthörige Kinder
- Patienten mit einem Restgehör, welches für das Sprachverständnis nicht ausreicht trotz optimal eingestellter Hörgeräte.

2.6. Cochlea-Implantat (CI)

2.6.1. Aufbau

Das Cochlea-Implantat-System fungiert als Hörprothese und schafft bei einer Unterbrechung der Reizweiterleitung Abhilfe durch elektrische Reizung des Hörnervs. Es besteht aus zwei Teilen, wovon einer implantiert ist und einer außerhalb des Ohres getragen wird. Abbildung 6 zeigt das gesamte Cochlea-System am Ohr.

Der äußere Teil besteht aus dem Gehäuse, welches hinter der Ohrmuschel getragen wird (1). Darin sind sowohl Mikrofone als auch Signalprozessor und Energieversorgung durch Knopfzellen-Batterien oder Akkus untergebracht. Daran schließt sich eine Sendespule (2) an, die mittels Magnetismus am Kopf gehalten wird. Unter dem Schädelknochen sitzt das eigentliche Implantat (3), bestehend aus einem Titan- oder Keramikgehäuse, welches eine Empfangsspule und weiterverarbeitende Elektronik enthält. Daran schließen sich die Referenzelektrode und der Elektronenstrang an, der in die Cochlea eingeführt wird und der elektrischen Reizung des Hörnervs dient (4) (Cochlear limited)[3].

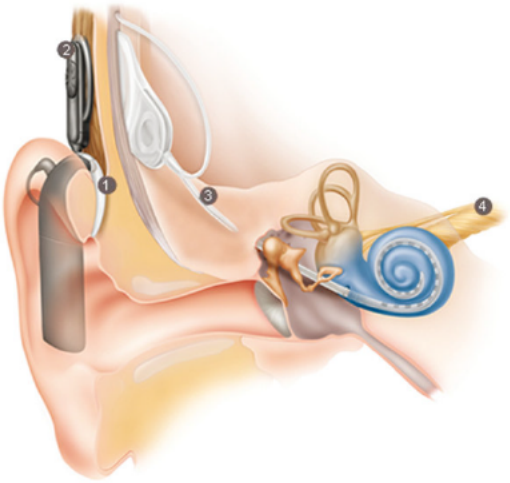

Abbildung 6: Das Hörorgan mit Cochlea-Implantat[4]

[3] Abgerufen am 06.09.2018
Hier verfügbar: https://www.cochlear.com/de/startseite/hoeren-und-hoerverlust/hoeren-und-hoerverlust/behandlungsmoeglichkeiten/elektrisch-akustische-implantate

[4] zur Verfügung gestellt von Cochlear limited: https://www.cochlear.com/de/startseite/hoeren-und-hoerverlust/hoeren-und-hoerverlust/behandlungsmoeglichkeiten/cochlea-implantate, Abgerufen 06.09.2018

2.6.2. Funktionsweise

Für die Schallaufnahme sind zwei Mikrofone verantwortlich, wodurch verschiedene Richtcharakteristiken ermöglicht werden. Wie in Abbildung 7 zu sehen, werden die akustischen Signale mittels des Analog-Digital (AD) Wandlers in digitale Signale umgewandelt, um vom Signalprozessor verarbeitet werden zu können. Im Signalprozessor finden unter anderem die automatische Situationserkennung sowie die Dynamikanpassung und die Aufteilung in mehrere Frequenzbänder durch Frequenzfilter statt. Über eine Kabelverbindung werden die elektrischen Signale an die Sendespule weitergeleitet und per Hochfrequenzübertragung transkutan, das bedeutet kabellos, an die implantierte Einheit übertragen. Die Empfangseinheit dekodiert das Signal und wandelt es in elektrische Pulse um, die den in der Cochlea platzierten Elektroden tonotopisch zugeteilt werden. So können, dem natürlichen Hören nachempfunden, hohe Frequenzen basal (an der Hörschneckenbasis) und tiefe Frequenzen apikal (an der Hörschneckenspitze) von den Spiralganglienzellen verarbeitet werden (Adunka, 2005). Voraussetzung dafür war die Entwicklung von mehrkanaligen Cochlea-Implantaten beispielsweise vom Unternehemen Cochlear mit 22 Elektroden auf dem in der Cochlea implementierten Elektrodenstrang. Dieser wird in der Scala Tympani platziert und stimuliert bei maximaler Einführtiefe somit ein Frequenzbereich von 100 Hz bis 10.000 Hz (Büchner, 2017). Entsprechend anders fällt die Hörwahrnehmung aus im Vergleich zu einem normalen Hörumfang (siehe 2.3).

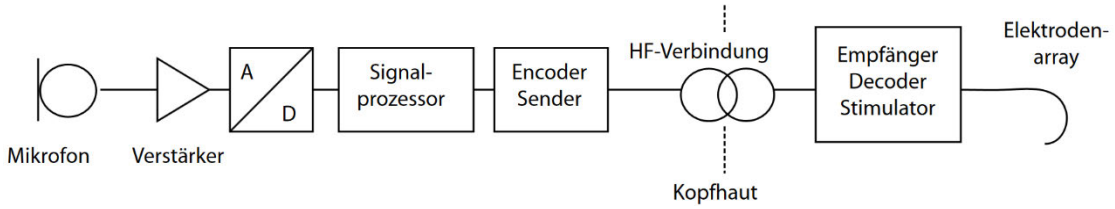

Abbildung 7: Schematische Darstellung eines Cochlea-Implantats (Mühler, 2010)

2.6.2.1. SCAN-Automatik

Die SCAN-Automatik des CI Soundprozessors CP900 unterscheidet zwischen den folgenden sechs akustischen Szenarien: Ruhe, Lärm, Sprache, Sprache im Störgeräusch, Wind, Musik. Der Algorithmus dafür basiert auf maschinellem

Lernen durch Training anhand von Beispielsets für verschiedene akustische Situationen. Etwa einmal pro Sekunde wird das Eingangssignal des Mikrofons auf Szenenzugehörigkeit überprüft. Ist die Szenenzugehörigkeit über mehrere Sekunden hin konstant, so wird das akustische Programm gewechselt und die Verarbeitungsstrategie an die akustische Szene angepasst, ohne dass eine Nachregelung durch den Nutzer erforderlich ist (Mauger, 2014). Die Verarbeitungsstrategien beziehen sich vor allem auf die Mikrofoncharakteristik (gerichtet oder omnidirektional), automatische Lautstärkeregelung, Störgeräuschunterdrückung sowie Sprachanhebung. Da sich diese Arbeit ausschließlich auf die Erkennung von Musik konzentriert, wird auf die Signalverarbeitungsstrategien nicht weiter eingegangen. Sie haben jedoch einen wesentlichen Einfluss auf den Höreindruck für den CI-Träger.

2.6.2.2. Datalogging

Gekoppelt an die Identifizierung akustischer Szenen durch die SCAN-Automatik ist das Datalogging. Dies ist ein aus der Hörgeräteindustrie übernommene Möglichkeit Hörsituationen automatisch zu protokollieren. Es wird angezeigt, wie lange das Gerät insgesamt getragen wird und in welchen akustischen Situationen sich der Träger aufhält, wie die Programmnutzung und Lautstärke Regelung durch den CI-Träger erfolgt und wie häufig die Spule vom Implantat getrennt wurde (Büchner, 2017). Solange Spule und Implantat verbunden sind, wird für die einzelnen, akustischen Szenen die Zeit, die sich der Nutzer darin aufhält, registriert und als Tagesdurchschnitt in der Software angegeben. So ergibt sich ein Bild zur akustischen Umgebung bzw. zum Alltag, in dem sich der Nutzer befindet. Dies haben sich auch die Studien zur Hörumgebung von CI Trägern von Busch et al. (2017) sowie von Oberhoffner (2018) zunutze gemacht. Dabei konnten anhand der Daten aus dem Datalogging unter anderem die Variation der akustischen Umgebung in Bezug auf das Alter ausgewertet werden. Ältere Menschen halten sich demnach weniger in musikalischen Umgebungen auf als Kinder. Mit der Überprüfung der SCAN-Automatik können entsprechend Rückschlüsse auf das Datalogging gezogen werden, wie akkurat die akustische Hörumgebung aufgezeichnet wird.

3. Methode zur Untersuchung der SCAN-Automatik

Musik in seiner akustischen Beschaffenheit ist sehr viel komplexer als Sprache oder Rauschen und dementsprechend schwieriger allumfassend zu detektieren. Im Folgenden wird die Musikauswahl vorgestellt, welche im Rahmen der Möglichkeiten eine möglichst große Musikdiversität widerspiegeln soll.

3.1. Auswahl der Musikbeispiele

Für die Auswahl der Musikstücke wurde auf das frei verfügbare *GTZAN Music Genre Dataset*[5] zurückgegriffen, das schon häufig zur Genreklassifikation genutzt worden ist (Sturm, 2013). Der Name leitet sich von Georg Tzanetakis ab, der den Datensatz zusammen mit Perry Cook erstellt hat. In ihrer Auswahl werden verschiedene Aspekte berücksichtigt. Die ausgewählten Musikstücke decken verschiedene Musikgenres ab, sind von gleicher Länge und liegen im selben Dateiformat vor. Es stehen insgesamt 1.000 Musikausschnitte zur Verfügung, mit jeweils 100 Titeln, die sich auf die folgenden zehn Genres aufteilen: Blues, Klassik, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae und Rock. Aus jedem Genre der GTZAN Datenbank werden jeweils zehn Titel genutzt, mit Ausnahme von Country Musik. Dieses Genre wird mit zehn eigens generierten Musiksamples der elektronischen Musik ersetzt, um die akustische Diversität zu steigern. Die ausgewählten Electro-Lieder weisen in ihrer Gesamtlänge teilweise Material von über einer Stunde auf. Das bedeutet, dass innerhalb eines Liedes starke Unterschiede in der Musikstruktur zu finden sind. Daher wurden hier verschiedene Musikausschnitte aus ein und demselben Musikstück generiert. Alle 100 ausgewählten Musikausschnitte sind jeweils 30 Sekunden lang und liegen im wav-Dateiformat mit 22050 Hz Mono als 16-bit Audio File vor. Der Vorteil von wav-Dateien ist, dass keine Datenkompression vorliegt. Dementsprechend ist die Rohdatei verlustfrei bezogen auf die Qualität der Aufnahme, aber auch relativ groß. Um eine Testzeit von 60 Sekunden zur Verfügung zu stellen, wurde jeder Titel doppelt abgespielt. Die Auswahl der Samples erfolgte willkürlich, es wurde jedoch darauf geachtet, eine möglichst hohe akustische Bandbreite zu generieren. Mittels Matlab wurden alle ausgewählten Titel sowie Rauschen und Sprache RMS (Root Mean Square)

[5] Abgerufen am 1.8.2018, hier verfügbar: http://marsyas.info/downloads/datasets.html

normalisiert. Dazu wurden Funktionen der *AK-Toolbox* des Fachbereiches Audiokommunikation und -technik der TU Berlin implementiert.[6]
Der Datensatz enthält keine Angaben zu Interpret und Titel. Diese Informationen wurden von Bob L. Sturm zusammengetragen und zugänglich gemacht[7], der im Jahr 2013 den Datensatz in einer Untersuchung analysierte. Zur besseren Übersicht wurden die Titel aus dem Paper herausgesucht und im Kapitel 3.1.2 aufgelistet.

3.1.1. Spektrogramme

Das Prinzip der Musikerkennung innerhalb der SCAN-Automatik ist unbekannt. Um dennoch ein Verständnis für ein mögliches Muster aufzubauen nach der die Musikerkennung arbeitet, sollen die ausgewählten Titel in Bezug auf ihre spektralen Eigenschaften näher betrachtet werden. Dazu wurde zu jedem Musikstück ein Spektrogramm erstellt, die sich auf dem beiliegenden Speichermedium befinden. Im Anhang finden sich die Spektrogramme für alle zehn Musikausschnitte pro Genre als Übersicht.
Ein Spektrogramm ist die Verbildlichung von Audiosignalen. Die X-Achse zeigt den Zeitverlauf, die Y-Achse die vorkommenden Frequenzanteile, und anhand der Farbe ist feststellbar, in welcher Intensität die Frequenzen zum jeweiligen Zeitpunkt auftreten. In den vorliegenden Dateien zeigt die Farbe Grau, dass keine Signalintensität vorliegt. Die Intensität steigert sich von Blau über Rot bis hin zu einem hellen Gelb, welches die stärkste Intensität verbildlicht. Ein Klangereignis besteht aus einem Frequenzgemisch. Klänge, die von der menschlichen Stimme oder einem Instrument erzeugt werden, weisen die Besonderheit auf eine Grundfrequenz und eine bestimmten Anzahl von spektralen Vielfachen dieser Grundfrequenz zu erzeugen, auch Obertöne oder Oberwellen genannt (Sengpielaudio)[8]. Die Grundfrequenz und die entsprechenden Vielfache passend zu Stimme oder Instrument und sind in Rot bzw. Gelb meist gut zu erkennen.

[6] Abgerufen am 10.8.2018,
hier verfügbar: https://www.ak.tu-berlin.de/menue/publications/open_research_tools/aktools/
[7] Abgerufen am 10.8.2018,
hier verfügbar: http://www.eecs.qmul.ac.uk/~sturm/research/GTZANindex.txt abgerufen am
[8]Abgerufen am 10.8.2018,
hier verfügbar: http://www.sengpielaudio.com/Rechner-harmonische.htm

Im Folgenden werden die verwendeten Musikausschnitte aufgeführt und im Zusammenhang mit den Spektrogrammen kurz analysiert und interpretiert. Die Abkürzung vor Interpret und Titel setzt sich zusammen aus dem Anfangsbuchstaben des Genres und der Titelnummer des GTZAN-Dataset. Das Genre Electro, welches aus eigenen Musikdateien generiert wurde, ist dagegen lediglich durchnummeriert.
Um die Beziehung zwischen den Musikausschnitten und der SCAN-Automatik zu verdeutlichen, werden einige Ergebnisse der Musikerkennungsrate der CI SCAN-Automatik vorweggenommen und hier mit aufgeführt.

3.1.2. Vorstellung der genrespezifischen Musikausschnitte

3.1.2.1. Blues

Für das Genre Blues wurden folgende Lieder ausgewählt:

- b09: John Lee Hooker: “It´s My Own Fault”
- b16: Robert Johnson: “Last Fair Deal Gone Down”
- b21: Robert Johnson: “When You Got A Good Friend”
- b29: Kelly Joe Phelps: “The House Carpenter”
- b35: Kelly Joe Phelps: “Capman Bootman”
- b37: Kelly Joe Phelps: “Piece By Piece”
- b40: Stevie Ray Vaughan: “Love Struck Baby”
- b45: Stevie Ray Vaughan: “Rude Mood”
- b69: Clifton Chenier: “I'm On My Way, Part 2”
- b83: Buckwheat Zydeco: “Drivin’ Old Grey”

Liederübergreifend sind Gitarre und Gesang als wesentliche Merkmale der oben ausgewählten Lieder zu erkennen. Eine oder zwei Gitarren begleiten den Gesang der ersten vier Musikausschnitte (b09 – b37). In b09 ist zusätzlich ein Piano und ein Bass wahrzunehmen, außerdem ein sparsam eingesetztes Percussionelement. In b37 übernimmt die Mundharmonika die Melodie. Die Lieder b40 bis b83 sind zusätzlich zur Gitarre mit Percussion begleitet. Dies ist im Spektrogramm gut durch die breitere Rotfärbung in hohen Frequenzen zu sehen.
Der Gesang erfolgt bei der Liederauswahl ausschließlich von männlichen Stimmen auf Englisch. Gut hörbar ist der Stilwandel zwischen alten und neuen

Stücken. Während im Jahr 1936 Robert Johnson (b16) seinem Gesang mehrere kurze Crescendi und Decrescendi verleiht und die Tonhöhe bluestypisch oft im Glissando von unten her erreicht, erklingt 1999 bei Kelly Joe Phelps (b29, b35) ein ebenmäßigerer Klang in Stimme und Gitarre. Diesen kontrastiert Joe Phelps in seinem Stück „Piece by Piece“ (b37) mit der Mundharmonika, deren Spielweise stark an die oben erwähnte Gesangsweise Robert Johnsons erinnert.

Mit dem zweiten Stück von Vaughan (b45) steht ein rein instrumentales Stück innerhalb des gesangsbetonten Genres zur Verfügung. Eindeutig auch zu erkennen am Spektrogramm, wo die Obertöne der Gitarre zwar noch zu erkennen sind, aber keine ausgeprägten Oberwellen wie beim Gesang zu sehen sind.

Von der SCAN-Automatik wurden alle Blues Musikstücke mindestens einmal im gesamten Testset als Musik klassifiziert.

3.1.2.2. Disco

Die Titel, welche das Genre Disco repräsentieren, stammen überwiegend aus den späten 1970er Jahren. Zwei Songausschnitte stammen aus den Jahren 1988 (d18) und 1990 (d23). Folgende Disco Samples wurden ausgewählt:

- d06: Alec R. Constandinos: “Romeo and Juliet“
- d09: Osibisa: “Get Up”
- d14: Bee Gees: “Staying Alive”
- d18: Sabrina: “S.E.X.”
- d23: Latoya Jackson: “Playboy (Be Me)”
- d35: The Gibson Brothers: “Better Do It Salsa”
- d42: Disco-Tex and Sex-O-Letters: “I Wanna Dance Wit’ Choo”
- d53: Amii Stewart: “Knock On Wood”
- d57: Boney M.: ”Auld Lang Syne”
- d74: KC and the Sunshine Band: “Get Down Tonight”

Prägend für Disco ist der eingängige Rhythmus, der zum Mitklatschen, -wippen, -tanzen animiert. Die rhythmusgebende Percussion ist in den Spektrogrammen sehr gut zu sehen, da sie eine senkrechte rote Linie von den tiefen bis zu den hohen Frequenzen zieht.

Ebenso typisch für das Genre ist der Gesang, der zum Refrain meist mehrstimmig auftritt. Wie beim Blues (b45) ist auch bei Disco ein gesangloses Musiksample integriert (d53). In d09 dominiert das Blasinstrument.
Neben den rhythmusgebenden Komponenten von Schlagzeug, Trommel etc. ist überwiegend das synthetische Piano vernehmbar (z.B. d06, d14, d18, d23, d35).
Die Stücke d06, d14 und d23 wurden von der SCAN-Automatik generell nicht als Musik klassifiziert, sondern dem Lärm zugeordnet.

3.1.2.3. Electro

Die Electro Musikausschnitte sind aus eigener Musikdatenbank generiert und dienen zur größeren akustischen Bandbreite.

- e01: Carsten Schorr: VA Minimal Techno Vol.3
- e02: The Chemical Brothers: In Dust We Trust
- e03: The Chemical Brothers: In Dust We Trust
- e04: Gregor Dambrovsky: Minimal House 2
- e05: Micha Moor: Space
- e06: The Chemical Brothers: Chemical Beats
- e07: Carsten Schorr: Minimal Techno Mix
- e08: The Chemical Brothers: Song to the Siren
- e09: Carsten Schorr: VA Minimal Techno Mix
- e10: Carsten Schorr: VA Minimal Techno Mix

In den vorliegenden Electro-Musikausschnitten leben die Kompositionen von Loops, Scratches sowie synthetischen (Percussion)-Sounds und verzerrten Klängen, denen keine eindeutige Tonhöhe mehr zuzuordnen sind. Gesangliche Teile sind bruchstückhaft enthalten, die Tonhöhe tritt generell, zugunsten der starken rhythmischen Entwicklungen, in den Hintergrund. Die Musikstücke grenzen an eine Abfolge von Geräuschen und Geräuschimitationen. Geloopte menschliche Stimmen (e01, e02, e06, e08) werden Teil eines rhythmischen Geflechts: Analog erzeugte Klänge, beispielsweise vom Drumset oder E-Bass, setzen sich oft von digital erzeugten Sounds ab. Die Beats sind durch die senkrecht von tiefen bis zu hohen Frequenzen verlaufenden roten Linienabfolgen gut auszumachen (z.B. e04, e07). Je mehr künstliche Klänge vermischt wurden, desto dichter ist das Spektrum mit roten Linien und gelben

Spots angereichert. Bei Klängen mit harmonischen Vielfachen (Oberwellen) ergeben sich zusammen mit dem Beat karierte bildliche Strukturen im Spektrogramm (z.B. e10). Klare diagonale gelbe Linien sind in e03 zu sehen, die mit fortschreitender Zeit von hohen zu niedrigen Frequenzen abfallen. Sie verbildlichen den prägnant herauszuhörenden Pfeifton, der schrittweise seine Tonhöhe verringert. Insgesamt zeichnet sich das Genre durch eine wiederholende Regelmäßigkeit der Beats bzw. Sounds aus und wird relativ gut erkannt (8 von 10 bei 80 dB).

3.1.2.4. Hip Hop

Bei diesen Musiksamples liegt der Fokus auf dem Beat und dem dazugehörigen Sprechgesang. Der Text wird als wesentliches Stilmittel eingesetzt und kann zuweilen gezielt aggressiv sein (z.B. h42).

- hh02: De la Soul: “Buddy“
- hh15: Beastie Boys: “Bodhisattva Vow”
- hh21: Beastie Boys: “The Negotiation Limerick File”
- hh26: 3LW: “No More (Baby I’ma Do Right)”
- hh37: Lil Bow Wow: “Bounce With Me”
- hh42: Wu-Tang Clan: “Triumph”
- hh45: Lauryn Hill: “Every Ghetto, Every City”
- hh46: A Tribe Called Quest: “Check The Rhime”
- hh59: Cypress Hill: “Hits From The Bong”
- hh89: Public Enemy: “Power To The People”

Die hier vertretenen Musikstücke stammen überwiegend aus den 1990er Jahren. Einige Lieder stammen aus den 2000ern (hh15, hh26, hh37). Es wird vorwiegend von Männern gerappt (z.B. hh02, hh15, h89), Frauenstimmen werden als Hintergrundstimme oder für den Refrain eingesetzt (hh26). Der Sprechgesang orientiert sich stark am vorgegebenen Beat und wechselt zwischen Wortkunst und Zwischenrufen passend zum Beat. Teilweise ist die Stimme auch verzerrt und vereint sich mit synthetischen Klängen (hh15) oder die Stimme ist als Sprechgesang über gescratchte Klänge gelegt (h89). Vereinzelt mischen sich mehrere Stimmen zusammen wie bei einem Gesprächsaustausch (z.B. h21).

In dem Titel von Lauryn Hill (hh45) ist ihre Stimme weniger als Rap, sondern überwiegend gesanglich eingesetzt, was von weiteren Frauenstimmen untermalt wird.
Die SCAN-Automatik klassifiziert die drei Stücke hh15, hh42 und hh89 generell nicht als Musik, sondern überwiegend als Lärm (hh15, hh89) und Sprache im Lärm (hh42). Einmal wird hh89 auch der Sprache zugeordnet. Möglicherweise ist die Klangstruktur zu dicht um Musik klar zu definieren. Erschwerend kommt dazu, dass die Anfänge der Musikausschnitte keinen klaren Rhythmus erkennen lassen oder dieser sehr schnell ist (hh89), so dass eine eindeutige Zuordnung zur Musik nicht erfolgt.

3.1.2.5. Jazz

Die ausgewählten Jazzstücke wurden überwiegend in den 90er Jahren aufgenommen, zwei Lieder sind aus den späten 80er Jahren (j89, j81). Das älteste Stück (j87) ist von 1959 und das neuste Werk (j75) von 2017.

- j04: James Carter: “The Intimacy Of My Woman´s Beautiful Eyes”
- j05: James Carter: “1944 Stomp”
- j23: Joe Lovano: “Blues Not To Lose“
- j32: Ranforld Marsalis Trio: “Beat’s Remark”
- j35: Coleman Hawkins: “April In Paris”
- j40: Coleman iHawkins: “She Is FunnyThat Way”
- j75: Dexter Gordon: “Broadway”
- j80: Tony Williams: “Life Of the Party”
- j81: Tony Williams: “Geo Rose”
- j87: Miles Davis: “So What”

Bei den Jazzausschnitten handelt es sich um verschiedene Besetzungen. Begleitet wird immer vom gezupften Kontrabass und vom Drumset, meist auch vom Klavier (außer bei j23, j32), gelegentlich tritt eine Bläsergruppe hinzu (Saxophone und Trompeten). Die Soli oder Leads (Hauptstimmen) werden mal vom Bass (j87), mal vom Tenor- (j75) bzw. Alt-Saxophon (j04, j05), von der Trompete - mit (j32) bzw. ohne Dämpfer (j81) - oder von einer Sängerin (j40) bestritten. In j23 ist ein angedeutetes Schlagzeugsolo zu vernehmen. In j80 spielen Saxophon und Trompete unisono im Oktavabstand, in j23 zwei Alt-Saxophone unisono.

In den Spektrogrammen sind die instrumentalen Oberwellen gelb zu rot deutlich erkennbar und auch das Vibrato des Saxophons gut zu sehen (z.B. j04). Genauso ist der Beat des Schlagzeuges aus den senkrechten Impulslinien zu erkennen. Dabei sind die Percussion-Klänge dominant in hohen Frequenzen mit häufiger Nutzung der Hi-Hat Schellen. Das Schlagzeug ist beim Jazz selten als basslastiges Instrument genutzt, sondern betont die Musik durch weichere Untermalung. Zusätzlich spielt der Kontrabass in diesen Beispielen nie con arco, sondern – wie im Jazz-Trio üblich – immer gezupft, was durchgehende Bassfrequenzen ebenfalls selten macht.
Lediglich ein Stück (j40) wird von Gesang begleitet, die anderen sind rein instrumental. Die frequenzübergreifend intensivste Anregung erfolgt in j75, welches als einziges Stück nicht als Musik klassifiziert, sondern überwiegend „Sprache im Störgeräusch“ zugeordnet wird. Möglicherweise liegt das an dem deklamierten Saxophon. Über den Zeitraum spektral am wenigsten Intensität ist bei j87 zu beobachten.

