Kompression der Erregungsmuster von Cochlea-Implantaten

Von der Fakultät für Elektrotechnik und Informatik

der Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des akademischen Grades

Doktor-Ingenieur

(abgekürzt: Dr.-Ing.)

angenommene

Dissertation

von

Reemt Hinrichs, M. Sc.

geboren am 14. Oktober 1986 in Hannover

2023

ABSTRACT

Cochlear implants are surgically implanted hearing aids, which can improve the sense of hearing in some people suffering from hearing loss. Its main components are a microphone, a signal processor as well as excitation electronics for the stimulation of the auditory nerve. Wireless transmission of audio signals from external devices like smartphones to the signal processor has become possible in the past years. There, minimal bitrate for the representation of the audio signals to be transmitted is desirable. These audio signals can be represented by the stimulation patterns, the current values derived from the audio signals by the cochlear implant.

This work investigates ways to compress the stimulation patterns of cochlear implants with the aim of reducing the bitrate at minimal algorithmic latency in the context of wireless transmission.

The submitted thesis covers five contributions on this topic: the first two contributions are concerned with the development and evaluation of a lossy compression algorithm for the stimulation patterns of cochlear implants, the Electrocodec. This codec is evaluated in cochlear implant users and compared to a state-of-the-art audio codec with respect to speech intelligibility and audio quality. The Electrocodec proves superior with regard to bitrate and algorithmic latency compared to the Opus audio codec. The Electrocodec achieves undegraded speech intelligibility and quality compared to the original stimulation patterns at a bitrate of 24.3 kbit/s and an algorithmic latency of 0 ms. The third contribution investigates lossless compression of the stimulation patterns using artificial neural networks. While this lossless compression achieves about 4-8 kbit/s higher bitrate than the Electrocodec, it still yields lower bitrates than Opus at lower latency. The fourth and fifth contribution are concerned with autoencoders for the lossy compression of the stimulation patterns. Using hyperparameter optimization as well as numerical methods for the approximation of gradients, equal or superior speech intelligibility compared to the Electrocodec is achieved, however, using an approximately 80 % reduced bitrate of only 4.67 kbit/s at equal algorithmic latency.

Keywords: cochlear implant, stimulation patterns, data compression

KURZFASSUNG

Cochlea-Implantate sind implantierte Hörgeräte, die das Hörvermögen mancher hörgeschädigten Menschen verbessern können. Hauptkomponenten sind ein Mikrophon, ein Signalprozessor sowie Anregungselektronik zur Stimulation des Hörnervs. Seit einigen Jahren ist die drahtlose Übertragung von Audiosignalen von externen Geräten wie Smartphones zum Signalprozessor möglich. Hierbei ist eine möglichst geringe Bitrate zur Repräsentation der zu übertragenen Audiosignale wünschenswert. Diese Audiosignale können durch die vom Cochlea-Implantat daraus abgeleiteten Erregungsmuster, Folgen von Stromwerten, repräsentiert werden. In dieser Arbeit wird die Kompression von Erregungsmustern von Cochlea-Implantaten betrachtet zwecks Reduktion der Bitrate bei minimaler algorithmischer Latenz im Kontext der drahtlosen Übertragung von Audiosignalen. Die vorgelegte Arbeit umfasst fünf Beiträge zu diesem Thema: In den ersten beiden Beiträgen wird ein verlustbehafteter Kompressionsalgorithmus für die Erregungsmuster von Cochlea-Implantaten entworfen und untersucht, der Electrocodec. Dieser wird in Hörtests mit Cochlea-Implantatträgern erprobt und mit einem aktuellen Audiocodec hinsichtlich der Sprachverständlichkeit und Audioqualität verglichen. Es zeigt sich eine Überlegenheit des Electrocodecs hinsichtlich Bitrate und algorithmischer Latenz im Vergleich zum Opus-Audiocodec. Der Electrocodec erzielt unveränderte Sprachverständlichkeit und Qualität im Vergleich zu den unkomprimierten Erregungsmustern bei einer Bitrate von 24,3 kbit/s sowie einer algorithmischen Latenz von 0 ms. Im dritten Beitrag wird eine verlustlose Kompression der Erregungsmuster auf Basis künstlicher neuronaler Netze untersucht. Diese erzielt eine um etwa 4-8 kbit/s höhere Bitrate als der Electrocodec, lieferte jedoch immer noch niedrigere Bitraten als Opus bei niedriger Latenz. Im vierten und fünften Beitrag werden Autoencoder zur verlustbehafteten Kompression der Erregungsmuster untersucht. Mittels Hyperparameteroptimierung sowie numerischer Methoden zur Approximation von Gradienten wird eine dem Electrocodec ebenbürtige oder überlegene Sprachverständlichkeit erzielt, jedoch mit einer bis zu etwa 80 % reduzierten Bitrate von nur noch $4,67\,\mathrm{kbit/s}$ bei gleicher algorithmischer Latenz.

Schlüsselbegriffe: Cochlea-Implantat, Erregungsmuster, Datenkompression

AKRONYME

ACE Advanced Combination Encoder.
AEC Autoencoder (ohne Rückkopplung).
AR(1) Autoregressiver Prozess der Ordnung 1.

CCITT Consultatif International Téléphonique et Télégraphique.
CI Cochlea-Implantat.
CSR Channel Stimulation Rate.

DFT Diskrete Fouriertransformation.
DPCM Differential Puls-Code Modulation.

EC Electrocodec.

FRAE Rückkopplungsautoencoder.

HSM Hochmair-Schulz-Moser-Satztest.

LGF Lautheitswachstumsfunktion.

NaN Not a Number.
NWKR Nadaraya-Watson Kernel Regressor.

PG Prädiktionsgewinn.

SDR Signal-Verzerrungs-Verhältnis.
SMAC Sequential Model-Based Algorithm Configuration.
SNR Signal-Rausch-Verhältnis.
SPSA Stochastic Perturbation Simultaneous Approximation.
SQAM Sound Quality Assessment Material.
STOI Short-Time Objective Intelligibility Measure.

VSTOI Vocoder STOI.

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

1

EINLEITUNG

Im Zentrum dieser Arbeit stehen Datenkompressionsalgorithmen für die Erregungsmuster von Cochlea-Implantaten (CI), spezielle Hörhilfen oder auch Reizprothesen, welche durch künstliche Stimulation des Hörnervs das Sprachverstehen von Menschen mit speziellen Hörschädigungen wiederherstellen oder verbessern können. In diesem Kapitel wird erläutert, inwiefern Cochlea-Implantate relevant sind und die Kompression von Erregungsmustern ein lohnenswertes Forschungsthema ist, gefolgt von der konkreten Problemstellung der vorgelegten Arbeit. Anschließend wird der Aufbau der Arbeit vorgestellt.

1.1 MOTIVATION

Der Hörsinn ist einer der fünf Sinne des Menschen, durch welchen es ihm möglich ist, seine Umwelt wahrzunehmen. Er gestattet einem Menschen insbesondere das Verstehen von gesprochener Sprache und ist damit von fundamentaler Bedeutung für das Leben in einer Gemeinschaft [Fri14]. Daher ist das Funktionieren des Hörsinns von eminenter Bedeutsamkeit und eine Beeinträchtigung oder gar Verlust ebendieses hat oftmals deutliche psychosoziale Folgen [HJH06].

Das Gehör des Menschen ist aus Sicht der Signalverarbeitung ein System, welches mechanische Schwingungen, sogenannte Schallwellen, in elektrische Signale umwandelt, die dann vom Gehirn verarbeitet und vom Menschen interpretiert werden können.

Bei gesunden Menschen liegt der wahrnehmbare, also hörbare, Frequenzbereich zwischen etwa 20 Hz bis zu etwa 20 kHz [PAF01], wobei letztere Grenze praktisch nur von jungen Menschen erreicht wird. Im Laufe des Lebens reduziert sich normalerweise insbesondere die obere Grenze des hörbaren Frequenzbereichs und auch die Empfindlichkeit in verschiedenen Frequenzbereichen [LY07]. Dies kann dazu führen, dass etwa Sprache in höherem Alter schwerer verstanden wird. In schwereren Fällen spricht man von einer Hörschädigung.

Nach Schätzungen der Vereinten Nationen beläuft sich die Zahl der als hörgeschädigt geltenden Menschen weltweit auf bis zu 20 % der Weltbevölkerung [CC17]. Da insbesondere ältere Menschen betroffen sind, lebt gerade in westlichen Ländern mit zunehmend älterer Bevölkerung eine große, wachsende Zahl an hörgeschädigten Menschen.

Neben Alterungserscheinungen treten Hörschädigungen des Weiteren als Folge von Unfällen und angeborenen sowie erworbenen Krankheiten auf [Egg17].

Um Betroffenen von Hörschädigungen das Leben zu erleichtern bzw. ihre Lebensqualität zu erhöhen, werden in verschiedensten Bereichen [MB15; Gla22] der Wissenschaft Methoden untersucht, Menschen den Hörsinn wiederzugeben oder diesen zu verbessern. Minimalziel ist im Allgemeinen eine Verbesserung oder vollständige Wiederherstellung des Sprachverstehens des Menschen in typischen - insbesondere sozialen - Situationen.

Im Bereich der Ingenieurswissenschaft kann man nichtinvasive und invasive Hörhilfen unterscheiden [SG22]. Erstere kommen ohne operativen Eingriff und Implantierung von elektrischen oder ähnlichen Gerätschaften aus, letztere nicht. Bekannte nichtinvasive Hörhilfen sind etwa das klassische Hörrohr, welches durch künstliche Erweiterung der Ohrmuschel das Hören unterstützt und vor Einzug der Elektrizität in den Alltag des Menschen Anwendung fand, oder, heutzutage eine verbreitete Variante, das konventionelle Hörgerät, welches im Kern aus einer Verstärkereinheit und einem kleinen Lautsprecher besteht, der in den Gehörgang eines Menschen von außen eingeführt wird.

Ansätze wie diese, welche nicht in die Umwandlung mechanischer Schwingungen in elektrische Signale eingreifen, kommen an ihre Grenzen, sobald eben dieser Mechanismus gestört ist. Ganz wesentlich am Hören beteiligt ist die Hörschnecke im Innenohr des Menschen, die sogenannte Cochlea. In diesem Organ kommt es zur Umwandlung von mechanischen Schwingungen in elektrische Impulse [Man+17]. Hierbei sind maßgeblich die sogenannten Haarzellen beteiligt, welche sehr zahlreich über die Oberfläche der Cochlea verteilt sind. Ihre (mechanische) Schwingung ist essentiell für die Umwandlung von akustischen, also mechanischen, Signalen in elektrische Signale, den Nervenimpulsen [Hud97].

Sind die Haarzellen beschädigt, so hat dies im Allgemeinen eine negative Konsequenz für den Hörsinn [WS19]. Bei massiver Schädigung der Haarzellen kommt es zu einer starken Beeinträchtigung des Hörens bis hin zur vollständigen Taubheit, zumindest auf dem betroffenen Ohr. Bei einer solchen Schädigung, bei welcher insbesondere der Hörnerv intakt bleibt, ist es nichtinvasiven Hörhilfen nicht möglich, den Hörsinn wiederherzustellen oder zu verbessern. Grund ist die gestörte Umwand-

lung mechanischer Signale in elektrische, was eine reine Verstärkung des mechanischen Signals nicht beheben kann.

Hier können invasive Hörhilfen eine Besserung bewirken. Ein bekannter, sehr erfolgreicher Vertreter dieser Klasse ist das sogenannte Cochlea-Implantat[1]. Bei diesem wird die Funktionalität der Haarzellen durch eine Kombination aus einem Mikrophon und Signalverarbeitung sowie in die Cochlea implantierten Elektroden (nebst Anregungselektronik) ersetzt, sodass das Hörverstehen teilweise oder vollständig wiederhergestellt werden kann. Die notwendige Signalverarbeitung wird hierbei durch den Signalprozessor des Cochlea-Implantats bewerkstelligt. Mit aktueller Technologie funktionieren Cochlea-Implantate typischerweise gut in Umgebungen mit geringem Hintergrundgeräuschpegel. Ihre Leistungsfähigkeit, gemessen durch das erzielte Sprachverstehen ihrer Träger, lässt jedoch in Umgebungen mit höherem Hintergrundgeräuschpegel, wie einem Restaurant, einer Feier, aber etwa auch dem Schulunterricht, stark nach [AM22]. Dieses Nachlassen des Sprachverstehens von Cochlea-Implantatträgern erfolgt dabei wesentlich schneller als das Nachlassen des Sprachverstehens von Normalhörenden, d.h. von Menschen mit intaktem Hörsinn.

Diese starke Reduktion des Hörverstehens in Gegenwart von Hintergrundrauschen fördert, ohne weitere Hilfsmittel, die soziale Isolation der Cochlea-Implantatträger, da das reduzierte Hörverstehen für sie soziale Interaktionen deutlich erschwert [NG03].

Ein wesentliches Ziel der gegenwärtigen Forschung ist es daher, insbesondere in diesen schwierigen Situationen mit höherem Hintergrundgeräuschpegel, das Sprachverstehen zu verbessern sowie auch die Lokalisierung von Sprechern im Raum zu verbessern, welche bei Cochlea-Implantatträgern ebenfalls eingeschränkt ist [Zei+15]. Ein weiterer Forschungspunkt befasst sich mit der Wahrnehmung von Musik, jedoch berücksichtigt die vorgelegte Arbeit lediglich das Sprachverstehen.

1.2 PROBLEMBESCHREIBUNG

Insbesondere das Sprachverstehen von Trägern von Cochlea-Implantaten ist durch Hintergrundrauschen, wie es in Restaurants oder anderen sozialen Situationen auftritt, stark beeinträchtigt [DG17].

Eine Vielzahl an Verfahren findet Anwendung, um insbesondere in diesen Situationen das Sprachverstehen zu verbessern. Diese Verfahren umfassen zum Beispiel Strahlenformung (engl. beamforming) [HGJ21; Sei+17], also die Schärfung der Richtcharakteristik des Mikrophons des

1 Genauer handelt es sich um eine sogenannte Reizprothese oder auch Hörprothese.

Cochlea-Implantats durch Kombination mehrerer Mikrophonsignale sowie externer Mikrophone (engl. *remote microphones*) [Mil+22], die durch Reduktion des Abstands zwischen Sprecher und Mikrophon das Signal eines Sprechers stärker auffängt und dadurch das Signal-Rausch-Verhältnis verbessert. Weitere Ansätze sind das kontralaterale Routing von Signalen [SBF22], bei dem Audiosignale von einem Ohr ohne Hörhilfe zum anderen mit Hörhilfe übertragen werden, sowie binaurale Signalverarbeitungsstrategien [GN18; Fum+21; GN22]. Letztere kombinieren auf geschickte Weise akustische Informationen beider Ohren, typischerweise bei Vorliegen eines Cochlea-Implantats in beiden Ohren, um das Sprachverstehen oder die Lokalisierungsfähigkeit zu verbessern.

Die genannten Verfahren haben gemein, dass eine drahtlose Übertragung von Audiosignalen notwendig oder gewünscht ist.

Neben den zuvor genannten Ansätzen existieren weitere Technologien, welche mittels drahtloser Übertragung von Audiosignalen das Leben von Cochlea-Implantatträgern zu verbessern suchen. Wesentlich sind hierbei insbesondere das drahtlose Streamen von Telefonanrufen [Hut+22c] direkt zum Signalprozessor des Cochlea-Implantats sowie das drahtlose Streamen des Audiosignals eines Fernsehers, welches das Fernsehen in Gesellschaft für Cochlea-Implantatträger erleichert [DWS16].

Bei allen in diesem Abschnitt angeschnittenen Ansätzen ist es vorteilhaft, die zu übertragene Datenmenge zu minimieren, d.h. die Zahl an zu übertragenden Bits pro Sekunde, die Bitrate, möglichst klein zu halten, aber gleichzeitig die Verständlichkeit und die Qualität der komprimierten Audiosignale möglichst wenig zu beeinträchtigen. Diese Bitratenreduktion reduziert zum einen im Allgemeinen die für die Übertragung der Audioinformation notwendige Energie [Bau17] und zum anderen wird hierdurch die Kapazität des verwendeten drahtlosen Kanals minimal belastet.

Aus diesen Gründen ist eine möglichst starke Datenkompression, also Reduktion der Bitrate, der zu übertragenden Audiosignale anzustreben.

1.3 IDEE UND ZIELSETZUNG

Kaum etwas ist über die von den Herstellern verwendeten Algorithmen zur Reduktion der Bitrate der Audiosignale in diesem Kontext bekannt. Oticon, einer der vier großen Hersteller, verwendet für kontralaterales Routing etwa den bekannten G.722 Audiocodec [Oti21]. Zwar ist bekannt, dass das Bluetooth-Protokoll z.B. für das Streamen von Telefonanrufen vielfach verwendet wird, jedoch ist keine Information über die genutzten Kompressionsalgorithmen öffentlich zugänglich. Eine Nutzung von Standardbluetooth-Audiocodecs wie dem Low Complexty Subband

Codec (SBC) oder dem Low Complexity Communications Codec (LC3) liegt nahe, lässt sich aber nicht belegen. Sicher kann man jedoch davon ausgehen, dass unmittelbar das Audiosignal komprimiert wird und keine auf Hörgeschädigte optimierte Audiokompression genutzt wird.

Jedoch ist es zur Optimierung von Kompressionsalgorithmen unabdingbar, den Wahrnehmer der dekomprimierten Signale zu berücksichtigen. Viele Verzerrungen, die etwa bei der Kompression von Audiosignalen entstehen, können von einem Menschen nicht wahrgenommen werden, sind also für diesen irrelevant. Es zeigt sich, dass es vorteilhaft ist, irrelevante Signalanteile zu identifizieren und diese gar nicht oder mit schlechter Qualität zu komprimieren [HD19b]. Dadurch kann im Allgemeinen die erzielbare Bitrate bei vorgegebener Qualität massiv reduziert werden.

Die Idee der vorgelegten Arbeit ist die Übertragung dieser Tatsache auf das Problem der Kompression von Audiosignalen (zwecks drahtloser Übertragung) im Kontext von Cochlea-Implantaten.

Die Idee ist dabei, zur Übertragung der oben genannten Audiosignale nicht das unmittelbare Audiosignal zu komprimieren, sondern dieses zunächst durch den Signalprozesser des Cochlea-Implantats zu verarbeiten. Hierdurch entstehen die sogenannten Erregungsmuster, eine Abfolge von Stromimpulsen, welche das akustische Signal für den Cochlea-Implantatträger codieren. In den Erregungsmustern ist ein Teil der Information des Eingangsaudiosignals nicht enthalten, welcher von Trägern von Cochlea-Implantaten nicht wahrgenommen werden kann. Dadurch ist es plausibel, dass eine Kompression der Erregungsmuster, bei gleicher subjektiver Qualität und Verständlichkeit, eine geringere Bitrate erzielen kann als eine Kompression der Audiosignale mit konventionellen Audiocodecs.

In der vorgelegten Arbeit wurden Kompressionsverfahren für diese Erregungsmuster des Cochlea-Implantats entwickelt und untersucht, welche für die drahtlose Übertragung von Audiosignalen an die Signalprozessoren von Cochlea-Implantaten verwendet werden können.

Hierbei wurde nicht nur die Bitrate möglichst stark gesenkt, sondern gleichzeitig auch die durch den Kompressionsalgorithmus induzierte algorithmische Latenz möglichst kleingehalten. Motivierend war, dass Träger von Cochlea-Implantaten sehr sensibel für relative zeitliche Verschiebungen zwischen gemeinsam auftretenden visuellen und akustischen Eindrücken sein können, wie sie etwa bei der Nutzung von externen Mikrophonen vorliegen. Ein Zeitversatz von 10 ms kann dabei bereits negativ bewertet werden [Eur13]. Um noch Raum für weitere Signalverarbeitung zu lassen, welche ihrerseits im Allgemeinen eine algorithmische Latenz größer Null aufweisen wird, wurde eine algo-

rithmische Latenz von unter 5 ms angestrebt. Tatsächlich wurden sogar ausschließlich Verfahren mit einer algorithmischen Latenz von 0 ms entwickelt. Dadurch können die entwickelten Kompressionsverfahren zu beliebigen Signalverarbeitungsketten ergänzt werden, ohne dass die algorithmische Latenz des Gesamtsystems erhöht wird. Ein weiterer Vorteil der Kompression der Erregungsmuster gegenüber der Kompression der korrespondierenden Audiosignale ist die mögliche Konvertierung der Audiosignale in Erregungsmuster auf einem externen Gerät vor der drahtlosen Übertragung. Hierdurch kann der Energieverbrauch des empfangenden Cochlea-Implantats reduziert werden, da eine Berechnung der Erregungsmuster entfällt.

Zusammengefasst ist das Ziel der Arbeit, die notwendige Bitrate für die Darstellung der zu komprimierenden Erregungsmuster möglichst kleinzuhalten und dabei gleichzeitig die assoziierte Verständlichkeit dieser Erregungsmusters möglichst wenig durch die Kompression zu beeinträchtigen. Des Weiteren wird eine möglichst geringe algorithmische Latenz angestrebt.

1.4 AUFBAU DER ARBEIT

Zunächst wird im Grundlagenkapitel die Funktionsweise des Gehörs des Menschen kurz erläutert, um anschließend den Aufbau und die Funktionsweise eines Cochlea-Implantats genauer zu betrachten. Hierbei wird insbesondere auf die Signalverarbeitung eingegangen, welche verwendet wird, um aus einem Audiosignal die Erregungsmuster abzuleiten. Des Weiteren wird auf typische Maße für das Hörverstehen aus dem Bereich der Forschung eingegangen. Anschließend werden die maßgeblichen angewendeten Verfahren wie Prädiktion, Quantisierung, neuronale Netze und andere erläutert und des Weiteren die verwendeten Methoden der Stochastik beschrieben, die unter anderem zur Auswertung der durchgeführten Hörtests Anwendung fanden.

Nachfolgend werden die im Rahmen der vorgelegten Arbeit entwickelten Kompressionsverfahren vorgestellt und die erzielten Ergebnisse präsentiert. Hierbei werden sowohl die Ergebnisse von durchgeführten Hörtests mit Cochlea-Implantatträgern als auch die Auswertung mit objektiven Maßen des Hörverstehens vorgestellt. Anschließend kommt es zur Untersuchung der Optimalität eines der Codierungsverfahrens.

Danach werden die Ergebnisse kritisch diskutiert und die entwickelten Verfahren schlussendlich miteinander verglichen. Den Abschluss der Arbeit bildet eine Zusammenfassung aller wesentlichen Ergebnissen.

2

GRUNDLAGEN

In diesem Kapitel werden die für die gesamte Arbeit relevanten Grundlagen erläutert, angefangen bei den wesentlichen Komponenten des menschlichen Gehörs über den Aufbau eines Cochlea-Implantats sowie der Leistungsfähigkeit aktueller Cochlea-Implantate hin zu den wesentlichen Elementen der Wahrscheinlichkeitstheorie, Statistik und Signalverarbeitung, die zum Verständnis der verwendeten Verfahren dieser Arbeit notwendig sind.

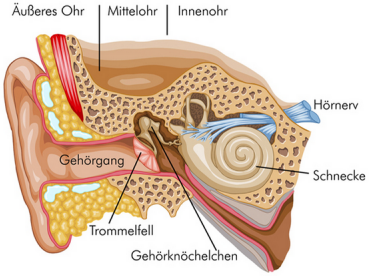
2.1 DAS GEHÖR DES MENSCHEN

Das Gehör des Menschen ist verantwortlich für die Umwandlung von Schallwellen als Träger von Information in elektrische Impulse, die vom menschlichen Gehirn verarbeitet werden können.

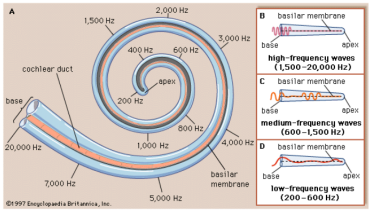
Makroskopisch besteht das Gehör des Menschen aus dem Außenohr, dem Mittelohr und dem Innenohr. Die Struktur ist in Abb. 2.1a dargestellt.

Das Außenohr besteht aus der Ohrmuschel und dem Gehörgang und dient dem Einfangen von Schallwellen sowie, aus der Perspektive der Signalverarbeitung, der Vorfilterung. Der Gehörgang hat eine Resonanzfrequenz von etwa 2700 Hz, und Frequenzen im Bereich von etwa 2 kHz bis etwa 8 kHz werden bis zu 15 dB verstärkt respektive die anderen Frequenzen gedämpft [Moo07].

Das Trommelfell separiert das Außenohr sowie das Mittelohr und dient mittelbar der Umwandlung der Luftschwingungen in Flüssigkeitsschwingungen in der Cochlea. Einfallende Schallwellen versetzen das Trommelfell in Vibration, wodurch wiederum die Gehörknochen, von welchen es drei Stück gibt, in Schwingungen versetzt werden. Diese Schwingungen wandern über den sogenannten Hammer, den ersten der drei Gehörknochen, über den Amboss zum sogenannten Steigbügel, dem letzten Gehörknochen. Der Steigbügel ist direkt mit der Hörschnecke, der Cochlea, über das ovale Fenster, die Öffnung zum Innenohr, ver-



(a) Aufbau des Gehörs.



(b) Orte der Amplitudenmaxima auf der Cochlea in Abhängigkeit von der Signalfrequenz.

Abbildung 2.1: (a) Makrostruktur des menschlichen Gehörs, bestehend aus Außenohr, Mittelohr und Innenohr. Einfallende Schallwellen versetzen das Trommelfell in Schwingung, was die Gehörknochen in Vibration versetzt. Die Gehörknochen versetzen wiederum die Flüssigkeit innerhalb der Cochlea, der Hörschnecke, in Schwingung. (b) Tonotopie der Cochlea. Hohe Frequenzen führen zu einer starken Auslenkung nahe dem ovalen Fenster, dem Eingang in die Cochlea (notiert als base). Tiefe Frequenzen führen nahe der Spitze der Schneckenform der Cochlea, dem Apex, zu einem Auslenkungsmaximum. Mit freundlicher Genehmigung durch Encyclopædia Britannica, Inc., copyright 2009. Nutzung mit Erlaubnis.

bunden. Dort wird die Umwandlung von mechanischer Schwingung in elektrische Impulse vollführt [Pic12]. Die Cochlea wird durch die Basilarmembran in zwei Gänge geteilt, welche in Abb. 2.1b zu sehen sind. Die über die Gehörknochen geleiteten Schallwellen versetzen das ovale Fenster in Schwingung. Diese Schwingungen führen zu Flüssigkeitswellen innerhalb der Cochlea und es kommt zu mechanischen Bewegungen der Basilarmembran. Man sagt, die Basilarmembran sei tonotopisch, d.h. dass zu jeder Frequenz ein bestimmter Ort auf der Cochlea respektive entlang der Basilarmembran existiert, an dem das Auslenkungsmaximum einer Schwingung dieser Frequenz liegt. Die ungefähre Lage dieser Auslenkungsmaxima ist in Abb. 2.1b dargestellt. Hohe Frequenzen haben ihr Auslenkungsmaximum nahe dem Eingang in die Cochlea, d.h. nahe dem ovalen Fenster. Tiefe Frequenzen wiederum haben ihr Auslenkungsmaximum nahe dem Ende der schneckenförmigen Struktur der Cochlea, d.h. nahe dem sogenannten Apex. Diese Tatsache motiviert die Positionierung der Elektroden eines Cochlea-Implantats.

Die Umwandlung in elektrische Impulse erfolgt innerhalb der Basilarmembran im sogenannten Corti'schen Organ. Dort führt die Bewegung der sogenannten inneren und äußeren Haarzellen zu Ladungsveränderungen und damit zu Stromimpulsen, die über den Hörnerv

zum Gehirn weitergeleitet werden. Es gibt etwa 11000 äußere und etwa 3500 innere Haarzellen, welche über die Cochlea verteilt sind [Ash08]. Man findet jedoch etwas variierende Angaben für die genaue Zahl an Haarzellen in der Fachliteratur. Durch die große Zahl an Haarzellen auf der vergleichsweise kleinen Oberfläche der Cochlea kommt es zu einer sehr feinen Auflösung der Frequenz.

Im Vergleich dazu stimulieren aktuelle Cochlea-Implantate typischerweise 22 Regionen der Cochlea durch ihre Elektroden.

2.1.1 *Hörschädigungen*

Während grundsätzlich jede Art von Beschädigung des im vorherigen Abschnitt grob skizzierten Übertragungssystems zu einer Beeinträchtigung des Hörsinns führen kann, interessieren im Rahmen dieser Arbeit jene Schädigungen, bei welchen der Hörnerv intakt bleibt, die Haarzellen jedoch derart geschädigt sind, dass sie die Umwandlung mechanischer in elektrische Signale nicht mehr leisten können oder diese zumindest sehr wesentlich beeinträchtigt ist. In diesen Fällen kann der Einsatz eines Cochlea-Implantats sinnvoll sein.

2.2 COCHLEA-IMPLANTATE

Eine der häufigsten Ursachen für eine Beeinträchtigung des Hörvermögens sind Beschädigungen der Haarzellen der Cochlea unter anderem bedingt durch Infektionen, Traumata sowie längere Exposition hoher Lärmpegel. Hörbeeinträchtigungen werden oftmals im Laufe des Lebens erworben, können aber auch von Geburt an bestehen. Insbesondere altersbedingter Hörverlust wird durch Beschädigungen der Haarzellen auf der Cochlea verursacht [Wu+20].

Sind die Haarzellen beschädigt, so ist die Umwandlung der mechanischen Schwingungen in der Cochlea in elektrische Signale/Pulse gestört. Teilweise können konventionelle Hörgeräte durch Verstärkung eingehender Schallsignale Abhilfe schaffen, jedoch gibt es einen Grad an Beschädigung der Haarzellen, ab welchem konventionelle Hörgeräte das Hörvermögen nicht mehr oder nicht mehr hinreichend verbessern können [Tur+10].

In diesem Fall, wenn die Haarzellen stark beschädigt sind, der Hörnerv jedoch im Wesentlichen intakt ist, kann durch direkte elektrische Stimulation des Hörnervs das Hörvermögen zumindest teilweise wiederhergestellt werden. Dies ist die Motivation für den Einsatz des sogenannten Cochlea-Implantats, welches die Funktion der Haarzellen übernehmen

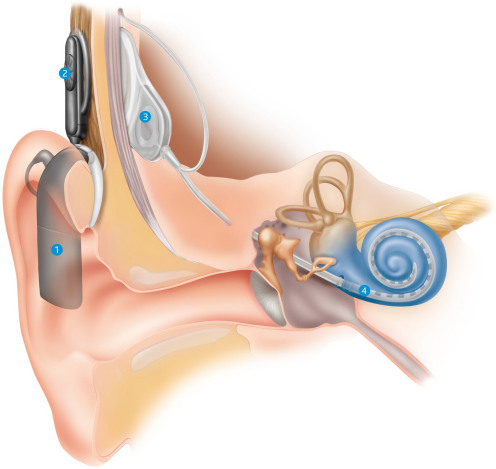


Abbildung 2.2: Schematik eines Cochlea-Implantats im Gehör. Hinter dem Ohr werden (1) Mikrophon und Signalprozessor getragen (zusammen mit der Batterie). Die vom Signalprozessor aus dem Mikrophonsignal abgeleiteten Stromwerte werden über die (2) Sendespule unter die Haut übertragen, wo sie von der (3) Empfängerspule aufgefangen werden. Diese gibt die empfangenen Signale an das (4) implantierte Elektrodenarray in der Cochlea ab. Dadurch wird der Hörnerv stimuliert und es entsteht ein Höreindruck. Bild mit freundlicher Genehmigung von Cochlear.

soll. Dazu müssen akustische Signale außen am Ohr durch ein Mikrophon aufgefangen und als elektrische Impulse über in die Cochlea implantierte Elektroden an den Hörnerv abgegeben werden. Maßgeblich für das hierdurch erzielbare Sprachvermögen ist zum einen die Zahl der Elektroden (sowie deren Platzierung) und der Algorithmus, welcher zur Umwandlung des akustischen Signals in elektrische Impulse genutzt wird. Diesen Algorithmus nennt man auch Stimulations- oder Codierungsstrategie (engl. stimulation strategy, aber auch sound coding strategy).

2.2.1 *Aufbau und Signalverarbeitung*

Die Hauptkomponenten eines Cochlea-Implantats sind heutzutage ein Mikrophon, welches akustische Signale auffängt, ein Signalprozesser, dessen Hauptaufgabe die Umwandlung des vom Mikrophon aufgefangenen Signals in elektrische Pulse ist, ein Spulenpaar zur drahtlosen Übermitt-

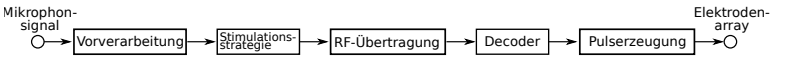


Abbildung 2.3: Blockdiagramm des vollständigen, allgemeinen Aufbaus eines Cochlea-Implantats. Adaptiert aus [WMF15]. Der Block RF-Übertragung stellt die Übertragung der abgeleiteten Stromwerte unter die Haut dar. Im Decoder wird eine für die tatsächliche Stromerzeugung passende Darstellungsform generiert.

lung dieser elektrischen Pulse unter die Haut, sowie Anregungselektronik und das Elektrodenarray, welches in die Cochlea implantiert ist, um an bestimmten Stellen die elektrischen Pulse an den Hörnerv abzugeben [WMF15]. Die Positionierung dieser Komponenten im Ohr ist in Abb. 2.2 gezeigt. Ähnlich wie auch bei konventionellen Hörgeräten wird der Signalprozessor samt Mikrophon hinter dem Ohr getragen (engl. behind-the-ear (BTE) hearing aid) und entsprechend sind die empfangenen Signale nicht von der Ohrmuschel und dem Gehörgang geformt. Abb. 2.3 zeigt das vollständige Blockdiagramm eines jeden Cochlea-Implantats vom Mikrophon bis hin zum Elektrodenarray. In der Vorverarbeitung kommt es z.B. zur Vorfilterung des Eingangsaudiosignals.

In der Anfangszeit des Cochlea-Implantats, d.h. in den 1950er Jahren, wurde lediglich eine Elektrode in die Cochlea implantiert [Rol+06]. Dadurch war kein Sprachverstehen herstellbar. Schnell wurden Mehrkanal-Cochlea-Implantate entwickelt und heutzutage haben Cochlea-Implantate bis zu 22 Elektroden, vereinzelt, nur vom Hersteller Nurotron, werden auch Cochlea-Implantate mit 24 Elektroden hergestellt [DJ17].

Es gibt eine Vielzahl an Stimulationsstrategien, wobei vier dominieren [WMF15], d.h. am weitesten in kommerziellen Signalprozessoren von Cochlea-Implantaten Anwendung finden. Diese vier sind der Advanced Combination Encoder (ACE), MP3000, Fine Structured Processing (FSP) sowie High Resolution (HiRes120). In dieser Arbeit wurde der Advanced Combination Encoder verwendet, der in der Mehrheit der Signalprozessoren von Cochlear Ltd. eingesetzt wird [SAH18]. Deswegen wird nur kurz auf die anderen Stimulationsstrategien eingegangen.

Eine vergleichende Übersicht ist in Abb. 2.4 gezeigt, wobei noch weitere Strategien, insbesondere das Continuous Interleaved Sampling (CIS), welches vor dem Aufkommen von ACE weit verbreitet war, und nach wie vor verwendet wird, dargestellt sind. Allen vier genannten Stimulationsstrategien ist gemein, dass sie das Audiosignal des Mikrophons mittels einer Filterbank in Subbänder zerlegen und, bis auf HiRes120, anschließend die Einhüllende für jedes Subband extrahieren. ACE und MP3000 nutzen zudem, anders als etwa CIS, eine Bandselektion (engl. channel selection), um in jedem Zeitschritt nur eine Teilmenge aller Elektroden

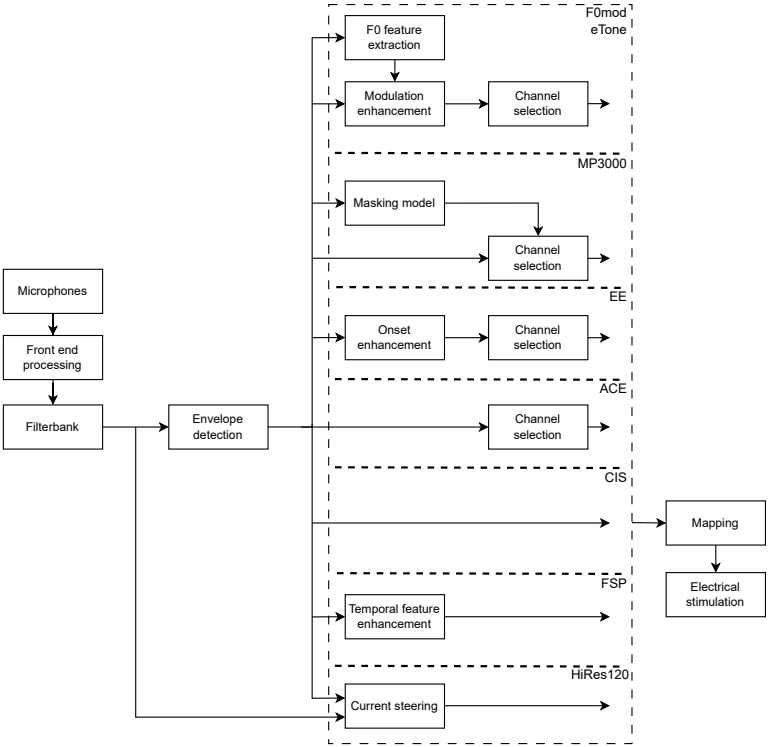


Abbildung 2.4: Vergleichende Übersicht aktuell verbreiteter Stimulationsstrategien von Cochlea-Implantaten. Abbildung adaptiert aus [WMF15].

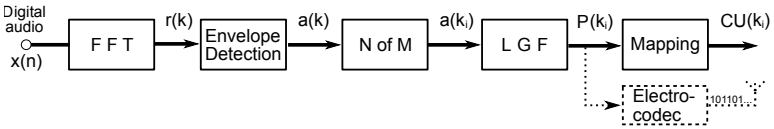


Abbildung 2.5: Blockdiagramm der Advanced Combination Encoder Stimulationsstrategie zusammen mit dem Electrocodec, einem im Rahmen der vorgelegten Arbeit entwickelten Kompressionssalgorithmus. Zunächst wird das digitalisierte Audiosignal mittels einer diskreten Fouriertransformation (FFT) in separate Subbänder aufgesplittet. Anschließend werden die Einhüllenden dieser Subbänder berechnet. Danach kommt es zur Bandselektion (N of M), und schließlich wird dieses akustische Signal in den elektrischen Bereich mittels der Lautheitswachstumfunktion (LGF) transformiert. Dieses Signal wird vom Electrocodec (und allen weiteren verlustbehafteten Verfahren) codiert. Im letzten Schritt wird das Ausgangssignal der LGF auf Stromwerte in klinischen Einheiten abgebildet (Mapping).

tatsächlich zu erregen. ACE nutzt als Selektionskriterium die Größe der Einhüllenden, wobei die N Bänder mit den betragsmäßig größten Einhüllenden selektiert werden. Im Gegensatz dazu verwendet MP3000 ein psychoakustisches Kriterium und nicht die Größe der Einhüllenden.

HiRes120 und FSP verzichten auf eine derartige Bandselektion. FSP versucht die Wahrnehmung von Tonhöhen mittels genauerer Darstellung der (zeitlichen) Feinstruktur des aufgefangenen akustischen Signals zu verbessern. HiRes120 verwendet sogenanntes current steering (zu deutsch etwa Stromsteuern), womit es möglich ist, virtuelle Elektroden zu erzeugen, d.h. scheinbar neue Bereiche der Cochlea, an denen keine Elektroden anliegen, anzuregen, indem die Strompulse der Elektroden geschickt gewählt werden.

2.2.1.1 *Der Advanced Combination Encoder*

Der Fokus der gesamten Arbeit stand auf der Kompression von Erregungsmustern, die mit dem Advanced Combination Encoder (ACE) generiert wurden. Zunächst soll ACE auf der Ebene der Signalverarbeitung erläutert und anschließend auf die konkret verwendete Implementierung in Matlab detailliert eingegangen werden. Diese wurde bereitgestellt von Cochlear Ltd. durch die Nucleus Matlab Toolbox (NMT) [Wil06].

Das Blockdiagramm der Forschungsimplementierung (engl. research implementation) von ACE ist in Abb. 2.5 zusammen mit dem Ansatzpunkt des Electrocodecs, eines im Rahmen der vorgelegten Arbeit entwickelten Kompressionsalgorithmus, dargestellt. ACE zerlegt zunächst

Tabelle 2.1: Mittenfrequenzen, Anzahl an DFT-Koffizienten, Startindizes sowie die Verstärkungsfaktoren je Subband des Cochlea-Implantats in der verwendeten Konfiguration. Dies entspricht den Standardeinstellungen. Adaptiert aus [Nogo8] mit korrigierten Verstärkungsfaktoren.

Bandnummer z	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Startindex k_{start_z}	3	4	5	6	7	8	9	10	11	13	15
#Koeffizienten N_z	1	1	1	1	1	1	1	1	1	2	2
Mittenfrequenz [Hz]	250	375	500	625	750	875	1000	1125	1250	1437	1687
Verstärkungsfaktoren g_z $[10^{-3}]$	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,68	0,68
Bandnummer z	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
Startindex k_{start_z}	17	19	22	25	28	32	36	41	45	51	58
#Koeffizienten N_z	2	2	3	3	4	4	5	5	6	7	8
Mittenfrequenz [Hz]	1937	2187	2500	2875	3312	3812	4375	5000	5687	6500	7437
Verstärkungsfaktoren g_z $[10^{-3}]$	0,68	0,68	0,65	0,65	0,65	0,65	0,65	0,65	0,65	0,65	0,65

das digitale Eingangssignal $x(n)$ mit Hilfe einer diskreten Fouriertransformation der Länge $L = 128$ in M Subbänder[1], typisch sind $M = 22$, der Standard von Cochlea-Implantaten von Cochlear Ltd., wobei jedes Subband einer Elektrode zugeordnet ist und einen gewissen Frequenzbereich abdeckt. Aus jedem Subbandsignal $r(k, m)$ wird anschließend die Einhüllende (engl. Envelope) bestimmt, wodurch die Signale $a(z, m)$ entstehen. Hierbei ist $z \in \{1, \ldots, M\}$ der Bandindex und m der Zeitindex. Die sogenannte N aus M Selektion (engl. N of M selection) oder auch Bandselektion wählt nachfolgend aus der Gesamtzahl M an Subbändern die N Subbänder mit der größten Amplitude aus. Nur diese werden weiterverarbeitet und alle anderen Subbänder tragen keinen Strom bzw. keinen Wert. Diese werden von der sogenannten Lautheitswachstumsfunktion (engl. loudness growth function (LGF)) zu Werten $P(z_i, m)$ im Intervall $[0, 1]$ abgebildet. Im letzten Verarbeitungsschritt von ACE werden die Werte $P(z_i, m)$ schließlich in Stromwerte $I(z_i, m)$ (engl. current level (CL)) in klinischen Einheiten (engl. clinical units (CU)) transformiert. Die Stromwerte sind natürliche Zahlen zwischen dem Schwellenwert (engl. threshold level (THR)) und der Obergrenze der angenehmen Lautheit (engl. most comfortable level (MCL)) und korrespondieren 1:1 mit Stromamplituden in Mikroampere [SDF17].

Im Folgenden wird die Signalverarbeitung von ACE mathematisch beschrieben, wobei sich die Darstellung und Nomenklatur eng an [Nogo8] anlehnt, hier jedoch etwas detaillierter ausfällt. Zunächst wird das Eingangssignal $x(n)$ abschnittsweise mit einem Hanning-Fenster gefenstert gemäß

$$x_w(l, m) = w(l)x(m \cdot N_s + l), l = 0, 1, \ldots, L-1, m \in \mathbb{Z}, \qquad (2.1)$$

1 In dieser Arbeit werden die Begriffe Subband und Band, sowie seltener auch Kanal, im Kontext von ACE synonym verwendet.

wobei N_s die sogenannte Blockverschiebung (engl. block shift) ist. Die Blockverschiebung ist die Zahl an Abtastwerten, um welche das Eingangssignal für jede Berechnung der Stromwerte verschoben wird. Sie ist gerade das Übersetzungsverhältnis der Abtastfrequenz f_s im Audiobereich zu der Zahl an Strompulsen pro Sekunde (engl. channel stimulation rate (CSR)), im Weiteren auch Stimulationsrate oder auch Kanalstimulationsrate genannt. Es ist also $N_s := [\frac{f_s}{CSR}]$, wobei $[\cdot]$ die Rundungsoperation darstellt. Bei einer CSR von 900 Pulsen pro Sekunde und der Abtastfrequenz $f_s = 16\,\text{kHz}$ ergibt sich $N_s = 18$. Aus den gefensterten Signalen $x_w(l, m)$ werden dann die Koeffizienten der diskreten Fouriertransformation (DFT) berechnet gemäß

$$r(k, m) = \sum_{l=0}^{L-1} x_w(l, m) e^{-j\frac{2\pi}{N}lk}. \tag{2.2}$$

Anschließend kommt es dann zur Bestimmung der Einhüllenden $a(z, m)$ gemäß

$$a(z, m) = \sqrt{\sum_{k=k_{start_z}}^{k_{start_z}+N_z-1} g_k |r(k, m)|^2}, \tag{2.3}$$

wobei die g_k die in Tabelle 2.1 angegebenen Verstärkungsfaktoren sind, die so gewählt sind, dass die Einhüllende den Wert 1 hat, wenn ein reiner Sinuston genau die zum Band $z \in \{1, \ldots, M\}$ gehörende Mittenfrequenz trifft [Nog08].

Die nachfolgende N aus M Selektion führt zu einer deutlichen Informationsreduktion. Die N Bänder mit den betragsmäßig größten Einhüllenden werden im Zeitschritt m ausgewählt, sodass nur für diese ein Stromwert berechnet wird, falls die zugehörige Einhüllende über dem Schwellenwert des Hörens s_{base}, dem sogenannten Basisniveau (engl. base level), liegt. Ein Band z wird selektiert, sofern

$$SV(a(z, m)) := \sum_{i \in \{1,\ldots,M\} \setminus \{z\}} \hat{u}(a(z, m), a(z_i, m)) \geqslant M - N \tag{2.4}$$

gilt, wobei

$$\hat{u}(a(z, m), a(z_i, m)) = \begin{cases} 1, & a(z, m) > a(z_i, m) \\ 1, & a(z, m) = a(z_i, m) \wedge z < z_i \\ 0, & sonst \end{cases} \tag{2.5}$$

ist, d.h. $\hat{u}$ ist eine modifizierte Sprungfunktion, die bei Gleichheit dem Band mit kleinerer Indizierung den Vorzug gibt.

Das Ausgangssignal der Bandselektion kann damit als Funktion

$$F(a(z,m)) = \begin{cases} a(z,m), & SV(a(z,m)) \geqslant M - N \wedge a(z,m) \geqslant s_{base} \\ -10^{-10}, & SV(a(z,m)) \geqslant M - N \wedge a(z,m) < s_{base} \\ NaN, & SV(a(z,m)) < M - N \end{cases} \quad (2.6)$$

geschrieben werden. Der Wert -10^{-10} dient als Platzhalterwert (engl. dummy value) in der NMT, da in jedem Zeitschritt N Bänder selektiert werden müssen. Jedoch führt der Wert -10^{-10} nicht zu einer Stimulation der Cochlea, d.h. er trägt keine Hörinformation. Der Wert NaN (not a number) wird für Bänder die nicht selektiert wurden genutzt. Er ist subjektiv, d.h. für den Höreindruck des Cochlea-Implantatträgers, identisch zum Wert -10^{-10}. In beiden Fällen wird kein Strom an die jeweilige Elektrode abgegeben.

Nachfolgend wird abkürzend $a(z_i, m)$ geschrieben für all jene Einhüllenden, für die $F(a(z,m)) = a(z,m)$ gilt.

Nach der Bandselektion werden die Einhüllenden $a(z_i, m)$ mittels der Lautheitswachstumsfunktion (engl. Loudness Growth Function (LGF)) in einen Bruchteil eines Stromwerts $P(z_i, m) \equiv P(a(z_i, m)) \in [0, 1]$ transformiert mittels

$$P(z_i, m) = \begin{cases} \frac{\log(1+\rho((a(z_i,m)-s_{base})/(s_{sat}-s_{base})))}{\log(1+\rho)}, & s_{base} \leqslant a(z_i,m) \leqslant s_{sat} \\ 1, & a(z_i,m) \geqslant s_{sat} \end{cases}, \quad (2.7)$$

wobei s_{sat} den sogenannten Sättigungsgrad (engl. saturation level) bezeichnet. Dies ist die lauteste akzeptable Amplitude. Eine höhere Amplitude würde zu einer unangenehmen Lautstärke für den Träger des Cochlea-Implantats führen. ρ steuert die Steilheit der LGF. Die entwickelten Kompressionsverfahren, mit Ausnahme eines verlustlosen Verfahrens, komprimieren genau die Werte $P(z_i)$.

Zum Schluss werden die $P(z_i)$ auf Stromwerte in sogenannten klinischen Einheiten abgebildet mittels der Formel

$$I(z_i, m) = THR(z_i) + [(MCL(z_i) - THR(z_i))P(z_i, m)]. \quad (2.8)$$

$[\cdot]$ ist erneut die Rundungsoperation. Die Stromwerte $I(z_i, m)$ korrespondieren über eine Exponentialfunktion mit Stromamplituden in μA [SDF17], welche über die in die Cochlea implantierten Elektroden nacheinander angelegt werden.

In dieser Arbeit wurden für die eingeführten Konstanten die Standardwerte von ACE verwendet mit einer CSR von 900 Pulsen pro Sekunde (PPS). Diese setzen $\rho = 416{,}2063$, $s_{base} = 4/256$ und $s_{sat} = 150/256$.

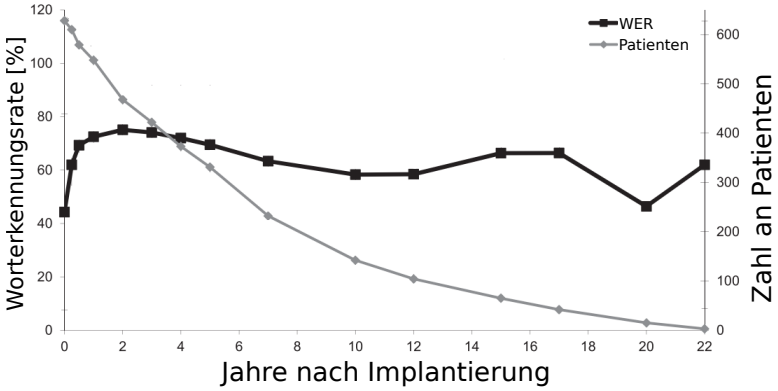


Abbildung 2.6: Mittlere Worterkennungsrate (WER) von Cochlea-Implantatträgern in Stille. Als Satztest wurde der Hochmair-Schulz-Moser Satztest (HSM) verwendet. Abbildung aus [Len+12] adaptiert.

Des Weiteren gilt durchweg $N = 8$ und $M = 22$, was weit verbreitete Einstellungen sind und ebenfalls den Standardeinstellungen der NMT entspricht. Insbesondere die Werte $\mathrm{THR}(z_i)$ und $\mathrm{MCL}(z_i)$ werden speziell auf den Träger des Cochlea-Implantats durch einen Audiologen abgestimmt [WGD21].

2.2.2 *Leistungsfähigkeit und Stand der Technik*

Mit aktueller Technologie ist es Nutzern von Cochlea-Implantaten typischerweise möglich, in ruhiger Umgebung Sprache gut zu verstehen [AM22] (und mäßig Schallquellen zu lokalisieren [And+22]). Hierbei ist jedoch die Definition von „gutem Sprachverstehen“ in der Fachliteratur relativ variabel, sodass je nach Definition $10 - 50\,\%$ der erwachsenen Nutzer von Cochlea-Implantaten ein schlechtes Sprachverstehen aufweisen [Mob+16]. Telefonie stellt für etwa ein Drittel der Nutzer von Cochlea-Implantaten ein großes Problem dar. Für sie ist Telefonie unmöglich, da der Sprecher nicht verstanden wird [Rum+15]. Bei diesem Problem kann moderne Technologie wie drahtloses Streamen von Telefonanrufen, sowie die Nutzung von Internettelefonie, eine Erleichterung verschaffen [Hut+22a].

Typischerweise werden von Cochlea-Implantatträgern Worterkennungsraten (engl. word recognition score (WRS)), das Verhältnis der Zahl an vorgespielten zur Zahl an verstandenen Worten, von etwa $80\,\%$ oder mehr in Stille erreicht [Vos+21; MG21], sofern ganze Sätze erkannt werden

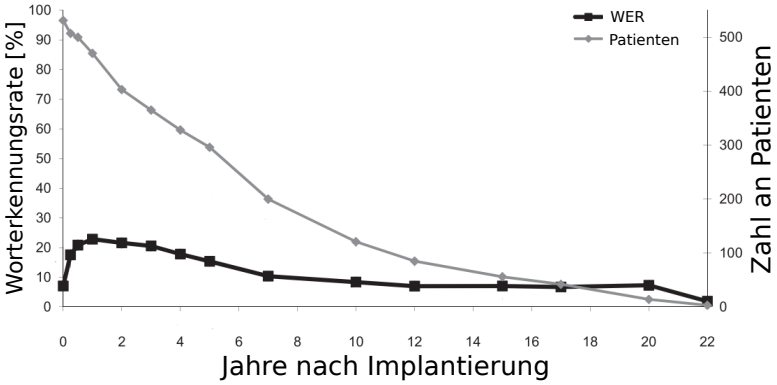


Abbildung 2.7: Mittlere Worterkennungsrate (WER) von Cochlea-Implantatträgern bei einem Signal-Rausch-Verhältnis von 10 dB. Als Satztest wurde der Hochmair-Schulz-Moser Satztest (HSM) verwendet. Abbildung aus [Len+12] adaptiert.

müssen, wobei das Sprachverstehen schnell im Vergleich zu Normalhörenden abnimmt, sobald Hintergrundrauschen vorliegt. Abb. 2.6 zeigt die mittlere Worterkennungsrate für eine große Kohorte von Cochlea-Implantatträgern, die in [Len+12] über Jahre nach der Implantierung ihres Cochlea-Implantats begleitet und wiederholt getestet wurden. Der starke Einfluss von Hintergrundrauschen ist im Vergleich von Abb. 2.6 und Abb. 2.7 zu sehen. Letztere zeigt das in [Len+12] gemessene Sprachverstehen bei einem Signal-Rausch-Verhältnis von 10 dB. Die mittlere Worterkennungsrate fiel dabei fast immer auf unter 20 %. Normalhörende würden hier eine Worterkennungsrate jenseits von 80 % erzielen. Hierbei ist zu beachten, dass mindestens zum Teil die Signalprozessoren der Probanden nicht aktualisiert wurden.

Genaue, allgemeingültige quantitative Angaben zur Worterkennungsrate in Abhängigkeit von etwa dem Signal-Rausch-Verhältnis sind schwer zu machen, da das in Untersuchungen erzielte Sprachverstehen von vielen konfundierenden Einflüssen wie Alter der Probanden, Stimulationsstrategie, akustisches Szenario, Jahren seit Implantierung des Cochlea-Implantats, Elektrodenposition, Art des Rauschens, Art des Sprachmaterials und weiteren Faktoren abhängt [MLN16]. Exemplarisch sei in Abb. 2.8, welche [AM22] entnommen ist, die Worterkennungsrate von Normalhörenden sowie Trägern von Cochlea-Implantaten in Abhängigkeit des Signal-Rausch-Verhältnisses dargestellt. Dieser Darstellung kann man gut das schlechtere Sprachverstehen von Cochlea-Implantatträgern entnehmen. Zum Erzielen einer Worterkennungsrate von 50 % wird gemäß

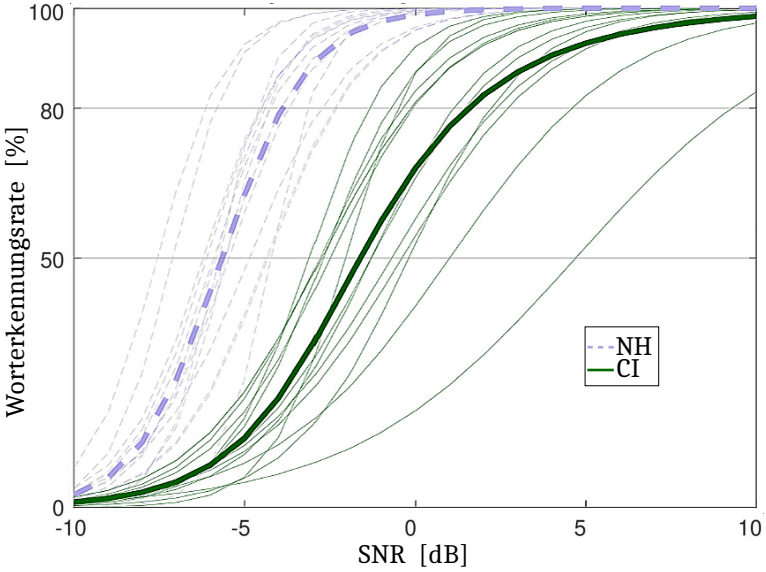


Abbildung 2.8: Gemitteltes Sprachverstehen (dicke Linien) für Normalhörende (NH; gestrichelt) und Cochlea-Implantatträgern (CI; durchgezogen) sowie individuelles Sprachverstehen (dünne Linien) einzelner Probanden in Abhängigkeit vom Signal-Rausch-Verhältnis (SNR). Abbildung angepasst aus [AM22].

dieser Arbeit ein um etwa 4 dB höheres Signal-Rausch-Verhältnis benötigt. Bei einer Worterkennungsrate von 80 % wird ein etwa 5 dB höheres Signal-Rausch-Verhältnis benötigt. Gut zu sehen ist in der Abbildung erneut die große Variabilität der einzelnen Cochlea-Implantatträger.

Eine große Zahl an Forschungsarbeiten [HGJ21] beschäftigt sich mit Rauschreduktionsalgorithmen mit dem Ziel, das Sprachverstehen von Cochlea-Implantatträgern in Situationen mit Hintergrundrauschen zu verbessern. Hierbei ist zwischen Algorithmen, die ein einzelnes Mikrophon voraussetzen, und Algorithmen, welche mehrere Mikrophone voraussetzen, zu unterscheiden. Beispiele für Algorithmen basierend auf einem Mikrophonsignal sind z.B. verbesserte Bandselektion des Cochlea-Implantats durch Selektion der Subbänder mit dem besten Signal-Rausch-Verhältnis [HL08], modifizierte, an Cochlea-Implantate angepasste Wiener-Filter [Gue+16] sowie Autoencoder zur Rauschentfernung [Wan+17]. Rauschreduktion auf Basis zweier oder mehr Mikrophonsignale wird meistens durch Strahlenformung erreicht [Vro+18; GR10], jedoch existieren auch andere Ansätze unter Verwendung von z.B.

blinder Quellenseparierung [KL08]. Durch diese und weitere Verfahren ist es im Allgemeinen möglich, die Worterkennungsrate um 10 % und mehr in anspruchsvollen akustischen Szenarien zu verbessern.

Weitere Technologien zur Verbesserung des Sprachverstehens von Cochlea-Implantatträgern, welche in ähnlicher Form von allen Herstellern angeboten werden, nutzen drahtloses Streaming von Audiosignalen von externen Geräten. Dies kommt etwa beim drahtlosen Streamen von Telefonanrufen [Wol+16; WDS16; Hut+22b] zum Einsatz, das jedes moderne Cochlea-Implantat unterstützt. Ferner wird Audio von externen Mikrophonen (engl. remote microphones) zum Cochlea-Implantat [Vro+17; WMS15] gestreamt sowie beim kontralateralen Routing (engl. contralateral routing of signals, CROS) [Nuñ+20], bei welchem Audio vom typischerweise nichtimplantierten, tauben Ohr drahtlos an das Cochlea-Implantat gestreamt wird. Weiterhin ist das Streamen von Audio für die Verwendung binauraler Strahlenformer nötig [Vro+18] sowie für binaurale Signalverarbeitungsstrategien.

2.2.3 *Objektive Maße des Hörverstehens*

Zur Messung des Hörvermögens von Hörgerätenutzern und insbesondere von Cochlea-Implantatträgern werden in der Forschung verschiedene Maße genutzt. Weit verbreitet ist die bereits erwähnte Worterkennungsrate. Praktisch wird dabei auf Satzlisten wie, im deutschsprachigen Bereich, den Hochmair-Schulz-Moser-Satztest (HSM) [Hoc+97] oder den Oldenburger Satztest (OLSA) [KKW99] zurückgegriffen. Der HSM, welcher zur Evaluierung eines im Rahmen der vorgelegten Arbeit entwickelten Kompressionsalgorithmus in Probanden verwendet wurde, besteht aus 30 Satzlisten á 20 Sätze, wobei jeder Satz aus fünf bis acht Worten besteht. Eine Reihe von Satzlisten wird zur Bestimmung der Worterkennungsrate Probanden vorgespielt, und diese geben durch Nachsprechen der verstandenen Satzteile Rückmeldung über ihr Verständnis des Sprachmaterials. Für jeden Satz werden vom Versuchsdurchführer die nicht erkannten Worte weggestrichen und die Zahl der verstandenen Worte notiert. Das Verhältnis der korrekt erkannten Worte zur Zahl der vorgespielten/präsentierten Worte in Prozent ist dann die Worterkennungsrate. Je höher die Worterkennungsrate ist, desto besser. Ein anderes, eng verwandtes Maß, ist die Sprachwahrnehmungsschwelle (engl. speech reception threshold). Bei dieser wird iterativ das Signal-Rausch-Verhältnis des Sprachmaterials reduziert bzw. erhöht bis eine mittlere Worterkennungsrate von 50 % erzielt wird. Das Signal-Rausch-Verhältnis in dB bei welcher diese Worterkennungsrate erzielt wird, ist dann die Sprachwahrnehmungsschwelle. Je niedriger die Sprachwahrnehmungsschwelle ist, desto besser.

Da es aufwändig ist, Hörtests durchzuführen, sind Verfahren sehr interessant, die algorithmisch für gegebenes Sprachmaterial das zu erwartende Hörverstehen von Nutzern von Cochlea-Implantaten bestimmen. Diese Art der objektiven Bestimmung des Hörverstehens hat eine lange Tradition im Bereich der Sprach- und Musiksignalverarbeitung [Thi+00] und wird seit einiger Zeit in Form spezialisierter Algorithmen im Bereich der Cochlea-Implantat Forschung genutzt.

Eine Vielzahl von Ansätzen zur objektiven Bestimmung des Sprachverstehens für gegebenes Sprachmaterial existiert für Cochlea-Implantate [WSS18b; Fal+15]. Hierbei hat sich das Short-Time Objective Intelligibility Measure (STOI) [Taa+10] als gutes Maß herausgestellt [Fal+15], welches vielfach zur Entwicklung oder Evaluierung von Algorithmen herangezogen wurde [HWL21; HMK18; Li+21]. Dieses wurde auch in der vorgelegten Arbeit für die Entwicklung von Kompressionsalgorithmen verwendet und soll im Folgenden kurz erläutert werden.

2.2.3.1 *Short-Time Objective Intelligibility Measure*

STOI vergleicht die Verständlichkeit eines Referenzsignal $ref(n)$ mit einer bearbeiteten oder verrauschten Kopie $y(n)$ durch Bestimmung des Korrelationskoeffizienten der Einhüllenden von speziellen Frequenzbändern. Genauer werden $J = 15$ Dritteloktavbänder verwendet, wobei diese Bänder durch Gruppierung der jeweiligen DFT-Koeffizienten gebildet werden. Die Norm des j-ten Dritteloktavbands, dies ist die jeweilige Einhüllende, wird in diesem Kontext berechnet gemäß

$$\mathrm{REF}_j(m) = \sqrt{\sum_{k=k_{j,1}}^{k_{j,2}} |\hat{ref}(k,m)|^2}$$

mit den unteren und oberen Bandgrenzen $k_{j,1}$ und $k_{j,2}$. $\hat{ref}(k,m)$ ist hierbei die Kurzzeitfouriertransformation von $ref(n)$. Analog wird $Y_j(m)$ bestimmt. Es ist $m = 1,\dots,M$, wobei M die Gesamtanzahl an Frames ist. Anschließend werden die $Y_j(m)$ modifiziert: Zum einen wird die Energie von $Y_j(m)$ normiert, sodass sie identisch mit jener des Referenzsignals ist. Des Weiteren werden die $Y_j(m)$ geclippt, um das Signal-zu-Verzerrungsverhältnis (engl. signal-to-distortion ratio) nach unten zu beschränken. Insgesamt ergeben sich diese modifizierten Signale $\hat{Y}_j(m)$ gemäß

$$\hat{Y}_j(m) = \max(\min(\alpha Y_j(m), \mathrm{REF}_j(m) + \Delta_m), \mathrm{REF}_j(m) - \Delta_m) \quad (2.9)$$

mit $\Delta_m := 10^{-\frac{\beta}{20}} \mathrm{REF}_j(m)$. α und β sind Konstanten, welche in [Taa+10] beschrieben sind.

