Aus der Klinik für Hals-Nasen-Ohrenheilkunde, Kopf- und Halschirurgie
der Medizinischen Fakultät Mannheim
(Direktorin: Prof. Dr. med. Nicole Rotter)

# Auswertung der Sprachaudiometrie mittels Spracherkennungssoftware, basierend auf dem Oldenburger Satztest

Inauguraldissertation
zur Erlangung des medizinischen Doktorgrades
der
Medizinischen Fakultät Mannheim
der Ruprecht-Karls-Universität
zu
Heidelberg

vorgelegt von
Christian Warken

aus
Mannheim
2021

Dekan: Herr Prof. Dr. med. Sergij Goerdt
Referentin: Frau Priv.-Doz. Dr. med. Angela Schell

# INHALTSVERZEICHNIS

Seite

# VERZEICHNISSE

## 1.1 Abkürzungsverzeichnis

DIN – deutsche Industrienorm

EZF – Echtzeitfaktor

HMM – Hidden-Markov-Modelle

HNO – Hals-Nasen-Ohrenheilkunde

HSM – Hochmair-Schulz-Moser Test

KI – Konfidenzintervall

MTA – medizinisch-technische Assistentin

OLSA – Oldenburger Satztest

PC – Personal Computer

SSD– single sided deafness (einseitige Taubheit)

WA – Wortakkuratheit

WR – Worterkennungsrate

## 1.2 Tabellen- und Abbildungsverzeichnis

# 2 EINLEITUNG

## 2.1 Bedeutung der menschlichen Sprache

Die menschliche Sprache ist ein wesentliches Element der Kommunikation und Interaktion innerhalb sozialer Gesellschaften. Eine gemeinsame Sprache ist zudem Teil der kulturellen Identität.

Die Diagnose, Quantifizierung und Auswertung von Hörminderungen, also den Störungen beim Empfangen und der kognitiven Verarbeitung der gesprochenen Sprache, schließt daher auch routinemäßig die Sprachtests, mit oder ohne Störgeräusch, ein. Im Hinblick auf die Begutachtung von verschiedenen Graden der Schwerhörigkeit gilt die Sprachaudiometrie als unentbehrlicher Bestandteil und als eine entscheidende Diagnostik für die quantitative Bemessung eines möglichen Hörschadens. Auch bei der Anpassung von Hörhilfen ist ein möglichst gutes Verstehen von Sprache der entscheidende Faktor für das Gelingen der Versorgung und die langfristige Akzeptanz einer Hörgeräteversorgung.

## 2.2 Bedeutung der Sprachaudiometrie in der Klinik

Sprachtests werden schon sehr lange verwendet. Die ersten Sprachtests wurden bereits 1804 durch Georg-Wilhelm Pfingsten in Kiel durchgeführt [1]. Sie haben eine hohe klinische Relevanz für die Versorgung mit Hörgeräten oder die Einschätzung von Graden der Arbeitsfähigkeit und Behinderung.

Zur Überprüfung der Sprachverständlichkeit stehen im deutschsprachigen Raum eine große Anzahl verschiedener Testverfahren zur Verfügung, die auf unterschiedlichen Testmaterialien basieren. Die wichtigsten Sprachverständlichkeitstests testen das Verstehen von Einsilbern, Mehrsilbern oder ganzer Sätze, diese können sinnhaft oder nicht sinnhaft seien. Viele dieser Sprachtests sind für Bedingungen in Ruhe und/oder im Störgeräusch geeignet.

In den Sprachwissenschaften bezieht sich die Sprachverständlichkeit auf die phonetisch-phonologische Struktur eines Worts oder eines Satzes, wohingegen

Sprachverständnis und -verstehen mit der Erfassung von Sinn und Bedeutung von Wörtern und Sätzen zu tun haben[2].

Die präzise Unterscheidung zwischen diesen Begriffen verdeutlicht die Vielschichtigkeit der genannten Messgrößen. Daraus wird ersichtlich, dass es nahezu unmöglich sein wird, ein Prüfverfahren zu finden, das alle Aspekte des korrekten Empfangs menschlicher Sprache und der kognitiven und assoziativen Verwertung von lautsprachlichen Signalen zu testen vermag.

### 2.2.1 Übersicht aktueller Sprachtests in der Klinik

Ein wesentlicher Fortschritt der deutschen Sprachaudiometrie erfolgte 1953 mit dem Freiburger Sprachtest durch Karl-Heinz Hahlbrock. Der Freiburger Sprachtest ist aktuell der in Deutschland am häufigsten verwendete Wörtertest (Freiburger Sprachaudiogramm), weshalb hier zunächst kurz auf diesen Test eingegangen wird auch wenn er in unserer Studie nicht verwendet wurde [3]. Die Freiburger Sprachaudiometrie besteht aus zwei Arten von Worten. Zum einen aus zweistelligen, meist viersilbigen Zahlwörtern (10 Gruppen zu je 10 Zahlen) und zum anderen aus einsilbigen Hauptwörtern (20 Gruppen zu je 20 Wörtern). Zahlwörter wie 22 oder 35 sind leichter verständlich, da es nur eine stark begrenzte Anzahl von ihnen gibt. Einsilbige Hauptwörter wie Hund oder Bach können nur dann korrekt nachgesprochen werden, wenn auch jeder Laut erkannt wurde.

Vorteile des Freiburger Wörtertests sind gut reproduzierbare Bedingungen, da die einzelnen Gruppen untereinander annähernd phonetisch gleich sind und das Testmaterial von einem geschulten Sprecher auf Band gesprochen wurde. Ein weiterer Vorteil ist die jahrzehntelange Erfahrung mit diesem Testverfahren. Er wurde zudem 1973 in einer Industrienorm fixiert (DIN 45621-1) [4].

Nachteilig sind die Beschränkung auf Einsilber und Zahlwörter und das begrenzte Testmaterial von 100 Zahlwörtern beziehungsweise 400 Einsilbern, die dadurch statistisch begrenzte Aussagekraft und die Unausgewogenheit des Testmaterials. Vorherige Studien zeigten signifikante Unterschiede der Test-Retest-Reliabität zwischen den verschiedenen 20 einsilbigen Wortlisten [5]. Die einzelnen Listen sind im Hinblick auf ihre Verständlichkeit nicht gleichwertig, sondern teilweise

unausgewogen [6]. Jedoch führten die verschiedenen Studien zu inkongruenten Ergebnissen und einzelne Listen können in dieser Hinsicht nicht pauschal als besser oder schlechter im Vergleich zu anderen Wortlisten bewertet werden. Weiterhin ist der Test nicht automatisiert und es existieren keine genauen Vorgaben zum Ablauf des Tests, wie beispielsweise die Geschwindigkeit der Wiedergabe. Es ist daher fraglich wie verlässlich die Aussagekraft der Freiburger Sprachaudiometrie zu bewerten ist.

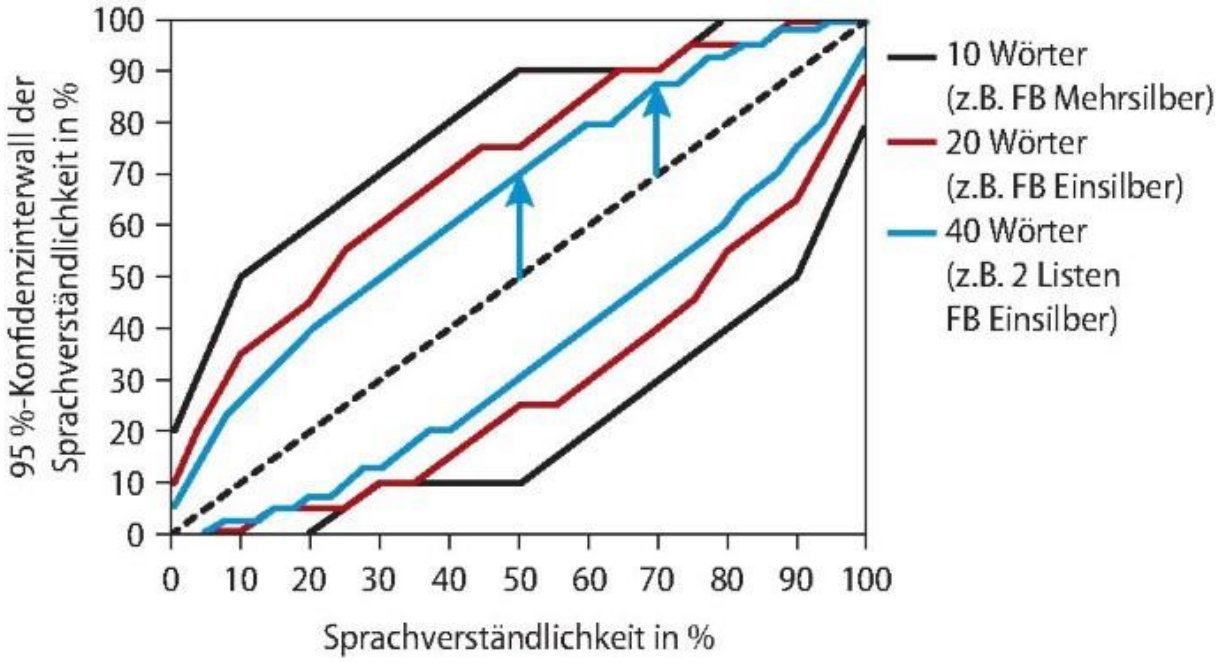


Abbildung 1: Grenzen für das 95 %-Konfidenzintervall für sprachaudiometrische Test-/Retest-Messungen in Abhängigkeit von der gemessenen Sprachverständlichkeit für Testlisten mit je 10, 20 oder 40 Testwörtern [7]

### 2.2.2 Vorteile des Oldenburger Satztests im Vergleich zu anderen Sprachtests

Hörgeschädigte Personen haben beim Sprachverständnis vor allem in lauter Umgebung mit Störgeräuschen weitaus größere Probleme als Normalhörenden. Um diesen Umstand realistisch zu erfassen, wird mit dem Oldenburger Satztest (OLSA) bei schwerhörigen Patienten die Sprachverständlichkeit mit oder ohne Hörgeräte im Störgeräusch geprüft, wobei ganze Sätze als Nutzschall dargeboten werden [8-10]. Als Störgeräusch wird ein sprachsimulierendes Rauschen mit einem festgelegten Schallpegel von 65 dB verwendet. Nutzschall und Störschall werden über getrennte Lautsprecher angeboten. Bei Messungen ohne Störgeräusch liegt der Sprachpegel zu Beginn bei 30 dB und passt sich nach einem vorgegebenen Schema an. Die Messungen werden als binaurale Freifeldmessungen durchgeführt.

Der OLSA ist ein adaptiver Sprachtest, der Pegel der Sprache wird hierbei entsprechend der Antwort der Testperson verändert. Das Sprachmaterial besteht aus 40 Testlisten von je 30 Sätzen mit einer zufälligen Kombination aus einem Inventar von insgesamt 50 Wörtern. Dadurch entsteht ein weitaus größeres Inventar des Testmaterials im Vergleich zum Freiburger Wörtertest. Es handelt sich um syntaktisch korrekte, aber semantisch unvorhersagbare Sätze, die immer einen Namen, ein Verb, ein Zahlwort, ein Adjektiv und ein Objekt in der genannten Reihenfolge beinhalten, z. B. „Thomas gibt acht schwere Schuhe“. Die Sätze sind meist sinnlos, sie sind somit nicht so leicht zu merken und es kann daher mit ihnen wiederholt gemessen werden. Das Sprachmaterial wird jeweils von einer Männerstimme gesprochen.

Satztests, wie der OLSA, sind grundsätzlich leichter zu verstehen als Einsilbertests, wie der Freiburger Wörtertest. In der klinischen Praxis wird bei einer diagnostizierten Schwerhörigkeit möglichst frühzeitig eine Therapie, meist die Versorgung mit Hörgeräten, angestrebt. Werden Messungen jedoch bei normaler Gesprächslautstärke durchgeführt, werden bei leichtgradiger und selbst mittelgradiger Schwerhörigkeit in der Regel fast 100% der Sätze eines Satztestes erkannt. Die entscheidende Fragestellung ist hierbei jedoch die Schwelle der 50%-Verständlichkeit. Daher wird der OLSA in der Praxis hauptsächlich für die adaptive Sprachaudiometrie im Störgeräusch verwendet. Eine weitere Problematik besteht darin, dass hochgradige hörgeschädigte Personen teilweise nicht in der Lage sind die 50%-Verständlichkeitsschwelle zu erreichen. Der Untersucher kann dies im Voraus häufig nicht sicher wissen. In solchen Fällen ist der Freiburger Wörtertest klar im Vorteil und meist der einzige sinnvoll anwendbare Sprachtest. Den perfekten Sprachtest gibt es also nicht und wird es wahrscheinlich auch in Zukunft nicht geben [11].

Zusammenfassend lässt sich sagen, dass der Freiburger Sprachtest durch seine häufige Verwendung und die Erwähnung in der audiometrischen Fachliteratur gut bekannt und einfach verfügbar ist. Zudem lassen sich seine Ergebnisse nach bekannten Zusammenhängen interpretieren. Da in den Universitäts-HNO-Kliniken aber modernere Sprachtest seit längerer Zeit verwendet werden, schlagen sich diese auch zunehmend in der Facharztausbildung nieder und gewinnen damit stetig an Bekanntheit. Das gilt mittlerweile für neue und in mancher Hinsicht bessere Testverfahren wie den Einsilber-Reimtest oder Satztests im Störgeräusch. Das ist einer der Gründe, weshalb der Fokus dieser Studie auf dem Oldenburger Satztest liegt.

Die Richtlinie des Gemeinsamen Bundesausschusses über die Verordnung von Hilfsmitteln in der vertragsärztlichen Versorgung sieht in § 21 Abs. 3 und § 22 Abs. 3 mittlerweile zudem auch Sprachtests wie den Göttinger oder Oldenburger Satztest im Störgeräusch für die Erfolgskontrolle der Hörgeräteversorgung optional vor [12]. Auch hier liegen einfach zu interpretierende Grenzwerte zugrunde, nämlich eine Verbesserung der Sprachverständlichkeitsschwelle im Störgeräusch um mehr als 2 dB - im Sinne einer Verringerung des „signal to noise ratio“ um 2 dB bei einer beidohrigen Hörgeräteversorgung. Aktuell sind für die Indikationsstellung einer Hörgeräteversorgung (noch) keine Satztests im Störgeräusch vorgesehen. Eine Hörgeräteindikation und Erfolgskontrolle allein aufgrund von Einsilbern in Ruhe vernachlässigen hierbei jedoch das wesentlichste Problem der meisten Innenohrschwerhörenden, nämlich das verminderte Sprachverstehen im Störgeräusch. Eine zuverlässige und reliable Interpretation der Ergebnisse im Hinblick auf die tatsächliche Einschränkung der Kommunikation im Alltag - der zentrale Grund, weswegen hörgeschädigte Personen Hilfe suchen - kann mit der Verständlichkeit von isoliert gesprochenen Einsilbern in perfekter Ruhe praktisch nicht erfolgen. Die für den Patienten so wichtigen Störgeräuschunterdrückungs-funktionen in den Hörgeräten lassen sich zudem auch nur im Störgeräusch zuverlässig bewerten und quantifizieren [12]. Bei den heute zur Verfügung stehenden, wie dem OLSA, kommen Störsignale zum Einsatz, die die Eigenschaften lebender Sprache aufweisen und somit einem Sprachverstehens im Stimmengewirr näherkommen.

Die Messgenauigkeit eines Sprachtests hängt wesentlich von der Steigung s50 der Diskriminationsfunktion in ihrem Wendepunkt. Je größer diese Zahl ist, desto genauer kann die Sprachverständlichkeitsschwelle L50 bestimmt werden und desto empfindlicher ist der Test für Änderungen dieser Schwelle. Die Steigung s50 beträgt 8 % pro dB für die Zahlwörter und 5 % pro dB für die Einsilber. Für den OLSA in Ruhe beträgt die Steigung 11 % pro dB, im Störgeräusch liegt dieser Wert bei 17% pro dB [13, 14]. Hey et. al zeigten 2003 die Unterschiede der Steigung der Diskriminationsfunktion von Hochmair-Schulz-Moser Test (HSM Satztest) und OLSA:

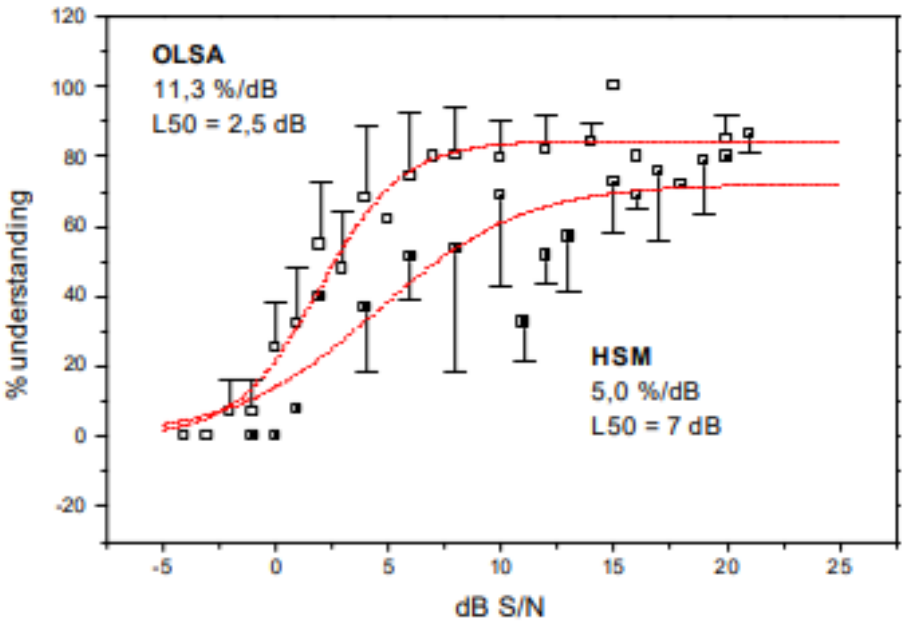


Abbildung 2: Mittelwert individueller Diskriminationsfunktionen für den OLSA und HSM Satztest mit Bestimmung des L50 und der Steigung am L50 [15]

Die Steigung spielt beispielsweise bei der Therapiekontrolle der Hörgeräteversorgung eine entscheidende Rolle. Mit Satztests, wie dem OLSA; können daher Unterschiede in der Störgeräuschunterdrückung bei Hörgeräten wesentlich genauer nachgewiesen werden als mit dem Einsilbertests. Die direkte Vergleichbarkeit verschiedener Sprachverstehenstests ist jedoch besonders bei unterschiedlichen Messweisen nur eingeschränkt möglich

Zufällige Einflüsse auf das Testergebnis, wie kurzzeitige Konzentrationsschwankungen, Ablenkungen oder die Vertrautheit mit bestimmten Testwörter führen dazu, dass nur große Ergebnisdifferenzen tatsächlich signifikant und nicht nur zufällig sind [16]. Grundsätzlich steigt die Messgenauigkeit im Sinne der Test-Retest-Reliabilität mit der Anzahl getesteter Testwörter an. Dies ist auf Abbildung 1 schematisch dargestellt, wo Abweichungen innerhalb des Konfidenzintervalls bis zu einem Unterschied des Sprachverständnis von 20% einschließen können. Einen bedeutenden Vorteil bieten aus diesem Grund Satztestverfahren, die wie im Fall des OLSA 150 Wörter verwenden und dementsprechend sehr viel genauere Ergebnisse liefern als Sprachtests die Einzelwörter abfragen. Zwar ist auch beim Freiburger Sprachtest die Verwendung von z. B. 100 Wörtern pro Messung möglich. Dann werden allerdings alle Zahlen und ein Viertel der Einsilber pro Messung „verbraucht“ und damit sind Verlaufskontrollen nicht mehr sinnvoll durchführbar. Gleichzeitig steigt der

Zeitbedarf und damit die Anforderung an die Konzentrationsfähigkeiten von Testperson und Testleiter an je länger ein Sprachtest dauert.