3.1.2.6. Klassik

Das klassische Genre in dieser Auswahl sticht musikalisch damit heraus, vorwiegend instrumental (größtenteils Streicher), ohne Gesang, Beats oder synthetische Klänge vertont zu sein. Die Auflistung der Musikausschnitte für das Genre Klassik ist wie folgt:

- k01: Johann Sebastian Bach: “Concert For Oboe, Bv 1053; I. Allegro”
- k06: Carl Philipp Emanuel Bach: “Oboe Concerto In E Flat Major, Wq 165, Adagio Ma Non Troppo”
- k14: Wolfgang Amadeus Mozart: “Symphony No.36 In C, K.435, Linz, Finale (Presto)”
- k23: Wolfgang Amadeus Mozart: “Symphony No.40 In G minor, K.550, Andante”
- k38: Henri Dutilleux: “Ainsi La Nuit For String Quartet I. Nocturne”
- k46: Karl Szymanowski: “Violin Concerto No. 1, Op. 35”
- k54: Guiseppe Verdi: “Rigoletto, Act 2, Tutte Le Feste Al Tiempo”
- k64: Franz Schubert: “Piano Sonata No.21 In B flat, D.960, Andante Sostenuto”
- k81: Pyotr Ilyich Tchaikovsky: “Symphony 6, Mvt. 2”
- k89: Antonio Vivaldi: “The Four Seasons, Op. 8 Spring Allegro I”

Bei der Betrachtung der Spektrogramme fällt auf, wie stark die Intensität der Frequenzanteile variiert, sowohl innerhalb eines Stückes (z.B. k14) als auch zwischen den einzelnen Stücken, z.B. k64 (Piano solo) im Vergleich zu k46 (Streichquartett mit experimenteller Spielweise). Das zeigt, welche Dynamik innerhalb des klassischen Musikgenres abgerufen werden kann. Spielen alle Instrumente eines Orchesters gleichzeitig, entsteht eine starke Intensität von verschiedenen Frequenzen, die zuweilen gewollt disharmonisch klingen (k46). Im Gegensatz dazu ist der Klang, welcher ausschließlich von einem Klavier erzeugt wird im Spektrogramm übersichtlicher (k64).
Der Musikausschnitt von Vivaldi (k89) ist der einzige Musikausschnitt, der über den längsten Zeitabschnitt hinweg die geringste Intensität zwischen 0 und 1 kHz aufweist, da die Streicher vorwiegend Töne über 1 kHz anregen.
Starke Intensitäten treten zudem beim Gesang auf (k54), wo das Vibrato der Stimme ebenfalls gut im Spektrogramm abgebildet ist. Ferner sind mehrere harmonische Vielfache der Instrumente zu erkennen und keine rhythmische Percussion auszumachen, wie in den beschriebenen Genres zuvor.
Es werden alle klassischen Stücke als Musik klassifiziert, womöglich durch die klar fixierte Tonhöhe sowie der dazu passenden, spektralen Vielfache und den zeitlich eindeutigen melodischen Verlauf.

3.1.2.7. Metal

Folgend sieht man die Auflistung der Musikausschnitte für Metal, überwiegend aus den Achtziger und Neunzigerjahren. Das neueste Stück (m18) ist als Einziges aus dem Jahr 2001.

- m03: Anthrax: “Metal Thrashing Mad“
- m07: Coroner: “Die By My Hand”
- m10: Dark Tranquillity: “My Faeryland Forgotten”
- m12: Dark Tranquillity: “Away, Delight, Away”
- m18: Slayer: “Darkness Of Christ”
- m26: Anthrax: “Across The River”
- m30: Megadeth: “Trust”
- m38: Metallica: “…And Justice For All”
- m67: ACDC: “Big Gun”
- m83: Metallica: “For Whom The Bell Tolls”

Bei Betrachtung der Spektrogramme zu den Musikausschnitten, sticht die Farbe Rot ins Auge. Alle Musikausschnitte weisen über den gesamten Zeitraum eine hohe Intensität von den tiefen bis zu den hohen Frequenzen auf und es sind wenig bis keine harmonischen Vielfachen auszumachen. Vergleicht man dies mit einem Spektrogramm für Rauschen, ist die Ähnlichkeit deutlich zu sehen.
In der folgenden Abbildung sind der Musikausschnitt m07 (oben) von Coroner und das Rauschen TP8000 aufgeführt (unten). Das Rauschen ist in seiner Struktur gleichmäßiger, doch genauso wie beim Musikausschnitt nimmt die Intensität zu höhen Frequenz ab. Im Musikausschnitt ist der Übergang unebener und eine Intensitätssteigerung um 3 kHz zu sehen, welche danach erst in die bläuliche Farbe abschwächt. Dies könnte auch erklären, warum das Genre Metal ausschließlich als Rauschen klassifiziert wurde.

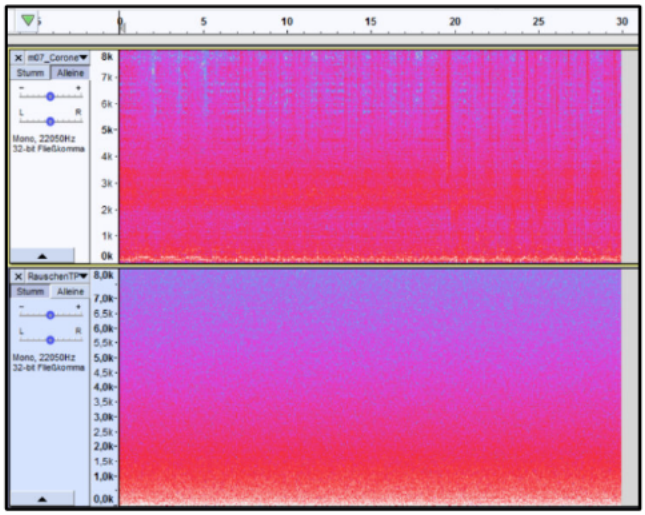

Abbildung 8: Vergleich der Spektren von Metal (m07) oben mit weißem Rauschen unten

Der Beat ist meist sehr schnell und im häufigen Wechsel. Das Instrument Gitarre ist verzerrt, genauso wie die Stimme (z.B. m18). Melodiöse sowie rhythmische, gleichmäßige Abschnitte sind nur wenige Sekunden lang, so dass die Musikausschnitte für die Musikerkennung des Cochlear Systems möglicherweise überwiegend strukturlos erscheinen und somit nicht greifbar sind.

3.1.2.8. Pop

Die Auswahl für das Genre Pop bewegt sich in der zeitlichen Einordnung überwiegend um die späten Neunzigerjahre und der Jahrtausendwende herum. Das älteste Stück (p65) ist von 1984, das neuste (p42) ist von 2007.

- p30: Britney Spears: “Lucky“
- p41: Peabo Bryson; Regina Belle: “A Whole New World”
- p42: Celine Dion: “That's The Way It Is”
- p45: Christina Aguilera: “Genie In A Bottle”
- p47: Christina Aguilera: “Lady Marmalade”
- p52: Destiny's Child: “Bootylicious”
- p53: Destiny's Child: ”Outro Amazing Grace”
- p65: Prince: “The Beautiful Ones”
- p78: Jennifer Lopez: “Play”
- p89: Mandy Moore: “Candy”

Bei diesen Titeln ist die Stimme stets im Vordergrund, was im Spektrogramm an den Oberwellen deutlich wird. Im Titel von Destiny`s Child "Outro Amazing Grace” (p53) sind mehrere Frauenstimmen sogar a capella zu hören.
Im Gegensatz zum Blues findet weniger instrumentaler Einsatz statt, stattdessen werden häufiger synthetische Klavierklänge verwendet. Die Pop-Titel sind rhythmisch gradliniger aufgebaut als die Musikausschnitte des Genres Disco. Durch den vorwiegend einfachen Aufbau sind die Titel eingängig und der Wert der Wiedererkennung hoch. Die Genres Klassik und Pop sind die beiden einzigen, welche von der Automatik zu 100% erkannt werden. Zu bemerken ist, dass bei beiden Genres stark ausgeprägte harmonische Oberwellen zu sehen sind: In der Klassik von den Instrumenten, beim Pop überwiegend vom Gesang.

3.1.2.9. Reggae

Die Musikausschnitte für das Genre Reggae sind aus den Siebziger- und Neunzigerjahren:

- re03: Bob Marley: “Three Little Birds”
- re23: Bob Marley: “Sun Is Shining”
- re28: Burning Spear: “Dub Creation”
- re36: Burning Spear: “Subjective Dub”
- re37: Burning Spear: “Friendly Dub”

- re41: Desmond Dekker: “Israelites”
- re44: Burning Spear: “Burning Reggae”
- re61: Bounty Killer: “Living Dangerously”
- re67: Dennis Brown: “Some Like It Hot”
- re87: Lauryn Hill; Bob Marley: “Turn Your Lights Down Low“

Rhythmisch zeichnet sich der aus jamaikanischen Wurzeln kommende Reggae durch seinen Off-Beat aus. Dabei wird die zweite und vierte Taktzeit von Keyboard oder Gitarre betont gespielt, die üblicherweise unbetont bleibt. In der Regel wird sonst mit dem Bass die betonte erste und dritte Taktzeit bespielt. Die Musik wird selbst dann als langsam wahrgenommen, wenn kleine, schnellere Percussionsounds beigemischt sind und strahlt eine gewisse Lässigkeit aus. Ausnahme dabei bildet re36, welches durch dichtere Instrumentierung und Klangverzerrungen im mittleren Frequenzbereich ein untypisches Klangbild ergibt.
Üblicherweise ist Reggae mit Gesang verbunden, eine Weiterentwicklung dazu bietet der Dub, welcher hier zweimal vorhanden ist (re28, re37). Der Dub wird ohne Gesang gespielt und im Tonstudio durch Klangmodulation erweitert. Diese sind im Spektrum gut zu sehen an den weniger intensiven Vielfachen der Oberwellen (überwiegend von der Gitarre). Im Gegensatz dazu sind bei den Musikausschnitten mit Gesang stark ausgeprägte Oberwellen vorzufinden (z.B. re61, re67, re87).
Bis auf re36 wurden alle Musikstücke von der SCAN-Automatik mindestens einmal der Kategorie Musik zugeordnet.

3.1.2.10. Rock

Für das Genre Rock erfolgte folgende Auswahl aus den Siebziger-, Achtziger- und Neunzigerjahren:

- r16: Queen: “Tie Your Mother Down”
- r23: Queen: “Bohemian Rhapsody”
- r33: The Rolling Stones: “Gimme Shelter”
- r38: Guns N' Roses: “Knockin' On Heaven's Door”
- r43: Led Zeppelin: “The Ocean”
- r47: Led Zeppelin: “The Wanton Song”
- r49: Simple Minds: “Real Life”

- r55: Simple Minds “African Skies”
- r87: Neil Young: “Rockin' In The Free World”
- r99: Stone Roses: “This Is The One”

Bei den Rockelementen hört man als wesentliches gemeinsames Merkmal die Gitarre, den Bass, das Schlagzeug und die Vocals heraus. Bei Queens “Bohemian Rhapsody“ und der Vertonung von “Knockin' On Heaven's Door” von Guns N' Roses ist das Klavier ebenfalls prägnant. Durch Instrumente und Gesang sind ausgeprägte Obertöne vorhanden, die im Spektrogramm sichtbar sind. Dort ist auch der Übergang gut zu sehen von melodiösen Sequenzen, denen ein Einsatz vom Schlagzeug folgt, verbildlicht durch die harte rote Kante, die sich senkrecht über das Frequenzspektrum zieht (r49, r99).
Anders als beim Metal, welches mit dem Rockgenre am nächsten verwandt ist, ist die Struktur der Spektrogramme bei der Rockauswahl weniger rauschähnlich, der Rhythmus gleichmäßiger und die Stimme melodiöser. Der Klang ist insgesamt wesentlich weicher als beim Metal, aber kraftvoller als beim Blues oder Pop.
Die Musikausschnitte r16 und r49 werden von der Musikautomatik dem Lärm statt der Musik zugeordnet, wobei r16 sogar dem Genre Metal zugeordnet werden könnte und womöglich deswegen nicht erkannt wurde.
Zusammenfassend scheint es so zu sein, dass melodisch erkennbare Struktur ein wichtiges Merkmal von Musik für das Musikerkennungssystem des Cochlea-Systems sind, während Rhythmus auch dem Lärm zugeordnet werden kann. Fehlt die melodische Struktur (also auch die erkennbare Tonhöhe, die eine gewisse Dauer hat und dann von einer anderen abgelöst wird), wird allenfalls noch Sprache detektiert, was möglicherweise an den rhythmischen Eigenschaften der dargebotenen sprachrelevanten Frequenzen liegt.

3.2. Messung

Die Messungen erfolgten in der Forschungsabteilung des Unfallkrankenhauses Berlin (UKB). Im Haus Nr. 49 steht ein schallarmer Anpassraum zur Einstellung von Hörprothesen zur Verfügung. Die Messkabine misst in Länge x Breite x Höhe: 190 cm x 181 cm x 214 cm.

3.2.1. Versuchsaufbau

In der folgenden Abbildung ist der schematische Versuchsaufbau erkennbar.

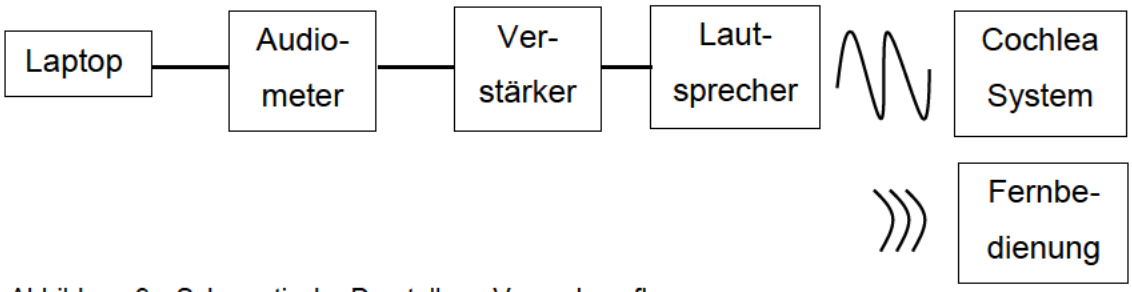

Abbildung 9: Schematische Darstellung Versuchsaufbau

Die Musikstücke für die Messung wurden über den Laptop der Klinik abgespielt. Es handelt sich um das Model Vaio PCG-9131M der Firma Sony. Das Betriebssystem ist Windows Vista. Als Mediaplayer wurde Audacity verwendet. Per Klinkenstecker wurde der Laptop mit dem Klinik-Audiometer KA450 der Firma Zeisberg Medizintechnik verbunden. Dort wurde das Musik- bzw. Sprachsignal in seiner Lautstärke reguliert sowie das Rauschsignal generiert. Über den Lphonics Verstärker (Gerät: PA500, Typ: 50, Jahr: 2011, Spannung: 230V/50 Hz, Leistung: 200 VA) wurde das Signal an den Canton Lautsprecher (CD220, 4...8 Ω) überführt. Der Lautsprecher beschallte das Cochlea-Implantat System, welches am Kunstkopf der Firma Sennheiser (Modell MZK 2002) angebracht war, um die Hörsituation so realistisch wie möglich zu gestalten.

Abbildung 10: Messaufbau – Beschallung von vorne

Das CI System kann modular zusammengebaut werden, so dass bei einem Defekt das entsprechende Ersatzteil schnell auszutauschen ist. Das Herzstück ist der Nucleus Prozessor CP910 (Baujahr hier: 2016). Daran schließt sich ein 28 cm langes Kabel an, welches mit der Spule verbunden ist. Mittels Magneten der Stärke 3 sind Spule und Implantat (CI422) verbunden. Das Implantat wurde zu Simulationszwecken in die Cochlea eines Meerschweinchens eingepflanzt. Zur Stromversorgung wurde an den Prozessor ein Batteriefach für zwei Batterien p675 der Firma Powerone angeschlossen. Das CI System wurde via Bluetooth mit der Fernbedienung CR230 verbunden. Diese zeigte die jeweilige akustische Situation an. Es gibt sechs verschiedene akustische Situationen, die erkannt werden können. Dazu gehören Ruhe, Sprache, Sprache im Störgeräusch, Lärm, Wind und Musik.

3.2.2. Messablauf

Es wurden zehn verschiedene Musikgenres mit jeweils zehn Musikstücken je 60 Sekunden in verschiedenen Testsituationen gemessen. Dazu wurde die Musik als Nutzsignal eingespielt und Sprache bzw. Rauschen fungierten als Störgeräusch. Zunächst erfolgte die Messung in ruhiger Umgebung mit 40 dB, 60 dB und 80 dB Nutzsignal von vorne (0°, Kunstkopf schaut in die Richtung der Signalquelle). Des Weiteren wurde mit jeweils 60 dB und 80 dB Nutzsignal von vorne und einem zusätzlich eingespielten Störsignal von 50 dB getestet. Das Störsignal wurde zusätzlich zum Nutzsignal jeweils von vorne (0°) bzw. hinten (180°) abgespielt (siehe
Abbildung 11). Beim ersten Störsignal handelt es sich um weißes Rauschen, beim zweiten Störsignal um ein Sprachsignal des MUSAN Datensatzes (siehe Abschnitt 5.1).

Abbildung 11: Messaufbau – Beschallung von vorne und hinten

Sowohl Stör- als auch Nutzsignal wurden vor der eigentlichen Messung auf ihren Lautstärkepegel hin untersucht. Das Messmikrofon Behringer ECM 8000 wurde an die Stelle platziert, an der sich das Cochlea-Implantat danach befand. Mittels des Mic Calibrator Type 4230 (1000 Hz / 94 dB) von Brüel und Kjaer wurde das Mikrofon kalibriert und der Wert in der Software „Spaichinger Schallpegelmesser“ von Dr. Ziegler hinterlegt. Das zu messende Signal wurde abgespielt und über das Interface UR 22 von Steinberg an die Software geleitet. Dort wurde der aktuell gemessene Schalldruckpegel mit dem hinterlegten Kalibrierwert verglichen und schließlich als dB(A)-Wert angezeigt.

Für die eigentliche Messung wurden die Musikausschnitte im jeweiligen Setting abgespielt und die erkannte akustische Situation sowie die Erkennungszeit notiert.

4. Untersuchungsergebnisse SCAN-Automatik

Die statistischen Tests wurden mithilfe des Programms IBM SPSS Statistics 22.0 durchgeführt und jeweils gegen ein Signifikanzniveau von $\alpha = 0{,}05$ getestet. Die SPSS Dateien sind auf dem dieser Arbeit beiliegenden Speichermedium zu finden.

4.1. Einfluss des Nutzpegels auf die Musikerkennung in ruhiger Umgebung

Die Automatikfunktion zur Musikerkennung des Cochlea-Implantat Pozessors wurde zunächst in ruhiger Umgebung für jeweils 100 Musikstücke aus zehn verschiedenen Genres bei einem Lautstärkepegel von 40 dB, 60 dB und 80 dB getestet.

Die Automatik des CI Prozessors setzt einen Mindestmusiklautstärkepegel voraus, um in einen anderen Modus als den Ruhemodus zu wechseln. Bei einem Musiklautstärkepegel von 40 dB wird keines der hundert getesteten Musikstücke als Musik erkannt, die Automatik verbleibt jeweils im Ruhemodus. Die Musikerkennungshäufigkeit liegt bei einem Musiklautstärkepegel von 60 dB dagegen bei 50 von hundert getesteten Musikstücken und steigt auf 73 von hundert Musikstücken bei einem Musiklautstärkepegel von 80 dB.

Die Messergebnisse zeigen überdies Unterschiede in der Erkennungshäufigkeit und Erkennungszeit einzelner Musikgenres auf (siehe Abbildung 12). Während bei einem Musiklautstärkepegel von 60 dB Klassik- und Pop-Musik mit acht von zehn getesteten Musikstücken die höchste Musikerkennungshäufigkeit aufweisen, wird Metal-Musik dagegen gar nicht von der SCAN-Automatik als Musik detektiert. Rock und Disco werden in sechs von zehn Fällen, Reggae, Blues und Jazz fünf von zehnmal, Electro viermal und Hip Hop dreimal als Musik erkannt. Die durchschnittlichen Musikerkennungszeiten schwanken je nach Genre zwischen 17,0 s für Jazz bis zu 44,5 s für Electro-Musik.

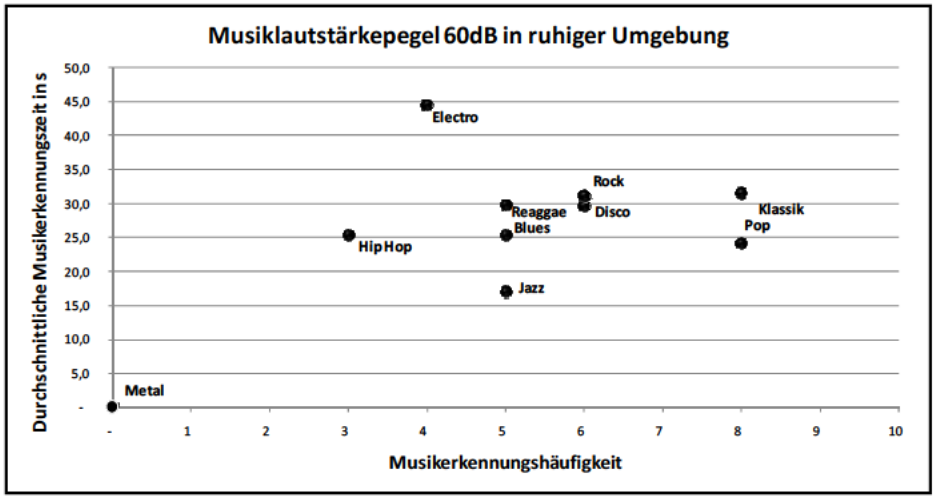

Abbildung 12: Musikerkennungshäufigkeit und –zeit bei 60dB Nutzpegel in ruhiger Umgebung

Generell nimmt mit zunehmendem Lautstärkepegel die Musikerkennungshäufigkeit unabhängig vom Genre zu, außer im Fall von Metal- sowie Disco-Musik (vgl. Abbildung 12 und Abbildung 13). Während bei einem Lautstärkepegel von 60 dB sechs von zehn getestete Disco-Musikstücke erkannt werden, sind es bei 80 dB nur fünf von zehn Musikstücken. Metal-Musik wird auch bei einem Musiklautstärkepegel von 80 dB nicht als Musik identifiziert, sondern ausschließlich von der Automatik als Lärm klassifiziert.

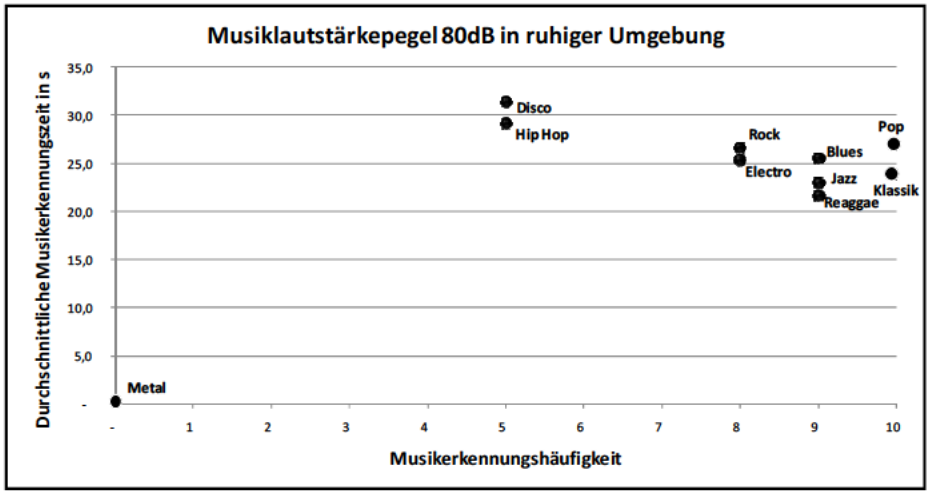

Abbildung 13: Musikerkennungshäufigkeit und –zeit bei 80dB Nutzpegel in ruhiger Umgebung

Die durchschnittlichen Musikerkennungszeiten bei einem Nutzpegel von 80 dB schwanken je nach Genre zwischen 21,7 s für Reggae bis zu 31,4 s für Disco-Musik. Der Einfluss des Lautstärkepegels auf die durchschnittliche Musikerkennungszeit ist je nach Genre unterschiedlich. Die

Musikerkennungszeit verringert sich mit zunehmender Lautstärke also nicht über alle Genres hinweg, sondern nur bei vier von neun relevanten Genres, wie der nachfolgenden Tabelle 1 entnommen werden kann.

Tabelle 1: Erkennungszeiten bei 60dB und 80dB Nutzpegel in ruhiger Umgebung

Lautstärkepegel	Durchschnittliche Erkennungszeiten je Musikgenre in s								
	Blues	Disco	Electro	Hip Hop	Jazz	Klassik	Pop	Reggae	Rock
60dB	25,4	29,7	44,5	25,3	17,0	31,5	24,1	29,8	31,2
80dB	25,6	31,4	25,4	29,2	23,0	23,8	27,0	21,7	26,6
Veränderung	**0,2**	**1,7**	**- 19,1**	**3,9**	**6,0**	**- 7,7**	**2,9**	**- 8,1**	**- 4,5**

Ein Chi-Quadrat-Test wurde durchgeführt, um zu prüfen, ob ein Zusammenhang zwischen Musikerkennung (Musik/keine Musik) und dem Nutzpegel besteht. Ein statistisch signifikanter Zusammenhang zwischen Musikerkennung und Lautstärkepegel kann nachgewiesen werden ($p = 0{,}001$). Dieser Effekt ist jedoch schwach, wie das Zusammenhangsmaß Phi zeigt ($\phi = 0{,}236$). Ebenfalls getestet wurde, ob ein statistisch signifikanter Zusammenhang zwischen Lärmerkennung (Lärm/kein Lärm) bzw. Ruheerkennung (Ruhe/keine Ruhe) und Nutzpegel vorliegt. Im Fall von Ruheerkennung kann dies bejaht werden ($p = 0{,}000$), nicht jedoch für Lärm ($p = 0{,}423$), wie die nachfolgende Tabelle 2 zeigt.

Tabelle 2: Einfluss Lautstärkepegel in ruhiger Umgebung auf die Erkennungshäufigkeit

Lautstärkepegel Musik	Erkennungshäufigkeiten der Automatik				
	Lärm	Musik	Ruhe	Sprache im Störgeräusch	Summe
40dB			100		100
60dB	29	50	19	2	100
	$p = 0{,}423$	$p = 0{,}001$	$p = 0{,}000$		
	$\phi = 0{,}057$	$\phi = 0{,}236$	$\phi = -0{,}300$		
80dB	24	73	1	2	100
Summe	**53**	**123**	**120**	**4**	**300**

Mittels einer zweifaktoriellen Varianzanalyse wurde ferner untersucht, welchen Einfluss die SCAN-Automatik (Lärm, Musik, Sprache im Störgeräusch) und der Nutzpegel (60 dB, 80 dB) auf die Erkennungszeit nehmen und ob eine

Interaktion zwischen beiden vorliegt.[9] Varianzhomogenität liegt jedoch nicht vor, wie der durchgeführte Levene-Test zeigt (p = 0,000). Die alternative Durchführung eines Welch-Tests ist nicht möglich, da dieser in SPSS für zweifaktorielle Varianzanalysen nicht implementiert ist. Die nachfolgenden Ergebnisse (siehe Tabelle 3) sind entsprechend eingeschränkt interpretierbar.
Das Gesamtmodell ist signifikant (p = 0,000). Es gibt einen Haupteffekt der SCAN-Automatik auf die Erkennungszeit (p = 0,000). Der CI Prozessor weist in Abhängigkeit vom erkannten Scanmodus unterschiedliche Erkennungszeiten auf. Für den Faktor Lautstärkepegel wird ebenfalls ein Haupteffekt ersichtlich (p = 0,028). Die Erkennungszeit ist demnach auch abhängig vom Lautstärkepegel. Der Interaktionsterm von SCAN-Automatik und Störgeräusch auf die Erkennungszeit ist ebenfalls signifikant (p = 0,014). Der Effekt der SCAN-Automatik ist also teilweise vom Lautstärkepegel abhängig.