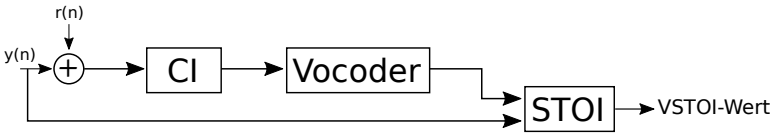


Abbildung 2.9: Das unbearbeitete, rauschfreie Signal $y(n)$ wird mit additivem Rauschen $r(n)$ gemischt und die Sprachverständlichkeit mittels des Short-Time Objective Intelligibility Measure (STOI) bestimmt, indem die Erregungsmuster des Cochlea-Implantats (CI) resynthetisiert werden, und das resultierende Signal mit dem unbearbeiteten Signal $y(n)$ verglichen wird.

Abschnittsweise ergibt sich das Sprachverständlichkeitsmaß, welches von STOI bestimmt wird, nun mittels

$$d_j(m) = \frac{\sum_m (REF_j(m) - \frac{1}{N}\sum_l REF_j(l))(\tilde{Y}_j(m) - \frac{1}{N}\sum_l \tilde{Y}_j(l))}{\sum_m (REF_j(m) - \frac{1}{N}\sum_l REF_j(l))^2 \sum_m (\tilde{Y}_j(m) - \frac{1}{N}\sum_l \tilde{Y}_j(l))^2}.$$

Bei dieser Formel handelt es sich um eine Schätzung des Korrelationskoeffizienten zwischen $REF_j(m)$ und $\hat{Y}_j(m)$. Der schlussendlich von STOI bestimmte Wert, der die Verständlichkeit des Signals $y(n)$ im Vergleich zu $ref(n)$ bemisst, ergibt sich dann via

$$d = \frac{1}{JM}\sum_{j,m} d_j(m) \tag{2.10}$$

mit J und M wie oben eingeführt. Es ist $d \in [0,1]$, wobei $d = 1$ die bestmögliche Verständlichkeit bezeichnet und $d = 0$ die schlechtmöglichste. Insbesondere im Kontext von Algorithmen, die die Erzeugung von Erregungsmustern von Cochlea-Implantaten tangieren, sowie Algorithmen, welche nachträglich die Erregungsmuster abändern, wie es im Falle der verlustbehafteten Kompression der Erregungsmuster geschieht, wird STOI zusammen mit einem Vocoder genutzt. Man spricht dann auch von Vocoder STOI (VSTOI). Hier werden die Erregungsmuster des Cochlea-Implantats, nachdem jegliche zusätzliche Algorithmik Anwendung gefunden hat, mittels eines Vocoders resynthetisiert, d.h. es wird ein Audiosignal aus den Erregungsmustern wiedergewonnen, welches dann mittels STOI mit dem unbearbeiteten, originalen Audiosignal verglichen wird. Beispielhaft ist die Berechnung dieser VSTOI-Werte (engl. VSTOI Scores) in Abb. 2.9 dargestellt. Das unbearbeitete Signal $y(n)$ wird mit additivem Rauschen $r(n)$ gemischt und die Verständlichkeit mittels STOI bestimmt, indem die Erregungsmuster des Cochlea-Implantats resynthetisiert werden und das resultierende Signal mit dem unbearbeiteten Signal $y(n)$ verglichen wird.

Der von STOI zurückgegebene Wert d kann nach [Taa+10] mittels der Funktion

$$f(d) = \frac{1}{1 + e^{a\,d+b}}$$

auf eine mittlere Worterkennungsrate abgebildet werden. Die reellen Parameter a und b sind dabei abhängig vom verwendeten Sprachmaterial. Insbesondere bedeutet dies, für negatives a wie in [Taa+10], dass die Worterkennungsrate streng monoton mit dem VSTOI-Wert wächst und somit zum Vergleich von Kompressionsalgorithmen verwendet werden kann. Ein höherer (V)STOI-Wert bedeutet also, auf dem selben Sprachmaterial und identischer CI-Konfiguration, höhere mittlere Sprachverständlichkeit. Auf Grund der Gleichung 2.9 ist STOI nicht differenzierbar und kann somit nicht unmittelbar zum Training künstlicher neuronaler Netze (oder anderer ableitbarer Algorithmen) mittels Gradientenverfahren genutzt werden. Diese Tatsache wird im Training von Autoencodern für die Kompression von Erregungsmustern aufgegriffen.

2.3 THEORETISCHE GRUNDLAGEN UND VERFAHREN DER DATENKOMPRESSION

Das Feld der Datenkompression, auch Quellencodierung genannt, befasst sich mit der Minimierung der zur Darstellung eines Signals respektive einer Quelle nötigen Informationsmenge, gemessen typischerweise in der notwendigen Zahl an Bits oder der Bits pro Sekunde, der sogenannten Bitrate. Verlustlose Datenkompression, auch Entropiecodierung genannt, befasst sich mit dem Problem der Minimierung der zur Darstellung nötigen Informationsmenge unter Gewährleistung perfekter Rekonstruktion, d.h. ein zu komprimierendes Signal ist nach Kompression und Dekompression unverändert und es tritt kein Informationsverlust auf. Verlustbehaftete Datenkompression lässt die Nebenbedingung der perfekten Rekonstruktion fallen und erlaubt einen gewissen Informationsverlust, es kommt zu irreversiblen Verzerrungen des Signals [Say96]. Dadurch ist jedoch das Erzielen einer weit besseren Kompression möglich, d.h. die zur Darstellung eines Signals notwendige Informationsmenge ist im Allgemeinen für verlustbehaftete Kompressionsalgorithmen deutlich kleiner als für verlustlose Kompressionsalgorithmen. Zum Verständnis der Methoden der Quellencodierung, die in dieser Arbeit eingesetzt wurden, sind zunächst einige Begriffe der Wahrscheinlichkeitstheorie notwendig.

2.3.1 *Grundlagen der Wahrscheinlichkeitstheorie*

Die moderne (axiomatische) Wahrscheinlichkeitstheorie fußt auf dem Ansatz von Kolmogorov, welcher Zufallsexperimente mittels Begriffen aus der Maß- und Integrationstheorie beschreibt [Beh13]. Zentral ist der Begriff des Wahrscheinlichkeitsraums, der das Tripel $(\Omega, \mathcal{E}, \mathbb{P})$ bestehend aus der Ergebnismenge Ω, einer auf der Ergebnismenge erklärten σ-Algebra[2] $\mathcal{E}$ sowie einem Wahrscheinlichkeitsmaß $\mathbb{P} : \mathcal{E} \to [0, 1]$ bezeichnet. Die Elemente von $\mathcal{E}$ werden Ereignisse genannt und $\mathbb{P}$ weist jedem Ereignis eine Zahl zu, die als Wahrscheinlichkeit dieses Ereignisses interpretiert wird. Für das Wahrscheinlichkeitsmaß fordert man die beiden Eigenschaften $\mathbb{P}(\Omega) = 1$ (Normiertheit) und $\mathbb{P}(\cup_i A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ für paarweise disjunkte Mengen $A_1, A_2, \ldots$ (σ-Additivität). Je nach Wahl des Wahrscheinlichkeitsraumes ergibt sich ein anderes Zufallsexperiment. Ist die Ergebnismenge endlich, d.h. $|\Omega| < \infty$, so kann man $\mathcal{E} = \mathrm{P}(\Omega)$ wählen, wobei $\mathrm{P}(\Omega)$ die Potenzmenge von Ω bezeichnet. Dies ist für $|\Omega| = \infty$ nicht immer möglich, da nicht messbare Teilmengen existieren können. In diesen Fällen wird standardmäßig die borelsche σ-Algebra verwendet, welche die kleinste σ-Algebra ist, die alle offenen Intervalle (im Falle reeller Zahlen) enthält [Beh13].

Insbesondere im Ingenieursbereich arbeitet man oftmals mit sogenannten Wahrscheinlichkeitsdichten (auch Dichtefunktion oder nur Dichte genannt). Dies sind nichtnegative Funktionen (es wird nur der eindimensionale Fall betrachtet) $f : A \subset \mathbb{R} \to \mathbb{R}_+$, für welche gilt $\int_A f(x)dx = 1$. Über eine solche Wahrscheinlichkeitsdichte kann man Wahrscheinlichkeitsmaße definieren mittels[3] $\mathbb{P}(A) := \int_A f(x)dx$. Der umgekehrte Fall, eine Dichte zu einem Wahrscheinlichkeitsmaß $\mathbb{P}$ (und einer assoziierten σ-Algebra $\mathcal{E}$) zu finden, sodass $\mathbb{P}(A) = \int_A f(x)dx$ gilt für alle $A \in \mathcal{E}$, wird wesentlich vom Satz von Radon-Nikodým behandelt, der die Existenz von Wahrscheinlichkeitsdichten für Zufallsvariablen und relativ allgemeine Wahrscheinlichkeitsmaße garantiert. Man darf also in den üblichen Fällen von der Existenz einer Dichtefunktion ausgehen [Els18].

Allgegenwärtig ist des Weiteren der Begriff der Zufallsvariable. Eine Zufallsvariable $\mathbf{X}$ ist eine messbare Funktionen[4] $\mathbf{X} : \Omega \to \mathbb{R}$. Eine Funktion $f : M \to N$ mit Messräumen $(M, \mathcal{M})$, $(N, \mathcal{N})$ heißt *messbar*, wenn für alle $A \in \mathcal{N}$ gilt, dass $f^{-1}(A) \in \mathcal{M}$ ist [Beh13]. Das heißt, dass Urbilder messbarer Mengen unter messbaren Funktionen messbar sind. Diese Bedingung ist notwendig, damit man Zufallsvariablen Wahrscheinlich-

2 Eine Mengenfamilie $\mathcal{E}$ definiert auf einer Grundmenge Ω heißt σ-Algebra, sofern gilt: 1) $\emptyset, \Omega \in \mathcal{E}$, 2) $\cup_{i=1}^{\infty} U_i \in \mathcal{E}, U_i \in \mathcal{E}$, 3) $U \in \mathcal{E} \Rightarrow U^c \in \mathcal{E}$.

3 Lebesgue-Integrierbarkeit von f über alle A vorausgesetzt.

4 Es sind alle stetigen Funktionen messbar (bezüglich einer gegebenen borelschen σ-Algebra).

keiten bzw. Wahrscheinlichkeitsmaße zuweisen kann. Betrachtet man nämlich die Menge $\{\omega \in \Omega | \mathbf{X}(\omega) = a \in \mathbb{R}\} \equiv \mathbf{X}^{-1}(\{a\})$, so kann dieser Menge durch das Wahrscheinlichkeitsmaß $\mathbb{P}$ des Wahrscheinlichkeitsraums $(\Omega, \mathcal{E}, \mathbb{P})$ eine Wahrscheinlichkeit gemäß $\mathbb{P}(X^{-1}(\{a\}))$ zugewiesen werden, was immer definiert ist aufgrund der Messbarkeit der Urbilder messbarer Funktionen. Gegeben eine Zufallsvariable $\mathbf{X} : \Omega \to \mathbb{R}$, so lässt sich nun das sogenannte induzierte Wahrscheinlichkeitsmaß $\mathbb{P}_{\mathbf{X}} : \mathcal{B} \to [0, 1]$ definieren über [Beh13]

$$\mathbb{P}_{\mathbf{X}}(A) = \mathbb{P}(\mathbf{X}^{-1}(A)), A \in \mathcal{B}, \tag{2.11}$$

wobei $\mathcal{B}$ die borelsche σ-Algebra ist, d.h. die kleinste σ-Algebra, die alle offenen Teilmengen von $\mathbb{R}$ enthält. Damit kann man entsprechend aus einem allgemeinen Wahrscheinlichkeitsmaß ein Wahrscheinlichkeitsmaß für Zufallsvariablen gewinnen. Für die Wahrscheinlichkeitsdichte $f_{\mathbf{X}}(\alpha)$ einer Zufallvariablen $\mathbf{X}$, oder auch kürzer $f(\alpha)$, falls der Kontext klar oder irrelevant ist, gilt dann entsprechend

$$\mathbb{P}_{\mathbf{X}}(A) = \int_A f_{\mathbf{X}}(\alpha) d\alpha. \tag{2.12}$$

Wesentlich für das Verständnis des Problems der Prädiktion, wie sie etwa in prädiktiven Codierungen verwendet wird, zu denen die in dieser Arbeit genutzte Differential Puls-Code Modulation gehört, ist der Begriff des Erwartungswerts.

Gegeben eine Wahrscheinlichkeitsdichte $f_{\mathbf{X}}(\alpha)$ einer Zufallsvariablen $\mathbf{X}$, so ist der Erwartungswert $\mathbf{E}(\mathbf{X})$ von $\mathbf{X}$ definiert als

$$\mathbf{E}(\mathbf{X}) := \int_{Dom(f_{\mathbf{X}})} \alpha f_{\mathbf{X}}(\alpha) d\alpha \tag{2.13}$$

mit dem Definitionsbereich $Dom(f_{\mathbf{X}})$ der Dichte $f_{\mathbf{X}}$. Für den Fall einer diskreten Zufallsvariablen auf einer endlichen Ergebnismenge $\Omega := \{\omega_1, \ldots, \omega_N\}$ ist der Erwartungswert definiert als [Beh13]

$$\mathbf{E}(\mathbf{X}) = \sum_{\omega \in \Omega} \mathbf{X}(\omega) \mathbb{P}(\{\omega\}) \equiv \sum_{i=1}^{N} x_i p_i \tag{2.14}$$

mit $x_i := \mathbf{X}(\omega_i)$ und $p_i := \mathbb{P}(\{\omega_i\})$.

Für (messbare) Funktionen $g(\mathbf{X})$ einer Zufallsvariable $\mathbf{X}$ ist

$$\mathbf{E}(g(\mathbf{X})) = \int_{Dom(f_{\mathbf{X}})} g(\alpha) f_{\mathbf{X}}(\alpha) d\alpha. \tag{2.15}$$

Der Erwartungswert ist ein lineares Funktional, d.h. für kontinuierliche und diskrete Zufallsvariablen $\mathbf{X}, \mathbf{Y}$ und $\alpha, \beta \in \mathbb{R}$ gilt

$$\mathbf{E}(\alpha\mathbf{X} + \beta\mathbf{Y}) = \alpha\mathbf{E}(\mathbf{X}) + \beta\mathbf{E}(\mathbf{Y}). \tag{2.16}$$

Um die im Allgemeinen komplizierte Struktur einer Wahrscheinlichkeitsdichte zu quantifizieren, werden verschiedene geläufige charaktierische Größen verwendet. Die vermutlich bekannteste neben dem Erwartungswert ist die Varianz $Var(\mathbf{X})$ einer Zufallsvariable, welche definiert ist als [Beh13]

$$Var(\mathbf{X}) := \mathbf{E}((\mathbf{X} - \mathbf{E}(\mathbf{X}))^2) = \mathbf{E}(\mathbf{X}^2) - \mathbf{E}(\mathbf{X})^2, \tag{2.17}$$

d.h. es handelt sich hierbei um die erwartete quadratische Abweichung einer Zufallsvariablen von ihrem Erwartungswert. Die ebenfalls sehr gebräuchliche Standardabweichung $Std(\mathbf{X})$ ist definiert als $Std(\mathbf{X}) := \sqrt{Var(\mathbf{X})}$ und ist ein besser interpretierbares Maß der Streuung einer Zufallsvariablen als die Varianz. Der Erwartungswert einer Zufallsvariablen ist gerade der Minimierer der mittleren quadratischen Abweichung, was leicht aus der Problemformulierung

$$E((\mathbf{X} - \alpha)^2) \to \min, \alpha \in \mathbb{R} \tag{2.18}$$

nebst Ableitung nach α folgt:

$$\frac{\partial}{\partial\alpha} E((\mathbf{X} - \alpha)^2) = \frac{\partial}{\partial\alpha}(E(\mathbf{X}^2) - 2\alpha E(\mathbf{X}) + \alpha^2) = -2E(\mathbf{X}) + 2\alpha \stackrel{!}{=} 0, \tag{2.19}$$

was offenbar die Lösung $\alpha = E(\mathbf{X})$ hat. Diese Eigenschaft wird zum Verständnis des Prädiktionsproblems helfen.

Wichtig für das Verständnis der Prädiktion ist zudem das Konzept der Abhängigkeit bzw. Unabhängigkeit von Zufallsvariablen oder Ereignissen. Gegeben einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{E}, \mathbb{P})$, so heißen zwei Ereignisse $A, B \in \mathcal{E}$ (statistisch) unabhängig genau dann, wenn $P(A \cap B) = P(A)P(B)$ gilt [Beh13]. Für eine Wahrscheinlichkeitsdichte $f_{\mathbf{X},\mathbf{Y}}(\alpha, \beta)$ zweier (statistisch) unabhängiger Zufallsvariablen $\mathbf{X}$ und $\mathbf{Y}$ entspricht dies $f_{\mathbf{X},\mathbf{Y}}(\alpha, \beta) = f_{\mathbf{X}}(\alpha)f_{\mathbf{Y}}(\beta)$, d.h. genau dann, wenn die Zufallsvariablen unabhängig sind, ist ihre Wahrscheinlichkeitsdichte als Produkt der Einzeldichten darstellbar. Hieraus folgt insbesondere, dass für statistisch unabhängige Zufallsvariablen $\mathbf{X}$ und $\mathbf{Y}$ die Gleichung

$$E(\mathbf{XY}) = E(\mathbf{X})E(\mathbf{Y}) \tag{2.20}$$

gilt.

2.3.2 *Dichten transformierter Zufallsvariablen*

Im Kontext der Untersuchung der Wahrscheinlichkeitsdichte der Erregungsmuster in Kapitel 5 stößt man auf das Problem, ausgehend von der Dichte $f_{\mathbf{X}}(\alpha)$ einer Zufallsvariablen $\mathbf{X}$, die Dichte $f_{g(\mathbf{X})}(\alpha)$ der transformierten Zufallsvariablen $\mathbf{Y} := g(\mathbf{X})$ mit bijektiver, differenzierbarer Funktion g zu bestimmen. Es lässt sich zeigen, dass die Dichte von $\mathbf{Y}$ gemäß

$$f_{\mathbf{Y}}(y) = f_{\mathbf{X}}(g^{-1}(y))|\frac{dg^{-1}(y)}{dy}| \tag{2.21}$$

bestimmt werden kann, was, wegen $\frac{d(g\circ g^{-1})(y)}{dy} = 1$ und $\frac{d(g\circ g^{-1})(y)}{dy} = g'(g^{-1}(y))\frac{dg^{-1}(y)}{dy}$, äquivalent ist zu[5] $f_{\mathbf{X}}(g^{-1}(y))|\frac{1}{g'(g^{-1}(y))}|$. Beide Varianten werden im Rahmen dieser Arbeit für die Berechnung genutzt. Ferner lässt sich leicht zeigen, dass für $c \in \mathbb{R}$

$$f_{c\cdot\mathbf{X}}(\alpha) = \frac{f_{\mathbf{X}}(\frac{\alpha}{c})}{|c|} \tag{2.22}$$

gilt [Beh13]. Relevant ist ferner, dass sich, mit den Dichten $f_{\mathbf{X}}$ und $f_{\mathbf{Y}}(\alpha)$ zweier statistisch unabhängiger Zufallsvariablen $\mathbf{X}$ und $\mathbf{Y}$, die Dichte $f_{\mathbf{X}+\mathbf{Y}}(\alpha)$ von $\mathbf{X}+\mathbf{Y}$ berechnen lässt gemäß [Rob78]

$$f_{\mathbf{X}+\mathbf{Y}}(\alpha) = \int_{\mathbb{R}} f_{\mathbf{X}}(u)f_{\mathbf{Y}}(\alpha-u)du. \tag{2.23}$$

Die Dichte der Summe zweier statistisch unabhängiger Zufallsvariablen ergibt sich also durch Faltung der jeweiligen Dichten.

2.3.3 *Korrelation*

Ein sehr wesentlicher Begriff der Stochastik ist jener der Korrelation. Im umgangssprachlichen Gebrauch ist unter Korrelation typischerweise eine allgemeine, nicht näher spezifizierte Abhängigkeit zwischen zwei Größen gemeint. Diese kann durch eine beliebige Funktion gegeben sein. Demgegenüber wird im engen Sinne mit dem Begriff Korrelation eine lineare statistische Abhängigkeit bezeichnet. Diese wird quantitativ durch Korrelationskoeffizienten bemessen. Vermutlich am weitesten verbreitet ist der Korrelationskoeffizient nach Pearson $\rho_{\mathbf{X},\mathbf{Y}}$, welcher definiert ist gemäß [Ibe14]

$$\rho_{\mathbf{X},\mathbf{Y}} := \frac{E(\mathbf{XY}) - E(\mathbf{X})E(\mathbf{Y})}{Std(\mathbf{X})Std(\mathbf{Y})} \tag{2.24}$$

5 $g'(x) \equiv \frac{dg(x)}{dx}$.

mit Zufallsvariablen $\mathbf{X}$ und $\mathbf{Y}$. Es gilt immer $-1 \leqslant \rho_{\mathbf{X},\mathbf{Y}} \leqslant 1$. Da für statistisch unabhängige Zufallsvariablen die Beziehung $\mathbb{E}(\mathbf{XY}) = \mathbb{E}(\mathbf{X})\mathbb{E}(\mathbf{Y})$ gilt, misst dieser Korrelationskoeffizient anschaulich, wie sehr sich zwei Zufallsvariablen in dieser Eigenschaft von statistisch unabhängigen Zufallsvariablen unterscheiden. Der Korrelationskoeffizient nach Pearson misst sogenannte lineare Abhängigkeiten [Bla93]. Gilt etwa $\mathbf{X} = \alpha\mathbf{Y}$ mit $\alpha \in \mathbb{R} \setminus \{0\}$, so gilt $|\rho_{\mathbf{X},\mathbf{Y}}| = 1$ [Beh13]. Die Umkehrung ist auch wahr, d.h. gilt $|\rho_{\mathbf{X},\mathbf{Y}}| = 1$, so ist[6] $\mathbf{X} = \alpha\mathbf{Y}$. Liegt $\rho_{\mathbf{X},\mathbf{Y}} = 0$ vor, so darf im Allgemeinen nicht geschlossen werden, dass die Zufallsvariablen statistisch unabhängig sind. Es kann dennoch eine deterministische Abhängigkeit bestehen, die vom Korrelationskoeffizienten nicht erfasst wird. Trotz dieser Schwäche ist der Korrelationskoeffizient als Maß für statistische Abhängigkeiten sehr beliebt aufgrund der Einfachheit seiner Berechnung.

Der Korrelationskoeffizient tritt im Rahmen dieser Arbeit in Form der normierten Autokorrelation auf. Ferner wird auf seiner Basis der in der vorgelegten Arbeit zur Modellbestimmung verwendete partielle Korrelationskoeffizient bestimmt. Dieser kann durch Fälle motiviert werden, in welchen tatsächlich eine Abhängigkeit der Art $\mathbf{X} = \alpha\mathbf{Z}$ sowie $\mathbf{Y} = \beta\mathbf{Z} + \eta$ vorliegt, wobei η statistisch unabhängig zu allen anderen Zufallsvariablen sein möge. Untersucht man nun die Abhängigkeit der Zufallsvariablen $\mathbf{X}$ von $\mathbf{Y}$, so ergibt sich eine große Korrelation (sofern $\mathrm{Std}(\eta)$ hinreichend klein ist), obwohl in Wirklichkeit $\mathbf{X}$ nur von $\mathbf{Z}$ abhängt. Dies wird durch die Abhängigkeit beider Zufallsvariablen $\mathbf{X}, \mathbf{Y}$ von $\mathbf{Z}$ bewirkt. Hält man den Wert von $\mathbf{Z}$ fest, so ändert sich $\mathbf{Y}$ gemäß der Zufallsvariablen η und hat somit keinen Einfluss auf $\mathbf{X}$. Man spricht von einer Scheinkorrelation. Um die Korrelation zwischen zwei Zufallsvariablen ohne den Einfluss weiterer Zufallsvariablen zu bestimmen, kann man den sogenannten partiellen Korrelationskoeffizienten nutzen, der für den Fall nur einer weiteren Zufallvariablen gemäß [MG45]

$$\rho_{\mathbf{X},\mathbf{Y}|\mathbf{Z}} = \frac{\rho_{\mathbf{X},\mathbf{Y}} - \rho_{\mathbf{X},\mathbf{Z}}\rho_{\mathbf{Y},\mathbf{Z}}}{\sqrt{1-\rho_{\mathbf{X},\mathbf{Z}}^2}\sqrt{1-\rho_{\mathbf{Y},\mathbf{Z}}^2}} \tag{2.25}$$

berechnet wird, wobei die Notation $\rho_{\mathbf{X},\mathbf{Y}|\mathbf{Z}}$ die Korrelation von $\mathbf{X}$ und $\mathbf{Y}$ unter Konstanthaltung von $\mathbf{Z}$ bezeichnet. Auch für den partiellen Korrelationskoeffizienten gilt immer $-1 \leqslant \rho_{\mathbf{X},\mathbf{Y}|\mathbf{Z}} \leqslant 1$.

2.3.4 *Stochastische Prozesse*

Gegeben einen zugrundeliegenden Wahrscheinlichkeitsraum $(\Omega, \mathcal{E}, \mathbb{P})$, so ist ein stochastischer Prozess eine Familie $\{\mathbf{X}_n | n \in I\}$ von Zufallsvaria-

6 Im strengen Sinne gilt die Gleichheit nur fast überall.

blen $\mathbf{X}_n$ (definiert auf demselben Wahrscheinlichkeitsraum), wobei im allgemeinsten Fall I eine beliebige Indexmenge ist. Hier wird jedoch lediglich $I = \mathbb{Z}$ betrachtet, was hier als diskrete Zeit interpretiert wird. Man nennt einen solchen stochastischen Prozess dann auch einen zeitdiskreten (Zufalls-)Prozess. Stochastische Prozesse dienen als mathematisches Modell nichtdeterministischer Systeme oder Prozesse in allen möglichen Bereichen der Wissenschaft [LRS13].

In der allgemeinen Form ohne weitere Einschränkungen, wie sie eingeführt wurden, ist dieses Modell jedoch noch nicht nützlich, da für die $\mathbf{X}_n$ keine weiteren Eigenschaften definiert sind. Man fordert daher typischerweise mindestens schwache Stationarität. Zur Einführung dieses Begriffs wird zunächst aber die sogenannte Autokorrelationsfunktion benötigt.

Diese ist von eminenter Bedeutung zur Beschreibung und Analyse von stochastischen Prozessen und ohne Normierung definiert gemäß [PP02]

$$\hat{\phi}_{X,X}(n,k) := \mathbf{E}(\mathbf{X}_n \mathbf{X}_{n+k}) \tag{2.26}$$

und mit Normierung auf den Wertebereich $[-1,1]$ als

$$\phi_{X,X}(n,k) = \frac{\hat{\phi}_{X,X}(n,k)}{\sigma_{\mathbf{X}_n}\sigma_{\mathbf{X}_{n+k}}} \tag{2.27}$$

mit $\sigma_{\mathbf{X}_n} := \mathtt{Std}(\mathbf{X}_n)$. Bei der normierten Autokorrelationsfunktion handelt es sich also lediglich um den Korrelationskoeffizienten nach Pearson (Mittelwertfreiheit angenommen), bei welchem mittels eines Index durch eine Menge von Zufallsvariablen iteriert werden kann.

Damit können nun schwach stationäre Prozesse definiert werden: Ein Zufallsprozess $(X_n)_{n\in\mathbb{Z}}$ heißt schwach stationär, wenn die folgenden Bedingungen erfüllt sind:

1) $\mathbf{E}(\mathbf{X}_n) = c \in \mathbb{R}, \forall n \in \mathbb{Z}$

2) $\phi_{X,X}(n,k) = \phi_{X,X}(0,k-n), \forall n,k \in \mathbb{Z}$

3) $\sigma_{\mathbf{X}_n} < \infty, \forall n \in \mathbb{Z}$.

Insbesondere ist also die Autokorrelationsfunktion unabhängig von der absoluten Zeit n und hängt ausschließlich von der Zeitdifferenz $n-k$ ab. Damit schreibt man die (normierte) Autokorrelationsfunktion eines schwach stationären Prozesses typischerweise als $\phi(k) := \frac{\mathbf{E}(\mathbf{X}_n \mathbf{X}_{n+k})}{\sigma^2_{\mathbf{X}_n}}$, wobei das gewählt n irrelevant ist. Aus der Eigenschaft 2 folgt, dass die Varianz $\sigma^2_{\mathbf{X}_n}$ für einen schwach stationären Prozess unabhängig von der Zeit ist.

2.3.5 *Quellencodierung*

Quellencodierung, der Fachbegriff für das Feld der Datenkompression, befasst sich mit dem Problem der verlustlosen und verlustbehafteten Kompression von Daten bzw. Signalen. Ziel ist die Minimierung der zur Darstellung eines Signals benötigten Menge an Bits [Say96]. Hierbei überführt der sogenannte Encoder das zu komprimierende Signal in eine Folge von Bits, deren Gesamtzahl kleiner sein sollte als die Gesamtzahl der Bits in der Ausgangsdarstellung. Diese Überführung ist im engeren Sinne die (Quellen-)Codierung bzw. (Daten-)Kompression und man spricht auch von der komprimierten Darstellung/Repräsentation. Der zum Encoder gehörende Decoder überführt diese komprimierte Darstellung in das ursprüngliche Signal, entweder, im Falle der verlustlosen Kompression, perfekt, d.h. ohne Verzerrung, oder, im Falle der verlustbehafteten Kompression, nicht perfekt, d.h. mit irgendeiner Art von Verzerrung. Das Ausgangssignal des Decoders wird auch Rekonstruktion oder rekonstruiertes Signal genannt.

Die Verzerrung eines Ursprungssignals x wird dabei über eine Bewertungsfunktion d gemessen, z.B. den mittleren quadratischen Fehler. Im Falle der verlustbehafteten Kompression eines Signals soll oftmals $d(x, \hat{x})$ für eine spezielle, an der Wahrnehmung des Menschen orientierte Bewertungsfunktion d, auch Verzerrungsmaß genannt, minimiert werden. $\hat{x}$ bezeichnet hierbei das rekonstruierte Signal nach Kompression und Dekompression. Durch die Berücksichtigung der Wahrnehmung ist es im Allgemeinen möglich eine Kompression zu verbessern, d.h. die Größe einer komprimierten Darstellung bei gleicher Qualität zu reduzieren [HD19a].

Verlustbehaftete Kompressionsalgorithmen kommen im Rahmen der vorgelegten Arbeit in Form der Differential Puls-Code Modulation sowie Autoencodern vor. Bei letzteren dient in dieser Arbeit als Bewertungskriterium zumeist die Verständlichkeit der rekonstruierten Erregungsmuster. Diese soll bei gleicher Bitrate möglichst hoch sein.

Grundsätzlich ist es interessant, theoretische Aussagen über die Komprimierbarkeit von Signalen zu machen, wie sehr also die zur Darstellung eines Signals bzw. Signale einer bestimmten Quelle nötige Zahl an Bits maximal reduziert werden kann. Dies hängt sowohl vom Verzerrungsmaß d, als auch von maximal zulässigen Verzerrung ab. Allgemein wird diese Frage von der sogenannten Raten-Verzerrungs-Funktion (engl. rate distortion function) für ein gegebenes Verzerrungsmaß d beantwortet. Jedoch ist die tatsächliche Berechnung für eine Signalquelle extrem schwierig und nur in ausgewählten, theoretischen Fällen sind Lösungen bekannt. Einfacher ist die Frage nach dieser Untergrenze für verlustlose Kompression.

Hier bildet die sogenannte Entropie einer Signalquelle die Untergrenze der erzielbaren Kompression [Say96]. Details des Gebiets der verlustlosen Kompression, auch Entropiecodierung genannt, sollen im Folgenden näher erläutert werden.

2.3.5.1 *Entropiecodierung*

Das Feld der Entropiecodierung befasst sich mit der verlustlosen Datenkompression. Zentral ist das Modell der (Daten-)Quelle $\mathbf{Q}$, einer Zufallsvariablen über einem Wahrscheinlichkeitsraum $(\mathcal{A}, \mathcal{E}, \mathbb{P})$, deren Werte als Symbole bezeichnet werden und aus der Menge[7] $\mathbf{Q}(\mathcal{A}) := \{s_0, \ldots, s_{N-1}\}$, dem sogenannten Alphabet, stammen [Mus15]. Praktisch ist es unerheblich, ob man $\mathbf{Q}(\mathcal{A})$ oder die Ergebnismenge $\mathcal{A}$ als Alphabet betrachtet. Eine Kernfrage der Entropiecodierung ist es, welche mittlere Informationsmenge zur Darstellung einer gegebenen Quelle $\mathbf{Q}$ notwendig ist, wobei der Informationsgehalt $I(s_k)$ des Symbols s_k, gemessen in Bit, definiert ist als

$$I(s_k) := -\log_2(\mathbb{P}(\mathbf{Q}^{-1}(\{s_k\}))). \tag{2.28}$$

Der erwartete Informationsgehalt einer Quelle $\mathbf{Q}$ wird als Entropie bezeichnet, notiert als $H(\mathbf{Q})$, und ist definiert als

$$H(\mathbf{Q}) := \mathbf{E}(I(Q)) = \sum_{k=0}^{N-1} p_k I(s_k), \tag{2.29}$$

wobei $p_k := \mathbb{P}(\mathbf{Q}^{-1}(\{s_k\}))$ verwendet wurde. Die Entropie wird als die kleinste Zahl an Bits interpretiert, mit der eine Quelle verlustlos codiert werden kann. Streng genommen wurde hier die unbedingte Entropie definiert, d.h. der erwartete Informationsgehalt von $\mathbf{Q}$ unter der Prämisse, dass keine statistischen Abhängigkeiten zwischen auftretenden Symbolen existieren[8].

Bei dieser Betrachtung wird ferner angenommen, dass die Quelle $\mathbf{Q}$ unendlich oft „abgefragt" wird, d.h. man betrachtet Folgen der Art $(\mathbf{Q}(\omega_i \in \mathcal{A}))_{i \in \mathbb{N}}$.

Eine verlustlose Codierung wird dabei von einem sogenannten (verlustlosen) Codec durchgeführt, einem Paar $(\mathtt{Enc}, \mathtt{Dec})$ von Abbildungen, wobei $\mathtt{Enc} : \mathcal{A} \to \{c_0, \ldots, c_{N-1}\}$ und $\mathtt{Dec} : \{c_0, \ldots, c_{N-1}\} \to \mathcal{A}$ ist mit den sogenannten Codewörtern $c_i \in \{0,1\}^{N_i}$ mit $N_i \in \mathbb{N}$ und $\mathtt{Dec} \circ \mathtt{Enc} = \mathrm{Id}_{\mathcal{A}}$, d.h. Dec ist das Linksinverse von Enc. Die Codewörter sind in der Praxis nichts anderes als Abfolgen von Nullen und

7 Es werden nur diskrete Quellen betrachtet.

8 Die Entropie lässt sich auch allgemeiner für Quellen mit beliebig langem Gedächtnis definieren und berechnen. Ihre Bedeutung ist auch dann dieselbe.

Einsen [GG91]. Für jeden verlustlosen Codec $(\mathrm{Enc}, \mathrm{Dec})$ stellt die Quellenentropie $H(\mathbf{Q})$ die untere Schranke der erwarteten Codewortlänge dar:

$$H(\mathbf{Q}) \leqslant \mathbf{E}(L(\mathrm{Enc}(\mathbf{Q}))), \tag{2.30}$$

wobei die Funktion L definiert ist als $L : \{c_0, \ldots, c_{N-1}\} \to \mathbb{N}, L(c_i) = N_i$. Für einen Codec $(\mathrm{Enc}, \mathrm{Dec})$ nennt man $\mathbf{E}(L(\mathrm{Enc}(\mathbf{Q}))) - H(\mathbf{Q}) \geqslant 0$ die Redundanz des Codecs bzw. des Codes [Mus15]. Im Allgemeinen ist die Entropie nicht ganzzahlig, weshalb, da Codierungen von Einzelsymbolen notwendigerweise mit einer ganzen Zahl an Bits je Einzelsymbol erfolgen, normalerweise eine gewisse Redundanz existiert.

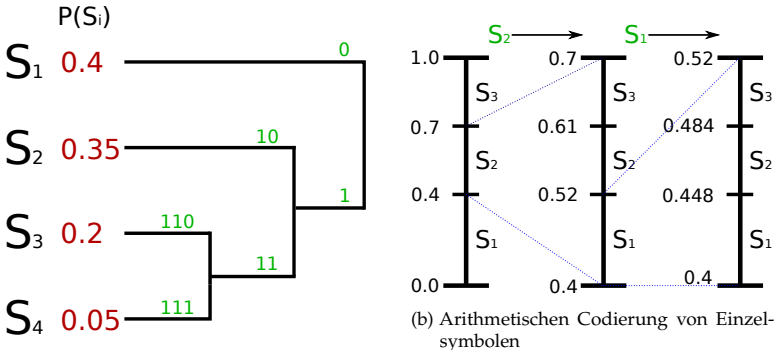
Bei der Codierung von Nachrichten, also Folgen $(s_k)_{k\in\mathbb{Z}}$ von Symbolen mit $s_k \in \mathcal{A}$, kann die Redundanz jedoch theoretisch auf Null reduziert werden, indem ein Codec gewählt wird, der die Symbole blockweise codiert, d.h. ein entsprechender Encoder Enc wäre dann definiert als $\mathrm{Enc} : \{s_0, \ldots, s_{N-1}\}^B \to \{\hat{c}_0, \ldots, \hat{c}_{\hat{N}-1}\}$ mit $B > 1$ (analog der Decoder). B bezeichnet man als Blocklänge. Ein Codec mit Blocklänge B sei notiert als $(\mathrm{Enc}_B, \mathrm{Dec}_B)$. Es lässt sich zeigen [Sha48], dass immer ein Codec existiert für den die Entropie für $B \to \infty$ erreicht wird, d.h. es gilt

$$H(\mathbf{Q}) = \lim_{B\to\infty} \mathbf{E}(L(\mathrm{Enc}_B(\mathbf{Q}))) \tag{2.31}$$

für einen Codec. Zwei weit verbreitete Verfahren der Codierung, die auch in dieser Arbeit Anwendung fanden, sind die sogenannte Huffman-Codierung und die arithmetische Codierung [Mus15]. Bei der Huffman-Codierung wird der sogenannte Huffman-Code erzeugt. Hierbei handelt es sich um einen Optimalcode für ein vorgegebenes Alphabet mit zugehörigen Symbolwahrscheinlichkeiten. Ein Beispiel ist in Abb. 2.10a dargestellt. Der Algorithmus zur Bestimmung des Huffmancodes geht wie folgt vor [Say96]:

Die Alphabetsymbole $\{a_1, \ldots, a_N\}$ werden ihrer Wahrscheinlichkeit nach geordnet. Die beiden Symbole mit niedrigster Wahrscheinlichkeit werden als Zweige eines Knotens des Codebaums notiert und ihre Wahrscheinlichkeiten addiert. Sie bilden zusammen ein neues „Pseudosymbol" und das Symbol mit der nächst höheren Wahrscheinlichkeit, im Beispiel, welches in Abb. 2.10a gezeigt ist, a_2, wird mit diesem Pseudosymbol über einen Knoten verbunden. Dies geht sukzessive weiter bis zum Symbol mit der höchsten Auftrittswahrscheinlichkeit. Zur Erstellung des Codebuchs wird nun der Baum von der Wurzel zu den Blättern durchwandert und jede Abzweigung erhält das Gewicht '1' oder '0'. Das einem Symbol assoziierte Codewort ist dann gegeben durch Verkettung der Gewichte aller Zweige, die von der Wurzel zum Symbolblatt durchwandert werden.

Dieser Algorithmus kann für beliebig große Alphabete mit beliebigen Auftrittswahrscheinlichkeiten durchgeführt werden. Prinzipiell ist der



(a) Huffmanbaum

(b) Arithmetischen Codierung von Einzelsymbolen

Abbildung 2.10: (a) Erstellung eines Huffman-Codebaums. In Rot sind die Symbolwahrscheinlichkeiten notiert, in Grün die Codeworte. (b) Codierung einer Symbolfolge mittels arithmetischer Codierung.

Huffman-Code optimal bezüglich der erzielten Redundanz unter der Voraussetzung statischer Auftrittswahrscheinlichkeiten der Symbole sowie Einzelsymbole. Die Codierung von Nachrichten ("Blockcodierung"), also Sequenzen von Einzelsymbolen, ist prinzipiell auch optimal mit der Huffmancodierung möglich, jedoch würde dafür ein mit der Nachrichtenlänge exponentiell wachsendes Codebuch benötigt werden. Eine Adaption an sich ändernde Auftrittswahrscheinlichkeiten der Symbole ist prinzipiell ebenfalls möglich, jedoch müsste für jeden Fall das Codebuch neu berechnet werden. Dadurch ist die Huffmancodierung für diese Fälle in der Praxis ungeeignet und es wird meistens auf arithmetische Codierung [Say96] oder verwandte Verfahren zurückgegriffen.

Die arithmetische Codierung ist als Blockcodierung konzipiert. Ihr Grundprinzip wird in Abb. 2.10b veranschaulicht. Es wird das Intervall $[0,1)$ betrachtet und jedem Alphabetsymbol ein Subintervall I_k zugewiesen, für das $|I_k| = |[l_k, r_k)| := r_k - l_k = P(a_k) \equiv P_{a_k}$ gilt, d.h. die Subintervallbreite ist gerade die Auftrittswahrscheinlichkeit des k-ten Symbols. Für eine beliebige Reihenfolge der Symbole ist die linke Seite l_k des Intervalls I_k gerade gegeben durch

$$l_k = \sum_{i=1}^{k-1} P_{a_i}. \tag{2.32}$$

Soll nun ein Symbol a_i codiert werden, so wird das Intervall I_i ausgewählt und die Subintervalle neu berechnet, nun mit I_i anstelle von $[0,1)$ als Grundlage. Dies kann für beliebig lange Nachrichten weitergeführt werden. Hat man auf diese Weise alle Symbole einer Nachricht codiert,

so stellt sich die Frage nach dem Codewort, welches das letzte Intervall, das sich nach Berücksichtigung des letztens Symbols einer Nachricht ergibt, eindeutig codiert. Denn über das finale Intervall ist eine Nachricht gemäß Konstruktion eindeutig festgelegt. Dies ist die Grundidee der arithmetischen Codierung. Eine übliche Lösung ist es, die linke Seite l_f des finalen Intervalls $I_f := [l_f, r_f)$ als Binärzahl zu codieren, d.h. man stellt l_f approximativ dar als

$$l_f^N = \sum_{n=1}^{N} b_n 2^{-n}, b_n \in \{0, 1\} \tag{2.33}$$

mit $l_f - l_f^N = \sum_{n=N+1}^{\infty} b_n 2^{-n}$, wobei N so gewählt sein muss, dass $l_f^N + 2^{-N} \in I_f$ ist. Dies ist immer gewährleistet, wenn gilt

$$N = \lceil -\log_2(|I_f|) \rceil + 1. \tag{2.34}$$

Als Codewort wird dann das N-Tupel $(b_1, b_2, \ldots, b_N)$ verwendet bzw. die Bitsequenz, welche sich durch die Verkettung dieser b_i ergibt. Somit ist N gerade die Codewortlänge. Die Gl. 2.34 ist eine hinreichende Bedingung für die Codewortlänge um Decodierbarkeit zu gewährleisten. Sie ist im Allgemeinen nicht notwendig und eine kleine Verbesserung, welche in dieser Arbeit zum Einsatz kam, wird in Abschnitt 2.3.6 kurz diskutiert.

Es lässt sich zeigen [Mus15], dass für eine Nachrichtenquelle **Q**, der mittleren Wortlänge $\overline{n}_{Huffman}$ der Huffmancodierung sowie der Blocklänge B

$$H(\mathbf{Q}) \leqslant \overline{n}_{Huffman} \leqslant H(\mathbf{Q}) + \frac{1}{B} \tag{2.35}$$

und für die mittlere Wortlänge $\overline{n}_{AC}$ der arithmetischen Codierung

$$H(\mathbf{Q}) \leqslant \overline{n}_{AC} \leqslant H(\mathbf{Q}) + \frac{2}{B} \tag{2.36}$$

gilt. Durch Grenzwertbildung $B \to \infty$ ergibt sich somit sowohl für die arithmetische als auch für die Huffmancodierung eine optimale mittlere Codewortlänge ohne Redundanz.

2.3.6 *Verbesserung der Arithmetischen Codierung*

Die Bestimmungsgleichung 2.34 für die Codewortlänge basiert auf der folgenden elementaren Überlegung: Ist $I := [a, b) \subset [0, 1)$ ein Intervall mit Intervallbreite $|I| = b - a$ sowie Schrittweite S und einem Punkt $p \in [a, \frac{a+b}{2})$, so garantiert $S \leqslant \frac{|I|}{2}$, dass $p + S \in I$ unabhängig von der genauen Wahl von p ist. Setzt man $|I| = 2^{-\hat{N}}$ mit $\hat{N} \in \mathbb{R}_+$, so muss $S \leqslant$

$2^{-(\hat{N}+1)}$ gelten. Gewiss erfüllt dann auch die Wahl $S = 2^{-(\lceil \hat{N} \rceil + 1)}$ die Bedingung $p + S \in I$ unabhängig vom exakten p. Da jedoch $\lceil \hat{N} \rceil - \hat{N} \geqslant 0$ ist, existiert ein Subintervall $\hat{I} := [a, c)$ mit $c < b$, sodass für $\forall \hat{p} \in \hat{I}$ gilt, dass $\hat{p} + 2^{-\lceil \hat{N} \rceil} \in I$ ist. In diesem Fall kann man also ein Bit sparen, da es sich bei $\lceil \hat{N} \rceil$ bzw. $\lceil \hat{N} \rceil + 1$ um die notwendige Codewortlänge handelt, um einen garantiert decodierbaren Code zu erhalten. Hierbei muss jedoch eines dieser $\hat{p}$ mit $\hat{N}$ Bit darstellbar sein. Zur Veranschaulichung sei das Intervall $I = [0, 2^{-(1+\varepsilon)})$ mit $0 < \varepsilon \ll 1$ betrachtet. Der exakte Mittelpunkt ist $2^{-2-\varepsilon}$. Hier ist $\hat{N} = 1 + \varepsilon$ und folglich $\lceil \hat{N} \rceil = 2$. Für $\varepsilon \approx 0$ ist folglich $2^{-\lceil \hat{N} \rceil}$ beliebig wenig vom wahren Mittelpunkt entfernt. Ergo gilt für nahezu alle $p \in [0, 2^{-2-\varepsilon})$, dass $p + 2^{-\lceil \hat{N} \rceil} \in I$ liegt. Ist folglich ein Punkt aus $[0, 2^{-(1+\varepsilon)})$ mittels $\lceil \hat{N} \rceil$ Bits darstellbar, so kann das zu I gehörende Codewort mit einem Bit weniger identifiziert werden. Nach Gleichung 2.34 wären hier drei Bits zur Codierung eines zu I gehörenden Symbols nötig. Tatsächlich ist es aber in diesem Beispiel möglich, das Intervall I mit der 0 zu identifizieren und die Codierung der 0 ($\in$ I) mittels zweier Bits unabhängig von weiteren Symbolen zu bewerkstelligen. Es ergibt sich also eine bessere Kompression. In der Praxis hat sich dadurch eine Bitratenreduktion von etwa 0,36 kbit/s im Kontext der Kompression der Bandselektion ergeben. Die Überlegungen dieses Abschnitts wurden vom Autor eigenständig getätigt, jedoch ergab eine nachträgliche Recherche, dass diese mögliche Bitersparnis grundsätzlich bekannt ist. Es konnte aber keine zitierfähige Quelle gefunden werden, weswegen keine Quelle angegeben ist.

2.3.7 *Prädiktion*

Betrachtet man einen schwach stationären Zufallsprozess $\mathbf{X}_n$, so lautet die typische Problemstellung der (Vorwärts-)Prädiktion eine Funktion f der Zufallsvariablen $\mathbf{X}_{n-1}, \mathbf{X}_{n-2}, \ldots, \mathbf{X}_{n-L}$ zu finden, welche die erwartete quadratische Abweichung minimiert, d.h. welche

$$E((\mathbf{X}_n - f(\mathbf{X}_{n-1}, \mathbf{X}_{n-2}, \ldots, \mathbf{X}_{n-L}))^2) \tag{2.37}$$

minimiert. Aus dem vorherigen Abschnitt ist bekannt, dass es sich bei f um den (bedingten) Erwartungswert handeln müsste. In der Tat ist der optimale Prädiktor gerade gegeben durch den bedingten Erwartungswert von $\mathbf{X}_n$ gegeben $\mathbf{X}_{n-1}, \mathbf{X}_{n-2}, \ldots, \mathbf{X}_{n-L}$, d.h. durch die Funktion $f(\mathbf{X}_{n-1}, \mathbf{X}_{n-2}, \ldots, \mathbf{X}_{n-L}) := E(\mathbf{X}_n | \mathbf{X}_{n-1}, \mathbf{X}_{n-2}, \ldots, \mathbf{X}_{n-L})$ [Kar93]. Dieser ist jedoch in der Praxis nicht oder nur mit sehr großem Aufwand berechenbar, und es ist üblich an dessen Stelle den besten linearen Prädiktor $f(\mathbf{X}_{n-1}, \ldots, \mathbf{X}_{n-L}) = \sum_{i=1}^{L} a_i \mathbf{X}_{n-i}$ der Ordnung L zu bestimmen.

Die optimalen Prädiktorkoeffizienten $a_i \in \mathbb{R}, i = 1, \ldots, L$ berechnet man dabei mittels Nullsetzen des Gradienten ∇ bezüglich der Koeffizienten a_i der erwarteten quadratischen Abweichung und lösen des entstehenden Gleichungssystems [Vaio7]:

$$\nabla E((\mathbf{X}_n - \sum_{i=1}^{L} a_i \mathbf{X}_{n-i})^2) \stackrel{!}{=} 0 \Leftrightarrow E(\mathbf{X}_n \mathbf{X}_{n-k}) = \sum_{i=1}^{L} a_i E(\mathbf{X}_{n-i} \mathbf{X}_{n-k})$$

für $k = 1, 2, \ldots, L$, wobei die rechte Seite mit Hilfe der (normierten) Autokorrelationsfunktion geschrieben werden kann als

$$\phi_{X,X}(k) = \sum_{i=1}^{L} a_i \phi_{X,X}(k-i), k = 1, 2, \ldots, L, \tag{2.38}$$

wodurch sich die sogenannte Wiener-Hopf Gleichung (auch Yule-Walker- oder Normalengleichung genannt) ergibt:

$$\underbrace{\begin{bmatrix} \phi_{X,X}(0) & \phi_{X,X}(1) & \cdots & \phi_{X,X}(N-1) \\ \phi_{X,X}(1) & \phi_{X,X}(0) & \cdots & \phi_{X,X}(N-2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{X,X}(N-1) & \phi_{X,X}(N-2) & \cdots & \phi_{X,X}(0) \end{bmatrix}}_{\mathbf{R}} \underbrace{\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}}_{\mathbf{a}} = \underbrace{\begin{bmatrix} \phi_{X,X}(1) \\ \phi_{X,X}(2) \\ \vdots \\ \phi_{X,X}(N) \end{bmatrix}}_{\mathbf{r}} \tag{2.39}$$

mit der Autokorrelationsmatrix $\mathbf{R}$ und dem Autokorrelationsvektor $\mathbf{r}$. Die optimalen Prädiktorkoeffizienten ergeben sich damit zu

$$\mathbf{a} = \mathbf{R}^{-1}\mathbf{r}, \tag{2.40}$$

wobei sich auf Grund der Toeplitz-Struktur von $\mathbf{R}$ schnelle Lösungsverfahren wie der Levinson-Durbin-Algorithmus für die Berechnung der Inversen $\mathbf{R}^{-1}$ anwenden lassen.

In einer praktischen Implementierung sind typischerweise N Abtastwerte $x(n-N), x(n-(N-1)), \ldots, x(n-1)$ verfügbar, die als Realisierung des zugrundeliegenden stochastischen Prozesses interpretiert werden, und es sind die optimalen Prädiktorkoeffizienten zu berechnen. Es gibt vier Hauptverfahren, um die optimalen Prädiktorkoeffizienten eines linearen Prädiktors zu berechnen: Die Autokorrelationsmethode, die Kovarianzmethode, die Burgmethode sowie die Maximum Likelihood Methode [Can20]. Zur Erläuterung der Unterschiede betrachte man den Prädiktionsfehler

$$e(n) = x(n) - \sum_{i=1}^{L} a_i x(n-i), \tag{2.41}$$

wobei $L < N$ gelten soll. Die Koeffizienten sollen das Summenquadrat $\sum_{n=1}^{N} e^2(n)$ des Prädiktionsfehlers minimieren. Nach Wahl der Methode erfolgt die Berechnung der optimalen Filterkoeffizienten wie zuvor mittels Gradientenbildung (mit Außnahme der Maximum Likelihood Methode). Ein Unterschied der Methoden rührt von der Behandlung des Anfangs der Prädiktion her. Zu Beginn existieren nicht für alle Filterkoeffizienten a_i vorherige Werte $x(n-i)$, da nur ein Ausschnitt zugänglich ist. In der Autokorrelationsmethode werden diese fehlenden Werte als Null angenommen. In der Kovarianzmethode wird statt dessen der Fehler über den kleineren Ausschnitt $x(n-(N-L)), x(n-(N-L-1)), \ldots, x(n-1)$ optimiert, sodass $x(n-N), x(n-(N-1)), \ldots, x(n-(N-L+1))$ am Anfang als vorherige genutzt werden können. In der Burgmethode wird der sogenannte Vorwärts- und Rückwärtsfehler gleichzeitig minimiert, wobei der Anfang der Prädiktion jeweils behandelt wird wie in der Autokorrelationsmethode. Die Maximum Likelihood Methode nutzt wiederum ein Wahrscheinlichkeitsmodell für den zugrunde liegenden stochastischen Prozess und kann dann, ähnlich der Autokorrelationsmethode, zu Beginn vorherige Werte als Null annehmen oder einen kürzeren Abschnitt der Daten zur Bestimmung der Prädiktorkoeffizenten berücksichtigen wie die Kovarianzmethode. Für $N \to \infty$ liefern alle Methoden dieselben Koeffizienten.

In der Praxis ist die Autokorrelationsmethode in der Codierung am weitesten verbreitet, obwohl die Kovarianzmethode etwas bessere Ergebnisse bezüglich des Prädiktionsfehlers zeigt. Der Grund für das Vorziehen der Autokorrelationsmethode ist die typische Notwendigkeit der Rekonstruktion des ursprünglichen Signals gegeben den (quantisierten) Prädiktionsfehler. Die Kovarianzmethode kann kein stabiles Filter garantieren, die Autokorrelationsmethode jedoch schon [Mak75]. In dieser Arbeit wurde für die Bestimmung der optimalen Prädiktorkoeffizienten ebenfalls die Autokorrelationsmethode gewählt. Zur Bewertung der Prädiktorgüte respektive der Prädiktionsgüte gibt es verschiedene Maße. In der Codierung ist unter anderem der sogenannte Prädiktionsgewinn PG geläufig, in dB definiert als

$$PG = 10 \cdot \log_{10}\left(\frac{\sigma_y^2}{\sigma_{y-\hat{y}}^2}\right) \tag{2.42}$$

mit der Varianz σ_y^2 eines zu prädizierenden Signals $y(n)$ und der Varianz $\sigma_{y-\hat{y}}^2$ des Prädiktionsfehlers $y(n) - \hat{y}(n)$. $\hat{y}$ ist die Prädiktion eines zu bewertenden Prädiktors. Der Prädiktionsgewinn misst wie sehr die Varianz des Prädiktionsfehlers im Vergleich zur Varianz des Ursprungssignals reduziert ist.

Die nachfolgend erläuterte Quantisierung von Signalen hängt maßgeblich von der Varianz der zu quantisierenden Signale ab. Je kleiner die Varianz, desto geringer der Fehler, der durch eine Quantisierung hervorgerufen wird. Wird die Quantisierung auf ein Prädiktionsfehlersignal angewendet, wie bei der noch erläuterten Differential Puls-Code Modulation, so führt ein höherer Prädiktionsgewinn ceteris paribus zu einem geringeren Fehler durch die Quantisierung. Daher ist er ein natürliches Maß für die Prädiktion im Kontext prädiktiver Kompressionsalgorithmen. Es gilt in diesem Kontext die grobe Faustregel, dass für je 6 dB Prädiktionsgewinn die Codebuchgröße eines Quantisierers, ohne Qualitätsverlust, um ein Bit reduziert werden kann [Mus15].

2.3.8 *Quantisierung*

Unter Quantisierung versteht man die Abbildung $Q : U \subset \mathbb{R}^n \rightarrow \{x_0, \ldots, x_{N-1}\} \subset \mathbb{R}^k$ mit $n, k \in \mathbb{N}$. Die x_i werden Codevektoren genannt und bilden zusammen das sogenannte Codebuch eines Quantisierers. Ein Quantisierer ist wiederum ein System, welches eine Quantisierung durchführt [GG91]. Man spricht von skalarer Quantisierung falls $n = 1$ ist und von Vektorquantisierung falls $n > 1$ ist. Meistens gilt $n = k$. $N \in \mathbb{N}$ bezeichnet die Größe des Codebuchs des Quantisierers bzw. die Anzahl der Codevektoren oder Quantisierungsstufen. Typisch für die Teilmenge U ist $U = \mathbb{R}^n$ sowie $(n = 1)$ $U = [a, b]$ mit $b > a$. Durch Quantisierung kommt es im Allgemeinen zu einem irreversiblen Informationsverlust und der Quantisierer ist das Element in einem Kompressionsalgorithmus, das den eigentlichen Informationsverlust herbeiführt.

Ziel beim Design eines Quantisierers ist die Minimierung des Ausdrucks

$$\mathbf{E}(d(Q(\mathbf{X}), \mathbf{X}) + \lambda H(Q(\mathbf{X}))) \tag{2.43}$$

mit $\lambda \geqslant 0$, d.h. die Minimierung einer gewichteten Summe der erwarteten Verzerrung $\mathbf{E}(d(Q(\mathbf{X}), \mathbf{X}))$ und des erwarteten Informationsgehalts $H(Q(\mathbf{X}))$. Für $\lambda = 0$ und $d(Q(\mathbf{X}), \mathbf{X}) = \mathbf{E}((Q(\mathbf{X}) - \mathbf{X})^2)$ sowie $n = 1$ wird der optimale (skalare) Quantisierer Lloyd-Max-Quantisierer genannt. Diesen findet man iterativ durch Anwendung des Lloyd-Max-Algorithmus, dessen Algorithmus aus einer abwechselnden Neuberechnung der Codevektoren $\{x_0, x_1, \ldots, x_{N-1}\}$ mittels

$$x_k^{i+1} = \frac{\int_{I_k^i} x f(x)}{\int_{I_k^i} f(x)} \tag{2.44}$$

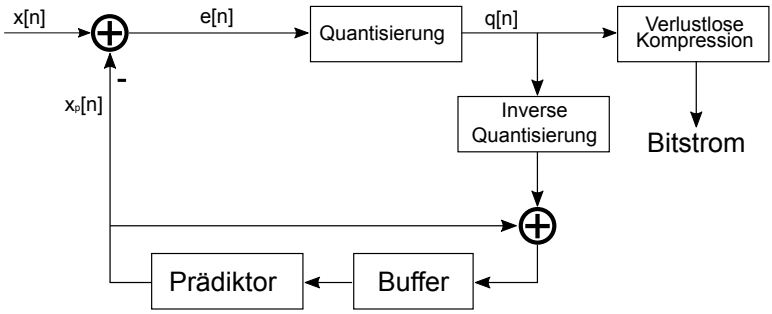


Abbildung 2.11: Blockschaltbild des Encoders einer Differential Puls-Code Modulation. Aus dem Eingangssignal $x[n]$ wird mit der Prädiktion $x_p[n]$ des Prädiktors der Prädiktionsfehler $e[n] := x[n] - x_p[n]$ berechnet. Dieser Prädiktionsfehler wird anschließend quantisiert, wodurch die Quantisierungsindizes $q[n]$ entstehen. Diese werden nachfolgend für eine optimale Kompression verlustlos komprimiert. Für eine konstante Bitrate kann die verlustlose Kompression der Quantisierungsindizes $q[n]$ entfernt werden.

sowie

$$I_k^{i+1} = [\frac{x_{k-1}^{i+1} + x_k^{i+1}}{2}, \frac{x_k^{i+1} + x_{k+1}^{i+1}}{2}] \tag{2.45}$$

besteht, wobei für $k = 0$ bzw. $k = N - 1$ die untere bzw. obere Intervallgrenze auf $-\infty$ bzw. $+\infty$ gesetzt wird [GG91]. Gestartet werden kann z.B. mit (mit auf die Ränder) gleich großen Intervallen. In praktischen Anwendungen hat man eine Stichprobe von Realisierungen gegeben, mit welchen man die Integrale in Gl. 2.44 schätzen muss. Analog geht man für den Fall $n > 1$ (Vektorquantisierung) vor.

2.3.9 *Differential Puls-Code Modulation*

Die Differential Puls-Code Modulation (DPCM) ist ein verlustbehaftetes Kompressionsverfahren und ist aus einem Prädiktor und einem Quantisierer aufgebaut [Say96]. Das Blockschaltbild des Encoders ist in Abb. 2.11 dargestellt. Der Prädiktor prädiziert das Eingangssignal $x(n)$ aus vergangenen, rekonstruierten Abtastwerten $\hat{x}(n-i)$ und es wird der Prädiktionsfehler $e(n) = x(n) + P(\hat{x}(n-1), \ldots, \hat{x}(n-L))$ berechnet, wobei $P(\hat{x}(n-1), \ldots, \hat{x}(n-L))$ die Prädiktion des Prädiktors zum Zeitpunkt n ist. Der Prädiktionsfehler wird dann vom Quantisierer quantisiert, wodurch der quantisierte Wert $\hat{e}(n)$ entsteht. Aus diesem wird dann das Signal rekonstruiert durch Addition der Prädiktion des Prädiktors, d.h.

$$\hat{x}(n) = \hat{e}(n) + P(\hat{x}(n-1), \ldots, \hat{x}(n-L)). \tag{2.46}$$

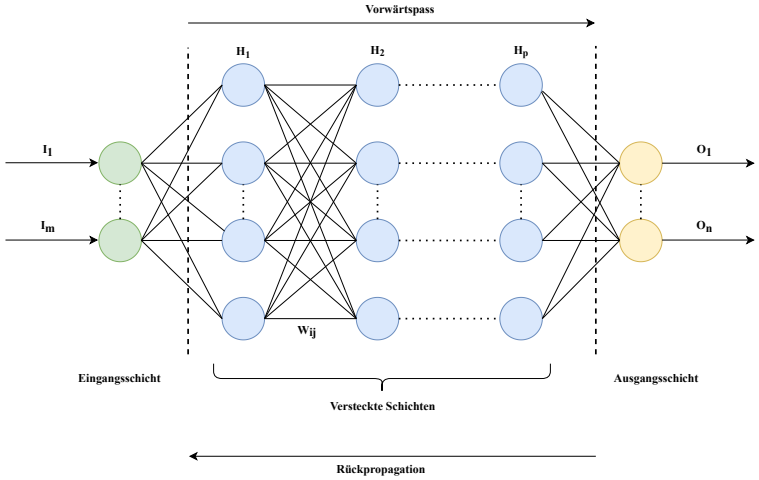


Abbildung 2.12: Schematischer Aufbau eines Vorwärtsnetzwerkes, d.h. eines künstlichen neuronalen Netzes ohne Rückführungen. Alle Signale fließen von links nach rechts.

Dem Decoder werden die Quantisierungsindizes $q(n)$ übermittelt, welche zu den quantisierten Prädiktionsfehlern $\hat{e}(n)$ gehören. Damit, zusammen mit einem identischen inversen Quantisierer, ist es dem Decoder möglich durch identische Prädiktion wie im Encoder das Ursprungssignal (verlustbehaftet) zu rekonstruieren.

Ein wesentlicher Grund für die Attraktivität der DPCM, insbesondere für diese Arbeit, ist die Möglichkeit, verzögerungsfrei zu komprimieren und zu dekomprimieren. Da der Prädiktor nur vergangene Werte zur Prädiktion verwendet und es möglich ist, den Prädiktor auf Basis vergangener, rekonstruierter Werte zu aktualisieren und zudem die Quantisierung keine Verzögerung bewirkt, kann die DPCM insgesamt ohne algorithmische Latenz betrieben werden. Die Notwendigkeit besonders niedriger Latenz der Signalverarbeitung für Cochlea-Implantatträger motivierte daher den Einsatz einer DPCM zur Codierung.

2.3.10 *Neuronale Netze und Autoencoder*

Künstliche neuronale Netze stellen einen weitverbreiteten Ansatz zur Parametrisierung stetiger Funktionen dar. Sie bestehen aus Verkettungen von Schichten sogenannter künstlicher Neuronen. Ein schematischer Ausschnitt eines sogenannten Vorwärtsnetzwerks ist in Abb. 2.12 dargestellt.

Die Eingangs-Ausgangs-Beziehung eines Neuron lautet [Agg18]

$$\mathrm{out}(x) = \sigma(x - b), \tag{2.47}$$

wobei b der sogenannte Biaswert des Neurons und $\sigma(x)$ die sogenannte Aktivierungsfunktion ist. Verbreitete Aktivierungsfunktionen sind die rectified linear unit (Relu), definiert als $max\{0, x\}$, Swish, definiert als $\frac{x}{1+e^{-x}}$, sowie tangens hyperbolicus, definiert als $\frac{e^x - e^{-x}}{e^x + e^{-x}}$.

