

---

# FÖRDERPREIS 2024

---

## **Evaluierung eines auf Basis der Worthäufigkeitsverteilung neu erstellten Sprachmaterials für den Freiburger Einsilbertest**

*Update and Evaluation of the Freiburg Monosyllabic Speech Test  
on the Basis of Word Frequency Distribution*

Masterarbeit im Rahmen des Studiengangs Hörakustik und Audiologische Technik  
der Universität zu Lübeck

vorgelegt von: Maren Harries  
ausgegeben und betreut von: Dr. Hendrik Husstedt  
2. Betreuung durch: Prof. Dr. rer. nat. habil. Jonas Obleser



Die Masterarbeit ist im Rahmen einer Tätigkeit im Deutschen Hörgeräte Institut, mit Unterstützung von Larissa Warkentin, entstanden.



UNIVERSITÄT ZU LÜBECK  
INSTITUT FÜR SIGNALVERARBEITUNG

# Evaluierung eines auf Basis der Worthäufigkeitsverteilung neu erstellten Sprachmaterials für den Freiburger Einsilbertest

Update and Evaluation of the Freiburg Monosyllabic Speech Test on the Basis  
of Word Frequency Distribution

## Masterarbeit

im Rahmen des Studiengangs  
**Hörakustik und Audiologische Technik**  
der Universität zu Lübeck

vorgelegt von  
**Maren Harries**

ausgegeben und betreut von  
**Dr. Hendrik Husstedt**

2. Betreuung durch  
**Prof. Dr. rer. nat. habil. Jonas Obleser**

Die Masterarbeit ist im Rahmen einer Tätigkeit im Deutschen Hörgeräte Institut, mit Unterstützung  
von Larissa Warkentin, entstanden.

Lübeck, den 29. Mai 2024

**Erklärung zur Master-Abschlussarbeit**

Ich versichere an Eides statt, die vorliegende Arbeit selbständig und nur unter Benutzung der angegebenen Quellen und Hilfsmittel angefertigt zu haben.

Lübeck, den 29. Mai 2024

.....

Maren Harries

Ich bin damit einverstanden, dass meine Arbeit veröffentlicht wird, insbesondere dass die Arbeit Dritten zur Einsichtnahme vorgelegt wird oder Kopien der Arbeit zur Weitergabe an Dritte angefertigt werden.

Lübeck, den 29. Mai 2024

.....