Tabelle 3: Tests der Zwischensubjekteffekte

Abhängige Variable: Zeit bis Erkennung der Scanautomatik

Quelle	Typ III Quadratsumme	df	Quadratischer Mittelwert	F	Sig.	Partielles Eta hoch zwei
Korrigiertes Modell	5890,605[a]	6	981,767	8,408	0,000	0,115
Konstanter Term	29235,082	1	29235,082	250,376	0,000	0,393
Scanautomatik	4670,641	3	1556,88	13,333	0,000	0,094
Nutzgeräuschindb	565,905	1	565,905	4,847	0,028	0,012
Scanautomatik * Nutzgeräuschindb	1011,994	2	505,997	4,333	0,014	0,022
Fehler	45187,954	387	116,765			
Gesamtsumme	355622	394				
Korrigierter Gesamtwert	51078,558	393				

a R-Quadrat = 0,115 (Angepasstes R-Quadrat = 0,102)

Der signifikante Haupteffekt SCAN-Automatik wurde weiter mittels Bonferroni-korrigierter Mehrfachvergleiche getestet (siehe Tabelle 4). Es kann gezeigt werden, dass sich Lärm und Musik (p = 0,001) sowie Musik und Sprache im Störgeräusch (p = 0,001) signifikant in der Erkennungszeit unterscheiden, nicht hingegen die anderen Modi der SCAN-Automatik. Für den Nutzpegel konnte

[9] Für alle Tests bezüglich der Erkennungszeiten wurde der Ruhemodus nicht weiter betrachtet. Zurückzuführen ist dies auf den Testaufbau. Die Grundeinstellung des CI Prozessors in allen Messreihen war Ruhe. Erkannte das CI Prozessor trotz Abspielen von Musik und ggf. Störgeräuschen keine Musik, Lärm, Sprache oder Sprache im Störgeräusch, verblieb es im Ruhemodus. Eine Aussage zur Erkennungszeit kann entsprechend nicht getroffen werden. Das Messen der Erkennungszeit setzt demnach eine Veränderung des Modus von Ruhe in einen anderen Modus voraus.

kein Post-hoc-Test durchgeführt werden, da nur die beiden Faktorstufen 60 dB und 80 dB vorliegen. Da für den Nutzpegel ein signifikanter Haupteffekt nachgewiesen werden konnte, müssen sich also 60 dB und 80 dB signifikant in der Erkennungszeit unterscheiden.

Tabelle 4: Mehrfachvergleiche

Abhängige Variable: Zeit bis Erkennung der Scanautomatik

Bonferroni

(I) Scanautomatik	(J) Scanautomatik	Mittelwert-differenz (I-J)	Standard-fehler	Sig.	95 % Konfidenzintervall	
					Untergrenze	Obergrenze
Lärm	Musik	-4,33*	1,112	0,001	-7,28	-1,38
	Sprache	-5,53	7,681	1,000	-25,89	14,84
	Sprache im Störgeräusch	6,83	2,991	0,138	-1,10	14,76
Musik	Lärm	4,33*	1,112	0,001	1,38	7,28
	Sprache	-1,20	7,682	1,000	-21,57	19,17
	Sprache im Störgeräusch	11,16*	2,995	0,001	3,22	19,10
Sprache	Lärm	5,53	7,681	1,000	-14,84	25,89
	Musik	1,20	7,682	1,000	-19,17	21,57
	Sprache im Störgeräusch	12,36	8,168	0,787	-9,30	34,02
Sprache im Störgeräusch	Lärm	-6,83	2,991	0,138	-14,76	1,10
	Musik	-11,16*	2,995	0,001	-19,10	-3,22
	Sprache	-12,36	8,168	0,787	-34,02	9,30

Basierend auf beobachteten Mittelwerten.

Der Fehlerterm ist Mittel der Quadrate(Fehler) = 116,765.

* die Mittelwertdifferenz ist auf der Stufe 0,05 signifikant.

Es wurde weiterhin getestet, ob sich die Mittelwerte der Musikerkennungszeiten bei 60 dB und 80 dB unterscheiden. Der Mittelwert der Musikerkennungszeit liegt bei einem Lautstärkepegel von 60 dB bei 29,2s sowie 25,5s bei einem Lautstärkepegel von 80 dB. Die Standardabweichung beträgt bei einem Lautstärkepegel von 60 dB 10,9 s sowie 9,9 s bei einem Lautstärkepegel von 80 dB. Normalverteilung liegt gemäß Kolmogorow-Smirnow-Test und Shapiro-Wilk-Test vor. Mithilfe des anschließend durchgeführten t-Tests kann gezeigt werden, dass der Lautstärkepegel keinen signifikanten Unterschied auf die Musikerkennungszeit hat (p = 0,054).

Tabelle 5: Ergebnisse des t-Tests für die Musikerkennungszeit bei 60dB und 80dB in ruhiger Umgebung

		Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
		F	Sig.	t	df	Sig. (2-seitig)	Mittelwert-differenz	Standard-fehler-differenz	95% Konfidenzintervall der Differenz	
									Unterer	Oberer
Zeit bis Er-kennung der Scanauto-matik	Varianzgleichheit angenommen	0,886	0,349	1,949	121	0,054	3,694	1,896	-0,059	7,447
	Varianzgleichheit nicht angenommen			1,916	99,045	0,058	3,694	1,928	-0,131	7,52

4.2. Einfluss des Nutzpegels auf die Musikerkennung unter Störgeräuschen

Ferner wurde der Einfluss des Lautstärkepegels unter verschiedenen Störsignalen mit 50 dB Störpegel untersucht. Grundsätzlich verschlechtert sich die Musikerkennungshäufigkeit unter Einfluss von Störgeräuschen. Während in störgeräuschfreier Umgebung bei einem Musiklautstärkepegel von 60 dB insgesamt 50 von hundert gemessenen Musikstücken als Musik von der Automatik erkannt werden, sind es unter Rauschen von vorne noch 31 von hundert Musikstücken, bei Sprache von hinten nur noch sechs. Bei einem Musiklautstärkepegel von 80 dB unter Einfluss von Störgeräuschen verringert sich die Musikerkennungshäufigkeit zwar ebenfalls, allerdings nicht so stark wie bei einem Nutzpegel von 60 dB.

Die Ergebnisse der durchgeführten Chi-Quadrat-Tests zeigen, dass unter Störgeräuschen jeweils ein signifikanter Zusammenhang zwischen Musikerkennung und Lautstärkepegel besteht (jeweils $p = 0{,}000$). Dieser Zusammenhang scheint zudem stärker ausgeprägt zu sein, als im Testszenario in ruhiger Umgebung, wie die höheren ϕ zeigen (siehe Tabelle 6).

Tabelle 6: Einfluss des Nutzpegels unter verschiedenen Störgeräuschen auf die Musikerkennungshäufigkeit

Lautstärke-pegel Musik	Erkennungshäufigkeiten der Automatik				
	Musik in ruhiger Umgebung	Musik bei Rauschen vorne	Musik bei Rauschen hinten	Musik bei Sprache vorne	Musik bei Sprache hinten
60dB	50	31	20	6	7
	$p = 0{,}001$	$p = 0{,}000$	$p = 0{,}000$	$p = 0{,}000$	$p = 0{,}000$
	$\phi = 0{,}236$	$\phi = 0{,}360$	$\phi = 0{,}483$	$\phi = 0{,}651$	$\phi = 0{,}578$
80dB	73	67	68	69	62
Summe	**123**	**98**	**88**	**75**	**69**

Des Weiteren wurde für die Testszenarien unter Störgeräusch untersucht, welchen Einfluss die SCAN-Automatik und der Lautstärkepegel auf die Erkennungszeit nehmen.

Die durchgeführten Levene-Tests sind – bis auf "Musik bei Sprache von hinten" – allesamt signifikant, so dass wiederum meist nicht von der Gleichheit der Fehlervarianzen ausgegangen werden kann. Entsprechend eingeschränkt interpretierbar sind die Ergebnisse der durchgeführten zweifaktoriellen Varianzanalysen (siehe Tabelle 7).

Das Gesamtmodell ist für alle Testszenarien signifikant. Es gibt jeweils einen Haupteffekt der SCAN-Automatik. Für den Faktor Nutzpegel kann nur ein Haupteffekt für "Musik bei Sprache von hinten" gezeigt werden (p = 0,004), ebenfalls weist nur in diesem Fall der Interaktionsterm eine Signifikanz auf (p = 0,000).

Tabelle 7: p-Werte für den Levene-Test und der Zwischensubjekteffekte

	p-Werte				
	Musik in ruhiger Umgebung	**Musik bei Rauschen vorne**	**Musik bei Rauschen hinten**	**Musik bei Sprache vorne**	**Musik bei Sprache hinten**
Levene-Test	0,000	0,002	0,045	0,149	0,016
Korrigiertes Modell	0,000	0,000	0,010	0,017	0,000
Scanautomatik	0,000	0,000	0,003	0,004	0,000
Nutzgeräuschindb	0,028	0,206	0,097	0,308	0,004
Scanautomatik * Nutzgeräuschindb	0,014	0,069	0,201	0,197	0,000

Die Post-hoc-Tests für die SCAN-Automatik zeigen für "Musik bei Rauschen von vorne" sowie für "Musik bei Sprache von hinten", dass sich lediglich die Erkennungszeiten von Lärm und Musik signifikant unterscheiden (p = 0,003 bzw. p = 0,005). Für "Musik bei Rauschen von hinten" kann kein statistisch signifikanter Zusammenhang zwischen den Modi der SCAN-Automatik gezeigt werden. Bei "Musik bei Sprache von vorne" kann kein Post-hoc-Test für die SCAN-Automatik durchgeführt werden, da die Faktorstufe "Sprache im Störgeräusch" weniger als zwei Fälle umfasst.

Die Mittelwerte der Musikerkennungszeiten sind unter Störgeräusch allesamt leicht höher als in ruhiger Umgebung. Die Tests bezüglich der Unterscheidung der Mittelwerte der Musikerkennungszeiten bei 60 dB und 80 dB Nutzpegel zeigen ein differenziertes Ergebnis. Für "Musik bei Sprache von vorne" kann gezeigt werden, dass es keinen signifikanten Einfluss des Lautstärkepegels auf die Musikerkennungszeit gibt. Für die Varianten "Musik bei Rauschen von vorne", "Musik bei Rauschen von hinten" sowie "Musik bei Sprache von hinten" besteht jedoch ein signifikanter Unterschied, wie die p-Werte in Tabelle 8 zeigen.

Tabelle 8: Einfluss des Lautstärkepegels auf die Musikerkennungszeit

	Musik in ruhiger Umgebung	Musik bei Rauschen vorne	Musik bei Rauschen hinten	Musik bei Sprache vorne	Musik bei Sprache hinten
Häufigkeit	**123**	**98**	**88**	**75**	**69**
60dB	50	31	20	6	7
80dB	73	67	68	69	62
Erkennungszeit (in s)	**26,97**	**30,73**	**29,82**	**28,09**	**27,93**
60dB	29,16	37,00	36,15	34,00	43,43
80dB	25,47	27,84	27,96	27,58	26,18
Standardabweichung (in s)	**10,444**	**12,731**	**12,493**	**9,842**	**11,036**
60dB	10,873	14,243	14,601	15,126	15,501
80dB	9,937	10,910	11,257	9,233	9,026
p-Wert (Musikerkennungszeit)	**0,054**	**0,002**	**0,029**	**0,300**	**0,005**

4.3. Einfluss von Rauschen von vorne auf die Musikerkennung

Die Musikerkennungshäufigkeit liegt bei einem Musiklautstärkepegel von 60 dB ohne Störgeräusch bei 50 von hundert Musikstücken und steigt auf 73 bei einem Lautstärkepegel von 80 dB (siehe Abschnitt 4.1). Unter Einfluss von Rauschen mit 50 dB von vorne verringert sich die Erkennungshäufigkeit bei einem Musiklautstärkepegel von 60 dB auf 31 erkannte Musikstücke bzw. auf eine Häufigkeit von 67 bei einem Musiklautstärkepegel von 80 dB.

Die Messergebnisse weisen auf Unterschiede in der Erkennungshäufigkeit und Erkennungszeit einzelner Musikgenres hin. Genreübergreifend werden bei einem Musiklautstärkepegel von 60 dB unter Rauschen von vorne weniger Musikstücke von der SCAN-Automatik als Musik klassifiziert als ohne Rauschen (vgl. Abbildung 12 auf Seite 35 mit Abbildung 14). Einzig Jazz kann die Erkennungshäufigkeit von fünf beibehalten, Metal wird weiterhin in keinem Fall erkannt.

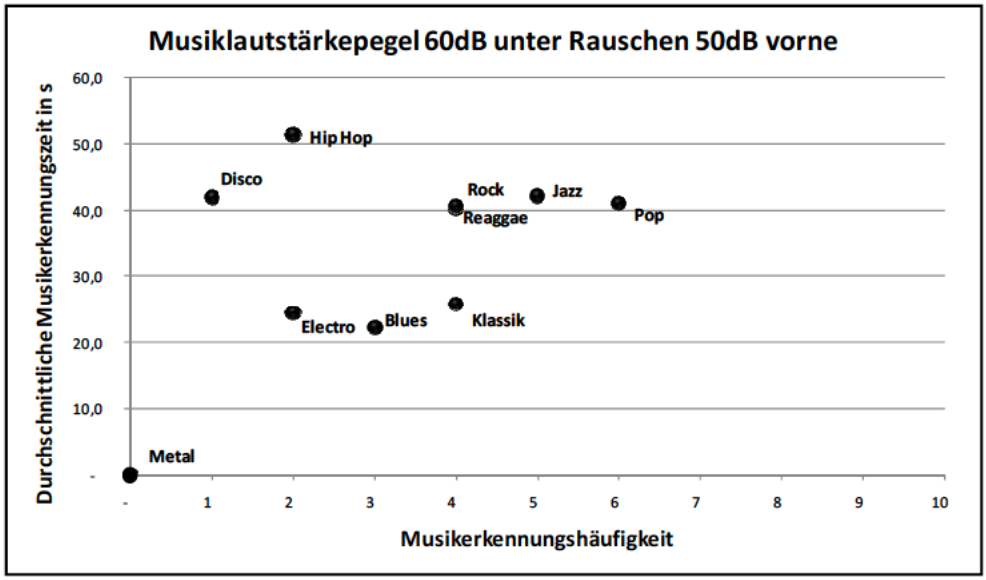

Abbildung 14: Musikerkennungshäufigkeit und –zeit bei 60dB Nutzpegel unter 50dB Rauschen von vorne

Die Erkennungszeiten bei einem Musiklautstärkepegel von 60 dB unter Rauschen von vorne verbessern sich sogar für Klassik, Electro und Blues (siehe Tabelle 9). Allerdings zeigen diese Genres auch deutlich geringere Häufigkeiten unter Rauschen von vorne auf als in ruhiger Umgebung. Für alle anderen Genres benötigt die SCAN-Automatik dagegen mehr Zeit im Durchschnitt für die Musikerkennung.

Tabelle 9: Durchschnittliche Erkennungszeit je Genre bei 60dB und 80dB Nutzgeräusch in ruhiger Umgebung und unter 50dB Rauschen

	Durchschnittliche Erkennungszeit in s								
Musiklautstärkepegel	**Blues**	**Disco**	**Electro**	**Hip**	**Jazz**	**Klassik**	**Pop**	**Reaggae**	**Rock**
60dB									
Ruhige Umgebung	25,4	29,7	44,5	25,3	17,0	31,5	24,1	29,8	31,2
Rauschen 50dB vorne	22,3	42,0	24,5	51,5	42,2	26,0	41,2	40,3	40,8
Veränderung	**- 3,1**	**12,3**	**- 20,0**	**26,2**	**25,2**	**- 5,5**	**17,0**	**10,5**	**9,6**
80dB									
Ruhige Umgebung	25,6	31,4	25,4	29,2	23,0	23,8	27,0	21,7	26,6
Rauschen 50dB vorne	32,8	34,0	24,4	31,4	22,9	20,1	29,7	23,2	36,0
Veränderung	**7,2**	**2,6**	**- 0,9**	**2,2**	**- 0,1**	**- 3,7**	**2,7**	**1,6**	**9,4**

Bei einem Musiklautstärkepegel von 80 dB unter Rauschen von vorne ist die Musikerkennungshäufigkeit im Vergleich zu störgeräuschfreier Umgebung genreübergreifend deutlich weniger eingeschränkt als bei 60 dB. Der Rückgang der Häufigkeit unter Rauschen von vorne ist je Genre zudem weniger stark als bei 60 dB. Pop, Jazz, Electro und Rock weisen jeweils eine um eins verringerte

Musikerkennungshäufigkeit im Vergleich zu störgeräuschfreier Umgebung auf, Klassik zeigt eine um zwei auf acht verringerte Häufigkeit. Reggae, Blues, Disco und Hip Hop weisen bei 80 dB Musiklautstärkepegel unter Rauschen von vorne keine Veränderung der Erkennungshäufigkeit gegenüber störgeräuschfreier Umgebung auf (vgl. Abbildung 13 auf Seite 35 mit Abbildung 15).

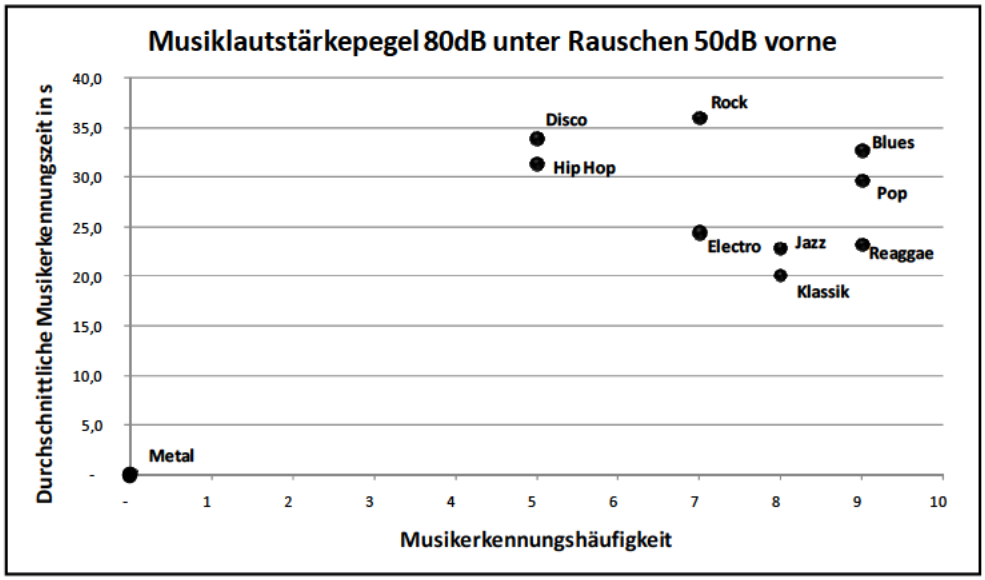

Abbildung 15: Musikerkennungshäufigkeit und –zeit bei 80dB Nutzpegel unter 50dB Rauschen von vorne

Die Erkennungszeiten verbessern sich – wie bereits unter 60 dB – auch bei einem Musiklautstärkepegel von 80 dB unter Rauschen von 50 dB von vorne für Klassik und Electro. Jazz bleibt annähernd stabil in der Erkennungszeit. Für alle anderen Genres benötigt die SCAN-Automatik im Durchschnitt mehr Zeit für die Musikerkennung als bei 80 dB Musiklautstärkepegel in störgeräuschfreier Umgebung (siehe Tabelle 9).

Die durchgeführten Chi-Quadrat-Tests zeigen (siehe Tabelle 10), dass bei 60 dB Musiklautstärkepegel ein statistisch signifikanter Zusammenhang zwischen Musikerkennung und Rauschen von vorne nachgewiesen werden kann ($p = 0{,}006$). Dieser Effekt ist jedoch eher schwach ausgeprägt ($\phi = -0{,}194$). Bei einem Musiklautstärkepegel von 80 dB kann dagegen kein statistisch signifikanter Zusammenhang zur Musikerkennung mit und ohne Rauschen von vorne festgestellt werden ($p = 0{,}355$).

Tabelle 10: Einfluss von Rauschen von vorne auf die Musikerkennungshäufigkeit

Einfluss Rauschen	Musikerkennungshäufigkeiten der Automatik	
	Musiklautstärkepegel 60dB	Musiklautstärkepegel 80dB
Ruhige Umgebung	50	73
	p = 0,006	p = 0,355
	ϕ = -0,194	ϕ = -0,065
Rauschen 50dB von vorne	31	67
Summe	**81**	**140**

Es wurde eine zweifaktorielle Varianzanalyse bezüglich des Einflusses der SCAN-Automatik und dem Einfluss von Rauschen (Rauschen 50 dB von vorne/ruhige Umgebung) auf die Erkennungszeit durchgeführt. Der Levene-Test ist bei 60 dB Musiklautstärkepegel jedoch signifikant, sodass keine Varianzhomogenität vorliegt und die Ergebnisse nur eingeschränkt interpretierbar sind.

Das Gesamtmodell ist jeweils bei 60 dB und 80 dB Musiklautstärkepegel signifikant (siehe Tabelle 11). Es zeigt sich jeweils ein Haupteffekt für die SCAN-Automatik. Für den Einfluss des Rauschens kann kein Haupteffekt nachgewiesen werden. Auch zeigen die Interaktionsterme jeweils keine Signifikanz.

Tabelle 11: Tests der Zwischensubjekteffekte

Abhängige Variable: Zeit bis Erkennung der Scanautomatik

60dB	Typ III Quadrat-summe	df	Quadratischer Mittelwert	F	Sig.	Partielles Eta hoch zwei
Korrigiertes Modell	5854,069[a]	5	1170,814	9,647	0,000	0,216
Konstanter Term	26349,047	1	26349,047	217,101	0,000	0,554
Scanautomatik	4854,085	2	2427,042	19,997	0,000	0,186
RichtungdesStörgeräuschs	124,038	1	124,038	1,022	0,313	0,006
Scanautomatik * RichtungdesStörgeräuschs	207,438	2	103,719	0,855	0,427	0,01
Fehler	21239,301	175	121,367			
Gesamtsumme	162521	181				
Korrigierter Gesamtwert	27093,37	180				

a R-Quadrat = ,216 (Angepasstes R-Quadrat = ,194)

Abhängige Variable: Zeit bis Erkennung der Scanautomatik

80dB	Typ III Quadrat-summe	df	Quadratischer Mittelwert	F	Sig.	Partielles Eta hoch zwei
Korrigiertes Modell	1404,936a	5	280,987	3,061	0,011	0,073
Konstanter Term	15467,699	1	15467,699	168,494	0,000	0,466
Scanautomatik	1039,268	2	519,634	5,661	0,004	0,055
RichtungdesStörgeräuschs	16,324	1	16,324	0,178	0,674	0,001
Scanautomatik * RichtungdesStörgeräuschs	67,971	2	33,986	0,37	0,691	0,004
Fehler	17717,365	193	91,8			
Gesamtsumme	146566	199				
Korrigierter Gesamtwert	19122,302	198				

a R-Quadrat = ,073 (Angepasstes R-Quadrat = ,049)

Die Post-hoc-Tests für den Haupteffekt SCAN-Automatik zeigen, dass sich jeweils Musik und Lärm signifikant in der Erkennungszeit unterscheiden. Bei 60 dB gibt es zudem noch einen signifikanten Effekt zwischen Musik und Sprache im Störgeräusch (siehe Tabelle 12).

Tabelle 12: Mehrfachvergleiche

Abhängige Variable: Zeit bis Erkennung der Scanautomatik
Bonferroni

60dB

(I) Scanautomatik	(J) Scanautomatik	Mittelwert-differenz (I-J)	Standard-fehler	Sig.	95 % Konfidenzintervall	
					Untergrenze	Obergrenze
Lärm	Musik	-8,39*	1,67	0,000	-12,43	-4,36
	Sprache im Störgeräusch	5,10	4,639	0,820	-6,11	16,31
Musik	Lärm	8,39*	1,67	0,000	4,36	12,43
	Sprache im Störgeräusch	13,49*	4,661	0,013	2,23	24,76
Sprache im Störgeräusch	Lärm	-5,10	4,639	0,820	-16,31	6,11
	Musik	-13,49*	4,661	0,013	-24,76	-2,23

Basierend auf beobachteten Mittelwerten.
Der Fehlerterm ist Mittel der Quadrate(Fehler) = 121,367.
* die Mittelwertdifferenz ist auf der Stufe ,05 signifikant.

80dB

(I) Scanautomatik (80dB)	(J) Scanautomatik	Mittelwert-differenz (I-J)	Standard-fehler	Sig.	95 % Konfidenzintervall	
					Untergrenze	Obergrenze
Lärm	Musik	-3,93*	1,525	0,032	-7,61	-0,25
	Sprache im Störgeräusch	6,42	4,962	0,591	-5,56	18,41
Musik	Lärm	3,93*	1,525	0,032	0,25	7,61
	Sprache im Störgeräusch	10,35	4,859	0,103	-1,38	22,08
Sprache im Störgeräusch	Lärm	-6,42	4,962	0,591	-18,41	5,56
	Musik	-10,35	4,859	0,103	-22,08	1,38

Basierend auf beobachteten Mittelwerten.
Der Fehlerterm ist Mittel der Quadrate(Fehler) = 91,800.
* die Mittelwertdifferenz ist auf der Stufe ,05 signifikant.

Es wurde weiterhin getestet, ob sich die Mittelwerte der Musikerkennungszeiten bei 60 dB und 80 dB Nutzpegel unter Rauschen von vorne im Vergleich zu ruhiger Umgebung unterscheiden (siehe Tabelle 13). Normalverteilung liegt nur bei 80 dB Musiklautstärkepegel vor. Der durchgeführte U-Test zeigt bei einem Musiklautstärkepegel von 60 dB, dass das Rauschen einen signifikanten Einfluss auf die Musikerkennungszeit nimmt (p = 0,013). Jedoch kann dieser Zusammenhang vom t-Test nicht für einen Musiklautstärkepegel von 80 dB festgestellt werden (p = 0,181).

Tabelle 13: Tests der Musikerkennungszeiten für den Einfluss von Rauschen

Einfluss Rauschen	Musikerkennungszeiten der Automatik in s	
	Musiklautstärkepegel 60dB	Musiklautstärkepegel 80dB
Ruhige Umgebung	29,16	25,47
	p = 0,013	p = 0,181
Rauschen 50dB von vorne	37,00	27,84

4.4. Einfluss von Sprache von vorne auf die Musikerkennung

Unter Einfluss von Sprache mit 50 dB von vorne verringert sich die Erkennungshäufigkeit bei einem Musiklautstärkepegel von 60 dB auf sechs erkannte Musikstücke (gegenüber 50 in ruhiger Umgebung) bzw. auf eine Häufigkeit von 69 (gegenüber 73 in ruhiger Umgebung) bei einem Musiklautstärkepegel von 80 dB.

Bei 60 dB Musiklautstärkepegel werden unter Sprache mit 50 dB von vorne überhaupt nur noch drei Genres, nämlich Pop, Rock und Klassik von der SCAN-Automatik erkannt (siehe Abbildung 16). Das Störgeräusch Sprache scheint bei 60 dB Musiklautstärkepegel demnach einen deutlich stärkeren Einfluss auf die allgemeine und genrespezifische Musikerkennung zu nehmen als Rauschen (vgl. Abbildung 14 auf Seite 42 mit Abbildung 16).