Das Eingangssignal x eines Neurons besteht dabei im Allgemeinen aus einer gewichteten Summe von Ausgangssignalen out_i vorheriger Neuronen, d.h. es ist $x = \sum_{i=1}^{N} w_i \mathrm{out}_i$. Insgesamt ergibt sich das Ausgangssignal $\mathrm{out}(x)$ eines Neurons, gegeben N Eingangssignale out_i vorheriger Neuronen, als

$$\mathrm{out}(x) = \sigma(\sum_{i=1}^{N} w_i \mathrm{out}_i - b). \tag{2.48}$$

Allgemeiner lässt sich die Ausgabe des j-ten Neurons einer Schicht des künstlichen neuronalen Netzes als

$$\mathrm{out}_j(x) = \sigma(\sum_{i=1}^{N} w_{ij} \mathrm{out}_i - b_j). \tag{2.49}$$

notieren, wobei das Gewicht w_{ij} das i-te Neuron der vorherigen Schicht mit dem j-ten Neuron der betrachteten Schicht verbindet respektive diese Verbindung gewichtet.

Die Gesamtheit dieser Gewichte sowie die Gesamtheit dieser Biaswerte sowie die Aktivierungsfunktion eines jeden Neurons bestimmen, welche Funktion das künstliche neuronale Netz ausführt.

Um die Gewichte und Biaswerte eines künstlichen neuronalen Netzes so zu wählen, dass eine gewünschte Funktion ausgeführt wird, ist ein Abstandsmaß notwendig, die sogenannte Verlustfunktion (engl. loss function oder auch nur loss) $\mathcal{L}$, welche eine Funktion des gewünschten Ausgangswerts y und des tatsächlichen Ausgangswerts $\hat{y}$ des künstlichen neuronalen Netzes ist. Sie misst, wie gut ein künstliches neuronales Netz die gewünschte Funktion ausführt. Ziel des sogenannten Trainings eines künstlichen neuronalen Netzes ist nun, die Gewichte so zu wählen, dass die Verlustfunktion $\mathcal{L}$ möglichst kleine Werte annimmt [Agg18].

Da $\hat{y} \equiv \hat{y}(\mathbf{x}, \mathbf{w})$ ist, also vom Eingangswert und den Gewichten[9] $\mathbf{w}$ abhängt, hängt auch die Verlustfunktion davon ab, d.h.

$$\mathcal{L} \equiv \mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}(\mathbf{x}, \mathbf{w})) \equiv \mathcal{L}(\mathbf{y}, \mathbf{x}, \mathbf{w}). \tag{2.50}$$

9 Die Biaswerte werden an dieser Stelle der Einfachkeit halber darunter subsummiert

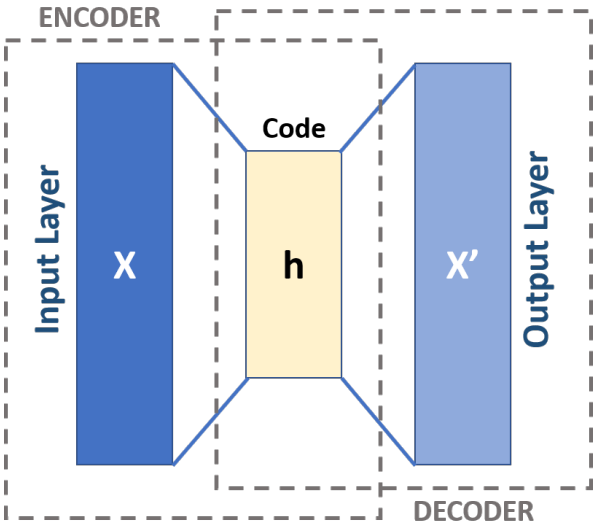


Abbildung 2.13: Grundstruktur eines Autoencoders bestehend aus Encoder, verjüngter verborgener Schicht oder Raum und dem Decoder. Abbildung aus [Mas19].

Ist $\mathcal{L} \in C^1(\mathbb{R}^N)$, d.h. stetig differenzierbar, so lässt sich ein künstliches neuronales Netz prinzipiell mittels Gradientenabstieg trainieren und die Gewichte können prinzipiell iterativ mittels [Agg18]

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \alpha \nabla_{\mathbf{w}} \mathcal{L} \tag{2.51}$$

bestimmt werden. α nennt man die Lernrate und sie kontrolliert die Schrittweite des Gradientenabstiegs. $\nabla_{\mathbf{w}}$ ist der Gradient bezüglich der Gewichte $\mathbf{w}$. Auf Grund der Schichtstruktur und des Aufbaus der künstlichen Neuronen ist $\hat{y}$ eine Verkettung von Funktionen der Art $\mathrm{out}(x) = \sigma(\sum_{i=1}^{N} w_i \mathrm{out}_i - b)$. Die Anwendung der Kettenregel der Differentialrechnung offenbart, dass die Ableitungen in Gl. 2.51 miteinander verknüpft sind. Dies kann zur schnellen Berechnung des Gradienten genutzt werden und ist der Kern des sogenannten Rückpropagationsalgorithmus, welcher in der Praxis typischerweise zur Beschleunigung des Trainings eines künstlichen neuronalen Netzes verwendet wird.

2.3.10.1 *Autoencoder*

In dieser Arbeit wurden Autoencoder zur Kompression der Erregungsmuster von Cochlea-Implantaten verwendet. Dies sind künstliche neuronale Netze mit einer um die mittlere Schicht, den sogenannten verborgenen Raum (engl. latent space), symmetrischen Struktur, bei denen die mittlere Schicht weniger Neuronen enthält als die Eingangs- und Ausgangsschicht [Agg18]. Diese Struktur ist schematisch in Abb. 2.13 dargestellt. Verwendet man dann eine Verlustfunktion, bei der Identität von Eingang- und Ausgangwerten den minimalen Wert der Verlustfunktion ergeben, so muss der Autoencoder im Allgemeinen einen Kompressionsalgorithmus erlernen, da in der mittleren versteckten Schicht[10] weniger Bits zur Darstellung existieren als am Eingang. Aufgrund dieser Tatsache sind Autoencoder ein beliebtes Werkzeug der Datenkompression [YMT22].

2.3.10.2 *Rückkopplungsautoencoder*

Ein Autoencoder überführt gegebene Eingangsdaten in eine verborgene Darstellung im verborgenen Raum und rekonstruiert aus dieser näherungsweise diese Eingangsdaten. Wird eine Sequenz von Eingangsdaten komprimiert, so sind sequentielle statistische Abhängigkeiten vom Autoencoder nicht für die Verbesserung der Kompression nutzbar. Da es sich bei den Stimulationsmustern der Cochlea-Implantate um Signale insbesondere mit einer zeitlichen Komponente handelt, ist ein Autoencoder wie zuvor beschrieben suboptimal[11]. Für den Fall der Kompression sequentieller Daten wurde in [Yan+19] ein sogenannter Rückkopplungsautoencoder vorgestellt. Dessen Blockdiagramm ist in Abb. 2.14 dargestellt. Anders als bei normalen Autoencodern werden vorherige, decodierte Frames $\hat{x}_{t-1}, \hat{x}_{t-2}, \ldots$ der Encoder- und Decoderschicht des Autoencoders zugeführt, sodass bei der Kompression und Dekompression statistische Abhängigkeiten berücksichtigt werden können und somit die notwendige Informationsmenge im verborgenen Raum reduziert werden kann. Die notwendige Bitrate lässt sich also verkleinern im Vergleich zu einem Autoencoder ohne Rückkopplung. Insgesamt kann die Struktur des Rückkopplungsautoencoders als eine nichtlineare, prädiktive Codierung angesehen werden.

10 Versteckte Schichten sind jene Schichten eines künstlichen neuronalen Netzes, die sich zwischen Eingangs- und Ausgangsschicht befinden.

11 Sofern latenzfrei, also ohne Blockbildung, codiert wird.

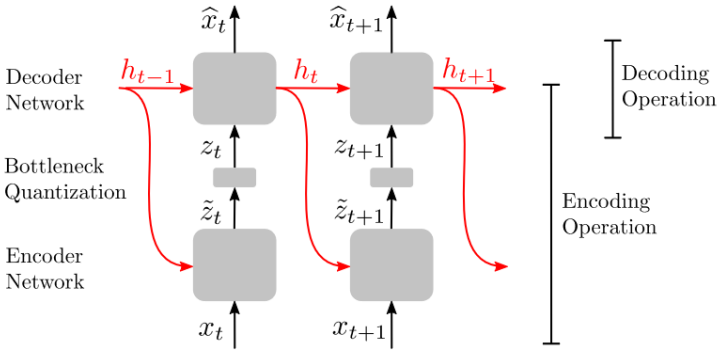


Abbildung 2.14: Blockdiagramm der Grundidee eines Rückkopplungsautoencoders. Für die Codierung im Zeitschritt t wird eine Zusammenfassung $h_{t-1} := (\hat{x}_{t-1-K}, \ldots, \hat{x}_{t-1-1})$ der letzten K decodierten Frames zum Encoder und Decoder zurückgeführt. Hierdurch lassen sich sequentielle Redundanzen ausnutzen. Abbildung angepasst aus [Yan+19].

2.3.11 *Audiocodierung: Stand der Technik*

Da die Kompression der Erregungsmuster als Alternative zur Kompression von Audiosignalen entwickelt und untersucht wurde, mussten im Rahmen der vorgelegten Arbeit Vergleiche mit modernen Audiocodecs vorgenommen werden.

Während es grundsätzlich eine riesige Zahl an Algorithmen zur Codierung von Audiosignalen gibt, welche oft zusätzlich in Sprach- und Musikcodecs unterteilt werden, ist die Zahl an Audiocodecs, die eine hinreichend niedrige algorithmische Latenz aufweisen, um einen sinnvollen Vergleich zu den in der vorgelegten Arbeit entwickelten Kompressionsalgorithmen darzustellen, sehr gering. Des Weiteren nutzen Cochlea-Implantate eine Eingangsabtastfrequenz von 16 kHz für die Audiosignale. Ohne Weiteres nutzbare Audiocodecs sollten also diese Abtastfrequenz unterstützen.

Ein mittlerweile nicht mehr moderner, jedoch dennoch aufgrund seiner geringen algorithmischen Latenz von etwa einer Millisekunde und Patentfreiheit verbreiteter Audiocodec ist der G.722 [BFM10]. Dieser zerlegt das zu komprimierende Audiosignal in zwei Frequenzbänder und komprimiert diese dann mittels prädiktiver Codierung. Standardmäßig erzielt der G.722 eine Bitrate von 64 kbit/s, kann jedoch mit speziellen Einstellungen bei geringerer Qualität mit bis zu 48 kbit/s betrieben werden.

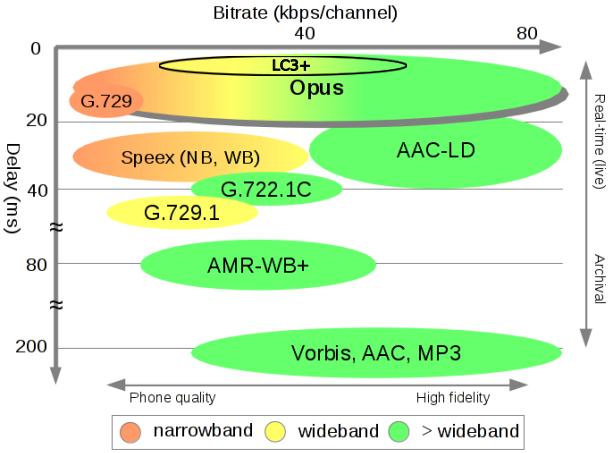


Abbildung 2.15: Ein (informeller) Vergleich von Opus mit Bezug auf algorithmische Latenz (Delay) sowie Bitrate mit anderen bekannten Audiocodecs adaptiert aus [IET]. Für den LC3+ Codec wurde der für die Arbeit relevante, ungefähre Arbeitsbereich ergänzt.

Der G.722 Audiocodec ist der einzige Audiocodec, welcher nachweislich [Oti21] beim drahtlosen Streaming von Audio für Cochlea-Implantate Anwendung findet. Sehr bekannte und verbreitete Audiocodecs sind der Enhanced Voice Service (EVS) Codec [Die+15], der für Internettelefonie entwickelt wurde, der Opus Audiocodec [Val+16], welcher unter anderem von WhatsApp verwendet wird, sowie in jüngster Zeit der Low Complexity Communication Codec (LC3) [Sch+21] bzw. LC3 Plus, der für Bluetooth und (Internet-)Telefonie entwickelt wurde. Der EVS-Codec besitzt jedoch eine algorithmische Latenz von 32 ms, was deutlich zu hoch für einen sinnvollen Einsatz für das drahtlose Streamen von Audio für Cochlea-Implantate ist. Der LC3+-Codec ist dafür grundsätzlich nutzbar mit einer minimalen algorithmischen Latenz von 5 ms [Eur18]. Die minimale Bitrate beträgt bei dieser Einstellung allerdings 64 kbit/s, was vergleichsweise hoch ist. Sie sinkt auf 32 kbit/s bei einer algorithmischen Latenz von 7,5 ms und auf 16 kbit/s bei einer algorithmischen Latenz von 12,5 ms, einer für den Anwendungsfall inakzeptabel hohen Latenz.

Ferner handelt es sich des Weiteren um einen kommerziellen Codec, welcher nicht frei verfügbar ist. Daher war es nicht möglich, im Rahmen der vorgelegten Arbeit diesen ohne Weiteres zu testen. Ferner war der Codec zur Durchführungszeit der im Rahmen der vorgelegten Arbeit durchgeführten Probandenstudie noch nicht öffentlich bekannt. Daher

konnten keine Tests an Probanden durchgeführt werden, weswegen er auch im späteren Verlauf der Forschungsarbeit nicht betrachtet wurde.

Der Opus-Audiocodec wurde 2012 in der ersten Version veröffentlicht und mit den Jahren weiter verbessert. Die in dieser Arbeit verwendete Version 1.3.1 wurde im Jahre 2019 veröffentlicht und stellt die aktuellste veröffentlichte Version dar. Opus verwendet eine Kombination von sogenannter Constrained Energy Lapped Transform (CELT) und dem Sprachcodec SILK [Val+16]. Opus ist extrem flexibel, sodass Audiosignale mit Abtastfrequenzen zwischen 8 kHz und 48 kHz, algorithmischen Latenzen zwischen 5 ms und 66,5 ms sowie Bitraten zwischen 6 kbit/s und 510 kbit/s codiert werden können. Diese sehr große Flexibilität, insbesondere in der Bitrate, zusammen mit der Möglichkeit mit sehr geringer algorithmischer Latenz zu codieren, ließ die Wahl als Hauptvergleichscodec im Rahmen der vorgelegten Arbeit auf den Opus-Audiocodec fallen. Insbesondere in den durchgeführten Hörtests war die Möglichkeit, die Bitrate relativ variabel einstellen zu können, von sehr großem Nutzen, wie bei der Beschreibung der Hörtests noch erläutert werden wird. Eine informelle Einordnung von Opus im Bezug zu einigen Vergleichscodecs ist in Abb. 2.15 dargestellt.

Vor kurzem wurde ein sogenannter neuraler Audiocodec [Déf+22] von Meta vorgestellt, der unter Verwendung künstlicher neuronaler Netze sehr geringe Bitraten um 3 kbit/s bei hoher Audioqualität erreichen soll. Jedoch weist dieser Codec, ebenso wie sein Konkurrent Lyra-2 [Goo22] von Google, eine algorithmische Latenz von mehr als[12] 13 ms bei einer Abtastfrequenz von 16 kHz auf, weswegen dieser Codec, ebenso wie Lyra-2, im Rahmen der vorgelegten Arbeit nicht berücksichtigt wurde.

2.3.12 *Stochastic Perturbation Simultaneous Approximation*

Zum Training von Autoencoderstrukturen zusammen mit nichtableitbaren Verlustfunktionen wurde auf die numerische Approximation von Gradienten mittels des sogenannten Stochastic Perturbation Simultaneous Approximation (SPSA)-Algorithmus [Spa92; CDP99], welcher zur Klasse der Kiefer-Wolfowitz-Algorithmen gehört, zurückgegriffen. Sucht man das Maximum $\underline{\omega} \in \mathbb{R}^N$ einer Funktion $f : \mathbb{R}^N \to \mathbb{R}$, so lautet die Iterationsgleichung des SPSA-Algorithmus

$$\underline{\omega}_{k+1} = \underline{\omega}_k + a_k \frac{(y^+_{k+1} - y^-_{k+1})}{c_k} \Delta_k \tag{2.52}$$

12 Tatsächlich müsste sich diese auf 20 ms belaufen.

mit $y_{k+1}^{\pm} = f(\underline{\omega}_k \pm c_k \Delta_k)$, binomialverteiltem Rauschen $\Delta_k \in \{-1, 1\}^N$ mit $\mathbb{P}(\Delta_k^i = \pm 1) = 0,5$, $a_k, c_k > 0$ sowie Folgen $(a_k)_{k\in\mathbb{N}}, (c_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ mit $a_k, c_k \to 0$. In dieser Arbeit wurde $a_k = \frac{a}{(A+k+1)^\gamma}$ mit $a = 1$ und $\gamma = 0,602$ sowie $c_k = \frac{c}{(k+1)^\beta}$ mit $\beta = 0,101$ verwendet. Die Parameter A und c wurden mittels Hyperparameteroptimierungsverfahren bestimmt. In konkreten Anwendungen in dieser Arbeit handelt es sich bei f um die Bewertung der Ausgabe eines Autoencoders mit und ohne zusätzliche Quantisierung durch das Short-Time Objective Intelligibility Measure (STOI). Insgesamt stellt dies eine nicht ableitbare Funktion dar.

2.3.13 *Sequential Model-Based Algorithm Configuration*

Die Hyperparameter von Algorithmen, insbesondere im Bereich des maschinellen Lernens, haben im Allgemeinen einen sehr großen Einfluss auf die Leistungsfähigkeit des Algorithmus. Ist die Lernrate beim Training eines künstlichen neuronalen Netzes zu klein oder zu groß gewählt, wird die Leistungsfähigkeit mangelhaft bleiben. Im Rahmen dieser Arbeit wurden Autoencoder für die Kompression von Erregungsmuster verwendet. Initial war die Leistungsfähigkeit der Autoencoder mangelhaft, da unter anderem keine sinnvolle Struktur für die Kompression von Erregungsmustern bekannt war. Nachdem eine kleine Gittersuche nicht erfolgreich war, wurde auf ein dediziertes Framework für die automatische Hyperparameteroptimierung zurückgegriffen, das sogenannte Sequential Model-Based Algorithm Configuration (SMAC) [Lin+21]. SMAC nutzt Bayessche Optimierung zur Optimierung von Blackbox Kostenfunktionen. SMAC ist insbesondere dann attraktiv, wenn ein Algorithmus relativ lange zur Ausführung benötigt, wie es oftmals beim Training eines künstlichen neuronalen Netzes der Fall ist. Mittels SMAC war es möglich innerhalb von nur etwa 50 Iterationen deutlich verbesserte Autoencoderstrukturen zu finden.

2.4 VERFAHREN DER STATISTIK

Zum Verständnis der statistischen Auswertung der durchgeführten Hörtests werden in diesem Abschnitt Hypothesentets sowie Varianzanalyse der Wilcoxon-Rangtest erläutert.

2.4.1 *Hypothesentests*

Hypothesentests, oder allgemeiner, statistische Tests, dienen der Ablehnbarkeit von Annahmen über den Wertebereich von statistischen

Parametern. Ein Standardbeispiel ist das Vorliegen zweier Stichproben $(x_i^k)_{i=1,\ldots,N,k=1,2}$, die aus zwei Zufallsexperimenten stammen, für welche entschieden werden soll, ob die Erwartungswerte μ_k gleich sind. Dies lässt sich mittels $\theta := \mu_1 - \mu_2$ umformulieren, als Prüfung von $\theta = 0$. Zur Beantwortung ist es notwendig, eine sogenannte Teststatistik zu konstruieren, nach welcher entschieden wird, ob die sogenannte Nullhypothese H_0 abgelehnt werden kann oder nicht. Im Falle der Ablehnung geht man von Wahrheit der Alternativhypothese H_1 aus. Es gibt keine allgemeine Regel, welche Fälle einer zu klärenden Frage als H_0 und welche als H_1 zu definieren sind [Beh13]. Bei Hypothesentests kann es zu vier Fällen kommen: Ist H_0 wahr und wird dann die Nullhypothese nicht abgelehnt (d.h. angenommen), so ist alles in Ordnung. Wird diese jedoch abgelehnt, so spricht man von einem falsch negativen Ergebnis und einem Fehler 2. Art. Ist H_0 nicht wahr und wird H_0 nicht abgelehnt (d.h. angenommen), so kommt es zu einem Fehler 1. Art. Anderenfalls wird korrekt entschieden. Ein idealer Test sollte nun die Wahrscheinlichkeiten α und β für einen Fehler 1. Art und 2. Art gegeben eine Stichprobe gleichzeitig minimieren. α wird auch Signifikanzniveau des Tests genannt. Trivial kann man einen Test entwerfen, welcher α (exklusiv) oder β auf Null bringt, indem man unabhängig von der vorgelegten Stichprobe immer H_0 annimmt oder ablehnt. Dies ist aber selten sinnvoll[13]. Das Neyman-Pearson Lemma weist die Existenz und Eindeutigkeit eines besten Tests nach, der zu einer gegebenen Wahrscheinlichkeit α eines Fehlers 1. Art die Wahrscheinlichkeit β eines Fehlers 2. Art minimiert [Beh13].

In dieser Arbeit wurde sogenannte Varianzanalyse sowie der Wilcoxon Rangtest verwendet, um zu entscheiden, ob die mittlere Sprachverständlichkeit von Gruppen von Probanden identisch ist oder nicht. Auf diese speziellen Tests wird in den nächsten beiden Abschnitten eingegangen werden.

Zum Vergleich verschiedener Codierungsstrategien musste nicht nur ein einzelner statistischer Test durchgeführt werden, sondern $N = 14$. Hat jeder dieser Tests eine Wahrscheinlichkeit für einen Fehler 1. Art von α, so ergibt sich eine Gesamtwahrscheinlichkeit P_{α_N} keinen Fehler 1. Art in den N Tests zu begehen von $(1-\alpha)^N$. Das heißt, mit zunehmender Zahl an Vergleichen bzw. Tests steigt unweigerlich die Wahrscheinlichkeit, einen Fehler 1. Art begangen zu haben. Um diesem Wachstum entgegenzuwirken, gibt es verschiedene Methoden. Eine weitverbreitete und sehr einfache ist die sogenannte Bonferronikorrektur. Bei dieser wird das Signifikanzniveau α mittels Division durch die Zahl an Vergleichen justiert, d.h. das neue Signifikanzniveau ist dann $\alpha_{Neu} = \frac{\alpha}{N}$. Ist α nämlich hinrei-

13 Eine Ausnahme wären Feuermeldungen bei der Feuerwehr.

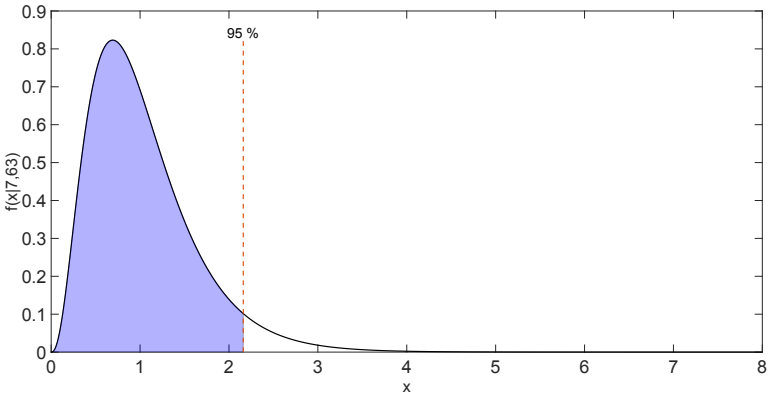


Abbildung 2.16: Wahrscheinlichkeitsdichtefunktion $f(x|6,67)$ der F-Verteilung mit den Freiheitsgraden 6 und 67, die im Rahmen der Hörtests implizit in Hypothesentests genutzt wird. Die gestrichelte Linie markiert die 95 % Grenze, d.h. 95 % aller Realisierungen liegen linksseitig dieser Linie.

chend klein, so ist $(1-\alpha)^N \approx 1 - N\alpha$ - eine Konsequenz des binomischen Lehrsatzes, wodurch sich die Bonferronikorrektur unmittelbar ergibt.

2.4.2 *Varianzanalyse*

Die Varianzanalyse (engl. Analysis of Variance (ANOVA)) dient der Untersuchung von Gruppenunterschieden [Laro8]. Grundsätzliche Idee ist, dass man, gegeben Stichproben S_i, durch das Verhältnis von Intragruppenvarianzen zur Intergruppenvarianz Aussagen über die Signifikanz von Unterschieden zwischen den Stichproben (typischerweise hinsichtlich der Erwartungswerte) treffen kann.

Es seien K Gruppen mit Erwartungswerten $\mu_1, \ldots, \mu_K$ sowie Beobachtungen $y_{ij} = \mu_j + \epsilon_{ij}$ mit $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ gegeben. Die Nullhypothese der Varianzanalyse ist nun

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K,$$

d.h. Gleichheit der Gruppenerwartungswerte. Die Alternativhypothese geht von der Verschiedenheit mindestens eines Erwartungswerts aus.

Die Teststatistik T der Varianzanalyse ist

$$T := \frac{MS_{Zw}}{MS_{In}} \tag{2.53}$$

mit

$$MS_{Zw} := \frac{\sum_{i=1}^{K} n_i(\overline{y}_i - \overline{y})}{K-1}$$

und

$$MS_{In} := \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)}{N-K}$$

mit $N = \sum_{i=1}^{K} n_i$ der Gesamtzahl an Probanden. Es ist $\overline{y}_j = \frac{\sum_{i=1}^{n_i} y_{ij}}{n_j}$ sowie $\overline{y} = \frac{\sum_{j=1}^{K} \overline{y}_j}{K}$.

Es lässt sich zeigen, dass $T \sim F(K-1, N-k)$ gilt, wobei F die sogenannte F-Verteilung ist, welche die Dichtefunktion

$$f(x \mid m, n) = m^{\frac{m}{2}} n^{\frac{n}{2}} \cdot \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}}, \quad x > 0 \tag{2.54}$$

besitzt. Für den Fall $F(6, 67)$, der bei der Analyse von im Rahmen der vorgelegten Arbeit durchgeführten Hörtests auftritt, ist die zugehörige Dichtefunktion in Abb. 2.16 dargestellt.

Die Intuition der Varianzanalyse, respektive der Teststatistik T, ist, dass MS_{Zw} ein Schätzer der Varianz σ^2 ist, sofern die Nullhypothese zutrifft, wohingegen MS_{In} ein Schätzer von σ^2 ist unabhängig ob H_0 oder H_1 zutrifft. Trifft H_0 nicht zu, so überschätzt MS_{ZW} den wahren Wert von σ^2 und es ist $T > 1$. Anderenfalls ist $T \approx 1$. Große Werte von T zeigen also tendenziell das Zutreffen der Alternativhypothese an. Genauer lehnt man die Nullhypothese H_0 zum 95 % Signifikanzniveau ab, sofern T rechts der eingezeichneten Senkrechten in Abb. 2.16 liegt, d.h. wenn gilt

$$\int_0^T f(x \mid K-1, N-k)dx > 0,95. \tag{2.55}$$

Analoges gilt für beliebige Signifikanzniveaus.

2.4.3 *Wilcoxon-Vorzeichen-Rangtest*

Gegeben Paare $(x_i, y_i) \in \mathbb{R}^2$, welche als Messergebnisse im i-ten Versuch einer Studie interpretiert werden können. x_i seien Realisierungen einer Zufallsvariablen $\mathbf{X}$, y_i seien Realisierungen einer Zufallsvariablen $\mathbf{Y}$. Der (zweiseitige) Wilcoxon-Vorzeichen-Rangtest prüft nun, ob die Mediane $\tilde{X}$ und $\tilde{Y}$ von $\mathbf{X}$ sowie $\mathbf{Y}$ identisch sind, d.h. ob $\tilde{X} = \tilde{Y}$ gilt. Das Vorgehen ist

hierbei wie folgt: Zunächst definiert man die Differenzen $d_i := x_i - y_i$ sowie den Rang[14] $r_i := \text{Rang}(|d_i|)$ der absoluten Differenzen.

Die Teststatistik $W := \min(W_-, W_+)$ mit $W_- := \sum_{i=1}^{N} I(x_i - y_i < 0) r_i$ sowie $W_+ := \sum_{i=1}^{N} I(x_i - y_i > 0) r_i$ ist dann approximativ normalverteilt für hinreichend großes N. Hierbei ist I die Indikatorfunktion. Im Falle der Gleichheit von Paaren (x_i, y_i) kommt es zu Korrekturen. Typischweise werden diese Paare im Test ignoriert, oder aber sie werden zu jeweils 50 % W_- und W_+ zugewiesen [Jea03]. Die Nullhypothese wird abgelehnt, sofern W einen kritischen Wert, der vom Signifikanzniveau α abhängt, unterschreitet.

Ein schulbuchmäßiges Vorgehen bei der Analyse von Untersuchungsergebnissen ist, zunächst die Ergebnisse mittels einer Varianzanalyse auf das Vorliegen von Gruppenunterschieden zu untersuchen. Dies liefert einem lediglich die Erkenntnis, ob es irgendeinen Unterschied gibt, aber nicht, zwischen welchen Gruppen. Nachfolgend wird dann bei einem festgestellten signifikanten Gruppenunterschied der Wilcoxon-Vorzeichen-Rangtest (oder vergleichbare Methoden) angewendet, um sukzessive eine Teilmenge der untersuchten Gruppen auf statistisch signifikante Unterschiede ihrer Mediane zu untersuchen.

14 Der Rang eines Werts x_k aus einer Stichprobe $x_1, \ldots, x_N$ ist im einfachsten Fall ohne Gleichheit definiert als $\text{Rang}(x_k) := \sum_{i=1, i\neq k}^{N} u(x_k - x_i)$ mit $u(x) = 1, x \geqslant 0$ und $u(x) = 0$ sonst. Der Rang ist also identisch mit der Platzierung eines Werts, wenn die Werte einer Stichprobe der Größe nach von unten nach oben geordnet sind.

3

ENTWICKELTE CODIERUNGSSTRATEGIEN, DATENSÄTZE SOWIE BESCHREIBUNG DES HÖRTESTS

Dieses Kapitel stellt die im Rahmen der vorgelegten Arbeit entwickelten Codierungsverfahren vor, erläutert die Generierung sowie Eigenheiten der genutzten Datensätze und schließt mit der Konzeption der durchgeführten Hörtests. Es wurden im wesentlichen zwei Ansätze zur Codierung der Erregungsmuster verfolgt: Zuerst wurde ein konventioneller Kompressionsalgorithmus auf Basis von Differential Puls-Code Modulation (DPCM) und arithmetischer Codierung entworfen, der sogenannte Electrocodec [Hin+19]. Dieser wurde in Hörtests mit Trägern von Cochlea-Implantaten untersucht [Hin+21a]. Anschließend wurde versucht, mittels künstlicher neuronaler Netze einen Codec zu entwickeln respektive zu lernen, welcher die Bitrate im Vergleich zum Electrocodec bei gleicher oder näherungsweise gleicher Verständlichkeit der codierten Erregungsmuster senken kann. Hierbei wurde zum einen ein verlustloser Codec entwickelt und untersucht [Hin+22a], und zum anderen mittels Autoencodern ohne [HOO22] und mit [HBO23] Rückkopplung die Kompression im Vergleich zum Electrocodec verbessert.

3.1 DER ELECTROCODEC

Der erste entwickelte Ansatz zur Codierung der Erregungsmuster, nachfolgend Electrocodec genannt, basierte auf der Verwendung von DPCMs für die Kompression der logarithmierten Einhüllenden gemäß Gl. 2.7 verbunden mit verlustloser Kompression der Bandselektion. Es wurden die logarithmierten Einhüllenden gemäß Gl. 2.7 und nicht die Stromwerte gemäß Gl. 2.8 codiert, da letztere von vielen individuell angepassten Parametern, den THR- und MCL-Werten je Band, abhängen, anders als die logarithmierten Einhüllenden. Durch die vorgeschaltete Bandselektion profitiert die Codierung der logarithmierten Einhüllenden bereits von ihrer deutlichen Irrelevanzreduktion.

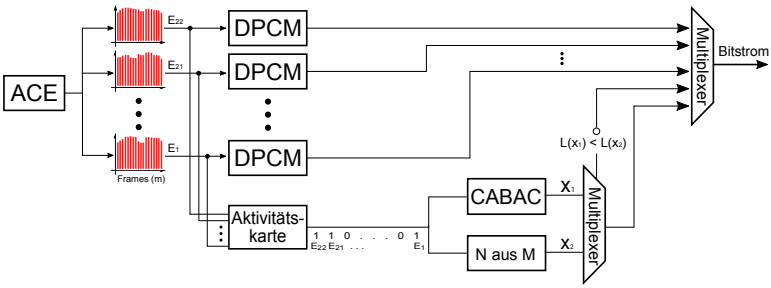


Abbildung 3.1: Struktur des Electrocodecs, welcher im Rahmen der vorgelegten Arbeit zur Kompression der Erregungsmuster von Cochlea-Implantaten entwickelt wurde. In jedem Zeitschritt m wird die Bandselektion der vom Advanced Combination Encoder (ACE) erzeugten Erregungsmuster, hier dargestellt in Rot, in die sogenannte Aktivitätskarte überführt. Die Aktivitätskarte wird dann entweder mit kontextadaptiver arithmetischer Codierung (CABAC) oder unter Ausnutzung der N aus M Eigenschaft von ACE komprimiert. Für jedes selektierte Band wird mittels einer Differential Puls-Code Modulation der zugehörige Stromwert codiert. Die so entstehende komprimierte Darstellung wird dann in der Darstellung nach Abb. 3.3 an den Decoder übermittelt.

Die Prädiktion der DPCMs wurde rückwärtsadaptiv gestaltet, um eine möglichst geringe algorithmische Latenz der Codierung zu erzielen. Die Kompression der Bandselektion erfolgte mit kontextadaptiver artihmetischer Codierung (engl. context-adaptive binary arithmetic coding (CABAC)). Als Kontexte dienten bei der Kompression der Bandselektion ausschließlich bereits codierte Werte aus dem aktuellen Frame. Ein Gesamtüberblick ist in Abb. 3.1 dargestellt.

Die Struktur des Encoders der verwendeten DPCM ist in Abb. 3.2 dargestellt. Anders als eine Standard-DPCM wie in Abb. 2.11 gezeigt wurde eine vorherige Differenzierung des Eingangssignals vorgenommen, in der Abbildung mit ID markiert. Die üblichen Algorithmen zur optimalen Bestimmung der linearen Prädiktorkoeffizienten gehen von einem mittelwertfreien Signal aus. Dies ist aber aufgrund der Abbildungsvorschrift nach Gl. 2.7 nicht der Fall für die zu komprimierenden Erregungsmuster. Ignoriert man diese Tatsache und verwendet Standardalgorithmen zur Berechnung der optimalen Koeffizienten, so sind die prädizierten Werte systematisch wesentlich kleiner als die wahren Werte. Zur Lösung wurde auf ein von integrierten autoregressiven Prozessen [BP70] bekanntes Verfahren zurückgegriffen, die Differenzierung eines Prozesses. Gilt $E(\mathbf{X}_n) \approx E(\mathbf{X}_{n+1})$, d.h. ist der Erwartungswert näherungsweise konstant, wobei $\mathbf{X}_n$ ein die Erregungsmuster darstel-

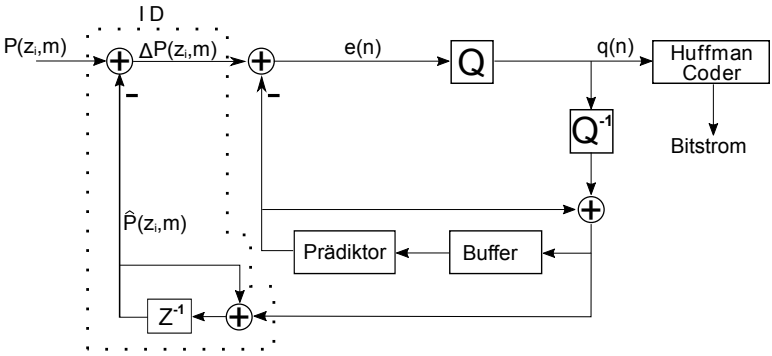


Abbildung 3.2: Encoder der in jedem Subband verwendeten Differential Puls-Code Modulation (DPCM). Die initiale Differenzierung (ID) dient der Entfernung des Mittelwertes. Auf der Differenz wird dann eine Standard-DPCM angewendet. Der Index des quantisierten Prädiktionsfehlers $q(n)$ wird nachfolgend mittels einer Huffmancodierung verlustlos komprimiert. $P(z_i, m)$ ist das Signal nach Gl. 2.7.

lender stochastischer Prozess sei, so folgt mit der Linearität des Erwartungswertes, dass $E(\mathbf{X}_n - \mathbf{X}_{n+1}) \approx 0$ ist. Anschließend können mit üblichen Algorithmen die optimalen Prädiktorkoeffizienten für das Signal $P(z_i, m) - \hat{P}(z_i, m-1)$ berechnet werden. $\hat{P}(z_i, m-1)$ bezeichnet das rekonstruierte Signal der DPCM. Als Quantisierer wurden Lloyd-Max-Quantisierer verwendet, deren Codevektoren vom Sprachmaterial des Sound Quality Assessment Material (SQAM) [Eur] stammen. Hierzu wurde der Electrocodec ohne Quantisierung auf das Sprachmaterial des SQAM angewendet und aus den sich ergebenen Prädiktionsfehlern dann, für jedes Subband separat, mittels des Lloyd-Max-Algorithmus Optimalquantisierer mit 1 Bit bis 8 Bit trainiert, d.h. je Subband standen anschließend acht Quantisierer zur Quantisierung des Prädiktionsfehlers zur Auswahl, mit welchen die Bitrate und Qualität steuerbar ist.

Separat dazu wurden Quantisierer für den ersten Wert einer Pulsfolge trainiert. Aufgrund der Bandselektion von ACE treten immer wieder Frames auf, in denen ein Band neu selektiert wird. Der Wert, der ohne Bandselektion auftreten würde, kann bei der Codierung nicht verwendet werden, da der Decoder keinen Zugriff darauf hat. Daher kann in diesen Fällen keine vorherige Prädiktion durchgeführt werden, weswegen ein separater Quantisierer respektive ein separates Codebuch notwendig ist, das das Intervall $[0, 1]$ abdeckt in denen die $P(z, m)$ liegen. Dieses Codebuch wurde ebenfalls für jedes Subband mittels des SQAM separat trainiert. Hierzu wurde für jedes Band aus den Erregungsmustern je-

Abbildung 3.3: Aufbau des vom Encoder zum Decoder übermittelten Bitstrings. Das Indikatorbit signalisiert, ob die Bandselektion entropiecodiert wurde, oder ob lediglich die N aus M Eigenschaft des Advanced Combination Encoders ausgenutzt wurde. Nachfolgend werden dann die Quantisierungsindizes übermittelt, welche mittels Huffmancodierung komprimiert wurden.

weils der erste Stromwert der Abschnitte mit Erregung des Subbands als Trainingsdatensatz des Lloyd-Max-Algorithmus verwendet. Das SQAM, eine detailliertere Beschreibung folgt in Abschnitt 3.5, wurde für das Training der genannten Komponenten verwendet, da es aus hochqualitativen Aufnahmen besteht. Es deckt den gesamten Frequenzbereich, der vom Cochlea-Implantat berücksichtigt wird, sehr gut ab, sodass möglichst vielfältige Daten für die beschriebene Schätzung von Wahrscheinlichkeiten existieren. Es wurde mit Blick auf die geplanten Hörtests davon abgesehen, verrauschte Signale im Training zu berücksichtigen, damit diesbezüglich kein Vorwurf der Verzerrung im Vergleich zu anderen Codecs gemacht werden kann.

Zur Prädiktion wurde in jedem Subband ein sample-adaptiver linearer Prädiktor mit einer (maximalen) Länge von fünf verwendet. Die Prädiktion gestaltete sich wegen der Bandselektion etwas schwierig. Wird ein Subband neu selektiert, so beginnt die lineare Prädiktion erst mit dem vierten Sample. Die Motivation war hierbei eine Mindeststichprobengröße für die Schätzung der optimalen linearen Prädiktorkoeffizienten zu gewährleisten. Während der ersten drei Zeitschritte nach Neuselektion eines Subbandes wird ausschließlich der rekonstruierte letzte Wert $\hat{P}(z_i, m)$ verwendet. Anschließend beginnt die zusätzliche lineare Prädiktion des Differenzsignals $e_i(m) := P(z_i, m) - \hat{P}(z_i, m-1)$. Die verwendete Bufferlänge wächst je Zeitschritt um eins, wurde jedoch auf den empirisch gefundenen Wert 50 begrenzt, ab welchem sie nicht weiter erhöht wird. Initial, mit Beginn der linearen Prädiktion des Differenzsignals $e(m)$, wird eine Bufferlänge von drei verwendet. Die maximale Bufferlänge, ermittelt für rauschfreie Sprachsignale, ist jedoch bei Vorliegen von Hintergrundrauschen von untergeordneter Bedeutung, da in den allermeisten Fällen deutlich vor Erreichen eben dieser das Subband deselektiert wird und es zum Neustart der Codierung des Subbandes kommt. An den Decoder wird das rekonstruierte Differenzsignal $\hat{e}_i(m) := Q(e_i(m) - P(\hat{e}_i(m-1), \ldots, \hat{e}_i(m-L)))$ übermittelt. Wird ein

Subband im Zeitschritt m deselektiert, jedoch im Zeitschritt $m+1$ erneut selektiert, so wird für den fehlenden Wert $\hat{e}_i(m)$ der prädizierte Wert $P(\hat{e}_i(m-1),\ldots,\hat{e}_i(m-L))$ eingesetzt. Die Codierung läuft dann ganz normal weiter. Ist ein Subband zwei aufeinanderfolgende Zeitschritte m und $m+1$ nicht selektiert, so wird die gesamte Codierung dieses Subbands neugestartet, d.h., dass insbesondere der Buffer des Prädiktors geleert wird. Motivation hierbei ist, dass eine Prädiktion mit zu vielen fehlenden Werten nicht hinreichend akkurat sein kann. Wird das Subband erneut selektiert, so beginnt die Codierung von vorne.

Zur Kompression der Bandselektion wird diese mittels Binärvektoren dargestellt, wobei ein selektiertes Band mit dem Wert '1' und ein nichtselektiertes Band mit dem Wert '0' dargestellt wird. Der Selektionszustand b_i aller Subbänder zum Zeitpunkt m ist also durch einen Vektor der Art

$$AM_m := (b_1^m,\ldots,b_M^m)^T, b_i^m \in \{0,1\} \tag{3.1}$$

zusammengefasst. AM_m, genannt Aktivitätskarte, wird dann mittels arithmetischer Codierung komprimiert, wobei die bedingten Wahrscheinlichkeiten

$$P(b_i^m | b_{i-1}^m,\ldots,b_{i-K}^m) \tag{3.2}$$

mit der Kontextlänge K für die Kompression der i-ten Komponente von AM_m verwendet wird, d.h., die Kontexte sind durch den Selektionszustand vorheriger Subbänder gegeben, welche zuvor codiert wurden und somit auch dem Decoder zur Verfügung stehen. Für Komponenten, für die $i-K<1$ gilt, wurde der verwendete Kontext soweit reduziert, dass sich $i-K=1$ ergab.

Die verwendeten Wahrscheinlichkeiten wurden ebenfalls auf Basis des SQAM-Datensatzes durch die relativen Häufigkeiten geschätzt. Es wurden Kontextlängen $K \in \{0,1,\ldots,12\}$ untersucht, wobei nur der Wechsel von $K=0$ auf $K=1$ eine sehr deutliche Reduktion der Bitrate lieferte. Die verwendeten Kontexte wurden bewusst nur aus dem aktuellen Frame gewählt, um die Interframeabhängigkeiten mit Blick auf Übertragungsfehler in drahtlosen Übertragungssystemen etwas zu reduzieren. Insbesondere eine Verzerrung der Bandselektion kann die Verständlichkeit von Erregungsmuster deutlich reduzieren [Qaz+13], weswegen zum einen die Bandselektion verlustlos codiert wurde und zudem die Abhängigkeiten nur auf Intraframebasis ausgenutzt wurden, sodass bei Verlust eines Frames, wie es im Rahmen einer drahtlosen Übertragung geschehen kann, die Bandselektion in den folgenden Frames dennoch korrekt decodiert werden kann.

Bei der Codierung der Bandselektion wurde die N aus M Eigenschaft von ACE ausgenutzt. Gilt bei der Codierung von AM_m^S, d.h. des Selektionszustands des S-ten Subbands, nämlich $\sum_{i=1}^{S-1} AM_m^i = N$, so ist

die Bandselektion vollständig festgelegt und somit kann die Codierung beendet werden.

Bei der Entwicklung des Electrocodecs, insbesondere für die Kompression der Bandselektion, wurden statische Wahrscheinlichkeiten verwendet mit Hinblick auf potentielle Übertragungsfehler. Hierdurch ist die Codierung der Bandselektion unabhängig von denkbaren Aktualisierungsmethoden, welche im Decoder und Encoder im Falle eines Übertragungsfehlers auseinanderlaufen könnten. Es hat sich gezeigt, dass bei der Codierung rauschbehafteter Signale minimal bessere Bitraten erzielbar sind, wenn eine Entropiecodierung der Bandselektion kombiniert wird mit der Option ohne Entropiecodierung und nur unter Ausnutzung der N aus M Eigenschaft von ACE zu codieren. Hierzu war die Ergänzung eines Indikatorbits, welches die im jeweiligen Zeitschritt gewählte Variante des Encoders an den Decoder übermittelt, notwendig. Die Entscheidung in jedem Zeitschritt, welche Variante gewählt wird, basiert auf der Länge der jeweils entstehenden Codewörter.

Der Grund, aus dem eine Kombination wie beschrieben leicht bessere Bitraten erzielt, ist in der fehlenden Adaption der Selektionswahrscheinlichkeiten sowie dem Erlernen der Selektionswahrscheinlichkeiten auf rauschfreien Daten begründet. Die Gründe, aus denen in diesem Punkten keine Adaption durchgeführt wurde, sind bereits dargelegt worden.

Der Bitstring, der in jedem Zeitschritt vom Encoder des Electrocodecs generiert wird, ist in Abb. 3.3 dargestellt.

3.2 VERLUSTLOSE KOMPRESSION DER ERREGUNGSMUSTER MITTELS KÜNSTLICHER NEURONALER NETZE

Als nächster, alternativer Ansatz zur verlustbehafteten Kompression der Erregungsmuster wie in Abschnitt 3.1 beschrieben, wurde die verlustlose Kompression mittels künstlicher neuronaler Netze untersucht. Hierbei wurde als das zu komprimierende Signal, anders als in allen anderen untersuchen Ansätzen, die Stromwerte $I(z_i, m)$ in klinischen Einheiten, welche gemäß Gl. 2.8 aus dem Ausgangssignal der Lautheitswachstumfunktion gebildet werden, gewählt. Nichtselektierte Bänder erhalten in der Nucleus Matlab Toolbox hierbei den Wert 0. Für eine verlustlose Kompression sind diese Signale besonders attraktiv, da der Dynamikbereich D_i dieser Stromwerte, definiert als

$$D_i = MCL(z_i) - THR(z_i) \tag{3.3}$$

für das Subband z_i, typischerweise zwischen 5 Bit und 7 Bit liegt. Die verwendete Codecstruktur ist stark von PAQ [Mah05], einem sehr bekannten verlustlosen Kompressionsverfahren, inspiriert und in Abb. 3.4

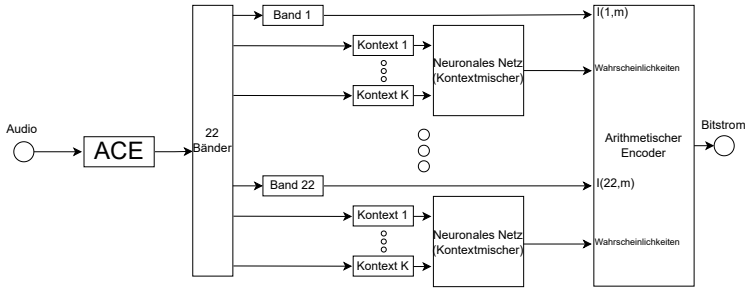


Abbildung 3.4: Struktur des entwickelten verlustlosen Codecs für die Erregungsmuster des Advanced Combination Encoders (ACE). Für jedes der $M = 22$ Subbänder werden K Kontextmodelle genutzt. Die von den Kontextmodellen geschätzten bedingten Auftrittswahrscheinlichkeiten der Stromwerte werden von Vorwärtsnetzwerken zu neuen Symbolwahrscheinlichkeiten kombiniert (Kontextmischung). Die Ausgabewahrscheinlichkeiten der Vorwärtsnetzwerke werden dann an die arithmetische Codierung weitergeleitet, welche die aktuellen Stromwerte $I(z, m)$ komprimiert.

dargestellt. Im entwickelten Verfahren wird für jedes Subband dabei eine Reihe von Kontexten berücksichtigt. Jedes Subband nutzt hierbei eigene, unabhängig arbeitende Kontexte. Die Aufgabe dieser Kontexte ist die Schätzung von bedingten Wahrscheinlichkeiten durch relative Häufigkeiten. Ein Kontext ist hierbei etwa durch den jeweils vorherigen Wert in einem Subband gegeben. Allgemeiner ist jedem Kontext ein Kontextmuster $\{(\Delta k_1, \Delta m_1), \ldots, (\Delta k_K, \Delta m_L)\}$ zugeordnet, wobei L die Kontextlänge bezeichnet. Ferner bezeichnet $\Delta k_i \in \mathbb{Z}$ eine Differenz im Subband und $\Delta m_i \in -\mathbb{N}$ eine Zeitdifferenz. D.h. gegeben das Band k im Zeitschritt m, so wird als Kontextwert, im Falle $L = 1$, der Wert $I(k + \Delta k_1, m + \Delta m_1)$ genutzt. Das aus dem Kontextwert $I(k + \Delta k_1, m + \Delta m_1)$ und dem Folgewert $I(k, m)$ bestehende Paar wird dann im Speicher des Kontexts abgelegt. Die bedingte Wahrscheinlichkeit $P(I(k, m)|I(k + \Delta k_1, m + \Delta m_1))$ im Zeitschritt m für das Band k ergibt sich dann als relative Häufigkeit basierend auf der beobachteten Häufigkeit der Folgewerte gegeben den aktuellen Kontext mit zugehörigem Kontextwert. Analog verhält es sich für Kontexte der Länge L und der Schätzung der bedingten Wahrscheinlichkeit $P(I(k, m)|I(k + \Delta k_1, m + \Delta m_1), \ldots, I(k + \Delta k_L, m + \Delta m_L))$. Eine Veranschaulichung der Funktionsweise der Kontexte ist in Abb. 3.5 dargestellt. Hierbei wird ein Kontext der Länge eins des zweiten Subbands mit einer beispielhaften Bufferlänge von fünf betrachtet, welcher den aktuellen Kontextwert 90 aufweist. Auf Basis der im Speicher

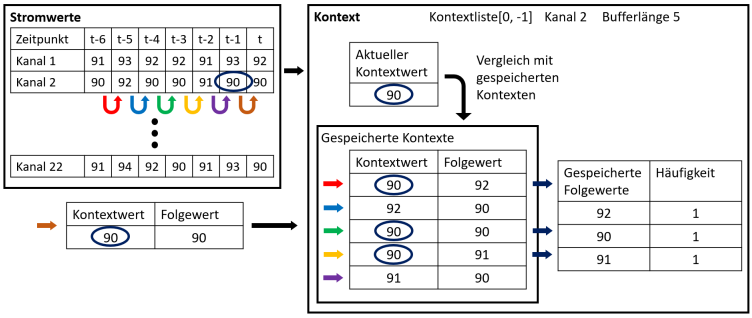


Abbildung 3.5: Funktionsweise der verwendeten Kontexte. Im Beispiel ist der Kontext [0,-1] dargestellt, d.h., derselbe Kanal im vorherigen Zeitschritt wird berücksichtigt. Die Bufferlänge entspricht hier fünf. Der aktuelle Kontextwert/Folgewertpaar 90, 90 wird in den Buffer des Kontextes eingetragen. Der älteste Eintrag, welcher hier durch das Paar aus dem Kontextwert 90 und dem Folgewert 92 besteht, wird entfernt. Zur Berechnung der bedingten Wahrscheinlichkeiten gegeben die gespeicherte Kontextwert/Folgewertpaare wird die relative Häufigkeit auf Basis des aktuellen Kontextwerts bestimmt.

des Kontextes vorliegenden Kontext-/Folgewertpaare wird die bedingte Wahrscheinlichkeit $P(I(2,m)|I(2,m-1))$ geschätzt.

Das aktuelle Kontext-/Folgewertpaar wird nun in den Speicher des Kontextes eingefügt, und die älteste Beobachtung wird entfernt. Hierdurch kommt es automatisch zu einer Anpassung an Änderungen in den Auftrittswahrscheinlichkeiten. Für nicht beobachtete Werte wird eine Mindestauftrittswahrscheinlichkeit verwendet, die in Kürze erläutert wird.

Tabelle 3.1 zeigt alle verwendeten Kontexte. Berücksichtigt wurden Kontexte der Länge eins bis drei. Ferner wurde zwischen Kurzzeit- und Langzeitkontexten unterschieden. Hierbei wurde ein Kontext als Langzeitkontext definiert, sofern mindestens drei Zeit- oder Subbandschritte zurück berücksichtigt wurden. Motiviert war die Nutzung von Langzeitkontexten durch die offensichtliche Existenz von Langzeitabhängigkeiten in den Erregungsmustern gewisser Subbänder. Dies wird im Unterabschnitt 3.6 gezeigt. Die Wahl der Kontexte wurde gesteuert durch die partiellen Korrelationskoeffizienten zwischen den Stromwerten $I(k,m)$ und $I(k+\Delta k, m+\Delta m)$ unter Konstanthaltung aller anderen Stromwerte zwischen einschließlich $m-8$ und m. Abb. 3.6 zeigt eine Messung dieser partiellen Korrelationen, durchgeführt auf dem TIMIT-Datensatz. Die angegebenen Werte stellen Mittelwerte der Absolutwerte der partiellen

[-5, -8]	0,002	[-5, -7]	0,004	[-5, -6]	0,003	[-5, -5]	0,004	[-5, -4]	0,007	[-5, -3]	0,007	[-5, -2]	0,005	[-5, -1]	0,024	[-5, 0]	0,012
[-4, -8]	0,002	[-4, -7]	0,005	[-4, -6]	0,004	[-4, -5]	0,004	[-4, -4]	0,007	[-4, -3]	0,008	[-4, -2]	0,005	[-4, -1]	0,024	[-4, 0]	0,018
[-3, -8]	0,002	[-3, -7]	0,005	[-3, -6]	0,005	[-3, -5]	0,005	[-3, -4]	0,008	[-3, -3]	0,010	[-3, -2]	0,004	[-3, -1]	0,027	[-3, 0]	0,037
[-2, -8]	0,004	[-2, -7]	0,006	[-2, -6]	0,005	[-2, -5]	0,006	[-2, -4]	0,009	[-2, -3]	0,010	[-2, -2]	0,005	[-2, -1]	0,022	[-2, 0]	0,042
[-1, -8]	0,005	[-1, -7]	0,008	[-1, -6]	0,008	[-1, -5]	0,009	[-1, -4]	0,014	[-1, -3]	0,018	[-1, -2]	0,007	[-1, -1]	0,044	[-1, 0]	0,331
[0, -8]	0,056	[0, -7]	0,076	[0, -6]	0,072	[0, -5]	0,084	[0, -4]	0,129	[0, -3]	0,105	[0, -2]	0,028	[0, -1]	0,682	[0, 0]	1,000
[1, -8]	0,004	[1, -7]	0,006	[1, -6]	0,008	[1, -5]	0,011	[1, -4]	0,016	[1, -3]	0,020	[1, -2]	0,008	[1, -1]	0,041		
[2, -8]	0,003	[2, -7]	0,006	[2, -6]	0,005	[2, -5]	0,006	[2, -4]	0,009	[2, -3]	0,010	[2, -2]	0,007	[2, -1]	0,022		
[3, -8]	0,003	[3, -7]	0,005	[3, -6]	0,005	[3, -5]	0,005	[3, -4]	0,009	[3, -3]	0,010	[3, -2]	0,005	[3, -1]	0,027		
[4, -8]	0,003	[4, -7]	0,005	[4, -6]	0,003	[4, -5]	0,004	[4, -4]	0,008	[4, -3]	0,009	[4, -2]	0,005	[4, -1]	0,024		
[5, -8]	0,002	[5, -7]	0,005	[5, -6]	0,003	[5, -5]	0,003	[5, -4]	0,007	[5, -3]	0,007	[5, -2]	0,006	[5, -1]	0,025		

Referenz	>1/4	>1/16	>1/32	>1/64	>1/128	<1/128

Abbildung 3.6: Wärmebild, das die partielle Korrelation für das sechste Subband zeigt. Die wichtigsten Kontexte für die Vorhersage sind dasselbe Subband im vorherigen Zeitschritt und das vorherige Subband im selben Zeitschritt. Die Notation $[\Delta k, \Delta m]$ ist kanalunabhängig, d.h. wenn z.B. das Subband k im Zeitschritt m kodiert werden soll, ist [-1,-2] der Kontext, der dem Subband $k - 1$ und dem Zeitschritt $m - 2$ entspricht.

Korrelationen aller Subbänder dar, für welche die gezeigten Kontexte gültig sind[1]. Es zeigte sich, dass für alle Subbänder die wichtigsten Kontexte durch den vorherigen Zeitschritt in demselben Subband und das vorherige Subband in demselben Zeitschritt gegeben sind.

Der partielle Korrelationskoeffizient, und nicht der gewöhnliche Korrelationskoeffizient, ist hierbei das sinnvollere Maß, da jeder berücksichtigte Kontext im Vergleich zu allen bereits berücksichtigten Kontexten möglichst viel neue Information liefern sollte. Dazu muss dieser möglichst statistisch unabhängig zu allen anderen Kontexten sein. Dies lässt sich einfach mittels des partiellen Korrelationskoeffizienten abschätzen. Die konkrete Auswahl der Kontexte wurde auf Basis der Kontexte erster Ordnung mit einer partiellen Korrelation größer als $\frac{1}{32}$ bestimmt. Zusätzlich wurde noch der Kontext $[0, -2]$ bei der Auswahl berücksichtigt, für den die nächst kleinere partielle Korrelation berechnet wurde. Aus dieser Menge wurden dann Kontexte zweiter und dritter Ordnung defininiert.

Zur eigentlichen Codierung der Stromwerte im Zeitschritt m wird über alle Subbänder iteriert und für jedes Subband werden die Schätzungen der bedingten Wahrscheinlichkeiten aller Kontexte abgefragt. Diese werden an ein Vorwärtsnetzwerk weitergeleitet, welches diese Schätzungen der Stromwertauftrittswahrscheinlichkeiten kombiniert und daraus neue, im Allgemeinen verbesserte, Auftrittswahrscheinlichkeiten berechnet. Diese werden dann jeweils für die Codierung des aktuellen Stromwerts im jeweiligen Subband verwendet. Hierbei wurde, zur Verhinderung extrem kleiner Wahrscheinlichkeiten im Falle seltener Symbole, sowie zur Berücksichtigung nicht beobachteter Symbole, eine Mindestauftrittswahr-

1 Etwa kann der Kontext [-1,0] nicht für das erste Subband genutzt werden, da hierzu ein nulltes Subband nötig wäre.

Tabelle 3.1: Auflistung der 38 verwendeten Kontexte je Subband des verlustlosen Codecs. Grundlegend wird dabei zwischen Kontexten erster, zweiter und dritter Ordnung unterschieden. Zusätzlich wurden Langzeitkontexte berücksichtigt, welche als Kontexte definiert wurden, die Kontextwerte mindestens drei Zeit- oder Subbandschritte entfernt berücksichtigen. Diese sollen besonders Langzeitabhängigkeiten erkennen. Hinsichtlich der Notation ist jedes Tupel, bestehend aus der Differenz zum Kanal und der zeitlichen Differenz, in eckigen Klammern dargestellt.

Kontexte						
1. Ordnung	[0, -1]	[-1, 0]	[0, -2]	[-2, 0]	[1, -1]	[-1, -1]
2. Ordnung	[0, -1], [-1, 0]		[0, -1], [0, -2]		[0, -1], [1, -1]	
	[0, -1], [-1, -1]		[-1, 0], [-1, -1]		[-1, 0], [-2, 0]	
	[-1, -1], [-2, 0]		[-1, -1], [0, -2]		[1, -1], [0, -2]	
3. Ordnung	[0, -1], [-1, 0], [-1, -1]			[0, -1], [-1, 0], [1, -1]		
	[0, -1], [-1, 0], [0, -2]			[0, -1], [-1, 0], [-2, 0]		
	[0, -1], [0, -2], [-1, -1]			[0, -1], [0, -2], [1, -1]		
	[0, -1], [-1, -1], [1, -1]			[0, -1], [-1, -1], [-2, 0]		
	[-1, 0], [-1, -1], [-2, 0]					
Langzeit	[-3,0]	[0,-3]	[0,-4]	[0,-5]	[0,-6]	[0,-7]
	[0,-8]	[0, -6], [0, -7], [0, -8]	[0, -2], [0, -3]	[0, -3], [0, -4]	[0, -4], [0, -5]	[0, -5], [0, -6]
	[0, -6], [0, -7]	[0, -7], [0, -8]				

scheinlichkeit $p_{min,k}$ für die Symbole eines jeden Subbands definiert. Diese ist gemäß

$$p_{min,k} = \frac{0,2}{D_k + 1} \tag{3.4}$$

festgelegt worden mit dem Dynamikbereich D_k nach Gl. 3.3. Im Nenner wird $D_k + 1$ verwendet, da D_k den Nullwert, welcher den Zustand ohne Stromwert codiert, nicht berücksichtigt. Für ein Subband k mit 32 Symbolen respektive Stromwerten - dies entspricht einer Auflösung von 5 Bit - ergibt sich $p_{min,k} = 0,625\,\%$. Der Vorfaktor $0,2$ wurde zunächst ad hoc mit dem Hintergedanken festgelegt, dass ein Fünftel der Auftrittswahrscheinlichkeit im Falle einer Gleichverteilung eine sinnvolle Untergrenze darstellt. In Kapitel 4 wird kurz eine durchgeführte Optimierung dieses Vorfaktors diskutiert, welche jedoch zu keiner Verbesserung geführt hat. Der Vorfaktor ist also in der Tat mit hoher Wahrscheinlichkeit sinnvoll gewählt worden. Musste eine Symbolwahrscheinlichkeit, die vom Vorwärtsnetzwerk ausgegeben wurde, gemäß der Untergrenze 3.4 angehoben werden, so war eine Normierung der Symbolwahrscheinlichkeiten nötig,

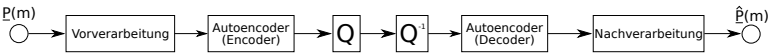


Abbildung 3.7: Signalfluss der Kompression der Erregungsmuster $\underline{P}(m) := (P(1,m),\ldots,P(M,m))$ mittels des Autoencoders. In der Vorverarbeitung werden alle NaN sowie Stellvertreterwerte -10^{-10} auf einen negativen Wert gesetzt, um diese von regulären Stromwerten zu unterscheiden. Nachfolgend werden die Erregungsmuster vom Autoencoder komprimiert, und diese komprimierte Darstellung quantisiert. Diese quantisierte Darstellung wird dann vom Decoder des Autoencoders rekonstruiert und in der Nachverarbeitung eine Selektion von maximal N der M Bänder sowie $P(z_i,M) \in [0,1]$ für selektierte Subbänder z_i sichergestellt.

sodass sich die Symbolwahrscheinlichkeiten zu eins addieren. Hierzu wurden die Symbolwahrscheinlichkeiten jeweils gemäß

$$\tilde{p}_{k,m,i} = \frac{p_{k,m,i}}{\sum_{i=1}^{D_k+1} p_{k,m,i}} \tag{3.5}$$

normiert. Hierbei ist $p_{k,m,i}$ die im Zeitschritt m vom Vorwärtsnetzwerk ausgegebene Wahrscheinlichkeit des i-ten Symbols im Subband k. Diese $\tilde{p}_{k,m,i}$ wurden dann von der arithmetischen Codierung zur eigentlichen Kompression der Erregungsmuster genutzt. Der Decoder ging vollständig analog vor.