### 2.2.3 Automatisierung in der Hördiagnostik

Die Automatisierung der Hördiagnostik hat sich in den letzten Jahren deutlich weiterentwickelt. Für die automatische (Reinton-)Audiometrie liegen bereits ausreichende Validierungsdaten vor und sie kann heutzutage bereits ein genaues Maß der Hörschwelle ermitteln. Für andere audiometrische Verfahren sind die Validierungsdaten weiterhin begrenzt und es bedarf weiterer Forschung [17]. Zudem besteht eine große Diskrepanz zwischen dem Bedarf und der Kapazität von Audiologen für die Durchführung von Hörtests. Durch eine Automatisierung könnten erhebliche Zeitersparnisse entstehen. Eine Ökonomisierung der Arbeitszeiten könnte dafür sorgen, dass mehr Menschen versorgt werden könnten [18]. Zurzeit müssen die Audiologen, die den Sprachtest durchführen, die Antwort des Patienten hören und verstehen, um diese dann in der Software als „falsch“ oder „richtig“ zu definieren. Die Überlegung ist, dies unterstützend von einer Spracherkennungssoftware durchzuführen zu lassen.

Ziel dieser Dissertation ist zu zeigen, ob die Auswertung des OLSA mit der Spracherkennungssoftware Dragon von Nuance Inc. mindestens ebenbürtig zu einer manuell durchgeführten Sprachaudiometrie ist. Zukünftig geplant ist die qualitative Weiterentwicklung und gegebenenfalls eine Automatisierung der Sprachaudiometrie. Erreicht werden soll, dass mehr Patienten von dieser Diagnostik profitieren können. Diese wäre unter Umständen genauer und weniger fehleranfällig als bisher, würde die Wartezeiten verkürzen, das Personal des Diagnostikteams entlasten und eventuell die Kosten für Leistungserbringer und Krankenkassen reduzieren.

## 2.3 Spracherkennung

### 2.3.1 Grundlagen

Die moderne Spracherkennung kann aktuell in zwei Unterarten aufgeteilt werden, sprecherunabhängige und sprecherabhängige Spracherkennung.

Charakteristisch für die sprecherunabhängige Spracherkennung ist die Eigenschaft, dass der Anwender direkt mit der Spracherkennung beginnen kann, ohne das System vor der ersten Verwendung trainieren zu müssen. Der Wortschatz ist dabei jedoch auf wenige tausend Wörter begrenzt. Sprecherunabhängige Sprach-erkennung wird daher vor allem dort erfolgreich verwendet, wo nur ein begrenzter Wortschatz benötigt wird, zum Beispiel in automatischen Dialogsystemen wie dem Tele-Banking. So erreichen Systeme zur Erkennung der gesprochenen englischen Ziffern von 0 bis 9 eine nahezu 100-%-Erkennungsquote [19].

Sprecherabhängige Spracherkennungsprogramme werden vom Benutzer vor oder während der Anwendung auf besondere Eigenarten der jeweiligen Aussprache trainiert. Ein zentrales Element ist die individuelle Interaktionsmöglichkeit mit dem Anwendungssystem, um ein optimales sprecherabhängiges Ergebnis zu erzielen. Ein Einsatz in Anwendungen mit häufig wechselnden Benutzern ist daher nicht sinnvoll. Die Größe des Wortschatzes ist im Vergleich zu sprecherunabhängigen Systemen weitaus ausgeprägter. So enthalten moderne Spracherkennungsprogramme, welche sprecherabhängig arbeiten, mehr als 300.000 Wörter und Wortformen. In Abhängigkeit von der Erfahrung des Anwenders und dem Umfang des Trainings mit dem System können auch mit sprecherabhängigen Spracherkennungsprogrammen sehr hohe Erkennungsquoten erreicht werden. Abhängig von der Fragestellung kann selbst eine Treffsicherheit von 95 Prozent als ungenügend empfunden werden, da zu viel Aufwand für die notwendige Nachbesserung bestehen würde [20]. Ausschlaggebend für das Outcome sprecherabhängiger Spracherkennung ist die Interaktion zwischen Nutzer und System. Die Entwicklung im Bereich der Spracherkennungssysteme schreitet aktuell sehr schnell voran. Die neuste Generation sprecherabhängiger Spracherkennungssysteme müssen nicht mehr zwangsläufig trainiert werden [21].

Moderne Systeme können beim Diktat von Fließtexten eine korrekte Erkennungsquoten von ca. 99 Prozent erreichen und können somit für viele Einsatzgebiete die benötigten Anforderungen erfüllen. Probleme und Fehler entstehen vor allem dort, wo der Verwender häufig neue, dem System noch nicht bekannte Worte oder Wortformen verwendet. Daher sollten auch sprecherabhängige Systeme vorzugsweise dort verwendet werden, wo ein (wenn auch größerer) standardisierter Fachwortschatz verwendet wird, z.B. in der Medizin.

Zusätzlich unterscheidet man zwischen Back-End-Systemen und Front-End-Systemen. In Front-End-Systemen erfolgt die Verarbeitung der gesprochenen Eingabe ohne nennenswerte Verzögerung. Die Umsetzung in einen Text erfolgt innerhalb weniger Sekunden und das Ergebnis kann sofort abgelesen werden. Die Umsetzung erfolgt in der Regel direkt auf dem Computer des Benutzers. Durch die direkte Interaktion zwischen Benutzer und den Spracherkennungsprogramm wird eine hohe Erkennungsqualität erzielt. In Back-End-Systemen wird die Umsetzung hingegen über einen externen Server durchgeführt und steht daher erst mit einer gewissen Zeitverzögerung zur Verfügung.

Neben dem Umfang und der Flexibilität des zugrunde liegenden Wortschatzes des Spracherkennungssystems, spielt auch die akustische Qualität des Inputs eine entscheidende Rolle. Bei Mikrofonen, die direkt vor dem Mund lokalisiert sind, wird eine signifikant höhere Erkennungsgenauigkeit erreicht als bei weiter entfernten Raummikrofonen [22]. Die wichtigste Voraussetzung ist hierbei zweifelslos eine präzise, deutliche und flüssige Aussprache, so dass Wortzusammenhänge und Wortfolgewahrscheinlichkeiten optimal in den Verarbeitungsprozess der Worterkennung einfließen können.

### 2.3.2 Verarbeitung und deren Herausforderungen

Bei der Verarbeitung von aufgenommener oder diktierter Sprache durch ein Spracherkennungssystem sind einige Herausforderungen zu beachten.

Zum einen muss primär erkannt werden, wann ein einzelnes Wort beginnt und endet. Daher muss das Spracherkennungssystem sicher sein, welche Silbe welchem Wort zuzuordnen ist. In der Alltagssprache werden einzelne Wörter üblicherweise ohne wahrnehmbare Pause ausgesprochen. Menschen können sich jedoch intuitiv an den Übergängen zwischen den Wörtern orientieren. Frühere Spracherkennungssysteme benötigten eine unterbrochene Sprache, auch diskrete Sprache genannt, bei der zwischen den Wörtern im normalen Redefluss unübliche Pausen gemacht werden. Moderne Systeme sind mittlerweile auch in der Lage eine fließende Aussprache, auch kontinuierliche Sprache genannt, zu erkennen und zu verarbeiten [23].

Zusätzlich spielt die Größe des verwendeten Wortschatzes für die Zuverlässigkeit der Spracherkennung eine wichtige Rolle. Durch die Flexion eines Wortes je nach grammatikalischer Funktion, können aus einem Wortstamm eine Vielzahl von Wortformen entstehen. Dies ist für die Größe des Wortschatzes von großer Bedeutung, da alle Wortformen bei der Spracherkennung als eigenständige Wörter betrachtet werden müssen, um sie sicher auswerten zu können.

Die Größe des Wörterbuchs hängt daher stark von der Sprache ab. Zum einen haben durchschnittliche deutschsprachige Sprecher mit circa 4.000 Wörtern einen deutlich größeren Wortschatz als englischsprachige mit rund 800 Wörtern. Außerdem ergeben sich durch die Flexion in der deutschen Sprache in etwa 40.000 Wortformen, wie in der englischen Sprache, wo nur etwa 3.000 Wortformen entstehen. Dies hängt vor allem mit der Anzahl und Verwendung von zusammengesetzten Kompositionswörtern zusammen [24]. Dies zeigt sich am Beispiel von “Spracherkennungssystem” im Vergleich zu “speech recognition system”, im deutschen wird aus den Worten “Sprache”, “Erkennung” und “System” ein neues Wort gebildet, während im englischen mit drei bereits vorhandenen Wörtern das gleiche benannt wird, ohne eine Wortneuschöpfung zu bilden.

In vielen Sprachen gibt es Wörter oder Wortformen, die gleich ausgesprochen werden, jedoch eine komplett andere Bedeutung haben. Hierfür gibt es in der deutschen Sprache viele Beispiele. So klingen die Wörter „Leere“ und „Lehre“ zwar identisch, haben jedoch trotzdem nichts miteinander zu tun. Diese Wortpaare nennt man Homophone. Da sprachkundige Menschen üblicherweise die Bedeutung der jeweiligen beiden Homophone kennen, können sie die Möglichkeiten anhand der Bedeutung oder dem Zusammenhang unterschieden. Heutige Spracherkennungs-systeme sind hierzu noch nicht in der Lage. Für die Unterscheidung von Groß- oder Kleinschreibung, gilt das gleiche. Beispiele hierfür sind das Substantiv "Lehre” und dem Verb “lehre” oder das Substantiv “Laut” und das Adjektiv “laut”. Daher sollte eine Kontextprüfungen erfolgen, um Homophone unterscheiden zu können. Anhand von Statistiken über die Häufigkeit bestimmter Wortkombinationen, kann man bei ähnlich oder gleich klingenden Wörtern relativ sicher entscheiden, welches tatsächlich gemeint ist.

Vokale sind mit einem Spracherkennungssystem leicht erkennbar. Dies hängt zum einen damit zusammen, dass die Anzahl der Vokale überschaubar ist. Zum anderen

konzentriert sich das Frequenzspektrum gesprochener Vokale typischerweise auf bestimmte unterschiedliche Frequenzen, die Formanten genannt werden. Bei der Spracherkennung spielt insbesondere die Lage der Formanten eine Rolle. Die tiefere Frequenz liegt im Bereich von 200 bis 800 Hertz, die höhere im Bereich von 800 bis 2400 Hertz. Über die Zuordnung zu diesen Frequenzen lassen sich die einzelnen Vokale normalerweise gut unterscheiden [25].

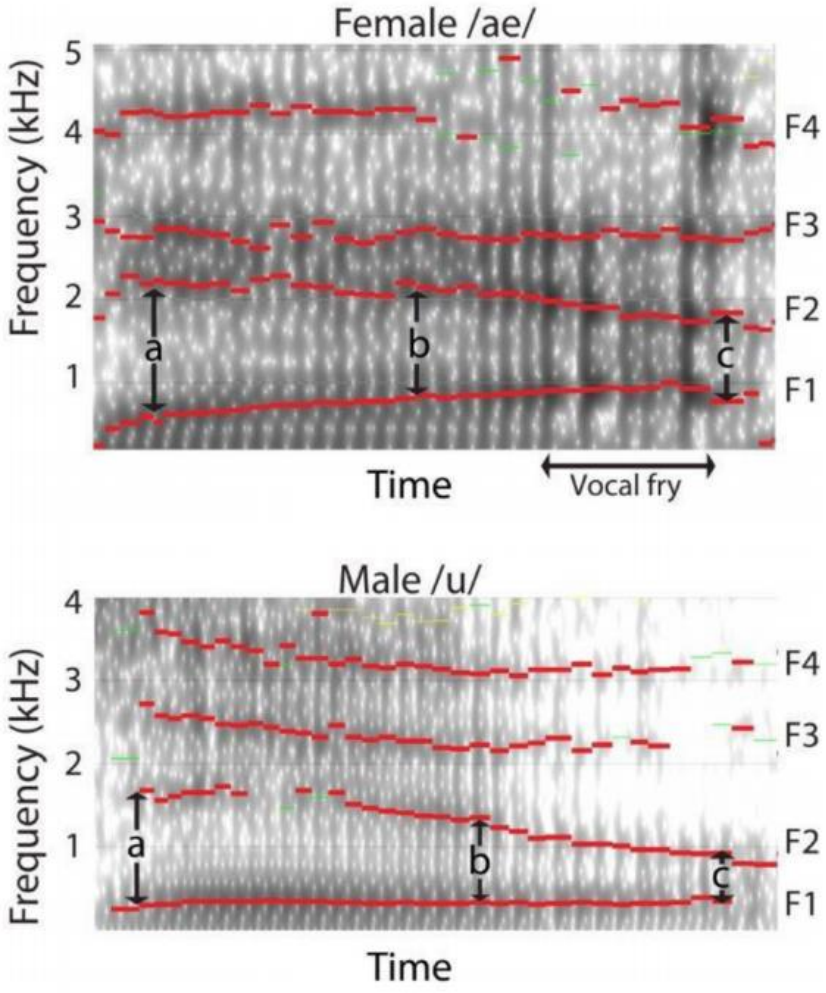


Abbildung 3: Spektrogramm der menschlichen Stimme. Die waagerechten Frequenzbänder sind die mit roten Linien markierten Formanten [26]

Die Konsonanten sind weitaus schwierig zu erkennen. Manche Konsonanten sind beispielsweise nur durch den Übergang zu den benachbarten Lauten feststellbar. Diese Konsonanten werden Plosivlaute genannt, zu ihnen gehören b, d, g, k, p und t. Andere Konsonanten sind im Gegensatz dazu leicht an charakteristischen spektralen akustischen Mustern erkennbar. So zeichnen Reibelaute durch einen hohen Energieanteil in höheren Frequenzbändern aus. Zu diesen Lauten zählen die Buchstaben c, f, h, r, s, v, x und z. Die relevanten Informationen zur Unterscheidung

mancher Reibelaute werden teilweise außerhalb des in Telefonnetzen übertragenen Spektralbereichs vermittelt liegt. Damit ist auch zu erklären, dass das Buchstabieren über Telefon auch in der Kommunikation zwischen zwei normalhörenden, sprachkundigen Menschen durchaus mühselig und fehleranfällig ist. Zudem kennt die deutsche Sprache noch Nasallaute (m, n), Approximanten (j, l) und Affrikaten (tz, pf), Verbindung eines Plosivlautes mit einem Reibelaut [27-29].

Grundsätzlich kann man Laute auf der Basis einer abnehmenden Verengung des Vokaltrakts folgendermaßen sortieren: Plosive, Affrikaten, Frikative, Approximanten und Vokale [30]. Je enger der Vokaltrakt, desto schwerer ist die Unterscheidung der möglichen Laute.

In der Regel sind Spracherkennungsprogramme auf eine Hochsprache, in Deutschland üblicherweise Hochdeutsch, eingestellt. Solange der Anwender die Hochsprache spricht, liegt die Erkennungsquote bei fast 100%. Dies bedeutet jedoch nicht, dass das verwendete Spracherkennungsprogramm jede Ausformung dieser Sprache verstehen kann. Bei der Verwendung von Dialekten und Soziolekten stoßen die Programme häufig an ihre Grenzen und die Erkennungsquote kann drastisch abfallen. Menschen sind in einer Situation, in der sie mit der Sprachform nicht vertraut sind, meist in der Lage, sich schnell auf die unbekannte Mundart ihres Gegenübers einzustellen oder Sprachinhalte aus dem Zusammenhang zu erkennen. Spracherkennungssoftwares ist dazu aktuell nicht ohne weiteres in der Lage. Dialekte müssen dem Programm hierfür erst in aufwendigen Prozessen beigebracht werden. Zusätzlich muss bedacht werden, dass sich Wortbedeutungen je nach Dialekt verändern können. Ein Mensch kann Probleme hierbei durch sein Hintergrundwissen leicht vermeiden. Softwares sind hierfür meist nicht in der Lage.

Sollte es in einem Gespräch zu Problemen mit dem Verständnis von Wörtern kommen, versuchen Menschen üblicherweise besonders laut zu sprechen oder missverstandene Begriffe zu ändern, zu beschreiben oder Synonyme zu verwenden. Dies kann jedoch bei der Verwendung von Spracherkennungsprogrammen zu weiteren Problemen führen, da diese auf eine normale Gesprächslautstärke trainiert ist und zum korrekten Erkennen von Worten eher mit Schlüsselwörtern arbeiten, als sinnvolle Zusammenhänge zu erfassen.