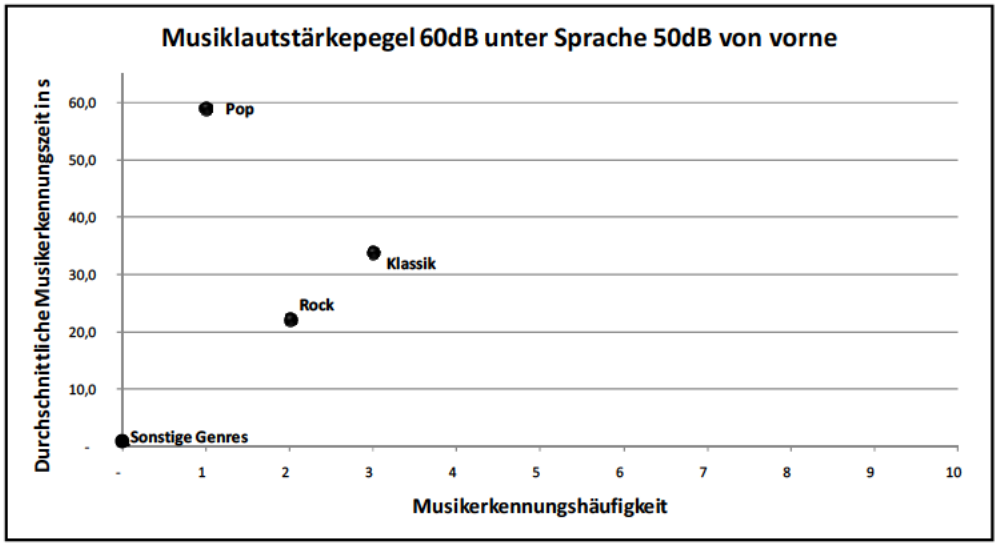

Abbildung 16: Musikerkennungshäufigkeit und –zeit bei 60dB Nutzpegel unter 50dB Sprache von vorne

Bei einem Musiklautstärkepegel von 80 dB unter Sprache von vorne ist ein solcher Effekt hingegen nicht feststellbar (vgl. Abbildung 15 auf Seite 43 mit Abbildung 17). Bei einem Musiklautstärkepegel von 80 dB unter Sprache mit 50 dB von vorne zeigt sich zudem, dass nur die zwei Genres Electro und Klassik eine geringere Häufigkeit aufweisen als in ruhiger Umgebung (vgl. Abbildung 13 auf Seite 35 mit Abbildung 17).

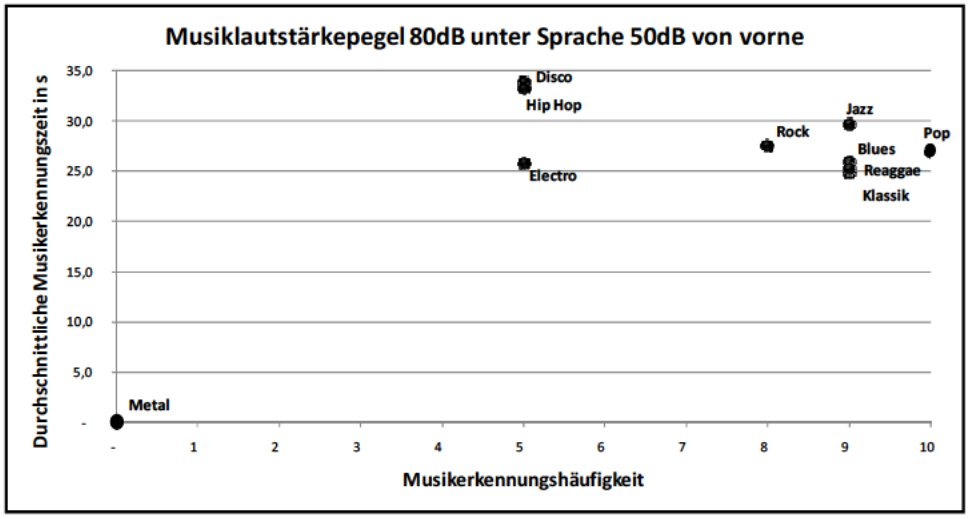

Abbildung 17: Musikerkennungshäufigkeit und –zeit bei 80dB Nutzpegel unter 50dB Sprache von vorne

Die durchschnittlichen Musikerkennungszeiten bei einem Musiklautstärkepegel von 80 dB sind in ruhiger Umgebung kleiner oder annähernd gleich groß wie unter Sprache von vorne (siehe Tabelle 14).

Tabelle 14: Durchschnittliche Erkennungszeit je Genre in ruhiger Umgebung und unter 50dB Sprache bei 60dB und 80dB Nutzgeräusch

	Durchschnittliche Erkennungszeit in s								
Musiklautstärkepegel	**Blues**	**Disco**	**Electro**	**Hip Hop**	**Jazz**	**Klassik**	**Pop**	**Reaggae**	**Rock**
60dB									
Ruhige Umgebung	25,4	29,7	44,5	25,3	17,0	31,5	24,1	29,8	31,2
Sprache 50dB vorne	-	-	-	-	-	33,7	59,0	-	22,0
Veränderung						**2,2**	**34,9**		**- 9,2**
80dB									
Ruhige Umgebung	25,6	31,4	25,4	29,2	23,0	23,8	27,0	21,7	26,6
Sprache 50dB vorne	25,9	33,8	25,8	33,2	29,7	24,8	26,9	25,2	27,5
Veränderung	**0,3**	**2,4**	**0,4**	**4,0**	**6,7**	**1,0**	**- 0,1**	**3,6**	**0,9**

Die durchgeführten Chi-Quadrat-Tests zeigen (siehe Tabelle 15), dass bei 60 dB Musiklautstärkepegel ein statistisch signifikanter Zusammenhang zwischen Musikerkennung und Sprache von vorne nachgewiesen werden kann ($p = 0{,}000$). Bei einem Musiklautstärkepegel von 80 dB kann dagegen kein statistisch signifikanter Zusammenhang zwischen Musikerkennung und Sprache von vorne festgestellt werden ($p = 0{,}533$).

Tabelle 15: Einfluss von Sprache von vorne auf die Musikerkennungshäufigkeit

Einfluss Sprache	Musikerkennungshäufigkeiten der Automatik	
	Musiklautstärkepegel 60dB	Musiklautstärkepegel 80dB
Ruhige Umgebung	50	73
	p = 0,000	p = 0,533
	ϕ = -0,490	ϕ = -0,044
Sprache 50dB von vorne	6	67
Summe	**56**	**140**

Die zweifaktorielle Varianzanalyse bezüglich des Einflusses der SCAN-Automatik und Sprache (Sprache von vorne/ruhige Umgebung) auf die Erkennungszeit zeigt, dass das Gesamtmodell jeweils bei 60 dB und 80 dB Musiklautstärkepegel signifikant ist (siehe Tabelle 16). Die Ergebnisse weisen jeweils auf einen Haupteffekt für die SCAN-Automatik hin. Für den Einfluss der Sprache kann kein Haupteffekt nachgewiesen werden. Auch zeigen die Interaktionsterme jeweils keine Signifikanz.

Tabelle 16: Tests der Zwischensubjekteffekte

Abhängige Variable: Zeit bis Erkennung der Scanautomatik

60dB	Typ III Quadratsumme	df	Quadratischer Mittelwert	F	Sig.	Partielles Eta hoch zwei
Korrigiertes Modell	2796,300a	4	699,075	5,819	0,000	0,184
Konstanter Term	13980,729	1	13980,729	116,368	0,000	0,53
Scanautomatik	1975,006	2	987,503	8,219	0,000	0,138
RichtungdesStörgeräuschs	225,829	1	225,829	1,88	0,173	0,018
Scanautomatik * RichtungdesStörgeräuschs	13,28	1	13,28	0,111	0,74	0,001
Fehler	12374,69	103	120,143			
Gesamtsumme	82621	108				
Korrigierter Gesamtwert	15170,991	107				

a R-Quadrat = ,184 (Angepasstes R-Quadrat = ,153)

80dB	Typ III Quadratsumme	df	Quadratischer	F	Sig.	Partielles Eta
Korrigiertes Modell	1429,978a	5	285,996	3,157	0,009	0,078
Konstanter Term	10022,272	1	10022,272	110,617	0,000	0,372
Scanautomatik	1223,886	2	611,943	6,754	0,001	0,067
RichtungdesStörgeräuschs	0,339	1	0,339	0,004	0,951	0
Scanautomatik * RichtungdesStörgeräuschs	20,742	2	10,371	0,114	0,892	0,001
Fehler	16942,768	187	90,603			
Gesamtsumme	139348	193				
Korrigierter Gesamtwert	18372,746	192				

a R-Quadrat = ,078 (Angepasstes R-Quadrat = ,053)

Die Post-hoc-Tests für den Haupteffekt SCAN-Automatik zeigen, dass sich Musik und Lärm jeweils signifikant in der Erkennungszeit unterscheiden (siehe Tabelle 17).

Tabelle 17: Mehrfachvergleiche

Abhängige Variable: Zeit bis Erkennung der Scanautomatik
Bonferroni

60dB

(I) Scanautomatik	(J) Scanautomatik	Mittelwert-differenz (I-J)	Standard-fehler	Sig.	95 % Konfidenzintervall	
					Untergrenze	Obergrenze
Lärm	Musik	-9,82*	2,133	0,000	-15,01	-4,63
	Sprache im Störgeräusch	-2,14	7,904	1,000	-21,38	17,10
Musik	Lärm	9,82*	2,133	0,000	4,63	15,01
	Sprache im Störgeräusch	7,68	7,888	0,998	-11,52	26,88
Sprache im Störgeräusch	Lärm	2,14	7,904	1,000	-17,10	21,38
	Musik	-7,68	7,888	0,998	-26,88	11,52

Basierend auf beobachteten Mittelwerten.
Der Fehlerterm ist Mittel der Quadrate(Fehler) = 120,143.
* die Mittelwertdifferenz ist auf der Stufe ,05 signifikant.

80dB

(I) Scanautomatik (80dB)	(J) Scanautomatik	Mittelwert-differenz (I-J)	Standard-fehler	Sig.	95 % Konfidenzintervall	
					Untergrenze	Obergrenze
Lärm	Musik	-5,22*	1,589	0,004	-9,06	-1,38
	Sprache im Störgeräusch	4,94	5,665	1,000	-8,75	18,62
Musik	Lärm	5,22*	1,589	0,004	1,38	9,06
	Sprache im Störgeräusch	10,16	5,553	0,207	-3,26	23,57
Sprache im Störgeräusch	Lärm	-4,94	5,665	1,000	-18,62	8,75
	Musik	-10,16	5,553	0,207	-23,57	3,26

Basierend auf beobachteten Mittelwerten.
Der Fehlerterm ist Mittel der Quadrate(Fehler) = 90,603.
* die Mittelwertdifferenz ist auf der Stufe ,05 signifikant.

Ferner wurde getestet, ob es signifikante Unterschiede in der Musikerkennungszeit unter Einfluss von Sprache gibt (siehe Tabelle 18). Bei einem Musiklautstärkepegel von 60 dB sowie bei 80 dB kann für Sprache von vorne gegenüber ruhiger Umgebung kein signifikanter Zusammenhang in der Musikerkennungszeit nachgewiesen werden.

Tabelle 18: Einfluss von Sprache von vorne auf die Musikerkennungszeit

Einfluss Sprache	Musikerkennungszeiten der Automatik in s	
	Musiklautstärkepegel 60dB	Musiklautstärkepegel 80dB
Ruhige Umgebung	29,16	25,47
	p = 0,466	p = 0,192
Sprache 50dB von vorne	34,00	27,58

4.5. Einfluss der Richtung des Rauschens auf die Musikerkennung

Die Musikerkennungshäufigkeit liegt bei einem Musiklautstärkepegel von 60 dB mit Rauschen von vorne bei 31 von hundert Musikstücken und steigt auf 67 bei einem Lautstärkepegel von 80 dB (siehe Abschnitt 0). Bei Rauschen von hinten verringert sich die Erkennungshäufigkeit bei einem Musiklautstärkepegel von

60 dB auf zwanzig erkannte Musikstücke bzw. erhöht sich auf 68 bei einem Musiklautstärkepegel von 80 dB.

Bei einem Nutzpegel von 60 dB kann genrespezifisch im Vergleich zu Rauschen von vorne nur Klassik seine Häufigkeit steigern. Alle anderen Genres werden dagegen weniger häufig erkannt. Disco, Electro und Hip Hop zeigen sogar – genau wie Metal – eine Häufigkeit von null (vgl. Abbildung 14 auf Seite 42 mit Abbildung 18).

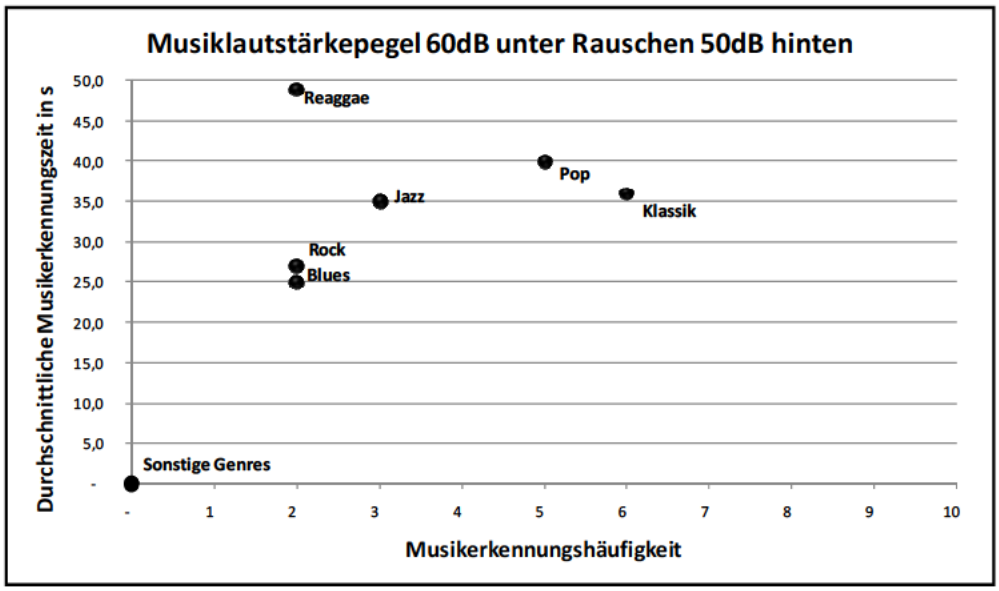

Abbildung 18: Musikerkennungshäufigkeit und –zeit bei 60dB Nutzpegel unter 50dB Rauschen von hinten

Die Musikerkennungszeiten bei einem Nutzpegel von 60 dB unter Rauschen von hinten erhöhen sich gegenüber Rauschen von vorne für Klassik, Reggae und Blues wohingegen Jazz, Pop und Rock niedrigere Erkennungszeiten aufweisen (siehe Tabelle 19).

Tabelle 19: Durchschnittliche Erkennungszeit je Genre bei Rauschen von vorne und hinten bei 60dB und 80dB Nutzpegel

	Durchschnittliche Erkennungszeit in s								
Musiklautstärkepegel	**Blues**	**Disco**	**Electro**	**Hip Hop**	**Jazz**	**Klassik**	**Pop**	**Reaggae**	**Rock**
60dB									
Rauschen 50dB vorne	22,3	42,0	24,5	51,5	42,2	26,0	41,2	40,3	40,8
Rauschen 50dB hinten	25,0	-	-	-	35,0	36,0	40,0	49,0	27,0
Veränderung	**2,7**	**-42,0**	**-24,5**	**- 51,5**	**- 7,2**	**10,0**	**- 1,2**	**8,8**	**-13,8**
80dB									
Rauschen 50dB vorne	32,8	34,0	24,4	31,4	22,9	20,1	29,7	23,2	36,0
Rauschen 50dB hinten	34,0	27,8	30,6	23,2	23,6	28,3	30,2	22,3	30,0
Veränderung	**1,2**	**- 6,3**	**6,2**	**- 8,2**	**0,8**	**8,2**	**0,5**	**- 0,9**	**- 6,0**

Bei einem Musiklautstärkepegel von 80 dB unter Rauschen von hinten ist die Musikerkennungshäufigkeit im Vergleich zu Rauschen von vorne genrespezifisch recht ähnlich. Hip Hop, Jazz, Reggae und Rock zeigen keine Veränderung, Blues und Disco haben eine um eins verringerte Häufigkeit, dagegen zeigen Electro, Klassik und Pop eine um eins erhöhte Häufigkeit gegenüber Rauschen von vorne (vgl. Abbildung 15 auf Seite 43 mit Abbildung 19).

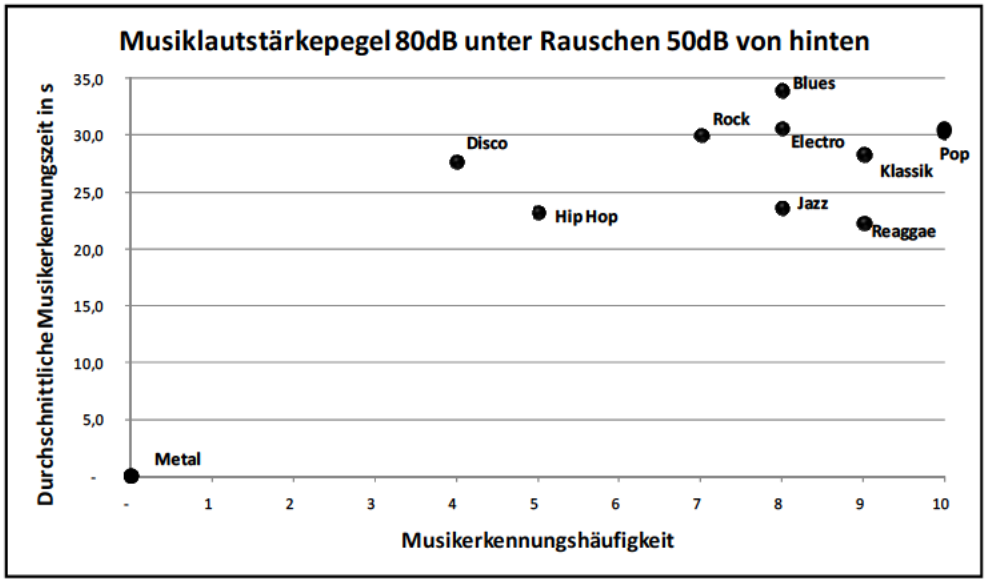

Abbildung 19: Musikerkennungshäufigkeit und –zeit bei 80dB Nutzpegel unter 50dB Rauschen von hinten

Die durchschnittlichen Erkennungszeiten sind bei 80 dB Nutzpegel mit Rauschen von hinten gegenüber Rauschen von vorne für die Genres Blues, Electro, Jazz, Klassik und Pop höher. Die anderen Genres weisen dagegen eine niedrigere Erkennungszeit auf (siehe Tabelle 19).

Es wurde untersucht, ob die Richtung des Rauschens Einfluss auf die Musikerkennungshäufigkeit nimmt. Die durchgeführten Tests haben gezeigt (siehe Tabelle 20), dass bei einem Musiklautstärkepegel von 60 dB kein statistisch signifikanter Zusammenhang zwischen der Richtung des Rauschgeräuschs und der Musikerkennungshäufigkeit besteht ($p = 0{,}074$). Ebenfalls kann kein Effekt bei einem Musiklautstärkepegel von 80 dB nachgewiesen werden ($p = 0{,}880$).

Tabelle 20: Einfluss der Richtung von Rauschen auf die Musikerkennungshäufigkeit

Einfluss Richtung Rauschen	Musikerkennungshäufigkeiten der Automatik	
	Musiklautstärkepegel 60dB	Musiklautstärkepegel 80dB
Rauschen 50dB vorne	31	67
	p = 0,074	p = 0,880
	ϕ = -0,126	ϕ = 0,011
Rauschen 50dB hinten	20	68
Summe	**51**	**135**

Mittels einer zweifaktoriellen Varianzanalyse wurde ferner untersucht, welchen Einfluss die SCAN-Automatik und die Richtung des Rauschens auf die Erkennungszeit nehmen und ob eine Interaktion zwischen beiden vorliegt. Varianzhomogenität liegt jedoch nicht vor für den Test bei 60 dB Nutzpegel. Die Ergebnisse sind entsprechend nur eingeschränkt aussagekräftig.

Bei einem Musiklautstärkepegel von 60 dB liegt ein signifikantes Gesamtmodell vor (siehe Tabelle 21). Weiterhin ist ein Haupteffekt für die SCAN-Automatik erkennbar (p = 0,000). Für die Richtung des Rauschens kann kein Effekt nachgewiesen werden (p = 0,492). Der Interaktionsterm ist ebenfalls nicht signifikant (p = 0,705). Das Gesamtmodell ist bei 80 dB Musiklautstärkepegel nicht signifikant. Die Haupteffekte werden daher auch nicht weiter betrachtet.

Tabelle 21: Tests der Zwischensubjekteffekte

Abhängige Variable: Zeit bis Erkennung der Scanautomatik

60dB	Typ III Quadrat-summe	df	Quadratischer Mittelwert	F	Sig.	Partielles Eta hoch zwei
Korrigiertes Modell	4688,367a	5	937,673	7,13	0,000	0,159
Konstanter Term	52006,681	1	52006,681	395,44	0,000	0,678
Scanautomatik	4433,705	2	2216,852	16,856	0,000	0,152
RichtungdesStörgeräuschs	62,397	1	62,397	0,474	0,492	0,003
Scanautomatik * RichtungdesStörgeräuschs	92,16	2	46,08	0,35	0,705	0,004
Fehler	24725,02	188	131,516			
Gesamtsumme	190659	194				
Korrigierter Gesamtwert	29413,387	193				

a R-Quadrat = ,159 (Angepasstes R-Quadrat = ,137)

Der Post-hoc-Test für den Haupteffekt SCAN-Automatik zeigt, dass sich bei 60 dB Musiklautstärkepegel Musik und Lärm (p = 0,000) sowie Musik und Sprache im Störgeräusch (p = 0,000) signifikant von einander unterscheiden.

Tabelle 22: Mehrfachvergleiche

Abhängige Variable: Zeit bis Erkennung der Scanautomatik
Bonferroni
60dB

(I) Scanautomatik	(J) Scanautomatik	Mittelwert-differenz (I-J)	Standard-fehler	Sig.	95 % Konfidenzintervall	
					Untergrenze	Obergrenze
Lärm	Musik	-10,22*	1,889	0,000	-14,78	-5,65
	Sprache im Störgeräusch	5,95	3,76	0,346	-3,13	15,03
Musik	Lärm	10,22*	1,889	0,000	5,65	14,78
	Sprache im Störgeräusch	16,17*	3,966	0,000	6,59	25,75
Sprache im Störgeräusch	Lärm	-5,95	3,76	0,346	-15,03	3,13
	Musik	-16,17*	3,966	0,000	-25,75	-6,59

Basierend auf beobachteten Mittelwerten.
Der Fehlerterm ist Mittel der Quadrate(Fehler) = 131,516.
* die Mittelwertdifferenz ist auf der Stufe ,05 signifikant.

Die Tests bezüglich des Einflusses der Richtung des Rauschens auf die Musikerkennungszeiten ergeben, dass es bei einem Musiklautstärkepegel von 60 dB sowie 80 dB keinen signifikanten Zusammenhang gibt (siehe Tabelle 23).

Tabelle 23: Einfluss der Richtung des Rauschens auf die Musikerkennungszeit

Einfluss Richtung Rauschen	Musikerkennungszeit der Automatik in s	
	Musiklautstärkepegel 60dB	Musiklautstärkepegel 80dB
Rauschen 50dB von vorne	37,00	27,84
	p = 0,728	p = 0,950
Rauschen 50dB von hinten	36,15	27,96

4.6. Einfluss der Richtung von Sprache auf die Musikerkennung

Die Musikerkennungshäufigkeit liegt bei einem Musiklautstärkepegel von 60 dB mit Sprache von vorne bei sechs Musikstücken und steigt auf 69 von 100 Musikstücken bei einem Lautstärkepegel von 80 dB (siehe Abschnitt 4.4). Bei Sprache von hinten liegt die Erkennungshäufigkeit mit einem Musiklautstärkepegel von 60 dB bei sieben Musikstücken bzw. bei 62 mit einem Musiklautstärkepegel von 80 dB.

Mit einem Nutzpegel von 60 dB unter Sprache von hinten wird im Vergleich zu Sprache von vorne ein Pop-Musikstück nicht erkannt. Klassik und Rock haben zudem eine um eins erhöhte Häufigkeit. Die Erkennungszeiten für beide Genres sind ebenfalls höher als unter Sprache von vorne (vgl. Abbildung 16 auf Seite 46 und Abbildung 20).

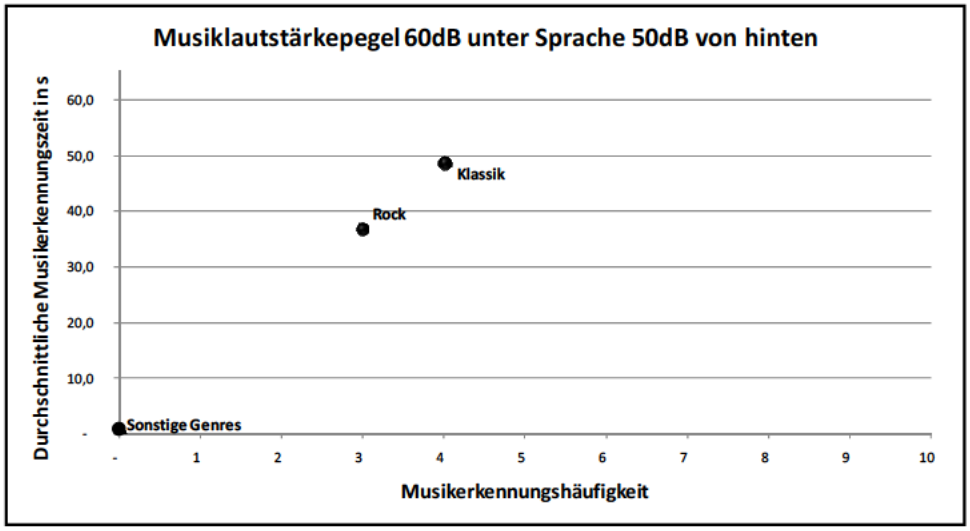

Abbildung 20: Musikerkennungshäufigkeit und –zeit bei 60dB Nutzpegel unter 50dB Sprache von hinten

Bei einem Musiklautstärkepegel von 80 dB unter Sprache von hinten ist die Musikerkennungshäufigkeit im Vergleich zu Sprache von vorne für die Genres Electro, Hip Hop, Pop, Reggae und Rock um eins niedriger. Jazz hat eine um zwei geringere Häufigkeit als bei 80 dB Nutzpegel mit Sprache von vorne. Alle anderen Genres haben keine veränderte Häufigkeit (vgl. Abbildung 17 auf Seite 47 und Abbildung 21).