3.3 AUTOENCODER

Zur weiteren Senkung der Bitrate wurden nach Vollendung des ersten Entwurfs des Electrocodecs Methoden des maschinellen Lernens verwendet, um einen verlustbehafteten Codec zu entwickeln, der bei gleicher Sprachverständlichkeit eine niedrigere Bitrate erzielen kann. Hierzu wurden unterschiedliche Autoencoderstrukturen untersucht. Eine Reihe von Tests waren notwendig, um ein Verfahren zu entwickeln, welches das Erreichen des gesteckten Ziels erlaubte. Eine speziell entworfene Verlustfunktion, die die Verzerrung der logarithmischen Einhüllenden und der Bandselektion separat bewertet, erlaubte eine Verbesserung der Verständlichkeit der codierten Erregungsmuster, jedoch konnte die Referenzverständlichkeit der uncodierten Erregungsmuster nicht erreicht werden. Die Motivation zur separaten Bewertung der Verzerrung der Bandselektion war wesentlich durch die Tatsache begründet, dass diese einen sehr wesentlichen Einfluss auf das Sprachverstehen von Cochlea-Implantatträgern hat [Qaz+13]. Ein weiterer Ansatz optimierte die Ge-

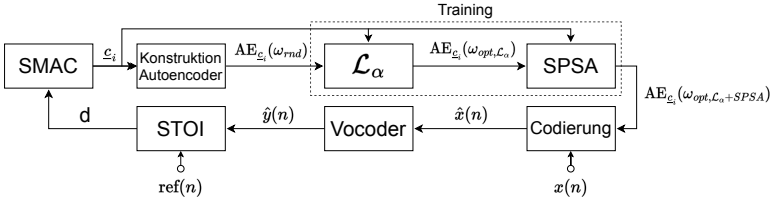


Abbildung 3.8: Verwendete Optimierungsschleife zur Optimierung der Autoencoderstruktur. In jeder Iteration werden Hyperparameter $\underline{c}_i$ des Sequential Model-Based Algorithm Configuration (SMAC) zur Konstruktion eines Autoencoders genutzt. Dieser wird anschließend unter Zuhilfenahme der Verlustfunktion $\mathcal{L}_\alpha$ nach Gl. 3.6 trainiert. Danach erfolgt eine weitere Optimierung mit dem Stochastic Perturbation Simultaneous Approximation (SPSA)-Algorithmus gemäß Gl. 2.52. Nach Abschluss dieses Trainingsschritts wird die Verständlichkeit der codierten Erregungsmuster im Vergleich zum Referenzaudiosignal `ref(n)` durch das Short-Time Objective Intelligibility Measure (STOI) bewertet.

wichte des Autoencoders gradientenfrei mittels eines evolutionären Algorithmus, wobei als Maß der Sprachverständlichkeit VSTOI verwendet wurde. Zwar konnte damit die Sprachverständlichkeit der codierten Erregungsmuster nach und nach verbessert werden, jedoch war das Verfahren zu langsam um in hinreichender Zeit die Referenz zu erreichen und zudem eine Optimierung der Autoencoderstruktur zuzulassen.

Schlussendlich wurde eine Kombination aus gradientenbasiertem und gradietenfreiem Training verwendet [HOO22], welche in einer ersten Stufe zur Optimierung der Hyperparameter des Autoencoders sowie weiterer Parameter genutzt wurde. Inbesondere wurden die Neuronenzahlen des Autoencoders optimiert. Danach, in einem zweiten Schritt, wurde die gefundene optimierte Autoencoderstruktur dann zunächst für viele Epochen gradientenbasiert und anschließend für viele Iterationen gradientenfrei[2] mit dem in Abschnitt 2.3.12 beschriebene SPSA-Algorithmus trainiert, um die Referenz bezüglich Verständlichkeit der codierten Erregungsmuster zu erreichen.

Für das gradientenbasierte Training des Autoencoders wurde die Verlustfunktion

$$\mathcal{L}_\alpha := (1-\alpha)\mathcal{L}_{\mathtt{LogEnv}} + \alpha\mathcal{L}_{\mathtt{Sel}} = (1-\alpha)\sum_{i\in \mathtt{Sel}}(P_i - \hat{P}_i)^2 + \alpha\sum_{i\in \mathtt{Sel}^C}\sigma(\hat{P}_i) \quad (3.6)$$

verwendet, wobei $\sigma(x) \equiv \mathrm{reLu}(x) = \max(x, 0)$, P_i die logarithmische Einhüllende und $\hat{P}_i$ die vom Autoencoder rekonstruierte logarithmische

2 Der Begriff „gradientenfrei“ wird hier etwas verkürzend für die Optimierung mittels numerischer Approximation zu berechnender Gradienten genutzt.

Einhüllende ist. Der Autoencoder komprimiert dabei zur Minimierung der algorithmischen Latenz in jedem Zeitschritt m genau ein Frame, d.h. jeweils einen Vektor der Art $(P_1^m, \ldots, P_M^m)^T$. Auf Grund der Erzeugung der Erregungsmuster mit der Nucleus Matlab Toolbox und dem Auftreten von NaN-Werten, siehe Abschnitt 2.2.1.1 sowie Gl. 2.6, war es notwendig die Daten vor- und nachzuverarbeiten. In diesem Schritt wurden alle Subbänder mit dem Stellvertreterwert -10^{-10} oder dem Wert NaN auf den Wert $-0,006$ gesetzt, welcher sich in Pilotuntersuchungen als sinnvoll herausgestellt hatte[3]. Dadurch wiesen alle Subbänder in jedem Zeitschritt nach der Vorverarbeitung Werte auf, welche vom Autoencoder verarbeitet werden können. Die Kennzeichnung der nichtselektierten Bänder durch einen negativen Wert motivierte die Verwendung der ReLu-Funktion für die Bewertung der Verzerrung der Bandselektion. Rekonstruiert der Autoencoder einen Wert größer Null für ein nichtselektiertes Band, so sollte die Verlustfunktion diesen Wert bestrafen. Ist der Wert jedoch kleiner Null, so sollte die Verlustfunktion diesen nicht bestrafen, egal welchen Betrag der rekonstruierte Wert hat, da die Selektion korrekt rekonstruiert werden kann. Beides zusammen lässt sich mit der ReLu-Funktion bewerkstelligen, wobei gehofft wurde, dass der Fall $\hat{P}_i = 0$ nicht auftritt.

Das gradientenfreie Training wurde mittels des SPSA-Algorithmus, beschrieben in Abschnitt 2.3.12, durchgeführt. Hierzu wurde ein mit der Verlustfunktion 3.6 vortrainierter Autoencoder verwendet, dessen Gewichte wie in Gl. 2.52 dargestellt iterativ zweimal zufällig pertubiert wurden. An die Stelle der Funktion f in Gl. 2.52 trat der VSTOI-Wert der vom Autoencoder, bestückt mit den Gewichten $\underline{\omega}_k \pm c_k \Delta_k$, codierten Erregungsmuster.

Die Struktur des Autoencoders wurde mittels SMAC bestimmt. Hierzu wurde eine Optimierungsschleife angewendet, die in Abb. 3.8 dargestellt ist. Mit einem einzelnen Erregungsmuster wurde, gegeben einen Satz von Hyperparametern $\underline{c}_i$, ein Autoencoder konstruiert und mit der Verlustfunktion 3.6 für 500 Epochen vortrainiert. Anschließend wurden 100 Iterationen des SPSA-Algorithmus durchgeführt und die Erregungsmuster dann mit dem Autoencoder codiert und bezüglich der Verständlichkeit mit STOI bewertet. Dessen Bewertung wurde dann an SMAC zurückgegegeben, wodurch die Optimierungsschleife geschlossen wurde.

Mittels dieser Kombination von gradientenbasiertem und gradientenfreiem Training konnten, wie in Abschnitt 4 gezeigt wird, Autoencoderstrukturen ermittelt werden, deren codierte Erregungsmuster die Referenzverständlichkeit erreichen konnten.

3 Der genaue Wert ist, sofern hinreichend klein, nicht besonders wichtig.

3.4 RÜCKKOPPLUNGSAUTOENCODER

Die zuvor beschriebene Methodik zur Bestimmung sowie des Trainings optimaler Autoencoderstrukturen wurde grundsätzlich für Rückkopplungsautoencoder übernommen, jedoch mit gewissen Abwandlungen.

Zunächst ist das Vorgehen aus Abschnitt 3.3 noch suboptimal, da die Gesamtstruktur bestehend aus Autoencoder und Quantisierer nicht gemeinsam optimiert wurde. Beim Rückkopplungsautoencoder wurde diese Tatsache berücksichtigt, so dass nach Hyperparameteroptimierung zunächst wie gehabt die Gewichte mittels SPSA-Algorithmus bezüglich STOI optimiert werden. Im Anschluss erfolgt das Training des Quantisierers mittels Lloyd-Max-Algorithmus. Bis hierhin ist das Vorgehen identisch zum Autoencoder ohne Rückkopplung. Jedoch kommt es nun zur erneuten Anwendung des SPSA-Algorithmus, der nun alle Parameter der gesamten Struktur, d.h. die Gewichte des Rückkopplungsautoencoders als auch die Vektoren der Codebücher, bezüglich STOI optimiert. Es zeigte sich, dass auf diese Weise die Gesamtstruktur bezüglich ihrer Leistungsfähigkeit erheblich verbessert werden kann. Hierdurch wurde ein allgemein nutzbares Verfahren für die Kompression der Erregungsmuster von Cochlea-Implantaten entwickelt, das automatisiert näherungsweise optimale Kompressionsalgorithmen entwirft. Das Vorgehen ist dabei so generisch, dass es für beliebige Signalverarbeitungsstrategien von Cochlea-Implantaten nutzbar sein sollte. Der Rückkopplungsautoencoder erwies sich im Training als deutlich instabiler als der gewöhnliche Autoencoder, sodass sich die Leistungsfähigkeit während des Trainings mitunter plötzlich deutlich reduzierte. Das Training gestaltete sich hierdurch etwas schwieriger. Um dieses Problem, welches insbesondere die Hyperparameteroptimierung beeinträchtigte, zu lösen, wurde in der Trainingsphase das aktuelle Modell nach jeder Trainingsepoche gespeichert und am Ende der Trainingsphase das beste Modell bzw. die besten Gewichte verwendet respektive dessen Leistungsfähigkeit gemessen durch die Verlustfunktion 3.6 oder STOI an SMAC zurückgegeben. Beispieltrainingskurven werden im Ergebnisteil zur Illustration dieses Sachverhalts gezeigt.

Desweiteren wurde frühes Trainingsstoppen (engl. early stopping) eingeführt, um das Training bzw. die Hyperparameteroptimierung zu beschleunigen, insbesondere aufgrund des im vorherigen Abschnitt beschriebenen schwierigen Trainingsverhaltens.

Eine letzte Neuerung, die beim Training des Rückkopplungsautoencoders in einem separaten Trainingsschritt erprobt wurde, ist eine Regularisierung der VSTOI-Werte mit der Absicht, die Kompressionsleistung unabhängiger vom Signal-Rausch-Verhältnis des akustischen Szenarios zu machen.

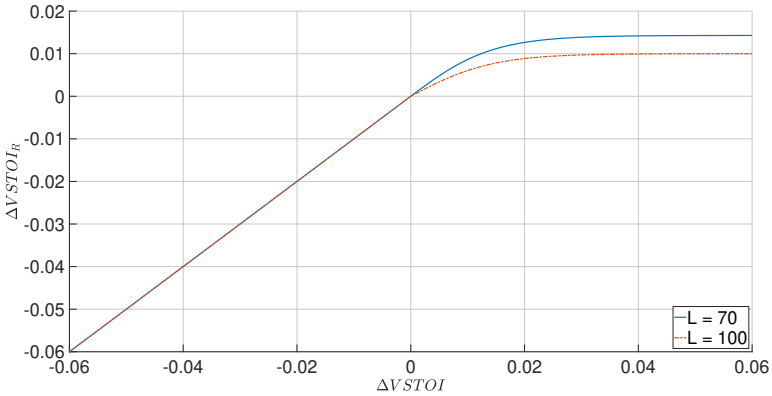


Abbildung 3.9: Visualisierungen der Gl. 3.7 für $L = 70$ und $L = 100$. Der glatte Übergang vom linearen in den nichtlinearen Bereich bei $\Delta VSTOI = 0$ ist gut zu erkennen.

Bereits beim Autoencoder - und ebenso beim Rückkopplungsencoder - zeigte sich, dass bei niedrigem Signal-Rausch-Verhältnis, d.h. bei eher starkem Rauschen, die VSTOI-Werte der codierten Erregungsmuster im Vergleich zur Referenz, d.h. die Erregungsmuster ohne Codierung, verbesserten. Der Autoencoder lernte mutmaßlich eine Art Rauschunterdrückung. Jedoch ging dies, zumindest nominell, marginal auf Kosten der Kompressionsleistung bei hohem Signal-Rausch-Verhältnis. Zur Erläuterung sei dies mathematisch formuliert: Sei $\mathrm{VSTOI}_{\mathrm{Coded}}$ der VSTOI-Wert eines decodierten Erregungsmusters des Autoencoders und $\mathrm{VSTOI}_{\mathrm{Ref}}$ der VSTOI-Wert des Referenzerregungsmusters, d.h., ohne dass dieses codiert worden wäre. Ferner sei $\Delta\mathrm{VSTOI} := \mathrm{VSTOI}_{\mathrm{Coded}} - \mathrm{VSTOI}_{\mathrm{Ref}}$. Bei niedrigem Signal-Rausch-Verhältnis zeigte sich, wie im Ergebnisteil gezeigt werden wird, dass im Mittel $\Delta\mathrm{VSTOI} > 0$ galt, mit unter relativ deutlich. Bei hohem Signal-Rausch-Verhältnis entsprechend im Mittel $\Delta\mathrm{VSTOI} < 0$. Da bei hohem Signal-Rausch-Verhältnis kaum ein Gewinn durch Rauschunterdrückung erzielbar ist, kann es für den Autoencoder mit Hinblick auf den mittleren VSTOI-Wert über alle Daten gemittelt vorteilhaft sein, die Codierung bei niedrigem Signal-Rausch-Verhältnis zu verbessern, auch wenn dafür die Codierung bei hohem Signal-Rausch-Verhältnis etwas vernachlässigt wird. Um dieser Eigenheit entgegen zu steuern, wurde wie folgt regularisiert: Im Training des Autoencoders

wird in jeder Epoche der modifizierte VSTOI-Wert $VSTOI^{L}_{Coded,mod}$ für jedes Erregungsmuster berechnet gemäß

$$VSTOI^{L}_{Coded,mod} := \begin{cases} VSTOI_{Ref} + \frac{1}{L}\tanh(L \cdot \Delta VSTOI), \Delta VSTOI > 0 \\ VSTOI_{Coded}, sonst \end{cases} \tag{3.7}$$

mit $L > 1$. Der Mittelwert aller $VSTOI^{L}_{Coded,mod}$ Werte wird dann als Gütekriterium der Optimierung mit dem SPSA-Algorithmus genutzt.

Der Ausdruck $\frac{1}{L}\tanh(L \cdot \Delta VSTOI)$ hat als Maximum den Wert $\frac{1}{L}$. Getestet wurden die Werte $L = 40, 70$ sowie $L = 100$. Für $\Delta VSTOI < 0$ ist der Verlauf von Gl. 3.7 linear. Ferner ist der Übergang glatt, was vorteilhaft für die Optimierung sein dürfte. Eine Abbildung des Verlaufs der Funktion nach Gl. 3.7 ist in Abb. 3.9 für $L = 70$ und $L = 100$ dargestellt.

Zum Training und zur Evaluierung der vorgestellten Codierungsansätze auf Basis künstlicher neuronaler Netze war eine hinreichend große und umfassende Datenbasis notwendig. Diese soll, zusammen mit den anderen im Rahmen der vorgelegten Arbeit genutzten Datensätze, im nächsten Abschnitt vorgestellt werden.

3.5 DATENSÄTZE

Im Rahmen der Arbeit wurde auf drei verschiedene Datensätze zurückgegriffen. Diese sind das Sound Quality Assessment Material (SQAM) [Eur], der Hochmair-Schulz-Moser-Satztest (HSM) [Hoc+97] sowie der TIMIT[4]-Sprachkorpus bzw. -Datensatz [ZSG90]. Das SQAM wurde für das Training des Electrocodecs und eine initiale Evaluation ebendieses im Vergleich zum G.722 Audiocodec verwendet. Der HSM wurde ebenso für die Evaluation des Electrocodecs genutzt und ferner, was seine Nutzung motivierte, in den durchgeführten Hörtests als Sprachmaterial verwendet. Der (augmentierte) TIMIT-Datensatz wurde für das Training und die Evaluierung der Codierungsstrategien auf Basis künstlicher neuronaler Netze genutzt.

Das SQAM ist ein hochqualitativer Datensatz von Audiosignalen mit einer Abtastfrequenz von 48 kHz, erstellt von der europäischen Rundfunkunion, der sowohl Testsignale wie reine Töne als auch Rauschen enthält, darüber hinaus aber auch verschiedene Aufnahmen von Instrumenten, Musik sowie Sprache in den Sprachen Englisch, Deutsch und Französisch. Es wurden im Rahmen der vorgelegten Arbeit nur die Sprachaufnahmen verwendet, welche insgesamt sechs Audioaufnahmen von einer Länge von jeweils etwa 30 Sekunden umfassen, wobei für jede Sprache eine

4 Der Name beruht mutmaßlich auf den Kürzeln der Organisationen, Texas Instruments und das Massachusetts Institute of Technology, welche den Sprachkorpus erstellt haben.

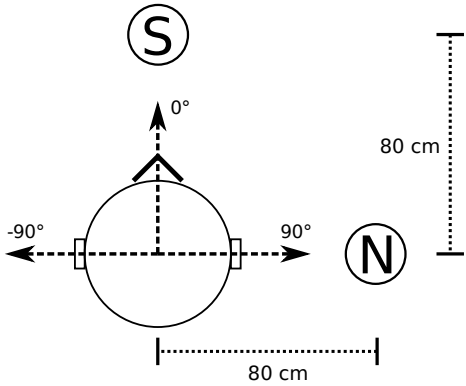
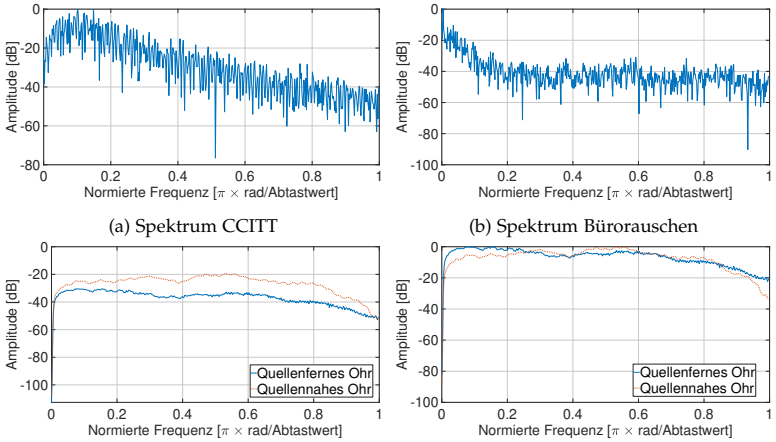


Abbildung 3.10: Positionierung des Sprechers (S) und des Rauschens (N) relativ zum Hörer im akustischen Szenario, welches in den Hörtests verwendet wurde. Die hierzu spiegelsymmetrische Anordnung wurde verwendet, wenn das bessere getestete Ohr das rechte war.

Sequenz einmal von einem Mann und einmal von einer Frau gesprochen wird.

Als bekannter deutschsprachiger Satztest aus der Forschung über Cochlea-Implantate diente der HSM als Datensatz für die ersten Erprobung von Codierungsstrategien. Er umfasst 30 Satzlisten á 20 Sätze, wobei jeder Satz einmal von einem Mann und einmal von einer Frau gesprochen wird. Jede Satzliste umfasst insgesamt genau 106 Worte. Insgesamt existieren also 1200 Aufnahmen, jede mit einer Abtastfrequenz von 44,1 kHz. Der offizielle Teil, welcher auch in der Forschung verwendet wird, ist jedoch jener mit männlichem Sprecher, weswegen nur diese 600 Aufnahmen verwendet wurden. Die Aufnahmen haben eine Dauer zwischen etwa einer Sekunde und etwa drei Sekunden. Der TIMIT-Sprachkorpus wurde als Datensatz hinzugezogen, nachdem Hörtests die Leistungsfähigkeit des Electrocodecs nachgewiesen hatten [Hin+21a]. Zur Weiterentwicklung der Codierung mittels künstlicher neuronaler Netze mit dem Ziel einer weiteren Reduktion der Bitrate bei gleicher Verständlichkeit der Erregungsmuster war es notwendig, die Datenbasis zu erweitern. Während der HSM lediglich einen (männlichen) Sprecher umfasst und das SQAM auch nur sechs, nutzt der TIMIT-Sprachkorpus 630 englischsprachige Sprecher, davon 438 männlich und 192 weiblich. Jeder Sprecher spricht insgesamt 10 Sätze, wobei jeder Satz im Mittel drei Sekunden lang ist. Der TIMIT-Datensatz ist insbesondere in einen dedizierten Trainings- und einen dedizierten Testdatensatz aufgeteilt,



(a) Spektrum CCITT

(b) Spektrum Bürorauschen

(c) Kopfübertragungsfunktion - unnormiert

(d) Kopfübertragungsfunktion - normiert

Abbildung 3.11: (a) Normiertes Spektrum des Consultatif International Téléphonique et Télégraphique (CCITT) [Int93] Rauschens, das im Rahmen der Hörtests als Hintergrundrauschen verwendet wurde. (b) Normiertes Spektrum des Bürorauschens, welches zur Augmentierung des TIMIT-Datensatzes verwendet wurde. (c) Unnormierter Amplitudengang der Kopfübertragungsfunktion zum quellennahen und quellenfernen (abgeschatteten) Ohr für einen Azimut von 90° wie im Falle des Rauschens in Abb. 3.10. (d) wie (c) aber mit Normierung. Die Abtastfrequenz betrug jeweils 16 kHz.

welche disjunkte Sprecher verwenden. Die Abtastfrequenz beträgt 16 kHz.

3.5.1 *Mischung mit Rauschen*

Alle genannten Datensätze umfassen Sprache in stiller Umgebung. In realen Umgebungen existiert fast immer ein gewisser Hintergrundpegel, d.h., Erregungsmuster, welche aus Mikrophonsignalen abgeleitet werden, enthalten in realen Situationen immer zumindest eine kleine Menge an Rauschen. Für die Durchführung von Hörtests sowie für die Entwicklung robuster Codierungsverfahren war die Anreicherung der Datensätze um Hintergrundrauschen daher unumgänglich. Zwar ist es zunächst trivial, rauschfreie Signale mit Rauschen zu mischen, jedoch führt ein naiver Ansatz nicht zu einem realistischen Höreindruck respektive Audiosignal. Grundsätzlich ist bei der Erstellung realistischer Audiosignale das akustische Szenario festzulegen, womit die Kombination aus den Reflek-

tionseigenschaften des umgebenden Raums, den Eigenheiten (Position, Lautstärke,...) der Schallquelle sowie den Eigenheiten der Rauschquelle(n) und dem Mikrophon gemeint ist.

Prinzipiell sind beliebig viele akustische Szenarien denkbar und entsprechend könnten auf beliebige Weise die Sprachaufnahmen mit Rauschen gemischt werden. Im Rahmen dieser Arbeit wurde sich jedoch auf typische Szenarien aus der Cochlea-Forschung beschränkt. Bei diesen ist typischerweise ein Sprecher und mindestens eine weitere Rauschquelle um den Träger des Cochlea-Implantats positioniert. Ein solches Szenario wurde auch in dem durchgeführten Hörtest verwendet. Es ist üblich, und wurde auch in der vorgelegten Arbeit so vollführt, diese akustischen Szenarien zu simulieren. Dies erfolgt unter Zuhilfenahme sogenannter kopfbezogener Übertragungsfunktionen (engl. head related transfer function (HRTF)), welche den Einfluss des Kopfes auf Schallwellen berücksichtigen. In der vorgelegten Arbeit wurden die Übertragungsfunktionen aus [Kay+09] verwendet. Diese offerieren fünf verschiedene akustische Umgebungen, namentlich einen reflexionsarmen Raum, ein Büro, zwei Cafeterien sowie einen Hof. In der vorgelegten Arbeit wurde die Hofumgebung nicht verwendet, da sie nur für sehr wenige Einfallswinkel kopfbezogene Übertragungsfunktionen anbietet. Diese kopfbezogenen Übertragungsfunktionen wurden zweimal zur Mischung von Audiosignalen verwendet: Einmal für den durchgeführten Hörtest und einmal für die Augmentierung des TIMIT-Datensatzes.

Im Rahmen des durchgeführten Hörtests wurde das in Abb. 3.10 dargestellte Szenario simuliert. Der Sprecher war hierbei in einem reflexionsarmen Raum vor dem Träger des Cochlea-Implantats positioniert, die Rauschquelle zu seiner Rechten. Der Abstand der Quellen zum Cochlea-Implantatträger wurde auf 80 cm festgelegt. In diesem Szenario ist das Signal-Rausch-Verhältnis des linken Ohrs auf Grund der Abschattung durch den Kopf besser, d.h. höher, als jenes am rechten Ohr. In einem solchen Fall ist z.B. die Anwendung des kontralateralen Routing des Signals des linken Ohrs zum rechten Ohr, inklusive drahtloser Übertragung, sinnvoll.

Als Rauschsignal wurde sogenanntes Consultatif International Téléphonique et Télégraphique (CCITT) Rauschen [Int93] verwendet. Dieses weist ein sprachartiges Spektrum auf und wird für die Modellierung sozialer Situationen verwendet, in welchen das Hintergrundrauschen wesentlich durch andere Sprecher gegeben ist. Das Amplitudenspektrum des CCITT-Rauschens sowie eines Bürorauschsignals nebst den Amplitudengängen der kopfbezogenen Übertragungsfunktionen ist in Abb. 3.11 dargestellt. Da das beschriebene akustische Szenario zu speziell gewählt ist für die Entwicklung einer robusten Codierungsstrategie, wurde für

Tabelle 3.2: Sprachazimut, Rauschazimut, Signal-Rausch-Verhältnis (SNR), Rauschart und akustisches Szenario für den Trainingsdatensatz und den Testdatensatz des augmentierten TIMIT-Sprachkorpus. Die Notation für Azimut und SNR bedeutet: Untergrenze:Schrittweite:Obergrenze. Bfr ist Hintergrundrauschen aus einem Restaurant und CCITT ist sprachähnliches Rauschen.

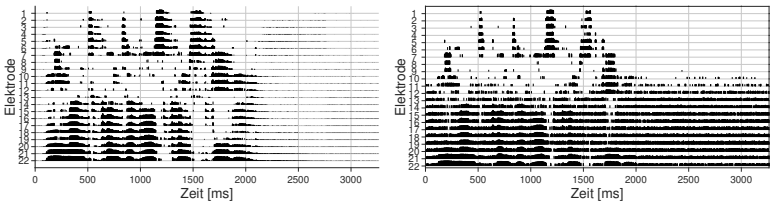
Label	Sprachazimut (°)	Rauschazimut (°)	SNR (dB)	Rauschart	Szenario
Train	-90:15:90	-90:15:90	-5:5:20, 30, 50	CCITT, Bus, Bfr, Büro	reflexionsarmer Raum, Büro
Test	-90:5:90	-90:5:90	-2.5:2.5:10, 20, 40	CCITT, Bus, Bfr, Büro	reflexionsarmer Raum, Büro, Cafeteria

die Entwicklung eines Codecs auf Basis künstlicher neuronaler Netze ein weiterer Datensatz erstellt. In diesem wurden Sprecher des TIMIT-Datensatzes mit verschiedenen Rauschsignalen gemischt, wobei sowohl das Rauschsignal, als auch die akustische Umgebung in den meisten Freiheitsgraden jeweils zufällig gewählt wurden, damit ein möglichst umfassender Datensatz zustande kam. Auf diese Art wurden ein dedizierter Trainingsdatensatz und ein dedizierter Testdatensatz erzeugt. Genauer wurden hierbei für jeden einzelnen Satz des TIMIT-Datensatzes das Signal-Rausch-Verhältnis, Einfallswinkel von Sprach- und Rauschsignal, Art des Rauschens sowie die akustische Umgebung (Cafeteria, Büro, ...) zufällig gewählt. Insbesondere wurde auch der Ausschnitt des Rauschens, welcher mit den jeweiligen Sprachsignalen vermischt wurde, welche deutlich kürzer waren als die Rauschsignale, zufällig gewählt. Tabelle 3.2 fasst die Auswahl der Werte genauer zusammen. Die Auflistung zeigt die Menge an Werten, aus denen pro TIMIT-Satz jeweils ein Wert zufällig gewählt wurde.

3.6 ANALYSE DER ERREGUNGSMUSTER

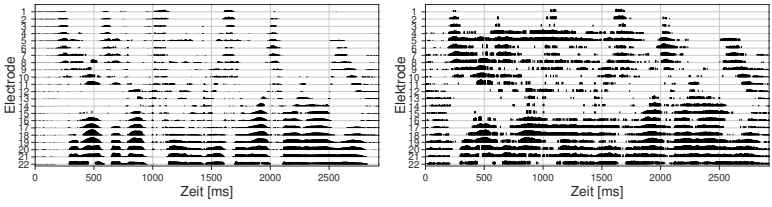
Da die Erregungsmuster, die der Signalprozessor eines Cochlea-Implantats generiert, außerhalb der unmittelbaren Forschung über Cochlea-Implantate praktisch unbekannt sind und keine Vorstellung davon bestehen dürfte, wie diese aussehen und welche Eigenheiten diese aufweisen, soll im Folgenden etwas auf Charakteristika von Erregungsmustern eingegangen werden.

Abbildung 3.12 zeigt Elektrodogramme jeweils eines Beispielsatzes des HSM sowie des TIMIT-Datensatzes in Ruhe sowie mit Hintergrundrauschen bei einem Signal-Rausch-Verhältnis von 10 dB. Elektrodogramme sind Visualisierungen der Stromwerte aller Subbänder oder Elektroden über die Zeit. Jeder Stromwert jeder Elektrode wird dabei durch einen einzelnen Balken dargestellt.



(a) HSM - Rauschfrei

(b) HSM - 10 dB

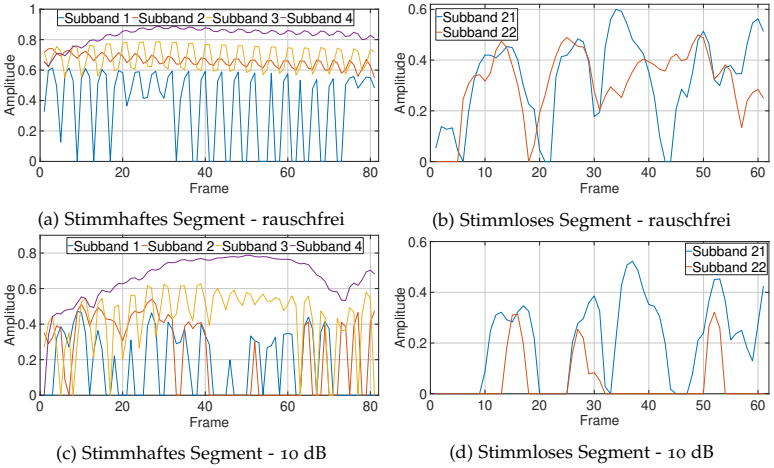


(c) TIMIT - Rauschfrei

(d) TIMIT - 10 dB

Abbildung 3.12: (a-b) Elektrodogramm des Satzes „Niemand konnte Auskunft geben“ aus dem HSM (a) ohne Hintergrundrauschen und (b) bei einem Signal-Rausch-Verhältnis von 10 dB und CCITT-Rauschen. (c-d) Elektrodogramm des Satzes “She had your dark suit in greasy wash water all year“ aus dem TIMIT-Datensatz (c) ohne Hintergrundrauschen und (b) bei einem Signal-Rausch-Verhältnis von 10 dB und Restaurantrauschen.

Für den HSM wurde zur Erzeugung der Grafik ein verwendetes Audiosignal respektive ein Satz des durchgeführten Hörtests verwendet. Sehr deutlich ist der Einfluss des Hintergrundrauschens in Abschnitten ohne oder mit sehr geringen Sprachanteilen zu beobachten, wie sie unmittelbar am Anfang für beide Sätze auftreten. Für das HSM wurde CCITT-Rauschen verwendet, dessen (geschätztes) Amplitudenspektrum in Abb. 3.11a gezeigt wird. Das Maximum des Amplitudenspektrums des CCITT-Rauschens wird bei etwa 800 Hz erreicht und nimmt anschließend um etwa 10 dB alle 800 Hz ab. Für das TIMIT-Beispiel wurde eine Aufnahme aus einem Restaurant genutzt. Während das CCITT-Rauschen stationär ist, und somit insbesondere für eine beliebige betrachtete Zeitspanne eine ähnliche Aussteuerung im Zeitbereich aufweist, ist die Aufnahme aus dem Restaurant stark instationär, sodass im Allgemeinen deutliche Unterschiede in der Aussteuerung im Zeitbereich auftreten. Dies ist die Ursache für die augenscheinlich intensivere Erregung in Abb. 3.12b im Vergleich zu Abb. 3.12d. Die gezeigten Elektrodogramme sind sinnvoll für einen groben Überblick und Eindruck der Erregung in Folge eines akustischen Szenarios, jedoch ist diese Art der Visualisierung wesentlich



(a) Stimmhaftes Segment - rauschfrei

(b) Stimmloses Segment - rauschfrei

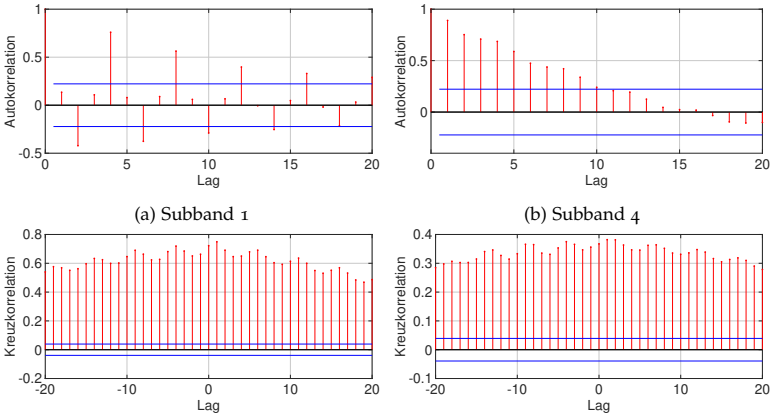
(c) Stimmhaftes Segment - 10 dB

(d) Stimmloses Segment - 10 dB

Abbildung 3.13: Ausschnitt des Zeitverlaufs der (a) Subbänder 1-4 für ein stimmhaftes Sprachsegment sowie der (b) Subbänder 21 und 21 für ein stimmloses Segment jeweils ohne Hintergrundrauschen. Die selben Segmente sind in (c) und (d) für den Fall mit Hintergrundrauschen dargestellt. Für beide Fälle wurde derselbe Satz wie in Abb. 3.12 verwendet.

zu grob für die Untersuchung und Entwicklung von Codierungsalgorithmen. Um einen etwas genaueren Eindruck der Erregungsmuster zu gewähren, wurde daher in Abb. 3.13 ein Segment der Erregungsmuster der ersten vier Subbänder dargestellt. Hierbei wurde ein stimmhafter[5] sowie ein stimmloser Ausschnitt des rauschfreien Signals aus Abb. 3.12 verwendet. Dargestellt sind für einen kleinen Zeitbereich der Verlauf der Stromwerte der Subbänder 1 bis 4 gemäß der Nummerierung aus Tabelle 2.1. Der Ausschnitt wurde gewählt, da er nahezu perfekt die für stimmhafte Laute überwiegende Periodizität der Erregungsmuster darstellt. Des Weiteren ist die Korrelation zwischen den Subbändern klar zu erkennen. Das Subband 1 springt in diesem Beispiel wiederholt auf den Wert Null. Dieser wurde im Beispiel zwecks Visualisierung an Stelle des Wertes NaN genutzt, und tritt im gezeigten Beispiel auf, da zu diesen Zeitpunkten ein anderes Subband einen größeren Stromwert aufweist, und daher das Subband 1 kurzzeitig nicht selektiert wird. Für das stimmhafte, rauschfreie Segment und die Subbänder 1 und 4 ist die normierte Autokorrelationsfunktion in Abb. 3.14, zusammen mit

5 Bei stimmhaften Lauten vibrieren qua Definition die Stimmbänder. Stimmhafte Laute umfassen insbesondere alle Vokale. Bei stimmlosen Lauten vibrieren die Stimmbänder nicht.



(a) Subband 1

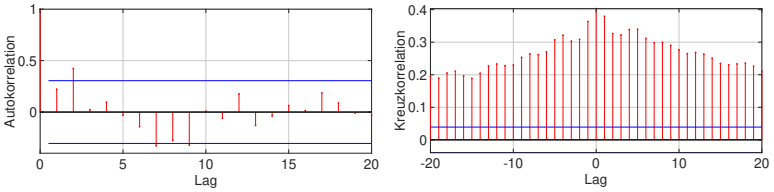
(b) Subband 4

(c) Subband 1 zu Subband 2

(d) Subband 1 zu Subband 4

Abbildung 3.14: Normierte Autokorrelationsfunktion für (a) das Subband 1 und (b) das Subband 4 der stimmhaften Segmente aus Abb. 3.13 sowie die Kreuzkorrelationsfunktion für (c) die Subbänder 1 und 2 sowie (d) die Subbänder 1 und 4. Es wurden jeweils die rauschfreien Segmente genutzt.

zugehörigen Kreuzkorrelationsfunktionen, dargestellt. In der Tat deutet die Kreuzkorrelation auf eine wesentliche Abhängigkeit zwischen den Subbändern hin, was genau dem Eindruck nach Abb. 3.13 entspricht. Exemplarisch ist in Abb. 3.15 der Effekt von Rauschen auf die Auto- und Kreuzkorrelation dargestellt. Hierbei wurde die Autokorrelation sowie Kreuzkorrelation für die diesmal rauschbehafteten, stimmhaften Segmente aus Abb. 3.13 berechnet. Sowohl die Autokorrelation als auch die Kreuzkorrelation sind deutlich gesunken. Dies ist wesentlich ein Effekt des Rauschens auf die Bandselektion. Hierdurch wechseln häufiger die selektierten Bänder und die statistischen Abhängigkeiten werden ausgeblendet. Des Weiteren haben die verwendeten Rauschsignale deutlich schwächere oder andersgelagerte statistische Abhängigkeiten, weswegen es zu solchen Änderungen in den Korrelationsfunktionen kommen kann. Je nach Rauschen und Signal-Rausch-Verhältnis kann es aber auch zum umgekehrten Fall kommen, dass die Bandselektion durch das Rauschen stabilisiert wird. Dies ist etwa bei niedrigem Signal-Rausch-Verhältnis und der Nutzung des CCITT-Rauschens teilweise der Fall. Die Frequenzen bis etwa 2400 Hz dominieren bei diesem Rauschsignal, wodurch es zu einer Fokussierung auf die unteren Subbänder kommt. Dieser Effekt kann in Abb. 3.12b am Ende des Erregungsmuster beobachtet werden.



(a) Subband 1 (b) Subband 1 zu Subband 2

Abbildung 3.15: (a) Normierte Autokorrelation des Subbands 1 sowie (b) die Kreuzkorrelation der Subbänder 1 und 2 für das stimmhafte Segment aus Abb. 3.13 bei einem Signal-Rausch-Verhältnis von 10 dB.

3.7 BESCHREIBUNG DER DURCHGEFÜHRTEN HÖRTESTS

Das Ziel der durchgeführten Hörtests war zum einen herauszufinden, ob und wann die durch den Electrocodec induzierten Verzerrungen zu einer Reduktion der Sprachverständlichkeit und der Sprachqualität führen, und gleichzeitig ein Vergleich mit einem state-of-the-art Audiocodec. Die durchgeführten Hörtests bestanden aus einem Sprachverständlichkeitstest und einem Sprachqualitätstest. Verständlichkeit und Qualität sind zwei im Wesentlichen unabhängige Größen, anhand derer Sprache bewertet werden kann. Etwa können roboterhafte Stimmen sehr gut verstanden werden und gleichzeitig nahezu beliebig schlechte Qualität aufweisen. Untersucht wurden in den Hörtests der Einfluss des in Abschnitt 3.1 eingeführten und beschriebenen Electrocodecs auf die Sprachverständlichkeit und -qualität der Stimulationsmuster von Cochlea-Implantaten im Vergleich zum Opus-Audiocodec, der in Abschnitt 2.3.11 beschrieben wurde. Im Rahmen dieser Hörtests wurden acht Testbedingungen definiert, die untersucht werden sollten. Diese sind in Tabelle 3.3 zusammengefasst. Die Bedingung **REF** bestand aus den Erregungsmustern, die durch Mischen von Sprach- und Rauschsignalen ohne weitere Verarbeitung der Signale erzeugt wurden. Hierbei lag also keinerlei Codierung der Erregungsmustern oder der Audiosignale vor.

Die Testbedingung **EC2** war der Elektrocodec mit einer Quantisierungsauflösung von 2 Bit, der vom Quantisierer jeder DPCM verwendet wurde. Analog dazu war die Testbedingung **EC3**, der Electrocodec mit einer Quantisierungsauflösung von 3 Bit, die vom Quantisierer verwendet wurde, und die Testbedingungen **EC4** und **EC7** wiesen einen 4-Bit- bzw. 7-Bit-Quantisierer auf. Die Testbedingungen **EC2** bis **EC4** wurden eingeführt, um die feineren Abhängigkeiten zwischen dem erzielten Sprachverständnis und der Auflösung des Quantisierers zu untersuchen. Die

Tabelle 3.3: Überblick über die in der Studie verwendeten Testbedingungen. Insgesamt wurden vier Bedingungen für den Electrocodec und drei Bedingungen für den Opus-Codec für die Bewertung ausgewählt. Eine Referenzbedingung, die dem Originalton ohne Anwendung eines Codecs entspricht, wurde ebenfalls einbezogen. Die angegebene Latenz ist die algorithmische Latenz. Die mittlere Bitrate wurde bei einem Signal-Rausch-Verhältnis von 0 dB auf dem HSM bestimmt.

Bezeichner	Beschreibung	Latenz
REF	Referenzbedingung. Unbearbeitetes Signal.	-
EC2	Electrocodec mit 4 Quantisierungsstufen in der DPCM (2 Bit). Mittlere Bitrate: 24,3 kbit/s	0 ms
EC3	Electrocodec mit 8 Quantisierungsstufen in der DPCM (3 Bit). Mittlere Bitrate: 30,6 kbit/s	0 ms
EC4	Electrocodec mit 16 Quantisierungsstufen in der DPCM (4 Bit). Mittlere Bitrate: 37,6 kbit/s	0 ms
EC7	Electrocodec mit 128 Quantisierungsstufen in der DPCM (7 Bit). Mittlere Bitrate: 53,5 kbit/s	0 ms
Opus16c	Opus mit konstanten 16 kbit/s. Mittlere Bitrate: 16 kbit/s	5 ms
Opus16v	Opus mit variablen 16 kbit/s. Mittlere Bitrate: 31 kbit/s	5 ms
Opus52v	Opus mit variablen 52 kbit/s. Mittlere Bitrate: 57,9 kbit/s	5 ms

Testbedingung **EC7** wurde als Rückfalltestbedingung eingeführt. Hätten die Bedingungen **EC2** bis **EC4** alle eine schlechte Verständlichkeit ihrer codierten Erregungsmuster gezeigt, hätte die Leistung der Testbedingung **EC7** verwendet werden können, um festzustellen, ob der gewählte Ansatz überhaupt funktionieren kann. In diesem Fall ist die durch die Kompression induzierte Verzerrung so gering, dass eine Reduktion der Sprachverständlichkeit oder -qualität a priori unwahrscheinlich anmutete. Des Weiteren ist die Bitrate der **EC7** Testbedingung ähnlich wie die der **Opus52v** Testbedingung, was einen besseren Vergleich ermöglicht.

Die Testbedingung **Opus16c** wurde eingeführt, um eine Codec-Einstellung zu untersuchen, bei der eine Verschlechterung des Sprachverständnisses zu erwarten ist. Um 16 kbit/s bei einer algorithmischen Latenz von 5 ms zu erreichen, wurde für diese Testbedingung eine konstante Bitrate erzwungen. Anderenfalls ist die tatsächliche Bitrate bei diesen Latenzen, aufgrund der standardmäßig genutzten variablen Bitrate von Opus, deutlich höher als die nominelle.

Die Testbedingung **Opus52v** ergab aufgrund der genutzten variablen Bitrate eine mittlere Bitrate für das Sprachmaterial des Hörtests zwischen 58 kbit/s und 60 kbit/s, abhängig vom Pegel des Hintergrundrauschens.

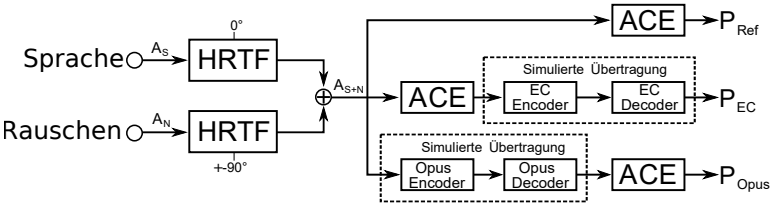


Abbildung 3.16: Blockschaltbild der Datengenerierung des Hörtests. Sprach- und Rausch(audio-)signale werden von kopfbezogenen Übertragungsfunktionen (HRTF) verarbeitet und anschließend gemischt. Die Erregungsmuster aller Testbedingungen werden auf Basis dieser Signalmischung erzeugt.

Es wurde erwartet, dass diese Testbedingung keine wahrnehmbare Verschlechterung der Verständlichkeit und Qualität bewirken würde.

Die Testbedingung **Opus16v** wurde nur bei den Probanden ID5 bis ID10 getestet (siehe Tabelle 3.4) und erreichte eine mittlere Bitrate von etwa 31 kbit/s. Dieser Codec wurde eingeführt, weil bei den ersten vier Probanden die **Opus16c** Testbedingung eine sehr schlechte Sprachleistung erbrachte, wobei viele Probanden nur wenige Wörter verstanden und die meisten eine Worterkennungsrate von unter 20 % erzielten. Der einzige Unterschied zwischen der **Opus16v** Testbedingung und der **Opus16c** Testbedingung besteht darin, dass die **Opus16v** Testbedingung auf 16 kbit/s mit variabler Bitrate eingestellt ist. Hierdurch wird die Wahl der tatsächlichen Bitrate zum Teil Opus überlassen, wodurch es zu einem deutlichen Anstieg der Bitrate und der Qualität kam. Es wurde erwartet, dass diese Testbedingung deutlich besser abschneiden würde als die **Opus16c** Testbedingung.

3.7.1 *Datengenerierung*

Die Datengenerierung ist genauer in Abb. 3.16 aufgeschlüsselt. Als Datenmaterial wurde der HSM-Satztest verwendet. Das ursprüngliche, unbearbeitete Audiosignal wurde hierbei mittels Kopfübertragungsfunktionen mit CCITT-Rauschen gemischt, welches sprachartiges Rauschen ist. Das Signal- zu Rauschverhältnis wurde an den jeweiligen Probanden angepasst. Es gibt drei verschiedene Erstellungspfade der Erregungsmuster. Die Referenzerregungsmuster P_{ref} wurden durch Anwendung des Advanced Combination Encoders auf das Geschmisch aus Sprache und Rauschen, welches mit Hilfe der Kopfübertragungsfunktionen erzeugt wurde, generiert. Es erfolgte keine Bearbeitung durch einen Audiocodec. Für die Erregungsmuster des Electrodocec wurde zunächst wie bei den

Tabelle 3.4: Demografische Daten der Probanden des durchgeführten Hörtests sowie das Signal-Rausch-Verhältnis, das bei den Tests verwendet wurde. Die getestete Seite war immer das bessere Ohr. Mindest-, Höchst- und Medianwerte des Dynamikbereichs der Probanden in klinischen Einheiten (CU) sind ebenfalls angegeben.

ID	Geschlecht (Alter)	Getestete Seite	Elektrodentyp	Zahl der aktiven Elektroden	Dynamikumfang (CU) (min/max/median)	Signal-Rausch-Verhältnis (dB)
ID01	M (82)	Rechts	CI512	22	31/48/44	20
ID02	M (66)	Rechts	CI24R (CA)	20	57/77/70.5	8
ID03	M (76)	Links	CI522	22	30/76/57	1
ID04	M (73)	Rechts	CI24RE	21	36/84/69	3
ID05	M (72)	Rechts	CI24RE	20	43/54/52	5
ID06	F (71)	Rechts	CI24RE	20	40/64/51.5	5
ID07	M (50)	Rechts	CI24RE	22	61/74/69.5	0
ID08	M (78)	Rechts	CI512	19	52/66/65	6
ID09	F (49)	Rechts	CI522	20	51/71/65.5	3
ID10	F (76)	Rechts	CI522	20	37/64/54	0

Referenzerregungsmustern vorgegangen, jedoch wurden im Anschluss diese Erregungsmuster mit dem Electrocodec komprimiert und dekomprimiert. Diese resultierenden Erregungsmuster P_{EC} würden bei einer drahtlosen Übertragung ebenso vom Electrocodec erzeugt werden. Analog wurde für Opus vorgegangen, jedoch wurden die Audiosignale, da es sich um einen Audiocodec handelt, mittels Opus komprimiert und dekomprimiert, und dieses dekomprimierte Signal wurde dann mittels des Advanced Combination Encoders in Erregungsmuster umgewandelt. Dies resultierte in den Erregungsmustern P_{Opus}. Die Erregungsmuster P_{Ref}, P_{EC} sowie P_{Opus} wurden, erzeugt mit den entsprechenden Codeceinstellungen bzw. Bitraten, dann direkt an den Signalprozessor der Probanden weitergeleitet.

3.7.2 *Probanden*

Insgesamt nahmen zehn Probanden mit Cochlea-Implantat an der Studie teil, von denen sieben Männer und drei Frauen waren. Das Durchschnittsalter der Teilnehmer betrug 69,3 Jahre. Bis auf einen Probanden war das bessere Ohr, im Sinne des erzielten Sprachverstehens, immer das rechte Ohr. Detaillierte Informationen über die Teilnehmer sind in Tabelle 3.4 aufgeführt. Alle Probanden gaben ihr Einverständnis zu dem Projekt, das vom institutionellen Überprüfungsgremium der Medizinischen Hochschule Hannover genehmigt wurde.

3.7.3 *Testprozedur*

Mit jedem Probanden wurden zwei Experimente durchgeführt. Das erste Experiment war der Sprachverständlichkeitstest und das zweite Expe-

riment war der Sprachqualitätstest. Der Sprachqualitätstest wurde als sogenannter MUSHRA-Test (multiple stimuli with hidden reference and anchor) durchgeführt [Int15].

Alle Tests wurden monaural durchgeführt, wobei im Falle von bilateralen Cochlea-Implantaten das Ohr mit der besten Leistung verwendet wurde. Wenn ein Proband eine andere Kanalstimulationsrate als die für den Hörtest festgelegte Kanalstimulationsrate von 900 Pulsen pro Sekunde verwendete, wurde zunächst eine Anpassung der Strompegel mit der neuen Kanalstimulationsrate durchgeführt. Bei der Anpassung wurden die Werte THR und MCL (siehe Abschnitt 2.2.1.1) zunächst um einen konstanten Wert verringert, sodass die Wahrnehmung der Stimuli zunächst kaum hörbar war, und anschließend schrittweise erhöht. Vor jeder Erhöhung wurde der Versuchsperson ein Beispielsatz aus dem HSM vorgespielt und die Versuchsperson gebeten, die wahrgenommene Lautstärke zu bewerten. Die wahrgenommene Lautstärke wurde auf einer Skala von 0 (Stille) bis 10 (schmerzhaft laut) Punkten eingestuft, wobei angestrebt wurde, etwa 6 Punkte zu erreichen, was einem angenehmen Lautstärkeniveau entspricht, bei dem der Proband keine Schwierigkeiten hat, das präsentierte Sprachmaterial zu verstehen. Nach der Anpassungsprozedur blieben die neu gefundenen Werte der THR und MCL in allen durchgeführten Experimenten fest und unverändert. Nur bei zwei Probanden war eine derartige Anpassungsprozedur notwendig.

3.7.4 *Sprachverständlichkeitstest*

Die in Abschnitt 2.2.3 beschriebene Worterkennungsrate wurde für jede Testbedingung ermittelt, um die Verständlichkeit der assoziierten Stimulationsmuster der jeweiligen Codecs zu testen. Für jede Testbedingung wurden zwei Listen des HSM verwendet. Die Listen wurden in zufälliger Reihenfolge präsentiert, ohne dass die Teilnehmer oder die Versuchsleiter wussten, welche Bedingung präsentiert wurde ("Doppelblind"). Zusätzliche Listen in der **REF** Testbedingung wurden verwendet, um den Teilnehmer zunächst zu trainieren. Initial kann ansonsten das Sprachverstehen in einer ungewohnten Situation sehr schlecht sein.

Dann wurde das Signal-Rausch-Verhältnis schrittweise verringert, um den Geräuschpegel zu ermitteln, bei dem die Probanden etwa 70 % der Wörter verstanden, um Boden- und Deckeneffekte zu vermeiden. Diese Listen wurden von dem eigentlichen Sprachverständnistest ausgeschlossen. Im Anschluss wurde dann der eigentliche Test wie beschrieben durchgeführt.

3.7.5 *Sprachqualitätstest*

Die Sprachqualität der Stimulationsmuster nach Codierung wurde mittels eines MUSHRA-Tests evaluiert. Bei einem MUSHRA-Test wird den Probanden ein Referenzsignal präsentiert und dazu eine gewisse Zahl an Testsignalen, welche verdeckt präsentiert werden, d.h., ohne dass bekannt ist, welcher Testbedingung das jeweilige Signal zugeordnet ist. Für das Referenzsignal wurden Sätze in der **REF** Testbedingung vorgespielt. Zur Prüfung einer ordnungsgemäßen Durchführung durch den Probanden sind den verdeckten Signalen das Referenzsignal selber als auch ein sogenanntes Ankersignal beigefügt. Letzteres ist das Referenzsignal in sehr schlechter Qualität. Dieses sollte vom Probanden als sehr schlecht bewertet werden. Das verdeckte Referenzsignal sollte entsprechend sehr gut bewertet werden. Aufgabe des Probanden ist dann, die verdeckten Signale im Vergleich zur Referenz hinsichtlich der Qualität zu bewerten. Die bestmögliche Bewertung ist hierbei kein wahrnehmbarer Unterschied zur Referenz. Alle Signale können vom Probanden beliebig oft abgespielt werden.

Der MUSHRA-Test wurde mit rauschfreiem Sprachmaterial durchgeführt. Der Test wurde rauschfrei durchgeführt, da eine Qualitätsverschlechterung schwer einzuschätzen ist, wenn auch erhebliche Hintergrundgeräusche vorhanden sind. Für den MUSHRA-Test wurden sechs Sätze aus dem HSM verwendet, die zuvor nicht im Sprachverständlichkeitstest präsentiert worden sind. Insgesamt wurden acht Testbedingungen getestet. Die vier Testbedingungen des Electrocodec, die **Opus16c** und **Opus52v** Testbedingungen des Opus-Codec, eine für die versteckte Referenz und eine für den versteckten Anker. Die **Opus16v** Testbedingung wurde nicht mit aufgenommen, da die Zahl der Testbedingungen des MUSHRA-Tests bereits sehr groß war und zudem nach den Ergebnissen der ersten vier Probanden kein Erkenntnisgewinn durch Ergänzung einer weiteren Testbedingung zu erwarten war. Der Nachteil der Einführung einer weiteren Testbedingung im Verlauf der Studie, die zu einer stärkeren Ermüdung der Probanden hätte führen können, überwog daher.

Das Ankersignal, verstanden als eigenständige Testbedingung und als solche bezeichnet mit **Anker**, wurde mit dem Electrocodec mit zwei Quantisierungsstufen (1 Bit) und anschließender Deaktivierung aller Teilbänder, die Frequenzen von 850 Hz und höher codieren, erstellt. Dadurch wurde eine sehr schlechte Qualität des Ankers sichergestellt. Der MUSHRA-Test wurde für jede Versuchsperson zweimal durchgeführt, sodass für jede Versuchsperson zwölf Bewertungen der wahrgenommenen Sprachqualität für jede Testbedingung aufgezeichnet wurden.

4

ERGEBNISSE

Dieses Kapitel stellt die Leistungsfähigkeit der entwickelten Codierungsverfahren vor. Zunächst wird die Evaluierung des Electrocodecs, dem einzigen entwickelten konventionellen Codec, inklusive einer detaillierten Diskussion der durchgeführten Hörtests vorgestellt. Anschließend werden die erzielten Ergebnisse der Codecs auf Basis künstlicher neuronaler Netze besprochen, beginnend mit dem einzigen verlustlosen Kompressionsalgorithmus. Nachfolgend wird die Leistungsfähigkeit von Autoencodern ohne und mit Rückkopplung erörtert.

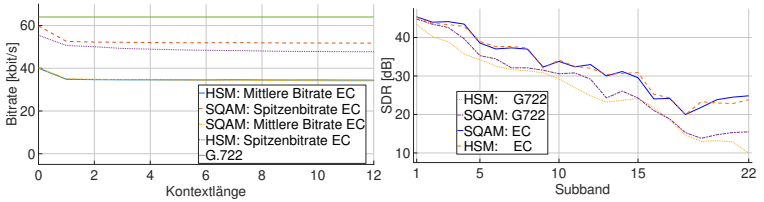
4.1 OBJEKTIVER VERGLEICH DES ELECTROCODECS MIT DEM G.722

Der Electrocodec wurde zunächst mit dem G.722, einem sehr bekannten Audiocodec, der bei der drahtlosen Übertragung von Audio im kontralateralen Routing von Signalen (CROS) zum Einsatz kommt und eine algorithmische Latenz von 1,3 ms aufweist, verglichen [Hin+19]. Als Vergleichsmaß wurde das Signal-Verzerrungs-Verhältnis (SDR) genutzt, berechnet gemäß

$$\mathrm{SDR}_C(z) = 10 \cdot \log_{10}\left(\frac{\sum_m P^2_{Clean}(z,m)}{\sum_m (P_{Clean}(z,m) - P_C(z,m))^2}\right) \tag{4.1}$$

mit $C = \{EC, G.722\}$ und dem Bandindex z. Die $P(z,m)$ entsprechen den Signalen aus Gl. 2.7. Das SDR ist die Signalleistung im Verhältnis zum Rekonstruktionsfehler des jeweiligen Codecs für das jeweilige Subband in dB. Als Sprachmaterial wurden sowohl die Sprachaufnahmen des SQAM-Datensatzes als auch der HSM-Satztest verwendet, die in diesem Vergleich ohne Rauschen verwendet wurden. Stille wurde in den Erregungsmustern am Anfang und am Ende entfernt.

Abb. 4.1a zeigt sowohl die mittlere als auch die Spitzenbitrate des Electrocodecs im Vergleich zum G.722 für den HSM-Datensatz sowie für den SQAM-Datensatz. Hierbei wurde für den Electrocodec die Bitrate



(a) Bitrate über Kontextlänge

(b) SDR pro Band des G.722 und des EC

Abbildung 4.1: (a) Bitrate des Electrocodecs (EC) und des G.722 Audiocodecs für das HSM- und SQAM-Sprachmaterial in Abhängigkeit von der gewählten Kontextlänge der arithmetischen Codierung. Eingezeichnet ist sowohl die Spitzenbitrate als auch die mittlere Bitrate. Der G.722 codiert mit konstanter Bitrate. Es zeigten sich über eine Kontextlänge von eins hinaus nur minimale Verbesserungen der Datenrate. (b) Signal-Verzerrungs-Verhältnis (SDR) des G.722 und des Electrocodecs (EC) für jedes der 22 Subbänder des Cochlea-Implantats. Die Verzerrung je Band war immer geringer für den Electrocodec als für den G.722.

in Abhängigkeit der Kontextlänge der arithmetischen Codierung, erläutert in Abschnitt 3.1, angegeben. Da der G.722 keine Entropiecodierung verwendet, ist seine Bitrate, anders als die des Electrocodecs, konstant weswegen, für einen faireren Vergleich, nicht nur die mittleren Bitraten, sondern ebenfalls die Spitzenbitraten abgebildet sind. Diese sind das Maximum der Bitrate über Ausschnitte einer Dauer von einer Sekunde.

Die Bitrate des G.722 beträgt konstante 64 kbit/s, wohingegen die mittlere Bitraten des Electrocodecs auf dem HSM- und SQAM-Datensatz für Kontextlängen der arithmetischen Codierung größer Eins in etwa 34,8 kbit/s beträgt. Die Spitzenbitrate war dagegen deutlich höher mit 59,9 kbit/s (HSM) und 55,4 kbit/s (SQAM) bei einer Kontextlänge von Null sowie 52,6 kbit/s (HSM) und 50,6 kbit/s (SQAM) für eine Kontextlänge von Eins. Grundsätzlich ergab sich, dass die arithmetische Codierung die Bitrate signifikant senken konnte. Insbesondere der Sprung von einer Kontextlänge Null zum Kontext der Länge Eins reduzierte die Bitrate deutlich. Kontexte von größerer Länge reduzierten die Bitrate nicht mehr substantiell.

Abb. 4.1b ist zu entnehmen, dass für jeden Kanal und jeden Datensatz die durch den Electrocodec induzierte Verzerrung der Erregungsmuster geringer ausfällt als jene des G.722. Insbesondere für Subbänder mit höherer Mittenfrequenz, d.h. die Subbänder 17 und höher, ergab sich eine deutliche Überlegenheit des Electrocodecs. Dies liegt am Codierungsprinzip des G.722, welcher ab etwa 4000 Hz Signalinformationen mit deutlich reduzierter Datenrate codiert, wodurch sich die Qualität ab diesem Punkt

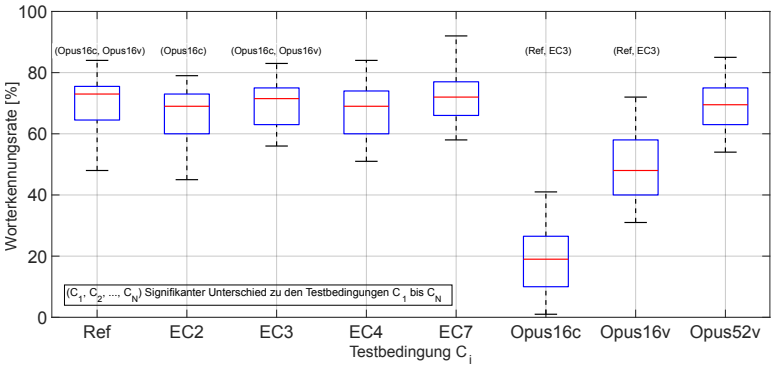


Abbildung 4.2: Worterkennungsraten für die getesteten Codecs sowie die Referenzbedingung (**REF**). Über den Boxplots sind die Testbedingungen notiert, zu denen ein statistisch signifikanter Unterschied durch einen Wilcoxon-Vorzeichen-Rangtest festgestellt wurde. Die **Opus16v** Testbedingung wurde nach dem vierten Probanden eingeführt, da die **Opus16c** Testbedingung eine sehr schlechte Sprachverständlichkeit zeigte.

verschlechtert. Insgesamt zeigte sich also, dass der Electrocodec sowohl geringere Bitraten als auch geringere Verzerrungen und somit eine bessere Leistung erzielt als der G.722 Audiocodec. In dieser ersten Version des Electrocodecs gab es zwei wesentliche Unterschiede zur nachfolgend an Probanden getesteten: Zum einen wurden, um den objektiven Vergleich mit dem G.722 zu ermöglichen, die Quantisierer der Subband-DPCMs mit unterschiedlichen Quantisierungsstufen durchgeführt. Die unteren Subbänder, welche mit niedrigen Frequenzen korrespondieren, nutzten 6-Bit-Quantisierung, während die höchsten 2-Bit-Quantisierung nutzten. Dazwischen wurde graduell abgestuft.

Zum anderen wurde mit dem zusätzlichen Bit B1 des Gesamtbitstroms des Electrocodecs (vergleiche Abb. 3.1), in dieser frühen Version des Electrocodecs codiert, ob sich die Bandselektion zum vorherigen Frame geändert hat. Ist dies nicht der Fall, so ist die Übertragung der Bandselektion nicht notwendig, wodurch einige Bits gespart werden können. Jedoch ist dieses Verfahren nur bei relativ hohem Signal-Rausch-Verhältnis, d.h. bei geringerem Rauschen, nützlich, da sich ansonsten die Bandselektion zu oft ändert und auf diese Weise selten Bits gespart werden können.

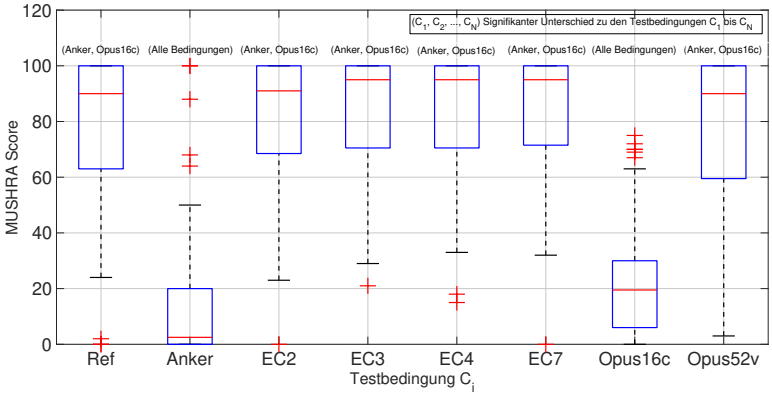


Abbildung 4.3: MUSHRA Scores der untersuchten Codecs sowie der Referenz- und Ankertestbedingung. Die verdeckte Referenz sowie der Anker wurden korrekt bewertet. Bis auf die **Opus16c** Testbedingung zeigte sich kein Unterschied der Codecs hinsichtlich der Sprachqualität. Über den Boxplots sind die Testbedingungen notiert, zu welchen ein signifikanter Unterschied besteht, festgestellt durch einen Wilcoxon-Vorzeichen-Rangtests.