Zusätzliche Probleme können auch durch sogenannte Polyseme, Homonyme oder Homophone entstehen. Polyseme Wörter sind mehrdeutig, beschreiben also mehrere mehr oder minder unterschiedliche Sachverhalte, die sich aus einem gemeinsamen Kontext entwickeln, z.B. “Schild”, das einerseits ein Informationszeichen und andererseits eine Schutzvorrichtung sein kann. “Polysemie gilt als natürlichsprachlicher Normalfall und als Ausdruck des sprachlichen Ökonomie-Prinzips” [31]. Der Begriff Homonym ist das Gegenstück zum Synonym, Homonyme stehen für denselben sprachlichen Ausdruck für verschiedene Begriffe, Synonyme stehen für verschiedene sprachliche Ausdrücke für denselben Begriff. Ein Beispiel hierfür wäre “Kiefer”, dies wäre entweder eine Baumart oder ein Teil des Schädels. Die Unterscheidung zwischen Polysemen und Homonymen ist nicht immer eindeutig. Homophone sind Wörter, die die gleiche Aussprache wie ein anderes mit unterschiedlicher Bedeutung hat, beispielsweise “Hertz” und “Herz” [32, 33]. Diese Herausforderungen können sowohl von Menschen als auch von Spracherkennungsprogrammen nur mit ausreichendem Training und Vorwissen bewältig werden.

### 2.3.3 Aufbau von Spracherkennungssystemen

Ein Spracherkennungssystem besteht stets aus mehreren Bestandteilen. Zunächst findet eine Vorverarbeitung statt, die die analogen Sprachsignale in die einzelnen Frequenzen zerlegt. Anschließend findet die tatsächliche Erkennung mit Hilfe akustischer Modelle, Wörterbücher und Sprachmodellen statt.

Die Vorverarbeitung besteht aus mehreren Schritten. Bei der Abtastung wird das analoge Signal digitalisiert, also in eine Bitfolge zerlegt. Dadurch ist es anschließend einfacher zu verarbeiten. Bei der Filterung wird meist die Energie des Signals herangezogen, um zwischen der Sprache und Störgeräuschen wie Hintergrundrauschen unterscheiden zu können. Dieser gefilterte Datensatz wird anschließend mittels schneller Fourier-Transformation transformiert [34]. Aus dem hierdurch ermittelten Spektrum der Frequenz, lassen sich die im Signal vorhandenen Frequenzanteile ablesen. Zur eigentlichen Spracherkennung wird anschließend ein Merkmalsvektor erstellt. Dies wird in der folgenden Abbildung 4 verdeutlich. Die Aufzeichnung der Sprache ist im roten Feld dargestellt. Durch die schnelle Fourier-

Transformation werden anhand der zunächst unförmigen Kurve eine Vielzahl von Sinuskurven modelliert, die übereinander gelegt wiederum die ursprüngliche Kurve ergeben. Das Spektrum der unterschiedlichen Frequenzen, hier im blauen Feld dargestellt, wird im nächsten Schritt zur eigentlichen Sprachanalyse verwendet.

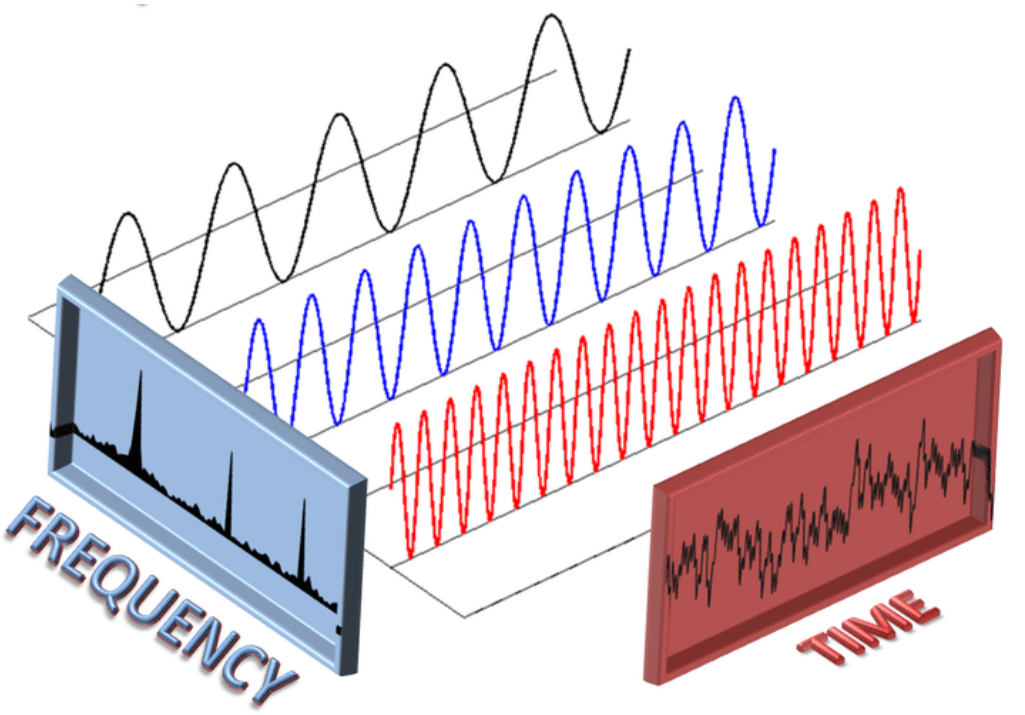


Abbildung 4: schnelle Fourier-Transformation [34]

Im weiteren Verlauf spielen lineare Hidden-Markov-Modelle (HMM), stochastische Modelle zu Wahrscheinlichkeitsketten, eine wichtige Rolle. Hierbei wird ermittelt wie wahrscheinlich es innerhalb eines Datensatzes, in der Sprachaudiometrie innerhalb eins Phonems als „kleinster“ Baustein der Sprache, ist, dass bestimmte Frequenzspektren aufeinander folgen. Im Umkehrschluss ermitteln diese Modelle anhand der bestimmten Abfolge verschiedener vorbekannter Frequenzspektren welches Phonem vorliegt [35, 36].

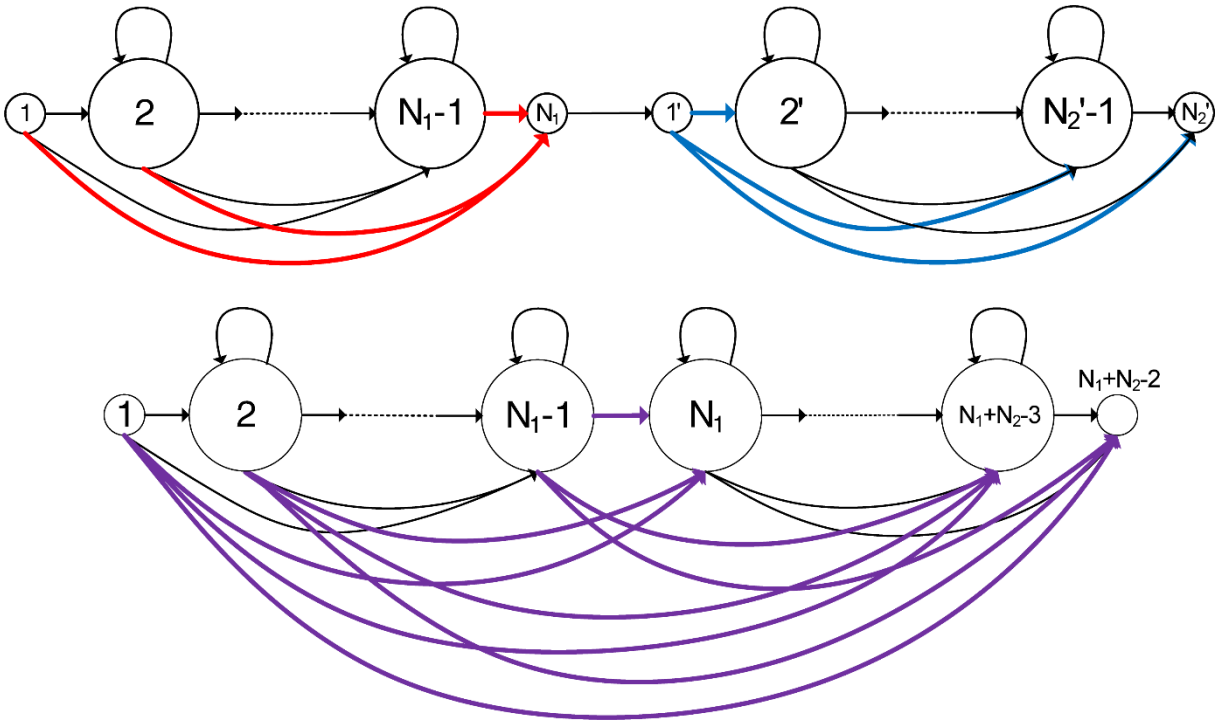


Abbildung 5: einfaches Markov-Modell [37]

Diese ermöglichen es also, die Phoneme zu finden, die am besten zu den Eingangssignalen passen. Dazu wird das im ersten Schritt erzeugte akustische Modell in verschiedene Teile zerlegt. Das Frequenzspektrum der Eingangssignale wird mit den Frequenzspektren der gespeicherten Teilstücke verglichen und anschließend mögliche Kombinationen gesucht [38]. Anhand der Abfolge der zuvor ermittelten Phoneme werden im Folgeschritt die daraus am wahrscheinlichsten vorliegenden Worte gebildet [39, 40].

Sprachmodelle versuchen anschließend, die Wahrscheinlichkeit bestimmter Wortkombinationen zu bestimmen und unwahrscheinlichere Hypothesen auszuschließen. Dazu werden in der Regel ebenfalls statistische Modelle verwendet. Daher ist es sehr wichtig diese Modelle an die Zielgruppe beziehungsweise im Idealfall an den spezifischen Anwender anzupassen. Eine N-Gramm-Statistik speichert die Auftrittswahrscheinlichkeit von Wortkombinationen aus mehreren Wörtern. Diese Statistiken werden aus großen Beispieltexten mit mehreren Millionen Worten gewonnen und repräsentative Wort- oder Buchstabenfolgen zu erhalten [41]. Dadurch können auch Homophone, sehr zuverlässig unterschieden werden. „Viel Glück“ wäre sehr viel wahrscheinlicher als „Fiel Glück“, obwohl die Aussprache identisch ist. Je mehr Worte ein N-Gramm enthält, desto genauer sind die korrekten Einschätzungen

der Auftrittswahrscheinlichkeiten der Wortkombinationen möglich. Allerdings müssen die Beispieltext-Datenbanken, aus denen beispielsweise Pentagramme, also Wortfolgen von fünf Wörtern, extrahiert werden, wesentlich größer sein als für Trigramme, Wortfolgen von drei Wörtern, denn es müssen sämtliche zulässigen Wortkombinationen aus fünf Wörtern in statistisch signifikanter Anzahl darin vorkommen. Dragon von Nuance Communications, welches ab der Version 12 auch Pentagramme verwendet, konnte die Erkennungsgenauigkeit des Systems dadurch signifikant steigert. Die Erkennungsgenauigkeit kann zusätzlich durch eine möglichst genaue Anpassung der Spracherkennungssoftware an Eigenschaften des Anwenders bezüglich des Wortschatzes, welcher beispielsweise durch den Bildungsgrad beeinflusst sein kann, der Häufigkeit der Verwendung einzelner Worte und Wortgruppen, welche in bestimmten Berufs- oder Gesellschaftsgruppen verändert sein kann, und möglicher Eigenheiten der Aussprache, etwa bei Dialekt oder anatomisch-funktionellen Problemen wie lispeln oder näseln. Das ist, neben der Möglichkeit das System aktiv umgestalten zu können, einer der wichtigsten Gründe, weshalb die meisten heute verwendeten Spracherkennungsprogramme sprecherabhängig sind und sich kontinuierlich verbessern.

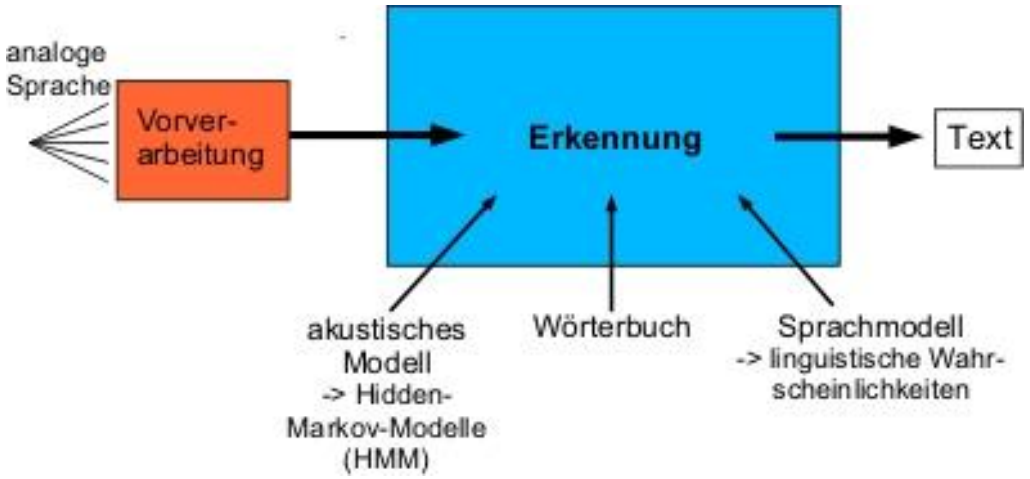


Abbildung 6: schematischer Aufbau eines modernen Spracherkennungssystems [42]

In den vergangenen Jahren wurden zunehmend Versuche unternommen, neuronale Netze für das akustische Modell zu verwenden. Hierbei ist der große Vorteil, dass diese Produkte nicht nur von der Verwendung und den Daten eines einzelnen Anwenders oder einer kleinen Anwendergruppe profitieren, sondern in kurzer Zeit mit einem großen Input von Daten versorgt werden. Dadurch können sich diese Systeme schneller an Veränderungen innerhalb einer Bevölkerungskohorte anpassen, welche

nicht nur bezüglich des Wortschatzes und der Wortwahl, sondern auch bezüglich der Aussprache bestehen. Moderne Spracherkennungssysteme, die auf Deep Learning aufsetzen verbessern ihr Outcome kontinuierlich und liefern Erkennungsraten, die schon im menschlichen Bereich liegen. Zudem gibt aber auch hybriden Ansätze, bei denen die aus der Vorverarbeitung gewonnenen Daten durch ein neuronales Netzwerk vor-klassifiziert werden sollen, und statt dem Rohdatensatz die verarbeiten Daten als Eingabe für die Hidden Markov Modelle genutzt werden [43].

Für die professionelle Verwendung von Spracherkennungssystemen gibt es vorgefertigte Datensätze mit dem zugrunde liegenden Vokabular, die die Arbeit mit dem Programm erleichtern sollen. Diese Vokabulare werden bei Dragon Datapack genannt. Je besser das Vokabular auf den vom Sprecher verwendeten Wortschatz und der Häufigkeit bestimmter Wortfolgen angepasst ist, desto geringer ist die Fehlerfrequenz. Ein Vokabular beinhaltet neben dem sprecherunabhängigen Fach- und Grundwortschatz auch ein auf den Anwender individuell angepasstes Wortfolgemodell, das sich durch die Verwendung auch im Laufe der Zeit ändert. Im Vokabular sind alle Wörter, die in die Software eingespeist wurden in korrekter Rechtschreibung und Aussprache, basierend auf der jeweiligen Hochsprache, hinterlegt. Dadurch wird ein gesprochenes Wort an seinem Klang durch das System erkannt. Wenn potenzielle Homophone vorliegen greift die Software auf das hinterlegte Wortfolgemodell zurück. Dort sind die Wahrscheinlichkeiten bestimmter Wortfolgen anwenderspezifisch definiert. Durch maschinelles Lernen werden die Wahrscheinlichkeiten stetig angepasst und optimiert. Neuere Systeme verzichten zunehmend auf die Vorgabe von starren Wortlisten da zusätzlich Kompositionswörter gebildet werden können.

Die Spracherkennung ist prinzipiell zu unterscheiden von der Stimm- bzw. Sprechererkennung, einem biometrischen Verfahren zur Personenidentifikation.

## 2.4 Fragestellung

Unsere Hypothese lautet, dass durch die Automatisierung und Modifizierung der Sprachaudiometrie eine mindestens gleichwertige Qualität hinsichtlich der Resultate, Test-Retest-Reliabilität und Interrater-Reliabilität erreicht werden kann als bei

konventioneller Durchführung. Hierzu wird standardmäßig der Oldenburger Satztest durchgeführt. Die Antworten der Probanden werden gespeichert und später ohne Beisein des Patienten mit Hilfe einer Spracherkennungs-Software mittels Dragon NaturallySpeaking Version 15.3 ausgewertet und mit der konventionellen Auswertung der Audiologen verglichen. Zusätzlich sollte getestet werden ob Patientengruppen mit bestimmten Merkmalen mehr von einer automatisierten Sprachaudiometrie profitieren als andere Gruppen. Hierfür wurde nach Unterschieden hinsichtlich Alter, Geschlecht und einseitiger oder beidseitiger Hörminderung getestet. Spracherkennungs-Softwares spielen heute schon eine bedeutende Rolle in den aktuellen Produkten vieler großer (amerikanischer) Technologieunternehmen, unter anderem bei Siri (Apple), Google Assistant, Cortana (Microsoft), Alexa/Echo (Amazon) und S Voice (Samsung). Daher erfolgte zusätzlich der Vergleich der Erkennungsraten mit Google Assistant und Microsoft Spracherkennung.

# 3 MATERIAL UND METHODEN

## 3.1 Studienkohorte

Um dieser Fragestellung nachzugehen, wurden erwachsene Patienten, die im Hörzentrum der Klinik für Hals-Nasen-Ohrenheilkunde, Kopf- und Halschirurgie des Universitätsklinikums Mannheim vorstellig wurden, getestet. Sie sollten sich auf Grund von einer Hörminderung in Behandlung befinden und mindestens auf einem Ohr unter einer hochgradigen Schwerhörigkeit, an Surditas grenzenden Schwerhörigkeit oder vollständigen Taubheit nach der 4-Frequenztabelle nach Röser 1973 leiden [44]. Zudem sollten die Studienteilnehmer der deutschen Sprache mächtig sein und eine verständliche, idealerweise akzentfrei, Aussprache haben. Dies wurde im Rahmen des Arztgespräches vor dem Studieneinschluss überprüft.

Ausgeschlossen wurden Patienten mit einem Alter von unter 18 Jahren und über 85 Jahren oder mit hochgradigen kognitiven Einschränkungen. Patienten die unter einer vollständigen Taubheit beidseits, einer angeborenen Taubheit, sowie einer beidseitigen Ertaubung vor der Sprachentwicklung litten wurden ebenfalls ausgeschlossen [44]. Zudem wurden Patienten ausgeschlossen, die mit starkem Akzent sprechen, sowie eine schlechte oder undeutliche Aussprache haben.