Maren Harries

## Zusammenfassung

Der Freiburger Einsilbertest (FET), entwickelt 1953 von Karl-Heinz Hahlbrock, ist im deutschsprachigen Raum einer der am weitesten verbreiteten standardisierten Sprachtests für die Audiometrie und Hörsystemanpassung. Trotz seiner häufigen Nutzung wurden in letzter Zeit verstärkt Kritikpunkte laut, die seine allgemeine Anwendbarkeit betreffen, insbesondere im Hinblick auf die Aktualität, die Artikulation sowie die phonemische und perzeptive Äquivalenz der Testlisten. Im Rahmen dieser Masterarbeit wird eine Überarbeitung des FET vorgestellt, welche kommerzielle Text-to-Speech (TTS)-Technologie einsetzt, um den Test durch eine automatisierte Auswahl aktuell häufig genutzter Einsilber an den modernen Sprachgebrauch anzupassen. Unter Beibehaltung der phonemischen Struktur nach Hahlbrock und des Vergleichs der Phonemverteilung mit statistischen Analysen der deutschen Sprache, strebt die Arbeit an, die Ausgewogenheit der phonemischen und perzeptiven Äquivalenz der neuen Testlisten zu evaluieren und optimieren. In einer Studie mit 27 normalhörenden Teilnehmenden wurden erste psychometrische Funktionen und Sprachverstehenswerte ermittelt. Die Auswertung der neu entwickelten Testlisten bestätigt größtenteils die Erkenntnisse von vorangegangenen Arbeiten bezüglich der Sprachverständlichkeitsschwelle in dB bei 50% (SRT50) und der Steigung. Dabei weist das neue synthetische Testmaterial ebenfalls eine leicht höhere Steigung als der ursprüngliche FET auf. Es wurden nur geringe Abweichungen der SRT50-Werte festgestellt, teilweise ohne signifikante Unterschiede zur aktuellen Literatur. Weiterhin zeigt die Untersuchung ein erhebliches Optimierungspotenzial hinsichtlich der phonemischen und perzeptiven Äquivalenz. Dabei scheint eine hohe Äquivalenz zwischen den Testlisten und eine Annäherung an die Phonemverteilung deutschsprachiger einsilbiger Substantive möglich. Eine Phonemverteilung wie im allgemeinen deutschen Sprachgebrauch ist aufgrund der unterschiedlichen phonemischen Verteilung zwischen ein- und mehrsilbigen Wörtern allerdings nur begrenzt realisierbar. Dies verdeutlicht die damit einhergehende Schwierigkeit, alle typisch deutschen Phonemanteile im richtigen Verhältnis in standardisierte Sprachtests zu integrieren. Die kontinuierliche Anpassung des FET an sprachliche Entwicklungen sollte durch den Einsatz automatisierter Verfahren und basierend auf diesen Forschungsergebnissen intensiviert werden. Die Ergebnisse der Arbeit zeigen, dass der Einsatz von TTS-Technologie zusammen mit einer Optimierung der phonemischen und perzeptiven Äquivalenz eine Aktualisierung des Sprachmaterials des FET möglich machen.

## **Abstract**

The Freiburger Monosyllabic Speech Test (FMST), developed in 1953 by Karl-Heinz Hahlbrock, is one of the most widely used standardized speech tests for audiometry and hearing aid fitting in German-speaking areas. Despite its widespread use, recent criticisms have raised concerns about its general applicability, especially regarding its actuality, articulation, and the phonemic and perceptual equivalence of its test lists. This master's thesis presents a Text-to-Speech-revision of the FMST, utilizing commercial TTS-technology to adapt the test by automatically selecting frequently used monosyllables to match modern language use. The objective of this study is to evaluate and optimize the phonemic and perceptual equivalence of the new test lists, while retaining the phonemic structure according to Hahlbrock and comparing the phoneme distribution as statistically analyzed by the German language. In a study comprising 27 participants with normal hearing, initial psychometric functions and speech recognition scores were determined. The evaluation of the newly developed test lists largely confirms the findings of previous studies and literature regarding the SRT50 and the slope. The new synthetic test material also shows a slightly higher slope than the original FMST. Only minor deviations in SRT50 values were found, some without significant differences from the current literature. Furthermore, the study shows considerable potential for optimization with regard to phonemic and perceptual equivalence. A high degree of equivalence between test lists and alignment with the phoneme distribution of German monosyllabic nouns seems possible. However, due to the different phonemic distributions between monosyllabic and polysyllabic words, a phoneme distribution similar to general German usage seems feasible only to a limited extent. The continuous adaptation of the FMST to linguistic developments should be intensified through the use of automated processes based on these research findings. The results of this work show that the use of TTS-technology, together with an optimization of phonemic and perceptual balance, allows an update of the speech material of the FMST.

## Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>1</b>
<b>2</b>	<b>Grundlagen</b>	<b>3</b>
2.1	Aufbau des Freiburger Einsilbertests . . . . .	3
2.2	Optimierungsbedarf des Freiburger Einsilbertests . . . . .	4
2.2.1	Aktualität der Einsilber . . . . .	4
2.2.2	Perzeptive Äquivalenz . . . . .	4
2.2.3	Phonemische Äquivalenz . . . . .	6
2.3	Psychometrische Funktion in der Sprachaudiometrie . . . . .	8
2.4	Aktuelle deutsche Sprachkorpora . . . . .	9
2.4.1	Dudenkorpus . . . . .	9
2.4.2	Datenbank für gesprochenes Deutsch (DSG) . . . . .	9
2.4.3	Leipzig Corpora Collection (LCC).....	10
2.4.4	Deutsche Referenz Wortlisten (DeReWo).....	10
2.4.5	Digitales Wörterbuch Deutscher Sprache DWDS.....	10
2.4.6	Centre for Lexical Information (CELEX).....	12
2.4.7	Wikipedia Einsilbersammlung.....	12
2.5	Phonemaufteilung.....	12
2.6	Text-to-Speech Synthese und Anwendung.....	14
<b>3</b>	<b>Methodik</b>	<b>16</b>
3.1	Übersicht zum Vorgehen der Testlisterstellung.....	16
3.2	Sichtung und Auswahl der Einsilber.....	17
3.3	Listenerstellung unter Berücksichtigung der Phonemstruktur.....	18
3.4	Synthetisierung der Einsilberauswahl.....	20
3.5	Subjektive Bewertung der synthetischen Einsilber.....	21
3.6	Evaluation der neuen Testlisten.....	22
3.6.1	Messaufbau.....	22
3.6.2	Probandenkollektiv.....	24
3.6.3	Studiendesign.....	25
3.7	Optimierungsziele für den Freiburger Einsilbertest.....	26
<b>4</b>	<b>Ergebnisse</b>	<b>27</b>
4.1	Beschreibung der Einsilbersammlung.....	27
4.2	Subjektive Bewertung der synthetisch erzeugten Einsilber.....	28
4.3	Beschreibung der neuen Testlisten.....	30
4.3.1	Perzeptive Äquivalenz.....	34
4.3.2	Korrelationsanalyse von Bewertung und Sprachverstehen.....	39
4.4	Optimierung.....	40

---

<b>5 Diskussion</b>	<b>43</b>
5.1 Auswahl der Einsilber und Vergleich zum Freiburger Einsilbertest .....	43
5.2 Listenzusammenstellung .....	44
5.3 Bewertung der Synthese des aktualisierten Einsilbertests.....	46
5.4 Psychometrische Funktionen der neuen Testlisten .....	46
5.5 Optimierung der phonemischen und perzeptiven Äquivalenz.....	47
<b>6 Fazit und Ausblick</b>	<b>51</b>
Literaturverzeichnis .....	55
<b>7 Anhang</b>	<b>56</b>
7.1 Psychometrische Funktionen aller Listen aus der Pilotstudie.....	56
7.2 Korrelation des Einzelwortverstehens mit der Tokenanzahl und der Phonemanzahl.....	57
7.3 Wortfrequenzfilterung der seltenen Einsilber im FET.....	58
7.4 Übersicht zur Wortverständlichkeit der aktualisierten Testlisten .....	59
7.5 Auflistung der erstellten Testlisten .....	62
7.6 Auflistung der perzeptiv optimierten Testlisten .....	65
7.7 IPA-Tabelle .....	68

**Abkürzungsverzeichnis**

<b>API</b>	Application Programming Interface
<b>CELEX</b>	Centre for Lexical Information
<b>BAS</b>	Bavarian Archive for Speech Signals
<b>DeReKo</b>	Deutsches Referenz Korpus
<b>DeReWo</b>	Deutsche Referenz Wortformenliste
<b>DGD</b>	Datenbank für gesprochenes Deutsch
<b>DNN</b>	Deep Neural Network
<b>DWDS</b>	Digitale Wörterbuch der deutschen Sprache
<b>DHI</b>	Deutsches Hörgeräte Institut
<b>FET</b>	Freiburger Einsilbertest
<b>FMT</b>	Freiburger Mehrsilbertest
<b>FMST</b>	Freiburger Monosyllabic Speech Test
<b>FOLK</b>	Forschungs- und Lehrkorpus Gesprochenes Deutsch
<b>G2P</b>	Grapheme to Phoneme
<b>GUI</b>	Graphical User Interface
<b>IPA</b>	International Phonetic Alphabet
<b>LCC</b>	Leipzig Corpora Collection
<b>RMS</b>	Root Mean Square
<b>SAMPA</b>	Speech Assessment Methods Phonetic Alphabet
<b>SRT</b>	Speech Reception Treshold (Sprachverständlichkeitsschwelle)
<b>SRT50</b>	Sprachverständlichkeitsschwelle in dB bei 50%
<b>SPL</b>	Sound Pressure Level (Schalldruckpegel)
<b>TTS</b>	Text-to-Speech