4.2 ERGEBNISSE DER HÖRTESTS

Der Electrocodec wurde in einem zweiten Schritt an CI-Probanden getestet [Hin+21a]. Die Untersuchungen wurden 2019 und 2020 im Deutschen Hörzentrum durchgeführt. Die Versuchsbeschreibung erfolgte in Abschnitt 3.7. Boxplots der Worterkennungsraten für die jeweiligen Einstellungen sind in Abb. 4.2 dargestellt, wobei die Ergebnisse aller Probanden genutzt wurden. Die **Opus16v** Testbedingung wurde nur in den letzten sechs Probanden evaluiert. Ein separater Boxplot, der die Ergebnisse nur für diese Probanden zeigt, ist im Anhang in Abb. 8.1 dargestellt. Qualitativ ergab sich für diese Probandenteilmenge kein Unterschied im Testergebnis. Die genauen Werte der mittleren erzielten Worterkennungsraten der Probanden für die jeweilige Einstellung, gemittelt über die jeweils zwei Satzlisten, ist im Anhang in Tabelle 8.1 angegeben. Boxplots der MUSHRA Scores sind in Abb. 4.3 dargestellt. Ein MUSHRA Score von 100 bedeutet maximale Qualität, d.h. keinen feststellbaren Unterschied zur Referenz. Ein MUSHRA Score von Null bedeutet minimale Qualität, d.h. maximalen Unterschied zur Referenz. Rote Kreuze markieren Ausreißer, welche jeweils die Bewertung eines einzelnen Satzes durch einen Probanden darstellen. Die Ankerbedingung wurde mit wenigen Ausnahmen sicher von den Probanden der Studie erkannt. Gleiches gilt

Tabelle 4.1: Ergebnisse des Wilcoxon-Vorzeichen-Rangtests für die Ergebnisse des Sprachverständlichkeitstests. Dargestellt sind die verglichenen Bedingungen in den mit *A* und *B* bezeichneten Spalten und die berechneten p-Werte in der mit *p* bezeichneten Spalte. Die fettgedruckten Werte zeigen signifikante Werte nach Anwendung der Bonferronikorrektur auf das Signifikanzniveau von $p < 0,05$.

Testbedingung		
A	B	p
EC2	**REF**	0,048
EC3	**REF**	0,575
EC4	**REF**	0,198
EC7	**REF**	0,614
Opus16c	**REF**	**< 0,001**
Opus16v	**REF**	**< 0,001**
Opus52v	**REF**	0,433

Testbedingung		
A	B	p
EC2	**Opus16c**	**< 0,001**
EC2	**Opus16v**	0,004
EC3	**Opus16v**	**0,002**
EC2	**Opus52v**	0,211
EC3	**Opus52v**	0,809
EC4	**Opus52v**	0,279
EC7	**Opus52v**	0,239

für die Referenzbedingung. Offensichtlich ist der deutliche Qualitätsunterschied aller Codectestbedingungen zur **Opus16c** Testbedingung, welche nahezu so schlecht wie die Ankerbedingung bewertet wurde. Im Median wurde die Qualität der **EC3**, **EC4** und **EC7** Testbedingungen besser bewertet als die Referenz. Dies gilt weniger stark auch für die **EC2** Testbedingung. Diese Unterschiede sind jedoch nicht signifikant und mutmaßlich auf mangelnde Sorgfalt der Probanden zurückzuführen. Eine einseitige Varianzanalyse mit wiederholten Messungen (ANOVA) sowie ein Wilcoxon-Vorzeichen-Rangtest wurden durchgeführt, um die Ergebnisse des Sprachverständnistests zu untersuchen. Die ANOVA ergab einen signifikanten Effekt der Testbedingungen mit $F(7, 63) = 71,98$ und $p < 0,001$. Vierzehn Wilcoxon-Vorzeichen-Rangtests wurden durchgeführt, um die Medianunterschiede zwischen den einzelnen Paaren der Testbedingungen zu untersuchen. Die Ergebnisse sind in Tabelle 4.1 dargestellt. Fettgedruckte Werte zeigen signifikante Unterschiede nach Anwendung der Bonferronikorrektur, die in Abschnitt 2.4.1 erläutert wurde, auf das Signifikanzniveau von $p < 0,05$. Der neue Schwellenwert für die Signifikanz nach Anwendung der Bonferronikorrektur betrug $\frac{p}{14} = 0,00357$. Demnach erzielte die **EC2** Testbedingung signifikant bessere Sprachverständlichkeit als die **Opus16c** Testbedingung. Ferner erzielte die **EC3** Testbedingung eine signifikant bessere Sprachverständlichkeit als die **Opus16v** Testbedingung. Es konnte knapp kein signifikanter Unterschied in der Sprachverständlichkeit zwischen der **EC2** und **Opus16v** Testbedingung festgestellt werden ($p = 0,004$ und Signifikanzniveau nach Bonferronikorrektur ca. $0,0036$). Ein signifikanter Unterschied könnte womöglich bei weiteren Untersuchungen festgestellt werden. Eine Tendenz ist aus Abb. 4.2 ersichtlich. Weitere signifikante Unterschiede konnten

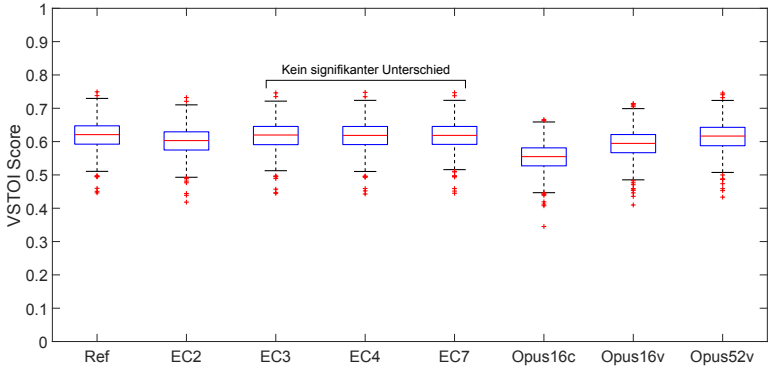
nach Bonferronikorrektur nicht festgestellt werden. Die Ergebnisse dieser Signifikanzuntersuchung sind über den jeweiligen Boxen in Abb. 4.2 eingetragen. Für den Sprachqualitätstest wurde ebenfalls eine ANOVA sowie ein Wilcoxon-Vorzeichen-Rangtest durchgeführt. Die ANOVA fand einen signifikanten Effekt der Testbedingungen $F(7, 63) = 264,1389$ mit $p < 0,001$. Der Wilcoxon-Vorzeichen-Rangtest fand signifikante Unterschiede ($p < 0,001$) zwischen der **Anker** Testbedingung und allen anderen Testbedingungen inklusive der **Opus16c** Testbedingung. Selbiges gilt für ebendiese, d.h., auch für die **Opus16c** Testbedingung fand der Wilcoxon-Vorzeichen-Rangtest signifikante Unterschiede hinsichtlich des MUSHRA Scores zu allen anderen Testbedingungen ($p < 0,001$).

Es wurde kein signifikanter Unterschied zwischen der **EC2** Testbedingung und der **REF** Testbedingung ($p > 0,05$) festgestellt.

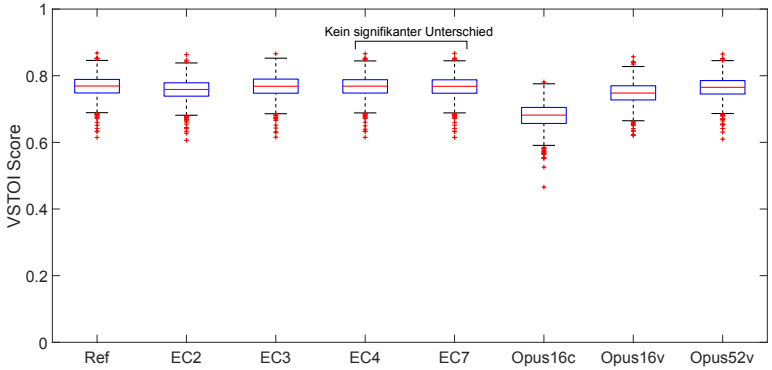
Kombiniert man dieses Ergebnis mit den mittleren Bitraten der jeweiligen Testbedingung aus Tabelle 3.3, so erreicht der Electrocodec folglich bessere oder gleiche Sprachverständlichkeit der codierten Stimulationsmuster bei niedriger oder gleicher Bitrate. Die Latenz des Electrocodecs ist zudem jener von Opus überlegen mit einer algorithmischen Latenz von 0 ms für den Electrocodec und einer minimalen algorithmischen Latenz von 5 ms für Opus.

Zur Nutzung von VSTOI für die Optimierung insbesondere weiterer Kompressionsalgorithmen wurden im Kontext der beschriebenen Hörtests die VSTOI-Werte der Testbedingungen für Signal-Rausch-Verhältnisse von 0 dB und 10 dB mittels Boxplots visualisiert. Hierdurch ist eine qualitative Einschätzung der Bewertung von VSTOI mit der erzielten Worterkennungsrate von CI-Trägern möglich. Abb. 4.4 zeigt Boxplots der VSTOI-Werte der jeweiligen Testbedingung für den gesamten HSM-Satztest. Die Median VSTOI-Werte für diese Einstellungen sind in Tabelle 4.2 notiert. Qualitativ fängt die Bewertung durch VSTOI den Einfluss der Codierung auf die Verständlichkeit der Erregungsmuster sehr gut ein. Der größte Abfall wird korrekt für die **Opus16c** Testbedingung angezeigt. VSTOI legt einen minimalen Unterschied zwischen der **REF** Testbedingung und der **EC2** Testbedingung nahe, was sich mit dem durchgeführten Hörtest dahingehend deckt, dass ein leicht negativer Effekt der **EC2** Testbedingung auf das Sprachverstehen zwar nicht statistisch nachgewiesen werden konnte, jedoch als Tendenz erkennbar war.

In Tabelle 4.2 sind zudem Worterkennungsraten angegeben, die über eine parametrisierte logistische Funktion aus den VSTOI-Werten abgeleitet worden sind. Hierzu wurden aus den Testergebnissen der Probanden in der **REF** Testbedingung die Parameter der logistischen Funktion gemäß dem Vorgehen aus [Taa+10] bestimmt. Da jedoch die Varianz der Ergebnisse in der **REF** Testbedingung nicht wesentlich verschieden waren,



(a) 0 dB



(b) 10 dB

Abbildung 4.4: VSTOI-Werte der von den untersuchten Codecs codierten Erregungsmuster sowie der Referenzeinstellung bei (a) 0 dB Signal-Rausch-Verhältnis und (b) 10 dB Signal-Rausch-Verhältnis. Qualitativ gibt der VSTOI-Wert die tatsächlich erzielte Sprachverständlichkeit der Codecs in den Hörtests wieder. Alle Unterschiede zwischen den Einstellungen waren signifikant, bis auf jene, für welche keine Signifikanz notiert ist.

Tabelle 4.2: Median VSTOI-Werte aller Testbedingungen über den gesamten HSM-Satztest für ein Signal-Rausch-Verhältnis von 0 dB und 10 dB. Die VSTOI-Werte wurden wie in [Taa+10] beschrieben auf Worterkennungsraten (WER) mittels der parametrisierbaren logistischen Funktion nach Gl. 2.2.3.1 abgebildet.

	0 dB		10 dB	
Testbedingung	Median VSTOI-Wert	WER [%]	Median VSTOI-Wert	WER [%]
REF	0,621	66,8	0,77	81,1
EC2	0,603	64,8	0,759	80,2
EC3	0,62	66,7	0,768	81,0
EC4	0,619	66,6	0,769	81,0
EC7	0,619	66,6	0,769	81,0
Opus16c	0,555	59,0	0,682	73,3
Opus16v	0,595	63,8	0,748	79,3
Opus52v	0,617	66,4	0,765	80,7

ist die Abbildung von mittleren VSTOI-Werten auf mittlere Worterkennungsraten mit großer Vorsicht zu genießen und dient lediglich der groben Übersicht und als grobe Richtschnur für die Interpretation von VSTOI-Unterschieden bei der Codierung der Erregungsmuster mittels Autoencoder, deren Ergebnisse in Abschnitt 4.4 vorgestellt werden. Jedoch kann man auch dieser Abbildung auf mittlere Worterkennungsraten eine starke Abhängigkeit von den VSTOI-Werten entnehmen, d.h. kleine Änderungen im VSTOI-Wert führen zu großen Änderungen in den zugehörigen mittleren Worterkennungsraten. Dies deckt sich mit anderen Veröffentlichungen [Taa+10; WSS18a]. Ausgehend von der **Opus52v** Testbedingung, deren Bitrate so hoch ist, dass keine Reduktion der Verständlichkeit zu erwarten ist, kann man schätzen, dass eine Reduktion des mittleren VSTOI-Werts von etwa 0,005 keinen messbaren Effekt auf die Sprachverständlichkeit der Erregungsmuster darstellt. Bis zu einer VSTOI-Differenz von 0,01 dürfte der Einfluss nahezu vernachlässigbar bleiben.

4.3 EVALUIERUNG DER VERLUSTLOSEN KOMPRESSION

Die verlustlose Kompression der Erregungsmuster mittels künstlicher neuronaler Netze wie in Abschnitt 3.4 beschrieben wurde auf dem TIMIT-Datensatz, dessen Generierung in Abschnitt 3.5 erläutert wurde, durchgeführt. Im weiteren Verlauf wird dieser entworfene, latenz- und verlustlose Kompressionsalgorithmus mit dem Kürzel ZDLLC (kurz für engl. zero delay lossless codec) bezeichnet. Es wurde der Einfluss der Netzstruktur, der Kontexte sowie der Bufferlänge auf die Kompressionsleistung untersucht. Ferner wurde der ZDLLC mit bekannten verlustlosen Kom-

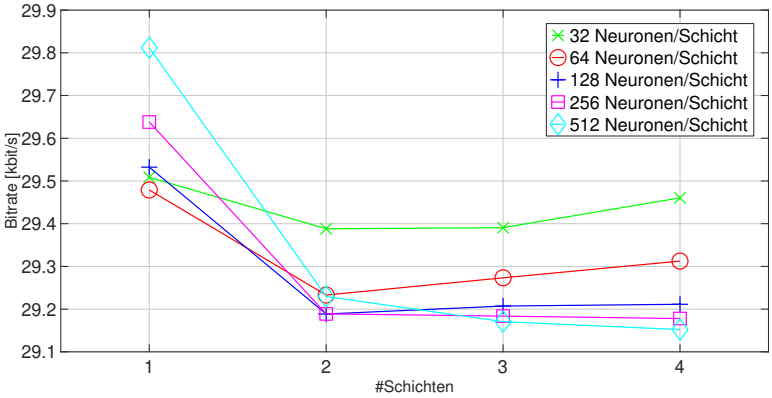


Abbildung 4.5: Erzielte mittlere Bitrate in kbit/s des ZDLLC auf dem Testdatensatz in Abhängigkeit von der Netzgröße. Genutzt wurden alle Kontexte aus Tabelle 3.1 exklusive der Langzeitkontexte. Der Dynamikumfang des Datensatzes betrug 6 Bit. Bis auf die niedrigste Zahl an Neuronen und dem Wechsel von einer auf zwei Schichten zeigt sich nur eine geringe Größenabhängigkeit.

pressionsverfahren verglichen. Hierzu wurden sowohl feste Auflösungen von 2 Bit/Symbol bis 6 Bit/Symbol je Elektrode gewählt, als auch echte mittlere Dynamikbereiche je Elektrode von CI-Trägern aus der Literatur verwendet, um die tatsächliche Kompressionsleistung abzuschätzen [Hin+22a]. Als Aktivierungsfunktion wurde in jedem Fall und jeder Schicht die Sigmoidfunktion genutzt. Trainiert wurden die künstlichen neuronalen Netze in jedem Fall auf dem Trainingsdatensatz. Die erzielte Bitrate auf einer Teilmenge von 200 Dateien des Testdatensatzes bei einem Dynamikumfang von 6 Bit/Symbol ist in Abb. 4.5 dargestellt. Ohne Kompression liegt die Bitrate bei 6 Bit/Symbol bei 118,8 kbit/s.

Hierbei wurde lediglich auf 200 Dateien evaluiert, da die Zahl der Untersuchungen, deren einzelne Berechungsdauer bereits erheblich gewesen ist, recht groß war. Hierdurch ließ sich die Bearbeitungsdauer deutlich reduzieren, ohne wesentliche Informationen zu verlieren. Es ergab sich eine minimale Bitrate von etwa 29,15 kbit/s unter Verwendung von vier Schichten und 512 Neuronen je Schicht. Dies ist jedoch nur marginal besser als eine Bitrate von etwa 29,19 kbit/s, welche mit nur zwei Schichten und 256 Neuronen je Schicht erzielt werden konnte. Insgesamt ergab sich nur beim Schritt von einer zu zwei Schichten ein wesentlicher Einfluss der Schichtanzahl. Bis auf die kleinste Konfiguration hatte die Zahl der Neuronen einen vernachlässigbaren Einfluss auf die erzielte Bitrate. Der Einfluss lag hierbei unterhalb von 0,1 kbit/s.

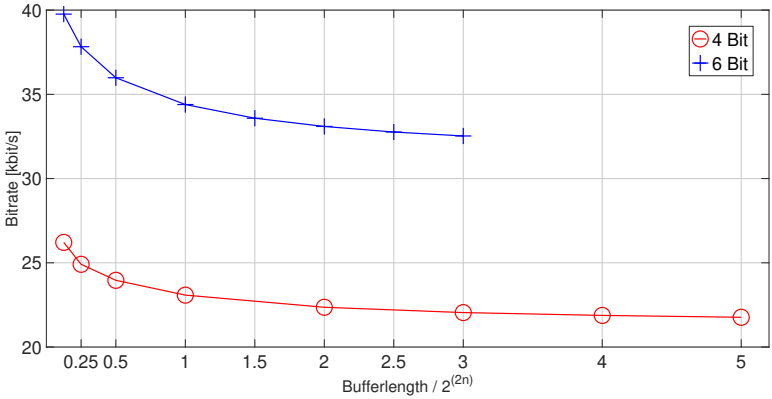


Abbildung 4.6: Erzielte mittlere Bitrate in kbit/s des untersuchten verlustlosen Codecs (ZDLLC) auf dem Trainingsdatensatz in Abhängigkeit von der Bufferlänge für Dynamikumfänge von 4 Bit und 6 Bit. Die Länge wurde zur gemeinsamen Darstellung auf 2^{2n} mit dem Dynamikumfang n normiert.

Zur Bestimmung einer sinnvollen Bufferlänge wurde in jedem Subband der Kontext [0,-1] genutzt, d.h., es ist unter Verwendung der bedingten Wahrscheinlichkeiten gegeben den zeitlich vorher gelegenen Wert in demselben Subband komprimiert worden. Hierbei wurde ohne Verwendung eines künstlichen neuronalen Netzes komprimiert, da nur die Güte der Schätzung der jeweiligen Wahrscheinlichkeiten untersucht werden sollte. Die erzielte Bitrate in Abhängigkeit der Bufferlänge ist in Abb. 4.6 für 4 Bit/Symbol und 6 Bit/Symbol dargestellt. Hierbei ist auf der Abszisse die Bufferlänge in der Größenordnung 2^{2n} dargestellt, wobei n die Bit je Symbol bezeichnet. Es zeigt sich, dass bis zu einer Bufferlänge von etwa 2^{2n}, d.h. 256 im Falle von 4 Bit je Symbol und 4096 im Falle von 6 Bit je Symbol, die Bitrate stark abnimmt und danach nur noch leicht sinkt. Auf Basis dieser Ergebnisse wurde für eine Auflösung von 4 Bit je Symbol eine Bufferlänge von $3 \cdot 2^{2 \cdot 4} = 768$ und für eine Auflösung von 6 Bit je Symbol eine Bufferlänge von $2 \cdot 2^{2 \cdot 6} = 8192$ gewählt.

Der Einfluss der jeweiligen Kontexte wurde untersucht, indem der Testdatensatz mit einer unterschiedlichen Zahl an Kontexten komprimiert und die resultierende Bitrate gemessen wurde. Das Ergebnis ist in Abb. 4.7 dargestellt, erneut für einen Dynamikumfang von 4 Bit/Symbol und 6 Bit/Symbol. Mit *Baseline* wurde der Fall eines einzelnen Kontextes bezeichnet, bei dem wie zuvor kein künstliches neuronales Netz für eine Kontextmischung genutzt wurde. Durch Vergleich mit den Bitraten anderer Fälle lässt sich hieran der Vorteil des Kontextmischens erkennen.

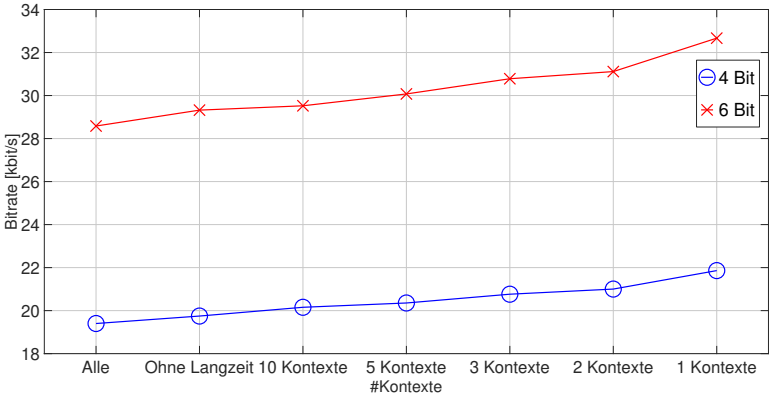


Abbildung 4.7: Bitrate des verlustlosen Kompressionsalgorithmus in Abhängigkeit von der Zahl der genutzten Kontexte, welche in Tabelle 3.1 aufgelistet sind. Als Baseline diente hierbei der je Subband einzelne Kontext [0,-1] ohne Kontextmischung.

Die vollständige Zahl an Kontexten bezieht sich auf die in der Tabelle 3.1 zusammengefassten Kontexte, welche insgesamt 38 umfasst. Im Vergleich zur Baseline lässt sich durch Kontextmischung die Bitrate von etwa 21,9 kbit/s bei einem Dynamikumfang von 4 Bit/Symbol auf etwa 19,5 kbit/s reduzieren. Dies entspricht einer Reduktion der Bitrate um etwa 11 %. Die Reduktion fällt für einen Dynamikumfang von 6 Bit/Symbol noch deutlicher aus, hier wurde eine Reduktion von etwa 32,7 kbit/s auf 28,4 kbit/s erzielt. Dies entspricht einer Reduktion der Bitrate um etwa 13, 2 %.

4.3.1 *Vergleich mit alternativen Kompressionsverfahren*

Der ZDLLC wurde sowohl mit verlustlosen als auch mit verlustbehafteten Kompressionsalgorithmen verglichen. Als verlustlose Vergleichscodecs wurden zum einen PAQ [Mah05] in der PAQ8N Version [Mah], das auf Grund seiner immensen und sehr vielseitigen Kompressionsfähigkeit oft als Referenz im Bereich der verlustlosen Kompression dient, und zum anderen Prediction by Partial Matching (PPM) [Say96], welches vor dem Aufkommen von PAQ im Bereich der verlustlosen Kompressionsalgorithmen Stand der Technik war, herangezogen. Für PPM wurde die Implementierung aus [Miu] genutzt. Beide Algorithmen, die nicht für das drahtlose Streamen von Daten konzipiert wurden, weisen eine algorithmische Latenz auf, die der Länge der zu codierenden Datei bzw. des

Tabelle 4.3: Mittlere Bitraten in kbit/s des untersuchten verlustlosen Codec (ZDLLC) und der Referenzalgorithmen PAQ sowie PPM, dem Opus Codec mit einer algorithmischen Latenz von 5 ms und 7,5 ms, sowie des Electrocodecs mit 2 Bit und 3 Bit Quantisiererauflösung je Subband (entspricht der **EC2** und **EC3** Testbedingung). Die Ergebnisse wurden über den gesamten Testdatensatz mit einer Symbolauflösung von 6 Bit erzielt. Separat sind die Ergebnisse auch für die Teilmenge des Testdatensatzes mit einem Signal-Rausch-Verhältnis kleiner gleich 5 dB tabuliert.

Datensatz\Codec	ZDLLC	PAQ	PPM	$\text{Opus}_{5\text{ms}}$	$\text{Opus}_{7,5\text{ms}}$	EC2	EC3
Testdatensatz	28,6	25,1	37,3	35,2	33,6	20,1	24,3
Testdatensatz ($\leqslant$5 dB)	32,6	30,4	44,3	35,2	33,6	22,7	27,8

Datenstroms entspricht. Daher ist der Vergleich nicht einwandfrei und dient lediglich der Einordnung der Kompressionsleistung des ZDLLC.
Neben den genannten verlustlosen Vergleichsverfahren wurden zwei verlustbehaftete Codecs zum Vergleich herangezogen. Zum einen der Opus-Audiocodec in zwei verschiedenen Latenzeinstellungen (5 ms und 7,5 ms). Zum anderen der Electrocodec in der 2 Bit und 3 Bit Variante, d.h. die Testbedingungen **EC2** und **EC3** wie in Abschnitt 3.7 eingeführt.

Für die Evaluierung wurde zum einen der gesamte Testdatensatz genutzt als auch die Teilmenge an Testdaten mit einem Signal-Rausch-Verhältnis von kleiner gleich 5 dB. Bei einem niedrigen Signal-Rausch-Verhältnis nimmt der Informationsgehalt der Erregungsmuster wesentlich zu, sodass die Bitrate der Codecs der Erregungsmuster jeweils deutlich ansteigt. Daher wurde die Auswertung im Hinblick auf diese Tatsache aufgegliedert. Bei etwa 5 dB saturiert der Zuwachs der Bitrate und es kommt nur noch zu einer kleinen Steigerung beim ZDLLC.

Tabelle 4.3 zeigt die mittleren Bitraten aller genannten Codecs bei einem Dynamikumfang von 6 Bit/Symbol. Der ZDLLC erzielt eine mittlere Bitrate über den gesamten Testdatensatz von 28,6 kbit/s und von 32,6 kbit/s für ein Signal-Rausch-Verhältnis von kleiner gleich 5 dB. Damit

Tabelle 4.4: Mittlere Dynamikbereiche der einzelnen Kanäle bzw. Subbänder für prä- und postlingual implantierte Cochlea-Implantatträger, entnommen aus [Zar+20]. Der Durchschnitt (AVG) der Dynamikbereiche der Subbänder ist ebenfalls angegeben.

Proband\Subband	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	AVG
Prälingual	56	58	57	58	57	57	57	57	57	58	58	57	55	56	57	59	58	58	58	58	59	59	58
Postlingual	38	38	39	40	39	40	41	40	41	40	40	40	39	38	38	37	36	35	34	32	30	29	37

Tabelle 4.5: Mittlere Bitraten in kbit/s auf dem gesamten TIMIT-Testdatensatz und der Teilmenge mit einem Signal-Rausch-Verhältnis kleiner gleich 5 dB des untersuchten verzögerungs- und verlustfreien Codecs (ZDLLC), sowie Prediction by Partial Matching (PPM) und PAQ unter Verwendung der in Tabelle 4.4 aufgeführten Dynamikbereiche von prälingual und postlingual implantierten CI-Nutzern. Die Datenraten von Opus und dem Electrocodec sind unverändert und daher hier nicht tabuliert.

Datensatz\Testbedingung	**Prälingual**			**Postlingual**		
	ZDLLC	PAQ	PPM	ZDLLC	PAQ	PPM
Testdatensatz	27,4	24,9	36,7	24,6	21,9	33,8
Testdatensatz ($\leqslant$ 5dB)	31,4	30,0	43,5	28,1	26,3	40,0

erzielt der ZDLLC eine deutlich geringere Bitrate als Opus in den untersuchten Latenzeinstellungen. Opus erzielt eine Bitrate von minimal 33,6 kbit/s, wobei diese Bitrate ziemlich invariant ist unter Änderungen des Hintergrundrauschens. Obwohl der ZDLLC verlustlos und Opus verlustbehaftet komprimiert, erzielt der ZDLLC nicht nur eine geringere Latenz, sondern auch eine niedrigere Bitrate als Opus. Ferner erreicht der ZDLLC eine deutlich geringere Bitrate als PPM, welches mit 37,3 kbit/s bzw. 44,3 kbit/s die schlechtesten Ergebnisse erzielt. Jedoch erreicht PAQ eine geringere Bitrate von 25,1 kbit/s auf dem gesamten Datensatz und 30,4 kbit/s für Signal-Rausch-Verhältnisse kleiner gleich 5 dB. Zu bedenken ist jedoch der deutliche Latenzunterschied der Kompressionsverfahren. Im Vergleich dazu erzielt der Electrocodec, der verlustbehaftet komprimiert, die niedrigsten Bitraten von bis zu 20,1 kbit/s für den **EC2** und 24,3 kbit/s für den **EC3**. Dies jedoch verlustbehaftet.

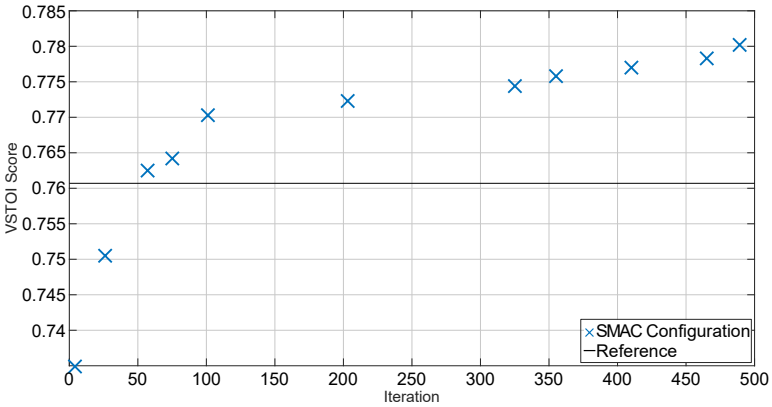
Für die Evaluierung auf realen Daten wurden mittlere Dynamikumfänge je Subband aus [Zar+20] entnommen. Hierbei gibt es wesentliche Unterschiede, je nachdem ob das Cochlea-Implantat vor (prälingual) oder nach (postlingual) Spracherwerb implantiert wurde. Die genutzten Dynamikumfänge je Elektrode sind in Tabelle 4.4 aufgeführt.

Tabelle 4.5 zeigt die erzielten Bitraten der untersuchten Codecs unter Nutzung realer Dynamikumfänge. Für Opus und den Electrocodec gelten dieselben Bitraten wie zuvor. Qualitativ zeigt sich das gleiche Bild wie zuvor in der Rangfolge der Codecs hinsichtlich der erzielten Bitrate. Für Dynamikumfänge von postlingual implantierten Cochlea-Implantatnutzern erzielt der ZDLLC eine mittlere Bitrate von 24,6 kbit/s auf dem Testdatensatz im Vergleich zu 24,3 kbit/s für den **EC3**. Das heißt, obwohl der **EC3** verlustbehaftet codiert, erzielt der verlustlose ZDLLC nahezu dieselbe Bitrate. In jedem Fall zeigt Opus hinsichtlich Latenz und Bitrate eine schlechtere Leistung als der ZDLLC mit einer

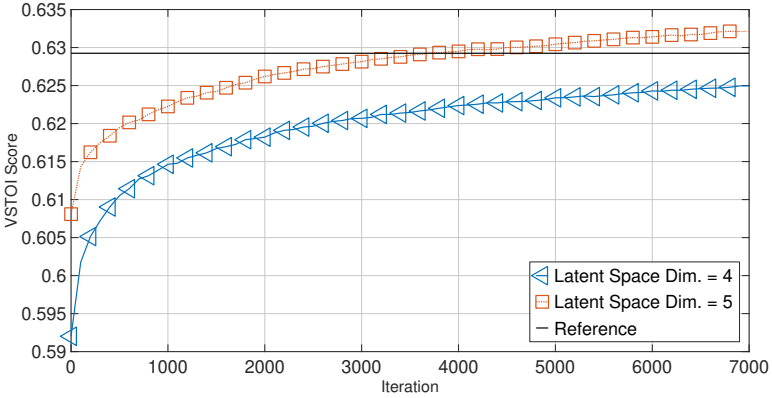
maximalen Bitrate von 31,4 kbit/s für den Testdatensatz bei einem Signal-Rausch-Verhältnis kleiner gleich 5 dB und unter Nutzung der prälingualen Dynamikumfänge. Das Verhalten von PAQ und PPM ist ähnlich wie zuvor. Die Bitrate von PAQ nimmt, ähnlich wie für den vorherigen Fall, mit abnehmendem Signal-Rausch-Verhältnis deutlich stärker zu als für den ZDLCC. Der Bitratenunterschied von PAQ und ZDLLC über den gesamten Testdatensatz liegt bei etwa 2,5 kbit/s, schrumpft aber auf 1,4 kbit/s bei niedrigem Signal-Rausch-Verhältnis. Noch deutlicher wird dies bei PPM. Insgesamt zeigt sich eine sehr gute Leistungsfähigkeit, die Opus aussticht und sogar in die Nähe von PAQ kommt, wobei PAQ eine wesentlich höhere algorithmische Latenz aufweist.

4.4 EVALUIERUNG DES AUTOENCODERS OHNE RÜCKKOPPLUNG

In diesem Abschnitt werden die Ergebnisse der Kompression der Erregungsmuster mittels Autoencoder ohne Rückkopplung vorgestellt [HOO22]. Zunächst wird die Hyperparameteroptimierung beschrieben, aus welcher optimale Autoencoderstrukturen sowie optimierte Parameter des SPSA-Algorithmus hervorgingen. Anschließend wird die eigentliche Kompressionsleistung auf dem TIMIT-Datensatz beschrieben. Der Verlauf der Hyperparameteroptimierung des Autoencoders sowie des SPSA-Algorithmus mit SMAC ist in Abb. 4.8a dargestellt. Hierbei wurde eine Latentdimension, was abkürzend die Dimension des verborgenen Raumes bezeichnet, von fünf gewählt. Die durchgezogene Linie ist die Referenz, d.h. der VSTOI-Wert des uncodierten Signals. Die Kreuze markieren den VSTOI-Wert des mit der gefundenen, optimierten Konfiguration trainierten Autoencoders respektive der damit codierten Erregungsmuster. Bereits nach 54 Iterationen wurde die Referenz übertroffen und es konnte sogar eine weitere Verbesserung, mutmaßlich durch Rauschunterdrückung, gefunden werden. Die final gefundene Konfiguration für eine Latentdimension von fünf ist im Anhang in Tabelle 8.2 zusammengefasst. „Log“ bezeichnet in dieser logarithmisches Stichprobenziehen, was im Falle von Größenordnungsunterschieden zwischen den Parametergrenzen Anwendung findet. Überraschend ist die Größe der ersten Schicht, welche größer ist als die Eingangsschicht von 22 Neuronen, und sich immer wieder reproduzieren ließ. Nach der Hyperparameteroptimierung wurde die gefundene Konfiguration verwendet, um den Autoencoder auf dem Trainingsdatensatz zu traininieren. Hierbei wurde zunächst gradientenbasiert über 500 Epochen unter Nutzung der Verlustfunktion gemäß Gl. 3.6 und dem Parameter α gemäß Hyperparameteroptimierung trainiert. Anschließend wurde der Autoencoder bezüglich STOI mittels des SPSA-Algorithmus trainiert. Dieser Trainingsverlauf ist in Abb. 4.8b



(a) Verlauf der Hyperparameteroptimierung



(b) Trainingsverlauf

Abbildung 4.8: (a) Verlauf der Hyperparameteroptimierung mit SMAC für eine Latentdimension von fünf mit zwei Layern. Tabelle 8.2 bietet eine Übersicht über die Hyperparameter und die finale Konfiguration. Die Referenz von 0,761 wurde nach 54 Iterationen übertroffen. (b) Verlauf der durchschnittlichen VSTOI-Werte auf dem Trainingsdatensatz bei der Optimierung der Gewichte der Autoencoder mit optimierten Konfigurationen für eine Latent Space Dimension von vier und fünf über 7000 Iterationen.

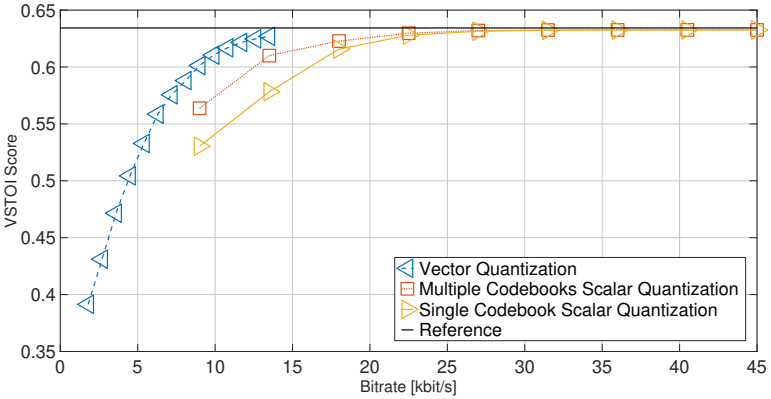


Abbildung 4.9: Mittlere VSTOI-Werte auf dem TIMIT-Testdatensatz in Abhängigkeit von der gewählten Bitrate für den Autoencoder einmal unter Verwendung von Vektorquantisierung und einmal unter Verwendung von skalarer Quantisierung. Bei der skalaren Quantisierung wurde, je Dimension des verborgenen Raums, zwischen separaten Codebüchern (Multiple) und identischen Codebüchern (Single) unterschieden.

dargestellt, wobei ebenfalls ein Autoencoder mit einer Latentdimension von vier gezeigt ist, dessen Struktur identisch mit SMAC optimiert wurde. Die Referenz, der Mittelwert der VSTOI-Werte der uncodierten Erregungsmuster, wurde nach etwa 3500 Iterationen vom Autoencoder mit einer Latentdimension von fünf erreicht. Die Referenz wurde auch nach 7000 Iterationen vom Autoencoder mit Latentdimension von vier knapp verfehlt.

Daher wurde die nachfolgende Auswertung mit einem Autoencoder der Latentdimension von fünf durchgeführt.

Die Datenrate des Autoencoders ohne Quantisierung der Latentvariablen beträgt bei einer Pulsrate von 900 PPS, einer Latentdimension von fünf und unter Verwendung von 32 Bit Fließkommazahlen (einfache Genauigkeit) 144 kbit/s, was für geläufige Audiocodecs ziemlich hoch ist. Daher wurden die Latentvariablen mittels skalarer und Vektorquantisierung quantisiert. Beide Quantisierer wurden mit Hilfe des Lloyd-Max-Algorithmus trainiert. Das Training der Quantisierer wurde mittels der Werte der Latentvariablen durchgeführt, welche bei Codierung des Trainingsdatensatzes in der verborgenen Schicht des Autoencoders auftraten. Um einen Eindruck der Verschiedenheit der Verteilungen der Latentvariablen zu erhalten, wurde die skalare Quantisierung einmal mit einem Codebuch für alle Latentvariablen durchgeführt (Single Co-

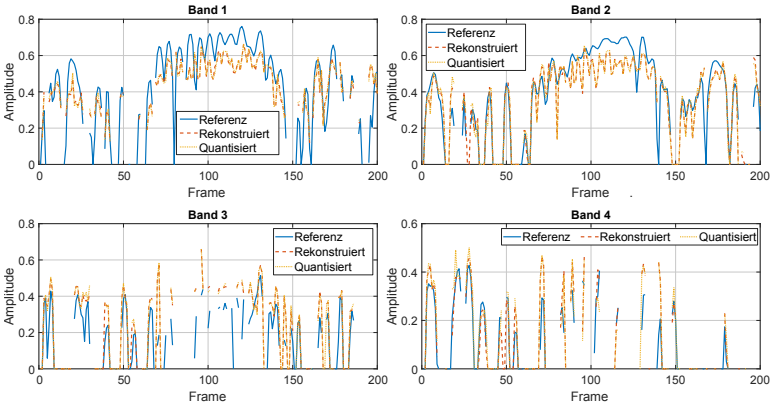


Abbildung 4.10: Beispielerregungsmuster der Bänder 1 bis 4 vor der Kompression mit dem Autoencoder (Referenz), nach der Kompression aber ohne Quantisierung des Latent Spaces (Rekonstruiert) und nach der Kompression inklusive Quantisierung des Latent Spaces (Quantisiert). In vielen Fällen ist der Einfluss der Quantisierung auf die Rekonstruktion vernachlässigbar, weswegen die Kurven oftmals kaum auseinander zu halten sind.

debook Scalar Quantization), und einmal mit einem Codebuch für jede Variable einzeln (Multiple Codebook Scalar Quantization). Der Qualitätsunterschied zwischen skalarer und Vektorquantisierung wiederum ist ein Proxy für die statistischen Abhängigkeiten zwischen den Latentvariablen. Die skalare Quantisierung wurde mit 2 Bit bis 10 Bit je Latentvariable durchgeführt, die Vektorquantisierung mit 0,25 Bit bis 3 Bit je Latentvariable. Die erzielten VSTOI-Werte auf dem Testdatensatz, wenn die Latentvariablen quantisiert wurden, ist in Abb. 4.9 dargestellt. Die Vektorquantisierung erzielte bei deutlich geringerer Bitrate näherungsweise den Referenz VSTOI-Wert, wohingegen die skalare Quantisierung eine deutlich höhere Bitrate benötigte. Ferner war der Unterschied zwischen Single und Multiple Codebook Scalar Quantization bei niedrigen Bitraten sehr deutlich ausgeprägt. Abb. 4.10 zeigt für vier Bänder exemplarische Stimulationsmuster vor der Kompression, nach der Kompression mittels des Autoencoders ohne Quantisierung und schlussendlich nach der Kompression mittels des Autoencoders inklusive Quantisierung des verborgenen Raums. Es ist zu erkennen, dass in den meisten Fällen die Quantisierung kaum einen Effekt auf die rekonstruierten Stimulationsmuster hat. Erkennbar ist ferner, dass bei den längeren Sequenzen eher der qualitative Verlauf rekonstruiert wird mit quantitativ größeren Abweichungen. Dies dürfte eine wesentliche Reduktion der Datenrate erlaubt haben, führt

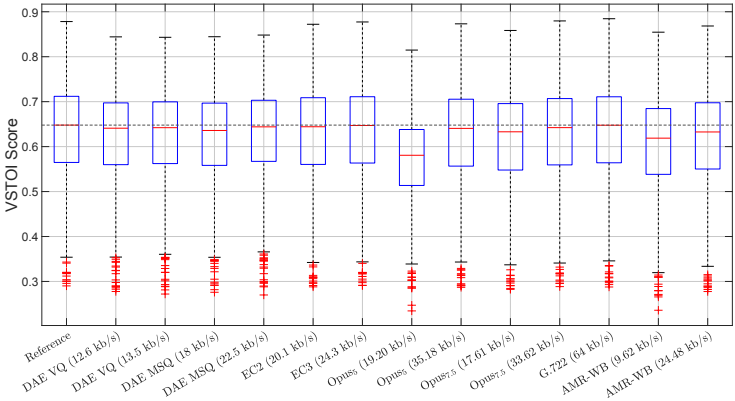


Abbildung 4.11: VSTOI-Werte des Autoencoders (DAE) sowie der Referenzsignale nebst einiger weiterer Codecs erzielt auf dem TIMIT-Testdatensatz für alle Signal-Rausch-Verhältnisse. Für den Opus-Audiocodec wurde die gewählte algorithmische Latenz als Index in Millisekunden notiert.

aber gemäß [Qaz+13] nicht oder nur marginal zu einer Beeinträchtigung des Sprachverstehens und ist daher eine sinnvolle Kompressionstechnik.

Der Autoencoder wurde dann mittels VSTOI mit weiteren Audiocodecs verglichen. Die Datengenerierung erfolgte grundsätzlich genauso wie bei dem Hörtest; Abb. 8.2 im Anhang zeigt den entsprechenden Signalfluss inklusive der Berechnung der VSTOI-Werte. Die erzielten VSTOI-Werte aller Audiocodecs für den gesamten TIMIT-Testdatensatz sind in Abb. 4.11 gezeigt.

Der Autoencoder erzielte mittlere Referenz-VSTOI-Werte bei einer Bitrate von 13,5 kbit/s, während Opus etwa 35,2 kbit/s für näherungsweise dieselben VSTOI-Werte bei einer algorithmischen Verzögerung von 5 ms benötigte. Der G.722 Audiocodec erzielte die Referenz ebenso, verwendete dafür jedoch 64 kbit/s und wurde nur zur Plausibilitätsprüfung ergänzt, da er der einzige Audiocodec ist, von dem bekannt ist, dass er tatsächlich im Kontext von Cochlea-Implantaten Anwendung findet. Der AMR-WB erzielte auch bei 24,5 kbit/s und einer algorithmischen Latenz von 20 ms nicht die Referenz. Ursache könnte hier die Implementierung von FFMPEG oder eine Limitierung von VSTOI sein. Der AMR-WB ist ein sehr bekannter Sprachcodec, und es wurde erwartet, dass, auch mit Rauschen, bei der höchsten Bitrate die Referenz erreicht werden würde. Dieses Ergebnis ist daher zweifelhaft. Die durch den AMR-WB hervorgerufenen Verzerrungen in den Audiosignalen schienen jedenfalls kaum

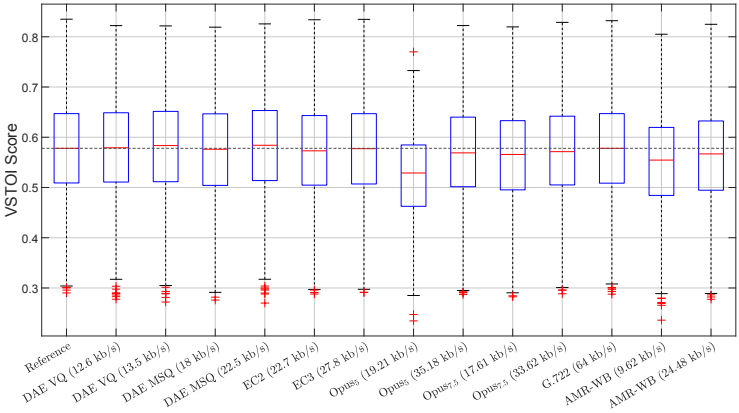


Abbildung 4.12: VSTOI-Werte des Autoencoders (DAE) sowie der Referenzsignale nebst einiger weiterer Codecs erzielt auf dem TIMIT-Testdatensatz für Signal-Rausch-Verhältnisse von 5 dB und weniger. Für den Opus-Codec wurde die gewählte algorithmische Latenz als Index in Millisekunden notiert.

wahrnehmbar zu sein. Der Abbildung entnimmt man eine scheinbare Verbesserung des VSTOI-Werts in Folge der Codierung mit dem Autoencoder für Signale mit niedrigem VSTOI-Wert. Ebenfalls entnimmt man eine scheinbare Verschlechterung der VSTOI-Werte in Folge der Codierung mit dem Autoencoder für Signale mit hohen VSTOI-Werten, welche tendenziell mit hohem Signal-Rausch-Verhältnis korrespondieren. Auf diese Verschlechterung wird in Abschnitt 4.4.1 genauer eingegangen. Niedrige VSTOI-Werte (vor Codierung) korrespondieren im Allgemeinen mit niedrigem Signal-Rausch-Verhältnis. Daher wurde zusätzlich die Subgruppe der VSTOI-Werte in Abb. 4.12 abgebildet, die zu Signalen mit einem Signal-Rausch-Verhältnis kleiner gleich 5 dB gehört. In der Tat erzielen die vom Autoencoder mit einer Bitrate von 13,5 kbit/s codierten Signale VSTOI-Werte, welche im Median um 0,005 über denen der Referenz liegen. Grund ist mit hoher Wahrscheinlichkeit eine erlernte Rauschunterdrückung. Probehören einiger vocodeter Signale bestätigte diesen Eindruck. In Tabelle 4.6 sind die Median VSTOI-Werte der Audiocodecs zusammen mit der Referenz über den gesamten Testdatensatz sowie die Teilmenge mit einem Signal-Rausch-Verhältnis kleiner gleich 5 dB zusammengefasst. Der Testdatensatz wurde gezielt so erstellt, dass er Einstellungen, d.h. nicht nur Sprecher, enthält, die im Trainingsdatensatz nicht enthalten sind. Dies gestattet die Generalisierung besser zu überprüfen. Es wurde daher eine Subgruppenanalyse

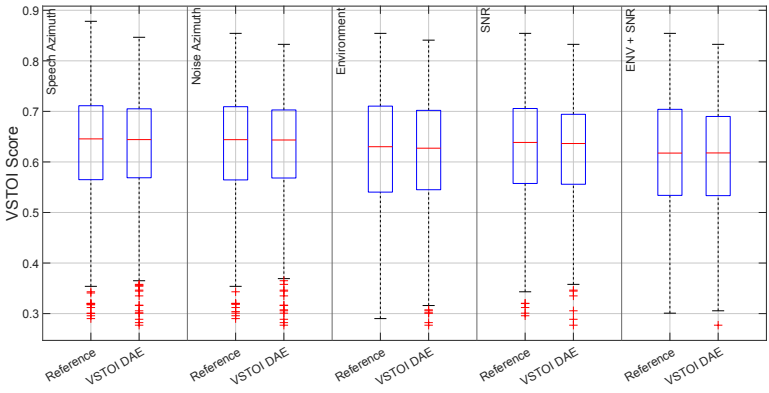


Abbildung 4.13: VSTOI-Werte des Autoencoders für Erregungsmuster mit „out-of-group“ -Einstellungen, d.h. für akustische Szenarien, die nicht im Trainingsdatensatz enthalten waren. Etwa wurden in der Kategorie „Speech Azimuth“ nur jene Erregungsmuster des Testdatensatzes berücksichtigt, welche Inzidenzwinkel des Sprachsignals nutzen, die nicht im Trainingsdatensatz verwendet wurden (siehe diesbezüglich auch Tabelle 3.2). Es ist keine Reduktion der Leistungsfähigkeit in Abhängigkeit von im Training nicht gesehenen akustischen Szenarien zu erkennen.

durchgeführt, welche die Leistung des Autoencoders nur für die Beispiele des Testdatensatzes auswertet und Einstellungen verwendet, die im Trainingsdatensatz nicht vorkommen, wie etwa ein Signal-Rausch-Verhältnis von 2,5 dB oder eine Cafeteria als akustisches Szenario. Das Ergebnis dieser Subgruppenanalyse ist in Abb. 4.13 dargestellt. Gezeigt sind separate Auswertungen für den Sprach- und Rauschazimut, die akustische Umgebung, sowie das Signal-Rausch-Verhältnis und die Kombination von Signal-Rausch-Verhältnis und akustischem Szenario. In keinem Fall

Tabelle 4.6: Median VSTOI-Werte des vektor- (VQ) und skalarquantisierten (SQ) Autoencoders (DAE) und anderer untersuchter Audiocodecs sowie der Referenzbedingung (Ref) über den gesamten Testdatensatz sowie über dessen Teilmenge mit einem Signal-Rausch-Verhältnis von kleiner gleich 5 dB. Die Werte in Klammern sind die jeweiligen Bitraten über den gesamten Testdatensatz in kbit/s.

Datensatz\Testbedingung	Ref	VQ-DAE (12,6)	VQ-DAE (13,5)	DAE MSQ (18)	DAE MSQ (22,5)	EC2 (20,1)	EC3 (24,3)
Testdatensatz	0,648	0,641	0,642	0,636	0,644	0,644	0,647
Testdatensatz ($\leqslant$ 5dB)	0,578	0,579	0,583	0,576	0,584	0,573	0,577
Datensatz\Testbedingung	$\text{Opus}_{5\,\text{ms}}$ (19,2)	$\text{Opus}_{5\,\text{ms}}$ (35,2)	$\text{Opus}_{7,5\,\text{ms}}$ (17,6)	$\text{Opus}_{7,5\,\text{ms}}$ (33,6)	G.722 (64)	AMR-WB (9,6)	AMR-WB (24,5)
Testdatensatz	0,581	0,641	0,633	0,642	0,648	0,619	0,633
Testdatensatz ($\leqslant$ 5dB)	0,529	0,569	0,566	0,571	0,578	0,555	0,567

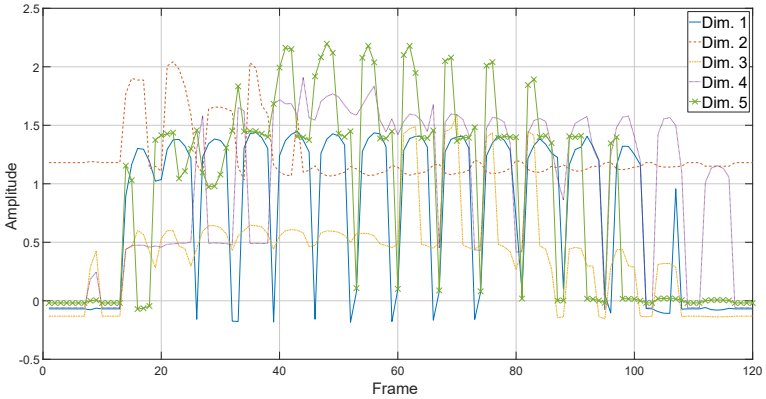


Abbildung 4.14: Beispielauszug aus dem verborgenen Raum des besten Autoencoders. Es ist klar eine zeitliche Abhängigkeit zu erkennen, weswegen die gefundene Autoencoderstruktur zwar sehr gut funktioniert, jedoch durch etwa prädiktive Codierungen noch verbessert werden könnte.

zeigte sich eine wesentliche Reduktion der Codierungsleistung. Die Kompressionsleistung generalisiert also sehr gut auf ungesehene Szenarien. Dies kann umgekehrt jedoch auch bedeuten, dass der Einfluss des Signal-Rausch-Verhältnisses, der Sprecher oder der akustischen Umgebung auf die statistischen Abhängigkeiten der Erregungsmuster nur marginal ist. Weitere Tests wurden durchgeführt, um die Reproduzierbarkeit der Hyperparameteroptimierung und eine mögliche Steigerungsfähigkeit der Codierungsleistung zu untersuchen, welche sich durch die Erhöhung der Schichtanzahl des Autoencoders ergeben könnte. Es wurde festgestellt, dass bei Erhöhung der Schichtanzahl von vier auf sechs auch bei einer Latentdimension von vier die Referenz erreicht werden kann, jedoch wurde keine Reduktion der Bitrate erzielt, da die Zahl an Bits je Latentvariable der Quantisierer erhöht werden musste, um die Referenz-VSTOI-Werte zu erzielen.

Es wurde der verborgene Raum/Latent Space des Autoencoders auf statistische Abhängigkeiten untersucht, um weitere Verbesserungsmöglichkeiten zu eruieren. Abb. 4.14 zeigt einen exemplarischen Ausschnitt einer Sequenz des Latent Space. Zum einen ist zu erkennen, dass der zeitliche Verlauf im Latent Space dem zeitlichen Verlauf der Erregungsmuster stark ähnelt. Als Vergleich möge Abb. 3.13 dienen. Zum anderen, damit zusammenhängend, ist eine offensichtliche, sehr wesentliche zeitliche Abhängigkeit zu erkennen. Daher sollte eine wesentliche Verbesserung des

Tabelle 4.7: Ergebnisse der Hyperparameteroptimierung des Rückkopplungsautoencoders (FRAE) mit Latentdimension von fünf. $\#\omega$ ist die Anzahl der Modellparameter, A und c sind SPSA Parameter, lr ist die Lernrate, r die Anzahl der Rückkopplungsschritte, α ist ein Parameter der Lossfunktion 3.6. VSTOI ist der finale VSTOI-Wert der FRAE-codierten Erregungsmuster, ΔVSTOI ist der Unterschied dieser VSTOI-Werte zum Referenzwert $\text{ref} = 0.79285$. Größere Werte sind besser.

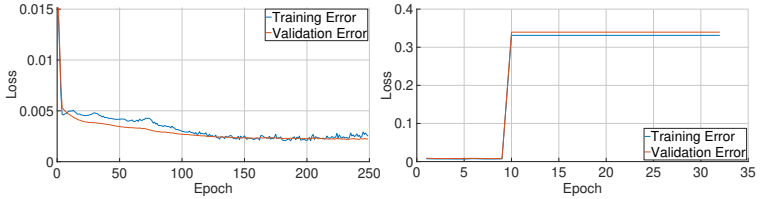
Bezeichner	$\#\omega$	A	c	lr	Neuronen pro Schicht			r	α	VSTOI	ΔVSTOI
FRAE-L5-H2-R1	2815	12037	0,009176	0,009837	17	24		1	0,484553	0,813458	0,02024
FRAE-L5-H2-R2	3655	14205	0,009725	0,00486	30	8		2	0,409898	0,812673	0,01946
FRAE-L5-H2-R3	4239	13713	0,011009	0,005749	27	9		3	0,297337	0,810086	0,01687
FRAE-L5-H2-R4	7239	37406	0,009783	0,00114	24	27		4	0,441055	0,799304	0,00609
FRAE-L5-H3-R1	4659	57498	0,004653	0,000959	21	21	30	1	0,293783	0,808049	0,01483
FRAE-L5-H3-R2	3497	48130	0,009887	0,001247	15	30	10	2	0,431882	0,801317	0,00810
FRAE-L5-H3-R3	5655	30233	0,003634	0,001421	29	20	10	3	0,278823	0,792756	-0,00045
FRAE-L5-H3-R4	3353	10273	0,020765	0,00479	9	28	10	4	0,280732	0,785826	-0,00738

Autoencoders möglich sein, z.B. mittels prädiktiver Codierungsansätze wie einer Differential Puls-Code Modulation (DPCM) im Latent Space diese Redundanz zu reduzieren.

Dies wurde auf verschiedene Weise untersucht und getestet, jedoch ohne eine Verbesserung der Kompressionsleistung zu erzielen. Die Ursachen des Ausbleibens einer Verbesserung sind nicht klar ersichtlich, das Gesamtsystem wird recht instabil sobald eine DPCM im Latent Space ergänzt wird. Daher gestaltete sich das Training entsprechend schwierig. Dies motivierte die Suche nach Alternativansätzen, die letztendlich im Rückkopplungsautoencoder mündete [HBO23].

4.4.1 *Evaluierung des Rückkopplungsautoencoders*

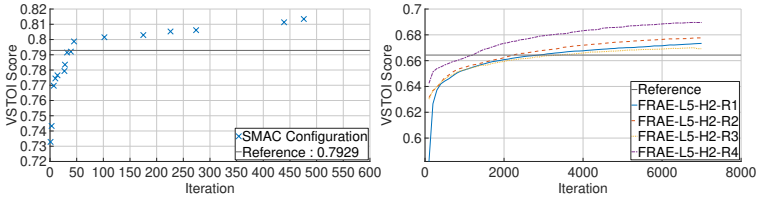
Die Hyperparameteroptimierung des Rückkopplungsautoencoders wurde mit einer Latentdimension von drei bis fünf, zwei und drei versteckten Schichten (jeweils von Encoder und Decoder) sowie ein bis vier Rückkopplungsschritten durchgeführt. Alle möglichen Kombinationen wurden hierbei berücksichtigt. Parallel wurde, da der bisherige Code von dem Python Framework Keras zu dem Python Framework Pytorch portiert wurde, eine erneute Hyperparameteroptimierung mit dem rückkopplungsfreien Autoencoder vorgenommen. Für den rückkopplungsfreien Autoencoder wurden ebenso Latentdimensionen von drei bis fünf als auch zwei und drei versteckte Schichten (jeweils von Encoder und Decoder) untersucht. Die Ergebnisse der Hyperparameteroptimierung des Rückkopplungsautoencoders, zusammen mit der Zahl der Gewichte des jeweiligen Netzes, für eine Latentdimension von fünf sind in Tabelle 4.7 angegeben. Beispielkurven, die den Verlauf der Verlustfunktion für



(a) Unauffälliger Trainingsverlauf (b) Pathologischer Trainingsverlauf

Abbildung 4.15: Trainingsverlauf eines Rückkopplungsautoencoders (a) eines unaufälligen Trainings und (b) eines pathologischen Trainings. Die Kompressionsleistung reduziert sich spontan bei Epoche 10.

das Gradiententraining mit der Verlustfunktion 3.6 für ein unauffälliges sowie ein „pathologisches“ Beispiel (siehe Abschnitt 3.4) wiedergeben, sind in Abb. 4.15 dargestellt. Das exemplarische pathologische Beispiel verschlechtert sich in der zehnten Iteration spontan massiv. Die Ursache dürfte ein instabiles Rekonstruktionsfilter des Decoders des Rückkopplungsautoencoders sein, wodurch die nachfolgenden Gradienten explodieren. Ein beispielhafter Verlauf der VSTOI-Werte über die Iterationen der Hyperparameteroptimierung ist in Abb. 4.16a dargestellt. Wie beim Autoencoder kommt es nach etwa 50 Epochen zum Übertreffen der Referenz und nachfolgend zu einer Verbesserung ebendieser. Abb. 4.16b zeigt den Trainingsverlauf der Optimierung mit dem SPSA-Algorithmus einiger Modelle im Anschluss an die Hyperparameteroptimierung. Hierbei wurde keine Quantisierung im Latent Space vorgenommen. Auf der Abszisse werden die mittleren VSTOI-Werte auf dem Trainingsdatensatz aufgetragen. In der Abbildung sind das gleiche Modell mit ein bis vier Rückkopplungsschritten abgebildet. Auffallend ist, dass der FRAE-L5-H2-R1, d.h. der Rückkopplungsautoencoder mit einer Latentdimension von fünf, zwei versteckten Schichten des Encoders und des Decoders sowie einem Rückkopplungsschritt, deutlich schlechter in das Training startet, jedoch diesen Nachteil sehr schnell wieder ausgleichen kann. Nach 7000 Iterationen zeigt sich im Wesentlichen das erwartete Bild, d.h. der mittlere VSTOI-Wert wächst mit zunehmenden Rückkopplungsschritten. Die einzige Ausnahme bildet der FRAE-L5-H2-R3, welcher schlechter als das Modell mit einem Rückkopplungsschritt abschneidet. Mutmaßlich hat hier die Hyperparameteroptimierung, die im Kern zufallsbasiert ist, etwas schlechtere Hyperparameter gefunden. Eine weitere Ursache könnte ein zufällig schlechter Startwert sein, jedoch war der FRAE-L5-H2-R3 praktisch gleichauf mit dem FRAE-L5-H2-R2 und der FRAE-L5-H2-R1 zeigt, dass schlechte Startwerte nicht automatisch



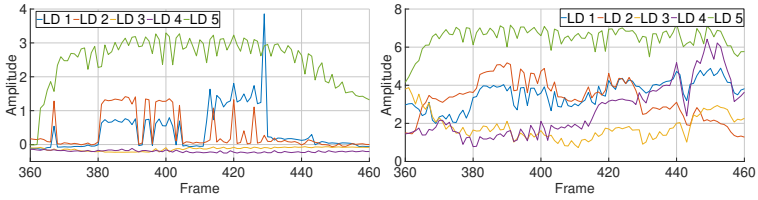
(a) Verlauf einer Hyperparameteroptimierung (b) Trainingsverlauf

Abbildung 4.16: (a) Verlauf der Hyperparameteroptimierung eines Rückkopplungsautoencoders (FRAE) mit SMAC für eine Latentdimension von fünf und zwei verborgene Schichten. Nach etwa 50 Iterationen wurde von einer gefundenen Konfiguration der Referenz VSTOI-Wert übertroffen. (b) Trainingsverlauf des FRAE über die Iterationen des SPSA-Algorithmus. Gestrichelt ist die Referenz über den Trainingsdatensatz eingezeichnet.

schlechte Endleistungen bedeuten. Tabelle 4.8 zeigt die mittleren VSTOI-Werte der Rückkopplungsautoencoder auf dem Trainingsdatensatz mit einer Latentdimension von fünf sowie Ergebnisse für den rückkopplungsfreien Autoencoder, zur Illustration bis hinab zu einer Latentdimension von drei. Diese Ergebnisse wurden ohne Quantisierung im Anschluss an die Hyperparameteroptimierung sowie das erste Training mit dem SPSA-Algorithmus über 7000 Iterationen erzielt. Es zeigt sich ein leichter Vorteil des Rückkopplungsautoencoders. An dieser Stelle wurde der verborgene Raum des Rückkopplungsautoencoders mit jenem des Autoencoders verglichen, um zu untersuchen, ob in der Tat eine prädiktive Codierung erlernt wurde. Diese sollte sich durch eine reduzierte zeitliche Abhängigkeit der Latentsignale offenbaren. Ein Ausschnitt des verborgenen Raums des Autoencoders ohne Rückkopplung (AEC) und des Rückkopplungsautoencoders (FRAE) ist in Abb. 4.17 dargestellt.

Jeweils wurde der gleiche Abschnitt eines Erregungsmusters codiert. Eine Messung der Autokorrelation bestätigte den visuellen Eindruck, dass die Autokorrelation der Latentsignale keineswegs vom FRAE reduziert wurde. An dieser Stelle wurde vermutet, dass die Datenrate ohne Quantisierung mit 144 kbit/s, wie zuvor berechnet, schlicht noch zu hoch ist, sodass eine prädiktive Codierung keinen wesentlichen Vorteil bietet. Daher wurde die Evaluierung des FRAE regulär fortgesetzt und die Kompressionsleistung in Abhängigkeit der Bitrate untersucht, wobei die Bitrate durch die Größe des Quantisierercodebuchs wie zuvor gesteuert wurde.