Zum Screening der Probanden wurde zuvor eine Reintonaudiometrie durchgeführt, falls innerhalb der vergangenen sechs Wochen keine Hördiagnostik erfolgte. Das zuletzt durchgeführte Reintonaudiogramm wurde nach der der 4-Frequenztabelle nach Röser 1973 [44] ausgewertet und der Patient in die entsprechende Art der Schwerhörigkeit (einseitige oder beidseitige Hörminderung) eingeordnet. Zudem wurde in den Vorarztbriefen und vor dem Aufklärungsgespräch nach möglichen Kontraindikationen gescreent.

Zur Durchführung der Studie lag das Ethikvotum 2018-528N-MA der Ethik-Kommission II der Universität Heidelberg vom 06.03.2018 vor. Die Testungen wurden von Dezember 2018 bis Januar 2020 durchgeführt. Die mittels klinischer Audiometer durchgeführte und manuell ausgewertete Sprachaudiometrie als Vergleichsverfahren ist standardisiert und wird breit angewandt. Sie dient somit als Goldstandard für die durchgeführten Vergleichsmessungen.

## 3.2 Versuchsaufbau

### 3.2.1 Oldenburger Satztest

Für die Durchführung unserer Studie wurde der OLSA in Ruhe verwendet, da das verwendete Spracherkennungsprogramm bei der Verwendung von nur einem Mikrofon nicht in der Lage gewesen wäre den Störschall vom Nutzschall zu separieren. Zudem wäre es für die Software schwierig gewesen, wenn sich die Lautstärke der gesprochenen Sprache nicht wesentlich von der Laustärke des Störschalls unterschieden hätte. Zudem wäre es im Nachhinein schwierig gewesen die richtig oder falsch erkannten Antworten der Patienten von der Auswertung der Sprachausgabe des OLSA-Testsystems zu unterscheiden.

Die Messungen wurden als binaurale Freifeldmessungen durchgeführt. Die Lautsprecher befanden sich direkt vor dem Patienten im Abstand von einem Meter zur Kopfmitte (in 0-Grad Azimuth). Die Probanden wurden gebeten, während der Messungen eine aufrecht sitzende Position einzunehmen und sich nicht mit dem Kopf den Lautsprechern zu nähern oder diesen zur Seite zu drehen. Die Probanden wurden dann gebeten, die Sätze zu wiederholen. Jedes Wort wurde entweder als richtig oder als falsch gewertet. Daraus konnte der absolute Wert und der Prozentsatz an korrektem Wortverständnis bei 30 Sätzen (entsprechend 150 Wörtern) berechnet werden. Es wurden nach dem Zufallsprinzip die Satzlisten 30_01, 30_02, 30_03 und 30_04 verwendet und der Patient sollte in vorangegangenen Testungen noch nicht mit der jeweils verwendet Satzliste getestet worden sein.

Der Sprachpegel liegt bei der Messung ohne Störgeräusch zu Beginn bei 30 dB und passt sich dem nachfolgenden Schema an:

| | Pegeländerung der Sprache | |
|---|---|---|
| Richtig verstandene Wörter im vorangegangenen Satz | Satz 2 bis 5 | Satz 6 bis 31 |
| 5 | -3 dB | -2 dB |
| 4 | -2 dB | -1 dB |
| 3 | -1 dB | 0 dB |
| 2 | +1 dB | 0 dB |
| 1 | +2 dB | +1 dB |
| 0 | +3 dB | +2 dB |

Tabelle 1: Adaptive Pegeländerungen für die manuelle Durchführung des OLSA, modifiziert nach „Oldenburger Satztest Bedienungsanleitung für den manuellen Test auf Audio-CD“ [45]

Die medizinisch-technischen Assistentinnen der Audiologie füllten parallel zur standardmäßigen Durchführung des OLSA auf einem vorgefertigten Formular aus, welche Worte der einzelnen Sätze richtig verstanden und falsch verstanden wurden. Dies sollte im Anschluss sicherstellen, ob die Auswertung durch das später verwendete Spracherkennungssystem die gleichen Worte als richtig oder falsch erkennen konnte.

HörTech empfiehlt vor der Durchführung des Tests die Patienten damit vertraut zu machen, indem Sie zwei 20er-Testlisten als Trainingslisten durchführen. Die erste Trainingsliste sollte dabei mit einem konstanten Pegel überschwellig durchgeführt werden, um den Patienten mit dem Sprachmaterial des OLSA vertraut zu machen. Die zweite Trainingsliste sollte adaptiv durchgeführt werden, um dem Patienten auch den Testablauf zu verdeutlichen. Bei hochgradig Schwerhörigen und Trägern von Cochlea-Implantaten sollten zum Kennenlernen der Wörter die Trainingslisten ohne Störgeräusch gemessen werden [45].

Die Aufnahme der Tondateien erfolgte mit der open source Digital Audio Workstation Audacity Version 2.4.2.

### 3.2.2 Technische Ausstattung

Die Messungen fanden in der großen Audiometriekabine (Raum 1.056) der Klinik für Hals-Nasen-Ohrenheilkunde, Kopf- und Halschirurgie, Universitätsklinikum Mannheim der Medizinischen Fakultät Mannheim statt. Die Durchführung der Messungen erfolgte durch geeignetes medizinisches Fachpersonal mit Kenntnis des Messablauf des OLSA. Die Kabine entspricht in Ausstattung und Schalldämmung der ISO DIN EN 8253-3 Norm. Im Einzelnen wurden folgende Geräte und Programme für die Untersuchung eingesetzt:

- klinisches Audiometer AT1000 von Auritec
- Visualisierungssoftware Avantgarde 5.0
- Digital Audio Workstation Audacity Version 2.4.2
- Spracherkennungssoftware Dragon Version 15.3
- Notebook (Acer), Betriebssystem Windows 10
- Textverarbeitungsprogramm Microsoft Office Word 2019
- Tabellenkalkulationsprogramm Microsoft Office Excel 2019
- Tonträger Digital Audiometer Disc: OLSA-CD 1 Track 1-12, ohne Verwendung des CCITT-Rauschens (Comité Consultatif International Téléphonique et Télégraphique) (CD 3), Copyright 2001 (KM-20091004-2, Version 1.0 vom 21.09.2011); HörTech gGmbH Marie-Curie-Str. 2, 26129 Oldenburg, Deutschland

### 3.2.3 Verwendete Spracherkennungssoftware

Im Rahmen unserer Studie wurde die Spracherkennungssoftware Dragon Version 15.3 (Nuance Communications Incorporation, Burlington, Massachusetts, USA) verwendet. Dragon (früher Dragon NaturallySpeaking) ist eine Software zur Spracherkennung am PC. Die Software setzt Äußerungen, die in ein mit dem Computer verbundenes Mikrofon gesprochen werden, in Text oder Steuerungsbefehle um. Es handelt sich um ein sprecherabhängiges Front-End-System, also eines, bei dem die Umsetzung der Sprache in Text auf dem Rechner des Nutzers erfolgt und unmittelbar nach dem Diktat der Äußerung sichtbar ist („what you say is what you see“). Im Verhältnis etwa zur Spracherkennungsfunktion von Smartphones, bei der die

Umsetzung der über das Internet gesendeten akustischen Informationen auf zentralen Servern erfolgt und der Text dann zurück übertragen wird, ergeben sich hierdurch deutliche Vorteile bei Geschwindigkeit und Genauigkeit der Umsetzung sowie der Möglichkeit zur Anpassung an Wortschatz und Bedürfnisse des Nutzers [46].

Die akustischen Signale werden zur Umsetzung digital abgetastet und im Rahmen eines „akustischen Modells“ nach Charakteristika eingeordnet, die eine ungefähre Zuordnung zu Lauten ermöglichen. Die Auswahl erfolgt statistisch unter Einsatz verschiedener Varianten von Hidden-Markov-Modellen. Ab der aktuellen Version 15 wird eine neue Spracherkennungsengine unter Einsatz von "Deep Learning" zu verwenden. Dieses akustische Modell wird bei einem anfänglichen Training und fortlaufend bei der Benutzung, insbesondere durch die Korrektur von Erkennungsfehlern, an die Stimme des jeweiligen Sprechers angepasst. Zu den „erkannten“ Lauten werden dann statistische Hypothesen über die jeweils am wahrscheinlichsten gesagten Worte angestellt. Bei ähnlich oder gleich klingenden Lauten/Worten entscheidet die Software somit anhand von Mehrwortfolgen innerhalb der Äußerung des Sprechers, welches Ergebnis als Text auf dem Bildschirm erscheint. Grundlage hierfür ist ein Sprachmodell (linguistisches Modell), welches diese Wahrscheinlichkeiten beschreibt. Der Erkennungsvorgang läuft so schnell im Hintergrund ab, dass der gesprochene Text fast sofort nach Beendigung der Äußerung auf dem Bildschirm erscheint.

Dragon verwendet bei der aktuellen Version das Sprachmodell „BestMatch IV“ mit Zusammenhängen von bis zu vier Wörtern, sogenannte Quadgramme. Das Sprachmodell funktioniert ausschließlich nach statistischen Methoden, nicht nach grammatikalischen Regeln. Die Erkennungsgenauigkeit ist aufgrund dieser Funktionsweise am besten, wenn zusammenhängende Äußerungen gesprochen werden, idealerweise vollständige und längere Sätze. Dementsprechend ist die Software auf die Erkennung von gut strukturierter Sprache ausgerichtet, nicht aber etwa für die Umsetzung von aufgezeichneten mündlichen Alltagsäußerungen mit vielen Satzbrüchen, Auslassungen und Füllworten.

Das Sprachmodell von Dragon baut auf einem mitgelieferten Wort-Lexikon auf, welches etwa 150.000 Wortformen im aktiven Vordergrundvokabular enthält. Da die Software keine grammatikalischen Regeln anwendet, sind im Vokabular nicht nur die Wortstämme, sondern alle einzelnen Wortformen hinterlegt. Um die Geschwindigkeit

der Umsetzung in einem akzeptablen Bereich zu halten, ist das Vokabular in verschiedene „Slots“ gegliedert, also ein Vordergrundvokabular, welches häufig vorkommende beziehungsweise häufig verwendete enthält, und ein Hintergrundvokabular. Das Sprachmodell der Software ist auf eine bestimmte Sprache ausgerichtet, das heißt, es ist nicht möglich, mit demselben Benutzerprofil Texte in verschiedenen Eingabesprachen zu diktieren.

Neben der Größe und Flexibilität des Wörterbuches spielt auch die Qualität der akustischen Aufnahme eine entscheidende Rolle. Bei Mikrofonen, die direkt vor dem Mund angebracht sind wird eine signifikant höhere Erkennungsgenauigkeit erreicht als bei weiter entfernten Raummikrofonen. Aus diesem Grund wurde der Abstand zwischen Sprecher und Mikrofon bei unserer Studie auf einen Meter standardisiert, um diese Fehlerquelle bestmöglich zu minimieren.

Der bis zur Version 13 verwendete Name der Software „NaturallySpeaking“ leitet sich von der Eigenschaft einer kontinuierlichen Spracherkennung ab. Hierbei muss der Sprecher zwischen den einzelnen Wörtern keine unnatürlichen Sprechpausen machen, sondern kann kontinuierlich sprechen. Müssen unnatürliche Sprachpausen gemacht werden, spricht man von diskreter Sprache. Die Software kann aus den Lautfolgen die wahrscheinlichen Wortgrenzen anhand der beschriebenen Methoden selbst ermitteln. Gleichwohl ist eine strukturierte, deutliche und flüssige Sprechweise der beste Erfolgsgarant. Nuance empfiehlt, sich an der Sprechweise von Nachrichtensprechern zu orientieren.

Die Erkennungsrate liegt bei einem gut eintrainierten Profil je nach Qualität der Hardware und Deutlichkeit der Sprechweise derzeit bei mehr als 98 Prozent. Aus Zeitgründen und Gründen der Praktikabilität wurde im Rahmen der Studie auf eine Trainingsphase verzichtet und die Messungen jeweils mit einem untrainierten Profil durchgeführt, unabhängig davon, ob der Proband zum ersten oder zweiten Mal zum Sprachtest erschien.

Zu Vergleichszwecken wurden die acht Testreihen, die die geringste Abweichung zwischen Dragon und der MTA zeigten, zudem die mit Google Assistant Spracherkennung Version 0.1.187945513 von (Google Limited Liability Company, Mountain View, Kalifornien, USA) und Microsoft Spracherkennung für Microsoft Office Word 2019 in Windows 10 (Microsoft Corporation, Redmond, Washington, USA).

Diese sind beide sprecherabhängige Spracherkennungssysteme. Auf die vom Hersteller empfohlene Trainingsphase wurde aus Gründen der Praktikabilität und besseren Vergleichbarkeit zu Dragon ebenfalls verzichtet.

## 3.3 Auswertung

### 3.3.1 Auswertung der Spracherkennung

Das Abspielen der aufgezeichneten Tondateien erfolgte erneut mit der Digital Audio Workstation Audacity Version 2.4.2. Die Tonspuren wurden über ein Kurzschlusskabel direkt vom Lautsprecherausgang auf den Mikrofoneingang übertragen. Zur Auswertung wurden Audacity, Word und Dragon, beziehungsweise bei den Kontrollmessungen Microsoft Spracherkennung oder Google Assistant parallel geöffnet.

Die Niederschrift der Ergebnisse der Spracherkennung erfolgte mit dem Textverarbeitungsprogramm Microsoft Office Word 2019 im Betriebssystem Windows 10 (beides Microsoft Corporation, Redmond, Washington, USA). Da Dragon stets versucht aus der Spracheingabe semantisch sinnvolle Texte zu generieren, wurden aufeinander folgende Sätze des OLSA von der Software zu einem zusammenhängenden Text fälschlicherweise umgeschrieben. Dies würde bei der Auswertung ein essenzielles Problem darstellen, da die Sätze semantisch nicht sinnvoll sind und abgesehen vom strukturellen Aufbau nichts miteinander zu tun haben. Selbst wenn also ein Satz komplett fehlerfrei erkannt wurde, würde er durch die Erkennung des anschließenden Satzes nachträglich fehlerhaft abgeändert werden, unabhängig davon, ob der zweite Satz richtig oder falsch ausgewertet wird. Da das Spracherkennungssystem nicht von allein in der Lage war, die 30 angebotenen Sätze voneinander abzugrenzen, wurde manuell nach jedem Satzpaar aus OLSA-Sprachausgabe und Antwort des Patienten eine neue Zeile und damit ein neuer „Text“ begonnen. Durch diese Modifikation der Spracherkennung konnten suffiziente Ergebnisse ermittelt werden.

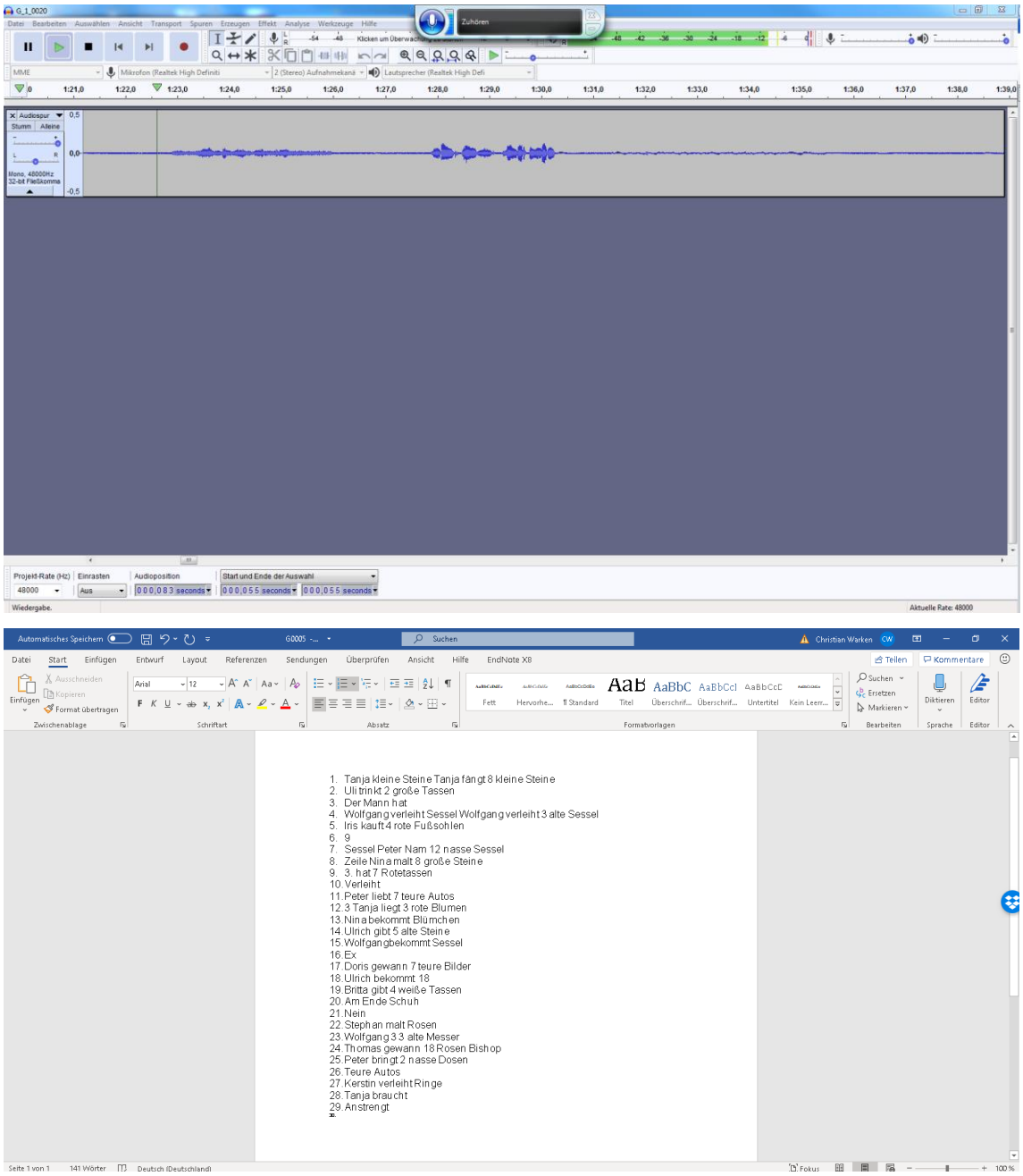


Abbildung 7: Auswertung mittels Audacity, Word und Dragon am geteilten Bildschirm

Die Auswertung mit Microsoft Spracherkennung und Google Assistant erfolgt analog zur Auswertung mittels Dragon. Die Sprachaufnahmen wurden über das Kurzschlusskabel vom Lautsprecherausgang direkt in den Mikrofoneingang eingespielt und die Textergebnisse in Echtzeit mit Word dokumentiert.