## Abbildungsverzeichnis

1	Differenz zwischen dem erwarteten Sprachverstehen (SV) von 50 % und dem erreichten SV der Listen in Ruhe mit zwei verschiedenen Kopfhörern und im Freifeld. Nachdruck von: I. Baljić et al. aus der HNO-Zeitschrift, 64:572–583, 2016 . . . . .	5
2	Differenz zwischen dem erwarteten Sprachverstehen (SV) von 50 % und dem erreichten SV der Listen im Störgeräusch. Die maximale Differenz von $\pm 6,6$ % ist als gestrichelte Linie dargestellt. Nachdruck von: A. Winkler et al. aus der HNO-Zeitschrift, 68:14–24, 2020 . . . . .	5
3	Prozentualer Anteil der Konsonanten in Speech Assessment Methods Phonetic Alphabet (SAMPA) im FET (rote Kreuze) über alle 20 Listen sowie nach der Literaturstatistik von Kohler (schwarze Kreise) als Mittelwerte und FET-Streubereich (rot umrahmt). Nachdruck von: M. Exter et al. aus der HNO-Zeitschrift, 64:557–563, 2016 . . . . .	7
4	Jede Liste des FET besteht aus 73 Phonemen mit Darstellung der Wortlängen und Vokalstellung in den einzelnen, phonemisch gleichen Gruppen des Verständnistests. Jedes Quadrat stellt ein Phonem dar, die Vokale sind schwarz markiert . . . . .	8
5	DWDS Wortfrequenzbarometer mit sieben Häufigkeitsstufen (0 sehr selten bis 6 sehr häufig), mit ihren jeweiligen Intervallen und Beispielen, Stand 21.10.2022.....	11
6	Prozess der Testlistenerstellung und Evaluation. ....	16
7	MATLAB Graphical User Interface (GUI) zur Bewertung der aktualisierten Wortsammlung für die 27 neuen Listen mit 540 Wörtern. ....	21
8	Messaufbau bestehend aus einem Lautsprecher in 1 Meter Entfernung bei 0 Grad. ....	22
9	Durchschnittliches Reintonhörvermögen der 27 Studienteilnehmer mit Fehlerbalken, welche die Standardabweichung pro Messfrequenz angeben. ....	24
10	MATLAB GUI zur Messung des aktualisierten Freiburger Einsilbertests.....	25
11	Vergleich der Wortfrequenzverteilung aus den Sprachkorpora und FET.....	28
12	Mittelwerte der Schulnotenbewertung der Synthesequalität von 540 Wörtern.....	29
13	Phonemklassenvergleich der Einsilber aus Sprachkorpora und FET.....	32
14	Beispielhaft Testliste 10 mit IPA-Phonemschriftdarstellung und Verteilung der Phoneme	33
15	Phonemische Verteilung der Einsilber-Testlisten im Vergleich mit den Literatur nach Kohler.....	34
16	Psychometrische Funktionen der 27 Listen für die Schallpegel 21,5 dB, 27,5 dB und 33,5 dB mit Angabe der besten Liste 8 (lila) und der schlechtesten Liste 4 (schwarz). ....	35
17	Psychometrische Funktionen der 27 Listen im Literaturvergleich. Für die psychometrische Funktion des aktualisierten synthetischen FET (pink) sind zusätzlich die 95 %-Konfidenzintervalle der Sprachverständlichkeit pro Messpegel gezeigt. ....	36
18	Abweichungen der Listenverständlichkeit vom Mittelwert (in Prozentpunkten). ....	37
19	Häufigkeitsverteilung des prozentualen Einzelwortverstehens der 540 Einsilber aus den 27 Listen mit N=27 Probanden.....	38