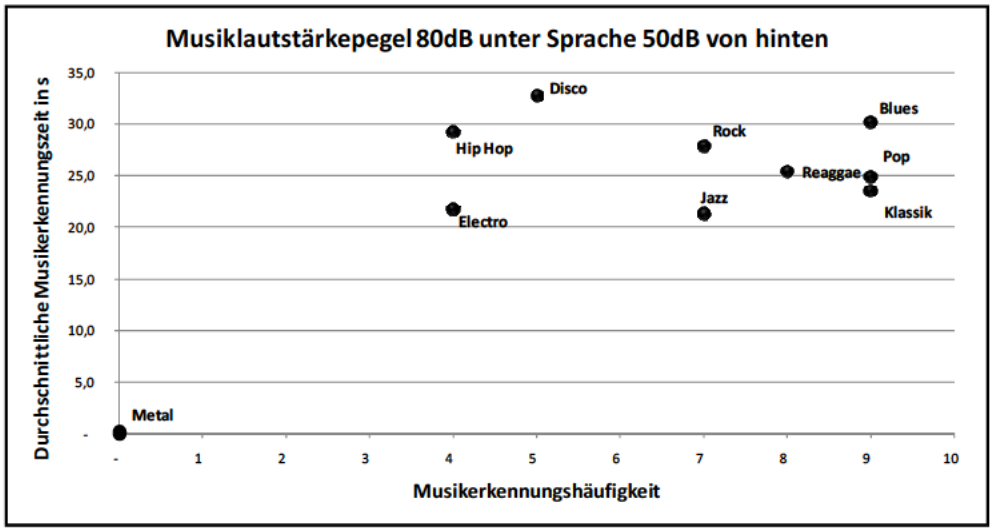

Abbildung 21: Musikerkennungshäufigkeit und –zeit bei 80dB Nutzpegel unter 50dB Sprache von hinten

Die durchschnittlichen Erkennungszeiten sind bei 80 dB Nutzpegel mit Sprache von hinten gegenüber Sprache von vorne für das Genre Blues höher, Reggae

und Rock sind annähernd stabil. Die anderen Genres weisen dagegen eine niedrigere Erkennungszeit auf (siehe Tabelle 24).

Tabelle 24: Durchschnittliche Erkennungszeit je Genre bei Rauschen von vorne und hinten bei 60dB und 80dB Nutzgeräusch

	Durchschnittliche Erkennungszeit in s								
Musiklautstärkepegel	**Blues**	**Disco**	**Electro**	**Hip Hop**	**Jazz**	**Klassik**	**Pop**	**Reggae**	**Rock**
80dB									
Sprache 50dB vorne	25,9	33,8	25,8	33,2	29,7	24,8	26,9	25,2	27,5
Sprache 50dB hinten	30,2	32,8	21,8	29,3	21,3	23,6	24,9	25,4	27,9
Veränderung	**4,3**	**- 1,0**	**- 4,1**	**- 4,0**	**- 8,4**	**- 1,2**	**- 2,0**	**0,2**	**0,4**

Die durchgeführten Chi-Quadrat-Tests bezüglich des Einflusses der Richtung von Sprache auf die Musikerkennungshäufigkeit zeigen sowohl bei 60 dB als auch 80 dB Nutzpegel keinen statistisch signifikanten Zusammenhang (siehe Tabelle 25).

Tabelle 25: Einfluss der Richtung von Sprache auf die Musikerkennungshäufigkeit

Einfluss Richtung Sprache	**Musikerkennungshäufigkeiten der Automatik**	
	Musiklautstärkepegel 60dB	**Musiklautstärkepegel 80dB**
Sprache 50dB von vorne	6	69
	p = 0,774	p = 0,298
	ϕ = 0,020	ϕ = -0,074
Sprache 50dB von hinten	7	62
Summe	**13**	**131**

Die durchgeführten zweifaktoriellen Varianzanalysen zeigen, dass das Gesamtmodell bei 80 dB Musiklautstärkepegel nicht signifikant ist. Weitergehend wird dieses Testergebnis daher nicht interpretiert. Bei 60 dB Musiklautstärkepegel ist das Gesamtmodell hingegen signifikant (siehe Tabelle 26). Es liegt jeweils ein Haupteffekt für die SCAN-Automatik vor. Für die Richtung der Sprache kann kein Haupteffekt nachgewiesen werden. Der Interaktionsterm ist jeweils nicht signifikant.

Tabelle 26: Tests der Zwischensubjekteffekte

Abhängige Variable:Zeit bis Erkennung der Scanautomatik

60dB	Typ III Quadrat-summe	df	Quadratischer Mittelwert	F	Sig.	Partielles Eta hoch zwei
Korrigiertes Modell	3498,532a	3	1166,177	7,783	0,000	0,327
Konstanter Term	34473,572	1	34473,572	230,064	0,000	0,827
Scanautomatik	3067,352	1	3067,352	20,47	0,000	0,299
RichtungdesStörgeräuschs	160,352	1	160,352	1,07	0,306	0,022
Scanautomatik * RichtungdesStörgeräuschs	278,572	1	278,572	1,859	0,179	0,037
Fehler	7192,468	48	149,843			
Gesamtsumme	44504	52				
Korrigierter Gesamtwert	10691	51				

a R-Quadrat = ,327 (Angepasstes R-Quadrat = ,285)

Der Post-hoc-Test für den Haupteffekt SCAN-Automatik bei 60 dB kann nicht durchgeführt werden, da die SCAN-Automatik hier nur die Faktorstufen Lärm und Musik annimmt und sich diese daher also signifikant voneinander unterscheiden müssen.

Die Tests bezüglich der Musikerkennungszeiten zeigen, dass unabhängig vom Nutzpegel kein Zusammenhang zwischen der Richtung der Sprache und der Musikerkennungszeit besteht (siehe Tabelle 27).

Tabelle 27: Einfluss der Richtung von Sprache auf die Musikerkennungszeit

Einfluss Richtung Sprache	Musikerkennungszeit der Automatik in s	
	Musiklautstärkepegel 60dB	Musiklautstärkepegel 80dB
Sprache 50dB von vorne	34,00	27,58
	p = 0,283	p = 0,382
Sprache 50dB von hinten	43,43	26,18

5. Musikdiskrimination

Am Fachbereich Audiokommunikation und –technik wurde ein Algorithmus zur Diskrimination von Musik, Sprache und Geräuschen erstellt, der zum Vergleich mit der Prozessortechnik herangezogen wird. Der Diskriminator stützt sich in seiner Logik und Featureauswahl auf ähnliche Algorithmen, die im Buch *Audio Content Analysis* von Alexander Lerch (2012) dargestellt sind. Da die Daten für die Musikdiskrimination ausschließlich digital verwertet werden, also keine Artefakte durch Schallaufnahme und –wandlung berücksichtigt werden müssen, entspricht es dem technischen Optimalfall zur Musikerkennung. Inwieweit die Implementierung des Algorithmus in eine Mikrotechnologie möglich ist, ist nicht Bestandteil dieser Arbeit.

Die Erstellung des Musikdiskriminators unterliegt der Disziplin des *Maschinellen Lernens*. Zum Aufgabenbereich des Maschinellen Lernens gehört unter anderem die Klassifizierung von Daten in vorgegebenen Kategorien anhand von automatisch extrahierten Informationen (Deskriptoren). Die Klassifizierungsalgorithmen lernen dabei mittels vorgegebener Daten die Klassifizierungsregeln und können sie dann als Modell auf andere, unbekannte Daten anwenden (*Überwachtes Lernen*). Es gibt verschiedene Modelle zur Datenklassifizierung, wie zum Beispiel das Hidden Markov Model oder Neuronale Netzwerke (Lerch, 2012). Im vorliegenden Anwendungsfall handelt es sich um Audiodaten, die genreübergreifend für Musik erkannt werden sollen, auch im Falle eines gleichzeitig auftretenden Störgeräuschs. Dafür bietet sich die Support Vector Machine als Klassifikator an, da dieses Modell relativ hochdimensionell, also mit vielen Merkmalsunterscheidungen arbeiten kann (Lerch, 2012). Im Wesentlichen sind für die Musikerkennung vier Bausteine notwendig, wie in der nachfolgenden schematischen Darstellung zur Musikerkennung zu sehen ist.

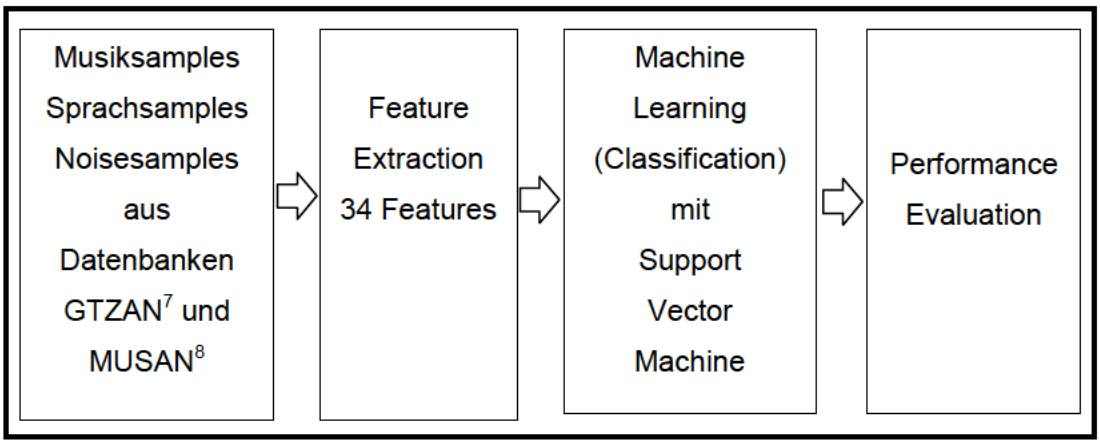

Abbildung 22: Schematische Darstellung zur Musikerkennung

Zunächst ist ein Datensatz notwendig, der die Anforderung erfüllt, verschiedene Audiodateien in einer einheitlichen Qualität zur Verfügung zu stellen, die möglichst viele unterschiedliche Musikgenres abdecken.

Mit dem an der TU-Berlin entwickelten Diskriminationsalgorithmus werden die Samples mittels 34 verschiedener Features auf ihre akustischen Eigenschaften hin untersucht und für jedes Feature ein gemittelter Wert über die Zeitspanne der Samples berechnet. Die eigentliche Klassifizierung findet schließlich mit der Support Vector Machine statt. Dafür wird der Datensatz unterteilt in Trainingssatz und Testsatz. Mit dem Trainingssatz lernt die Support Vektor Machine die verschiedenen Daten zu separieren und eine Vorhersage über die Trennschärfe zwischen den Klassen zu tätigen (siehe Abschnitt 5.3). Mit dem Testsatz wird schließlich überprüft, wie gut die Klassifizierung funktioniert. Bei binärer Klassifizierung erhält man als Ergebnis eine 2x2 Array Matrix, in der die richtig und falsch erkannten Samples aufgeführt sind.

In der Performance Evaluation wird die Validität der Diskriminationsergebnisse mit dreifacher Kreuzvalidierung überprüft, um beispielsweise Overfitting zu vermeiden. Beim Overfitting leitet der Klassifikator die Regel sehr spezifisch ab, kann aber auf Testdaten nicht generalisieren. Demnach erzielt er ausschließlich für die vorliegenden Trainingsdaten sehr gute Ergebnisse, eine generalistische Auswertung von beliebigen unbekannten Daten wäre jedoch unzureichend.

5.1. Auswahl des Trainingsmaterials

Zu Trainingszwecken des Musik/Sprach-Diskriminators werden unterschiedliche Daten aus dem MUSAN[10] Datensatz generiert. Der „Music, Speech and Noise Corpus“ bzw. MUSAN Datensatz beinhaltet insgesamt etwa 109 Stunden Audiomaterial. Davon entfallen ungefähr 60 Stunden auf Sprache, 42,5 Stunden auf Musik und der Rest verteilt sich auf 929 Dateien für Geräusche (Snyder, 2015). Die Sprachsamples liegen ausschließlich in englischer Sprache vor. Ein Drittel etwa (20 Stunden 21 Minuten) sind vorgelesene Sprache von der digitalen Bibliothek „Librivox“ mit allgemein zugänglichen Hörbüchern. Zwei Drittel (40 Stunden 1 Minute) sind öffentlich zugängliche Regierungsdebatten aus den Vereinigten Staaten von Amerika (Snyder, 2015).

Um den Diskriminator mit Samples zu trainieren, wurden 100 rauschartige Hintergrundgeräusche je eine Minute sowie 100 einminütige Sprachproben aus dem MUSAN Datensatz extrahiert. Die 100 Musiksamples wurden aus dem GTZAN-Datensatz generiert. Die Mixsamples wurden aus den beiden oben erwähnten Datensätzen mit Matlab erstellt. So liegen insgesamt folgende Daten vor:

- 100 Musiksamples aus dem GTZAN Datensatz
- 100 Sprachsamples aus dem MUSAN Datensatz
- 100 Noisesamples aus dem MUSAN Datensatz
- 100 Mixsamples (Musik und Rauschen), beides gleich gewichtet
- 100 Mixsamples (Musik und Rauschen) mit der Gewichtung: Musik ist 10 dB lauter als Rauschen
- 100 Mixsamples (Musik und Rauschen) mit der Gewichtung: Musik ist 30 dB lauter als Rauschen
- 100 Mixsamples (Musik und Sprache), beides gleich gewichtet
- 100 Mixsamples (Musik und Sprache) mit der Gewichtung: Musik ist 10 dB lauter als Sprache
- 100 Mixsamples (Musik und Sprache) mit der Gewichtung: Musik ist 30 dB lauter als Sprache

5.2. Verwendete Features

Zur Musikdiskrimination wurden insgesamt 34 Features verwendet (nach Logan, 2000; Calvo et al., 2007; Duda et al. 2010), die folgend aufgelistet sind:

[10] Abgerufen am 3.7.2018, hier verfügbar: https://arxiv.org/abs/1510.08484

- 13 verschiedene Mel Frequency Cepstral Coefficents (MFCC)
- 13 verschiedene Shifted Delta Cepstral Coefficients (SDCC)
- Spectral Flux
- Spectral Flatness
- Spectral Centroid
- Spectral Spread
- Spectral Power Tonal Ratio
- Zero Crossing Rate
- Root Mean Square (RMS) Energy
- Time Peak Envelope

Alle verwendeten Features wurden über die gesamte Länge der Aufnahme extrahiert. Um einen Feature-Wert pro Feature für die Länge der Aufnahme zu erhalten, wurde der Musikausschnitt mit einem Framing versehen, welches per Hamming-Fenster mit einer Länge von 11,6 ms (256 Samples bei einer Samplingrate von 22.050 Hz) und 75 % Overlap funktioniert. Von den resultierenden Frames wurde schließlich der Durchschnitt der Werte kalkuliert. Im Falle der spektralen und cepstralen Features bezieht sich das Framing auf das Spektrogramm, in den anderen Fällen bezieht es sich auf die diskreten Signale.
Pro Trainings- und Testsatz wurden die Features für die Klassifizierung z-normalisiert. Die vorliegenden Ergebnisse sind der Durchschnitt über 3 Folds nach der 3-fold Cross-Validation (siehe Abschnitt 5.4).
Zunächst werden hier die einzelnen Features erläutert und im Anschluss auf die Support Vector Machine und die dreifache Kreuzvalidierung näher eingegangen.

5.2.1. 13 Mel Frequency Cepstral Coefficents

Mel Frequency Cepstral Coefficents (MFCC) werden nach der Computerwissenschaftlerin Beth Logan hauptsächlich in folgenden fünf Schritten vom Zeitsignal bis hin zu den MFCCs berechnet (Logan, 2000):

1) Signal fenstern
2) Amplituden Spektrum mittels diskreter Fouriertransformation berechnen
3) Amplituden Spektrum logarithmieren
4) Konvertierung ins Mel Spektrum
5) Diskrete Cosinus Transformation

Die Signalfensterung im ersten Schritt erfolgt meist mittels Hammingfenstern (siehe Abschnitt 2.2.1) und wird mit einer Filterbank von Dreiecksfiltern gewichtet, wie sie in Abbildung 23 zu sehen ist.

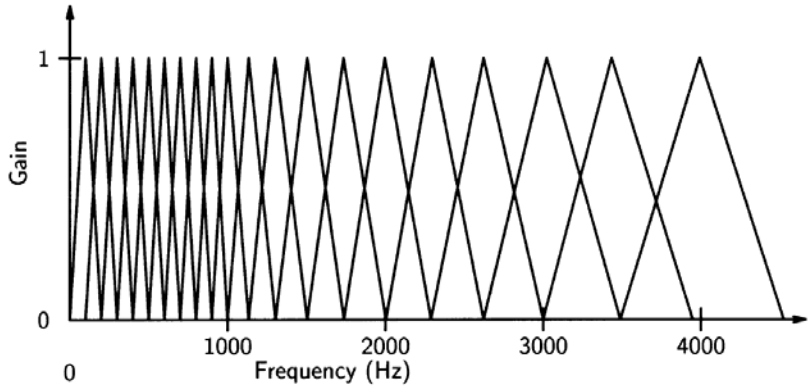

Abbildung 23: Filterbank mit Dreiecksfiltern zur mel frequency cepstrum coefficient (MFCC) Generierung nach Davis et al. (1980)

Für jedes Fenster wird ein Cepstrum-Vektor generiert. Im nächsten Schritt wird die diskrete Fouriertransformation auf jedes einzelne Fenster angewendet, um das Amplitudenspektrum zu erhalten. Angelehnt an die annähernd logarithmische Lautheitsempfindung (siehe Abschnitt 2.4.1) wird das Amplitudenspektrum logarithmiert.

Nach einer Glättung des Signals im vierten Schritt werden beispielsweise 256 gewonnene, spektrale Komponenten innerhalb von 40 Frequenz-Bins gebündelt (Logan, 2000). Angelehnt an die auditive Verarbeitung des Menschen für Tonheit erfolgt die Aufteilung der Frequenz-Bins Mel-skaliert, das heißt zu höheren Frequenzen logarithmisch (siehe Abschnitt 2.4.2).

Daraus resultieren für jedes Fenster hochkorrelierende Komponenten der Mel-Spektrum-Vektoren, die im letzten Schritt mittels diskreter Cosinus Transformation dekorreliert werden. Schließlich erhält man nach der Transformation für jedes Fenster 13 cepstrale Features (Logan, 2000).

5.2.2. 13 Shifted Delta Cepstral Coefficients

Shifted Delta Cepstral Coefficients (SDCC) dienen der Spracherkennung. Sie basieren auf der spektralen Verarbeitung von Signalen und orientieren sich

zeitlich am Phonem- und Silbenbereich (Calvo et al., 2007). Ihre Berechnung erfolgt laut Calvo et al. mit Hilfe eines Sets aus vier Parametern:

1) N: Nummer der cepstralen Koeffizienten in jedem Rahmen
2) d: Zeit im Voraus und als Verzögerung für die Delta-Berechnung
3) P: Zeitverschiebung zwischen den aufeinanderfolgenden Blöcken
4) k: Anzahl der Blocks, deren Koeffizienten verkettet sind, um den finalen Vektor zu formen.

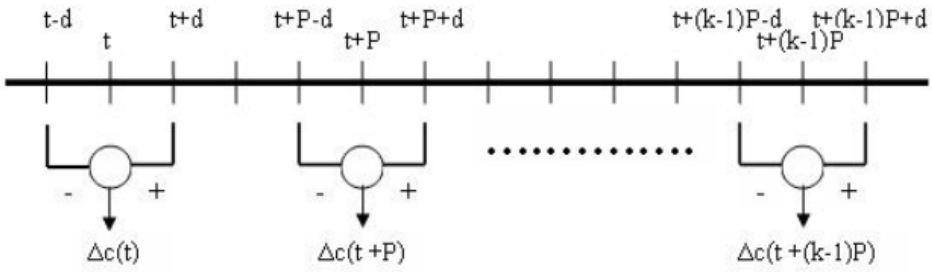

Abbildung 24: Berechnung der Shifted Delta Cepstral (Calvo et al., 2007)

In Abbildung 24 ist zu sehen, wie erst Δc in Abhängigkeit von der Zeit innerhalb der Zeitspanne d ermittelt wird. Für den zweiten Block wird zusätzlich die Zeitverschiebung P zwischen den Blöcken mit einbezogen. Dies wird für alle Blöcke der Anzahl k wiederholt. Je nach Dimension der cepstralen Koeffizienten verändert sich die Kombination der vier Parameter (*N, d, P, k*) (Calvo et al., 2007).

5.2.3. Spectral Flux

Mittels des Spectral Flux kann erkannt werden, wie stark sich ein Signalspektrum verändert. Dafür wird das Signal gefenstert, um anschließend die einzelnen aufeinanderfolgenden Frames der Short Time Fourier Transform (STFT) miteinander vergleichen zu können. Als Ergebnis erhält man den *Euklidischen Abstand* ausgedrückt in einem Wert zwischen 0 und A, mit A als die höchstmögliche Amplitude des Signals. Ist das Signal relativ stetig, gibt es einen niedrigen Ausgangswert. Sobald sich die Amplitude des Signals verändert, steigt auch der Wert zum Spectral Flux an (Lerch, 2012, S.44f). Da die Berechnungen vektorbasiert sind, bezieht sich der euklidische Abstand auf

den n-dimensionalen Vektorraum. Als Formel ausgedrückt gilt für den Spectral Flux (Lerch, 2012):

$$v_{SF}(n) = \frac{\sqrt{\sum_{\kappa=0}^{\kappa/2-1}(|X(k,n)| - |X(k,n-1)|)^2}}{\kappa/2}. \qquad (11)$$

Der Spectral Flux kann auch als Annäherung an die Empfindung der Rauigkeit aufgefasst werden (Lerch, 2012). Nach Zwicker und Fastl (2007) beschreibt die Rauigkeit die zeitliche Veränderung eines Signals und ist bei amplituden- und frequenzmodulierten Signalen wahrnehmbar. Definiert ist sie als 1 asper = Rauigkeit eines amplitudenmodulierten 1 kHz Tons bei 60 dB mit einer Modulationsfrequenz von 70 Hz und einem Modulationsgrad m = 1 (Zwicker und Fastl, 2007).

5.2.4. Spectral Flatness

Die Spectral Flatness bietet ein Maß für die Tonalität eines Geräusches im Vergleich zu seinem Rauschanteil. Sie ist definiert als Verhältnis von geometrischem Mittelwert zu arithmetischem Mittelwert (Lerch, 2012):

$$v_{TF}(n) = \frac{\sqrt[\kappa/2]{\prod_{k=0}^{\kappa/2-1}|X(k,n)|}}{2/\kappa \cdot \sum_{k=0}^{\kappa/2-1}|X(k,n)|} = \frac{\exp\left(2/\kappa \cdot \sum_{k=0}^{\kappa/2-1} log(|X(k,n)|)\right)}{2/\kappa \cdot \sum_{k=0}^{\kappa/2-1}|X(k,n)|}. \qquad (12)$$

Der resultierende Wert bewegt sich zwischen null und der maximalen spektralen Größe. Je höher der Wert desto flacher bzw. verrauschter ist das Spektrum. Je niedriger der Wert, desto tonaler ist das Spektrum. Weist das Signal an einer Stelle eine Nullstelle auf, so ist auch der Wert für die Spectral Flatness gleich null, was nicht als sehr hohe Tonalität ausgelegt werden darf und daher bei Signalpausen einer besonderen Berücksichtigung bedarf. Eine Möglichkeit, dieses Problem zu vermeiden, ist die Berechnung des arithmetischen Mittels benachbarter spektraler Koeffizienten zu berechnen, also einen gleitenden Durchschnitts-Filter zu nehmen. Um das abnehmende Frequenz-Auflösungs-Verhalten des menschlichen Ohres zu berücksichtigen,

werden die Filter mit zunehmender Frequenz länger. Meist wird ein Frequenzspektrum von 250 Hz - 16 kHz in die Berechnung mit einbezogen, um weitere nützliche Informationen herauszuziehen. Das Frequenzspektrum ist in diesem Fall in 24 überlappende Frequenzbänder unterteilt und das Ergebnis der STFT ein Vektor (Lerch 2012).

5.2.5. Spectral Centroid

Spectral Centroid gibt an, wo sich der Mittelpunkt der spektralen Energie des gesamten Spektrums befindet. Mit dem Spectral Centroid lassen sich Aussagen über die Klangfarbe in Bezug auf Helligkeit und Schärfe machen (Lerch, 2012). Der resultierende Wert zwischen null und eins zeigt den Frequenzanteil an. Niedrige Werte stehen für signifikant niedrige Frequenzen und entsprechend wenig Klangschärfe. Damit kann gut indiziert werden, ob Sprache oder Musik vorliegt, da Musik durch Percussion-Elemente meist mehr hochfrequente Geräusche aufweist (Scheirer, 1997). Mathematisch definiert ist es als frequenzgewichtete Summe über die frequenzungewichtete Summe des Energiespektrums (Lerch 2012, S.45). Bei Signalpausen ist v_{SC} nicht definiert und wird bei niedrigen Signalleveln sehr hoch.

$$v_{SC}(n) = \frac{\sum_{k=0}^{\kappa/2-1} k \cdot |X(k,n)|^2}{\sum_{k=0}^{\kappa/2-1} |X(k,n)|^2}. \qquad (13)$$

5.2.6. Spectral Spread

Dieses Feature ist eng verbunden mit dem Spectral Centroid und äquivalent zu diesem zu berechnen (Lerch 2012):

$$v_{ss}(n) = \sqrt{\frac{\sum_{k=0}^{\kappa/2-1} (k - v_{sc}(n))^2 \cdot |X(k,n)|^2}{\sum_{k=0}^{\kappa/2-1} |X(k,n)|^2}}. \qquad (14)$$

Es gibt die durchschnittliche Bandbreite des Spectral Centroid an. Auch hier gilt wie bei der Spectral Flatness, dass niedrigere Werte für niedrige Konzentration

spektraler Energie stehen und damit tonale Geräusche kennzeichnen. Wie das Spectral Centroid ist das Spectral Spread nicht definiert für Signalpausen und resultiert in diesem Fall in sehr hohe Werte (Lerch, 2012).

5.2.7. Spectral Tonal Power Ratio

Die Spectral Tonal Power Ratio detektiert die Tonalität eines Audiosignals. Im einfachsten Fall ist sie zu berechnen mit dem Verhältnis der tonalen Energie zur gesamten Energiedichte des Signals:

$$v_{Tpr} = \sqrt{\frac{E_T(n)}{\sum_{i=0}^{K/2-1} |X(k,n)|^2}} \,. \qquad (15)$$

Das Ergebnis liegt zwischen null und eins. Je höher das Ergebnis, desto höher die Indikation für ein tonales Spektrum (Lerch, 2012, S.57).

5.2.8. Zero Crossing Rate

Die Nulldurchgangsrate definiert, wie viele Nulldurchgänge ein Signal in einem definierten Zeitraum aufweist, entsprechend sein Vorzeichen von positiv zu negativ und umgekehrt ändert. Je häufiger das Vorzeichen gewechselt wird, desto hochfrequenter erscheint das Signal. Weiterhin kann die Rate zwischen den Blöcken verglichen werden. Je ähnlicher die Rate, desto periodischer wird das Signal geschätzt. Für eine stille Umgebung wäre der resultierende Wert entsprechend null, für weißes Rauschen sehr hoch und für einen Sinuston eine periodische Wiederholung des Wertes vorhanden (Lerch 2012, S.62).

5.2.9. Root Mean Square (RMS) Energy

Der Root Mean Square eines Audiosignals kalkuliert den Effektivwert des Signals:

$$v_{RMS}(n) = \sqrt{\frac{1}{\kappa} \sum_{i=i_s(n)}^{i_e(n)} x(i)^2} \qquad (16)$$

Dabei werden die höheren Signalwerte (Frequenzen) stärker gewichtet als niedrigere (Lerch 2012). Damit wird das menschliche Lautstärkeempfinden abgebildet, welches hohe Frequenzen bei niedrigeren Lautstärkepegeln als gleichlaut empfindet wie tiefe Frequenzen (siehe 2.4.1).