### 3.3.2 Auswertung der Ergebnisse

Die Qualität eines Spracherkennungssystems lässt sich durch verschiedene Parameter bestimmen. Neben der Erkennungsgeschwindigkeit – meist als Echtzeitfaktor (EZF) angegeben – lässt sich die Erkennungsgüte ermitteln. Die Erkennungsgeschwindigkeit spielte für unsere Studie keine relevante Rolle, da bei manueller Durchführung des OLSA nicht von relevanten Unterschieden auszugehen ist.

Die Erkennungsgüte eines Spracherkennungssystems lässt sich als Wortakkuratheit (WA) oder Worterkennungsrate (WR) messen. Die Wortakkuratheit errechnet sich aus der Formel:

$$WA\ [\%] = 100 * [1-(N_{sub}+N_{del}+N_{ins})/N_{ges}].$$

$N_{sub}$: Anzahl der vom Erkennungssystem durch andere Wörter ersetzten, d.h. „verwechselten“, Wörter (Substitutionen)
$N_{del}$: Anzahl der nicht erkannten Wörter (Deletionen)
$N_{ins}$: Anzahl der fälschlicherweise eingefügten Wörter (Insertionen)
$N_{ges}$: Anzahl aller gesprochenen Wörter

Die Worterkennungsrate (engl. „word recognition rate“, WR), wird ähnlich wie die Wortakkuratheit berechnet, allerdings ohne Berücksichtigung der fälschlicherweise eingefügten Wörter ($N_{ins}$).

Die Worterkennungsrate errechnet sich also aus der Formel:

$$WR\ [\%] = 100 * [1-(N_{sub}+N_{del})/N_{ges}].$$

Der Maximalwert von Wortakkuratheit und Wortkorrektheit beträgt 100%. Der mögliche Minimalwert der Wortkorrektheit ist 0%, während die Wortakkuratheit bei großem $N_{ins}$ auch negativ werden kann.

Zur Auswertung unserer Studie berücksichtigten wir den Vergleich der Worterkennungsraten zwischen dem Spracherkennungsprogramm und der manuellen Auswertung durch das audiologische Personal. Der Vergleich der Wortakkuratheit wäre hierfür nur bedingt geeignet gewesen, da das Prinzip des OLSA auf der Anzahl der korrekt verstandenen Worte des aus fünf Worten bestehenden Satzes besteht.

Werte unter 0 sind hierbei eben so wenig vorgesehen, wie auch Worte, die nichts direkt mit dem Satztest zu tun haben. Fälschlicherweise eingefügten Wörter würden überproportional häufig vorkommen, da die Patienten häufig Satzteile wie „nicht verstanden“, „vielleicht“ oder „weiß nicht“ einfügen, die das Spracherkennungssystem nicht von den auszuwertenden Worten unterscheiden kann und „fälschlicherweise inserierte Wörter“ bei der manuellen Auswertung schlicht nicht berücksichtigt werden. Zur Bestimmung der Wortakkuratheit müssten daher die Teile, die nicht zur Wiedergabe des abgespielten Textes gehörten, herausgeschnitten werden. Dies ist zum einem technisch sehr aufwendig, zudem ist die Unterscheidung zwischen substituierten und inserierten Worten nicht immer klar definiert, wenn gleichzeitig nicht erkannte Worte vorkommen. Auch wären anschließend die Ergebnisse der WA und WR nahezu identisch, da die jeweiligen Sätze aus lediglich fünf Worten bestehen.

Da unausgesprochene Wörter nicht im Sinne einer Worterkennungsrate ausgewertet werden können, wurden die ausgesprochenen und verarbeiteten Wörter in vier Kategorien eingeteilt: 1.) korrekt als richtig erkannt (entspricht WR), 2.) korrekt als falsch erkannt (entspricht den nicht oder vom Patienten falsch ausgesprochen Worten), 3.) nicht korrekt als richtig erkannt (entspricht Deletionen und Substitutionen) und 4.) nicht korrekt als falsch erkannt (entspricht Insertionen). Da die Insertionen, die nicht mit dem OLSA in Zusammenhang stehen nicht ausgewertet wurden entfiel in der Auswertung die vierte Kategorie. Es wurde also die Anzahl der korrekten Items in „N/150“ ausgewertet.

### 3.3.3 Statistische Auswertung

Die Datenauswertung erfolgte mit Microsoft Office Excel 2019 (Microsoft Corporation, Redmond, Washington, USA). Die Statistik erfolgte mit SAS v9.4 (SAS Institute, Cary, North Carolina, USA).

Unsere Studienkohorte war als nicht normalverteilte Population anzusehen. Daher erfolgte die statistische Auswertung mit dem Wilcoxon-Mann-Whitney-Test. Der Wilcoxon-Rangsummentest findet als Alternative zum t-Test für unabhängige Stichproben Anwendung, wenn dessen Voraussetzungen verletzt sind. Da die verschiedenen untersuchten Stichproben nicht als normalverteilt angesehen werden

konnten, war in unserer Studie der t-Test zu verwerfen. Folglich wurden die Analysen nach Alter (unter 60 Jahren und 60 Jahre oder mehr), Geschlecht (männlich und weiblich), Art der Hörminderung (einseitige und beidseitige Schwerhörigkeit), sowie die Kombination von jeweils zwei dieser Eigenschaften der Wilcoxon-Rangsummentest verwendet. Da davon auszugehen war, dass jüngere Probanden (<60 Jahre) durchschnittlich ein besseres Outcome als ältere Patienten (60+ Jahre) und einseitig ertaubte (SSD) durch ihr normalhörendes Ohr ein besseres Outcome als beidseitig ertaubte Probanden habe werden, wurde die Statistik für diese beiden Eigenschaften mit einseitigen Tests ausgewertet [47]. Da für die Auswertung nach Geschlecht nicht im Voraus festzulegen war ob Männer oder Frauen ein besseres Outcome haben, wurde hierfür ein zweiseitiger Test durchgeführt. Da dieser Rangsummentest nur für die statistische Auswertung von zwei unabhängigen Stichproben geeignet ist, musste für die Statistik der vier OLSA-Satzlisten ein anderer Test verwendet werden. Daher wurden die Daten der vier OLSA-Satzlisten mit dem Kruskal-Wallis-Test für die Analyse von mehr als zwei unabhängigen Stichproben angewandt. Es wurde jeweils ein Signifikanzniveau von <5% angestrebt.

# 4 ERGEBNISSE

## 4.1 Studiendurchführung

### 4.1.1 Epidemiologie der Studienkohorte

Eingeschlossen wurden 44 Patienten, davon waren 18 weiblich (40,9%) und 26 männlich (59,1%). Das Alter reichte von 23 Jahren bis 84 Jahren, mit einem Median von 61 Jahren. Das Durchschnittsalter lag bei 58,4 Jahren. 19 Patienten (43,2%) waren unter 60 Jahre alt, 25 Patienten (56,82%) waren 60 Jahre oder älter. 30 Personen (68,2%) litten unter einer beidseitigen hochgradigen Schwerhörigkeit, 14 Personen (31,8%) hatten eine unilaterale hochgradige Schwerhörigkeit und hörten auf dem anderen Ohr normal oder hatten maximal eine leichtgradige Hörminderung basierend auf der 4-Frequenztabelle nach Röser 1973 [44]. Alle Studienteilnehmer gaben Deutsch als Muttersprache an. Zudem gaben zwei Studienteilnehmer Türkisch als zusätzliche Muttersprache an.

| Tonhörverlust dB | 500 Hz | 1000 Hz | 2000 Hz | 4000 Hz |
|---|---|---|---|---|
| 10 | 0 | 0 | 0 | 0 |
| 15 | 2 | 3 | 2 | 1 |
| 20 | 3 | 5 | 5 | 2 |
| 25 | 4 | 8 | 7 | 4 |
| 30 | 6 | 10 | 9 | 5 |
| 35 | 8 | 13 | 11 | 6 |
| 40 | 9 | 16 | 13 | 7 |
| 45 | 11 | 18 | 16 | 8 |
| 50 | 12 | 21 | 18 | 9 |
| 55 | 14 | 24 | 20 | 10 |
| 60 | 15 | 26 | 23 | 11 |
| 65 | 17 | 29 | 25 | 12 |
| 70 | 18 | 32 | 27 | 13 |
| 75 | 19 | 32 | 28 | 14 |
| 80 | 19 | 33 | 29 | 14 |
| ab 85 | 20 | 35 | 30 | 15 |

Abbildung 8: 4-Frequenztabelle nach Röser 1973 [44]

### 4.1.2 Durchführung der Hördiagnostik

Die Untersuchungen wurden vom Dezember 2018 bis Juli 2019 durchgeführt. 14 Patienten (31,8%) erschienen zweimalig zum Oldenburger Satztest, davon waren 8 männlich (57,1%) und 6 weiblich (42,9%). Von diesen Personen wurde jeweils der

Mittelwert beider Messungen ausgewertet. Die Satzlisten 30_01, 30_02, 30_03 und 30_04 wurden jeweils bei elf Probanden verwendet.

## 4.2 Resultate der automatisierten Spracherkennung

### 4.2.1 Worterkennungsrate

Die Abweichung der Auswertung durch das Spracherkennungssystem zur Auswertung durch audiologisches Fachpersonal (Goldstandard) betrug durchschnittlich 33,98 Wörter (95%-Konfidenzintervall (KI): 24,32-43,63 Wörter), was einem Unterschied von 22,65% (95%-KI: 14,40-28,91%) entspricht. Dies entspricht einem signifikanten Unterschied ($p<0,0001$). Der Median unterschied sich um 20,5 Wörter, dies entspricht 13,67%. Nur in einer einzigen Messreihe konnte das Spracherkennungssystem mehr Worte (N=1) erkennen als das audiologische Fachpersonal. In einer weiteren Messreihe war das Ergebnis von Dragon und dem Menschen exakt gleich. Bei 42 Messreihen war das Spracherkennungssystem (teilweise deutlich) schlechter als das Fachpersonal.

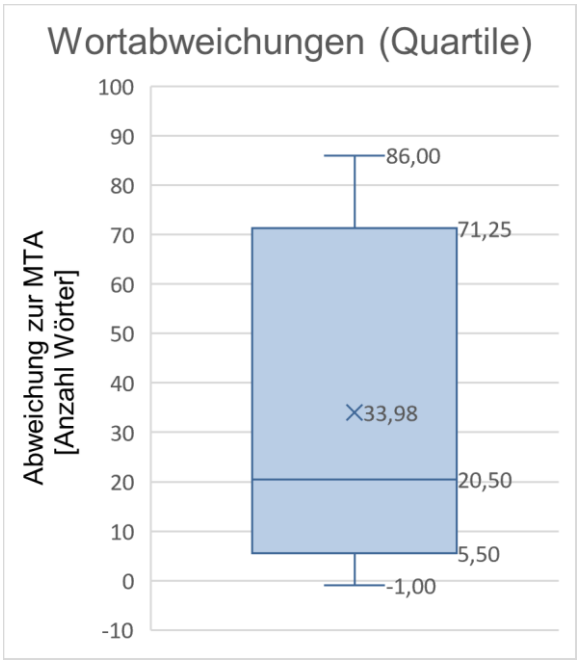


Abbildung 9: Boxplot zur Gesamtkohorte

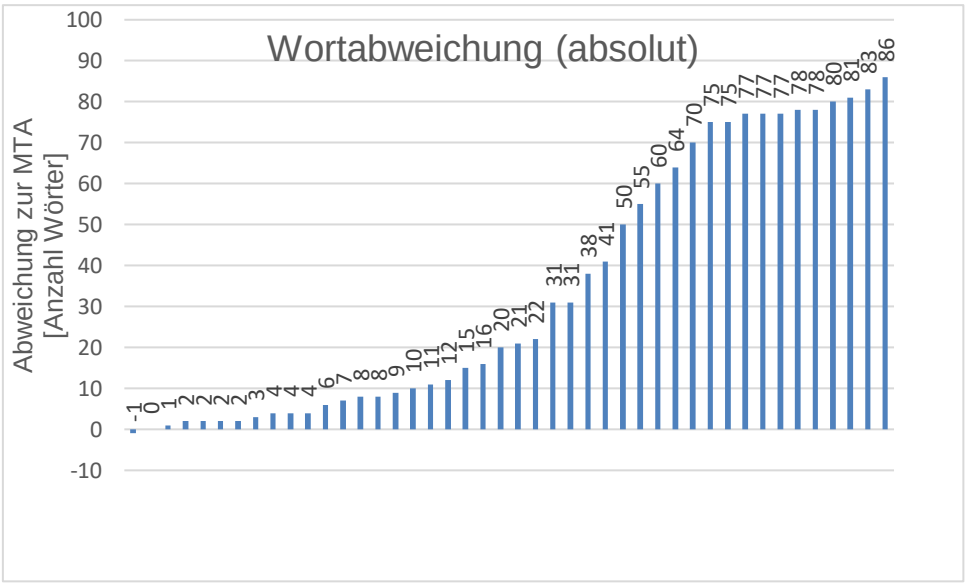


Abbildung 10: Rangsummen

Probanden unter 60 Jahren hatten mit einer medianen Abweichung von 10,0 Wörtern und einer mittleren Abweichung von 24,37 Wörtern (95%-KI: 10,45-38,28) eine signifikant geringere Fehlerquote im Vergleich zur Auswertung durch die MTA als Probanden ab 60 Jahren mit einer medianen Fehlerquote von 31,0 Wörtern und einem Mittelwert von 41,28 Wörtern (95%-KI: 27,88-54.68) (p<0,05; 0.0339).

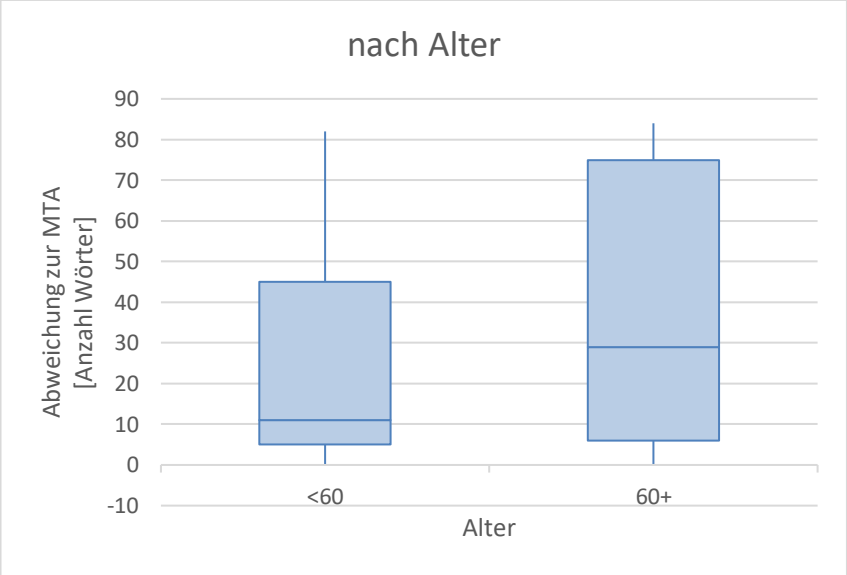


Abbildung 11: Boxplot zur Auswertung nach Alter

Eine Subgruppenanalyse zeigte, dass sich der signifikante Unterschied der Altersgruppen lediglich beim männlichen Geschlecht zeigte. Männer unter 60 Jahren hatten mit einer medianen Abweichung von 9,0 Wörtern und einer mittleren Abweichung von 17,00 Wörtern (95%-KI: 0,02-33,98) eine signifikant geringere Fehlerquote zum Goldstandard als Männer ab 60 Jahren mit einem Median von 36,00 Wörtern und einem Mittelwert von 41,00 Wörtern (95%-KI: 24,36-57.64) (p<0,05; 0.0224). Bei Frauen bestand kein signifikanter Unterschied zwischen diesen beiden Altersgruppen. Frauen unter 60 Jahren hatten mit einer medianen Abweichung von 12,00 Wörtern und einer mittleren Abweichung von 32,56 Wörtern (95%-KI: 7,09-58.02) keine signifikant geringere Fehlerquote als Frauen ab 60 Jahren mit einem Median von 22,00 Wörtern und einem Mittelwert von 41,78 Wörtern (95%-KI: 13,74-69,82) (p>0,05; 0,2825). Nicht signifikant war ebenfalls die Subgruppenanalyse der Altersgruppen nach einseitiger Hörminderung <60 Jahren (Median: 10,5; Mittelwert: 20,00; 95%-KI: 2,65-37,35 Wörter) und ab 60 Jahren (Median: 45,0; Mittelwert: 41,75; 95%-KI: (-19,9)-103,4 Wörter) (p>0,05; 0,0893) und beidseitiger Hörminderung <60

Jahren (Median: 8,0; Mittelwert: 29,22; 95%-KI: 3,01-55,44 Wörter) und ab 60 Jahren (Median: 31,0; Mittelwert: 41,19; 95%-KI: 26,52-55,86 Wörter) (p>0,05; 0,1649).