Das Ergebnis auf dem Testdatensatz ist für die drei besten FRAE-Modelle, sowie die besten zwei AEC in Abb. 4.18 dargestellt. Hierbei



(a) AEC-L5-H2 (b) FRAE-L5-H3-R4

Abbildung 4.17: Latentdimensionen (LD) des AEC-L5-H2 sowie des FRAE-L5-H3-R4 für einen Ausschnitt der Datei DR2_FCYLo_SX37, jeweils ohne Quantisierung. Auffällig ist, dass noch deutliche zeitliche Abhängigkeiten im verborgenen Raum des FRAE vorhanden sind. Daraus lässt sich schließen, dass der erlernte Kompressionsalgorithmus noch nicht so gut wie möglich ist.

wurde noch keine Entropiecodierung der Quantisierungsindizes vorgenommen. Die beiden besten FRAE, FRAE-L5-H3-R4 und FRAE-L5-H3-R2 zeigen eine deutlich bessere Kompressionsleistung als der beste AEC, welcher immerhin in guter Näherung mit dem FRAE-L5-H2-R4 gleichauf ist. Der Abbildung entnimmt man des Weiteren einen deutlichen Vorteil von drei Schichten im Vergleich zu zwei beim FRAE. Für den AEC ließ sich das Übertreffen der Referenz bei einer Bitrate von 13,5 kbit/s reproduzieren. Der beste FRAE übertrifft jedoch bei 9 kbit/s die Referenz und ist dem AEC somit deutlich überlegen.

Vergleicht man die Leistungen der FRAE, wie in Abb. 4.18 dargestellt, mit der Leistung vor Quantisierung, wie in Tabelle 4.8 zusammengefasst, so stellt man interessanterweise fest, dass der beste FRAE nach Quantisierung nicht der beste FRAE vor Quantisierung ist. Die Ursache ist nicht klar, offenbar erzeugt der beste FRAE nach Quantisierung, der FRAE-L5-H3-R4, eine deutlich stärkere Kreuzkorrelation im latenten Raum, sodass die Vektorquantisierung effektiver ist. Auf diesen Punkt wird am Ende des Abschnitts zurückgekommen.

Zur Maximierung der Leistungsfähigkeit der Autoencoder, FRAE und AEC, wurden nun die besten Modelle inklusive Vektorquantisierung mittels des SPSA-Algorithmus optimiert. Erst dann ist ein einwandfreier Vergleich von FRAE und AEC möglich. Hierzu wurde erneut für 7000 Iterationen trainiert. Dabei wurden nur ausgewählte Codebuchgrößen berücksichtigt, um die Dauer der Untersuchungen etwas zu reduzieren. Tabelle 4.9 zeigt die VSTOI-Werte auf dem Testdatensatz, inklusive Quantisierung, vor und nach diesem letzten Trainingsschritt. 6- und 7-Bit-Quantisierung wurde für den AEC nicht berücksichtigt, da es als

Tabelle 4.8: Mittlere VSTOI-Werte $\overline{VSTOI}$ der untersuchten Rückkopplungsautoencoder (FRAE) und rückkopplungsfreien Autoencoder (AEC) auf dem Trainingsdatensatz ohne Quantisierung. ΔVSTOI wie in Abschnitt 3.4 definiert. Der mittlere Referenzwert für VSTOI betrug 0,6643.

FRAE	$\overline{VSTOI}$	ΔVSTOI	AEC	$\overline{VSTOI}$	ΔVSTOI
FRAE-L5-H2-R1	0,67329	0,00895	AEC-L3-H2	0,66260	-0,00174
FRAE-L5-H2-R2	0,67762	0,01328	AEC-L3-H3	0,64235	-0,02199
FRAE-L5-H2-R3	0,66925	0,00491	AEC-L4-H2	0,67106	0,00672
FRAE-L5-H2-R4	**0,68955**	**0,02522**	AEC-L4-H3	0,66440	0,00006
FRAE-L5-H3-R1	0,68511	0,02077	AEC-L5-H2	0,67952	0,01518
FRAE-L5-H3-R2	0,68688	0,02254	AEC-L5-H3	**0,68001**	**0,01567**
FRAE-L5-H3-R3	0,67851	0,01417			
FRAE-L5-H3-R4	0,68546	0,02112			

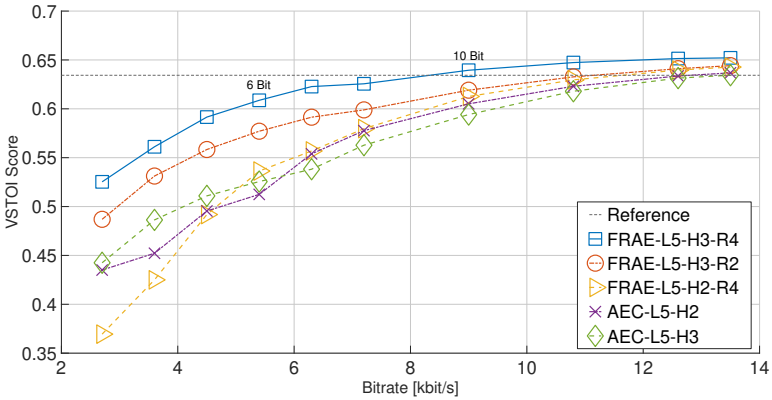


Abbildung 4.18: Mittlere VSTOI-Werte der drei besten Rückkopplungsautoencoder (FRAE) sowie der beiden besten Autoencoder ohne Rückkopplung (AEC) auf dem TIMIT-Testdatensatz in Abhängigkeit von der Bitrate. Es zeigt sich ein klarer Vorteil des FRAE im Vergleich zum AEC. Die gestrichelte Linie zeigt die Referenz, d.h. den mittleren VSTOI-Wert der Erregungsmuster ohne Codierung.

Tabelle 4.9: Mittlere VSTOI-Werte der leistungsstärksten Modelle auf dem Testdatensatz vor und nach der Optimierung der Autoencoder einschließlich der Quantisierer mittels SPSA. VSTOI-Werte, die die mittleren VSTOI-Werte der Referenz übertreffen, sind durch Fettdruck hervorgehoben. Der mittlere VSTOI-Wert der Referenz betrug 0,6343.

Modell\Bit	6		7		8		10	
	Vorher	Nachher	Vorher	Nachher	Vorher	Nachher	Vorher	Nachher
FRAE-L5-H3-R2	0,5770	0,6184	0,5914	0,6257	0,5988	0,6296	0,6190	**0,6423**
FRAE-L5-H2-R4	0,5363	0,5919			0,5797	0,6142	0,6126	0,6296
FRAE-L5-H3-R4	0,6087	0,6328	0,6227	**0,6416**	0,6255	**0,6431**	**0,6395**	**0,6499**
AEC-L5-H2					0,5776	0,6073	0,6051	0,6218
AEC-L5-H3					0,5627	0,5866	0,5941	0,6049

extrem unwahrscheinlich angesehen wurde, was die Ergebnisse bestätigten, dass die AEC die Referenz bei diesen Einstellungen erreichen können. Ebenfalls wurde aus diesem Grund für den FRAE-L5-H2-R4 auf eine Codebuchgröße von 7 Bit verzichtet. Die Tabelle zeigt, dass durch die Optimierung der Gesamtstruktur eine massive Verbesserung der Codierungsleistung erzielt werden kann. Im Extremfall des FRAE-L5-H2-R4 und einer Codebuchgröße von 6 Bit wurde eine Verbesserung des mittleren VSTOI-Werts um mehr als 0,05 erzielt. Dies entspricht, geht man grob nach den Werten aus Tabelle 4.2, einer mutmaßlichen Verbesserung der korrespondierenden mittleren Worterkennungsraten um etwa 10 %, nach [WSS18b] sogar noch mehr. Da der Testdatensatz, vergleiche Tabelle 3.2, tendenziell niedrigere Signal-Rausch-Verhältnisse enthält, wurde hierbei der Korrespondenz nach Tabelle 4.2 bei 0 dB der Vorzug gegeben. Die zugehörigen Bitraten, nach Optimierung der Gesamtstruktur inklusive Vektorquantisierung, sind in Tabelle 4.10 zusammengefasst. Hierbei wurden die Quantisierungsindizes mittels Huffmancodierung komprimiert, sodass eine näherungsweise optimale Kompressionsleistung erzielt wurde. Das Codebuch der Huffmancodierung wurde auf

Tabelle 4.10: Bitrate in kbit/s der besten Modelle auf dem TIMIT-Testdatensatz nach Optimierung der Autoencoder inklusive Quantisierer mittels des SPSA und Huffmancodierung der Quantisierungsindizes.

Modell\Codebuchgröße [Bit]	6	7	8	10
FRAE-L5-H3-R2	4,80 kbit/s	5,56 kbit/s	6,36 kbit/s	7,98 kbit/s
FRAE-L5-H2-R4	4,74 kbit/s		6,33 kbit/s	7,92 kbit/s
FRAE-L5-H3-R4	4,67 kbit/s	5,54 kbit/s	6,51 kbit/s	8,10 kbit/s
AEC-L5-H2			6,16 kbit/s	7,82 kbit/s
AEC-L5-H3			6,44 kbit/s	8,12 kbit/s

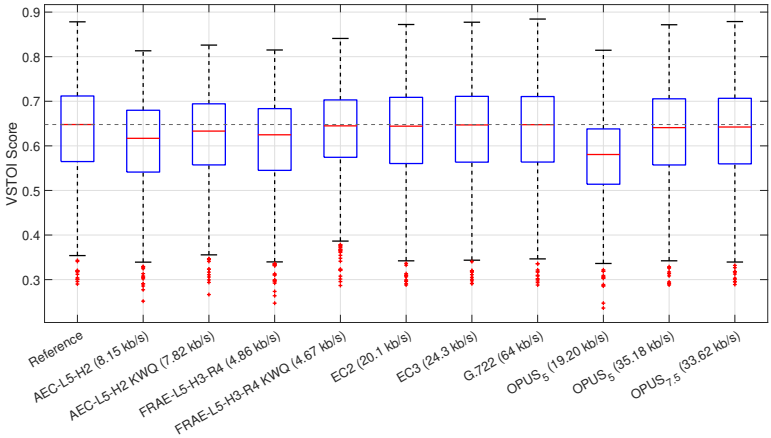


Abbildung 4.19: Vergleich der leistungsstärksten Autoencoder mit dem Electrocodec, Opus und dem G.722-Audiocodec auf dem TIMIT-Testdatensatz. Die Ergebnisse nach Optimierung des gesamten Modells einschließlich des Quantisierers mittels SPSA-Algorithmus werden mit KWQ bezeichnet. Für Opus wird die algorithmische Latenz als Index in Millisekunden angegeben.

Basis des Trainingsdatensatzes erzielt. Die geschätzte Entropie der Quantisierungsindizes ist im Anhang in Tabelle 8.3 angegeben. Es zeigt sich eine geringe Redundanz von unter 0,1 Bit. Abschließend wurde erneut ein Vergleich mit herkömmlichen Audiocodecs durchgeführt. Abb. 4.19 zeigt Boxplots der erzielten VSTOI-Werte des AEC sowie des FRAE, einmal vor und einmal nach Optimierung der Gesamtstruktur mittels des SPSA-Algorithmus. Die Ergebnisse nach Optimierung der Gesamtstruktur sind mit dem Kürzel KWQ versehen. Des Weiteren werden die Autoencoder wie schon in Abschnitt 4.4 mit dem Electrocodec, dem G.722 sowie Opus verglichen. Opus wurde hierbei mit 5 ms sowie 7,5 ms algorithmischer Latenz konfiguriert, was in der Abbildung als Index notiert wurde. Die enorme Verbesserung in Folge der Optimierung von Autoencoder und Quantisierer wird unmittelbar offenbar. Erneut zeigt sich beim AEC, nach Optimierung der Gesamtstruktur, und weniger ausgeprägt auch beim FRAE, eine leichte Verschlechterung bei hohen VSTOI-Werten und eine Verbesserung bei niedrigen VSTOI-Werten. Deswegen wurde erneut die Testdatensatzteilmenge mit einem Signal-Rausch-Verhältnis von kleiner gleich 5 dB in Abb. 4.20 betrachtet. Hier zeigt sich eine deutliche Verbesserung im Median für den FRAE, aber auch den AEC, im Vergleich zur Referenz. Über den gesamten Testdatzsatz erreicht der FRAE im

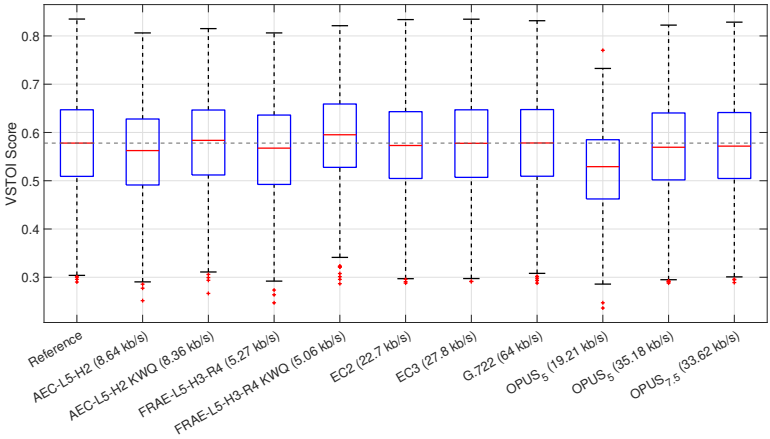
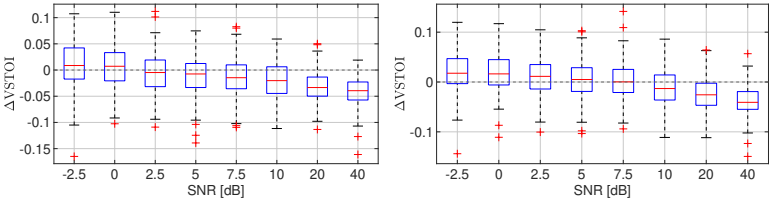


Abbildung 4.20: Vergleich der leistungsstärksten Autoencoder mit verschiedenen Audiocodecs auf der Teilmenge des TIMIT-Testdatensatzes mit einem Signal-Rausch-Verhältnis von kleiner gleich 5 dB. Die Ergebnisse nach der Optimierung des gesamten Modells einschließlich des Quantisierers mittels SPSA werden mit KWQ bezeichnet.

Median die Referenz bei einer Bitrate von 4,67 kbit/s im Gegensatz zu 7,82 kbit/s für den AEC und einem im Vergleich deutlich reduzierten Median VSTOI-Wert. Die Datenraten steigen bei niedrigem Signal-Rausch-Verhältnis zwar etwas an, sind aber bei kleiner gleich 5 dB mit 5,06 kbit/s für den FRAE und 8,36 kbit/s für den AEC noch immer sehr gering. Im Vergleich dazu benötigt der Electrocodec mindestens eine Bitrate von 20,1 kbit/s zum Erreichen der Referenz. Opus erreicht diese erst bei etwa 33,6 kbit/s. Zudem beträgt Opus' algorithmische Latenz in diesem Fall 7,5 ms. Der G.722 wurde hier nur zum Vergleich und als Plausibilitätsprüfung inkludiert. Die Spitzenbitrate des besten FRAE auf dem Testdatensatz, berechnet als maximale summierte Codewortlängen in einem Zeitfenster von einer Sekunde, beträgt 7,64 kbit/s. Die minimale Bitrate beträgt nahezu exakt 2 kbit/s. Zur detaillierten Aufschlüsselung der Verständlichkeit der codierten Erregungsmuster sind in Abb. 4.21 die ΔVSTOI-Werte der beiden besten Autoencoder, jeweils nach Optimierung der Gesamtstruktur, in Abhängigkeit vom Signal-Rausch-Verhältnis dargestellt. Die erwähnte, und in Abb. 4.21 gut sichtbare, scheinbar reduzierte Verständlichkeit der codierten Erregungsmuster gemäß VSTOI-Werte bei hohem Signal-Rausch-Verhältnis muss jedoch mit Vorsicht interpretiert werden. Bei einem Signal-Rausch-Verhältnis von 10 dB oder mehr können bereits Deckeneffekte auftreten [WSS18b; AM22], sodass eine nominelle Reduk-



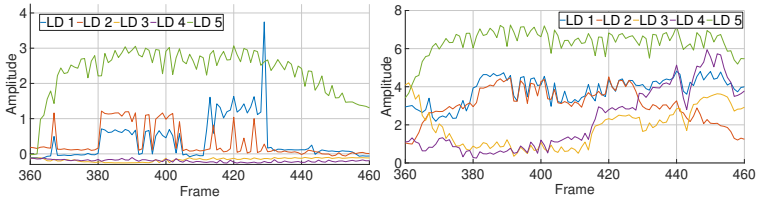
(a) AEC-L5-H2 10 Bit

(b) FRAE-L5-H3-R4 6 Bit

Abbildung 4.21: ΔVSTOI-Werte des (a) besten Autoencoders ohne Rückkopplung AEC-L5-H2 inklusive Quantisierung und (b) besten Rückkopplungsautoencoders FRAE-L5-H3-R4 inklusive Quantisierung auf dem TIMIT-Testdatensatz in Abhängigkeit vom Signal-Rausch-Verhältnis (SNR). Jeweils nach Optimierung der Gesamtstruktur inklusive des Quantisierers.

tion des VSTOI-Wertes hierbei ggf. kaum einen messbare Reduktion der Sprachverständlichkeit darstellt. Bei den durchgeführten Hörtests etwa, musste für alle Probanden bis auf einen zur sicheren Verhinderung von Deckeneffekten das Signal-Rausch-Verhältnis unter 10 dB gesenkt werden, zumeist deutlich.

Nach Abschluss dieser letzten Trainingsphase wurde erneut der verborgene Raum des AEC mit jenem des FRAE verglichen. Abbildung 4.22 zeigt den gleichen Ausschnitt wie zuvor des verborgenen Raums des AEC bzw. FRAE nach Optimierung der Gesamtstruktur bestehend aus AEC bzw. FRAE und Vektorquantisierung mit einer Codebuchgröße von 10 Bit bzw. 6 Bit. Dies sind die Einstellungen niedrigster Bitrate bei der noch die Referenz erreicht wird (FRAE) bzw. die Autoencoder ihr am nächsten sind (AEC). Anders als erwartet zeigt sich noch immer eine deutliche zeitliche Abhängigkeit der Latentsignale. Eine Untersuchung und ein Vergleich der Kreuzkorrelationen vor und nach der Optimierung, für den Ausschnitt aus Abb. 4.22 in Tabelle 4.11 zusammengefasst, offenbarte jedoch, dass die Kreuzkorrelationen der Latentsignale angestiegen ist im Vergleich zu vorher, insbesondere für den FRAE. Stichprobenhaft zeigte sich dieser Anstieg der Kreuzkorrelation durchweg. Dies motiviert die folgende Hypothese: Der FRAE nutzt die Rückkopplung zur Erhöhung der Kreuzkorrelationen, wodurch, stärker als beim AEC, die Vektorquantisierung effektiver wird. Der AEC ist dazu nur eingeschränkt in der Lage aufgrund der fehlenden Rückkopplung. Dadurch ist die Codierung insgesamt besser. Die augenscheinlichen zeitlichen Abhängigkeiten sind nach Optimierung vermehrt Schein(auto-)korrelationen, welche tatsächlich durch Kreuzkorrelationen maßgeblich erzeugt werden.



(a) AEC-L5-H2 10 Bit

(b) FRAE-L5-H3-R4 6 Bit

Abbildung 4.22: Latent Space der Datei DR2_MDEM0_SI1868 codiert durch AEC-L5-H2 und FRAE-L5-H2-R4, nachdem die Gesamtstruktur inklusive Quantisierer mittels des SPSA optimiert wurde. Auffällig ist, dass noch deutliche zeitliche Abhängigkeiten im Latent Space des FRAE vorhanden sind. Daraus lässt sich schließen, dass die erlernte Komprimierung noch nicht so gut wie möglich ist.

Zum Verständnis stelle man sich ein eindimensionales Signal mit beliebig starker zeitlicher Abhängigkeit vor. Kopiert man dieses nun einige Male und betrachtet die Signale wie hier gemeinsam, so scheint jedes Signal eine starke zeitliche Abhängigkeit aufzuweisen. Die zeitlichen Abhängigkeiten der Kopien sind jedoch eigentlich die zeitlichen Abhängigkeiten des Originalsignals. Verschwinden diese im Originalsignal, so auch in den Kopien. Ähnlich, nur nicht ganz so deterministisch, ist es beim FRAE. Durch die Erhöhung der Kreuzkorrelation verbessert sich die Vektorquantisierung, jedoch wird ein weiterer möglicher Zugewinn durch verbesserte zeitliche Prädiktion respektive Ausnutzung zeitlicher Abhängigkeiten ebenfalls reduziert. Kurz gesagt transformiert der FRAE zeitliche Abhängigkeiten in „dimensionale“ Abhängigkeiten.

Zum Abschluss soll nun noch kurz auf die Regularisierung des Rückkopplungsautoencoders, die in Abschnitt 3.4 beschrieben wurde, zur Reduktion der Disparitäten der VSTOI-Werte in Abhängigkeit vom Signal-Rausch-Verhältnis eingegangen werden. Mit dieser kann die Codierungsleistung etwas ausgeglichener gestaltet werden.

4.4.2 *Regularisierung*

Tabelle 4.13 zeigt Median-VSTOI-Werte mit und ohne Regularisierung für $L = 40,70$ sowie $L = 100$ des FRAE-L5-H2-R4 nach weiteren 7000 Iterationen Optimierung. Wie erwartet kommt es durch weiteres Training inklusive Regularisierung zu einer Reduktion der Median-VSTOI-Werte bei niedrigem Signal-Rausch-Verhältnis. Dafür jedoch steigen die Medianwerte bei hohen Signal-Rausch-Verhältnis und die Codierungsleistung

Tabelle 4.11: Kreuzkorrelation der Latentdimensionen (LD) untereinander vom (a) rückkopplungsfreien Autoencoder AEC-L5-H2 und (b) des Rückkopplungsautoencoders FRAE-L5-H3-R4 für den Ausschnitt aus Abb. 4.17 vor Einführung der Quantisierung. Es zeigt sich noch kein klarer Unterschied in der Kreuzkorrelation. (c) und (d) zeigen für denselben Ausschnitt die Kreuzkorrelation nach Einführung der Quantisierer nebst Optimierung der Gesamtstruktur. Während die Kreuzkorrelationen für den AEC näherungsweise gleich geblieben sind, haben diese sich für den FRAE zumeist erhöht, teilweise deutlich.

	LD 1	LD 2	LD 3	LD 4	LD 5
LD 1	1,00	0,37	-0,01	-0,21	0,43
LD 2	0,37	1,00	-0,53	0,39	0,34
LD 3	-0,01	-0,53	1,00	-0,45	-0,31
LD 4	-0,21	0,39	-0,45	1,00	-0,28
LD 5	0,43	0,34	-0,31	-0,28	1,00

(a) AEC-L5-H2

	LD 1	LD 2	LD 3	LD 4	LD 5
LD 1	1,00	-0,02	0,12	0,71	0,25
LD 2	-0,02	1,00	-0,49	-0,52	0,60
LD 3	0,12	-0,49	1,00	0,31	-0,55
LD 4	0,71	-0,52	0,31	1,00	0,06
LD 5	0,25	0,60	-0,55	0,06	1,00

(b) FRAE-L5-H3-R4

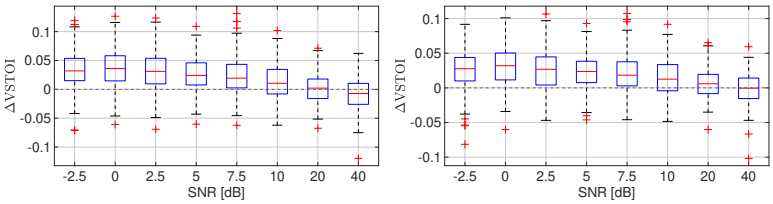
	LD 1	LD 2	LD 3	LD 4	LD 5
LD 1	1,00	0,33	-0,06	-0,32	0,43
LD 2	0,33	1,00	-0,59	0,30	0,37
LD 3	-0,06	-0,59	1,00	-0,36	-0,45
LD 4	-0,32	0,30	-0,36	1,00	-0,31
LD 5	0,43	0,37	-0,45	-0,31	1,00

(c) AEC-L5-H2 10 Bit

	LD 1	LD 2	LD 3	LD 4	LD 5
LD 1	1,00	0,35	0,13	0,46	0,49
LD 2	0,35	1,00	-0,64	-0,36	0,72
LD 3	0,13	-0,64	1,00	0,72	-0,53
LD 4	0,46	-0,36	0,72	1,00	0,03
LD 5	0,49	0,72	-0,53	0,03	1,00

(d) FRAE-L5-H3-R4 6 Bit

des Rückkopplungsautoencoders wird insgesamt etwas ausgeglichener. Bis hinab zu einem Signal-Rausch-Verhältnis von 10 dB kommt es zu einer leichten Verbesserung der Median-VSTOI-Werte. Darunter reduzieren diese sich, was eine erwartete Folge der Regularisierung ist. Abb. 4.23 zeigt Boxplots der ΔVSTOI-Werte in Abhängigkeit von dem Signal-Rausch-Verhältnis ohne Regularisierung und mit Regularisierung für $L = 70$. Es ist zum einen die Änderung der Medianwerte und zudem eine Reduktion der Streuung in den einzelnen Signal-Rausch-Verhältnisgruppen offensichtlich. Etwa lag der minimale ΔVSTOI-Wert, der kein Ausreißer war, ohne Regularisierung bei einem Signal-Rausch-Verhältnis von 40 dB bei etwa -0,075. Dieser Wert stieg auf etwas über -0,05 nach Regularisierung an. Ähnliche Änderungen zeigen sich für jedes Signal-Rausch-Verhältnis. Insgesamt kommt es, mit wenigen Ausnahmen, zu einer Reduktion von Extremwerten, sowohl im positiven als auch negativen Bereich. Optimal erscheint ein Wert von L um 70. Durch weitere Trainingsiterationen kann die Codierungsleistung nominell zumindest noch etwas ausgeglichener werden, jedoch ist die Relevanz aufgrund der erwähnten Möglichkeit von Deckeneffekten bei hohem Signal-Rausch-Verhältnis nicht klar, weswegen weitere Hörtests nötig wären, insbesondere auch für Ausreißer, die leider im Rahmen der vorgelegten Arbeit nicht durchgeführt werden konn-



(a) Ohne Regularisierung

(b) L = 70

Abbildung 4.23: Änderung der ΔVSTOI-Werte auf dem TIMIT-Testdatensatz durch Regularisierung. Die Werte sind gruppiert nach den Signal-Rausch-Verhältnissen der Dateien. In beiden Abbildungen werden die Ergebnisse des Modells FRAE-L5-H2-R4 ohne Quantisierung dargestellt.

ten. Zur Erzielung einer ausgeglicheneren Codierungsleistung wurde auch versuchsweise mit einem Trainingsdatensatz gearbeitet, welcher ausschließlich aus Dateien mit einem hohen Signal-Rausch-Verhältnis bestand. Jedoch zeigte sich auch in diesem Fall ein Abfall für ein hohes Signal-Rausch-Verhältnis, nun jedoch mit insgesamt, also im Mittel für alle Signal-Rausch-Verhältnisse, deutlich reduzierter Codierungsleistung. Hieraus lässt sich ableiten, dass die untersuchte Regularisierung zumindest gewissen Datenaugmentierungen, mittels welcher eine ausgeglichenere Codierungsleistung angestrebt werden könnte, überlegen sein dürfte.

Tabelle 4.13: Mediane der ΔVSTOI-Werte des FRAE-L5-H2-R4 in Abhängigkeit vom Regularisierungskoeffizienten L und Signal-Rausch-Verhältnissen (SNR) der Dateien auf dem TIMIT-Testdatensatz. Die Modelle mit Regularisierung wurden für zusätzliche 7000 Iterationen durch den SPSA-Algorithmus optimiert.

Regularisierung\SNR [dB]	-2,5	0	2,5	5	7,5	10	20	40	$\overline{\text{VSTOI}}$
Ohne Regularisierung	0,03196	0,03605	0,03112	0,02413	0,01932	0,01044	0,00211	-0,00719	0,6549
L = 40	0,03065	0,03416	0,02865	0,02497	0,02058	0,01352	0,00631	-0,00121	0,6553
L = 70	0,02764	0,03201	0,02682	0,02365	0,01828	0,01256	0,00595	-0,00046	0,6533
L = 100	0,02671	0,03032	0,02553	0,02203	0,01844	0,01218	0,00556	-0,00155	0,6528

5

ÜBERLEGUNGEN ZUR OPTIMALITÄT

In diesem Kapitel wird zum Electrocodec zurückgekehrt und die Optimalität des gewählten Codierungsansatzes untersucht. Es stellt sich im Kontext der Kompression der Erregungsmuster von Cochlea-Implantaten z.B. die Frage, ob nicht etwa ein konventioneller Audiocodec, der an die Besonderheiten von Cochlea-Implantatträgern bzw. von Cochlea-Implantaten angepasst ist, gleiche oder bessere Leistungsfähigkeit erzielen kann als ein Codec auf Basis der Erregungsmuster. Eine andere Frage ist, in wie weit die Komponenten des Electrocodecs, also etwa die lineare Prädiktion, optimal sind. Grundsätzlich, bei jedem Kompressionsproblem, ist es interessant, Aussagen über die Optimalität eines Kompressionsansatzes respektive einer erzielten Kompressionsleistung zu gewinnen. Für verlustlose Kompressionsalgorithmen kann dies durch die Entropie einer Datenquelle oft zumindest approximativ bewerkstelligt werden. Allgemeiner, wie in Abschnitt 2.3.5 erläutert, liefert entsprechendes die sogenannte Raten-Verzerrungs-Funktion für verlustbehaftete Kompressionsalgorithmen.

Initial sollte in diesem Kapitel die Raten-Verzerrungs-Funktion des Advanced Combination Encoders beim Anliegen Gaußscher Prozesse bestimmt werden. Man könnte dann Optimalcodierungen Gaußscher Prozesse, für welche die Raten-Verzerrungs-Funktion, unter Verwendung des mittleren quadratischen Fehlers als Verzerrungsmaß, bekannt ist, bezüglich der assoziierten Verzerrung der vom Advanced Combination Encoder aus diesen codierten Gaußschen Prozessen erzeugten Erregungsmuster mit vom Electrocodec codierten Erregungsmustern, die ebenfalls aus diesem Gaußschen Prozess generiert wurden, vergleichen. Um hier jedoch zu einem Ergebnis zu kommen, müsste ein Verzerrungsmaß für die Erregungsmuster festgelegt werden. Jedoch liefern die subjektiven Untersuchungen, die im Rahmen der vorgelegten Arbeit durchgeführt wurden, nur grobe Anhaltspunkte für den Zusammenhang der Verzerrung der Erregungsmuster mit der assoziierten Sprachverständlichkeit. Es ist unklar, wie genau etwa eine Verzerrung der Bandselektion ge-

wichtet werden sollte, deren großer Einfluss aus [Qaz+13] bekannt ist. Verwertbare, quantitative Erkenntnisse diesbezüglich existieren in der Literatur nicht.

Eine weitere Hürde für eine umfassende Untersuchung ist durch das Fehlen von adäquaten Signalmodellen für stimmhafte Sprache gegeben. Gaußsche Prozesse stellen ein sehr gutes Modell für stimmlose Sprache, also Sprachanteile ohne Schwingung der Stimmbänder, dar [KA93]. Sie sind jedoch ungeeignet für die Modellierung stimmhafter Sprache [Att11], welche eine starke Periodizität aufweisen, und bei welchen ein einfaches Histogramm eine bimodale univariate Verteilung offenbart, sodass ein Gaußscher Prozess ein schlechtes Modell darstellen muss. Eine umfassende theoretische Untersuchung würde ein vollständiges statistisches Modell von Sprachsignalen benötigen, welches zum derzeitigen Stand der Wissenschaft nicht existiert.

Es wurde daher von diesem Vorhaben Abstand genommen, da insbesondere die Leistungsfähigkeit der vorgestellten Autoencoder nahe gelegt hat, welch niedrige Datenraten erzielbar sind, sofern auf Sprachverständlichkeit als Zielgröße optimiert wird.

An Stelle der Raten-Verzerrungs-Funktion wird in diesem Kapitel die Prädiktion der Erregungsmuster untersucht, um zu eruieren, ob der Electrocodec durch nichtlineare Prädiktion verbessert werden kann. Des Weiteren erfolgt eine Abschätzung der Entropie der Bandselektion, um zu untersuchen, inwieweit die Datenrate des Electrocodecs durch eine Verbesserung der verlustlosen Kompression der Bandselektion reduziert werden könnte. Das Ergebnis ist eine Abschätzung der theoretischen Kompetitivität des konventionellen Codierungsansatzes des Electrocodecs mit der Codierung der Erregungsmuster durch Autoencoder.

Auch für die genannten Untersuchungen ist ein statistisches Signalmodell der Eingangssignale des Cochlea-Implantats notwendig. Wie erwähnt stellen Gaußsche Prozesse ein gutes Modell für stimmlose Sprache dar. Daher beschränkt sich dieses Kapitel auf stimmlose Sprache und es werden Gaußsche Prozesse als Eingangssignale betrachtet.

Die Struktur dieses Kapitels ist nun wie folgt:

- Zunächst wird durch einigen Rechenaufwand plausibel gemacht, dass eine analytische Lösung der Fragestellungen nicht möglich ist
- Anschließend wird mittels spezieller Schätzverfahren der Optimalprädiktor, also der bedingte Erwartungswert, auf Basis einer großen Datenmenge für die Erregungsmuster bestimmt. Auf dieser Basis kann abgeschätzt werden, ob nichtlineare Prädiktion eine bedeutende Verbesserungsmöglichkeit des Electrocodecs darstellt.

- Zum Schluss wird auf Basis einer großen Datenmenge die Entropie der Bandselektion bestimmt. Diese wird dann mit der mittleren Wortlänge der komprimierten Darstellung der Bandselektion des Electrocodecs verglichen. Hierdurch ist das Verbesserungspotential der verlustlosen Kompression des Electrocodecs abschätzbar.

5.1 GRUNDSÄTZLICHES VORGEHEN

Die Forschungsimplementierung des Advanced Combination Encoders besteht aus den Schritten Diskrete Fouriertransformation (inklusive Fensterung), der Einhüllendenberechnung, der Bandselektion sowie der Lautheitswachstumfunktion (LGF). Der letzte Schritt, die Abbildung auf Stromwerte in klinischen Einheiten, ist an dieser Stelle nicht relevant, da der Electrocodec das LGF-Ausgangssignal nutzt.

Ist ein statistisches Eingangssignalmodell gegeben, so kann auf Basis der bekannten Folge von Rechenoperationen, aus denen die Erregungsmuster hervorgehen, ein statistisches Modell der Ausgangsdaten, der Erregungsmuster, gewonnen werden. Jedoch stößt man bereits beim vermutlich einfachsten Fall von statistisch unabhängigem, normalverteilten Rauschen, dem einfachsten Gaußschen Prozess, an analytische Grenzen, wie im Folgenden gezeigt wird.

5.2 WAHRSCHEINLICHKEITSDICHTEFUNKTIONEN NACH DER DFT

Gegeben mittelwertfreies, weißes Gaußsches Rauschen x_n mit der Varianz σ^2, d.h.

$$x_n \sim \mathcal{N}(0, \sigma^2), \forall n \in \mathbb{Z} \tag{5.1}$$

und einer DFT der Länge L berechnet gemäß

$$X(k) = \sum_{n=0}^{L-1} x_n e^{-j\frac{2\pi n k}{L}}. \tag{5.2}$$

Real- und Imaginärteil von X_k sind gegeben zu

$$\Re\{X_k\} := X_{Re}(k) = \sum_{n=0}^{L-1} x_n \cos\left(\frac{2\pi n k}{L}\right) \tag{5.3}$$

sowie

$$\Im\{X_k\} := X_{Im}(k) = \sum_{n=0}^{L-1} x_n \left(-\sin\left(\frac{2\pi n k}{L}\right)\right). \tag{5.4}$$

Es lässt sich zeigen, siehe Abschnitt 8.1 im Anhang für eine ausführliche Rechnung, dass die DFT-Koeffizienten paarweise statistisch unabhängig sind und einer komplexen Normalverteilung folgen. Genauer gilt

$$\begin{aligned} X_{Im}(k) &= 0 \\ X_{Re}(k) &\sim \mathcal{N}\left(0, \sigma^2 L\right) \end{aligned} \tag{5.5}$$

für $k = l \cdot \frac{L}{2}, l \in \mathbb{Z}$ sowie für alle anderen DFT-Koeffizienten

$$\begin{aligned} X_{Im}(k) &\sim \mathcal{N}\left(0, \sigma^2 \tfrac{L}{2}\right) \\ X_{Re}(k) &\sim \mathcal{N}\left(0, \sigma^2 \tfrac{L}{2}\right) \end{aligned} . \tag{5.6}$$

Der nächste Schritt in der Berechnung der Erregungsmuster ist die Bestimmung der Einhüllenden $a(z)$.

5.3 DIE WAHRSCHEINLICHKEITSDICHTE DER EINHÜLLENDEN

Die nachfolgende Rechnung gilt für alle Bänder bis auf das 22. Band, da dieses gemäß Tabelle 2.1 den DFT-Koeffizienten $L/2$ mit $L = 128$ beinhaltet, welcher eine höhere Varianz als die übrigen Koeffizienten aufweist. Zudem hängt die Einhüllende des 22. Bandes sowohl vom 63. als auch 65. DFT-Koeffizient ab, welche aufgrund der Symmetriebeziehung $X(k) = \overline{X(N-k)}$ voneinander abhängen. Sie sind nicht statistisch unabhängig. Die Änderungen für dieses Band werden am Ende dieses Abschnitts diskutiert.

Für das Quadrat der Einhüllenden $a(z)$ des Bands z gilt

$$a^2(z) = \sum_{n=n_{start_z}}^{n_{end_z}-1} g_z \cdot \left(X_{Re}(n)^2 + X_{Im}(n)^2\right). \tag{5.7}$$

g_z ist der Gain-Faktor von Band z, n_{start_z} ist der erste DFT-Koeffizient in Band z, während $n_{end_z} - 1$ der letzte DFT-Koeffizient in Band z ist. Die Anzahl der verwendeten DFT-Koeffizienten in Band z ist $N_z = n_{end_z} - n_{start_z} + 1$.

Zur übersichtlichen Berechnung werden die Hilfszufallsvariablen

$$R(n) \sim \mathcal{N}(0,1) \tag{5.8}$$

sowie

$$I(n) \sim \mathcal{N}(0,1) \tag{5.9}$$

eingeführt. Es gilt dann die Beziehung

$$X_{Re}(n) = \sigma\sqrt{\frac{L}{2}} R(n) \tag{5.10}$$

sowie

$$X_{Im}(n) = \sigma\sqrt{\frac{L}{2}}I(n). \tag{5.11}$$

Es folgt für $a^2(z)$:

$$\begin{aligned} a^2(z) &= \sum_{n=n_{start_z}}^{n_{end_z}-1} g_z\sigma^2\frac{L}{2}\cdot\left(R(n)^2+I(n)^2\right) \\ &= \underbrace{g_z\sigma^2\frac{L}{2}}_{:=\sigma_z^2}\underbrace{\sum_{n=n_{start_z}}^{n_{end_z}-1}(R(n)^2+I(n)^2)}_{:=Y} \\ &= \sigma_z^2 Y \end{aligned} \tag{5.12}$$

Die Zufallsvariable Y ist die Summe von $2N_z$ unabhängigen, quadrierten, standardnormalverteilten Zufallsvariablen und ist damit χ^2-verteilt, d.h.

$$Y \sim \chi^2(2N_z). \tag{5.13}$$

Für die Wahrscheinlichkeitsdichte von $a^2(z)$ folgt, mit der Dichte der χ^2-Verteilung sowie des allgemeinen Verhaltens der Wahrscheinlichkeitsdichte skalierter Zufallsvariablen gemäß Gl. 2.22, die Dichte

$$f_{a^2(z)}(x) = \begin{cases} \dfrac{\left(\frac{1}{2\sigma_z^2}\right)^{N_z}}{\Gamma(N_z)}x^{N_z-1}e^{-\left(\frac{1}{2\sigma_z^2}\right)x}, & x > 0\,. \\ 0, & x \leqslant 0 \end{cases} \tag{5.14}$$

Wegen $N_z \in \mathbb{N}$, und damit[1] $\Gamma(N_z) = (N_z-1)!$, entspricht $f_{a^2(z)}(x)$ der Dichte der Erlang-Verteilung $Erl(n,\lambda)$ mit $n = N_z$ sowie $\lambda = \frac{1}{2\sigma_z^2}$ [Ibe14]. Die Verteilungsfunktion der Erlang-Verteilung wird in Kürze benötigt. Diese ist gegeben zu [Ibe14]

$$F_{Erl}(x|n,\lambda) = \begin{cases} 1-e^{-\lambda x}\sum_{k=0}^{n-1}\frac{(\lambda x)^k}{k!} & x \geqslant 0 \\ 0 & x < 0 \end{cases}. \tag{5.15}$$

Mittels des Transformationssatzes für Wahrscheinlichkeitsdichten 2.21 kann man dann, wegen $a(z) \geqslant 0$, mit $a(z) = \sqrt{a^2(z)} := g(a^2(z)) \equiv g(Y)$

1 $\Gamma(x)$ ist die sogenannte Gammafunktion, definiert zu $\int_0^\infty t^{x-1}e^{-t}\,dt$. Für $n \in \mathbb{N}\setminus\{0\}$ gilt $\Gamma(n) = (n-1)!$.

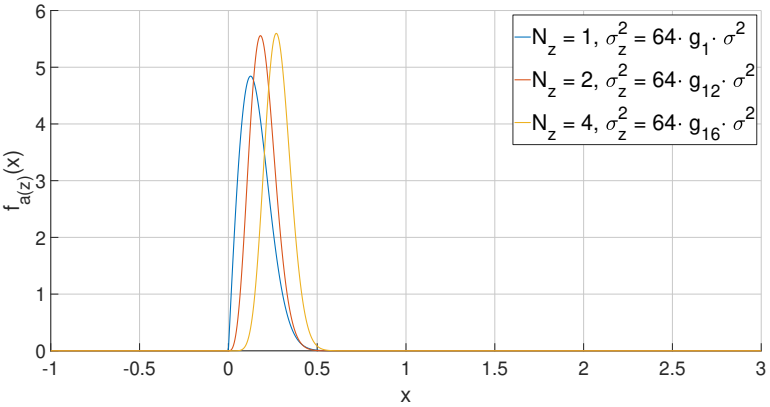


Abbildung 5.1: Die Dichte $f_{a(z)}(x)$ der zentralen Chi-Verteilung nach Gl. 5.16 für verschiedene Beispielparameter mit $\sigma = 0,5$.

sowie $g'(y) = -\frac{1}{2\sqrt{y}}$ leicht die Dichte von $a(z)$ bestimmen. Diese berechnet sich zu

$$f_{a(z)}(x) = \begin{cases} \frac{x^{2N_z-1}}{2^{N_z-1}(\sigma_z^2)^{N_z}\Gamma(N_z)} e^{-\frac{x^2}{2\sigma_z^2}}, & x > 0 \\ 0, & x \leqslant 0 \end{cases} \tag{5.16}$$

und ist die Dichte $f(x|\sigma, L)$ der zentralen Chi-Verteilung [AV16] mit den Parametern $\sigma = \sigma_z$ und $L = N_z$. Beispielvisualisierungen dieser Dichte sind in Abb. 5.1 für verschiedene Werte von N_z und g_z, welche zu den Bändern 1,12 sowie 16 nach Tabelle 2.1 gehören, bei einer Standardabweichung $\sigma = 0,5$ zusammengefasst.

5.3.1 *Wahrscheinlichkeitsdichte der Einhüllenden des 22. Bandes*

Wie erwähnt muss die Dichte von $a^2(z)$ für $z = 22$, d.h. dem 22. Band, separat betrachtet werden, da nach Gl. 5.5 die Verteilung des $\frac{L}{2}$-ten DFT-Koeffizienten anders berechnet wird als für alle anderen Koeffizienten, und gemäß Tabelle 2.1 dieser Koeffizienten in die Berechnung von $a^2(22)$ eingeht. Genauer werden die DFT-Koeffizienten 58,59,...,63, 64,65 für diese Berechnung verwendet. Für die Koeffizienten 63 und 65 gilt $X(63) = \overline{X(L-63)} = \overline{X(65)}$, d.h. der Wert des einen Koeffizienten legt den Wert des anderen fest. Das Analogon von Gl. 5.12 für das Band 22 ist

$$a^2(22) = \sum_{n=1}^{5} g_{22}\sigma^2\tfrac{L}{2}\left(R(n)^2 + I(n)^2\right) + g_{22}\sigma^2 L \cdot R^2(64) + 2 \cdot g_{22}\sigma^2\tfrac{L}{2}\left(R(63)^2 + I(63)^2\right) \tag{5.17}$$

Der Vorfaktor 2 des letzten Terms stammt von der Zusammenfassung des Betragsquadrats des 63. und des 65. DFT-Koeffizienten. Das Betragsquadrat des 65. DFT-Koeffizienten ist immer identisch zum 63., weswegen dieser mittels einer Verdopplung des Betragsquadrats des 63. DFT-Koeffizienten berücksichtigt werden kann. Gl. 5.17 lässt sich etwas zusammenfassen zu

$$a^2(22) = g_{22}\sigma^2\frac{L}{2}\underbrace{\sum_{n=1}^{5}\left(R(n)^2 + I(n)^2\right)}_{=:A} + g_{22}\sigma^2 L\underbrace{\left(R(63)^2 + I(63)^2 + R^2(64)\right)}_{=:B}.$$

Der Ausdruck A ist $\chi^2(10)$ verteilt, B entsprechend $\chi^2(3)$ verteilt. Um nun die Dichte von $a^2(22)$ zu bestimmen, wird zunächst die Dichte von $\frac{A}{2}$ bestimmt, d.h. es wird der Vorfaktor $\frac{1}{2}$ von A berücksichtigt. Anschließend wird mittels der Bestimmungsformel 2.23 für die Summe zweier statistisch unabhängiger Zufallsvariablen die Dichte von $\frac{A}{2} + B$ bestimmt. Schlussendlich wird noch der Vorfaktor $g_{22}\sigma^2 L$ eingepreist. Die Dichte $f(x)$ einer $\chi^2(k)$ verteilten Zufallsvariablen mit Freiheitsgrad k ist

$$f(x) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}, x \geqslant 0$$

und sonst Null[2]. Folglich ist die Dichte $f_A(x)$ von A gegeben zu

$$f_A(x) = \frac{1}{2^5\Gamma(5)}x^4 e^{-\frac{x}{2}}, x \geqslant 0$$

und die Dichte $f_B(x)$ von B ist gegeben zu

$$f_B(x) = \frac{1}{2^{\frac{3}{2}}\Gamma(\frac{3}{2})}x^{\frac{1}{2}}e^{-\frac{x}{2}}, x \geqslant 0.$$

Die Dichte von $f_{\frac{A}{2}}(x)$ bestimmt sich gemäß Gl. 2.22 zu

$$f_{\frac{A}{2}}(x) = 2f_A(2x) = 2\left(\frac{1}{2^5\Gamma(5)}(2x)^4 e^{-\frac{2x}{2}}\right) = \frac{x^4}{\Gamma(5)}e^{-x}, x \geqslant 0.$$

2 Dies wird im Folgenden nicht mehr extra erwähnt werden.

Die Dichte $f_{\frac{A}{2}+B}(x)$ von $\frac{A}{2}+B$ lässt sich mittels der Faltungsformel 2.23, unter Berücksichtigung des Trägers der Dichten, berechnen gemäß

$$\begin{aligned} f_{\frac{A}{2}+B}(x) &= \int_{-\infty}^{\infty} f_{\frac{A}{2}}(\alpha) f_B(x-\alpha)\,d\alpha = \int_0^x f_{\frac{A}{2}}(\alpha) f_B(x-\alpha)\,d\alpha \\ &= \int_0^x \left(\frac{\alpha^4}{\Gamma(5)} e^{-\alpha}\right)\left(\frac{\sqrt{x-\alpha}}{2^{\frac{3}{2}}\Gamma(\frac{3}{2})} e^{-\frac{x-\alpha}{2}}\right) d\alpha \\ &= \underbrace{\frac{1}{2^{\frac{3}{2}}\Gamma(5)\Gamma(\frac{3}{2})}}_{=:a} \underbrace{\int_0^x \alpha^4 e^{-\alpha}\sqrt{x-\alpha} e^{-\frac{x-\alpha}{2}}\,d\alpha}_{=:I(x)}. \end{aligned}$$

Es ist $a = \frac{1}{24\sqrt{2\pi}}$, was aus $\Gamma(\frac{3}{2}) = \frac{\pi}{2}$ folgt. Letzteren Wert kann man üblichen Tafelwerken entnehmen. Das Integral $I(x)$ ergibt

$$\begin{aligned} I(x) &= \int_0^x \alpha^4 e^{-\alpha}\sqrt{x-\alpha} e^{-\frac{x-\alpha}{2}}\,d\alpha \\ &= 2e^{-\frac{x}{2}}\left(\sqrt{x}(x^3 + 13x^2 + 105x + 945)\right. \\ &\quad \left. - \sqrt{2}(x^4 + 12x^3 + 90x^2 + 420x + 945) F\left(\sqrt{\frac{x}{2}}\right)\right) \end{aligned}$$

mit $F(x) = e^{-x^2}\int_0^x e^{u^2}\,du$. Dieses Integral wurde mit Hilfe des Computeralgebrasystems Wolfram Alpha bestimmt und hat aufgrund der Funktion F keine analytische Lösung. Es ist $I(0) = 0$ sowie $\lim_{x\to\infty} I(x) = 0$, letzteres wegen $e^{-\frac{x}{2}} F(\sqrt{\frac{x}{2}}) \leqslant e^{-\frac{x}{2}}\sqrt{\frac{x}{2}}$.

Insgesamt ergibt sich dann die Dichte $f_{a^2(22)}(x)$ zu

$$f_{a^2(22)}(x) = f_{g_{22}\sigma^2 L(\frac{A}{2}+B)}(x) = \frac{f_{\frac{A}{2}+B}(\frac{x}{g_{22}\sigma^2 L})}{g_{22}\sigma^2 L} = \frac{a}{g_{22}\sigma^2 L} I\left(\frac{x}{g_{22}\sigma^2 L}\right).$$

Die zugehörige Verteilungsfunktion existiert zwar, besitzt jedoch ebenfalls keinen geschlossen Ausdruck und sei hier formal angegeben als $F_{a^2(22)}(x) = \int_{-\infty}^x f_{a^2(22)}(u)du$. Sie ist für $x \leqslant 0$ identisch Null. Die Dichte von $f_{a(22)}$ ergibt sich wie für die anderen Bänder aus $f_{a^2(22)}(x)$ mittels des Transformationssatzes für Wahrscheinlichkeitsdichten 2.21 und $a(22) = \sqrt{a^2(22)} := g(a^2(22)) \equiv g(Y)$ sowie $g'(y) = -\frac{1}{2\sqrt{y}}$. Es ist also

$$f_{a(22)}(x) = f_{a^2(22)}(x^2)2x = 2x\frac{a}{g_{22}\sigma^2 L} I\left(\frac{x^2}{g_{22}\sigma^2 L}\right).$$

Diese Dichte, zusammen mit einer Approximation, ist für $\sigma = 0,5$ in Abb. 5.2 dargestellt. Die Approximation wurde auf Basis der zentralen Chi-Verteilung nach Gl. 5.16 wie am Ende des vorigen Unterabschnittes berechnet, jedoch mit $N_z = 8$ und $g_{22} = 0,65 \cdot 10^{-3}$ aus Tabelle 2.1.

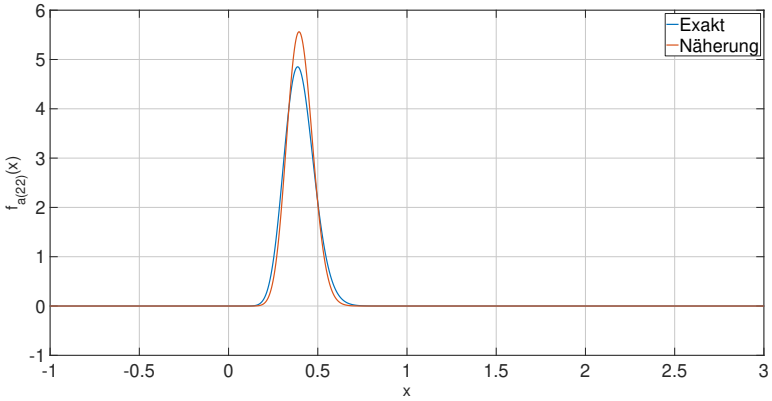


Abbildung 5.2: Die Dichte $f_{a(22)}(x)$ einmal exakt nach Gl. 5.3.1 mit $\sigma = 0,5$ und einmal genähert durch eine zentrale Chi-Verteilung wie in Abb. 5.1. Diese Näherung wurde mit g_{22} aus Tabelle 2.1, $N_z = 8$ sowie $\frac{L}{2} = 64$ berechnet.

5.4 WAHRSCHEINLICHKEITSDICHTE NACH DER LAUTHEITSWACHSTUMSFUNKTION

Mit der Dichte $f_{a(z)}(x)$ kann man nun die Dichte des Ausgangssignals der Lautheitswachstumsfunktion 2.7 berechnen. Das Ausgangssignal der Lautheitswachstumsfunktion des Bands z sei im Folgenden mit $l(z)$ notiert. Für den Fall $a(z) < s_{base}$ gibt es keine Ausgabe. Die Wahrscheinlichkeit $P(l(z) = \text{„keine Ausgabe“})$ bestimmt sich gemäß

$$P(l(z) = \text{„keine Ausgabe“}) = P(a(z) < s_{base}) = \int_{-\infty}^{s_{base}} f_{a(z)}(x)dx = F_{a(z)}(s_{base}) \tag{5.18}$$

mit der Verteilungsfunktion $F_{a(z)}(x)$ von $a(z)$. Analog erhält man

$$P(l(z) = 1) = P(a(z) \geqslant m_{sat}) = \int_{m_{sat}}^{\infty} f_{a(z)}(x)dx = 1 - F_{a(z)}(m_{sat}). \tag{5.19}$$

Die Dichte für $l(z) \in [0, 1)$ kann erneut über die Transformationsformel für Wahrscheinlichkeitsdichten bestimmt werden. Dazu wird die Umkehrfunktion der Lautheitswachstumsfunktion für diesen Werteberech benötigt. Die Umkehrfunktion ist gegeben zu

$$a(z) = \frac{1}{\rho}(m_{sat} - s_{base})\left((1+\rho)^{l(z)} - 1\right) + s_{base} \tag{5.20}$$

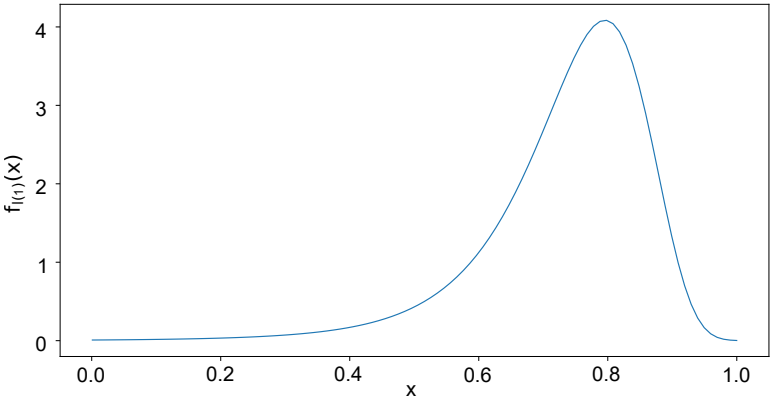


Abbildung 5.3: Dichte $f_{l(1)}(x)$ des Ausgangssignals der Lautheitswachstumsfunktion von Band 1 gemäß Gl. 5.22 für $\sigma = 0,5$.

und damit folgt für die Ableitung nach $l(z)$

$$\left|\frac{da(z)}{dl(z)}\right| = \frac{1}{\rho}\left(m_{sat} - s_{base}\right)(1+\rho)^{l(z)} \ln(1+\rho). \tag{5.21}$$

Damit ergibt sich die Dichte $f_{l(z)}(l(z))$ für $l(z) \in [0,1)$ sowie $z \neq 22$ zu

$$\begin{aligned} f_{l(z)}(l(z)) &= f_{a(z)}\left(\frac{1}{\rho}\left(m_{sat} - s_{base}\right)\left((1+\rho)^{l(z)} - 1\right) + s_{base}\right)\left|\frac{da(z)}{dl(z)}\right| \\ &= \frac{2^{1-N_z}}{\Gamma(N_z)\,\sigma_z^{2N_z}}\left(\frac{1}{\rho}\left(m_{sat} - s_{base}\right)\left((1+\rho)^{l(z)} - 1\right) + s_{base}\right)^{2N_z-1} \\ &\cdot \exp\left(-\frac{\left(\frac{1}{\rho}\left(m_{sat} - s_{base}\right)\left((1+\rho)^{l(z)} - 1\right) + s_{base}\right)^2}{2\sigma_z^2}\right) \\ &\cdot \frac{1}{\rho}\left(m_{sat} - s_{base}\right)(1+\rho)^{l(z)} \ln(1+\rho) \end{aligned} \tag{5.22}$$

Abb. 5.3 zeigt diese Dichte für das Band 1 mit $\sigma = 0,5$.

Für das 22. Band erfolgt die Rechnung analog, jedoch unter Verwendung der Wahrscheinlichkeitsdichte $f_{a(22)}(x)$.

Nun folgt die Bestimmung der Wahrscheinlichkeitsverteilung der Bandselektion, welche für die Bestimmung der Entropie ebendieser nötig ist.

5.5 WAHRSCHEINLICHKEITSVERTEILUNG DER BANDSELEKTION

Die Bandselektion wählt in jedem Zeitschritt N der $M = 22$ Bänder mit den höchsten Amplituden aus. Die ausgewählten Bänder seien im Folgenden $z_1, \ldots, z_N$ und die nicht ausgewählten $z_{N+1}, \ldots, z_M$, wobei zusätzlich zur Vereinfachung $z_i = i$ angenommen werden soll. Wegen $a < b \Rightarrow a^2 < b^2$ kann die weitere Berechnung auf Basis von $a^2(z)$ durchgeführt werden.

Zunächst wird eine Zufallsvariable Y definiert, die das Minimum der ausgewählten Bänder beschreibt:

$$Y := \min\left\{a^2(z_1), a^2(z_2), ..., a^2(z_N)\right\}. \tag{5.23}$$

Die Wahrscheinlichkeit dafür, dass die Bänder $z_1, z_2, ..., z_N$ ausgewählt werden, ist durch

$$\begin{aligned} P(z_1, z_2, ..., z_N \text{ ausgewählt}) &= \int_{-\infty}^{\infty} P\left(z_1, z_2, ..., z_N \text{ ausgewählt}, Y = x\right) dx \\ &= \int_{-\infty}^{\infty} f_Y(x)\, P\left(z_1, z_2, ..., z_N \text{ ausgewählt} | Y = x\right) dx \end{aligned}$$

gegeben. Dabei ist $f_Y(x)$ die Wahrscheinlichkeitsdichtefunktion von Y. $f_Y(x)$ lässt sich über die Ableitung der Verteilungsfunktion F_Y bestimmen:

$$\begin{aligned} P(Y > x) &= P\left(a^2(z_1) > x, a^2(z_2) > x, ..., a^2(z_N) > x\right) \\ &= P\left(a^2(z_1) > x\right) \cdot P\left(a^2(z_2) > x\right) \cdot ... \cdot P\left(a^2(z_N) > x\right) = 1 - F_Y(Y \leqslant x), \end{aligned}$$

also $F_Y(Y \leqslant x) = 1 - \prod_{i=1}^{N}\left(1 - F_{a^2(z_i)}(x)\right)$. Mit dem allgemeinen Zusammenhang $f_X(\alpha) = \frac{dF_X(\alpha)}{d\alpha}$ folgt dann

$$f_Y(x) = \frac{d}{dx}F_Y(x) = \sum_{i=1}^{N} f_{a^2(z_i)}(x) \prod_{k=1, k\neq i}^{N}\left(1 - F_{a^2(z_k)}(x)\right). \tag{5.24}$$

Es fehlt noch die Berechnung von $P(z_1, z_2, ..., z_N \text{ ausgewählt} | Y = x)$. Da $Y = \min\left\{a^2(z_1), a^2(z_2), ..., a^2(z_N)\right\}$, sind $a^2(z_1), a^2(z_2), ..., a^2(z_N) \geqslant x$. $a^2(z_{N+1}), ..., a^2(z_M)$ sind statistisch unabhängig von Y und unabhängig voneinander. Damit $z_1, z_2, ..., z_N$ ausgewählt werden, müssen $a^2(z_{N+1}), ..., a^2(z_M) \leqslant x$ sein. Daraus folgt

$$P(z_1, z_2, ..., z_N \text{ ausgewählt} | Y = x) = \prod_{j=N+1}^{M} F_{a^2(z_j)}(x). \tag{5.25}$$

Die Wahrscheinlichkeit dafür, dass die Bänder $z_1, z_2, ..., z_N$ ausgewählt werden, ist also

$$P(z_1, z_2, ..., z_N \text{ ausgewählt}) = \int_{-\infty}^{\infty} \sum_{i=1}^{N} f_{a^2(z_i)}(x) \prod_{k=1, k\neq i}^{N} \left(1 - F_{a^2(z_k)}(x)\right) \prod_{j=N+1}^{M} F_{a^2(z_j)}(x) dx \tag{5.26}$$

Durch Einsetzen der Dichtefunktion 5.14 und Verteilungsfunktion von $a^2(z)$ mit den Parametern $\lambda = \lambda_{z_i} = \frac{1}{2\sigma_{z_i}^2}$ und $n = N_{z_i}$ gemäß Gleichung erhält man

$$P(z_1, z_2, ..., z_N \text{ ausgewählt}) = \int_0^{\infty} \sum_{i=1}^{N} \frac{\lambda_{z_i}^{N_{z_i}} x^{N_{z_i}-1} e^{-\lambda_{z_i} x}}{\Gamma(N_{z_i})} \prod_{j=1, j\neq i}^{N} (1 - F_{Erl}(x|N_{z_j}, \lambda)) \left(\prod_{l=N+1}^{M-1} F_{Erl}(x|N_{z_l}, \lambda)\right) F_{a^2(22)}(x) dx$$

Das Einsetzen der Verteilungsfunktionen führt dann auf

$$P(a(z_1), a(z_2), ..., a(z_N) \text{ ausgewählt}) = \int_0^{\infty} \sum_{i=1}^{N} \frac{\lambda_{z_i}^{N_{z_i}} x^{N_{z_i}-1} e^{-\lambda_{z_i} x}}{\Gamma(N_{z_i})} \prod_{j=1, j\neq i}^{N} \left(e^{-\lambda_{z_j} x} \sum_{m=0}^{N_{z_j}-1} \frac{(\lambda_{z_j} x)^m}{m!}\right) \prod_{l=N+1}^{M-1} \left(1 - e^{-\lambda_{z_l} x} \sum_{n=0}^{N_{z_l}-1} \frac{(\lambda_{z_l} x)^n}{n!}\right) \cdot F_{a^2(22)}(x) dx$$

Zwar existiert dieses Integral, jedoch kann keine geschlossene Lösung gefunden werden, da die Verteilungsfunktion $F_{a^2(22)}(x)$, formal definiert in Abschnitt 5.3.1, keine geschlossene Lösung besitzt. Nutzt man für $F_{a^2(22)}(x)$ eine Approximation auf Basis der zentralen Chi-Verteilung, wie sie in Abb. 5.2 eingezeichnet und in Gl. 5.15 gegeben ist, so erhält man das Integral

$$\int_0^{\infty} \sum_{i=1}^{N} \frac{\lambda_{z_i}^{N_{z_i}} x^{N_{z_i}-1} e^{-\lambda_{z_i} x}}{\Gamma(N_{z_i})} \prod_{j=1, j\neq i}^{N} \left(e^{-\lambda_{z_j} x} \sum_{m=0}^{N_{z_j}-1} \frac{(\lambda_{z_j} x)^m}{m!}\right) \prod_{l=N+1}^{M} \left(1 - e^{-\lambda_{z_l} x} \sum_{n=0}^{N_{z_l}-1} \frac{(\lambda_{z_l} x)^n}{n!}\right) dx.$$

Dieses Integral ist zwar prinzipiell geschlossen lösbar, denn es lässt sich auf einen Ausdruck der Art

$$\sum_{i=1}^{N} \frac{\alpha_i \lambda_i^{N_i}}{\Gamma(N_i)} \int_0^{\infty} x^{N_i - 1} e^{-\Lambda_i x} P_i(x) dx \tag{5.27}$$

mit $\Lambda_{z_i}, \alpha_i \in \mathbb{R}_+$ und Polynomen $P_i(x)$ der Ordnung L_i

$$P_i(x) := \sum_{k=0}^{L_i} p_k^i x^k \tag{5.28}$$

herunterbrechen. Hierbei ist das Superskript i von p_k^i keine Potenz sondern ein Index. Für eine geschlossene Lösung muss jedoch wiederholt die Cauchy-Produktformel (für den Spezialfall von Polynomen) sowie auf das Multinomtheorem zurückgegriffen werden. Durch die Cauchy-Produktformel kommt es zu iterierten Faltungen von Polynomkoeffizienten, was zu einer nicht mehr überschaubaren Lösung führt, welche zusätzlich durch die Anwendung des Multinomtheorems zur Vereinfachung des zweiten Produkts noch deutlich unhandlicher wird. Das Problem ist jedoch noch komplexer, da zur Bestimmung der Entropie der Bandselektion zusätzlich alle so berechneten Wahrscheinlichkeiten summiert und logarithmiert werden müssen. Mit sehr viel Rechnen erhält man dann für $P(a(z_1), a(z_2), ..., a(z_N)$ ausgewählt$)$ den Ausdruck

$$\sum_{i=1}^{N} \sum_{\underline{k} \leqslant 1} \sum_{s=0}^{Q(\underline{k})} \sum_{m=0}^{N^{*i}} (-1)^{|\underline{k}|} \binom{N_i+s+m-1}{N_i-1} (s+m)! r_s^{\underline{k}} c_m^{*i} \frac{\lambda_i^{N_i}}{((\sum_{j=1}^{N} \lambda_j) + <\lambda_{N,M}, \underline{k}>)^{(N_i+s+m)}}$$

mit dem Multiindex $\underline{k} = (k_{N+1}, \ldots, k_M), k_i \in \{0, 1\}$ sowie Konstanten $Q(\underline{k}), N^{*i} \in \mathbb{N}$, dem Vektor $\lambda_{N,M} := (\lambda_{N+1}, \ldots, \lambda_M)$ und dem Skalarprodukt $< \lambda_{N,M}, \underline{k} >:= \sum_{i=N+1}^{M} \lambda_i k_i$. Es ist $\underline{k} \leqslant 1 := \{\underline{k} | k_{N+1} \leqslant 1, \ldots, k_M \leqslant 1\}$. Insbesondere der Binomialkoeffizient, da nur selten geschlossene Lösungen für Summen existieren, sowie die Koeffizienten $r_s^{\underline{k}}$ und c_m^{*i}, die aus sehr vielen Faltungen anderer Koeffizienten hervorgehen, verunmöglichen eine geschlossene Lösung für die Entropie der Bandselektion.