---

20	Korrelation von Wortbewertung der TTS-Synthese und Sprachverstehen der 540 einsilbigen Substantive. ....	39
21	Phonemische Äquivalenz pro Liste vor (blau) und nach (rot) der Optimierung. ....	40
22	Psychometrische Funktionen der 27 Listen nach Optimierung der perceptiven Äquivalenz. ....	41
23	Darstellung der phonemische Verteilung inkl. Mittelwertlinie der 27 Listen für jede der neun Phonemklassen. ....	42
24	Psychometrische Funktionen der 27 Listen (Messschallpegel 28,5 dB und 35,5 dB) mit acht Probanden, Pilotstudienresultate für Pegelermittlung bei 20 %, 50 %, 80 %.....	56
25	Korrelation zwischen dem Einzelwortverstehen in Prozent und der Tokenanzahl. ....	57
26	Korrelation zwischen dem Einzelwortverstehen in Prozent und der Phonemanzahl.....	57
27	Balkendiagramm zum Einzelwortverstehen mit Wortnamen in Prozent sortiert nach Testlisten, Liste 1 bis 4.....	59
28	Balkendiagramm zum Einzelwortverstehen mit Wortnamen in Prozent sortiert nach Testlisten, Liste 5 bis 8.....	59
29	Balkendiagramm zum Einzelwortverstehen mit Wortnamen in Prozent sortiert nach Testlisten, Liste 9 bis 12.....	60
30	Balkendiagramm zum Einzelwortverstehen mit Wortnamen in Prozent sortiert nach Testlisten, Liste 13 bis 16.....	60
31	Balkendiagramm zum Einzelwortverstehen mit Wortnamen in Prozent sortiert nach Testlisten, Liste 17 bis 20.....	61
32	Balkendiagramm zum Einzelwortverstehen mit Wortnamen in Prozent sortiert nach Testlisten, Liste 21 bis 24.....	61
33	Balkendiagramm zum Einzelwortverstehen mit Wortnamen in Prozent sortiert nach Testlisten, Liste 25 bis 27.....	62
34	International Phonetic Alphabet (IPA)-Tabelle, Ausgabe 2020, mit den phonetischen Symbolen und ihrer Beschreibung.....	68

## Tabellenverzeichnis

1	Phonemverteilung in der deutschen Sprache nach K. J. Kohler.....	19
2	Auflistung der verwendeten Geräte.....	23
3	Tabellarische Übersicht über die Anzahl an automatisch extrahierten einsilbigen Substantiven aus den verschiedenen Sprachkorpora sowie die Gesamtsumme aller Korpora. ....	27
4	Bewertung TTS-Synthese der 26 schlecht synthetisierten (Note > 3) Einsilber vor und nach der Korrektur der Synthesequalität. ....	30
5	Tabellarische Übersicht von der finalen Phonemaufteilung, der neu extrahierten Einsilbersammlung (600) aus den verschiedenen Sprachkorpora. ....	31
6	Vergleich der Verteilung der Phonemanzahl in der neuen listenbezogenen Einsilbersammlung (540) mit den Vorgaben von Hahlbrock. ....	31
7	Vergleich der gemittelten Sprachverständlichkeitsschwelle und Steigung der psychometrischen Funktionen. ....	35
8	Auflistung der 80 veralteten Einsilber des FET der Wortfrequenz 1 und 2, sortiert nach Tokenmenge. ....	58
9	1. Version der Testlisten: Liste 1 bis Liste 9.....	62
10	1. Version der Testlisten: Liste 10 bis Liste 18.....	63
11	1. Version der Testlisten: Liste 19 bis Liste 27.....	64
12	2. Version: Testlisten perzeptiv optimiert: Liste 1 bis Liste 9.....	65
13	2. Version: Testlisten perzeptiv optimiert: Liste 10 bis Liste 18.....	66
14	2. Version: Testlisten perzeptiv optimiert: Liste 19 bis Liste 27.....	67