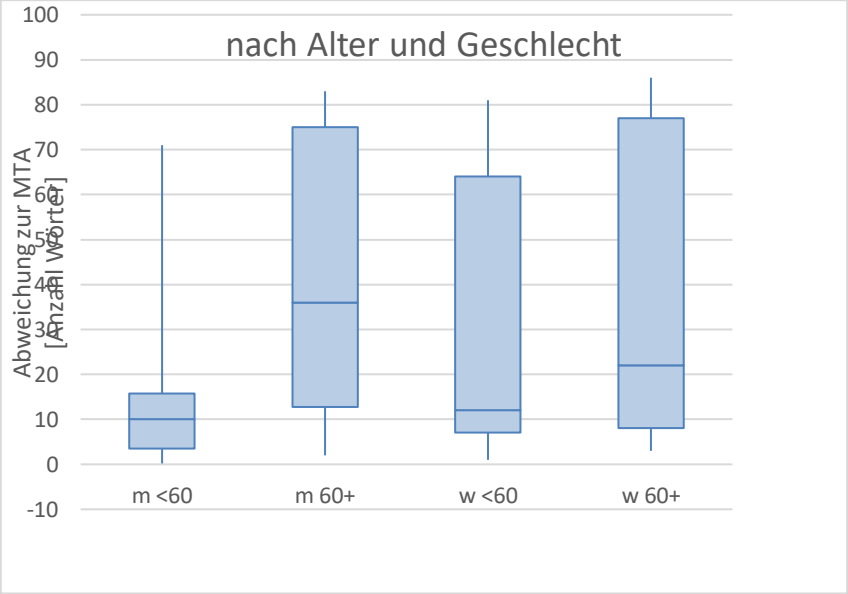


Abbildung 12: Boxplot zur Auswertung nach Alter und Geschlecht

Es zeigte sich ein deutlicher Unterschied der Fehleranzahl zwischen Probanden mit einer single sided deafness (SSD, einseitiger Taubheit) (Median: 11,5; Mittelwert: 26,21; 95%-KI: 9,30-43,12 Wörter) und Patienten mit einer beidseitigen Hörminderung (Median: 26,5; Mittelwert: 37,60; 95%-KI: 25,40-49,80 Wörter). Der Unterschied stellte sich allerdings als nicht signifikant heraus (p>0,05; 0,0827). Auch Subgruppenanalysen nach Personen unter 60 Jahre mit SSD (Median: 10,5; Mittelwert: 20,00; 95%-KI: 2,65-37,35 Wörter) und beidseitiger Hörminderung (Median: 8,0; Mittelwert: 29,22; 95%-KI: 3,01-55,44 Wörter) (p>0,05; 0,3266), sowie Personen ab 60 Jahre mit SSD (Median: 45,0; Mittelwert: 41,75; 95%-KI: (-19,9)-103,4 Wörter) und beidseitiger Jahren (Median: 31,0; Mittelwert: 41,19; 95%-KI: 26,52-55,86 Wörter) (p>0,05; 0,3147). Ebenso zeigte eine Auswertung nach Männern mit SSD (Median: 13,0; Mittelwert: 30,70; 95%-KI: 7,24-54,16 Wörter) und binauraler Hörminderung (Median: 25,5; Mittelwert: 32,44; 95%-KI: 16,42-48,45 Wörter) (p>0,05; 0,2987), sowie Frauen mit SSD (Median: 10,5; Mittelwert: 15,00; 95%-KI: (-10,50)-40,49 Wörter) und bilateraler Schwerhörigkeit (Median: 43,0; Mittelwert: 43,50; 95%-KI: 22,92-64,08 Wörter) (p>0,05; 0,1318) keinen signifikanten Unterschied.

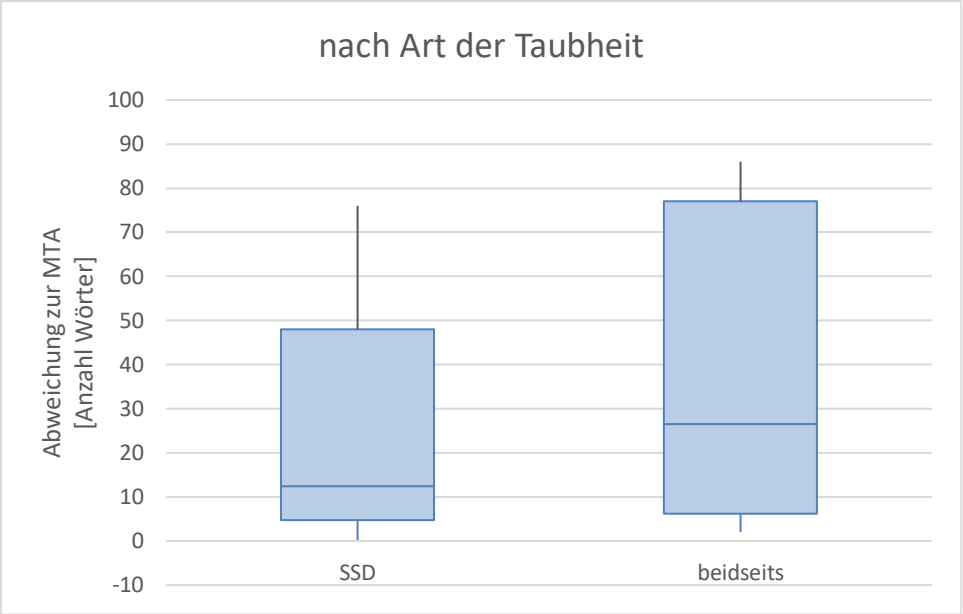


Abbildung 13: Boxplot zur Auswertung nach Art der Taubheit

Es zeigte sich kein signifikanter Unterschied der Fehleranzahl zwischen Männern (Median: 18,0; Mittelwert: 31,77; 95%-KI: 19,45-44,09 Wörter) und Frauen (Median: 21,5; Mittelwert: 37,17; 95%-KI: 20,19-54,14 Wörter) (p>0,05; 0,4662). Auch Subgruppenanalysen nach Männern (Median: 9,0; Mittelwert: 17,00; 95%-KI: 0,02-33,98 Wörter) und Frauen (Median: 12,0; Mittelwert: 32,56; 95%-KI: 7,09-58.02 Wörter) unter 60 Jahre (p>0,05; 0,2883), sowie Männer (Median: 36,0; Mittelwert: 41,00; 95%-KI: 24,36-57,64 Wörter) und Frauen (Median: 22,0; Mittelwert: 41,78; 95%-KI: 13,74-69,82 Wörter) ab 60 Jahre (p>0,05; 0,7768). Ebenfalls nicht signifikant waren Unterschiede zwischen Männern (Median: 13,0; Mittelwert: 30,70; 95%-KI: 7,24-54,16 Wörter) und Frauen (Median: 10,5; Mittelwert: 15,00; 95%-KI: (-10,50)-40,49 Wörter) mit unilateraler Schwerhörigkeit (p>0,05; 0,6202), sowie Männern (Median: 25,5; Mittelwert: 32,44; 95%-KI: 16,42-48,45 Wörter) und Frauen (Median: 43,0; Mittelwert: 43,50; 95%-KI: 22,92-64,08 Wörter) mit bilateraler Taubheit (p>0,05; 0,3933) zeigten keinen signifikanten Unterschied.

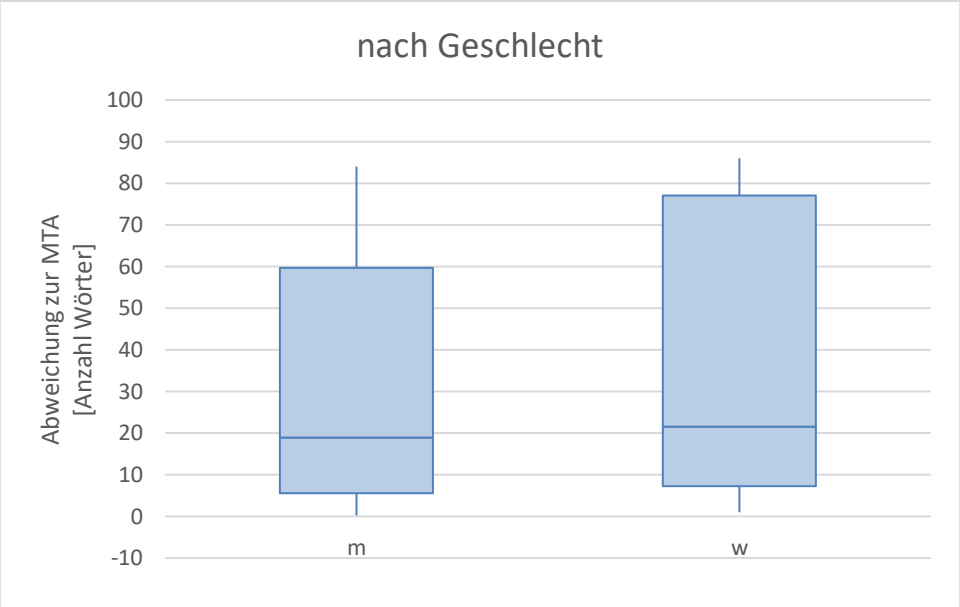


Abbildung 14: Boxplot zur Auswertung nach Geschlecht

Die Anzahl der Abweichung zwischen der Auswertung der MTA und der Auswertung durch das Spracherkennungsprogramm war in der Satzliste 4 mit einem Median von 55,0 Wörtern und einer mittleren Differenz von 43,91 Wörtern (95%-KI: 19,45-68,36) mit Abstand am größten. Die Fehleranzahl der Satzlisten 1-3 zeigten nur einen geringen Unterschied mit in absteigender Reihenfolge Satzliste 2 (Median: 22,0; Mittelwert: 34,45; 95%-KI: 13,33-55,58 Wörter), Satzliste 1 (Median: 16,0; Mittelwert: 28,82; 95%-KI: 9,04-48,60 Wörter) und Satzliste 3 (Median: 12,0; Mittelwert: 28,73; 95%-KI: 7,62-49,83 Wörter). Es zeigte sich jedoch kein statistisch signifikanter Unterschied zwischen den verschiedenen Satzlisten (p>0,05; 0,6584).

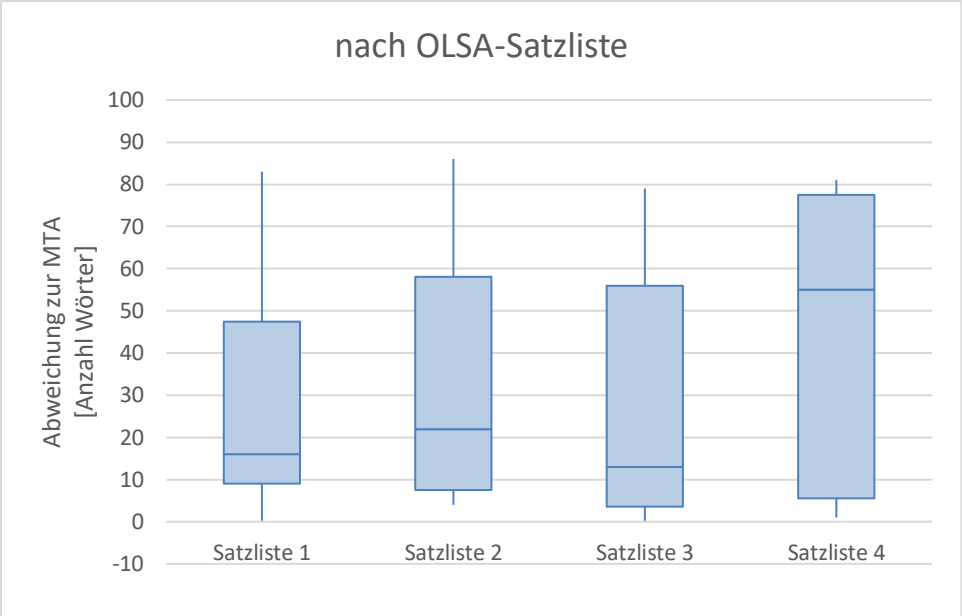


Abbildung 15: Boxplot zur Auswertung nach OLSA-Satzliste

### 4.2.2 Vergleichsmessungen

Die Kontrollmessungen mit Microsoft Spracherkennung und Google Assistant ergaben keine relevanten Ergebnisse. Google Assistant konnte in den acht getesteten Tonaufnahmen kein einziges Wort erkennen. Microsoft Spracherkennung konnte in den gleichen acht Aufzeichnungen jeweils zwischen einem und drei Worte korrekt erkennen. Da keine der Kontrollmessungen zu einem plausiblen Ergebnis führten und die Abweichung zu Dragon und der manuellen Auswertung bei fast 100% lagen, wurde hierbei auf eine statistische Auswertung verzichtet.

# 5 DISKUSSION

## 5.1 Ergebnisse

Die Auswertung der gesamten Studienkohorte ergab ein sehr heterogenes Bild. Während bei manchen Testreihen die Auswertung durch die Spracherkennungssoftware (fast) deckungsgleich zur Auswertung durch das die Messung durchführende Personal waren, bestanden bei anderen Testreihen Abweichungen von über 50%. Wenn man bedenkt, dass sowohl die korrekt verstanden als auch die richtigerweise als falsch, beziehungsweise nicht verstanden Worte in der statistischen Auswertung berücksichtigt wurden, ist vermutlich von noch höheren Fehlerquoten auszugehen. In insgesamt dreizehn Testreihen wurden weniger als zehn Prozent der korrekt verstandenen Worte auch von dem Spracherkennungssystem als korrekt verstanden erkannt. Nicht verstandene Worte sind hierbei weitaus einfacher zu erkennen da der Proband entweder mitteilte einen Teil der Worte nicht verstanden zu haben oder nichts antwortete.

Im Rahmen der Auswertung zeigte sich, dass die Sätze, in denen die Versuchsperson entweder alle fünf Worte oder kein einziges Wort verstand vom Spracherkennungssystem wesentlich häufiger richtig verarbeitet werden konnten als die Sätze, die im Bereich der Sprachverständlichkeitsschwelle L50 lagen, also bei zwei bis drei korrekten Worten. Eine Erklärung hierfür ist das Bestreben von Dragon einen sprachlich sinnvollen Text beziehungsweise Satz zu bilden. Ein Satz bestehend aus Subjekt-Verb-Zahlwort-Adjektiv-Substativ ergibt logischerweise mehr Sinn als ein lückenhafter Satz ohne Zusammenhang. Ein leerer Satz ist hierbei selbsterklärend.

Einen Nachteil, den die Spracherkennungssoftware in unserer Studie hatte, war die fehlende Interaktion. Ein menschlicher Testleiter kann jederzeit nachfragen, wenn er sich bei der Antwort der Testperson unsicher ist, um eine Wiederholung der Antwort bitten oder der Testperson mitteilen laut und deutlich zu sprechen. Die Spracherkennung kann dies, zumindest in unserem Versuchsaufbau, nicht. Das wäre auch eine mögliche Erklärung für die deutlich abweichende Testgenauigkeit von Dragon im Vergleich zum Fachpersonal. Eventuell war die Gesprächslautstärke des Probanden für das audiologische Personal ausreichend, für das Spracherkennungsprogramm jedoch nicht. Eine weitere Fehlerquelle im Hinblick hierauf kann auch im Versuchsaufbau liegen. Die Probanden hatten einen Anstand

von einem Meter zum Mikrofon und wurden gebeten direkt hineinzusprechen. Bei einer Antwort direkt in Richtung der MTA, einer Änderung der Sitzposition oder einer versehentlichen Positionierung der Arme vor dem Mikrofon können diese Vorgaben gegebenenfalls nicht erfüllt worden sein. Diese Tatsache kann auch vom Testleiter nur bedingt überprüft werden, da während der Durchführung des Sprachtestes der Monitor das Mikrofon, teilweise auch den Probanden, verdeckt und die MTA zeitgleich den OLSA durchführen, die Audiodaten aufzeichnen und den Auswertungsbogen ausfüllen muss.

## 5.2 Unterschiede beim Vergleich verschiedener Gruppen

Im Rahmen unserer Studie konnte keine Personengruppe gefunden werden, für die eine automatisierte Durchführung des OLSA eine adäquate Alternative zum etablierten Goldstandard ist. Jedoch zeigte sich, dass die eingeschränkte Aussagekraft bei manchen Subgruppen schwächer ausgeprägt war als bei anderen Gruppen. Wie zu erwarten war funktionierte das Spracherkennungssystem bei jüngeren Probanden zuverlässiger, allerdings auch hier signifikant schlechter als bei der manuellen Auswertung. Eine mögliche Erklärung hierfür könnte sein, dass kognitive Prozesse, wozu die Sprachverarbeitung zu zählen ist, im Alter langsamer ablaufen [48, 49] Wir gehen nicht davon aus, dass dies primär auf eine höhere Technikaffinität zu erklären ist, da die Probanden lediglich sprechen mussten und die Verwendung der Software erst im Nachhinein erfolgte [47]. Eine mögliche Erklärung könnte der Einfluss der Dauer der Hörminderung auf die Qualität der Sprache sein. Da auch Patienten mit kongenitalen Hörschädigungen und Patienten mit lange zurückliegenden Ohroperationen eingeschlossen wurden ist daher davon auszugehen, dass die durchschnittliche Dauer der Hörminderung beziehungsweise Taubheit länger war und dies einen negativen Einfluss auf die Funktionalität des Spracherkennungssystems hatte [50-53]. Gegen diese Annahme spricht allerdings, dass die Funktionalität auch bei einer unilateralen Schwerhörigkeit tendenziell besser war als bei einer bilateralen Schwerhörigkeit. Bei einer unilateralen Taubheit mit adäquaten Restgehör auf der Gegenseite und einer binauralen Freifeldmessung ist die Dauer der Taubheit nicht vorhanden und es wäre von einer größeren Differenz zu Patienten mit beidseitiger Taubheit und teilweise jahrelanger Surditas zu erwarten. Es muss also noch weitere Gründe geben, die in unserer Studie möglicherweise nicht betrachtet wurden. Dazu

gehören die Größe des Wortschatzes, der Umfang des Sprachgebrauchs im Alltag oder der sozio-ökonomische Status.

## 5.3 Stärken und Schwächen

Wie bereits beim Design der Studie vermutet wurde, ist nach heutigem Stand der Technik ein Spracherkennungssystem kein adäquater Ersatz für den Menschen. Hierfür zeigen sich mehrere Gründe auf die in den folgenden Abschnitten eingegangen wird.

### 5.3.1 Eigenschaften der Sprache

Problematisch bei der Bewertung der Antwort durch den Untersucher bleiben regionaltypische Aussprachevarianten wie „Teig-Teich“. Eine weitere Schwierigkeit besteht in der Verwendung von Neologismen und Lehnwörtern aus anderen Sprachen. Zudem kann es je nach regionalem Dialekt für eine und dieselbe Sache diverse vollkommen verschiedene Worte geben, wie Portemonnaie, Geldbeutel oder (Geld-)Börse. Andererseits kann der gleiche Begriff innerhalb eines Sprachgebiets auch für vollkommen verschiedene Dinge stehen. Eine Person aus Süddeutschland würde unter einem „Pfannkuchen“ etwas völlig anderes verstehen als eine Person aus Berlin. Im Gegensatz dazu würde eine Muttersprachlerin aus Österreich beide Dinge nicht „Pfannkuchen“ nennen. Außerdem ist, übrigens für alle Sprachtests, bei der Übertragung des Untersuchungsbefundes auf das Sprachverstehen des Probanden im Alltag zu beachten, dass die Erhöhung des Sprachpegels am Audiometer etwas ganz anderes ist als das laute Sprechen in der zwischenmenschlichen Kommunikation, bei dem sich nicht nur der Pegel, sondern auch die Frequenzzusammensetzung ändert. Wie bereits zuvor beschrieben war, ist für das Sprachekennungsprogramm das Erkennen spezifischer Frequenzzusammensetzungen für die Identifikation einzelner Phoneme essenziell. Das Programm wird also wahrscheinlich Schwierigkeiten bekommen, wenn für identische Phonem unterschiedliche Frequenzanteile vorliegen und damit steigt die Wahrscheinlichkeit von Fehlern. Ein unter Laborbedingungen erhaltenes Testergebnis ist in Bezug auf seine Relevanz für den Alltag noch immer unvollkommen.