5.6 WEITERE NOTWENDIGE SCHRITTE EINER THEORETISCHEN ANALYSE

Die Rechnungen in den vorangegangenen Abschnitten illustrieren die Komplexität der notwendigen Rechnungen bereits für den einfachsten Fall von weißem Rauschen sowie unbedingten Wahrscheinlichkeiten und unter Nutzung einer Approximation. Hierbei wurde zudem zusätzlich vereinfachend die Fensterung des Advanced Combination Encoders ignoriert, wobei diese sich hierbei nur durch Multiplikation der Varianz des Rauschens mit einem Skalar erkennbar macht [Ric]. Für zumindest etwas realistische Untersuchungen muss ein Gaußscher Prozess höherer Ordnung als Eingangssignal fungieren, außerdem müssen bedingte Wahrscheinlichkeitsdichten bestimmt werden, um den bedingten Erwartungswert der Erregungsmuster zu berechnen. Aus den vorangegangenen Rechnungen sollte plausibel geworden sein, dass dieses Unterfangen hoffnungslos ist. Aus diesem Grund wird im Folgenden eine empirische Untersuchung der Entropie der Bandselektion sowie des Optimalprädik-

tors der Erregungsmuster auf Basis sehr großer Datenmengen vorgestellt. Das genaue Vorgehen wird im nächsten Abschnitt erläutert.

5.7 DIE EMPIRISCHE BESTIMMUNG DES OPTIMALPRÄDIKTORS UND DER ENTROPIE DER BANDSELEKTION.

Die Schätzung der Entropie der Bandselektion ist leicht durchführbar. Mit einem Signalmodell werden hinreichend viele Eingangswerte und aus diesen Erregungsmuster generiert. Aus diesen Erregungsmuster ist dann die Bandselektion ableitbar. Diese ist, wie in den Erläuterungen zum Electrocodec in Abschnitt 3.1 erwähnt, als Binärvektor $B_n = (b_n^1, \ldots, b_n^M)$, die Aktivitätskarte, mit der diskreten Zeit n, darstellbar. Zur Schätzung der Entropie werden diese Binärvektoren als Dualzahlendarstellung interpretiert und so gemäß

$$I_n = \sum_{i=0}^{M-1} b_n^i 2^i \tag{5.29}$$

auf ganze Zahlen abgebildet. Die Entropie der Bandselektion wird dann als die Entropie einer Datenquelle, welche die I_n erzeugt, verstanden und über die relativen Häufigkeiten dieser I_n geschätzt. Zwar ist grundsätzlich auch die bedingte Entropie $H(I_n|I_{n-1}, \ldots, I_{n-k})$ beliebiger Ordnung auf diese Art bestimmbar, jedoch ist oftmals bereits für die bedingte Entropie erster Ordnung eine den RAM typischer Computersysteme sprengende Datenmenge nötig. Daher wird sich auf die unbedingte Entropie beschränkt.

Für die Schätzung des Optimalprädiktors wird auf Kernel Regression zurückgegriffen. Betrachtet man die bedingte Erwartung $E(Y|X)$ einer Zufallsvariablen Y gegeben eine weitere Zufallsvariable X, so sucht man ein $m(X)$ mit

$$E(Y|X) = m(X). \tag{5.30}$$

Der sogenannte Nadaraya–Watson Kernel Regressor (NWKR) schätzt nun $m(X)$ gemäß

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^{n} y_i K_h(x - x_i)}{\sum_{i=1}^{n} K_h(x - x_i)}, \tag{5.31}$$

wobei K_h der sogenannte Kernel mit Bandbreite $h > 0$ ist. Im Rahmen der vorgelegten Arbeit wurde der Gaußkernel $K_h(x) = \frac{1}{h} e^{-\frac{x^2}{2h^2}}$ verwendet bzw. dessen multivariate Verallgemeinerung, sodass die Bandbreite h als die Standardabweichung verstanden werden kann. Die Güte der Schätzung des bedingten Erwartungswerts durch die Kernel Regression

hängt, neben der Stichprobengröße, hauptsächlich von einer sinnvollen Wahl der Bandbreite h ab [Sco92]. Im Allgemeinen ist h ein Vektor, d.h. $h = (h_1, \ldots, h_N)$. Es gibt Daumenregeln zur Wahl von h, jedoch liefern diese für reale, nicht normalverteilte Daten typischerweise stark suboptimale Ergebnisse. Deswegen ist es typisch, die optimale Bandbreite durch Kreuzvalidierung unter Auslassung jeweils eines Datenpunktes (engl. leave one out cross validation) zu bestimmen [Sco92]. Hierzu wurde im Rahmen dieser Arbeit für eine gegebene, geordnete Stichprobe $(x_1, \ldots, x_L)$ L-mal über diese Stichprobe iteriert, jeweils bei der i-ten Iteration der i-te Datenpunkt weggelassen und mit den restlichen Datenpunkten auf Basis der Bestimmungsgleichung 5.31 die Prädiktion für den i-ten Datenpunkt bestimmt. Mittels gradientenfreier Optimierung wurde dann die optimale Bandbreite bestimmt, indem der Prädiktionsgewinn der Prädiktion, definiert in Gl. 2.42, maximiert wurde. Aufgrund der Kreuzvalidierung kommt es hierbei nicht zu einer Überanpassung.

5.8 DATENGRUNDLAGE

Zur empirischen Bestimmung des Optimalprädiktors sowie der Entropie der Bandselektion werden zunächst Daten benötigt. Diese sollten zumindest einen Teilaspekt der Sprache möglichst realistisch repräsentieren, um einigermaßen sinnvolle Ergebnisse zu erhalten. Hierzu wurde auf Aufnahmen zurückgegriffen, welche im Rahmen einer anderen Forschungsarbeit aufgenommen wurden. Fünf Probanden, drei Männer und zwei Frauen, produzierten dabei wiederholt, möglichst konstant, d.h. ohne Variation der Lautstärke oder des Klangs, verschiedene Phoneme über eine Dauer von fünf bis etwa 20 Sekunden. Die berücksichtigten Phoneme waren /a:/, /e:/, /i:/, /o:/, /u:/, /n/ und /ng/ sowie die stimmlosen Phoneme /f/, /s/ und /ʃ/. Für diese Arbeit wurden lediglich die stimmlosen Phoneme verwendet, da bekannt ist, dass ein Gaußscher Prozess ein sehr gutes Signalmodell für diese darstellt. Die Aufnahmen der stimmlosen Phoneme, durchgeführt mit einer Abtastrate von 48 kHz, wurden in Segmente mit einer Länge von 10.000 Abtastwerten zerlegt, wobei Einschwingvorgänge aus den Aufnahmen zunächst entfernt wurden.

5.8.1 *Stationaritätsprüfung*

Für diese Segmente dieser Aufnahmen wurde dann die Stationarität dieser Aufnahmen geprüft. Hierzu wurden adaptive lineare Prädiktoren der Länge 1 bis 15 für die Prädiktion dieser Segmente genutzt und mit

einem optimalen, statischen linearen Prädiktor verglichen. Die Adaption der Prädiktorkoeffizienten wurde in jedem Zeitschritt vorgenommen und die Optimalkoeffizienten jeweils über die Wiener-Hopf-Gleichung 2.39 bestimmt. Die Idee hierbei ist die folgende: Für einen (schwach) stationären Prozess gilt zwingend, dass kein adaptiver linearer Prädiktor einen geringeren mittleren quadratischen Prädiktionsfehler als ein statischer Prädiktor auf Basis der wahren Autokorrelationsfunktion aufweisen kann. Je instationärer ein Prozess ist, umso mehr sollte eine Adaption der Prädiktorkoeffizienten einen Vorteil gegenüber einem statischen Prädiktor haben. Daher kann als ad-hoc Maß der (schwachen) Stationarität eines Prozesses die Differenz des mittleren quadratischen Prädiktionsfehlers, oder wahlweise des sogenannten Prädiktionsgewinns, optimaler adaptiver linearer Prädiktoren und eines optimalen statischen linearen Prädiktors dienen. Je kleiner diese Differenz, desto stationärer der Prozess. Diese Bewertung der Stationarität wurde im Rahmen dieser Arbeit auf Basis des Prädiktionsgewinns PG, definiert in Gl. 2.42, durchgeführt. Die berücksichtigten Prädiktorordnungen speisten sich dabei aus typischen Parameterzahlen des sogenannten linear predictive coding, einem klassischen Verfahren der Sprachcodierung. Hierbei werden üblicherweise 10 bis 15 Prädiktorkoeffizienten verwendet. Ein Segment wurde nun als hinreichend stationär bewertet, sofern die adaptiven linearen Prädiktoren aller betrachteten Ordnungen maximal eine Verbesserung des Prädiktionsgewinns des statischen linearen Prädiktors gleicher Ordnung um weniger als $0,2\,\mathrm{dB}$ erzielten. Dieser Wert wurde als Grenze gewählt, da er sich gut mit der subjektiven Einschätzung deckte, d.h., für Signale, die als sehr stationär angesehen wurden, war diese Verbesserung typischerweise kleiner als $0,2\,\mathrm{dB}$.

5.9 GENERIERUNG KÜNSTLICHER ERREGUNGSMUSTER

Für jedes als stationär bewertete Segment wurde ein Gaußscher Prozess 15. Ordnung definiert, dessen Parameter an das jeweilige Segment mittels der Methode der kleinsten Quadrate optimal angepasst wurden. Dank dieses Prozessmodells konnten nun beliebige Datenmengen mit einer festen Wahrscheinlichkeitsverteilung generiert werden. Dies gestattete die Generierung einer beliebigen Menge an Erregungsmustern für ein gegebenes Eingangssignal. Diese künstlichen Eingangssignale des Advanced Combination Encoders, auf Basis derer die Erregungsmuster berechnet wurden, wurden jeweils spitzennormalisiert, sodass die Spitzenwerte jeweils bei Eins lagen. Anderenfalls kommt es schnell zum Übersteuern im elektrischen Bereich. Die Spitzennormalisierung entspricht der Festsetzung der Standardabweichung der Eingangsprozesse auf in etwa

1/3, da bei dieser Standardabweichung ca. 99,7 % der Realisierungen im Intervall [-1,1] liegen.

Da die Schätzung der Entropie der Bandselektion relativ geradlinig erfolgen kann - lediglich die benötigte Datenmenge für eine gute Schätzung ist iterativ zu bestimmen - wohingegen die Schätzung des bedingten Erwartungswertes einiges an „Fingerspitzengefühl" benötigt und die Vertrauenswürdigkeit ex-ante nicht gegeben ist, wird im Folgenden auf die Validierung der genutzten Methodik zur Bestimmung des Optimalprädiktors detailliert eingegangen.

5.10 VALIDIERUNG DER SCHÄTZUNG DES BEDINGTEN ERWARTUNGSWERTS

Zur Validierung wurden Modelle stochastischer Prozesse benötigt, deren Optimalprädiktion bekannt ist. Diese werden im folgenden Unterabschnitt erläutert. Danach wird das Vorgehen der eigentlichen Validierung beschrieben.

5.10.1 *Prozessmodelle*

Erneut wurde hierbei zum einen auf Gaußsche Prozesse zurückgegriffen. Diese wurden mittels autoregressiver Prozesse realisiert, d.h., das grundlegende Prozessmodell der Ordnung P war

$$y(n) = \sum_{i=1}^{P} a_i y(n-i) + \epsilon(n) \tag{5.32}$$

mit $\epsilon(n) \sim \mathcal{N}(0,1)$ sowie $a_i \in \mathbb{R}$. Der Optimalprädiktor dieses Prozesses ist der lineare Prädiktor und der maximal erzielbare Prädiktionsgewinn PG_{max} berechnet sich, dies gilt allgemein für jedes Prozessmodell, zu

$$PG_{max} = 10 \cdot \log_{10}(\frac{\sigma_y^2}{\sigma_\epsilon^2}). \tag{5.33}$$

Es ist jedoch nicht hinreichend, nur Gaußsche Prozesse zu berücksichtigen, da es denkbar ist, dass das genutzte Schätzverfahren für den bedingten Erwartungswert in lokalen Minima stecken bleibt. Ein naheliegendes lokales Minimum wäre der lineare Prädiktor respektive dessen Prädiktionsgewinn. Daher wurde ein weiteres Prozessmodell erdacht, dessen Optimalprädiktor nichtlinear ist und daher im Folgenden zur Unterscheidung auch als nichtlineares Prozessmodell bezeichnet wird. Dieses Modell war

$$y(n) = k \cdot e^{-|y(n-1)|} + \epsilon(n) \tag{5.34}$$

mit $\epsilon(n)$ wie zuvor sowie $k \in \{1, 5, 10\}$. Grundgedanke war, einen möglichst einfachen stabilen Prozess mit nichtlinearem Optimalprädiktor zu definieren. Die Exponentialfunktion mit negativem Exponenten liefert eine Möglichkeit, große Werte auf bequeme Weise „abzudämpfen". Auch mit linearen Prädiktoren unendlicher Länge kann der Optimalprädiktor dieses Prozesses nicht beliebig genau approximiert werden. Der Optimalprädiktor $\hat{y}(n)$ ist naheliegenderweise $\hat{y}(n) = k \cdot e^{-|y(n-1)|}$. Die genutzten Werte von k führen zu Realisierungen mit deutlich unterschiedlicher Datenpunktdichte für beliebige Intervalle durch deutlich veränderte Streubreite sowie zu einem deutlichen Unterschied hinsichtlich des maximal erzielbaren Prädiktionsgewinns.

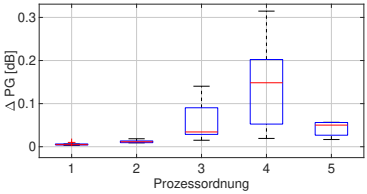
5.10.2 *Validierung*

Die eigentliche Validierung der Schätzung des bedingten Erwartungswertes mittels des Nadaraya-Watson Kernel Regressors erfolgte nun wie folgt:

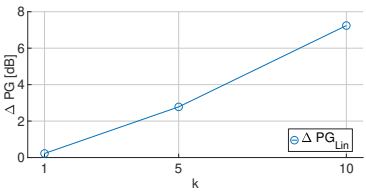
Für die Gaußschen Prozesse und Modellordnungen $N = 1, \ldots, 5$ wurden jeweils fünf mal N zufällig gewählte Modellpole $p_i \in (-1, 1)$ bestimmt, aus welchen die Modellparameter $a_1, \ldots, a_N$ berechnet wurden. Die Parameterberechnung erfolgte dabei über den z-Bereich durch die Gleichung

$$\prod_{i=1}^{N}(1 - p_i z^{-1}) = 1 - \sum_{i=1}^{N} a_i z^{-i}. \tag{5.35}$$

Dieses indirekte Vorgehen garantiert die Stabilität des resultierenden Gaußschen Prozesses. Gegeben eins dieser zufällig bestimmten Modelle, so wurden 20.000 Datenpunkte generiert, auf deren Basis dann die Bandbreiten $h_1, \ldots, h_N$ des Nadaraya-Watson Kernel Regressors mittels Kreuzvalidierung, wie in Abschnitt 5.7 beschrieben, durch Optimierung bestimmt wurden. Dieser Ansatz impliziert, dass die Ordnung des Ansatzes des bedingten Erwartungswertes immer der tatsächlichen Modellordnung entspricht. Als Optimierungskriterium wurde der erzielte Prädiktionsgewinn des geschätzten bedingten Erwartungswertprädiktors genutzt. Als Optimierungsalgorithmus wurde ein Innere-Punkte-Verfahren verwendet. Als Initialisierung der Bandbreiten wurde $h_i = 0{,}01$ genutzt, was sich in Vorarbeiten für eine recht große Zahl an Testfällen als sinnvollen Startpunkt erwiesen hatte. Der Suchbereich wurde durch eine Ober- und Untergrenze eingeschränkt. Als Obergrenze h_i^o wurde $h_i^o = \frac{max y}{4}$ gewählt, wobei y das betrachtete Signal ist. Als Untergrenze h_i^u wurde $h_i^u = 0{,}0001$ gewählt. Beide haben sich empirisch für eine große Zahl an Testfällen als sinnvoll erwiesen. Die Optimierung wurde zehnmal



(a) ΔPG für die autoregressiven Prozesse



(b) ΔPG für den nichtlinearen Prozess

Abbildung 5.4: Differenz des Prädiktionsgewinns ΔPG des optimalen Prädiktors und (a) dem Nadaraya-Watson Kernel Regressor für autoregressive Prozesse sowie (b) für den nichtlinearen Prozess und den linearen Prädiktor. Für den nichtlinearen Prozess und den Nadaraya-Watson Kernel Regressors sind die korrespondierenden Ergebnisse in Abb. 5.5 dargestellt.

mit zufälliger Initialisierung, außer beim ersten Mal, wiederholt und die Bandbreiten, die zum höchsten Prädiktionsgewinn geführt haben, ausgewählt. Für diese wurde dann der Prädiktionsgewinn berechnet. Identisch wurde für das nichtlineare Prozessmodell vorgegangen. Hierbei wurde für jeden Wert des Parameters a nur eine Realisierung betrachtet. Grund ist, dass eine wesentliche Abhängigkeit der Ergebnisse von der Realisierung schon beim autoregressiven Prozess beobachtbar sein sollte und daher nicht erneut untersucht werden musste.

Mittels des nichtlinearen Prozessmodells wurde zusätzlich der Fall untersucht, bei dem der Optimalprädiktor eine andere Ordnung hat als die Ordnung des Nadaraya-Watson Kernel Regressors. Im Allgemeinen ist die Ordnung nicht bekannt, daher wäre es fatal, wenn sich sinnvolle Ergebnisse, d.h. solche nahe am Optimum, nur bei Übereinstimmung dieser beiden Prädiktorordnungen ergeben würden. Für diese Untersuchung wurde ein Nadaraya-Watson Kernel Regressor der Ordnung drei verwendet und für $k = 1$, $k = 5$ sowie $k = 10$, wie zuvor, jeweils eine Realisierung generiert und der bedingte Erwartungswert geschätzt.

Die Ergebnisse dieser Validierung werden nachfolgend vorgestellt und anschließend wird zu den Ergebnissen der Anwendung auf die Erregungsmuster übergegangen.

5.10.3 *Ergebnisse der Validierung*

Abb. 5.4a zeigt die Differenz des Prädiktionsgewinns $\Delta PG := PG_{Opt} - PG_{NWKR}$ des Prädiktionsgewinns des optimalen Prädiktors PG_{Opt} und des Prädiktionsgewinns des Nadaraya-Watson Kernel Regressors PG_{NWKR} für die autogressiven Prozesse. Abb. 5.4b zeigt die Differenz

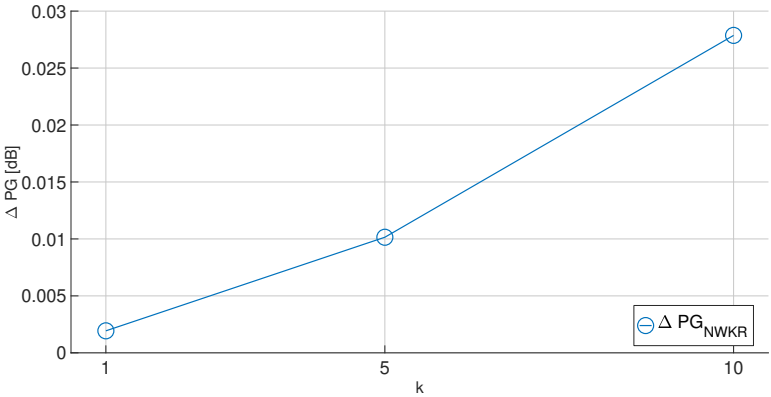


Abbildung 5.5: Differenz der Prädiktionsgewinne ΔPG des optimalen Prädiktionsgewinns und des Prädiktionsgewinns des Nadaraya-Watson Kernel Regressors für den nichtlineare Prozess gemäß Gl. 5.32 in Abhängigkeit vom Prozessparameter k.

des Prädiktionsgewinns des optimalen Prädiktors und des Prädiktionsgewinns des linearen Prädiktors für den nichtlinearen Prozess. Aus Gründen der Übersichtlichkeit zeigt Abb. 5.5 das entsprechende Ergebnis für den Nadaraya-Watson Kernel Regressor.

Zur besseren Interpretation der Ergebnisse aus Abb. 5.4a sind die optimalen Prädiktionsgewinne für die jeweilige Prozessordnung in Tabelle 5.1 tabuliert. In jedem Fall erzielte der Nadaraya-Watson Kernel Regressor Prädiktionsgewinne sehr nahe dem Optimum. Es ist eine leichte Zunahme der Differenz zum optimalen Prädiktionsgewinns mit zunehmender Ordnung von Prozess und Kernel Regressor zu beobachten. Generell ist eine Abnahme der Güte der Schätzungen mit zunehmender Ordnung des Kernel Regressors zu erwarten, da exponentiell mehr Daten für eine gute Schätzung notwendig werden. Darüber hinaus ist die scheinbar etwas schlechtere Leistung für eine Modellordnung von vier und etwas weniger ausgeprägt für eine Modellordnung von drei, zum einen durch eine, ob der Polauswahl, zufällig deutlich höhere Varianz des zu prädizierenden Signals y zu erklären, womit ein erhöhter maximaler Prädiktionsgewinn einhergeht. Eine höhere Varianz bei fester Stichprobengröße führt zu einer geringeren Wertedichte (Datenpunkte je Einheitsintervall) und hierdurch zu einer schlechteren Schätzerleistung. Des Weiteren kann der Zufall, der in der Optimierung der Bandbreiten über die Initialisierung eine Rolle spielt, einen kleineren Teil der Unterschiede erklären.

Tabelle 5.1: Maximal möglicher, also optimaler, Prädiktionsgewinn für jede Modellordnung N und jede Realisierung der (a) autoregressiven sowie (b) nichtlinearen Prozesse. Per Zufall war der maximale Prädiktionsgewinn der autoregressiven Prozesse fünfter Ordnung, auf Grund der jeweiligen, zufällig bestimmten Pole, vergleichsweise niedrig.

(a) Optimaler Prädiktionsgewinn der autoregressiven Prozesse

N	Opt. Prädiktionsgewinn [dB]				
1	2,93	1,34	5,5	2,24	0,54
2	2,61	0,13	1,26	0,68	7,39
3	0,58	2,11	10,73	9,0	4,06
4	21,64	8,56	3,14	8,65	6,47
5	2,82	1,52	2,57	0,99	1,88

(b) Optimaler Prädiktionsgewinn des nichtlinearen Prozesses

k	Opt. Prädiktionsgewinn [dB]
1	0,27
5	4,84
10	9,89

In jedem Fall erzielte der Nadaraya-Watson Kernel Regressor in sehr guter Näherung den Prädiktionsgewinn des Optimalprädiktors. Des Weiteren kam es niemals zu einer Überanpassung, welche aufgrund der Kreuzvalidierung unmöglich sein sollte.

Gleiches ergab sich für den Nadaraya-Watson Kernel Regressor der Ordnung drei, der für dasselbe nichtlineare Prozessmodell zur Prädiktion genutzt wurde. Die wahre Prozessordnung war eins, sodass nun die Prädiktorordnung nicht zur Prozessordnung passte. Dennoch war die Differenz zwischen dem optimalen Prädiktionsgewinn und dem Prädiktionsgewinn des Kernel Regressors nahezu identisch mit z.B. einem Prädiktionsgewinn von $9,92\,\mathrm{dB}$ des Kernel Regressors sowie einem Prädiktionsgewinn von $9,95\,\mathrm{dB}$ für den Optimalprädiktor für $k = 10$.

Aufgrund des Wechselspiels von Datenmenge, Prädiktorordnung und Güte der erzielbaren Schätzung, die im Rahmen der bisher vorgestellten Validierung wie erwartet beobachtet worden ist, wurde schlussendlich, zur Abschätzung der nötigen Datenmenge für ein mit hoher Wahrscheinlichkeit nahezu optimales Schätzergebnis, mit unterschiedlich großen Datenmengen der Optimalprädiktor mittels des Nadaraya-Watson Kernel Regressors für die Erregungsmuster geschätzt. Hierbei wurde lediglich weißes, Gaußsches Rauschen als Eingangssignal verwendet. Es wurden 20.000, 40.000 sowie, für den Fall einer Prädiktorordnung von fünf, auch 100.000 Ausgangswerte der Lautheitswachstumsfunktion für jedes Subband betrachtet. Abb. 5.6 zeigt für das Band 22 die Differenz der Prädiktionsgewinne des optimalen linearen Prädiktors und des Nadaraya-Watson Kernel Regressors. Qualitativ identische Ergebnisse wurden für die anderen Bänder erzielt. Für Prädiktorordnungen $N = 1$ und $N = 2$ sind bereits nahezu perfekte Ergebnisse mit einer Stichprobengröße von 20.000 erzielbar, was an minimalen Verbesserungen in Folge der Stich-

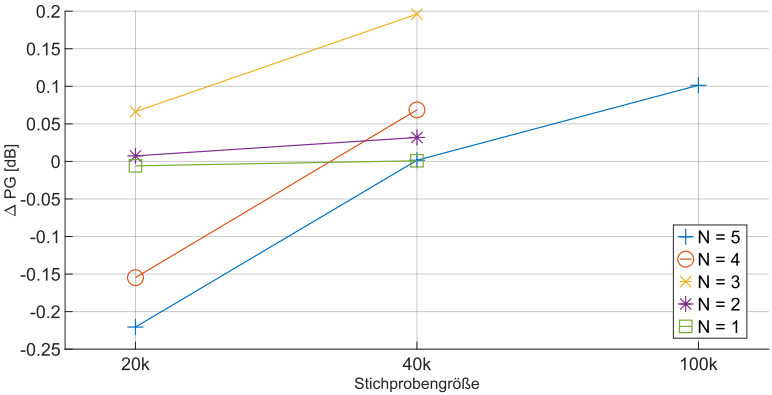


Abbildung 5.6: Differenz der Prädiktionsgewinne ΔPG des Nadaraya-Watson Kernel Regressors und des optimalen linearen Prädiktors auf Erregungsmustern unter Verwendung einer Stichprobengröße von 20.000, 40.000 und 100.000. Die Erregungsmuster wurden aus weißem, gaußschem Rauschen berechnet. Ein positiver Wert zeigt eine Überlegenheit des Kernel Regressors an. Nur für $N = 5$, dem kritischsten Fall, wurde eine Stichprobengröße von 100.000 evaluiert. Für $N = 1$ und $N = 2$ sind bereits 20.000 Datenpunkte hinreichend.

probenvergrößerung zu erkennen ist. Alle anderen Prädiktorordnungen können eine gewisse Verbesserung durch Erhöhung der Datenmenge erzielen. Auf Basis dieser Resultate wurde für die eigentliche Untersuchung des Optimalprädiktors der Erregungsmuster eine Stichprobengröße von 100.000 verwendet. Des Weiteren wurde maximal eine Prädiktorordnung von vier untersucht. Zumindest bis zu dieser sollten sich so in den meisten Fällen sehr gute Ergebnisse ergeben.

5.11 DER OPTIMALPRÄDIKTOR DER ERREGUNGSMUSTER FÜR STIMMLOSE SPRACHE

Da die Optimierung mit 100.000 Datenpunkten und für eine größere Anzahl an Bändern des Advanced Combination Encoders sowie eine größere Anzahl an Eingangsprozessen, aus welchen dann die zugehörigen Erregungsmuster abgeleitet werden, extrem lange dauert, wurde sich auf eine kleine Auswahl an Eingangsprozessen beschränkt. Für jedes der Phoneme /f/, /s/ und /ʃ/ wurden jeweils sechs autoregressive Prozesse auf Basis von jeweils separaten Sprecheraufnahmen generiert und mit diesen jeweils Erregungsmuster generiert. Bei der Auswahl der Prozesse wurde

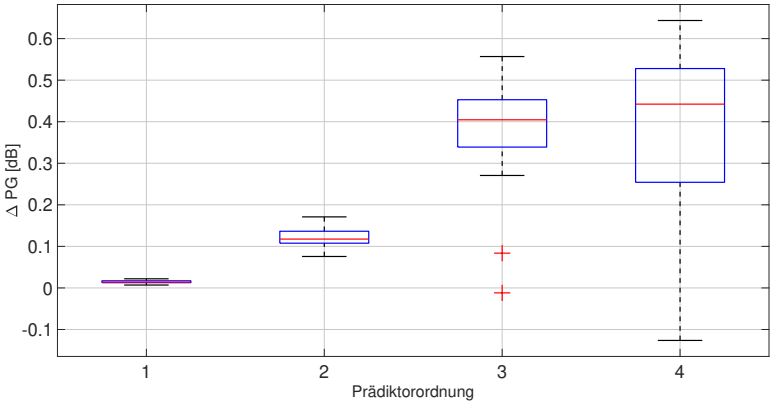


Abbildung 5.7: Differenz der Prädiktionsgewinne ΔPG des Prädiktionsgewinns (PG) des Nadaraya-Watson Kernel Regressors und des optimalen linearen Prädiktors über alle Bänder in Abhängigkeit von der Prädiktorordnung. Durch zufällige ungünstige Initialisierung kommt es teilweise zu suboptimalen Ergebnissen.

auf eine gute Anpassung, gemessen durch den mittleren quadratischen Fehler, an die jeweiligen Sprachaufnahmen geachtet. Ferner wurde darauf geachtet für jedes Phonem mindestens einen Mann und eine Frau als Sprecher zu berücksichtigen. Des Weiteren wurden zur weiteren Reduktion der Rechenzeit die Optimierungen zur Bestimmung der optimalen Bandbreiten des Nadaraya-Watson Kernel Regressors lediglich viermal durchgeführt. Ferner ist anzumerken, dass, da stimmlose Phoneme recht hochfrequent sind, vornehmlich die Bänder 16 bis 22, welche in etwa den Frequenzbereich von 3000 Hz bis 8000 Hz abdecken (vergleiche Tabelle 2.1), oberhalb des Basisniveaus s_{base} liegen. Die Bänder, welche zu niedrigeren Frequenzen gehören, wechseln häufig zwischen Werten oberhalb und unterhalb des Basisniveaus, sodass mitunter größere Lücken in den Erregungsmustern existieren, da Werte unterhalb des Basisniveaus auf den Platzhalterwert -10^{-10} abgebildet werden, siehe dazu Gl. 2.6. Hierdurch wird die Untersuchung stark erschwert, weshalb sich auf die Bänder 16 bis 22 fokussiert wurde, die in den meisten Fällen durchgängig selektiert waren. Die Alternative, das Basisniveau abzusenken, erschien wenig zielführend, da nicht klar ist, wie die dann veränderte Lautheitswachstumfunktion die statistischen Abhängigkeiten im Vergleich zur gängigen Standardeinstellung verändert. Abb. 5.7 stellt die Differenz der Prädiktionsgewinne ΔPG des jeweiligen Nadaraya-Watson Kernel Regressors und des jeweils optimalen linearen Prädiktors für alle berück-

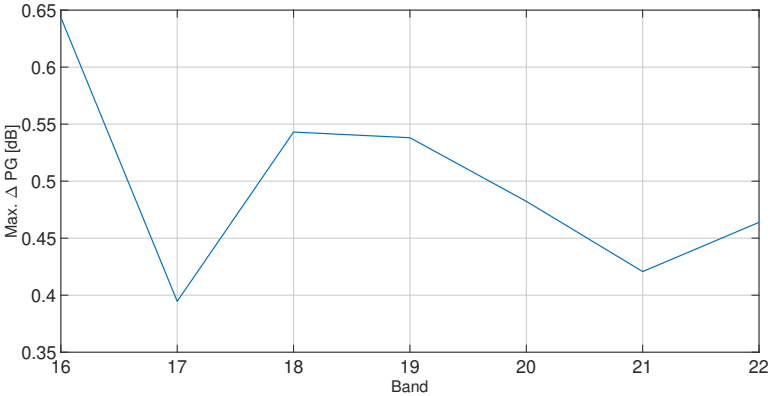
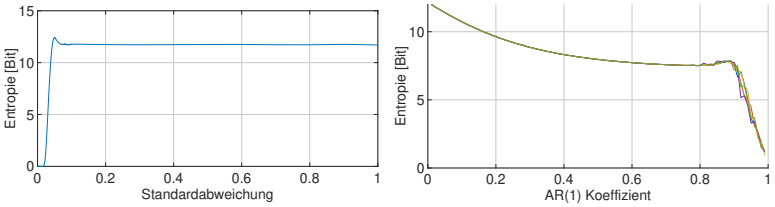


Abbildung 5.8: Maximum der Prädiktionsgewinne ΔPG des Nadaraya-Watson Kernel Regressors und des optimalen linearen Prädiktors pro Band für die Bänder 16 bis 22 der Differenz. Es ist eine leichte Tendenz in Richtung höherer Nichtlinearität für niedrigere Bänder zu erkennen.

sichtigten Bänder in Abhängigkeit der Prädiktorordnung dar. Zwar ist in der Tat eine Verbesserung der linearen Prädiktion möglich, jedoch betrug selbst die höchste gefundene Differenz lediglich $0,64$ dB, d.h., der Prädiktionsgewinn des Nadaraya-Watson Kernel Regressors lag maximal $0,64$ dB über dem Prädiktionsgewinn des zugehörigen optimalen linearen Prädiktors. Diese Verbesserung gegenüber der linearen Prädiktion erscheint für Codierungszwecke vernachlässigbar, da eine typische Abschätzung von einer Reduktion um etwa 1 Bit je 6 dB Prädiktionsgewinn ausgeht. Bei weniger als 1 dB Verbesserung ist keine relevante Reduktion der Datenrate durch Prädiktoroptimierung für stimmlose Phoneme zu erwarten. Abb. 5.8 stellt das Maximum der Differenz der Prädiktionsgewinne ΔPG des jeweiligen Nadaraya-Watson Kernel Regressors und des jeweils optimalen linearen Prädiktors in Abhängigkeit vom betrachteten Band dar. Es ist eine leichte Tendenz in Richtung stärkerer Nichtlinearitäten für untere Bänder zu erkennen. Aufgrund der oben beschriebenen Lücken in den Erregungsmustern für niederfrequente Bänder konnte jedoch nicht untersucht werden, ob sich dieser Trend fortsetzt.

5.12 DIE ENTROPIE DER BANDSELEKTION

Um eine erste Vorstellung der Entropie der Bandselektion zu erhalten wurde zunächst erneut weißes Gaußsches Rauschen als Eingangssignal



(a) Entropie der Bandselektion in Abhängigkeit der Standardabweichung des Eingangsrauschens.

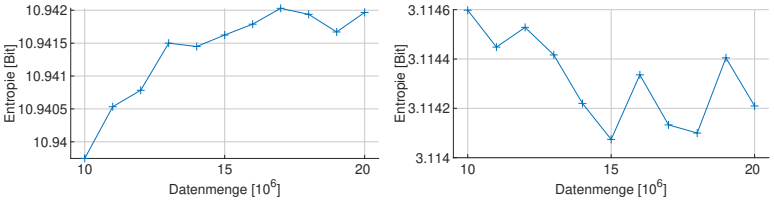
(b) Entropie der Bandselektion in Abhängigkeit des Koeffizienten eines autoregressiven Prozesses der Ordnung 1 (AR(1)).

Abbildung 5.9: Entropie der Bandselektion des Advanced Combination Encoders für (a) weißes Gaußsches Rauschen als Eingangssignal mit variabler Standardabweichung sowie (b) einen autoregressiven Prozess der Ordnung eins als Eingangsignal. Auf Grund der beobachteten Varianz für größere Werte des Prozesskoeffizienten wurden jeweils fünf Wiederholungen, jeweils mit einer neuen Realisierung, durchgeführt.

verwendet aus dem dann die Erregungsmuster gewonnen wurden. Hierbei wurden unterschiedliche Rauschvarianzen bzw. -standardabweichungen genutzt, um die Abhängigkeit von der Eingangsleistung abzuschätzen. Dazu wurden zahlreiche Realisierungen mit Rauschstandardabweichungen zwischen 0 und 1 erzeugt und jeweils 10^7 Erregungsmusterwerte generiert, auf deren Basis dann die Entropie der Bandselektion bestimmt wurde. Des Weiteren wurde ein autoregressiver Prozess $y(n)$ der Ordnung eins (AR(1)) verwendet, um für diesen einfachen Fall, bei dem im Gegensatz zu vorher durch die Wahl des Prozesskoeffizienten ein Ungleichgewicht im Leistungsdichtespektrum erzielt werden kann, die Auswirkung von dominanten Frequenzen zu untersuchen. Das Prozessmodell war hierbei entsprechend

$$y(n) = -ay(n-1) + \epsilon(n) \tag{5.36}$$

mit $\epsilon(n) \sim \mathcal{N}(0,1)$ sowie $a \in [0,01; 0,99]$, wobei a in 0,01 Schritten erhöht wurde. Da für größere Werte von a eine nicht vernachlässigbare Varianz der Entropieschätzung beobachtet wurde, wurden in diesem Falle für jeden Wert von a jeweils fünf Realisierungen erzeugt, aus diesen Erregungsmuster berechnet und für diese dann die Entropie der Bandselektion bestimmt. Die Ergebnisse sind in Abb. 5.9 dargestellt. Für den Fall des weißen, Gaußschen Rauschens wie in Abb. 5.9a ist gut die Sensitivität des Advanced Combination Encoders zu beobachten. Ab einer Standardabweichung von ca. 0,015 hat das Rauschen eine hinreichende Leistung, um zu ersten Erregungen zu führen. Innerhalb eines sehr schmalen Bereichs steigt die Entropie der Bandselektion von 0 auf den Spitzenwert



(a) Entropie der Bandselektion für ein /f/ Phonem.

(b) Entropie der Bandselektion für ein /s/ Phonem.

Abbildung 5.10: Geschätzte Entropie der Bandselektion in Abhängigkeit von der verwendeten Datenmenge beispielhaft für ein /f/ und ein /s/ Phonem. Zwar steigt die Entropieschätzung teilweise auch bei der Verwendung von 10^7 Werten, jedoch so minimal das eine weitere Erhöhung der Datenmenge unnötig erschien.

von 12,44 Bit, den es bei einer Standardabweichung von nahezu 0,05 annimmt. Danach sinkt die Entropie leicht und pendelt sich um den Wert von 11,72 Bit ein. Die Entropie der Bandselektion für den Fall des AR(1) Eingangsprozesses des Advanced Combination Encoders in Abhängigkeit vom Parameter a aus Gl. 5.36 nimmt mit steigenden Werten von a zunächst ab und erreicht ein lokales Minimum von etwa 7,55 Bit für $a = 0,8$. Anschließend steigt die Entropie kurz auf 7,85 Bit für $a = 0,89$ und sinkt danach rapide bis auf 1,29 Bit für $a = 0,99$. Das Verhalten für Werte von a nahe von 1 ist plausibel, da dann hohe Frequenzen mehr und mehr verstärkt werden, sodass der AR(1) Prozess als Filterung von weißem Rauschen mit einem sehr schmalbandigen Hochpass interpretiert werden kann. Im Extremfall sind dann nur noch die beiden höchsten Subbänder des Advanced Combination Encoders selektiert. Um die Güte der Schätzung der Entropie der Bandselektion stimmloser Phoneme zu beurteilen, wurden mittels der in Abschnitt 5.8 und folgenden beschriebenen Gaußschen Prozesse 10^7 bis $2 \cdot 10^7$ Erregungsmusterwerte erzeugt, und jeweils die Entropie der Bandselektion geschätzt. Abb. 5.10 zeigt diese Schätzung auf Basis zweier Beispielprozesse, welche zum Phoneme /f/ sowie /s/ gehören. Zwar ist selbst bei diesen Datenmengen noch ein minimaler Zuwachs der Entropieschätzung zu beobachten, jedoch fällt dieser so gering aus, dass im Weiteren mit einer Stichprobengröße von 10^7 gearbeitet wurde.

Für die drei betrachteten Phoneme sind geschätzte Entropien der Bandselektion in Abb. 5.11 dargestellt. Insgesamt wurden 18 Realisierungen, sechs je Phoneme, berücksichtigt. Es ist ein klarer Unterschied im Informationsgehalt für die Phoneme zu erkennen. /s/ und /sh/ mit im Median zwischen 3,99 Bit und 5,33 Bit liegen bei Werten, die der Entropie

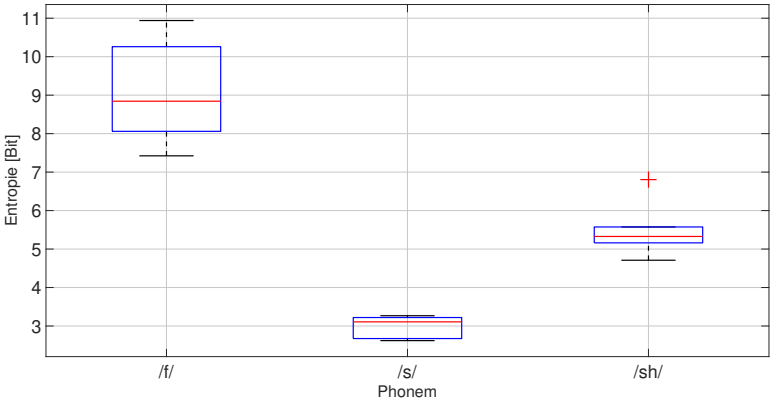


Abbildung 5.11: Verteilung der geschätzten Entropien der Bandselektion für die Phoneme /f/, /s/ sowie /ʃ/ (in der Abb. als /sh/ bezeichnet). Es ist ein deutlicher Unterschied zwischen den hochfrequenten Phonemen /s/ und /ʃ/ sowie dem niederfrequenteren /f/ zu erkennen.

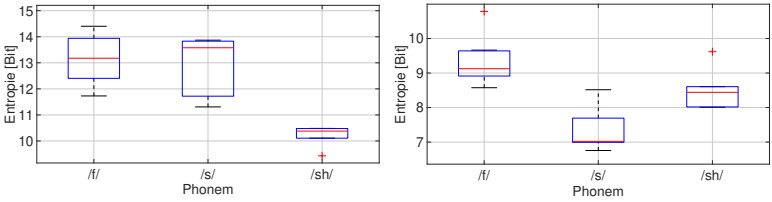
der Bandselektion des AR(1) Prozesses für große Werte von a ähneln. Dies ist sinnvoll, da der recht scharfe Charakter dieser zwei Phoneme auf die hohe Energie in den höheren Frequenzen zurückzuführen ist. Dahingegen ist das /f/ Phoneme deutlich weniger höhenlastig und hat ein etwas ausgeglicheneres Leistungsdichtespektrum. Hierdurch ist die deutlich höhere Medianentropie von 8,84 Bit zu erklären. Ein Unterschied zwischen Männern und Frauen wurde hierbei nicht festgestellt.

Um nun zu untersuchen, wie weit die Kompression der Bandselektion für diese Fälle von dem Ideal der Entropie entfernt ist, wurden dieselben Erregungsmuster mit dem Electrocodec codiert, und die mittleren Codewortlängen, welche bei der Kompression der Bandselektion, dargestellt über die Aktivitätskarte (vergleiche Abschnitt 3.1), anfallen, bestimmt. Zur Verfikation wurden zudem die Erregungsmuster, die aus den realen Aufnahmen berechnet wurden, aus denen die Gaußschen Prozessen extrahiert worden sind, mit dem Electrocodec codiert, und ebenfalls die mittleren Codewortlängen, die bei der Kompression dieser Bandselektionen anfielen, bestimmt. Die Medianergebnisse sind für die drei berücksichtigten Phoneme in Tabelle 5.2 zu entnehmen. $\hat{H}$ notiert die Schätzung der Entropie, $\overline{n}_{T,\mathrm{Electrocodec}}$ ist die mittlere Codewortlänge des Electrocodecs wie eben beschrieben für die künstlich erzeugten Erregungsmuster und $\overline{n}_{R,\mathrm{Electrocodec}}$ ist die mittlere Codewortlänge, welche sich bei der Codierung der aus den realen Aufnahmen abgeleiteten Erregungsmuster ergibt. Es zeigt sich eine sehr große Redundanz von

Tabelle 5.2: Median der geschätzten Entropie $\hat{H}$ der Bandselektionen für die Phoneme /f/, /s/ sowie /ʃ/. Dazu sind die mittleren Wortlängen des vom Electrocodec erzeugten Codes für die Aktivitätskarten eben dieser Bandselektionen, einmal auf Basis dieser synthetischen Daten ($\overline{n}_{T,Electrocodec}$), abgeleitet aus den autoregressiven Prozessen und einmal, zur Verifikation, auf Basis der ursprünglichen, realen Sprachaufnahmen, aus denen unmittelbar die Erregungsmuster und daraus die Aktivitätskarte bestimmt wurden ($\overline{n}_{R,Electrocodec}$). Alle Zahlenwerte sind in Bit.

Phonem	$\hat{H}$	$\overline{n}_{T,Electrocodec}$	$\overline{n}_{R,Electrocodec}$
/f/	8,84	18,46	18,69
/s/	3,99	15,49	15,42
/ʃ/	5,33	14,41	14,38

9 - 12 Bit. Ferner sind die mittleren Codewortlängen $\overline{n}_{T,Electrocodec}$ und $\overline{n}_{R,Electrocodec}$ nahezu identisch, was die Modellierung der Phoneme durch Gaußsche Prozesse zusätzlich sinnvoll erscheinen lässt. Zwar ist die Größe der Redundanz in der Tat erstaunlich, aber, aufgrund des begründeten Entwurfs der Entropiecodierung der Bandselektion wie in Abschnitt 3.1 beschrieben, auch nicht ganz überraschend. Im Falle der betrachteten stimmlosen Phoneme sind sehr wesentlich höherfrequente Bänder des Advanced Combination Encoders beteiligt. Dadurch lässt sich die N aus M Eigenschaft, welche im Normalfall bereits eine sehr wesentliche Reduktion der Datenrate gestattet, kaum gewinnbringend ausnutzen, da oftmals die Aktivitätskarte bis zu den letzten Bändern codiert werden muss, bis das N-te Band selektiert wurde und die Codierung beendet werden kann. Des Weiteren ist etwa für das /f/ Phonem eine Selektion von bis zu drei niederfrequenten Bändern, typischerweise direkt die ersten drei, typisch mit einer darauf folgenden „Lücke“ in der Bandselektion, d.h. die nächsten Bänder sind nicht selektiert. Erst höherfrequente Bänder werden dann wieder selektiert. Dies spielt dem erlernten Wahrscheinlichkeitsmodell des Electrocodecs zuwider, welches größere Lücken in der Bandselektion als unwahrscheinlich erachtet. Tiefe Ursache hierfür sind die Gesamtanteile von stimmloser und stimmhafter Phoneme an der Sprache. Stimmhafte Anteile überwiegen deutlich und stimmlose Anteile sind zudem deutlich kürzer. Dies erklärt auch, wieso trotz dieser hohen Redundanz die erzielte Bitrate des Electrocodecs nicht wesentlich höher ist, zumindest für rauschfreie Signale. Die hohe Redundanz für die untersuchten stimmlosen Phoneme dürfte zu einer Erhöhung der Spitzenbitrate führen, aber nur unwesentlich zu einer Erhöhung der mittleren Bitrate. Dies passt gut zu den ersten Ergebnissen



(a) 0 dB (b) 10 dB

Abbildung 5.12: Geschätzte Entropie der Bandselektion für die Phoneme /f/, /s/ sowie /ʃ/ bei einem Signal-Rausch-Verhältnis von (a) 0 dB und (b) 10 dB. Es wurde CCITT-Rauschen verwendet (siehe Abschnitt 3.5). Es ist eine deutliche Erhöhung im Vergleich zum rauschfreien Fall zu erkennen.

die in Abb. 4.1a dargestellt wurden. Die Spitzenbitrate lag etwa 20 kbit/s über der mittleren Bitrate.

Eine Schlussfolgerung aus der Untersuchung ist die Möglichkeit der Reduktion zumindest der Spitzenbitrate des Electrocodec, etwa durch Umschalten des Wahrscheinlichkeitsmodells im Falle von stimmlosen Segmenten oder durch Anpassung der Codierungsrichtung der Bandselektion. Der Electrocodec codiert diese immer von Band 1 aufwärts bis zu Band 22 bzw. bis N selektierte Bänder codiert wurden. Eine Anpassung der Reihenfolge der Codierung an die zuletzt beobachtete Bandselektion, d.h. man startet mit den im letzten Zeitschritt selektierten Bändern, könnte die Codierung verbessern. Auf Grund der N aus M Eigenschaft kann man so bei adaptiver, sinnvoller Wahl der Codierungsreihenfolge die Kompression der Bandselektion für die betrachteten stimmlosen Phoneme oftmals, mit Ausnahme des /s/ Phonemes, bei dem nur wenige Bänder selektiert sind, deutlich verbessern.

Jedoch geht der betrachtete Fall von rauschfreien Signalen aus. Abb. 5.12 zeigt dieselbe Untersuchung, diesmal jedoch mit CCITT-Hintergrundrauschen. Es kommt im Vergleich zum rauschfreien Fall, also Abb. 5.11, zu einer erheblichen Steigerung der Entropie der Bandselektion auf durchweg über 10 Bit bei 0 dB und immerhin mehr als 6,5 Bit für 10 dB. Ähnliche Werte ergeben sich für andere Arten von Rauschen und eine Schätzung der Entropie der Bandselektion für den gesamten TIMIT-Testdatensatz, der in dieser Arbeit verwendet wurde, liefert eine Entropie von 10,54 Bit. Diese Zahl ist etwas mit Vorsicht zu genießen, da der gesamte Datensatz sicher eine zeitabhängige Quelle darstellt, jedoch passt die Größenordnung gut mit den Werten für 0 dB und 10 dB zusammen. Vergleicht man nun diese Entropien mit der mittleren Codewortlänge des besten Rückkopplungsautoencoders, welche bei 4,67 kbit/s und einer

Stimulationsrate von 900 Pulsen pro Sekunde in etwa 5,2 Bit beträgt, so ist klar, dass eine Verbesserung der verlustlosen Kompression der Bandselektion des Electrocodecs nicht zu einer vergleichbaren Bitrate wie jene der Autoencoder führen wird. Allein für die Kompression der Bandselektion wären Bitraten notwendig, die jene der besten Autoencoder deutlich übersteigt. Die Autoencoder codieren bei diesen Datenraten jedoch bereits die gesamte Information der Erregungsmuster.

Diese Ergebnisse machen deutlich, dass zwar noch etwas Luft für Verbesserung der Bitrate des Electrocodecs existiert, speziell bei der Spitzenbitrate, jedoch kann ohne signifikante Änderung der Codierungsstrategie keine den Autoencodern ebenbürtige Codierungsleistung erzielt werden. Insbesondere erscheint eine Verzerrung der Bandselektion, die zumindest hin und wieder bei den Autoencodern auftritt, unerlässlich.

6

DISKUSSION

Im Rahmen der vorgelegten Arbeit wurden Codierungsverfahren für die Erregungsmuster von Cochlea-Implantaten entwickelt und untersucht. Es wurden dabei zwei Ansätze verfolgt: Zuerst wurde ein konventioneller Kompressionsalgorithmus auf Basis von Differential Puls-Code Modulation (DPCM) und arithmetischer Codierung entworfen, der sogenannte Electrocodec, und in Hörtests mit Trägern von Cochlea-Implantaten untersucht. Anschließend wurde versucht, mittels künstlicher neuronaler Netze Codecs zu entwickeln respektive zu lernen, welche die Bitrate im Vergleich zum Electrocodec senken können bei gleicher oder näherungsweise gleicher Sprachverständlichkeit. Hierbei wurde zum einen ein verlustloser Codec entworfen und zum anderen zwei verlustbehaftete Codecs auf Basis von Autoencodern. Tabelle 6.1 zeigt eine Zusammenfassung der Codierungsleistung der entwickelten Kompressionsalgorithmen, erzielt auf dem TIMIT-Testdatensatz, dessen Erzeugung und Zusammensetzung in Abschnitt 3.5 beschrieben wurde. ΔVSTOI ist die Differenz zwischen den mittleren VSTOI-Werten der Erregungsmuster ohne Codierung (Referenz) und den mittleren VSTOI-Werten der Erregungsmuster nach Codierung. Für den Electrocodec wurde die Codierungsleistung für zwei verschiedene Bitraten angegeben, da in den Hörtests für die 2-Bit-Einstellung nur knapp kein signifikanter Unterschied zur Referenz festgestellt wurde. Der Electrocodec erzielt eine minimale Bitrate von 20,1 kbit/s bei einem mittleren ΔVSTOI-Wert von $-0,003$ (2 Bit) bzw. eine Bitrate von 24,3 kbit/s bei einem mittleren ΔVSTOI-Wert von $-0,001$ (3 Bit). Der verlustlose ZDLCC Codec erreicht eine Bitrate von 28,6 kbit/s,

Tabelle 6.1: Übersicht der Codierungsleistung der im Rahmen der Arbeit entwickelten Codecs. Alle Bitraten sind mittlere Bitraten.

Codec	Electrocodec (2 Bit)	Electrocodec (3 Bit)	ZDLCC	Autoencoder	Rückkopplungsautoencoder
Bitrate	20,1 kbit/s	24,3 kbit/s	28,6 kbit/s	10,98 kbit/s	4,67 kbit/s
ΔVSTOI	-0,003	-0,001	0	-0,001	-0,002
Latenz	0	0	0	0	0

der Autoencoder ohne Rückkopplung erzielt, mit einer Codebuchgröße von 14 Bit, eine Bitrate von 10,98 kbit/s bei einem ΔVSTOI-Wert von -0,001 und der Rückkopplungsautoencoder erzielt das beste Ergebnis mit einer Bitrate von 4,67 kbit/s und einem ΔVSTOI-Wert von -0,002.

Mit Hilfe der Shannon-Hartley-Formel [Sha49] für die Kapazität C in Bit/Sekunde eines Kanals

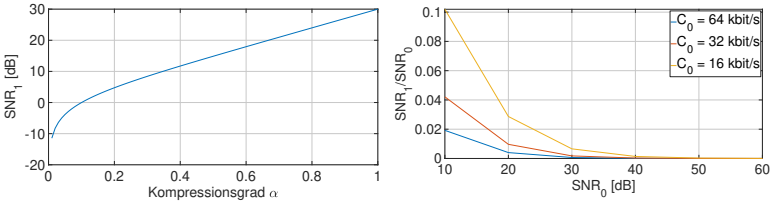
$$C = B \cdot \log_2(1 + \frac{P}{N}) \tag{6.1}$$

mit der Bandbreite B, der Signalleistung P und der Rauschleistung N eines additiven, weißen, Gaußschen Störrauschens lässt sich ein Wert für die theoretische Energie- bzw. Leistungsersparnis infolge einer verbesserten Kompression der Erregungsmuster bzw. von Audiosignalen im Kontext der drahtlosen Übertragung für Cochlea-Implantate bestimmen. Die Shannon-Hartley-Formel gibt die maximale Datenrate an, mit der über einen gegebenen Kommunikationskanal, charakterisiert durch die Bandbreite und das Signal-Rausch-Verhältnis, fehlerfrei kommuniziert werden kann. Die Annahme im Folgenden ist jeweils, dass das Signal-Rausch-Verhältnis $\frac{P}{N}$ soweit reduziert ist, dass die jeweilige Kanalkapazität genau der zur Übertragung von Audioinformationen nötigen Bitrate entspricht. Ferner sei das Rauschen bzw. dessen Leistung konstant. Ausgehend von einer Referenzkanalkapazität C_0, einem Kompressionsverhältnis $\alpha \in (0, 1)$ und einem Referenz Signal-Rausch-Verhältnis $SNR_0 := \frac{P_0}{N_0}$ ergibt sich eine mögliche Reduktion des Signal-Rausch-Verhältnisses, bei konstanter Bandbreite, durch Reduktion auf die Kanalkapazität $C_1 = \alpha C_0$ auf den Wert

$$SNR_1 = (1 + SNR_0)^\alpha - 1. \tag{6.2}$$

Bei konstanter Rauschleistung entspricht eine Reduktion des Signal-Rausch-Verhältnisses direkt einer gleichgroßen Reduktion der zur Datenübertragung nötigen Signalleistung.

Abb. 6.1a zeigt exemplarisch die Abhängigkeit von SNR_1 vom Kompressionsverhältnis α mit $SNR_0 = 30\,\mathrm{dB}$. Zur besseren Einschätzung der Größenordnungen: Für eine Bandbreite von $B = 8\,\mathrm{kHz}$ sowie einer Kanalkapazität $C = 64\,\mathrm{kbit/s}$ ergibt sich ein nötiges Signal-Rausch-Verhältnis von nahezu exakt 24 dB. Abb. 6.1b zeigt das Verhältnis $\frac{SNR_1}{SNR_0}$, hier wurden jeweils lineare Einheiten verwendet, in Abhängigkeit von SNR_0 mit $\alpha = \frac{4{,}67\,\mathrm{kbit/s}}{C_0}$ und $C_0 = 64\,\mathrm{kbit/s}$, $32\,\mathrm{kbit/s}$ sowie $C_0 = 16\,\mathrm{kbit/s}$. $4{,}67\,\mathrm{kbit/s}$ ist die Bitrate des besten Rückkopplungsautoencoders, 64 kbit/s entspricht der Bitrate des G.722, 32 kbit/s entspricht in etwa der Bitrate von Opus. 16 kbit/s wurden für eine weitere Veranschaulichung hinzugefügt. Im Vergleich zum G.722 ergeben sich mögliche Reduktionen



(a) (b)

Abbildung 6.1: Dargestellt sind in (a) das Signal-Rausch-Verhältnis SNR_1 in Abhängigkeit vom Kompressionsverhältnis $\alpha \in (0, 1)$ bei einem ursprünglichen Signal-Rausch-Verhältnis $SNR_0 = 30\,dB$ sowie zugehörigen Kanalkapazitäten $C_1 = \alpha C_0$ und identischer Bandbreite und in (b) das Verhältnis der Signal-Rausch-Verhältnisse vor (SNR_0) und nach (SNR_1) Reduktion der Kanalkapazität in Folge verbesserter Datenkompression in Abhängigkeit vom Ausgangssignal-zu-Rauschverhältnis SNR_0. Die drei Kurven entsprechen drei verschieden starken Kanalkapazitätsreduktionen, wobei jeweils von der Kanalkapazität C_0 auf die Kanalkapazität $C_1 = 4,67\,kbit/s$, der Bitrate des besten Rückkopplungsautoencoders, reduziert wird.

der Signalleistung auf $1/50$ bis weniger als $1/100$ der Ursprungssignalleistung. Im Vergleich zu Opus ergeben sich mögliche Reduktionen auf etwa $1/25$ bis weniger als $1/100$ der Ursprungssignalleistung. Selbst bei einer Referenzkapazität von 16 kbit/s ergäbe sich eine mögliche Reduktion auf mindestens in etwa $1/10$ der Ursprungssignalleistung. Natürlich stellt dies nur eine theoretische Einschätzung der Energieeinsparung dar. Real hängt die Einsparung von der tatsächlich eingesetzten Kanalcodierung sowie der Übertragungstechnik inklusive des genutzten Protokolls, welches z.B. definiert, wie viele Nutzbits je Zeitschritt übertragen werden können, ab. Eine umfassende Betrachtung sprengt den Rahmen der vorgelegten Arbeit. Für den Electrocodec mit 3-Bit-Codierung, welcher sicher keine Reduktion der Sprachverständlichkeit bewirkt, ergibt sich bei analoger Rechnung und $SNR_0 = 20\,dB$ im Vergleich zum G.722 eine mögliche Reduktion der Signalleistung auf etwa $1/20$. Im Vergleich zu Opus ergibt sich eine mögliche Reduktion der Signalleistung auf etwa $1/3$. Die sich durch die Codierung der Erregungsmuster ergebenden Bitraten können also eine starke Reduktion der benötigten Übertragungsleistung, und damit Gesamtenergie, eines drahtlosen Übertragungssystems bewirken.

Die Rechenkomplexität der jeweiligen Codierungsansätze wurde in dieser Arbeit beim Entwurf der Algorithmen nicht berücksichtigt. Hier sollen die entwickelten Verfahren kurz hinsichtlich ihrer Rechenkomplexität grob verglichen werden. Ein genauer Vergleich mit anderen Verfahren,

etwa Opus, ist schwierig, da die genaue Komplexität, gemessen durch z.B. die Zahl der Multiplikationen und Additionen, nicht in der Literatur dokumentiert ist. Die Rechenkomplexität ist für die Autoencoder relativ gering, da sie aus nur wenigen tausend Gewichten bestehen und einer entsprechenden Zahl an Multiplikationen und, etwas weniger, Additionen bedürfen. Die Vektorquantisierung des besten Rückkopplungsautoencoders FRAE-L5-H3-R4, welcher 3353 Gewichte aufweist, nutzte ein 6 Bit großes Codebuch. Die assoziierte Komplexität der Codevektorauswahl entspricht damit unter Verwendung der Formel aus [SG11] in etwa 1200 Additionen und Multiplikationen. Insgesamt ergeben sich in etwa 7900 Multiplikationen und Additionen für den besten Rückkopplungsautoencoder. Zusätzlich sind weitere Rechenoperationen für die Anwendung der Aktivierungsfunktionen nötig, die an dieser Stelle vernachlässigt werden.

Der Electrocodec nutzt bei der Selektion von acht Subbändern acht DPCMs. Für jede dieser DPCM muss im schlechtesten Fall je Zeitschritt ein neuer Prädiktor mittels der Wiener-Hopfgleichung bestimmt werden. Die hierbei durchzuführende Matrixinversion kann mit weniger als $5,67 \cdot n^{\log_2(7)}$ arithmetischer Operationen (Multiplikation und Addition) bestimmt werden [Str69], wobei n die Ordnung der jeweiligen Matrix ist. Für einen Prädiktor der Ordnung fünf ist eine 5x5 Matrix zu invertieren, welche folglich weniger als $5,67 \cdot 5^{\log 2(7)} \approx 520$ arithmetische Operationen benötigt. Für acht Subbänder werden also für die Matrixinversion weniger als in etwa 4160 arithmetische Operationen benötigt. Die eigentliche Berechnung der Prädiktorkoeffizienten, Auswahl der Codevektoren sowie die Schätzung der Autokorrelation sind vergleichsweise vernachlässigbar. Einige wenige Rechenoperationen fallen noch durch die arithmetische Codierung an. Jedoch ist auch diese aufgrund der geringen Subbandzahl vernachlässigbar. Durch großzügiges Aufrunden käme man auf insgesamt (Encoder + Decoder) 8500 arithmetische Operationen je Zeitschritt für den Electrocodec. Dies ist die gleich Größenordnung wie der oben genannte Rückkopplungsautoencoder. Anders verhält es sich mit Hinblick auf den entwickelten verlustlosen ZDLLC Codec. Dieser nutzt größere künstliche neuronale Netze in jedem Subband sowie ein ziemlich rechenintensives Verfahren für die Schätzung von Auftrittswahrscheinlichkeiten. Selbst bei zwei Schichten mit jeweils 64 Neuronen ergeben sich für Encoder und Decoder zusammen mehr als 100.000 Parameter und entsprechend viele Multiplikationen und Additionen. Hinsichtlich der Rechenkomplexität ist dieser folglich mit Abstand der schlechteste Codec aller untersuchten Verfahren.

Hinsichtlich des nötigen Speicherplatzes scheint der Electrocodec einen Vorteil gegenüber den anderen Verfahren zu haben, da nur wenige Werte

(Codebücher und Kontextwahrscheinlichkeiten) dauerhaft vorrätig gehalten werden müssen. Der Speicher ist in Hörgeräten typischerweise sehr begrenzt [GPB22]. Für den Rückkopplungsautoencoder müsste zumindest die Hälfte der Parameter im Speicher des Empfängerprozessors abgelegt werden. Hier könnten Kompressionsverfahren für neuronale Netze selbst Abhilfe schaffen, um den nötigen Speicherplatz zu reduzieren [Hin+22c].