5.2.10. Time Peak Envelope

Der Time Peak Envelope zeigt ähnliche Ergebnisse an wie der RMS. Er bewegt sich als Wert zwischen null und eins. Null repräsentiert Stille und eins die vollskalierten Signale. Der einfachste Weg eine Einhüllende zu extrahieren ist es, das absolute Maximum zu finden. Mit Hilfe eines PPM (Peak Programme Meter) wird die Einhüllende Sample für Sample generiert. Dabei spielen Ein- und Ausschwingzeiten eine Rolle. Sie werden auch *„attack“*- und *„release-“* time genannt. Die attack time ist wesentlich kürzer als die release time (Lerch 2012).
Die 34 Features decken verschiedene Signal- und Spektraleigenschaften ab, die für die Diskriminierung von Sprache, Musik und Rauschen hilfreich sind. Spectral Centroid beispielsweise wird voraussichtlich andere Werte bei Sprache liefern (die sich hauptsächlich im mittleren Frequenzbereich konzentriert) als bei Musik (die bezüglich der Frequenz einen sehr viel breiteren Bereich beansprucht). Genauso verhält es sich mit dem Spectral Flux, der eine hohe Spektrumsvarietät bei Sprache ausmacht, während das Spektrum für Rauschen relativ stetig ausfällt und somit einen anderen Wert liefert.

5.3. Support Vector Machine

Die Support Vector Machine arbeitet nach dem Prinzip, eine gradlinige Entscheidungsgrenze bzw. Hyperplane zu finden, die die unterschiedlichen Datensammlungen für eine beliebige Anzahl an Dimensionen separiert. Der einfachste Fall wäre zweidimensional. Man unterscheidet dabei, ob die Datenpunkte linear eindeutig trennbar sind, wie in Abbildung 26 zu sehen ist,

oder ob die Trennlinie nicht linear ist, wie in Abbildung 25 links zu sehen ist. Die Trennlinie ist in diesem Fall für die unterschiedlichen Klassen im zweidimensionalen Raum keine Gerade, sondern eine Kurve (Burgers, 1998).

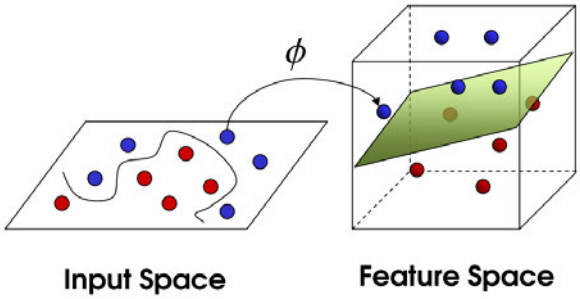

Abbildung 25: Lineare Datentrennung zweidimensional links und dreidimensional rechts[11]

Um eine gradlinige Trennung zu erhalten, ist eine Transformation der Datenpunkte in eine höhere Ebene, in diesem Fall in eine dreidimensionale Ebene nötig. So können die neu angeordneten Datenpunkte mittels einer geraden Hyperplane getrennt werden, sind entsprechend klassifiziert und können anschließend wieder zurücktransformiert werden. Die Lösung ergibt sich, indem der Raum der Datenzustände $\mathbb{R}^n$, welcher alle Datenpaare aufführt, die von Systemeingang und -ausgang erfasst werden, verlassen wird. Man begibt sich stattdessen in einen höherdimensionalen Merkmalsraum *H* (Heinert, 2010). Durch die folgende Transformation wird jedes Muster durch die Transformation auf einen Vektor x in der Hyperebene abgebildet:

$$\phi: X \rightarrow H, so\ dass\ x \rightarrow \boldsymbol{x} = \phi\ (x). \qquad (17)$$

Mit dem *Kernel-Trick* funktioniert unter Beachtung des Mercer Theorems die Rücktransformation in den Zustandsraum der Datengebung nach der Klassenseparation. Der Satz von Mercer drückt vereinfacht aus, dass jede Hyperebene im höherdimensionellen Raum einem Repräsentanten im ursprünglichen Datenraum zugeordnet werden kann. Dies ist eine stetige, symmetrische, positiv definierte Kernfunktion $K\ (x_i,x_j)$, die sich als Skalarprodukt $(xi,xj)=\langle\phi(xi),\phi(xj)\rangle$ ausdrücken lässt (Heinert, 2010). Das Mercer

[11]Abgerufen am 18.10.2018: https://stackoverflow.com/questions/9480605/what-is-the-relation-between-the-number-of-support-vectors-and-training-data-and/9484824#9484824

Theorem beschreibt demnach, welche Kernfunktion geeignet ist. Darunter zählt auch die hier verwendete Kreuzkorrelationsfunktion.
Die optimale Hyperplane liegt im gleichmäßigen Abstand zu den Datenpunkten der verschiedenen Klassen, die die kleinste Differenz zueinander aufweisen. Somit entsteht eine größtmögliche Spannweite zwischen den Klassen (maximum Margin, siehe
Abbildung **26**), um Klassifizierungsfehler so gering wie möglich zu halten.

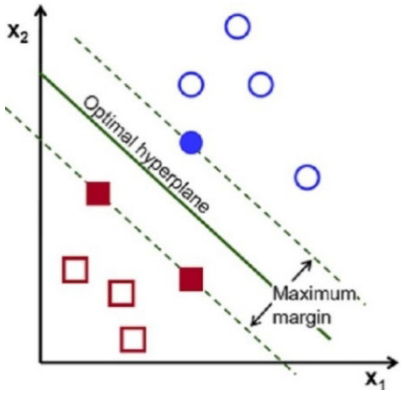

Abbildung 26: Trennung der Klassen durch optimale Hyperplane[12]

Die Datenpunkte, die am nächsten zur Hyperplane liegen, dienen auch als Support Vektoren und Namensgeber des Klassifikators, wie in Abbildung 26 als ausgefüllte Quadrate bzw. Kreise zu sehen (Burgers, 1998).

5.4. Dreifache Kreuzvalidierung

Zur Sicherstellung valider Ergebnisse wird die dreifache Kreuzvalidierung auf Trainings- und Testsatz angewendet. Das bedeutet, dass der Datensatz von jeweils 100 Samples in drei etwa gleich große Teile (Folds) zerlegt wird. Ein Fold davon wird als Testsatz definiert, die anderen beiden Folds als Trainingssatz. Diese drei Folds werden dreimal gegeneinander getestet, so dass der Testsatz alle Werte einmalig angenommen hat (Lerch, 2012). Als Ergebnis erhält man den Durchschnittswert über alle drei Folds für die Accuracy der Diskrimination. Durch das kreuzweise Testen wird die Varianz der Accuracy

[12]Abgerufen am 18.10.2018: https://stackoverflow.com/questions/9480605/what-is-the-relation-between-the-number-of-support-vectors-and-training-data-and/9484824#9484824

zur Diskrimination verringert und die Ergebnisse erhalten eine stabilere Aussagekraft.

6. Ergebnisse des Musikdiskriminators

Um die Testsituation des Cochlea-Implantats nachzustellen, wurden verschiedene Szenarien mit dem Diskriminator geprüft. Einerseits wurden die spezifischen Genres gegenüber Rauschen bzw. Sprache unterschieden. Andererseits wurden Musik im Allgemeinen sowie Sprache, Noise und einzelne Genres (Hip Hop, Klassik, Metal) verschiedenen Mix-Klassen gegenübergestellt. Diese Mix-Klassen empfinden die Messszenarien mit dem CI System nach. Das Cochlea-Implantat nimmt aus einem Gemisch von Rauschen und Musik bzw. Sprache und Musik eine Klassifizierung vor. Entsprechend soll die Situation für den Diskriminator erfolgen. Das bedeutet, dass zwischen zwei Klassen die Trennschärfe ermittelt wurde. Klasse 1 beinhaltet Musik, Noise oder Sprache und Klasse 2 hat als Bestandteil Musik und Noise bzw. Sprache in drei verschiedenen Gewichtungen inne:

1) Gleiche Gewichtung
2) Musik 60 dB, Noise oder Sprache 50 dB
3) Musik 80 dB, Noise oder Sprache 50 dB

Aus zeitlichen Gründen wurden für das Szenario mit der gemischten Klasse drei statt zehn Genres beispielhaft verglichen. Die Ergebnisse zur Trennschärfe des Musikdiskriminators stellen den Optimalfall zu Unterscheidung von Musik gegenüber einem anderen Signal dar. Die Daten sind ohne Umwege über Lautsprecher und Mikrofon direkt zur Auswertung verfügbar und die Berechnung konzentriert sich ausschließlich auf zwei Klassen. Entsprechend fallen die Ergebnisse anders aus als bei dem CI Prozessor, der entscheiden muss bei welcher der sechs Szenen zur SCAN-Automatik die Wahrscheinlichkeit am höchsten liegt, dass diese Situation gerade vorliegt.

Für 100 Musiksamples gegenüber 100 Noisesamples in gleicher Gewichtung liegt die Trennschärfe des Diskriminators bei 89 %. Er steigert die Trennschärfe bei der Zuordnung von Musik und Sprache auf 99,5 %.

Im Vergleich dazu betrug die Erkennungshäufigkeit des Cochlea-Implantats in Ruhe für alle Musikstücke bei 80 dB Lautstärke 73 von hundert Musikstücken. Mit zusätzlichem Rauschen von 50 dB reduzierte sich die Erkennungshäufigkeit des Cochlea-Implantats auf 67 von hundert Musikstücken bei einem Nutzpegel

von 80 dB. Bei einem Sprachpegel von 50 dB und einem Nutzsignal von 80 dB liegt die Musikerkennung bei 69 von hundert Musikstücken.

6.1. Musik-Genre und Noise

Die allgemeinen Ergebnisse spiegeln sich in den genrespezifischen Ergebnissen wider. Genre und Noise bzw. Sprache sind gleichwertig gewichtet. Am höchsten fällt die Separierungsgenauigkeit von Genre gegenüber Noise bei Pop aus mit 98 %, gefolgt von Blues mit 97 % und Klassik mit 96 %. Die restlichen Genres bewegen sich zwischen 93 % und 95 % (siehe Tabelle 28). Die Spannweite, in der sich die Diskriminationsgenauigkeit bewegt, ist nicht höher als 10% und in jedem Fall höher als 90 %, was einer eindeutigen Zuordnung sehr nahe kommt. Im Vergleich zur Musik und Noise im Allgemeinen verringert sich hier die Anzahl der Elemente im Testsatz auf 10 Stücke pro Genre, wodurch sich die Trennschärfe verbessert, allerdings muss man hier auf Overfitting achten.

Tabelle 28: Separierungsgenauigkeit für jedes Genre gegenüber Rauschen

Klasse 1	Klasse 2	Trennschärfe in %
Blues	Noise	**97,0**
Disco	Noise	**95,5**
Electro	Noise	**95,0**
Hip Hop	Noise	**94,5**
Jazz	Noise	**92,5**
Klassik	Noise	**96,0**
Metal	Noise	**98,0**
Pop	Noise	**98,0**
Reggae	Noise	**93,0**
Rock	Noise	**93,5**

6.2. Musik-Genre und Sprache

Die genrespezifischen Ergebnisse der Separierungsgenauigkeit liegen für Musik und Sprache etwas höher als für Musik und Noise. Bewegten sich die Ergebnisse für Musik und Rauschen zwischen 92,5 % und 98 %, so reduziert sich die Spannweite der Trennschärfe für Musik und Sprache auf 1,5 % und liegt im schlechtesten Fall bei 98% für das Genre Disco und bei 99,5 % für die

besten Genres Jazz und Electro. In Tabelle 29 sind die Ergebnisse einzeln aufgeführt.

Tabelle 29: Separierungsgenauigkeit für jedes Genre gegenüber Rauschen

Klasse 1	Klasse 2	Trennschärfe in %
Blues	Sprache	**99,0**
Disco	Sprache	**98,0**
Electro	Sprache	**99,5**
Hip Hop	Sprache	**98,0**
Jazz	Sprache	**99,5**
Klassik	Sprache	**99,0**
Metal	Sprache	**99,0**
Pop	Sprache	**99,0**
Reggae	Sprache	**99,0**
Rock	Sprache	**98,5**

6.3. Musik und Mix Musik-Noise im gleichen Verhältnis

Für die gemischten Klassen und Musik verringert sich die Genauigkeit der Differenzierung deutlich um 18% im Vergleich zur Differenzierung Musik und Noise alleine (siehe Tabelle 30). Für die drei Genres Metal, Klassik und Hip Hop ist die Genauigkeit sehr ähnlich mit etwa 90 % und liegt damit 4-8 % niedriger als die Trennschärfe für Musik und Noise. Am besten wird Sprache mit 98 % diskriminiert, Noise dagegen unterliegt einer Trennschärfe von 84 % (siehe Tabelle 30).

Tabelle 30: Separierungsgenauigkeit für Musik gegenüber der Mix-Klasse Musik und Noise gleich gewichtet

Klasse 1	Klasse 2	Trennschärfe in %
Musik	Mix Musik 60dB Noise 60dB	**69,0**
Metal	Mix Musik 60dB Noise 60dB	**90,0**
Klassik	Mix Musik 60dB Noise 60dB	**90,5**
Hip Hop	Mix Musik 60dB Noise 60dB	**90,5**
Sprache	Mix Musik 60dB Noise 60dB	**98,5**
Noise	Mix Musik 60dB Noise 60dB	**84,0**

6.4. Musik und Mix Musik-Noise im Verhältnis 60 dB zu 50 dB

Verändert man das Verhältnis in der Mix-Gruppe dahingehend, dass Musik 10dB lauter ist als Noise, steigt die Separierungsgenauigkeit für Musik im Allgemeinen von 69 % auf 76 % (siehe Tabelle 31). Sprache und Noise steigern sich ebenfalls. Für Metal sinkt die Accuracy auf 88,5 %, für Klassik und Hip Hop bleibt sie gleich wie in folgender Tabelle aufgeführt.

Tabelle 31: Separierungsgenauigkeit für Musik gegenüber der Mix-Klasse Musik 60dB und Noise 50dB

Klasse 1	Klasse 2	Trennschärfe in %
Musik	Mix Musik 60dB Noise 50dB	**76,5**
Metal	Mix Musik 60dB Noise 50dB	**88,5**
Klassik	Mix Musik 60dB Noise 50dB	**90,5**
Hip Hop	Mix Musik 60dB Noise 50dB	**90,5**
Sprache	Mix Musik 60dB Noise 50dB	**99,5**
Noise	Mix Musik 60dB Noise 50dB	**86,0**

Die Musikerkennung der SCAN-Automatik liegt bei 31 Musikstücken unter der Voraussetzung 60 dB Musik und 50 dB Rauschen. Am deutlichsten ist der Unterschied zwischen Separierungsgenauigkeit des Diskriminators und der Klassifizierung mittels SCAN-Automatik des CI Prozessors am Genre Metal auszumachen. Während es vom CI Prozessor in keinem Szenario erkannt wird, wird es vom Diskriminator mit 88,5 % Genauigkeit ausgemacht. Für Klassik und Hip Hop werden vom CI Prozessor weniger als die Hälfte der Stücke unter Rauschen erkannt (Hip Hop 2 von 10, Klassik 4 von 10). Die Trennschärfe des Diskriminators liegt bei 90,5 % für Hip Hop sowie Klassik.

6.5. Musik und Mix Musik-Noise im Verhältnis 80 dB zu 50 dB

Das Verhältnis der Mix-Gruppe liegt nun bei 80 dB Musik zu 50 dB Rauschen. Es ist zu sehen, dass die Genauigkeit des Diskriminators mit zunehmendem Musiklautstärkepegel für Musik im Allgemeinen abnimmt. Ebenso verhält es sich für Metal. Für die übrigen Klassen wird die Performance gesteigert (siehe Tabelle 32).

Tabelle 32: Separierungsgenauigkeit für Musik gegenüber der Mix-Klasse Musik 80dB und Noise 50dB

Klasse 1	Klasse 2	Trennschärfe in %
Musik	Mix Musik 80dB Noise 50dB	**72,0**
Metal	Mix Musik 80dB Noise 50dB	**85,0**
Klassik	Mix Musik 80dB Noise 50dB	**92,5**
Hip Hop	Mix Musik 80dB Noise 50dB	**92,5**
Sprache	Mix Musik 80dB Noise 50dB	**99,5**
Noise	Mix Musik 80dB Noise 50dB	**87,0**

Die Ergebnisse verhalten sich gegensätzlich zu denen der SCAN-Automatik des CI Prozessors, wo alle Genres öfter erkannt werden, also Musik im Allgemeinen bei 80 dB Musik und 50 dB Rauschen mit 67 von 100 Stücken besser klassifiziert wird als bei 60 dB Musik und 50 dB Rauschen.

6.6. Musik gegenüber Mix Musik-Sprache im gleichen Verhältnis

Wie bei den Ergebnissen mit dem Noise-Mix verringert sich die Genauigkeit der Differenzierung für alle getesteten Klassen. Für Musik im Allgemeinen verringert sich die Trennschärfe von 99 % auf 72 % und weist damit einen geringeren Abfall auf als die Differenzierung für Musik gegenüber dem Noise-Mix in gleicher Gewichtung. Es schneiden nicht alle Klassen in dieser Differenzierung besser ab als die Differenzierung für den Noise-Mix. Für Metal ist das Ergebnis mit 88 % niedriger. Alle anderen Klassen sind im Vergleich zum Noise-Mix besser oder gleichwertig (siehe Tabelle 33).

Tabelle 33: Separierungsgenauigkeit für Musik gegenüber der Mix-Klasse Musik und Sprache gleich gewichtet

Klasse 1	Klasse 2	Trennschärfe in %
Musik	Mix Musik 60dB Sprache 60dB	**72,0**
Metal	Mix Musik 60dB Sprache 60dB	**88,0**
Klassik	Mix Musik 60dB Sprache 60dB	**91,5**
Hip Hop	Mix Musik 60dB Sprache 60dB	**91,5**
Sprache	Mix Musik 60dB Sprache 60dB	**98,5**
Noise	Mix Musik 60dB Sprache 60dB	**85,5**

6.7. Musik und Mix Musik-Sprache 60 dB zu 50 dB

Wird das Verhältnis von Musik und Sprache um 10dB zugunsten der Musik verändert, bleiben die Ergebnisse weitestgehend gleich, bis auf Metal, das sich um über drei Prozentpunkte verbessert und damit sogar Hip Hop und Klassik übertrifft (siehe Tabelle 34).

Tabelle 34: Separierungsgenauigkeit für Musik gegenüber der Mix-Klasse Musik 60dB und Sprache 50dB

Klasse 1	Klasse 2	Trennschärfe in %
Musik	Mix Musik 60dB Sprache 50dB	**73,0**
Metal	Mix Musik 60dB Sprache 50dB	**91,5**
Klassik	Mix Musik 60dB Sprache 50dB	**91,0**
Hip Hop	Mix Musik 60dB Sprache 50dB	**91,0**
Sprache	Mix Musik 60dB Sprache 50dB	**98,5**
Noise	Mix Musik 60dB Sprache 50dB	**85,5**

Die Musikerkennungshäufigkeit des Cochlea-Systems liegt bei 60 dB Musiklautstärkepegel und 50 dB Sprachlautstärkepegel bei 6 von 100 Stücken. Davon sind drei Musikausschnitte aus dem Genre Klassik, zwei aus dem Genre Rock und eins aus dem Genre Pop.

6.8. Genre und Mix Musik-Sprache 80 dB zu 50 dB

Das Verhältnis der Mixgruppe beläuft sich auf 80 dB Musik und 50 dB Sprache. Während die Ergebnisse für das vorangegangene Verhältnis 60 dB zu 50 dB größtenteils gleich bleiben, verschlechtern sie sich für die einzelnen Genres (siehe Tabelle 35). Vor allem für Hip Hop verringert sich die Trennschärfe von rund 91 % auf 75,5 %. Für Metal liegt sie bei 85 % und für Klassik bei 86 %. Die Separierungsgenauigkeit für Musik im Allgemeinen bleibt mit 72,5 % relativ stabil, genauso wie bei Sprache mit 98 %. Die Diskrimination für Noise verbessert sich auf 89 %.

Tabelle 35: Separierungsgenauigkeit für Musik gegenüber der Mix-Klasse Musik 80dB und Sprache 50dB

Klasse 1	Klasse 2	Trennschärfe in %
Musik	Mix Musik 80dB Sprache 50dB	**72,5**
Metal	Mix Musik 80dB Sprache 50dB	**86,0**
Klassik	Mix Musik 80dB Sprache 50dB	**85,5**
Hip Hop	Mix Musik 80dB Sprache 50dB	**75,5**
Sprache	Mix Musik 80dB Sprache 50dB	**98,0**
Noise	Mix Musik 80dB Sprache 50dB	**89,0**

Im Gegensatz dazu verbessert sich die Klassifizierung beim Cochlea-Implantat erheblich mit 69 von 100 Stücken. Bis auf Metal werden alle Genres zumindest zur Hälfte erkannt. Pop liegt sogar bei 10 von 10, dahinter Jazz, Blues, Reggae und Klassik mit 9 von 10 sowie Rock mit 8 von 10 Musikausschnitten.

7. Zusammenfassung und Diskussion

Die Möglichkeit Musik zu hören und zu genießen, ist womöglich ein wesentlicher Faktor, die Zufriedenheit von CI-Patienten zu verbessern und die Akzeptanz für das Gerät an sich weiter zu erhöhen. Neue, an Musik angepasste Signalverarbeitungsstrategien könnten helfen, den Musikgenuss von CI-Trägern zu erhöhen. Damit der Prozessor eine verbesserte Musikverarbeitungsstrategie anwenden kann, muss er zunächst jedoch erkennen, dass es sich bei den Umgebungsgeräuschen um Musik handelt. Diese Arbeit behandelt daher, wie akkurat Musik mit dem gegenwärtigen Stand der Technik detektiert werden kann und welche Faktoren Einfluss hierauf nehmen.

Die eingangs aufgestellten Annahmen zum Einfluss von bestimmten Faktoren auf die Musikerkennung der CI SCAN-Automatik konnten nur zum Teil bestätigt werden. Die Ergebnisse zeigen zusammenfassend, dass

- die Musikerkennungshäufigkeit der CI SCAN-Automatik genrespezifisch variiert
- sich mit zunehmendem Musiklautstärkepegel die Musikerkennungshäufigkeit der CI SCAN-Automatik sowohl in ruhiger Umgebung als auch unter Einfluss von Störgeräuschen verbessert
- der Musiklautstärkepegel in ruhiger Umgebung keinen Einfluss auf die Musikerkennungszeit hat, unter Störgeräuschen aber teilweise ein signifikanter Effekt nachgewiesen werden kann
- Störgeräusche von 50 dB Lautstärkepegel einen Einfluss auf die Musikerkennungshäufigkeit für einen Musiklautstärkepegel von 60 dB, nicht jedoch für einen Musiklautstärkepegel von 80 dB haben
- ein Effekt auf die Musikerkennungszeit nur für das Störgeräusch Rauschen mit 50 dB unter 60 dB Musiklautstärkepegel zu sehen ist
- die Richtung des Störgeräuschs die Musikerkennungshäufigkeit und -zeit der CI SCAN-Automatik - entgegen der anfänglichen Vermutung - nicht beeinflusst.

Bei näherer Betrachtung der Untersuchungsergebnisse zur Musikerkennung der SCAN-Automatik und des Musikdiskriminators können diese dahingehend gedeutet werden, dass Musik nur unter bestimmten Voraussetzungen mit der gegenwärtigen Technik hinreichend genau, d.h. zum weit überwiegenden Teil,

erkannt wird. Die SCAN-Automatik erreicht bestenfalls eine Musikerkennungshäufigkeit von 73 von hundert getesteten Musikstücken. Die Ergebnisse zur Trennschärfe des Musikdiskriminators zeigen, dass Separierungsgenauigkeiten für Musik von über 90 % technisch möglich sind. Dabei muss jedoch berücksichtigt werden, dass der Musikdiskriminator die Daten ohne Umwege über Lautsprecher und Mikrofon direkt als digitale Daten verarbeiten kann und die Situationserkennung zudem nur zwischen zwei Klassen anstelle der sechs Modi der SCAN-Automatik zu unterscheiden hat. Möglicherweise weist die SCAN-Automatik allein aus diesem Grund bereits eine allgemein niedrigere Musikerkennungsrate aus.

Die Untersuchungsergebnisse legen zudem nahe, dass das abgespielte Genre die Musikerkennung beeinflusst. Die SCAN-Automatik scheint allgemein Pop- und Klassikstücke deutlich häufiger zu erkennen als andere Musikrichtungen. Dies liegt möglicherweise an den klar identifizierbaren Obertönen von Gesang bzw. Instrumenten innerhalb des Frequenzspektrums. Metal dagegen wird in keinem Fall als Musik klassifiziert, möglicherweise da Metal eine stark rauschähnliche Struktur aufweist und damit eher von der SCAN-Automatik als Lärm klassifiziert wird. Beim Musikdiskriminator sticht keines der getesteten Genres besonders hervor. Die Separierungsgenauigkeit liegt meist bei über 70 %. Zudem wird auch Metal Musik erkannt.

Weiterhin kann gezeigt werden, dass der Lautstärkepegel der Musik Einfluss auf die Musikerkennung nimmt. Allgemein steigt mit zunehmendem Lautstärkepegel der Musik unter sonst gleichen Bedingungen die Musikerkennungshäufigkeit durch die SCAN-Automatik an. Der Effekt des Musiklautstärkepegels scheint zudem unter Störgeräuschen noch größer zu sein. Während bei einem Musiklautstärkepegel von 80 dB auch unter Nebengeräuschen mit 50 dB fast annähernd gute Musikerkennungshäufigkeiten der SCAN-Automatik wie in ruhiger Umgebung erreicht werden, sinken diese bei 60 dB Musiklautstärkepegel unter Nebengeräuschen von 50 dB deutlich unter die Werte in ruhiger Umgebung. Weist die SCAN-Automatik also eine höhere Erkennungsrate auf je höher der Lautstärkepegel der Musik unter gleich laut bleibendem Störpegel ist, sinkt im Gegenzug dazu die Zuordnungsgenauigkeit des Diskriminators überwiegend. Dies liegt

möglicherweise daran, dass - wenn Musik in der Mischgruppe stärker gewichtet wird - es sich für den Diskriminator schwieriger gestaltet, eine Trennlinie zwischen den Ebenen zu ziehen, da die Datenpunkte in beiden Kategorien stark vertreten sind. Beide Arbeitsweisen zur Musikerkennung scheinen daher Vor- und Nachteile zu haben. Zwar weist der Diskriminator allgemein eine hohe Trennschärfe auf und erkennt auch Musikrichtungen wie Metal, jedoch scheint es dem Diskriminator - bedingt durch die Arbeitsweise des Algorithmus - weniger gut zu gelingen, bei höheren Musiklautstärkepegeln unter gleichlaut bleibenden Störpegel, die Zuordnungsgenauigkeit zu erhalten. Das Feature Spectral Flatness des TU-Diskriminators beispielsweise garantiert eine relativ hohe Zuordnungsgenauigkeit zwischen Geräuschen und Musik, aber nicht zwischen zwei verschiedenen Musikrichtungen. Die stärkere Gewichtung der Musik ist wiederum von Vorteil für die SCAN-Automatik, da die Musikeigenschaften prägnanter aus der akustischen Situation hervortreten und Musik entsprechend einfacher erkannt werden kann.