Zudem hängt das Ergebnis einer konventionellen Sprachaudiometrie maßgeblich vom Sprachverständnis des Untersuchers, welches üblicherweise nicht vorab überprüft wird, und der Aussprache der zu untersuchenden Person ab, weshalb von einer nicht zu vernachlässigenden Interrater-Reliabilität auszugehen ist. Diesen Umstand beschrieb Johann Wolfgang von Goethe bereits vor über 200 Jahren: „Niemand hört als was er weiß, niemand vernimmt als was er empfinden, imaginieren und denken kann.“ [54].

Der Zwang zum Nachsprechen der Testwörter kann zu negativen Einflüssen auf die Reliabilität und vor allem auf die Objektivität der Testergebnisse führen. Durch dialektale Färbung, ungenaue Aussprache oder auch durch Aufmerksamkeitsschwankungen entstehen zufällige Fehler mit entsprechender Konsequenz für das Testergebnis. Zusätzlich können auch eine Schwerhörigkeit oder Konzentrationsstörungen des Testleiters zu fehlerhaften Resultaten führen. Um diese Unsicherheiten so gut es geht zu vermeiden bieten sich geschlossene Testverfahren an, die mit einer schriftlichen oder bildgestützten Vorlage von Antwortalternativen ein sprachfreies Auswählen des gehörten Wortes durch den Patienten ermöglichen. Dadurch können Übertragungsfehler von der Testperson zum Testleiter weitestgehend vermieden werden.

Geschlossene Verfahren erfordern eine gewisse Mitarbeit des Patienten. Es hat sich zwar gezeigt, dass auch ältere und computerunerfahrene Personen keine Probleme mit diesem Antwortformat haben. Dennoch ist zu bedenken, dass die selbständige Durchführung nicht bei allen Probanden ohne weiteres möglich ist. Probleme sind beispielsweise zu erwarten, wenn Personen gravierende Sehbeeinträchtigungen oder ein motorisches Handicap haben. Ebenso könnten erhebliche Lese-Rechtschreib-Defizite oder auch Demenzerkrankungen sich nachteilig auf die Validität der Testergebnisse auswirken. Zudem muss die Testperson die jeweilige Testsprache beherrschen, da die gehörten Begriffe nicht nur wiederholt werden müssen, sondern auch das passende Symbol ausgewählt werden muss. Zusätzlich muss die Ratewahrscheinlichkeit einkalkuliert werden, da Patienten auch einen zufälligen Begriff auswählen oder zwischen sinnvollen Alternativen wählen können, ohne das Wort wirklich verstanden zu haben.

### 5.3.2 Spracherkennunssystem

Für die gewünschte Fragestellung ist die Auswahl des geeigneten Spracherkennungssystem essenziell. Das „perfekte“ Spracherkennungsprogramm existiert nicht. Das von uns verwendete Dragon Version 15.3 ist ein sprecherabhängiges System, das für medizinisches Fachpersonal konzipiert wurde. In der Wortdatenbank sind daher neben vielen Worten aus dem deutschen Standardwortschatz auch Fachbegriffe, wie „Femur“, enthalten die im allgemeinen Sprachgebrauch üblicherweise verwendet werden, im klinischen Alltag aber gehäuft Verwendung finden. Viele dieser Begriffe sind aus anderen Sprachen wie Latein, Altgriechisch oder Englisch abgeleitet. Da es sich um ein modernes System handelt, dass idealerweise auch ohne vorherige Trainingsphase funktioniert, waren die Ergebnisse dennoch bei einem Teil der Probanden zufriedenstellend. Eine Voraussetzung hierfür ist eine klare, deutliche Aussprache in normaler Gesprächslautstärke und die Verwendung von Worten, die auch im Wortschatz des Systems verzeichnet sind und eine hohe Wahrscheinlichkeit haben in allgemeinen Sprachgebrauch vorzukommen. Vor allem in den Testreihen mit einer hohen Diskrepanz zwischen den Ergebnissen der Spracherkennungssoftware und der manuellen Auswertung wurden häufig Begriffe, wie „Lymphknoten“ oder „Resorption“ dokumentiert, die man eher in einem Artikel in einer medizinischen Fachzeitschrift als in einem Alltagsgespräch unter Nachbarn vermuten würde. Diese Begriffe sind auch nicht im Wortrepertoire des OLSA enthalten. Auch bestimmte Fehler aus der Alltagssprache kamen in der Auswertung gehäuft vor: „dritter“ oder „Bretter“ statt „Britta“, „mal“ statt „malt“, „L“ statt „elf“ oder „bei“ statt „zwei“. Dies lässt sich darauf zurückführen, dass die elektronische Spracherkennung mit Wahrscheinlichkeits-modellen arbeitet. Es ist einfach wahrscheinlicher, dass ein Patient einen „dritte[n]“ Tag im Krankenhaus verbringt, im Vergleich zur Wahrscheinlichkeit, dass der Vorname „Britta“ lautet.

### 5.3.3 Trainingsphase

Durch eine Trainingsphase mit einem Spracherkennungsprogramm kann die Qualität und damit die Verwertbarkeit der Ergebnisse signifikant verbessert werden. Zum einen erkennt die Spracherkennungssoftware die individuellen Eigenschaften der Sprache

des jeweiligen Sprechers, zum anderen wird der Sprecher zunehmend mit dem verwendeten System vertraut [55, 56]. Für Folgestudien wäre es daher von großem Vorteil die Spracherkennung bereits während der beiden Testreihen, die als Trainingsphase vor der eigentlich ausgewerteten Satzliste verwendet werden, zu aktivieren und damit das Spracherkennungssystem individuell an die zu testende Person anzupassen. Zusätzlich wird dann die zu testende Personen nicht nur mit dem Aufbau und Ablauf des OLSA vertraut gemacht, sondern auch mit der Funktionalität der Spracherkennungssoftware.

### 5.3.4 Statistische Einschränkungen

Es bestanden teilweise sehr kleine Subgruppen, beispielsweise nur vier Probanden in der Gruppe der einseitig tauben Frauen und ebenso wenige in der Gruppe der einseitig tauben Patienten im Alter von 60 Jahren oder mehr. Daher ist davon auszugehen, dass manche Ergebnisse aufgrund der zu geringen Gruppengröße nicht signifikant sind. Größere Gruppengrößen, zum Beispiel durch den gezielten Einschluss von Patienten mit bestimmten Kombinationen von Eigenschaften (SSD, weiblich, 60+) könnten zeigen, dass auch für andere Subgruppen als männlich und unter 60 Jahren signifikante Unterschiede bestehen. Zudem wären bei einer deutlich größeren Anzahl von Probanden auch Analysen unter Einbeziehung von mehr als zwei Eigenschaften möglich. Dadurch könnte sich zeigen, ob eine automatisierte Spracherkennung mit dem OLSA für ganz speziell ausgewählte Patientengruppen im klinischen Alltag zuverlässig und gleichwertig zum aktuellen Goldstandard möglich ist.

## 5.4 Mögliche Verbesserungen

In unserer Studie fehlte die Auswertung der größten Vergleichsgruppe, nämlich Menschen ohne Hörschädigung. Dieser Vergleich könnte zeigen, ob eine automatisierte Spracherkennung bei der Gruppe mit den vermeintlich günstigsten Grundvoraussetzungen einen Mehrwert zeigt. Sprachkundige hörgesunde Menschen, ohne kognitive Einschränkungen, sollten über eine adäquate Aussprache verfügen. Dies vereinfacht einem Spracherkennungssystem die korrekte Auswertung von präsentierten Spracheingaben. Da der OLSA im klinischen Alltag vor allem für die

Diagnostik bei Patienten mit höhergradig eingeschränktem Gehör verwendet wurde, wurden normalhörende Personen in unserer Studie nicht berücksichtigt. Hieraus würde sich daher aus unserer Sicht keine Relevanz ergeben zur Erleichterung der Durchführung, der Möglichkeit mehr Menschen von dieser Diagnostik profitieren lassen zu können oder der Senkung der Kosten. Allerdings könnte die Einbeziehung von Hörgesunden ein wichtiger Zwischenschritt sein, eine verbesserte automatisierte Spracherkennung zu etablieren.

Manche Sprachtests versuchen die Wahrscheinlichkeit von durch den Untersucher falsch verstandenen oder vom Untersuchten nicht korrekt nachgesprochenen Sprachlaute zu minimieren. Beim Reimtest nach Sotscheck markiert der Proband nach einem 5 AFC-Verfahren („five alternatives forced choice“) nach der akustischen Darbietung des einsilbigen Worts das vermeintlich gehörte Wort auf einer visuellen Anzeige. Die 5 Antwortalternativen sind nach dem Prinzip von „Minimalpaaren“ zusammengestellt. Die angezeigten Begriffe unterschieden sich nur im Anlaut, im Inlaut oder im Auslaut (Beispiel im Anlautteil: Sinn–hin–bin–Zinn–Kinn). Ein weiteres, hiervon unabhängiges Merkmal dieses Sprachtestes ist der automatisierte Ablauf. Dadurch muss der Untersucher nicht über die komplette Dauer der Untersuchung anwesend sein [57].

In unserer Studie wurde nicht überprüft, ob die Qualität der Aussprache mit der Herkunft oder dem Wohnort, der Art des Schulabschlusses oder der (früheren) beruflichen Tätigkeit des Probanden korreliert. Wie zuvor erwähnt sind Spracherkennungssysteme üblicherweise für die hochdeutsche Aussprache und einen bestimmten Wortschatz trainiert, in unserem Fall die medizinische Fachsprache. Es ist daher davon auszugehen, dass die Qualität unserer Messungen bei Menschen, die eher einen Dialekt sprechen, einen anderen Fachwortschatz verwenden oder einen niedrigen Bildungsabschluss haben, sukzessive abnimmt. Zusätzlich könnte der sozio-ökonomische Status Einfluss auf das Outcome haben. Ebenso wurde nicht berücksichtigt zu welcher Tageszeit die Messung durchgeführt wurde und wie hoch der Grad der Erschöpfung des Probanden war. Vorherige Studien zeigten bereits, dass es bei vermehrter Müdigkeit und vermindertet Aufmerksamkeit zu Problemen mit der Aussprache kommen kann [58]. Die Messungen unserer Patienten wurden meistens am Ende eines Termins in unserer Klinik durchgeführt. Die Anwesenheitszeit betrug in der Regel vier bis sechs Stunden. Somit ist damit zu rechnen, dass viele Patienten

zum Zeitpunkt des Testes ermüdeten und über eine verminderte Konzentrationsfähigkeit verfügten. Für Folgestudien könnte es somit von großem Vorteil sein, die Messungen zu einer definierten Uhrzeit direkt nach Erscheinen in der Testkabine durchzuführen, unnötige Verzögerungen zu vermeiden und die Qualität der Messungen nicht künstlich zu vermindern.

Um festzustellen ob mögliche Abweichungen zwischen dem Ergebnis des Spracherkennungssystem und dem Ergebnis der MTA nur durch die (möglicherweise) fehlerhafte Aussprache des Probanden zu erklären sind, sollte auch die Sprachausgabe des OLSA direkt mittels Dragon ausgewertet werden.

## 5.5 Ausblick

### 5.5.1 Evolution der Spracherkennungssysteme

In den vergangenen Jahren wurden zunehmend Versuche unternommen, neuronale Netze für das akustische Modell zu verwenden. Moderne Spracherkennungs-systeme, die auf Deep Learning aufsetzen verbessern ihr Outcome kontinuierlich und liefern Erkennungsraten, die schon in Bereichen liegen, die auch ein gesunder Mensch erreichen würde [59]. Zudem gibt aber auch hybriden Ansätze, bei denen die aus der Vorverarbeitung gewonnenen Daten durch ein neuronales Netzwerk vor-klassifiziert werden sollen, und die Ausgabe des Netzes als Eingabe für die Hidden Markov Modelle genutzt werden.

Deep-Learning-Technologien findet man heute in vielen Bereichen der Medizin von der Bild- und Spracherkennung bis hin zur Computer- und Systembiologie. Deep-Learning ist ein Teilgebiet des maschinellen Lernens, welches mit künstlichen neuronalen Netzwerken arbeitet. Um die Arbeitsweise einer künstlichen Intelligenz zu verstehen, muss man sich bewusst sein, dass Maschinen mit Algorithmen arbeiten, während Menschen viele Aufgaben durch Intuition und Erfahrung lösen. Es gibt somit Fragestellungen, die für eine Maschine sehr leicht zu lösen, für einen Menschen jedoch sehr schwer. Hierunter fallen typischerweise komplizierte Berechnungen, die sich mit mathematischen Formeln leicht lösen lassen. Beispielsweise kann ein einfacher Taschenrechner eine komplizierte Potenzrechnung in Sekundenbruchteilen lösen, während selbst ein spezialisierter Mensch hierfür ohne Hilfsmittel eine gewisse

Zeit benötigt [19]. Es existieren jedoch auch gegenteilige Fragestellungen, die sich auch von Menschen mit geringen Grundkenntnissen lösen lassen. Hierunter fallen insbesondere die Bild- oder Spracherkennung. Sogar Kleinkinder können Bilder oder Geräusche von ihnen bekannten Tieren verlässlich zuordnen, sofern sie dies zuvor erlernt haben. Ein Mensch würde das Foto eines Hundes sehr leicht erkennen, während für einen Computer schwierig ist die sensorischen Eingangsdaten zu verstehen. So existieren zunächst nur wahllose Bildpunkte mit unterschiedlichen Informationen (Farbton, Helligkeitsstufe) welche kategorisiert werden müssen. Die Überführung einer Menge von Bildpunkten in eine Kette von Ziffern (0 und 1) ist sehr kompliziert. Komplexe Muster müssen aus Rohdaten extrahiert werden. Deep Learning durchläuft hierbei einen selbstadaptiven Prozess und nutzt eine Reihe hierarchischer Schichten und Konzepte. Dazu werden künstliche neuronale Netze konstruiert, die wie das menschliche Gehirn gebaut sind, wobei die Neuronen wie ein Netz miteinander verbunden sind. Die erste Schicht verarbeitet die Rohdateneingabe und leitet ihre Ausgaben anschließend an die nächste Schicht weiter. Diese verarbeitet die Informationen der vorherigen Schicht und gibt das Ergebnis ebenfalls an die nächste Schicht weiter. Die enthaltenen Merkmale werden zunehmend abstrakter. Die zugrunde gelegten Konzepte Modell bestimmen, welche Folgerungen die jeweilige Ebene aus den Resultaten der Vorebene zieht. Die Eingabeschicht und Ausgabeschicht sind dabei einsehbar, während dazwischen mehrere verborgene Schichten (hidden layers) existieren, die vom Menschen nicht einsehbar und auch nicht manipulierbar sind. Die Korrektheit eines Deep Learning Modell kann also dadurch beurteilt werden, ob aus den Eingabedaten die richtigen Schlüsse gezogen werden, z.B. ob das Bellen eines Hundes als Hund verstanden wird. Hierbei besteht die Fähigkeit aus Erfahrungen zu lernen und Konzepte zu verstehen. Die wird versucht in Abbildung 16 vereinfacht darzustellen [60].

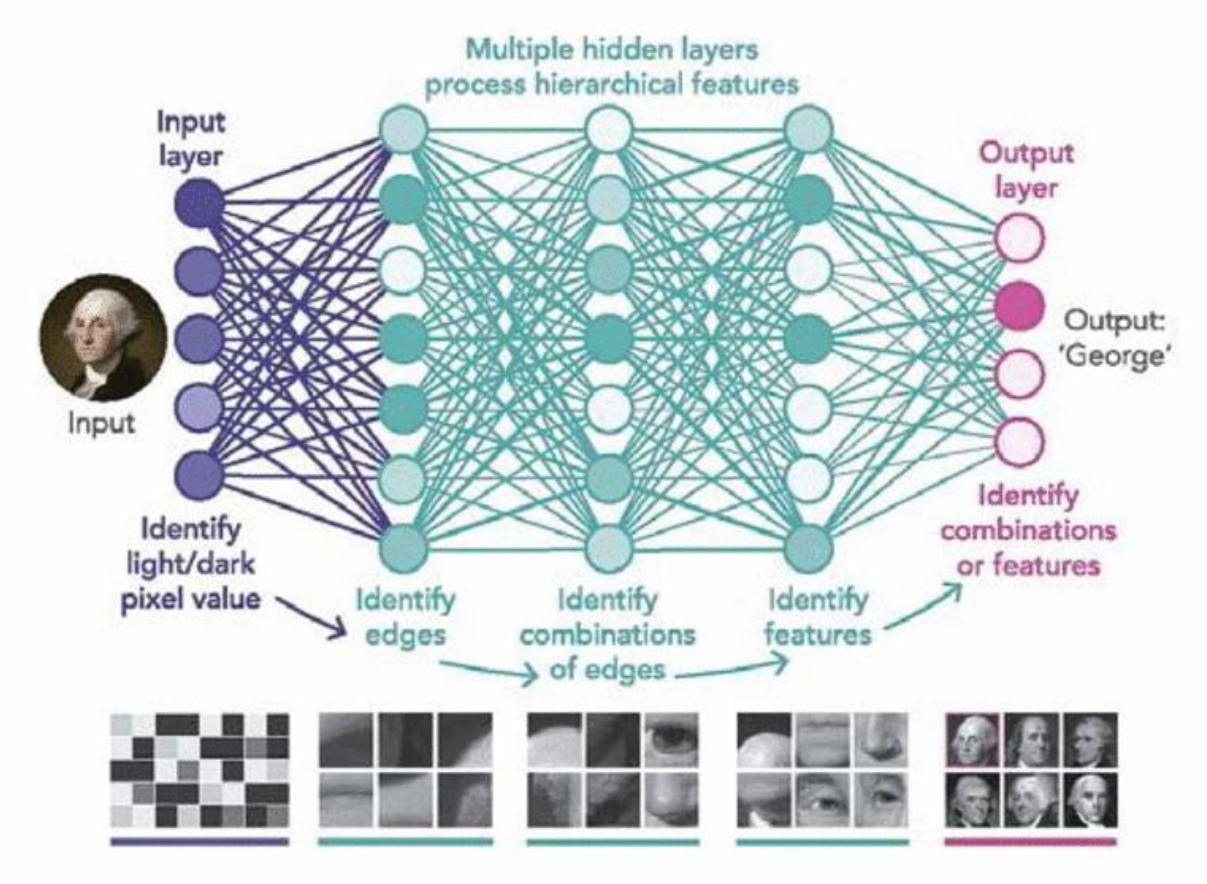


Abbildung 16: Aufbau eines Deep-Learning-Systems [61]

In der aktuellen Forschung zu Deep-Learning-Systemen in der Spracherkennung werden verschiedene stochastische Modelle modifiziert und miteinander verknüpft, z.B. „feed-forward neural networks“ und HMM. Hierdurch werden statische Mustererkennungen und Sequenzmustererkennungen verknüpft [62, 63]. Dies ist primär ein Forschungsgebiet der Bioinformatik und nur sekundär der Humanmedizin ist. Die weitere Entwicklung hat jedoch auch einen hohen Stellenwert für die zukünftige medizinische Versorgung.