## 1 Einleitung

Der Freiburger Einsilbertest (FET), standardisiert nach der Norm DIN 45621-1:1995-08 [1], findet im deutschsprachigen Raum auch nach über 70 Jahren eine breite Anwendung [2]. Der FET umfasst 20 Listen mit jeweils 20 einsilbigen deutschen Substantiven und dient zur Ermittlung des prozentualen Sprachverstehens sowie zur Bewertung des Nutzens von Hörsystemen. Die Durchführung kann sowohl über Kopfhörer als auch über Lautsprecher erfolgen, wobei das Sprachverstehen in Ruhe oder im Störgeräusch als Prozentsatz korrekt wiederholter Wörter berechnet wird [1]. Die Ergebnisse des FET sind entscheidend für die Indikationsstellung einer Hörsystemversorgung und der Kostenübernahme durch die Krankenkassen. Validität und Reproduzierbarkeit von Sprachtests spielen deshalb in der audiologischen Diagnostik eine entscheidende Rolle.

In den letzten Jahren häufte sich Kritik aufgrund der Verwendung von unpopulären und veralteten Einsilbern. Zudem wurde die phonemische und perzeptive Äquivalenz der Testlisten in Frage gestellt [3]. Untersuchungen von Steffens haben gezeigt, dass rund 45 % der Einsilber, wie z.B. Lump oder Knecht, nicht mehr im allgemeinen Sprachgebrauch vorkommen [4]. Des Weiteren legen seine Forschungen nahe, dass die vermehrte Nutzung der Einsilber in der Schriftsprache möglicherweise zu einer Verzerrung der Messergebnisse bei gebildeten Personen mit hoher Lesekompetenz führen könnte. Auffällig erschien auch, dass einige Testlisten, wie die Listen 5 und 15, eine größere Häufigkeit in der deutschen Sprache aufweisen als andere, was jedoch nicht konsequent zu einer verbesserten Sprachverständlichkeit führte [4].

Veraltete Wörter können heutzutage problemlos anhand der Wortfrequenz in der deutschen Sprache identifiziert und aussortiert werden, allerdings stellt sie nach Winkler et al. keine eindeutige Korrelation bezüglich der Verständlichkeit her [5]. In Reaktion auf die geäußerte Kritik wurden bereits verschiedene Ansätze zur Überarbeitung des FET vorgeschlagen. So hat Felix Hahn in seiner Bachelorarbeit durch das Weglassen kritisch betrachteter Wörter und der Reduktion der Listenlängen eine Anpassung vorgenommen [6]. Diese Maßnahmen zielten darauf ab, veraltete oder kaum genutzte Wörter zu eliminieren und die Wortlisten durch Kürzung zu modernisieren. Allerdings führt diese Vorgehensweise zu einer erhöhten Messgenauigkeit und einer gesteigerten Wahrscheinlichkeit von Lerneffekten durch das wiederholte Testen derselben Wörter. Für eine Weiterentwicklung des Tests ist es deshalb sinnvoll, dem Test aktualisierte einsilbige Substantive hinzuzufügen und nicht zu reduzieren [6].