Die Ergebnisse, welche zeigen, dass über ein Viertel aller Titel in ruhiger Umgebung nicht erkannt werden, bestätigen die Vermutung von Busch et al. (2017) hinsichtlich des Dataloggings. Musik wird abhängig von Genre und akustischer Situation nicht zu 100 % akkurat vom Szenenklassifizierer identifiziert und somit auch nicht als musikalische Situation im Datalogging gespeichert wird. Entsprechend ist es möglich, dass der akustische Alltag von CI Trägern mehr Musik beinhaltet als über das Datalogging aufgezeichnet wird.

Beobachtet werden konnte für die SCAN-Automatik ferner, dass die Situationserkennung Zeit benötigt. Es wurde deutlich, dass Lärm durchschnittlich schneller erkannt wird als Musik. Dies liegt vermutlich daran, dass die Struktur für Lärm weniger komplex ist als für Musik und daher einfacher für die SCAN-Automatik zu erkennen ist. In ruhiger Umgebung liegt die durchschnittliche Musikerkennungszeit bereits bei 29 s bei 60 dB bzw. 25,5 s bei 80 dB Nutzpegel. Unter Einfluss von Störgeräuschen steigen die durchschnittlichen Musikerkennungszeiten teilweise deutlich an. Dies deckt sich mit der beschriebenen Arbeitsweise der SCAN-Automatik, die je nach akustischer Situation entscheidet, welche Situation vorliegt. Treten vielfältige akustische Informationen gleichzeitig auf, so muss die Situation häufiger

überprüft werden, um alle akustischen Informationen richtig einordnen zu können.

Der Vorteil der schnellen Lärmidentifizierung schlägt sich in mehreren Studien zur Spracherkennung in ruhiger und geräuschvoller Umgebung nieder in denen das Sprachverstehen mit fortschreitender Technik stetig verbessert wurde (Ceulaer et al. (2014), Müller-Deile et al., (2009)). Wird Lärm gut identifiziert, kann er auch effektiv herausgefiltert werden, um Sprache hervorzuheben. Das lässt vermuten, dass die Differenzierung von Lärm und Sprache mit jeder neuen Prozessor-Generation zuverlässiger funktioniert, sonst könnten die signalverarbeitenden Strategien nicht effektiv reagieren und sich das Sprachverstehen der CI-Träger nicht verbessern. Leider gibt es keine dem Autor bekannte Veröffentlichung zu Erkennungsrate und -zeit von Sprache, Musik oder Lärm in Hörgeräten oder Cochlea-Implantaten[13]. Somit können lediglich durch die Performance der Hörgeschädigten zum Sprachverstehen, wie in den aufgeführten Studien gezeigt, Rückschlüsse auf die Erkennungszuverlässigkeit gezogen werden. Spracherkennungs- und Sprachverarbeitungsstrategien sind im Ansatz schon seit der Cochlear Freedom Generation von 2005 mit SmartSound® verfügbar (Gilden et al., 2016), konnten entsprechend schon mehr Entwicklungszeit erfahren und vermutlich höhere Erkennungsraten erzielen als für Musik. Die Musikerkennung wurde bei der Firma Cochlear mit der intelligenten Szeneneklassifizierung SmartSound iQ® im Nucleus 6 System 2013 eingeführt (Ceulaer, 2014) und ist somit in der zweiten Sound-Prozessor-Generation und daher acht Jahre jünger im Vergleich zu sprachverarbeitenden Strategien.

Zusammenfassend zeigt sich, dass Musik nicht nur häufiger von der SCAN-Automatik in ruhiger Umgebung erkannt wird, sondern auch schneller als unter Störgeräuschen. Die Erkennungsraten sind möglicherweise jedoch zu gering und die Erkennungszeiten zu hoch, um darauf aufbauend mit einer veränderten Musikverarbeitungsstrategie den Musikgenuss bei CI-Patienten allgemein steigern zu können. Lediglich für ausgewählte Musikrichtungen, wie Pop und Klassik, die recht gut von der SCAN-Automatik erkannt werden, könnte dies

[13] Stand 01.03.2019, unter der Schlagwortsuche auf Deutsch und Englisch: Spracherkennungsrate in Hörgeräten/CI, Musikerkennungsrate in Hörgeräten/CI vorverarbeitende Strategien des Signalprozessors, Lärmidentifizierung bei Hörgeräten/CI, Automatische Szenenklassifikation, Sprachdetection HG/CI, u.a.

womöglich gelingen. Der CI-Träger müsste sich dabei jedoch in ruhiger Umgebung befinden und der Musiklautstärkepegel entsprechend hoch eingestellt sein. Bestimmte musikalische Ereignisse wie ein Festival mit lautem Publikum könnte der CI-Träger voraussichtlich nur unter Einschränkungen genießen. Möchte man als CI-Träger Musik für sich allein genießen bietet die Firma Cochlear allerdings auch das direkte Streamen der Musik vom iPhone auf die CI Systeme, welches den Hörgenuss von Musik steigern kann (Dölle, 2018).

In einer weitergehenden Probandenstudie wäre es interessant zu untersuchen, ob beim Cochlear CI System genrespezifisch Vorteile für den CI-Träger hinsichtlich des Musikgenusses feststellbar sind. Inwiefern wirkt sich die an die Musikerkennung anschließende Verarbeitungsstrategie vorteilhaft auf die Bewertung zum Musikgenuss seitens des CI-Trägers aus und ist ein Unterschied zwischen verschiedenen Musikrichtungen feststellbar?

Die durchgeführten Messungen bezüglich der Arbeitsweise der SCAN-Automatik könnten in einer weitergehenden Untersuchung zudem durch feinere Abstufungen des Musiklautstärkepegels (z.B. zusätzlich 50 dB, 70 dB und 90 dB) erweitert werden. Die Untersuchungen lassen erkennen, dass 80 dB Lautstärkepegel gegenüber 60 dB vorteilhaft für die Musikerkennung ist, nicht aber ob eine weitergehende Erhöhung eine weitere Verbesserung mit sich bringt bzw. ob sich gar ein optimaler Lautstärkepegel identifizieren lässt.

Für zukünftige Studiendesigns wäre es jedoch von Vorteil den Messumfang zu erhöhen. Der Messumfang in der vorliegenden Arbeit ist möglicherweise zu gering, um allgemeine, statistisch valide Aussagen hinsichtlich der Zusammenhänge je Genre zu tätigen. Zudem erhielten die Aussagen zur Musikerkennungszeit allgemein größere Bedeutung, da die erkannten Häufigkeiten von Musik in bestimmten Testsituationen teilweise sehr niedrig waren. Entsprechend ist die Aussagekraft zu den durchschnittlichen Musikerkennungszeiten möglicherweise durch Ausreißer geprägt.

8. Literaturverzeichnis

Adunka, Oliver und Jan Kiefer (2005): „Wie funktioniert der Sprachprozessor in Cochlea-Implantaten?“ In: *Central European Media Enterprises Ltd.* Laryngo−Rhino−Otol 84, S. 841-850, Georg Thieme Verlag KG Stuttgart, New York.

Battmer, Rolf-Dieter (2009): „25 Jahre Cochlear-Implantat in Deutschland– eine Erfolgsgeschichte mit Perspektiven: Indikationserweiterung, Reliabilität der Systeme“ In: *Cochlear Implantat heute*, S.1-9.

Békésy, Georg V. (1944): „Über die Frequenzauflösung in der menschlichen Schnecke" In: *Acta Oto-Laryngologica*, 32(1), S. 60-84.

Büchner, Andreas und L. Gärtner, (2017): „Technische Entwicklung bei Cochleaimplantaten - Stand der Technik“ In: *HNO 2017*, 65, S.276-289, Springer Medizin Verlag GmbH.

Burgers, Christopher J.C. (1998): „A Tutorial on Support Vector Machines for Pattern Recognition” In: *Data Mining and Knowledge Discovery 2*, S.121-167.

Busch, Tobias; Vanpoucke, Filiep, und Astrid van Wieringena (2017): „Auditory Environment Across the Life Span of Cochlear Implant Users Insights From Data Logging" In: *Journal of Speech*, Language and Hearing Research, 60, S. 1362–1377.

Buyens, Wim; Van Dijk, Bas; Moonen Marc und Jan Wouters (2017): „Evaluation of a stereo music pre-proocessing scheme for cochlear implant users" In: *Journal of the American Academy of Audiology*, S.1–9.

Davis S.B. und P. Mermelstein (1980): „Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences.” In: *IEEE Trans. ASSP,* 28, S.357-366.

De Ceulaer, Geert; Swinnen, Freya; Pascoal, David; Philip, Birgit et al. (2014): „Conversion of adult Nucleus 5 cochlear implant users to the Nucleus 6 system" In: *Cochlear Implants International*, S.1-11.

Dölle, Mirko (2018): „Ohrenspitzer Hörgeräte und Implantat - Soundprozessoren mit Bluetooth" In: *c´t Magazin für Computertechnik*, 22.

Euler, Manfred (2004): „Kreative Uhren in unserem Kopf” In: *Biologie unserer Zeit*, 34(4) S.249-255,Wiley-VCH Verlag, Weinheim.

Gilden, Jan; Lewis, Kristen; Grant, Ginger und Jillian Crosson (2016): „Improved hearing in noise using new signal processing algorithms with the Cochlear™ Nucleus® 6 sound processor" In: *Journal of Otology 10,* S. 51-56.

Fastl, Hugo und Eberhard Zwicker (2007): *Psychoacoustics Facts and Models*, Springer Verlag Heidelberg.

Heinert, Michael (2010): „Support Vector Machines Teil 1: Ein theoretischer Überblick“ In: *Zeitschrift für Geodäsie,* 135, S. 1-11.

Hessel, Horst (2014): „Musikhören mit dem Cochlea-Implantat – Grenzen und Möglichkeiten.“ In: *Cochlear Limited.*

Houtsma, Adrianus J.M. (1995) „Pitch Perception”, In: *Hearing - Handbook of Perception and Cognition*, CA: Academic Press, S. 267-295.

Hughes, Michelle (2013): *Objective Measures in Cochlear Implants Core Clinical Concepts in Audiology*, Plural Publishing.

Janssen, Thomas (2014): *Hören – Audiologische Diagnostik*, Technische Universität München.

Kohlberg, Gavriel D.; Mancuso, Dean M; Chari, Divya A. und Anil K. Lalwani (2015): „Music Engineering as a Novel Strategy for Enhancing Music Enjoyment in the Cochlear Implant Recipient" In: *Hindawi Publishing Corporation, Behavioural Neurology*, S.1-7

Kong, Ying-Yee; Cruz, Rachel; Jones, Ackland J. and Fan-Gang Zeng (2004): „Music Perception with Temporal Cues in Acoustic and Electric Hearing" In: *Ear & Hearing, Lippincott Williams & Wilkins*, S.1-10

Leal, Mariana C.; Shin, Young Je; Laborde, Marie-Laurence, Lugardon Stephanie; Andrieu Sandrine und Bernard Fraysse et al. (2003): „Music perception in adult cochlear implant recipients" In: *Acta Otolaryngol*, 123, S.1 -10.

Lenarz, Thomas; Roland Laszig (2012): *Cochlea-Implantat Versorgung und zentral-auditorische Implantate Leitlinie der Deutschen Gesellschaft für Hals-Nasen-Ohren-Heilkunde, Kopf- und Hals-Chirurgie e.V., Bonn*, Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF).

Lerch, Alexander (2012): *Audio Content Analysis*, John Wiley & Sons, Inc.

Logan, Beth (2000), „Mel frequency cepstral coefficients for music modeling” In: *Proc. Int. Symp. Music Information Retrieval (ISMIR).*

Looi, Valerie (2008): „The effect of cohlear implantation on music perception" In: *Otorinolaringol,* 58, S.169-190.

Looi, Valerie; Wong, Yuhan and Jenny H. Y. Loo (2016):„The Effects of Training on Music Perception and Appreciation for Cochlear Implant Recipients" *Hindawi Publishing Corporation, Advances in Otolaryngology*, S.1-12 .

Mauger, Stefan J.; Warren, Chris D.; Knight, Michelle R.; Goorevich, Michael und Esti Nel (2014): „Clinical evaluation of the Nucleus® 6 cochlear implant system: Performance improvements with SmartSound iQ”, In: *International Journal of Audiology*, 53, S. 564-576.

McDermott, Hugh J. (2004): „Music Perception with Cochlear Implants: A Review" In: *Trends in Amplification*, 8 (2), S.59-82

Migirov, Lela; Kronenberg, Jona, und Yael Henkin(2009):„Self-Reported Listening Habits and Enjoyment of Music Among Adult Cochlear Implant Recipients" In: *Annals of Otology.* Rliinology & Uiryngnlcay 118(5):35-355.

Möller, Sebastian (2016): *Sprachkommunikation –Sprachsignalverarbeitung und Sprachtechnologie*–, Technische Universität Berlin.

Möser, Michael (2012): *Technische Akustik*, Springer Verlag Heidelberg.

Mrowinski, Dieter und Günther Scholz (2006): *Audiometrie: Eine Anleitung für die praktische Hörprüfung*, 3. Auflage, Georg Thieme Verlag Stuttgart – New York.

Mühler, Roland und Michael Ziese (2010): „Technischer Leitfaden Cochlea Implantat" In: *Universtitätsklinikum Magdeburg*, S.1-26

Müller-Deile, J.; Kortmann, T.; Hoppe, H.; Hessel, H. und A. Morsnowski (2009): „Verbesserung der Sprachverständlichkeit durch neuen Cochlear-Implant-Sprachprozessor" In: *HNO Springer Medizin Verlag*, 57, S. 567–574.

Oppenheim, Alan V.; Ronald Schafer W. und John R. Buck (2004): *Zeitdiskrete Signalverarbeitung*, 2. Auflage, Pearson Studium.

Petersen, Björn; Mortensen, Malene Vejby; Hansen, Mads und Peter Vuust (2012): „Singing in the Key of Life: A Study on Effects of Musical Ear Training After Cochlear Implantation" In: *Psychomusicology: Music, Mind, and Brain American Psychological Association*, 22 (2), S. 134–151.

Scheirer, E. und M. Slaney (1997): „Construction and evaluation of a robust multifeature speech/music discriminator” In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* S. 1331–1334.

Snyder, David; Guoguo Chen und Daniel Povey (2015): „MUSAN: A Music, Speech, and Noise Corpus” In: *The Johns Hopkins University*, Baltimore.

Sturm, Bob (2013): „The GTZAN dataset Its contents, its faults, their effects on evaluation, and its future“ In: *Journal of Intelligent Information System*, 41, S. 371-406.

Sturm, Bob (2014): „The state of the art ten years after a state of the art: Future research in music information retrieval" In*: J. New Music Research*, 43 (2), S.147–172.

Tzanetakis, George und Perry Cook (2002): „Musical Genre Classification of Audio Signals“ In: *IEEE Transactions on Speech and Audio Processing,* 10 (5), S.293-302.

Ulrich, Jens und Eckhard Hoffmann (2007): *Hörakustik Theorie und Praxis*, Median Verlag.

Weinzierl, Stefan (2008): *Handbuch der Audiotechnik*, Springer Verlag Heidelberg.

Wilson, Blake S.; Lawson, Dewey T.; Müller, Joachim M; Tyler, Richard S.; und Jan Kiefer (2003): „Cochlear Implants: Some Likely Next Steps", In: *Annual Review Biomed. Eng*. , 5:207–49

Internetquellen

Cochlear Deutschland GmbH & Co. KG (2918), Cochlea-Implantate https://www.cochlear.com/de/startseite/hoeren-und-hoerverlust/hoeren-und-hoerverlust/behandlungsmoeglichkeiten/cochlea-implantate, Zugriff 06.09.2018

Sengpielaudio: http://www.sengpielaudio.com/Rechner-harmonische.htm, abgerufen am: 10.9.2018