# 6 ZUSAMMENFASSUNG

Sprache ist ein wesentliches Element der Kommunikation und Interaktion innerhalb sozialer Gesellschaften. Die Sprachaudiometrie gilt als entscheidende Diagnostik für die quantitative Bemessung eines möglichen Hörschadens. Zur Überprüfung der Sprachverständlichkeit stehen im deutschsprachigen Raum eine große Anzahl verschiedener Testverfahren zur Verfügung, die das Verstehen von Einsilbern, Mehrsilbern oder ganzer Sätze testen.

Hörgeschädigte Personen haben beim Sprachverständnis vor allem in lauter Umgebung weitaus größere Probleme als Normalhörenden. Um diesen Umstand realistisch zu erfassen, wird mit dem Oldenburger Satztest (OLSA) bei schwerhörigen Patienten im Störgeräusch geprüft. Der OLSA ist ein adaptiver Sprachtest, der Pegel der Sprache wird hierbei entsprechend der Antwort der Testperson verändert.

Die automatische Reintonaudiometrie liefert ein genaues Maß der Hörschwelle, aber Validierungsdaten für andere audiometrische Verfahren sind weiterhin begrenzt. Zudem besteht eine große Diskrepanz zwischen dem Bedarf und der Kapazität von Audiologen für die Durchführung von Hörtests. Durch eine Automatisierung könnten Zeitersparnisse entstehen und mehr Patienten von der Diagnostik profitieren.

Sprecherabhängige Spracherkennungsprogramme werden vom Benutzer auf Eigenarten der Aussprache trainiert. In Abhängigkeit von der Erfahrung und dem Trainingsumfang können sehr hohe Erkennungsquoten erreicht werden. Die neuste Generation dieser Systeme müssen nicht mehr zwangsläufig trainiert werden. Neben dem Umfang und der Flexibilität des zugrunde liegenden Wortschatzes des Spracherkennungssystems, spielt auch die akustische Qualität des Inputs eine entscheidende Rolle. Die wichtigste Voraussetzung ist hierbei zweifelslos eine präzise, deutliche und flüssige Aussprache.

Wir konnten zeigen, dass die Auswertung des OLSA mit der Spracherkennungssoftware Dragon von Nuance Inc. signifikant schlechter war als die manuelle Auswertung durch geschultes Fachpersonal ($p<0,0001$). Die Auswertung der gesamten Studienkohorte ergab ein heterogenes Bild, mit Testreihen die deckungsgleich zur Auswertung durch die MTA waren und Testungen mit Abweichungen von über 50%. Limitierend für die statistische Auswertung waren die

teilweise zu kleinen Subgruppen. Für Folgestudien sollte sich auf genau definierte Probandengruppen konzentriert werden, um ausreichende Gruppengrößen rekrutieren zu können.

Für die gewünschte Fragestellung ist die Auswahl des geeigneten Spracherkennungssystem essenziell. Das „perfekte“ Spracherkennungsprogramm existiert nicht. Durch eine Trainingsphase kann die Qualität und Verwertbarkeit der Ergebnisse signifikant verbessert werden. In unserer Studie fehlte die Auswertung der größten Vergleichsgruppe, nämlich Menschen ohne Hörschädigung.

In den vergangenen Jahren wurden zunehmend Versuche unternommen, neuronale Netze für das akustische Modell zu verwenden. Moderne Spracherkennungssysteme, die auf Deep Learning aufsetzen, verbessern ihr Outcome kontinuierlich und liefern Erkennungsraten, die in Bereichen liegen, die auch ein gesunder Mensch erreichen würde. In der aktuellen Forschung zu Deep-Learning-Systemen in der Spracherkennung werden verschiedene stochastische Modelle modifiziert und miteinander verknüpft.

# 7 LITERATURVERZEICHNIS


1.) Gehörmesser zur Untersuchung der Gehörfähigkeit galvanisirter Taubstummen in besonderer Rücksicht auf die Erlernung der artikulirten Tonsprache. Daf. 1804
2.) Fowler, C.A. (1995). "Speech production". In J.L. Miller; P.D. Eimas (eds.). Handbook of Perception and Cognition: Speech, Language, and Communication. San Diego: Academic Press
3.) Hahlbrock KH. Uber Sprachaudiometrie und neue Wörterteste. Arch Ohren Nasen Kehlkopfheilkd. 1953;162(5):394-431.
4.) Deutsches Institut für Normung (1995) DIN 45621-1 Sprache für Gehörprüfung – Teil 1: Ein- und mehrsilbige Wörter. Beuth, Berlin
5.) Winkler, A., Holube, I. Test-Retest-Reliabilität des Freiburger Einsilbertests. HNO 64, 564–571 (2016)
6.) Kiessling J. Moderne Verfahren der Sprachaudiometrie Laryngorhinootologie. 2000 Nov;79(11):633-5.
7.) Kießling, J. Probleme erkennen und Fehler vermeiden. HNO Nachrichten 50, 28–41 (2020)
8.) Wagener KC, Kühnel V, Kollmeier B (1999) Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. Z Audiol 38:4–15
9.) Wagener K, Brand T, Kollmeier B (1999) Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests. Z Audiol 38:44–56
10.) Wagener K, Brand T, Kollmeier B (1999) Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests. Z Audiol 38:86–95
11.) Stellungnahme der Bundesärztekammer zum Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Hilfsmittel-Richtlinie: Freiburger Einsilbertest im Störschall vom 24. November 2016
12.) BAnz AT 12.06.2020 B3
13.) Kollmeier B (2008) Aktuelle und zukünftige Entwicklungen der Hörgerätetechnik. In: Kießling J, Kollmeier B, Diller G (Hrsg) Versorgung und Rehabilitation mit Hörgeräten. Georg Thieme, Stuttgart, S 131–152
14.) Kollmeier B, Lenarz T, Winkler A, Zokoll MA, Sukowski H, Brand T, Wagener KC (2011) Hörgeräteindikation und -überprüfung nachmodernen Verfahren der Sprachaudiometrie im Deutschen. HNO59(10):1012–1021
15.) Hey M., Vorwerk W., Langer J., Vorwerk K., Begall K.. 2003. Vergleich von Satztest im Störschall bei Cochlea Implantat Patienten. 6. Jahrestagung der Deutschen Gesellschaft fuer Audiologie, 1–3
16.) Hornsby BW, Naylor G, Bess FH. A Taxonomy of Fatigue Concepts and Their Relation to Hearing Loss. Ear Hear. 2016 Jul-Aug;37 Suppl 1(Suppl 1):136S-44S
17.) Mahomed F, Swanepoel de W, Eikelboom RH, Soer M. Validity of automated threshold audiometry: a systematic review and meta-analysis. Ear Hear. 2013 Nov-Dec;34(6):745-52.
18.) Margolis RH, Morgan DE. Automated pure-tone audiometry: an analysis of capacity, need, and benefit. Am J Audiol. 2008 Dec;17(2):109-13.
19.) L. Deng and D. Yu. Foundations and Trends in Signal Processing Vol. 7, Nos. 3–4 (2013) 197–387

20.) https://www.nuance.com/healthcare/provider-solutions/speech-recognition/dragon-medical-one.html [ausgerufen am 07.11.2020]
21.) L. Lamel, J.-L. Gauvain: Speech Recognition. Oxford Handbooks Online (Vol. 14). Oxford University Press, 2005.
22.) Pommergaard HC, Huang C, Burcharth J, Rosenberg J. Voice recognition software can be used for scientific articles. Dan Med J. 2015 Feb;62(2): A5012.
23.) https://shop.nuance.de/store/nuanceeu/de_DE/Content/pbPage.dragon_professional_individual?currency=EUR&pgmid=95401000&utm_source=google&utm_medium=cpc&utm_campaign=EHK_AO_2020_DragonPC_Ecom%2B%2F%2BMXD%2B%2F%2BB%2B%2F%2BBrand%2B%2F%2B_%2B%2F%2BPure%2BBrand%2B%2F%2BDE%2B_%2BDE%2B%2F%2B_%2B%2F%2BExact%2B%2F%2BDesktop&keyword=nuance_e&gclid=EAIaIQobChMIssmf24jB7QIVdIBQBh3RJQTkEAAYASAAEgJ9VPD_BwE [aufgerufen am 07.11.2020]
24.) Karl-Heinz Best: Längen von Komposita im Deutschen, in: Glottometrics 23, 2012, S. 1-6
25.) Stumpf C (1919) Zur Analyse geflüsterter Vokale. Beiträge zur Anatomie, Physiologie, Pathologie und Therapie des Ohres, der Nase und des Halses Bd. 12.
26.) Kent RD, Vorperian HK. Static measurements of vowel formant frequencies and bandwidths: A review. J Commun Disord. 2018 Jul-Aug;74:74-97.
27.) IPA. 1999. Handbook of the International Phonetic Association. Cambridge: Cambridge University Press.
28.) Kohler KJ (1995) Einführung in die Phonetik des Deutschen, 2. Aufl. Erich Schmidt, Berlin
29.) Siebs T (1930) Deutsche Bühnenaussprache, Hochsprache. Ahn, Köln
30.) Caroline Féry: Phonologie des Deutschen. Universität Potsdam, Potsdam 2001, S. 45f
31.) Heidrun Pelz: Linguistik. 1996, S. 216
32.) Alfred Raab, Homophone der deutschen Sprache, Nürnberg: rab-Verlag, 1971
33.) Meibauer: Einführung in die germanistische Linguistik. 2. Aufl. 2007, S. 193
34.) Kalhara, P., et al. (2017). TreeSpirit: Illegal logging detection and alerting system using audio identification over an IoT network.
35.) Rose RC, Juang BH. Hidden Markov models for speech and signal recognition. Electroencephalogr Clin Neurophysiol Suppl. 1996; 45:137-52.
36.) Schuster-Böckler B, Bateman A. An introduction to hidden Markov models. Curr Protoc Bioinformatics. 2007 Jun;Appendix 3:Appendix 3A
37.) Lee LM, Le HH, Jean FR. Improved model adaptation approach for recognition of reduced-frame-rate continuous speech. PLoS One. 2018 Nov 7;13(11)
38.) Poritz, A. B., "Linear Predictive Hidden Markov Models and the Speech Signal," Proc. ICASSP '82, pp. 1291-1294, Paris, France, May 1982.
39.) Levinson, S. E., Rabiner, L. R., and Sondhi, M. M., "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," B.s. T.}., Vol. 62, No.4, Part 1, pp.1035-1074, April 1983.
40.) Juang, B. H., and Rabiner, l. R., "Mixture Autoregressive Hidden Markov Models for Speech Signals," Transactions on IEEE/Acoustics, Speech, and Signal Processing, Vol. ASSP-33, No.6, pp. 1404-1413, Dec. 1985.
41.) Helmut Meier: Deutsche Sprachstatistik. Zweite erweiterte und verbesserte Auflage. Olms, Hildesheim 1967, S. 336–339

42.)https://wiki.infowiss.net/Datei:VorgehensweiseSpracherkennung.jpg [ausgerufen am 07.11.2020]
43.)O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP). 2012
44.)Boenninghaus,H.-G., D. Röser: Neue Tabellen zur Bestimmung desprozentualen Hörverlustes für das Sprachgehor. Z. Laryng. Rhino!.52 (1973) 153
45.)Oldenburger Satztest Bedienungsanleitung für denmanuellen Test auf Audio-CD (KM-20091004-2, Version 1.0 vom 21.09.2011) Copyright © 2011 HörTech gGmbH Oldenburg
46.)Mandel MA. A commercial large-vocabulary discrete speech recognition system: DragonDictate. Lang Speech. 1992 Jan-Jun;35 ( Pt 1-2):237-46.
47.)Hoppe U, Hast A, Hocke T. Sprachverstehen mit Hörgeräten in Abhängigkeit vom Tongehör. HNO. 2014 Jun;62(6):443-8
48.)Salthouse TA. Selective review of cognitive aging. J Int Neuropsychol Soc. 2010 Sep;16(5):754-60.
49.)Gray WD, Hills T. Does cognition deteriorate with age or is it enhanced by experience? Top Cogn Sci. 2014 Jan;6(1):2-4.
50.)Kurz A, Grubenbecher M, Rak K, Hagen R, Kühn H. The impact of etiology and duration of deafness on speech perception outcomes in SSD patients. Eur Arch Otorhinolaryngol. 2019 Dec;276(12):3317-3325
51.)Cohen SM, Svirsky MA. Duration of unilateral auditory deprivation is associated with reduced speech perception after cochlear implantation: A single-sided deafness study. Cochlear Implants Int. 2019 Mar;20(2):51-56.
52.)Müller A, Hocke T, Hessel H, Mir-Salim P. Audiologische Ergebnisse bei transkranieller CROS-Versorgung in Abhängigkeit von der Ertaubungsdauer. Laryngorhinootologie. 2017 May;96(5):293-298.
53.)Beyea JA, McMullen KP, Harris MS, Houston DM, Martin JM, Bolster VA, Adunka OF, Moberly AC. Cochlear Implants in Adults: Effects of Age and Duration of Deafness on Speech Recognition. Otol Neurotol. 2016 Oct;37(9):1238-45.
54.)Goethe JW (1817) Paralipomenon zu Deutsche Sprache. I 41:466
55.)Mohr DN, Turner DW, Pond GR, Kamath JS, De Vos CB, Carpenter PC. Speech recognition as a transcription aid: a randomized comparison with standard transcription. J Am Med Inform Assoc. 2003 Jan-Feb;10(1):85-93.
56.)Erdel T, Crooks S. Speech recognition technology: an outlook for human-to-machine interaction. J Healthc Inf Manag. 2000 Summer;14(2):13-21.
57.)Sotscheck, 1.: (1982), »Ein Reimtest für Verständlichkeitsmessungen mit deutscher Sprache als ein verbessertes Verfahren zur Bestimmung der Sprachübertragungsgüte«, Der Fernmelde-Ingenieur 36, Heft 4/ 5, I
58.)Alhanbali S, Dawes P, Lloyd S, Munro KJ. Hearing Handicap and Speech Recognition Correlate with Self-Reported Listening Effort and Fatigue. Ear Hear. 2018 May/Jun;39(3):470-474.
59.)Zhang L, Zhao Z, Ma C, Shan L, Sun H, Jiang L, Deng S, Gao C. End-to-End Automatic Pronunciation Error Detection Based on Improved Hybrid CTC/Attention Architecture. Sensors (Basel). 2020 Mar 25;20(7):1809
60.)L. Deng and J. Chen. Sequence classification using the high-level features extracted from deep neural networks. In Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP). 2014

61.) Deep Learning in Multimedia. Available from: https://www.fkt-online.de/archiv/artikel/2020/fkt-5-2020/27807-deep-learning-in-multimedia/ [aufgerufen am 05.11.2020]
62.) Y. Bengio. Artificial neural networks and their application to sequence recognition. Ph.D. Thesis, McGill University, Montreal, Canada, 1991
63.) O. Abdel-Hamid, L. Deng, D. Yu, and H. Jiang. Deep segmental neural networks for speech recognition. In Proceedings of Interspeech. 2013

# 8 LEBENSLAUF

## PERSONALIEN

| | |
|---|---|
| Name und Vorname: | Warken, Christian |
| Geburtsdatum: | 14.02.1990 |
| Geburtsort: | Mannheim |
| Familienstand: | Verheiratet |
| Vater: | Warken, Horst |
| Mutter: | Warken, Regine |

## SCHULISCHER WERDEGANG

| | |
|---|---|
| 2000 – 2009 | Peter-Petersen-Gymnasium Mannheim |
| (22.06.2009) | Abitur (Note: 1,6) |

## UNIVERSITÄRER WERDEGANG

| | |
|---|---|
| WS 2010/2011 | Beginn des Studiums der Humanmedizin<br>An der Medizinischen Fakultät der Universität Heidelberg |
| (07.09.2012) | 1. Abschnitt der Ärztlichen Prüfung (M1) |
| 2012 – 2015 | Hauptstudium |
| WS 2013/2014 | Beginn des Studiums in Health Economics<br>Am Mannheim Institute for Public Health (MIPH) |
| (15.10.2015) | 2. Abschnitt der Ärztlichen Prüfung (M2) |
| 2015 – 2016 | Praktisches Jahr |
| (09.11.2016) | 3. Abschnitt der Ärztlichen Prüfung (M3), Gesamtnote: 1,83 |
| 08.01.2018 | Master of Science in Health Economics, Note: 1,7 |

# 9 DANKSAGUNG

Meinen Eltern, Regine und Horst, danke ich von ganzem Herzen. Ohne sie wäre ich nicht da, wo ich jetzt bin.

Frau Prof. Dr. med. Nicole Rotter, danke ich für die Annahme als Doktorand, die Überlassung des Themas und die Möglichkeit ihre Klinik für die Durchführung der Arbeit nutzen zu können.

Außerdem gilt mein Dank Frau Sylvia Büttner für die großartige Unterstützung bei der statistischen Auswertung.

Ich danke auch den freiwilligen Probanden für ihre Geduld und Teilnahme an den Tests, ohne die diese Arbeit nicht möglich gewesen wäre.

Ebenso gebührt mein Dank Johannes Burkart und Dr. Tobias Balkenhol für die fachliche Betreuung und ihr stets offenes Ohr bei Fragen.

Ein außerordentlicher Dank gilt PD Dr. med. Angela Schell für die stete Unterstützung und Hilfe, nicht nur bei Anfertigung dieser Arbeit.

Besonderen Dank schulde ich meiner geliebten Melanie, die mir jeden Tag hilft ein besserer Mensch zu werden und mir den nötigen Antrieb gab diese Arbeit zu vollenden.