Insgesamt zeigen sich die Autoencoder als den beiden anderen entworfenen Verfahren überlegen. Ihre erzielte Bitrate ist deutlich geringer bei gleicher algorithmischer Latenz und näherungsweise gleicher (Electrocodec) oder deutlich geringerer (ZDLLC) Komplexität sowie trotz allem moderater Speichernutzung. Die flexible Optimierung der Autoencoder gestattet ferner, dies ist ein sehr großer Vorteil insbesondere gegenüber dem Electrocodec, eine einfache Anpassung an eine Vielzahl an Signalverarbeitungsstrategien von Cochlea-Implantaten, sodass eine weite Anwendung unmittelbar möglich erscheint. Einziges Manko ist gegebenfalls der leichte Abfall der nominellen Verständlichkeit, gemessen durch das Short-Time Objective Intelligibility Measure auf Basis vocodeter Signale (VSTOI), bei hohen Signal-Rausch-Verhältnissen, wie in Abb. 4.23 zu sehen. Ob dieser tatsächlich messbar und nicht etwa durch Deckeneffekte ohne Bedeutung ist, müsste durch Hörtests eruiert werden. Die Ergebnisse des Hörtests und der in diesem Kontext unter anderem bestimmte Median VSTOI-Wert der **EC2** sowie **Opus52v** Testbedingung, siehe hierzu Tabelle 4.2, lassen es plausibel erscheinen, dass ein Abfall um $0,005$ bis $0,01$ kaum einen Einfluss auf die Verständlichkeit haben dürfte. Daher ist zu erwarten, dass auch bei hohem Signal-Rausch-Verhältnis die Rückkopplungsautoencoder eine oft unveränderte Verständlichkeit der codierten Erregungsmuster erzielen.

6.1 ZUKÜNFTIGE ARBEITEN

Ein wesentliches Hauptkriterium beim Entwurf der entwickelten Codierungsverfahren war eine möglichst geringe algorithmische Latenz. Tatsächlich weisen alle Verfahren eine algorithmische Latenz von 0 ms auf. Eine Latenz von weniger als 10 ms ist zwar relevant für drahtlose Audioübertragung im Kontext direkter Kommunikation, d.h. von Angesicht zu Angesicht, die etwa bei kontralateralem Routing von Signalen oder Distanzmikrophonen auftritt [Eur13]. Sie dürfte jedoch bei Anwendungen wie dem drahtlosen Übertragen von Telefonanrufen keine allzu große Bedeutung haben. Bei VOIP-Telefonaten etwa ist eine Latenz von über 50 ms durchaus noch akzeptabel [DQ19] und es ist nicht ersichtlich, wieso diese Latenz für Träger von Cochlea-Implantaten geringer

ausfallen müsste. Eine derartige Latenz würde generative Kompressionsansätze gestatten, die typischerweise, für einen gewissen Qualitätsbereich, niedrigere Datenraten erzielen, jedoch höhere Latenzen benötigen.

Das im Rahmen der vorgelegten Arbeit immer wieder genutzte objektive Sprachverständlichkeitsmaß Short Time Objective Intelligibility Measure (STOI) wurde für Normalhörende entwickelt, wird jedoch aufgrund seiner Leistungsfähigkeit auch im Bereich der Cochlea-Forschung eingesetzt. Jedoch gibt es sicherlich keine optimale Bewertung der Verständlichkeit von Erregungsmustern. Qazi et al. [Qaz+13] haben festgestellt, dass das Sprachverstehen im Wesentlichen unbeeinflusst bleibt, wenn der Dynamikumfang jeder Elektrode eines Cochlea-Implantatträgers auf den Minimalwert 1 gesetzt wird. STOI würde diese Dynamikumfangreduktion jedoch negativ bewerten. Ein auf Cochlea-Implantatträger ausgelegtes objektives Sprachverständlichkeitsmaß würde eine weitere Reduktion der Bitrate bei gleichbleibender Sprachverständlichkeit ermöglichen. Eine größere Zahl an Hörtests mit Trägern von Cochlea-Implantaten hätte ferner dabei geholfen, die entwickelten Kompressionsverfahren weiter zu verbessern, jedoch waren diese aus Kapazitätsgründen nicht durchführbar.

Weiterhin wurde im Rahmen der vorgelegten Arbeit der Einfluss einer Dynamikkompression der Eingangsaudiosignale, welche im Kontext der Forschungsimplementierung des Advanced Combination Encoders typischerweise ignoriert wird [Nog+05], nicht untersucht. Da die akustischen Szenarien recht stationär waren, und, bei Überschreiten der maximal zulässigen Amplitude im elektrischen Bereich, die Signale zur Vermeidung von Abschneidungen passend skaliert wurden, dürfte der Einfluss jedoch recht gering ausfallen. Zukünftige Arbeiten könnten dennoch der Vollständigkeit halber eine Dynamikkompression berücksichtigen.

Bei der drahtlosen Übertragung von Audiosignalen respektive Daten kommt es im Allgemeinen zu Übertragungsfehlern. Diese würden ohne weitere Maßnahmen zu einer Reduktion der Audioqualität oder gar dem Scheitern der Übertragung als Ganzes im Falle etwa eines komplexeren Kompressionsalgorithmus führen. Zur drastischen Reduktion der Auftrittswahrscheinlichkeit von sogenannten Paketverlusten werden standardmäßig Kanalcodierungsverfahren verwendet, welche durch geschickte Erhöhung der Redundanz eine Vielzahl an Übertragungsfehlern korrigierbar werden lässt. Jedoch ist zur Optimierung des Verbundes aus Kompressionsalgorithmus und Übertragungskanal das Zusammenspiel dieser beiden Komponenten zu optimieren. Hierbei wäre zu untersuchen, ob die Fehlertoleranz des Kompressionsalgorithmus gegenüber Paketverlusten verbesserbar ist. Dann könnte gegebenenfalls die durch Kanalcodierung erhöhte Redundanz auf Kosten einer dann mutmaß-

lich erhöhten Bitrate des Kompressionsalgorithmus reduziert werden. In Summe könnte sich so ein leistungsfähigeres System ergeben. Dieser Aspekt wurde im Rahmen dieser Arbeit nicht betrachtet, wobei jedoch teilweise auf eine gewisse Fehlerrobustheit geachtet wurde respektive diese intrinsisch vorhanden war. Der Electrocodec hat absichtlich die Bandselektion unabhängig von vorherigen Zeitschritten komprimiert, damit im Falle des Verlustes eines Pakets die Bandselektion dennoch in nachfolgenden Zeitschritten korrekt decodierbar bleibt. Ferner ist der entwickelte Autoencoderansatz ohne Rückkopplung intrinsisch robust gegenüber Übertragungsfehlern, da er nur Frequenzabhängigkeiten betrachtet und keine zeitlichen Abhängigkeiten nutzt, um die Kompression zu verbessern. Hierdurch geht bei Verlust eines Datenframes auch nur dieses Datenframe verloren. Andere Zeitschritte sind davon nicht berührt. Zukünftige Arbeiten sollten den Aspekt des möglichen Verlustes an Datenframes mit berücksichtigen und das Gesamtsystem optimieren.

7

ZUSAMMENFASSUNG

In der vorgelegten Arbeit wurden Kompressionverfahren entwickelt und für die Kompression der Erregungsmuster von Cochlea-Implantaten untersucht. Den Kontext bildet die drahtlose Übertragung von Audiosignalen von externen Geräten an den Signalprozessor von Cochlea-Implantaten. Als Signalverarbeitungsstrategie des Cochlea-Implantats wurde im Rahmen der vorgelegten Arbeit der Advanced Combination Encoder (ACE) verwendet. Dabei wurde ein konventionelles Kompressionsverfahren, der Electrocodec, welches ohne künstliche neuronale Netze auskommt, sowie drei Kompressionsverfahren auf Basis künstlicher neuronaler Netze vorgeschlagen. Das Augenmerk lag hier neben einer möglichst geringen Datenrate auf möglichst niedriger Latenz. Daher weisen alle entwickelten Verfahren eine algorithmische Latenz von 0 ms auf.

Die vom Electrocodec codierten Erregungsmuster wurden hinsichtlich ihrer Sprachverständlichkeit und Qualität an Trägern von Cochlea-Implantaten getestet. Es wurde hierbei bei einer mittleren Bitrate von 24,3 kbit/s keine Reduktion der Sprachverständlichkeit sowie Sprachqualität festgestellt, wodurch eine bessere Leistungsfähigkeit als die des Opus-Audiocodecs erzielt werden konnte. Ferner wurde die Bitrate des G.722 Sprachcodecs, dem einzigen Codec von dem eine Anwendung im Rahmen von Cochlea-Implantaten bekannt ist, von 64 kbit/s deutlich unterboten bei niedriger algorithmischer Latenz. Mittels des Nadaraya-Watson Kernel Regressors konnten für stimmlose Sprache relevante experimentelle Ergebnisse erzielt werden, die nahelegen, dass die vom Electrocodec genutzte lineare Prädiktion näherungsweise optimal ist.

Der einzige entwickelte verlustlose Kompressionsalgorithmus auf Basis der Erregungsmuster erzielte eine Bitrate von 28,6 kbit/s auf dem mit Rauschen gemischten Timitdatensatz im Vergleich zu minimal 33,6 kbit/s für Opus bei gleichzeitig signifikant niedriger algorithmischer Latenz. Auf demselben Datensatz erzielte der Electrocodec, mit verlustbehafteter Kompression, eine Birate von bis zu 20,1 kbit/s bei gleicher algorithmi-

scher Latenz. Mittels Autoencodern sowie Hyperparameteroptimierung und dem Simultanuous Perturbation Stochastic Approximation (SPSA) Algorithmus zur numerischen Approximation des Gradienten des Short-Time Objective Intelligibility Measures (STOI), einem Algorithmus zur objektiven Bewertung der Verständlichkeit von Sprachsignalen, war es schlussendlich möglich, die Bitrate wesentlich unter 10 kbit/s zu reduzieren. Autoencoder ohne Rückkopplung erzielten dabei die Referenzverständlichkeit auf dem mit Rauschen gemischten Timitdatensatz bei einer Bitrate von 10,98 kbit/s. Autoencoder mit Rückkopplung der decodierten Ausgangssignale zu den Eingangsschichten des Encoders und Decoders erzielten oder übertrafen die Referenzverständlichkeit auf dem selben Datensatz bei einer Bitrate von 4,67 kbit/s. Jeweils steigt die Bitrate nur unwesentlich bei einer Erhöhung der Hintergrundrauschleistung der Testdaten.

Das wichtigste Resultat der vorgelegten Arbeit ist ein Verfahren zur automatischen Entwicklung näherungsweise optimaler Kompressionsalgorithmen durch eine Kombination aus Bayescher Optimierung zur Hyperparameteroptimierung, numerischen Approximationsverfahren zur Optimierung der verwendeten Quantisierer und der Autoencoder bezüglich der Sprachverständlichkeit der decodierten Signale, sowie Entropiecodierung der Quantisierungsindizes zur Minimierung der schlussendlichen Bitrate. Dieses Vorgehen ist derart generisch, dass es für beliebige Signalverarbeitungstrategien, die in Cochlea-Implantaten Anwendung finden, nutzbar sein dürfte.

LITERATUR

[AM22] K. H. A. Abdel-Latif und H. Meister. „Speech Recognition and Listening Effort in Cochlear Implant Recipients and Normal-Hearing Listeners“. In: *Frontiers in Neuroscience* 15 (2022). ISSN: 1662-453X. DOI: 10.3389/fnins.2021.725412. URL: https://www.frontiersin.org/article/10.3389/fnins.2021.725412.

[Agg18] C. Aggarwal. *Neural Networks and Deep Learning: A Textbook*. Springer Cham, Jan. 2018. ISBN: 978-3-319-94462-3. DOI: 10.1007/978-3-319-94463-0.

[AV16] S. Aja-Fernández und G. Vegas-Sánchez-Ferrero. „Statistical analysis of noise in MRI“. In: *Switzerland: Springer International Publishing* (2016).

[And+22] S. R. Anderson, R. Jocewicz, A. Kan, J. Zhu, S. Tzeng und R. Y. Litovsky. „Sound source localization patterns and bilateral cochlear implants: Age at onset of deafness effects“. In: *PLOS ONE* 17.2 (Feb. 2022), S. 1–30. DOI: 10.1371/journal.pone.0263516.

[Ash08] J. Ashmore. „Cochlear Outer Hair Cell Motility“. In: *Physiological Reviews* 88.1 (2008). PMID: 18195086, S. 173–210. DOI: 10.1152/physrev.00044.2006. eprint: https://doi.org/10.1152/physrev.00044.2006. URL: https://doi.org/10.1152/physrev.00044.2006.

[Att11] V. Atti. „Predictive Modeling of Speech“. In: *Algorithms and Software for Predictive and Perceptual Modeling of Speech*. Cham: Springer International Publishing, 2011, S. 9–52. ISBN: 978-3-031-01516-8. DOI: 10.1007/978-3-031-01516-8_2. URL: https://doi.org/10.1007/978-3-031-01516-8_2.

[Bau17] J.-Y. Baudais. „Fundamental energetic limits of radio communication systems“. In: *Comptes Rendus Physique* 18.2 (2017). Energy and radiosciences, S. 144–155. ISSN: 1631-0705. DOI: https://doi.org/10.1016/j.crhy.2016.11.009. URL: https://www.sciencedirect.com/science/article/pii/S1631070516301621.

[Beh13] E. Behrends. *Elementare Stochastik*. Erste Edition. Vieweg+Teubner Verlag Wiesbaden, 2013.

[Ben91] R. A. Bentler. „The resonance frequency of the external auditory canal in children.“ In: *Ear and hearing* 12 2 (1991), S. 89–90.

[Bla93] R. Blacher. „Higher Order Correlation Coefficients“. In: *Statistics* 25 (1993), S. 1–15.

[BFM10] E. Böhmler, J. Freudenberger und M. Müller. „Comparison of SBC and G.722 speech codecs for Bluetooth wideband speech transmission“. In: Jan. 2010.

[BP70] G. E. P. Box und D. A. Pierce. „Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models“. In: *Journal of the American Statistical Association* 65.332 (1970).

[Can20] Ç. Candan. „Making linear prediction perform like maximum likelihood in Gaussian autoregressive model parameter estimation“. In: *Signal Processing* 166 (2020), S. 107256. ISSN: 0165-1684. DOI: https://doi.org/10.1016/j.sigpro.2019.107256. URL: https://www.sciencedirect.com/science/article/pii/S0165168419303081.

[CC17] S. Chadha und A. Cieza. „Promoting global action on hearing loss: World Hearing Day“. In: *International Journal of Audiology* (Feb. 2017), S. 1–3. DOI: 10.1080/14992027.2017.1292431.

[CKC17] D. Chaudhuri, S. Kundu und N. Chattoraj. „Harvesting energy with zinc oxide bio-compatible piezoelectric material for powering cochlear implants“. In: *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*. 2017, S. 1–5. DOI: 10.1109/IPACT.2017.8245043.

[CDP99] H. Chen, T. Duncan und B. Pasik-Duncan. „A Kiefer-Wolfowitz algorithm with randomized differences“. In: *IEEE Transactions on Automatic Control* 44.3 (1999), S. 442–453. DOI: 10.1109/9.751340.

[DQ19] S. Daoud und Y. Qu. „A Comparison Research on DSCP Marking's Impact to the QoS of VoIP-based and SS7-based Phone Calls“. In: *2019 7th International Conference on Information, Communication and Networks (ICICN)*. 2019, S. 66–71. DOI: 10.1109/ICICN.2019.8834943.

[Déf+22] A. Défossez, J. Copet, G. Synnaeve und Y. Adi. *High Fidelity Neural Audio Compression*. 2022. DOI: 10.48550/ARXIV.2210.13438. URL: https://arxiv.org/abs/2210.13438.

[DJ17] A. Dhanasingh und C. Jolly. „An overview of cochlear implant electrode array designs“. In: *Hearing Research* 356 (2017), S. 93–103. ISSN: 0378-5955. DOI: https://doi.org/10.1016/j.heares.2017.10.005. URL: https://www.sciencedirect.com/science/article/pii/S0378595517302940.

[Die+15] M. Dietz u. a. „Overview of the EVS codec architecture“. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, S. 5698–5702. DOI: 10.1109/ICASSP.2015.7179063.

[DG17] M. F. Dorman und R. H. Gifford. „Speech Understanding in Complex Listening Environments by Listeners Fit With Cochlear Implants“. In: *Journal of Speech, Language, and Hearing Research* 60.10 (2017), S. 3019–3026. DOI: 10.1044/2017_JSLHR-H-17-0035. eprint: https://pubs.asha.org/doi/pdf/10.1044/2017_JSLHR-H-17-0035. URL: https://pubs.asha.org/doi/abs/10.1044/2017_JSLHR-H-17-0035.

[DWS16] M. M. Duke, J. Wolfe und E. C. Schafer. „Recognition of Speech from the Television with Use of a Wireless Technology Designed for Cochlear Implants.“ In: *Journal of the American Academy of Audiology* 27 5 (2016), S. 388–94.

[Egg17] J. J. Eggermont. *Hearing Loss: Causes, Prevention, and Treatment*. First Edition. Academic Press, 2017.

[Els18] J. Elstrodt. *Maß- und Integrationstheorie*. 8. Aufl. Springer Spektrum Berlin, Heidelberg, Jan. 2018. ISBN: 978-3-662-57939-8. DOI: https://doi.org/10.1007/978-3-662-57939-8.

[Eur] European Broadcasting Union. *EBU SQAM CD - Sound Quality Assessment Material Recordings for Subjective Tests*. online - last access 24.11.2022. URL: https://tech.ebu.ch/publications/sqamcd.

[Eur13] European Telecommunications Standards Institute. *Electromagnetic compatibilityand Radio spectrum Matters (ERM); System Reference Document; Short Range Devices (SRD); Technical characteristics of wireless aids for hearing impaired people operating in the VHF and UHF frequency range*. Techn. Ber. ETSI TR 102 791 V1.2.1 (2013-08). ETSI Technical Committee Electromagnetic compatibility und Radio spectrum Matters, Aug. 2013, S. 32. URL: https://www.etsi.org/deliver/etsi_tr/102700_102799/102791/01.02.01_60/tr_102791v010201p.pdf.

[Eur18] European Telecommunications Standards Institute. *DiDigital Enhanced Cordless Telecommunications (DECT); Low Complexity Communication Codec plus (LC3plus)*. Techn. Ber. ETSI TS 103 634 V1.3.1. Letzter Zugriff: 07.06.2023. 2018.

[Fal+15] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates und S. Scollie. „Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and limitations of existing tools“. In: *IEEE Signal Processing Magazine* 32.2 (2015), S. 114–124. DOI: 10.1109/MSP.2014.2358871.

[Fri14] E. Friauf. „Hearing“. In: *e-Neuroforum* (2014). DOI: https://doi.org/10.1007/s13295-014-0062-8.

[Fum+21] M. J. Fumero, A. Eustaquio-Martín, J. M. Gorospe, R. Polo López, M. A. Gutiérrez Revilla, L. Lassaletta, R. Schatzer, P. Nopp, J. S. Stohl und E. A. Lopez-Poveda. „A state-of-the-art implementation of a binaural cochlear-implant sound coding strategy inspired by the medial olivocochlear reflex“. In: *Hearing Research* 409 (2021), S. 108320. ISSN: 0378-5955. DOI: https://doi.org/10.1016/j.heares.2021.108320. URL: https://www.sciencedirect.com/science/article/pii/S0378595521001544.

[GN18] T. Gajecki und W. Nogueira. „A Synchronized Binaural N-of-M Sound Coding Strategy for Bilateral Cochlear Implant Users“. In: *Speech Communication; 13th ITG-Symposium*. 2018, S. 1–5.

[GN22] T. Gajecki und W. Nogueira. „Deep Latent Fusion Layers for Binaural Speech Enhancement“. In: (Okt. 2022). DOI: 10.36227/techrxiv.21215378.v1. URL: https://www.techrxiv.org/articles/preprint/Deep_Latent_Fusion_Layers_for_Binaural_Speech_Enhancement/21215378.

[GPB22] L. Gerlach, G. Payá-Vayá und H. Blume. „A Survey on Application Specific Processor Architectures for Digital Hearing Aids“. In: *Journal of Signal Processing Systems* 94.11 (Nov. 2022), S. 1293–1308. ISSN: 1939-8115. DOI: 10.1007/s11265-021-01648-0.

[GG91] A. Gersho und R. M. Gray. *Vector Quantization and Signal Compression*. Springer New York, NY, 1991. DOI: 10.1007/978-1-4615-3626-0.

[GAH22] R. Ghosh, H. Ali und J. H. L. Hansen. „CCi-MOBILE: A Portable Real Time Speech Processing Platform for Cochlear Implant and Hearing Research“. In: *IEEE Transactions on Biomedical Engineering* 69.3 (2022), S. 1251–1263. DOI: 10.1109/TBME.2021.3123241.

[GR10] R. Gifford und L. Revit. „Speech Perception for Adult Cochlear Implant Recipients in a Realistic Background Noise: Effectiveness of Preprocessing Strategies and External Options for Improving Speech Recognition in Noise“. In: *Journal of the American Academy of Audiology* 21 (Juli 2010), 441–51, quiz 487. DOI: 10.3766/jaaa.21.7.3.

[Gla22] G. Glantz. „Gene Therapy for Hearing Loss on the Horizon“. In: *The Hearing Journal* 75.1 (2022). ISSN: 0745-7472. URL: https://journals.lww.com/thehearingjournal/Fulltext/2022/01000/Gene_Therapy_for_Hearing_Loss_on_the_Horizon.1.aspx.

[Goo22] Google. *Lyra 2 Audio Codec*. Letzter Zugriff: 13.11.2022. 2022. DOI: 10.48550/ARXIV.2210.13438. URL: https://opensource.googleblog.com/2022/09/lyra-v2-a-better-faster-and-more-versatile-speech-codec.html.

[Gue+16] N. Guevara, A. Bozorg-Grayeli, J.-P. Bebear, M. Ardoint, S. Saaï, D. Gnansia, M. Hoen, P. Romanet und J.-P. Lavieille. „The Voice Track multiband single-channel modified Wiener-filter noise reduction system for cochlear implants: patients' outcomes and subjective appraisal“. In: *International Journal of Audiology* 55.8 (2016). PMID: 27108635, S. 431–438. DOI: 10.3109/14992027.2016.1172267. eprint: https://doi.org/10.3109/14992027.2016.1172267. URL: https://doi.org/10.3109/14992027.2016.1172267.

[HMK18] F. Hajiaghababa, H. R. Marateb und S. Kermani. „The design and validation of a hybrid digital-signal-processing plug-in for traditional cochlear implant speech processors“. In: *Computer Methods and Programs in Biomedicine* 159 (2018), S. 103–109. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2018.03.003. URL: https://www.sciencedirect.com/science/article/pii/S0169260717306582.

[HJH06] A.-S. Helvik, G. Jacobsen und L. R.-M. Hallberg. „Psychological well-being of adults with acquired hearing impairment“. In: *Disability and Rehabilitation* 28.9 (2006). PMID: 16690583, S. 535–545. DOI: 10.1080/09638280500215891. eprint: https:

//doi.org/10.1080/09638280500215891. URL: https://doi.org/10.1080/09638280500215891.

[HGJ21] F. Henry, M. Glavin und E. Jones. „Noise Reduction in Cochlear Implant Signal Processing: A Review and Recent Developments“. In: *IEEE Reviews in Biomedical Engineering* PP (Juli 2021), S. 1–1. DOI: 10.1109/RBME.2021.3095428.

[HD19a] J. Herre und S. Dick. „Psychoacoustic Models for Perceptual Audio Coding—A Tutorial Review“. In: *Applied Sciences* 9.14 (2019). ISSN: 2076-3417. DOI: 10.3390/app9142854. URL: https://www.mdpi.com/2076-3417/9/14/2854.

[HD19b] J. Herre und S. Dick. „Psychoacoustic Models for Perceptual Audio Coding—A Tutorial Review“. In: *Applied Sciences* 9.14 (2019). ISSN: 2076-3417. DOI: 10.3390/app9142854. URL: https://www.mdpi.com/2076-3417/9/14/2854.

[Hoc+97] I. J. Hochmair-Desoyer, E. Schulz, L. M. Moser und M. Schmidt. „The HSM sentence test as a tool for evaluating the speech understanding in noise of cochlear implant users.“ In: *The American journal of otology* 18 6 Suppl (1997).

[HL08] Y. Hu und P. Loizou. „A new sound coding strategy for suppressing noise in cochlear implants“. In: *The Journal of the Acoustical Society of America* 124 (Juli 2008), S. 498–509. DOI: 10.1121/1.2924131.

[HWL21] E. H.-H. Huang, C.-M. Wu und H.-C. Lin. „Combination and Comparison of Sound Coding Strategies Using Cochlear Implant Simulation With Mandarin Speech“. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29 (2021), S. 2407–2416. DOI: 10.1109/TNSRE.2021.3128064.

[Hud97] A. J. Hudspeth. „How hearing happens“. en. In: *Neuron* 19.5 (Nov. 1997), S. 947–950.

[Hut+22a] M. E. Huth, R. L. Boschung, M. D. Caversaccio, W. Wimmer und M. Georgios. „The effect of internet telephony and a cochlear implant accessory on mobile phone speech comprehension in cochlear implant users“. en. In: *Eur. Arch. Otorhinolaryngol.* 279.12 (Dez. 2022), S. 5547–5554.

[Hut+22b] M. E. Huth, R. L. Boschung, M. D. Caversaccio, W. Wimmer und M. Georgios. „The effect of internet telephony and a cochlear implant accessory on mobile phone speech comprehension in cochlear implant users“. In: *European Archives of Oto-Rhino-Laryngology* (Apr. 2022). ISSN: 1434-4726. DOI:

10.1007/s00405-022-07383-x. URL: https://doi.org/10.1007/s00405-022-07383-x.

[Hut+22c] M. E. Huth, R. L. Boschung, M. D. Caversaccio, W. Wimmer und M. Georgios. „The effect of internet telephony and a cochlear implant accessory on mobile phone speech comprehension in cochlear implant users“. In: *European Archives of Oto-Rhino-Laryngology* (Apr. 2022). ISSN: 1434-4726. DOI: 10.1007/s00405-022-07383-x. URL: https://doi.org/10.1007/s00405-022-07383-x.

[Ibe14] O. Ibe. *Fundamentals of Applied Probability and Random Processes, Second Edition*. 2. Aufl. Academic Press, 2014. ISBN: 9780128010358.

[IET] IETF-Codec-Arbeitsgruppe. *Opus - Latenz vs. Bitrate*. https://opus-codec.org/comparison/. Letzter Zugriff: 13.11.2022.

[Int93] International Telecommuncation Union. „ITU Recommendation G.227“. In: (1993). last access 10.09.2019. URL: https://www.itu.int/rec/T-REC-G.227-198811-I/en.

[Int15] International Telecommuncation Union. „ITU-RBS.1534-0 (Method for the subjective assessment of intermediate quality levels of coding systems)“. In: (2015). last access 01.12.2019. URL: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-1-200301-S!!PDF-E.pdf.

[Jea03] S. C. Jean Dickinson Gibbons. *Nonparametric Statistical Inference*. 4th ed., rev. and expanded. Statistics, textbooks and monographs 168. Marcel Dekker, 2003. ISBN: 9780824740528; 0824740521.

[Kar93] A. F. Karr. *Probability*. Erste Edition. Springer New York, NY, 1993.

[Kay+09] H. Kayser, S. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann und B. Kollmeier. „Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses“. In: *EURASIP Journal on Advances in Signal Processing* 2009 (Dez. 2009), S. 6. DOI: 10.1155/2009/298605.

[KAZ00] F. Keiler, D. Arfib und U. Zölzer. „Efficient linear prediction for digital audio effects“. In: *proc. DAFX*. Citeseer. 2000.

[KL08] K. Kokkinakis und P. C. Loizou. „Using blind source separation techniques to improve speech recognition in bilateral cochlear implant patients“. en. In: *J. Acoust. Soc. Am.* 123.4 (Apr. 2008), S. 2379–2390.

[KA93] G. Kubin und B. S. Atal. „Linear autoregressive modeling of unvoiced speech“. In: *The Journal of the Acoustical Society of America* 93.4Supplement (Apr. 1993), S. 2354–2354. DOI: 10.1121/1.406213. eprint: https://pubs.aip.org/asa/jasa/article-pdf/93/4_Supplement/2354/12089911/2354_1_online.pdf. URL: https://doi.org/10.1121/1.406213.

[KKW99] V. Kuehnel, B. Kollmeier und K. Wagener. „Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests“. In: *Zeitschrift für Audiologie* 38 (Sep. 1999), S. 4–15.

[Lar08] M. G. Larson. „Analysis of Variance“. In: *Circulation* 117.1 (2008), S. 115–121. DOI: 10.1161/CIRCULATIONAHA.107.654335. eprint: https://www.ahajournals.org/doi/pdf/10.1161/CIRCULATIONAHA.107.654335. URL: https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.107.654335.

[Len+12] M. Lenarz, H. Sönmez, G. Joseph, A. Büchner und T. Lenarz. „Long-Term Performance of Cochlear Implants in Postlingually Deafened Adults“. In: *Otolaryngology–Head and Neck Surgery* 147 (2012), S. 112–118.

[Li+21] L. Li, J.-Y. Han, W.-Z. Zheng, R.-J. Huang und Y.-H. Lai. „Improved Environment-Aware–Based Noise Reduction System for Cochlear Implant Users Based on a Knowledge Transfer Approach: Development and Usability Study“. In: *Journal of Medical Internet Research* 23 (Okt. 2021), e25460. DOI: 10.2196/25460.

[Lin+21] M. Lindauer, K. Eggensperger, M. Feurer, A. Biedenkapp, D. Deng, C. Benjamins, T. Ruhkopf, R. Sass und F. Hutter. „SMAC3: A Versatile Bayesian Optimization Package for Hyperparameter Optimization“. In: *ArXiv: 2109.09831*. 2021. URL: https://arxiv.org/abs/2109.09831.

[LRS13] G. Lindgren, H. Rootzén und M. Sandsten. *Stationary stochastic processes for scientists and engineers*. CRC press, 2013.

[LY07] X. Liu und D. Yan. „Ageing and hearing loss“. In: *The Journal of Pathology* 211.2 (2007), S. 188–197. DOI: https://doi.org/10.1002/path.2102. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/path.2102. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/path.2102.

[LK54] E. Lukacs und E. P. King. „A Property of the Normal Distribution“. In: *The Annals of Mathematical Statistics* 25.2 (1954), S. 389–394. DOI: 10.1214/aoms/1177728796. URL: https://doi.org/10.1214/aoms/1177728796.

[MG45] K. M.G. *The advanced theory of statistics Volume 1*. 2. Aufl. Kendall's Advanced Theory of Statistics 1. Charles Griffin Company, 1945. ISBN: 0340614307; 9780340614303. URL: libgen.li/file.php?md5=95aa1a9ebfee10852372b9922619407b.

[Mah] M. Mahoney. *PAQ8N*. http://mattmahoney.net/dc/paq.html#paq8. zuletzt abgerufen am 07.04.2022.

[Mah05] M. V. Mahoney. „Adaptive weighing of context models for lossless data compression“. In: 2005.

[Mak75] J. Makhoul. „Linear prediction: A tutorial review“. In: *Proceedings of the IEEE* 63.4 (1975), S. 561–580. DOI: 10.1109/PROC.1975.9792.

[Man+17] G. Manley, A. Gummer, A. Popper und R. Fay. *Understanding the Cochlea*. Springer Cham, Sep. 2017. ISBN: 978-3-319-52073-5. DOI: 10.1007/978-3-319-52073-5.

[MG21] K. V. C. Martins und M. V. S. Goffi-Gomez. „The influence of stimulation levels on auditory thresholds and speech recognition in adult cochlear implant users“. In: *Cochlear Implants International* 22.1 (2021). PMID: 32972324, S. 42–48. DOI: 10.1080/14670100.2020.1822495. eprint: https://doi.org/10.1080/14670100.2020.1822495. URL: https://doi.org/10.1080/14670100.2020.1822495.

[Mas19] M. Massi. *Autoencoder schema*. Online; Letzter Zugriff: 19.06.2023. 2019. URL: https://commons.wikimedia.org/wiki/File:Autoencoder_schema.png.

[Mil+22] S. Miller, J. Wolfe, S. Neumann, E. C. Schafer, J. Galster und S. Agrawal. „Remote microphone systems for cochlear implant recipients in small group settings“. en. In: *J. Am. Acad. Audiol.* 33.3 (März 2022), S. 142–148.

[Miu] H. Miura. *PPMD*. https://github.com/miurahr/ppmd-cffi. zuletzt abgerufen am 07.04.2022.

[MLN16] A. C. Moberly, J. H. Lowenstein und S. Nittrouer. „Word Recognition Variability With Cochlear Implants: "Perceptual Attention"Versus Äuditory Sensitivity
“. In: *Ear Hear* 37.1 (2016), S. 14–26.

[Mob+16] A. C. Moberly, C. Bates, M. S. Harris und D. B. Pisoni. „The enigma of poor performance by adults with cochlear implants“. en. In: *Otol. Neurotol.* 37.10 (Dez. 2016), S. 1522–1528.

[Moo07] B. Moore. *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues*. 2. Aufl. Wiley Series in Human Communication Science. Wiley-Interscience, 2007. ISBN: 047051633X; 9780470516331; 9780470518182; 0470518189.

[Mos+17] I. Mosnier, N. Mathias, J. Flament, D. Amar, A. Liagre-Callies, S. Borel, E. Ambert-Dahan, O. Sterkers und D. Bernardeschi. „Benefit of the UltraZoom beamforming technology in noise in cochlear implant users“. In: *European Archives of Oto-Rhino-Laryngology* 274 (2017), S. 3335–3342.

[MB15] U. Müller und P. G. Barr-Gillespie. „New treatment options for hearing loss“. en. In: *Nat. Rev. Drug Discov.* 14.5 (Mai 2015), S. 346–365.

[Mus15] H.-G. Musmann. *Quellencodierung*. Institut für Informationsverarbeitung, Leibniz Universität Hannover, Vorlesungsskript, 2015.

[NG03] J. G. Nicholas und A. E. Geers. „Personal, social, and family adjustment in school-aged children with a cochlear implant“. In: *Ear and hearing* 24.1 (2003), 69S–81S.

[Nog08] W. Nogueira. „Design and Evaluation of Signal Processing Strategies for Cochlear Implants based on Psychoacoustic Masking and Current Steering“. Diss. Leibniz Universität Hannover, 2008.

[Nog+05] W. Nogueira, A. Büchner, T. Lenarz und B. Edler. „A psychoacoustic „NofM“-type speech coding strategy for cochlear implants“. In: *EURASIP Journal on Advances in Signal Processing* 2005 (2005), S. 1–16.

[Nuñ+20] F. Nuñez-Batalla, A. Fernández-Junquera, L. Súarez-Villanueva, E. Díaz-Fresno, I. Sandoval-Menéndez, J. Martínez und J. Llorente-Pendás. „Application of Wireless Contralateral Routing of Signal (CROS) Technology in Unilateral Cochlear Implant Users“. In: *Acta Otorrinolaringologica (English Edition)* 71 (Nov. 2020), S. 333–342. DOI: 10.1016/j.otoeng.2019.10.004.

[Orf07] S. Orfanidis. *Optimum Signal Processing*. McGraw-Hill Publishing Company, 2007.

[Oti21] Oticon. *Oticon CROS PX miniRITE R - Datasheet*. Letzter Zugriff: 01.06.2023. 2021. URL: https://www.hearingaid.org.uk/files/brandsproducts/oticon/oticon-cros-px-consumer-brochure.pdf.

[OK14] A. Oxenham und H. Kreft. „Speech Perception in Tones and Noise via Cochlear Implants Reveals Influence of Spectral Resolution on Temporal Processing". In: *Trends in hearing* 18 (Mai 2014). DOI: 10.1177/2331216514553783.

[PP02] A. Papoulis und U. Pillai. „Probability, Random Variables, and Stochastic Processes, Fourth Edition". In: (Jan. 2002).

[Pic12] J. O. Pickles. *An Introduction to the Physiology of Hearing*. 4th edition 2012. Emerald Group Publishing Limited, 2012. ISBN: 1780521669; 9781780521664.

[Pri12] D. S. J. D. Prince. *Computer Vision: Models, Learning, and Inference*. Cambridge University Press, 2012. ISBN: 9781107011793.

[PAF01] D. Purves, G. Augustine und D. Fitzpatrick. *Neuroscience*. Second Edition. Oxford University Press Inc, 2001.

[Qaz+13] O. Qazi, B. van Dijk, M. Moonen und J. Wouters. „Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility". In: *Hearing Research* 299 (Feb. 2013), S. 79–87. DOI: 10.1016/j.heares.2013.01.018.

[Ric] M. A. Richards. *The Discrete-Time Fourier Transform and Discrete Fourier Transform of Windowed Stationary White Noise*. https://radarsp.weebly.com/uploads/2/1/4/7/21471216/dft_of_noise.pdf. Letzter Zugriff: 28.04.2023.

[Rob78] A. T. C. Robert V. Hogg. *Introduction to mathematical statistics*. 4th ed. Macmillan, 1978. ISBN: 9780023557101; 0023557109.

[Rol+06] J. Roland, T. C. Huang, A. J. Fishman, S. Waltzman und J. Roland. „Cochlear implant electrode history, choices, and insertion techniques". In: *Cochlear implants* 124 (2006).

[Rum+15] C. Rumeau, J. Frère, B. Montaut-Verient, A. Lion, G. Gauchard und C. Parietti-Winkler. „Quality of life and audiologic performance through the ability to phone of cochlear implant users". In: *Archives of Oto-Rhino-Laryngology* 272 (Dez. 2015), S. 3685–3692. DOI: 10.1007/s00405-014-3448-x.

[SAH18] J. N. Saba, H. Ali und J. H. L. Hansen. „Formant priority channel selection for an "n-of-m" sound processing strategy for cochlear implants". In: *The Journal of the Acoustical Society of America* 144.6 (2018), S. 3371–3380. DOI: 10.1121/1.5080257.

[Say96] K. Sayood. *Introduction to Data Compression*. First Edition. Burlington: Morgan Kaufmann, 1996. ISBN: 978-1558603462.

[Sch+21] M. Schnell, E. Ravelli, J. Büthe, M. Schlegel, A. Tomasek, A. Tschekalinskij, J. Svedberg und M. Sehlstedt. „Lc3 and Lc3plus: The new audio transmission standards for wireless communication". In: *Journal of the Audio Engineering Society* (2021).

[SG22] J. Schuster-Bruce und E. Gosnell. „Conventional Hearing Aid Indications And Selection". en. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, Jan. 2022.

[Sco92] D. W. Scott. „Multivariate Density Estimation: Theory, Practice, and Visualization". In: (1992).

[Sei+17] C. Seifert, J. Thiemann, L. Gerlach, T. Volkmar, G. Payá-Vayá, H. Blume und S. van de Par. „Real-time implementation of a GMM-based binaural localization algorithm on a VLIW-SIMD processor". In: *2017 IEEE International Conference on Multimedia and Expo (ICME)*. 2017, S. 145–150. DOI: 10.1109/ICME.2017.8019478.

[SG11] E. P. Seraco und J. G. R. Gomes. „Computation of the complexity of vector quantizers by affine modeling". In: *Signal Processing* 91.5 (2011), S. 1134–1142. ISSN: 0165-1684. DOI: https://doi.org/10.1016/j.sigpro.2010.10.015. URL: https://www.sciencedirect.com/science/article/pii/S0165168410004020.

[Sha48] C. E. Shannon. „A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), S. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[Sha49] C. Shannon. „Communication in the Presence of Noise". In: *Proceedings of the IRE* 37.1 (1949), S. 10–21. DOI: 10.1109/JRPROC.1949.232969.

[Spa92] J. Spall. „Multivariate stochastic approximation using a simultaneous perturbation gradient approximation". In: *IEEE Transactions on Automatic Control* 37.3 (1992), S. 332–341. DOI: 10.1109/9.119632.

[SDF17] D. Spirrov, B. van Dijk und T. Francart. „Optimal gain control step sizes for bimodal stimulation“. In: *International Journal of Audiology* 57.3 (Nov. 2017), S. 184–93. DOI: 10.1080/14992027.2017.1403655.

[Str69] V. Strassen. „Gaussian elimination is not optimal“. In: *Numerische Mathematik* 13 (1969), S. 354–356.

[SBF22] H. C. Stronks, J. J. Briaire und J. H. M. Frijns. „Residual hearing affects contralateral routing of signals in cochlear implant users“. en. In: *Audiol. Neurootol.* 27.1 (2022), S. 75–82.

[Taa+10] C. H. Taal, R. C. Hendriks, R. Heusdens und J. Jensen. „A short-time objective intelligibility measure for time-frequency weighted noisy speech“. In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2010, S. 4214–4217. DOI: 10.1109/ICASSP.2010.5495701.

[Thi+00] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. g. Beerends und C. Colomes. „PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality“. In: *Journal of the Audio Engineering Society* 48.1/2 (Jan. 2000), S. 3–29.

[Tur+10] C. W. Turner, B. J. Gantz, S. Karsten, J. Fowler und L. A. Reiss. „Impact of hair cell preservation in cochlear implantation: combined electric and acoustic hearing“. en. In: *Otol. Neurotol.* 31.8 (Okt. 2010), S. 1227–1232.

[Vai07] P. Vaidyanathan. „The Theory of Linear Prediction“. In: *Synthesis Lectures on Signal Processing* 2 (Jan. 2007). DOI: 10.2200/S00086ED1V01Y200712SPR03.

[Val+16] J.-M. Valin, G. Maxwell, T. B. Terriberry und K. Vos. *High-Quality, Low-Delay Music Coding in the Opus Codec*. 2016. DOI: 10.48550/ARXIV.1602.04845. URL: https://arxiv.org/abs/1602.04845.

[Vos+21] T. G. Vos u. a. „Influence of Protective Face Coverings on the Speech Recognition of Cochlear Implant Patients“. In: *The Laryngoscope* 131.6 (2021), E2038–E2043. DOI: https://doi.org/10.1002/lary.29447. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/lary.29447. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/lary.29447.

[Vro+18] J. Vroegop, N. Homans, A. Goedegebure, G. Dingemanse, T. van Immerzeel und M. van der Schroeff. „The Effect of Binaural Beamforming Technology on Speech Intelligibility in Bimodal Cochlear Implant Recipients“. In: *Audiology and Neurotology* 23 (Juni 2018), S. 32–38. DOI: 10.1159/000487749.

[Vro+17] J. L. Vroegop, J. G. Dingemanse, N. C. Homans und A. Goedegebure. „Evaluation of a wireless remote microphone in bimodal cochlear implant recipients“. In: *International Journal of Audiology* 56.9 (2017). PMID: 28395552, S. 643–649. DOI: 10.1080/14992027.2017.1308565. eprint: https://doi.org/10.1080/14992027.2017.1308565. URL: https://doi.org/10.1080/14992027.2017.1308565.

[WS19] E. L. Wagner und J. B. Shin. „Mechanisms of Hair Cell Damage and Repair“. In: *Trends Neurosci* 42.6 (Juni 2019), S. 414–424.

[Wan+17] S.-S. Wang, Y. Tsao, H.-L. S. Wang, Y.-H. Lai und L. P.-H. Li. „A deep learning based noise reduction approach to improve speech intelligibility for cochlear implant recipients in the presence of competing speech noise“. In: *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2017, S. 808–812. DOI: 10.1109/APSIPA.2017.8282144.

[WGD21] J. Wathour, P. Govaerts und N. Deggouj. „Variability of fitting parameters across cochlear implant centres“. In: *European Archives of Oto-Rhino-Laryngology* 278 (Dez. 2021), S. 1–9. DOI: 10.1007/s00405-020-06572-w.

[WSS18a] G. Watkins, B. Swanson und G. Suaning. „An Evaluation of Output Signal to Noise Ratio as a Predictor of Cochlear Implant Speech Intelligibility“. In: *Ear and Hearing* 39 (Feb. 2018), S. 1. DOI: 10.1097/AUD.0000000000000556.

[WSS18b] G. D. Watkins, B. A. Swanson und G. J. Suaning. „An Evaluation of Output Signal to Noise Ratio as a Predictor of Cochlear Implant Speech Intelligibility“. In: *Ear and Hearing* 39 (2018), S. 958–968.

[Wil06] J. Wilber. *Der Klang der Innovation bei Cochlear Limited*. https://de.mathworks.com/company/newsletters/articles/the-sound-of-innovation-at-cochlear-limited.html. Letzter Zugriff: 01.06.2023. 2006.

[WDS16] J. Wolfe, M. Duke und E. Schafer. „Speech Recognition of Bimodal Cochlear Implant Recipients Using a Wireless Audio Streaming Accessory for the Telephone“. In: *Otology and neurotology: official publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology* In press (Okt. 2016). DOI: 10.1097/MAO.0000000000000903.

[Wol+16] J. Wolfe, M. M. Duke, E. Schafer, G. Cire, C. Menapace und L. O'Neill. „Evaluation of a wireless audio streaming accessory to improve mobile telephone performance of cochlear implant users“. In: *International Journal of Audiology* 55.2 (2016). PMID: 26681229, S. 75–82. DOI: 10.3109/14992027.2015.1095359. eprint: https://doi.org/10.3109/14992027.2015.1095359. URL: https://doi.org/10.3109/14992027.2015.1095359.

[WMS15] J. Wolfe, M. Morais und E. Schafer. „Improving Hearing Performance for Cochlear Implant Recipients with Use of a Digital, Wireless, Remote-Microphone, Audio-Streaming Accessory“. In: *Journal of the American Academy of Audiology* 26 (Juni 2015), S. 532–9. DOI: 10.3766/jaaa.15005.

[WMF15] J. Wouters, H. J. McDermott und T. Francart. „Sound Coding in Cochlear Implants: From electric pulses to hearing“. In: *IEEE Signal Processing Magazine* 32.2 (2015), S. 67–80. DOI: 10.1109/MSP.2014.2371671.

[Wu+20] P.-z. Wu, J. T. O'Malley, V. de Gruttola und M. C. Liberman. „Age-Related Hearing Loss Is Dominated by Damage to Inner Ear Sensory Cells, Not the Cellular Battery That Powers Them“. In: *Journal of Neuroscience* 40.33 (2020), S. 6357–6366. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.0937-20.2020. eprint: https://www.jneurosci.org/content/40/33/6357.full.pdf. URL: https://www.jneurosci.org/content/40/33/6357.

[Yan+19] Y. Yang, G. Sautière, J. J. Ryu und T. S. Cohen. „Feedback Recurrent AutoEncoder“. In: *CoRR* abs/1911.04018 (2019). arXiv: 1911.04018. URL: http://arxiv.org/abs/1911.04018.

[YMT22] Y. Yang, S. Mandt und L. Theis. *An Introduction to Neural Data Compression*. 2022. arXiv: 2202.06533 [cs.LG].

[Zar+20] A. Zarowski, A. Molisz, L. Coninck, A. Vermeiren, T. Theunen, L. Theuwis, T. Przewoźny, J. Siebert und E. Offeciers. „Influence of the Pre- or Postlingual Status of Cochlear Implant Recipients on Behavioural T/C-levels“. In: *International Journal of Pediatric Otorhinolaryngology* 131 (Jan. 2020), S. 109867. DOI: 10.1016/j.ijporl.2020.109867.

[Zei+15] D. M. Zeitler, M. F. Dorman, S. J. Natale, L. H. Loiselle, W. A. Yost und R. H. Gifford. „Sound Source Localization and Speech Understanding in Complex Listening Environments by Single-sided Deaf Listeners After Cochlear Implantation“. In: *Otology & Neurotology* 36 (2015), S. 1467–1471.

[ZSG90] V. Zue, S. Seneff und J. Glass. „Speech database development at MIT: Timit and beyond“. In: *Speech Communication* 9.4 (1990), S. 351–356. ISSN: 0167-6393. DOI: https://doi.org/10.1016/0167-6393(90)90010-7. URL: https://www.sciencedirect.com/science/article/pii/0167639390900107.

VERÖFFENTLICHUNGEN

ZEITSCHRIFTENARTIKEL

[Hin+21a] R. Hinrichs, T. Gajęcki, J. Ostermann und W. Nogueira. „A subjective and objective evaluation of a codec for the electrical stimulation patterns of cochlear implants." In: *The Journal of the Acoustical Society of America* 149.2 (Feb. 2021), S. 1324–1337. DOI: 10.1121/10.0003571.

[Hin+22b] R. Hinrichs, K. Gerkens, A. Lange und J. Ostermann. „Convolutional Neural Networks for the Classification of Guitar Effects and Extraction of the Parameter Settings of Single and Multi-Guitar Effects from Instrument Mixes". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2022.1 (Okt. 2022), S. 28. DOI: 10.1186/s13636-022-00257-4.

KONFERENZBEITRÄGE

Vorträge

[HBO23] R. Hinrichs, J. Bilsky und J. Ostermann. „Vector-Quantized Feedback Recurrent Autoencoders for the Compression of the Stimulation Patterns of Cochlear Implants at Zero Delay". In: *24th International Conference on Digital Signal Processing (DSP 2023)* (2023).

[HDO21] R. Hinrichs, J. Dunkel und J. Ostermann. „Mixing Time-Frequency Distributions for Speech Command Recognition Using Convolutional Neural Networks". In: *2021 6th International Conference on Frontiers of Signal Processing (ICFSP)* (2021), S. 6–11. DOI: 10.1109/ICFSP53514.2021.9646416.

[Hin+22a] R. Hinrichs, L. Ehmann, H. Heise und J. Ostermann. „Lossless Compression at Zero Delay of the Electrical Stimulation Patterns of Cochlear Implants for Wireless Streaming of Audio Using Artificial Neural Networks". In: *7th International Conference on Frontiers of Signal Processing* (2022).

[Hin+19] R. Hinrichs, T. Gajęcki, J. Ostermann und W. Nogueira. „Coding of Electrical Stimulation Patterns for Binaural Sound Coding Strategies for Cochlear Implants“. In: (2019), S. 4168–4172. DOI: 10.1109/EMBC.2019.8857271.

[HGO22] R. Hinrichs, K. Gerkens und J. Ostermann. „Classification of Guitar Effects and Extraction of Their Parameter Settings from Instrument Mixes Using Convolutional Neural Networks“. In: *11th International Conference on Artificial Intelligence in Music, Sound, Art and Design (EvoMUSART) 2022* (2022). Hrsg. von T. Martins, N. Rodríguez-Fernández und S. M. Rebelo, S. 101–116.

[Hin+24] R. Hinrichs, N. Jiang, R. Beltran, T. Krause, M. Käding, A. Lange, B. Schmidt, J. Ostermann und S. Marx. „Analysis of the Repeatability of the Pencil Lead Break in Comparison to the Ball Impact and Electromagnetic Body-Noise Actuator“. In: *20th World Conference on Non-Destructive Testing (WCNDT 2024)* (2024). Akzeptiert.

[Hin+20] R. Hinrichs, N. Jiang, T. Krause, A. Lange und J. Ostermann. „Analysis of the repeatability of the pencil lead break artificial sound source“. In: *59th Annual British Conference on Non-Destructive Testing*. 2020.

[Hin+18] R. Hinrichs, T. C. Krause, M. Käding, J. Ostermann und S. Marx. „Measurement-Based Model of Structural Sound Transmission in a Concrete Specimen“. In: *12th European Conference on Non-Destructive Testing (ECNDT 2018)* (2018).

[HOO22] R. Hinrichs, F. Ortmann und J. Ostermann. „Vector-Quantized Zero-Delay Deep Autoencoders for the Compression of Electrical Stimulation Patterns of Cochlear Implants using STOI“. In: *IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)* (2022).

[Hin+21b] R. Hinrichs, A. Schmidt, J. Koslowski, B. Bergmann, B. Denkena und J. Ostermann. „Analysis of the impact of data compression on condition monitoring algorithms for ball screws“. In: *Procedia CIRP* 102 (2021). 18th CIRP Conference on Modeling of Machining Operations (CMMO), Ljubljana, Slovenia, June 15-17, 2021, S. 270–275. ISSN: 2212-8271. DOI: 10.1016/j.procir.2021.09.046.

[HSO23] R. Hinrichs, A. J. Y. Sitcheu und J. Ostermann. „Continuous Sign-Language Recognition using Transformers and Augmented Pose Estimation“. In: *12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)*. 2023, S. 672–678. DOI: 10.5220/0011709100003411.

[JHO20] H. Jürgens, R. Hinrichs und J. Ostermann. „Recognizing Guitar Effects and Their Parameter Settings“. In: *23rd International Conference on Digital Audio Effects (DAFx-20)* (2020).

Poster

[Hin+22c] R. Hinrichs, K. Liang, Z. Lu und J. Ostermann. „Improved Compression of Artificial Neural Networks through Curvature-Aware Training“. In: *2022 International Joint Conference on Neural Networks (IJCNN)* (2022), S. 1–8. DOI: 10.1109/IJCNN55064.2022.9892511.

Fachberichte

[Nog+23] W. Nogueira, T. Gajecki, J. Ostermann und R. Hinrichs. „Signal Coding for Binaural Signal Processing in Cochlear Implants“. In: *Unimagazin* 1 (Juni 2023). Kein Peer-Review.

[OH19] J. Ostermann und R. Hinrichs. „Signal Coding for Binaural Signal Processing in Cochlear Implants“. In: *Binaire* (Okt. 2019). Kein Peer-Review.

[OH20] J. Ostermann und R. Hinrichs. „Links und rechts verbinden“. In: *Unimagazin* 1 (Juni 2020). Kein Peer-Review.

8

ANHANG

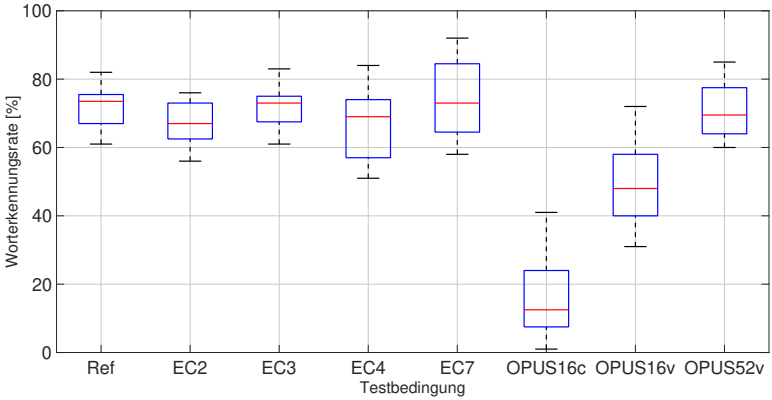


Abbildung 8.1: Worterkennungsraten für alle Testbedingungen derjenigen Probanden des Hörtests, welche die später eingeführte **Opus16v** Testbedingung beurteilt haben. Dies waren die letzten sechs der insgesamt zehn Probanden.

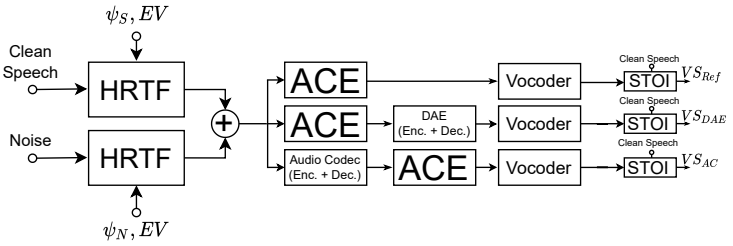


Abbildung 8.2: Blockdiagramm der Erzeugung des TIMIT-Datensatzes inklusive der Berechnung der zugehörigen VSTOI-Werte (VS).

Tabelle 8.1: Worterkennungsraten in Prozent für alle Testbedingungen und Probanden des durchgeführten Hörtests. Es handelt sich um Mittelwerte über die jeweils verwendeten zwei Satzlisten. Die genannte Probandenbezeichnung entspricht jener aus Tabelle 3.4. Die **Opus16v** Testbedingungen wurde lediglich in sechs Probanden untersucht.

	Testbedingung							
Proband	**REF**	**EC2**	**EC3**	**EC4**	**EC7**	**Opus16c**	**Opus16v**	**Opus52v**
ID01	60,5	76,5	58	64,5	72	31,5	N/A	73
ID02	80	70	66,5	74	71	30	N/A	70,5
ID03	73,5	62	65	61,5	69,5	19	N/A	64
ID04	63	48	77	67,5	72	21,5	N/A	65,5
ID05	69	68	73	61	81	4,5	34,5	77,5
ID06	68,5	60	63	61	75	27	44	65
ID07	67	73,5	79	69	71	9,5	57,5	67,5
ID08	78	60,5	67,5	79	68,5	20	67,5	82,5
ID09	74	70,5	77	67	74	27,5	51,5	68,5
ID10	75	70,5	72,5	65	72	7	38	63

Tabelle 8.2: Grenzen, Standardwerte und optimierte Werte der Hyperparameteroptimierung des Autoencoders ohne Rückkopplung aus Abschnitt 4.4 für eine Latentdimension von fünf mit zwei Schichten. Log gibt an, ob die Werte logarithmisch gezogen wurden. Als Aktivierungsfunktion wurde in allen Schichten $Swish$ verwendet.

Parameter	Standard	Untergrenze	Obergrenze	Log	Optimiert
#Neuronen (L1)	16	16	30	X	30
#Neuronen (L2)	8	6	16	X	14
α	0.5	0.1	0.9	X	0.46834
lr	0.001	0.0001	0.1	✓	0.0016
A	100000	10000	200000	✓	11337
c	0.001	0.0001	0.1	✓	0.0129

Tabelle 8.3: Entropie in Bit der untersuchten Autoencodermodell mit (FRAE) und ohne (AEC) Rückkopplung auf dem Testdatensatz nach Optimierung der Gesamtstruktur inklusive des Quantisierers. Zu den Entropiewerten sind die zugehörigen Bitraten angegeben.

Modell\Codebuchgröße [Bit]	6	7	8	10
FRAE-L5-H3-R2	5,2440 4,72 kbit/s	6,0572 5,45 kbit/s	6,9225 6,23 kbit/s	8,6624 7,80 kbit/s
FRAE-L5-H2-R4	5,1566 4,64 kbit/s		6,8753 6,19 kbit/s	8,6680 7,80 kbit/s
FRAE-L5-H3-R4	5,0965 4,59 kbit/s	6,0280 5,43 kbit/s	7,0393 6,34 kbit/s	8,8154 7,93 kbit/s
AEC-L5-H2			6,6956 6,03 kbit/s	8,5009 7,65 kbit/s
AEC-L5-H3			6,9856 6,29 kbit/s	8,8214 7,94 kbit/s

8.1 VERTEILUNG DER DFT-KOEFFIZIENTEN VON GAUSSSCHEM RAUSCHEN

Die Verteilung der DFT-Koeffizienten von Gaußschem Rauschen aus 5.2 kann unmittelbar aus der Berechnungsvorschrift der DFT bestimmt werden: Der k-te DFT-Koeffizient $X(k)$ berechnet sich gemäß

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}nk},$$

wobei hier $x(n) \sim \mathcal{N}(0, \sigma^2)$ ist.

Dies lässt sich als

$$X(k) = \sum_{n=0}^{N-1} x(n) cos(\frac{2\pi}{N}nk) - j\sum_{n=0}^{N-1} x(n) sin(\frac{2\pi}{N}nk) = X_{Re}(k) - jX_{Im}(k)$$

schreiben. Es gilt für normalverteilte, statistisch unabhängige Zufallsvariablen $x_1, \ldots, x_L$ mit $x_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ sowie $b_i \in \mathbb{R}$ die Beziehung [Pri12]

$$\sum_{i=1}^{L} b_i x_i \sim \mathcal{N}(\sum_{i=1}^{L} b_i \mu_i, \sum_{i=1}^{L} b_i^2 \sigma_i^2).$$

Folglich gilt

$$X_{Re}(k) \sim \mathcal{N}(0, \sigma^2 \sum_{n=0}^{N-1} cos(\frac{2\pi}{N}nk)^2), X_{Im}(k) \sim \mathcal{N}(0, \sigma^2 \sum_{n=0}^{N-1} sin(\frac{2\pi}{N}nk)^2).$$

Wegen

$$\sum_{n=0}^{N-1} \cos\left(\frac{2\pi nk}{N}\right)^2 = \begin{cases} N, & k = l \cdot \frac{N}{2}, l \in \mathbb{Z} \\ \frac{N}{2}, & sonst \end{cases}$$

sowie

$$\sum_{n=0}^{N-1} \sin\left(\frac{2\pi nk}{N}\right)^2 = \begin{cases} 0, & k = l \cdot \frac{N}{2}, l \in \mathbb{Z} \\ \frac{N}{2}, & sonst \end{cases}$$

folgt damit insgesamt

$$\begin{array}{l} X_{Im}(k) = 0 \\ X_{Re}(k) \sim \mathcal{N}\left(0, \sigma^2 N\right) \end{array} \quad \text{für } k = l \cdot \frac{N}{2}, l \in \mathbb{Z}$$

sowie für alle anderen k

$$\begin{array}{l} X_{Im}(k) \sim \mathcal{N}\left(0, \sigma^2 \frac{N}{2}\right) \\ X_{Re}(k) \sim \mathcal{N}\left(0, \sigma^2 \frac{N}{2}\right) \end{array}.$$

Damit ist die Verteilung der einzelnen DFT-Koeffizienten bestimmt. Insbesondere ist diese also, bis auf den nullten und $L/2$-ten Koeffizienten, unabhängig vom Parameter k. Der Real- und Imaginäranteil sind ferner jeweils statistisch unabhängig. Dies folgt aus [LK54] nebst der Identität

$$\sum_{n=0}^{N-1} \sin\left(\frac{2\pi nk}{N}\right)\cos\left(\frac{2\pi nk}{N}\right) = 0$$

und gilt für alle $k \in \{0,\dots,N-1\}$. Bis auf die Symmetrie $X(k) = \overline{X(N-k)}$ der DFT-Koeffizienten, die für die DFT reeller Signale gilt und ein deterministischer Zusammenhang zwischen $X(k)$ und $X(N-k)$ ist, sind die DFT-Koeffizienten paarweise statistisch unabhängig. Dies ergibt sich wiederum aus der Darstellung

$$\underbrace{\begin{pmatrix} X_{Re}(3) \\ X_{Im}(3) \\ \vdots \\ X_{Re}(\frac{N}{2}) \end{pmatrix}}_{=:Y} = \underbrace{\begin{pmatrix} \cos(\frac{2\pi}{N}0\cdot 3), \cos(\frac{2\pi}{N}1\cdot 3), \dots, \cos(\frac{2\pi}{N}(N-1)\cdot 3) \\ \sin(\frac{2\pi}{N}0\cdot 3), \sin(\frac{2\pi}{N}1\cdot 3), \dots, \sin(\frac{2\pi}{N}(N-1)\cdot 3) \\ \vdots \\ \cos(\frac{2\pi}{N}0\cdot \frac{N}{2}), \cos(\frac{2\pi}{N}1\cdot \frac{N}{2}), \dots, \cos(\frac{2\pi}{N}(N-1)\cdot \frac{N}{2}) \end{pmatrix}}_{=:A} \underbrace{\begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix}}_{=:x}.$$

Hierbei wurde $X_{Im}(\frac{N}{2})$ nicht berücksichtigt, da diese Komponente identisch Null ist und, erweiterte man die Matrix A (und dann auch Y) um eine entsprechende Zeile, A nicht mehr vollen Rang haben würde, was eine technische Konsequenz für die Anwendung der gleich folgenden Aussage hätte. Des Weiteren wurde mit $k = 3$ begonnen, da dies der erste vom Advanced Combination Encoder berücksichtigte DFT-Koeffizient ist.

Es gilt nun folgende Aussage [Pri12]: Ist Z ein $Kx1$ multivariat normalverteilter Zufallsvektor und B eine LxK Matrix mit vollem Rang, so ist $T = BZ$ ein multivariat normalverteilter Zufallsvektor mit Kovarianzmatrix[1] $Var(Y) = BVar(X)B^T$.

Folglich ist Y wie oben definiert multivariat normalverteilt mit Kovarianzmatrix $AVar(x)A^T$. $Var(x)$ ist eine Diagonalmatrix, da die x_i nach Definition statistisch unabhängig und gleichverteilt sind, sodass $AVar(x)A^T = Var(x)AA^T$ ist. Für gerades N, also eine gerade DFT-Länge, die im Falle des Advanced Combination Encoders gegeben ist, ist AA^T eine Diagonalmatrix und somit auch die Kovarianzmatrix $Var(Y)$ von Y. Damit folgt aus der Normalverteiltheit von Y die statistische Unabhängigkeit der Komponenten von Y.

Möchte man diese Aussage noch um $X_{Im}(\frac{N}{2})$ erweitern, so kann man die Verbunddichte $f_Y(y_1,\dots,y_{\frac{N}{2}-2})$ von Y mit $\delta(u)$, der Deltadistribution, multiplizieren, um die Verbunddichte von $(Y, X_{Im}(\frac{N}{2}))$ zu erhalten.

1 Der Erwartungswert ist hier nicht relevant

Qua Konstruktion ergibt sich damit die statistische Unabhängigkeit von $(Y, X_{Im}(\frac{N}{2}))$ wegen der Faktorisierung der Verbundwahrscheinlichkeit. Möchte man ohne Nutzung der Deltadistribution auskommen, so müsste eine gemischte Verteilung betrachtet werden, d.h. mit kontinuierlichem und diskreten Anteil. Dies ist aber nur eine Frage der Rigorosität.

Reemt Hinrichs

Personal Data

Year of Birth:	1986
Place of Birth:	Hannover, Germany
Email:	hinrichs@tnt.uni-hannover.de

Work Experience

2018–	Wissenschaftlicher Mitarbeiter Institut für Informationsverarbeitung, Leibniz University Hannover
2012–2017	Studentische Hilfskraft Institut für Informationsverarbeitung, Leibniz University Hannover
2016–2017	Studentische Hilfskraft Institut für Fertigungstechnik und Werkzeugmaschinen, Leibniz University Hannover

Ausbildung

2018-2023	Dr.-Ing. Leibniz Universität, Hannover
2015-2017	Mechatronik, M. Sc. Leibniz Universität, Hannover
2007-2015	Mechatronik, B. Sc. Leibniz Universität, Hannover
2006	Abitur Goetheschule, Hannover