9. Anhang

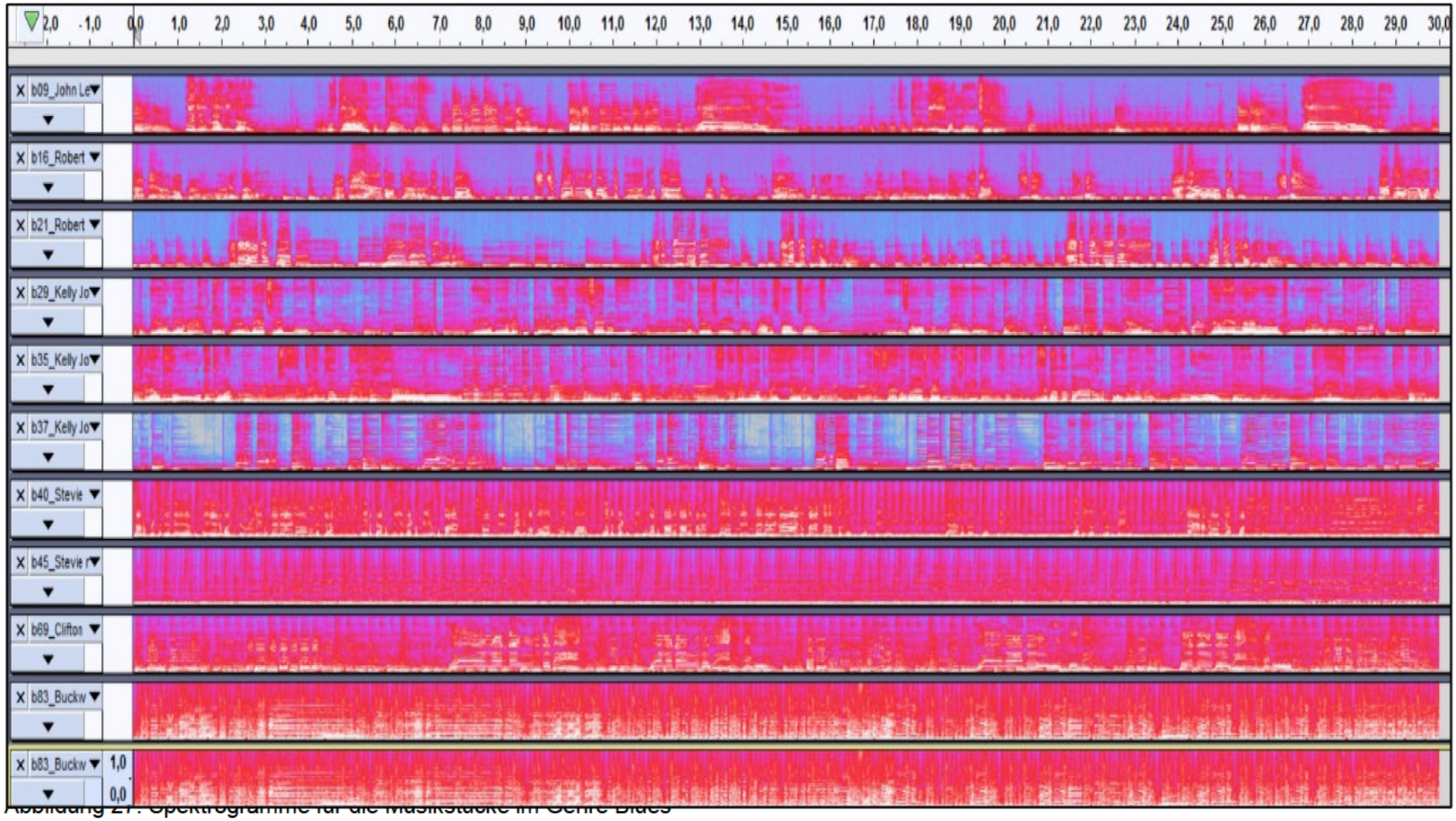

Abbildung 27: Spektrogramme für die Musikstücke im Genre Blues

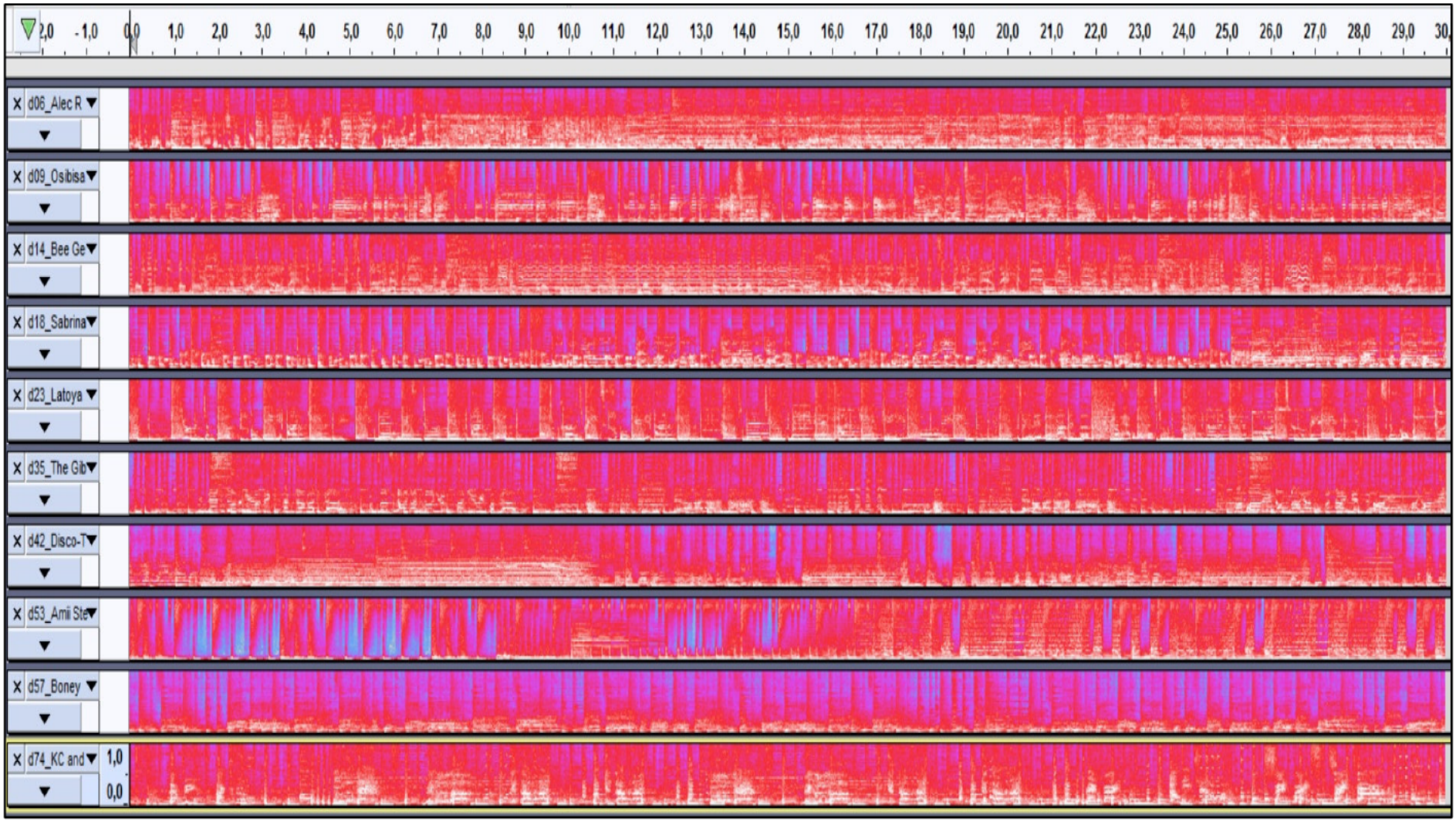

Abbildung 28: Spektrogramme für die Musikstücke im Genre Disco

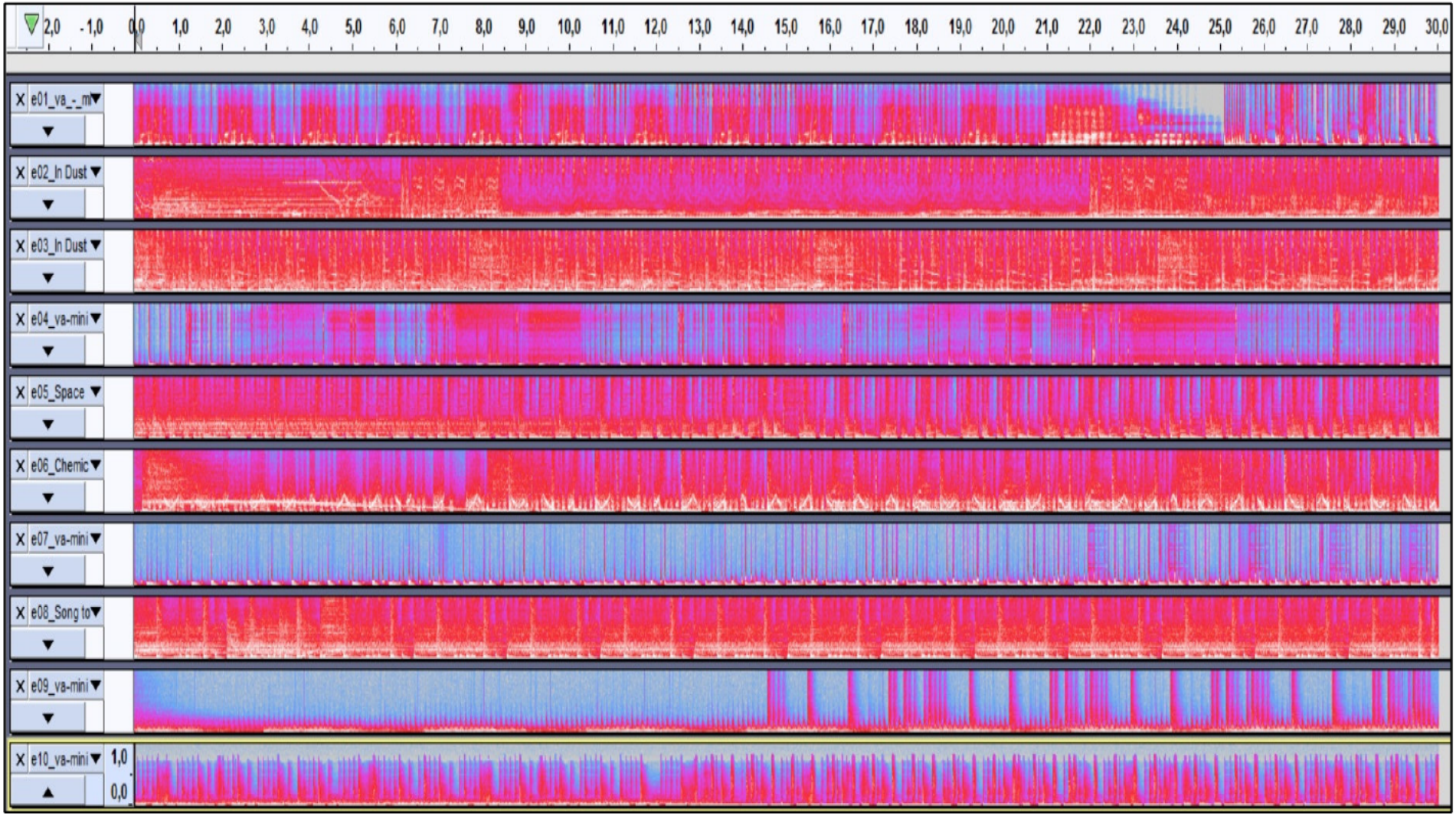

Abbildung 29: Spektrogramme für die Musikstücke im Genre Electo

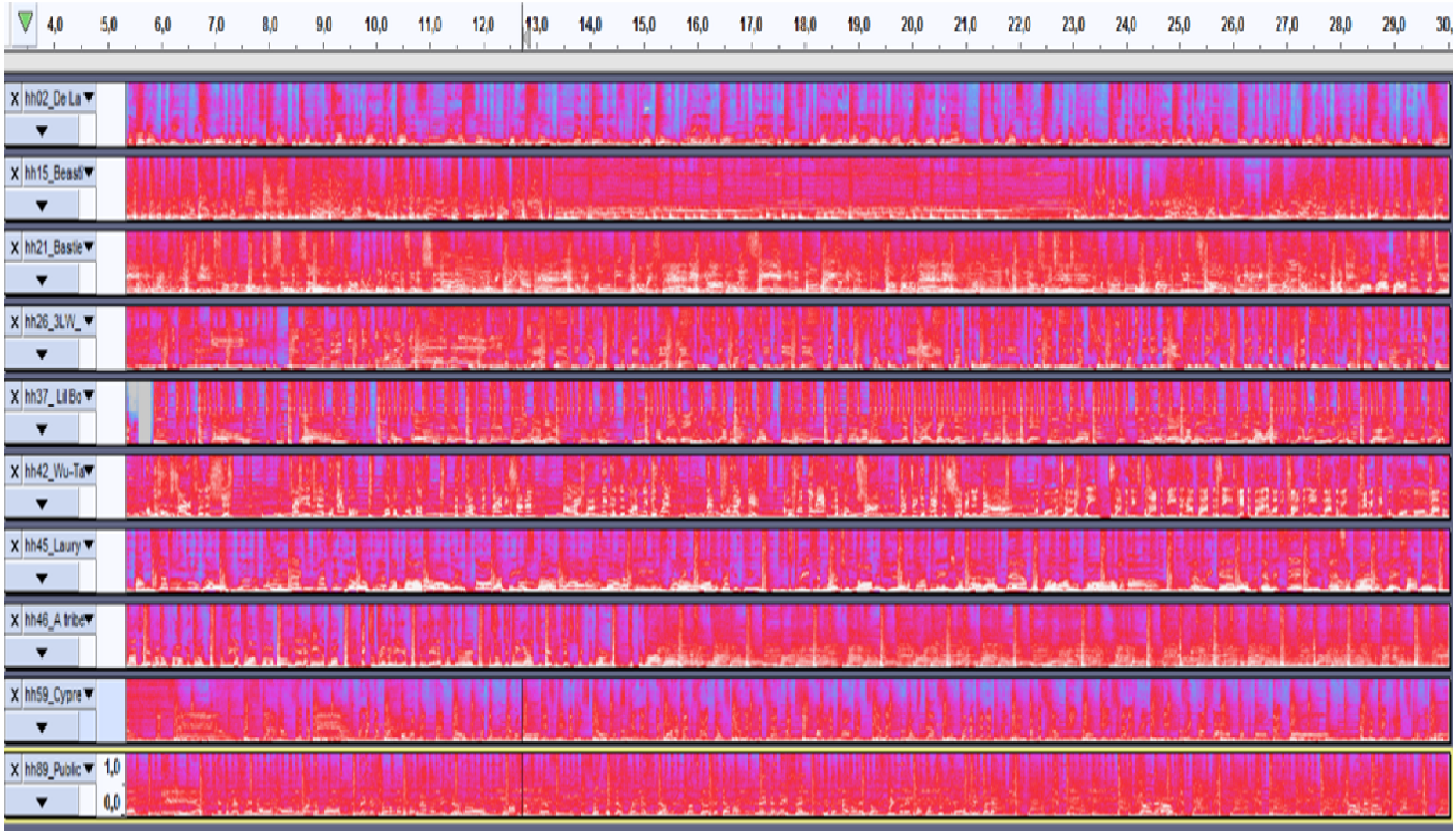

Abbildung 30: Spektrogramme für die Musikstücke im Genre Hip Hop

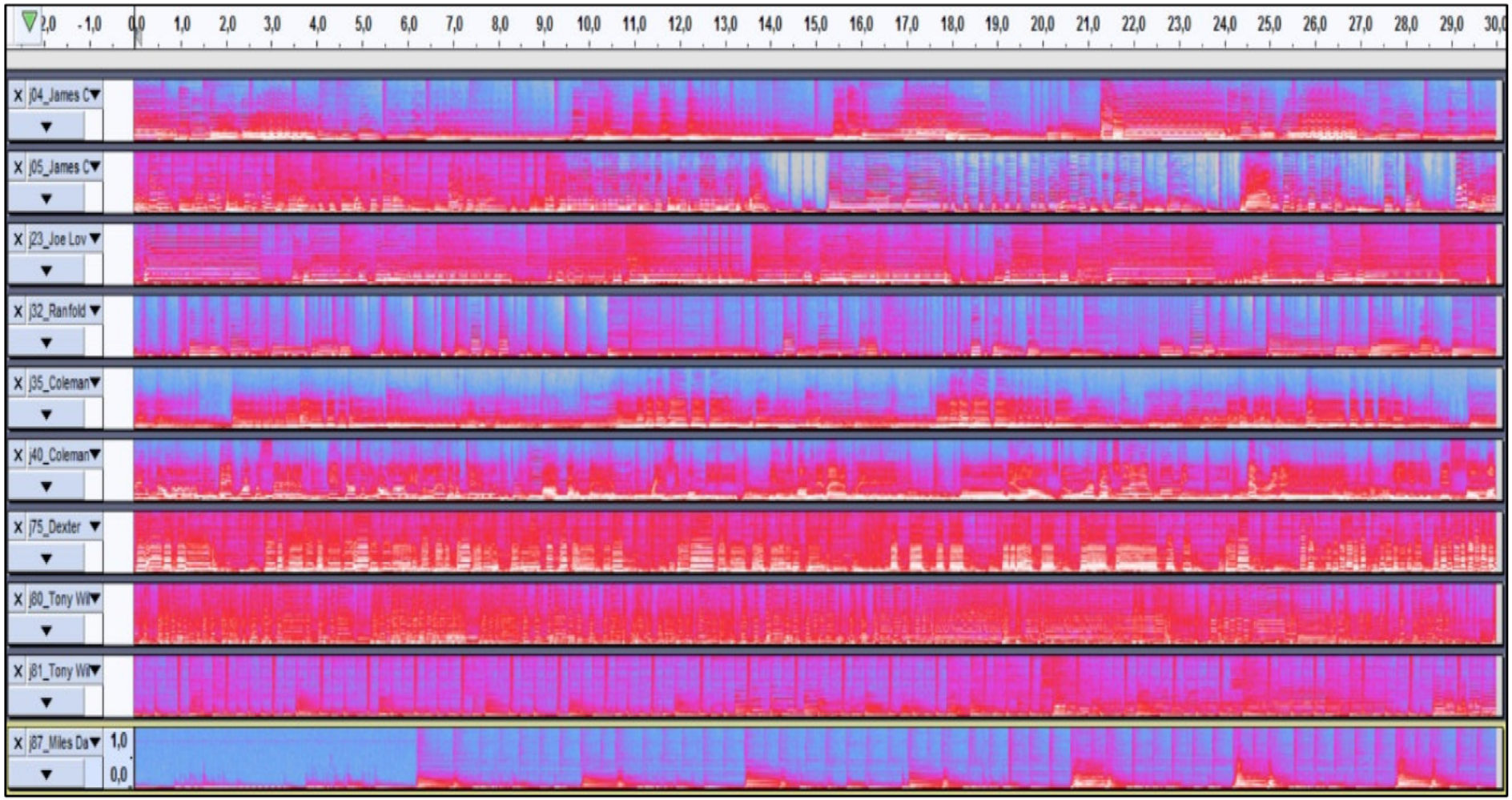

Abbildung 31: Spektrogramme für die Musikstücke im Genre Jazz

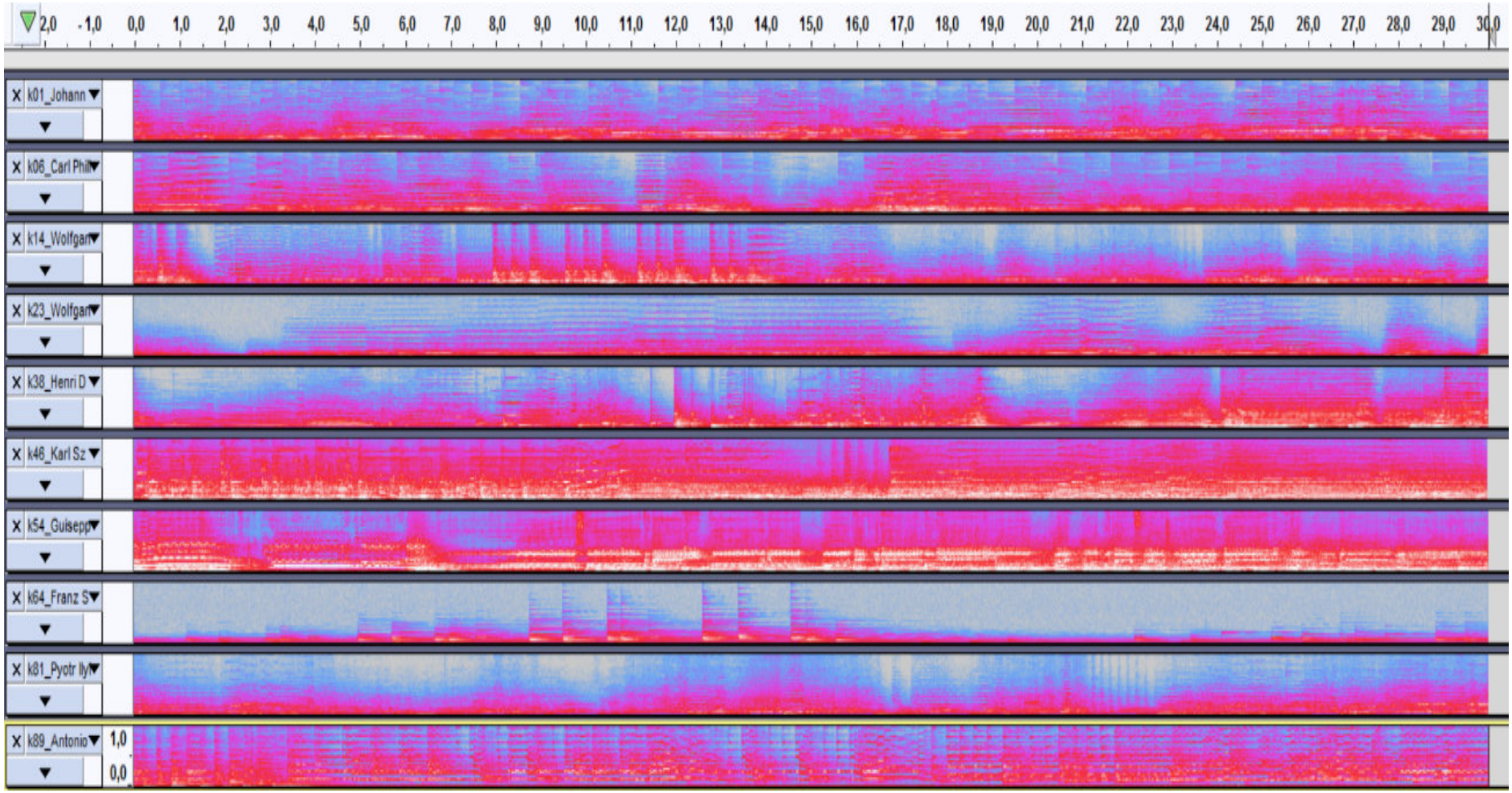

Abbildung 32: Spektrogramme für die Musikstücke im Genre Klassik

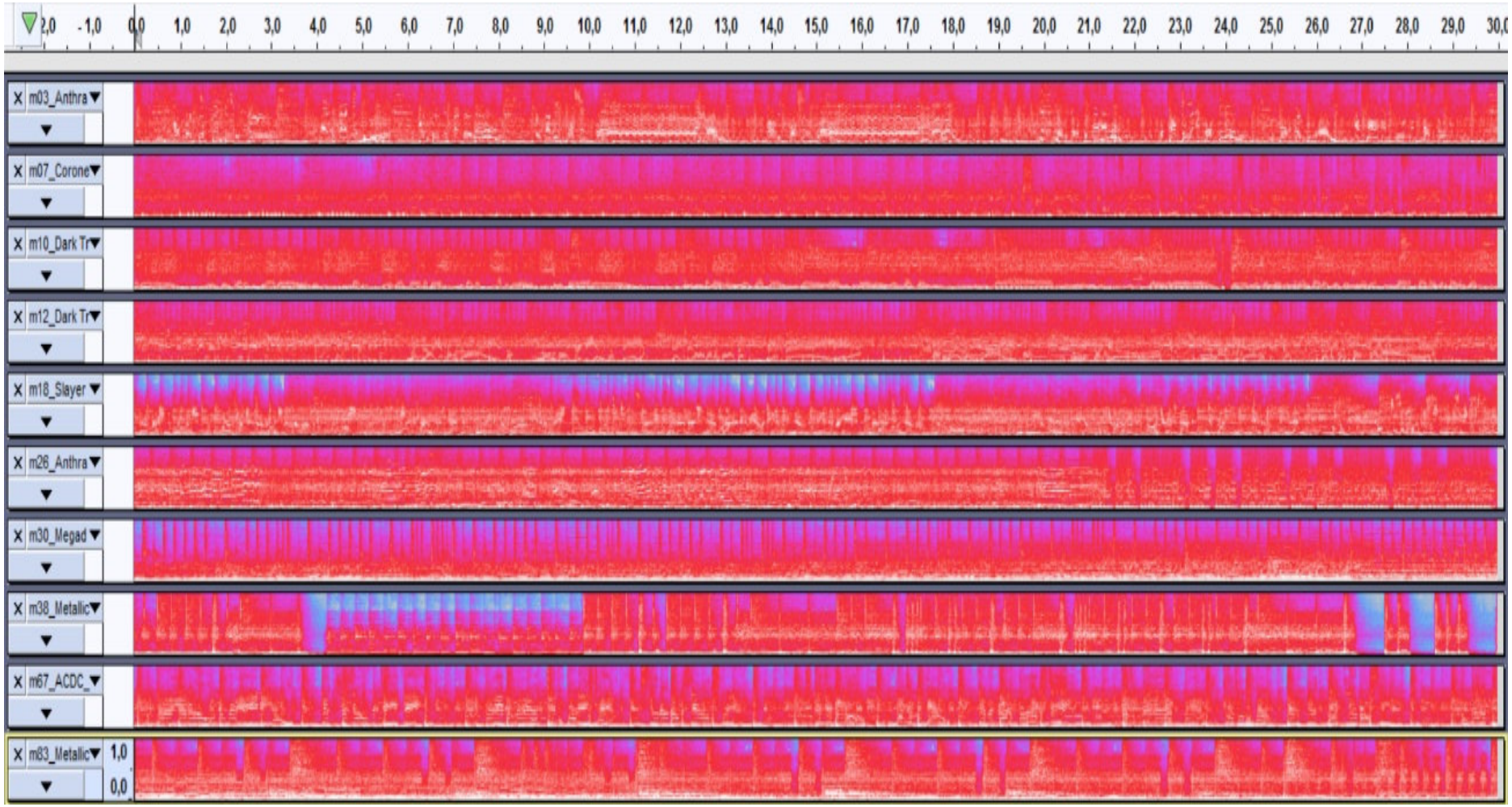

Abbildung 33: Spektrogramme für die Musikstücke im Gerne Metal

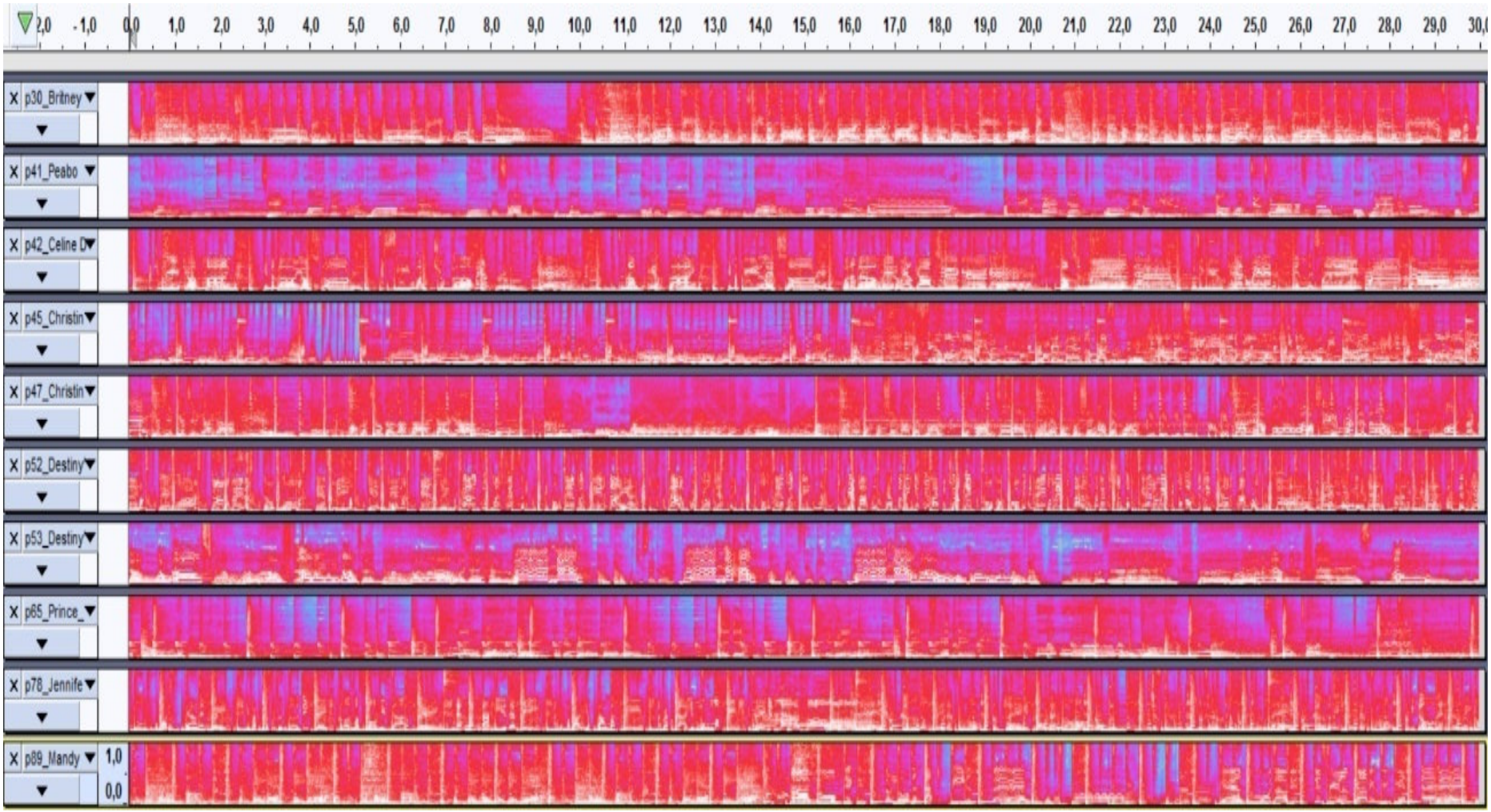

Abbildung 34: Spektrogramme für die Musikstücke im Genre Pop

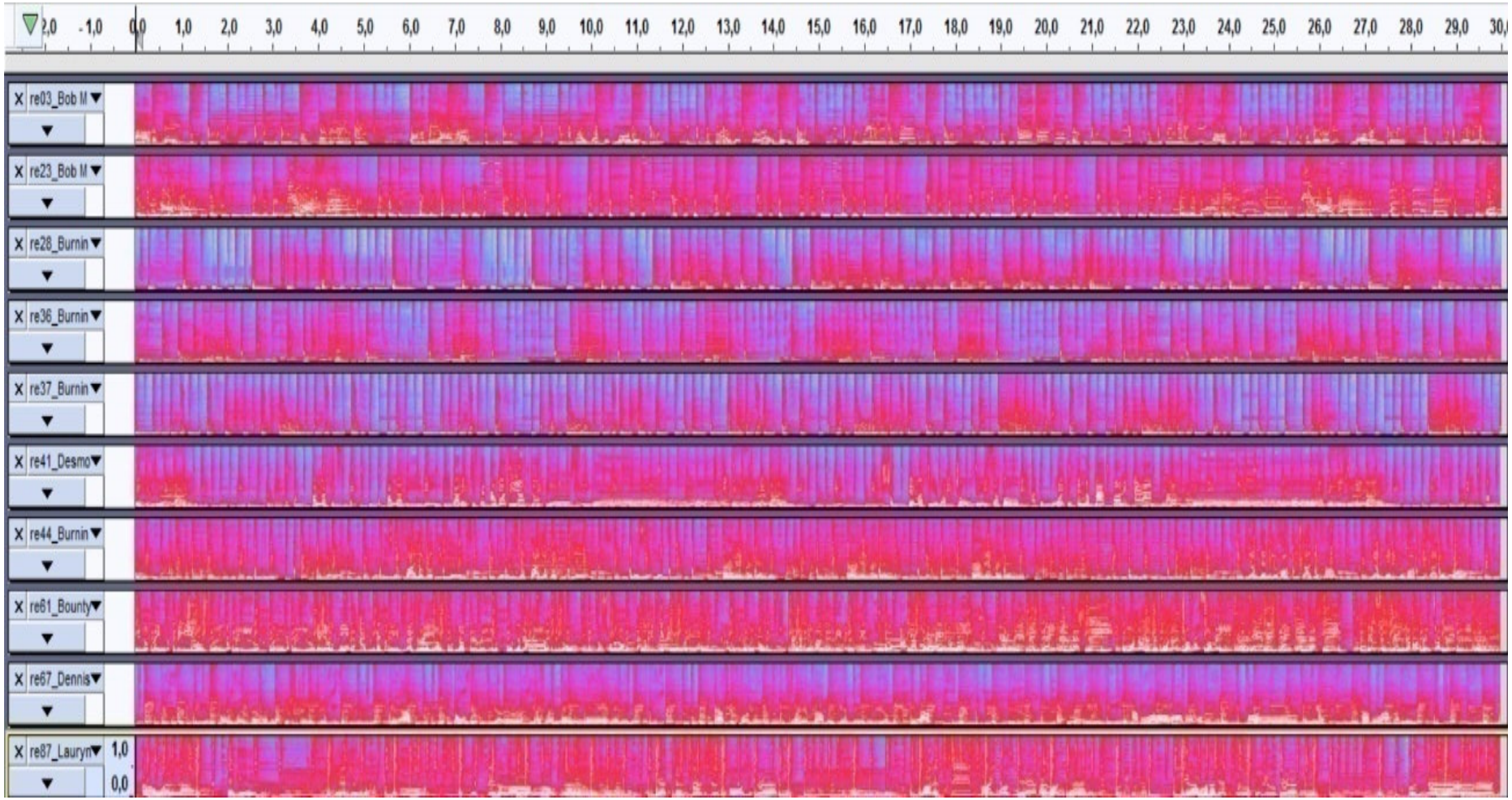

Abbildung 35: Spektrogramme für die Musikstücke im Genre Reggae

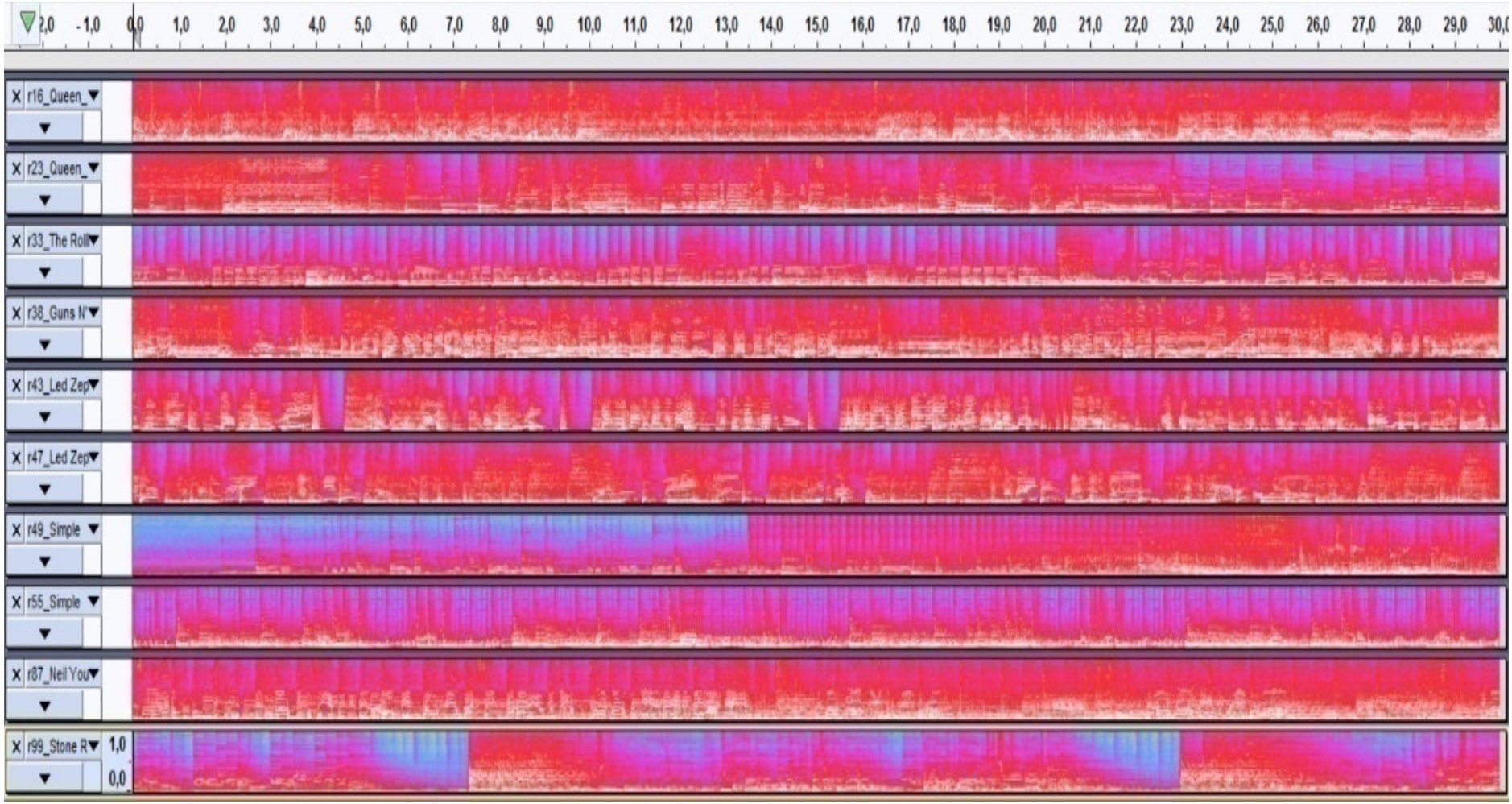

Abbildung 36: Spektrogramme für die Musikstücke im Genre Rock