Außerdem wurde eine Neuaufnahme eines deutschsprachigen Einsilbertests 2007 durch Mahfoud [7] und dessen Evaluation 2010 durch Qualen [8] in Würzburg durchgeführt. Bei der Zusammenstellung des Materials wurde im Unterschied zum FET jegliche Form von einsilbigen Wörtern, wie Verben und Adjektive, genutzt. Der original FET hingegen arbeitet ausschließlich mit einsilbigen deutschen Substantiven [9]. Eine kostengünstige und nachhaltigere Variante zur Optimierung des FET kann in Zukunft ein Sprachtest sein, der synthetisch über ein Text-to-Speech Programm, im Folgenden TTS genannt, erstellt wird.

In einer Studie vom Deutschen Hörgeräte Institut durch Schwarz et al. wurde ein mit den original Listen exakt übereinstimmend TTS-generierter FET hinsichtlich der Sprachverständlichkeit mit einer Probandenstudie geprüft und mit dem ursprünglichen Freiburger Einsilbertest, eingesprochen

von Claus Wunderlich, verglichen [10]. Der Vergleich zwischen den psychometrischen Funktionen des FET mit synthetischem Testmaterial zeigte bei der mittleren Sprachverständlichkeitsschwelle (engl. Speech Reception Treshold, SRT) und der Steigung in einer Folgeuntersuchung keinen signifikanten Unterschied mehr zu den psychometrischen Funktionen des Originals [10]. Außerdem wurden in einer ergänzenden Praktikumsprojektarbeit von Schwarz et al. Einsilber aus drei Sprachkorpora Leipzig Corpora Collection (LCC), Datenbank für gesprochenes Deutsch (DGD) und Deutsches Referenz Korpus Deutsches Referenz Korpus (DeReKo) extrahiert und manuell nach Kategorien wie seltene Wörter, Namen, Regionalismen und Anglizismen gefiltert [11]. Basierend auf diesem systematischen Vorgehen, wurden vorbereitend für den FET, neue einsilbige Substantive mit Hilfe der Wortfrequenz, systematisch und zusätzlich automatisiert selektiert. Die möglichst vollständige Sammlung aller deutschen einsilbigen Substantive wurde aus verschiedenen gesprochenen und geschriebenen Korpora über eine in MATLAB integrierte Application Programming Interface (API) automatisch extrahiert, um die Objektivität zu erhöhen und die bisher umfangreiche Bearbeitungszeit und manuelle Fehler zu reduzieren. Auf dieser Basis wurden veraltete Wörter reduziert und neue Wörter ergänzt. Für die anschließende Erstellung neuer Testlisten, musste die phonemische Struktur umfangreich analysiert werden. Nur so war zu Beginn der Arbeit die vergleichende Literaturanalyse und Listengruppierung nach dem Phonemschema von Hahlbrock möglich [9].

Der Fokus der Masterarbeit liegt darauf, die automatisiert extrahierten, häufigen einsilbigen Substantive für die Testlistenstellung zu nutzen und den Aufbau der Phonemverteilung von Hahlbrock unverändert zu belassen. Dafür werden so viele Testlisten wie möglich mit je 20 Wörtern und jeweils 73 Phonemen pro Liste [9] erzeugt. Ausgehend davon soll im zweiten Schritt der Arbeit das cloudbasierte Acapela-TTS zur Signalerzeugung genutzt werden. Eine abschließende Evaluierung mit Probandenmessungen soll erfolgen, um Vergleiche und erste Optimierungsansätze zu erörtern. Die Forschungsfrage dieser Arbeit richtet sich darauf, inwiefern nach der Evaluation eine Verbesserung der perzeptiven Äquivalenz der Testlisten für den Freiburger Einsilbertest erreicht werden kann. Ein wesentliches Ziel ist es, die erlangten Ergebnisse für die Optimierung der perzeptiven und phonemischen Äquivalenz gemäß DIN EN ISO 8253-3 zukünftig zu nutzen [12] und eine Basis zu schaffen, die für die Weiterentwicklung hilfreich ist. Die Herausforderung in dieser Arbeit besteht darin, eine automatisierte Methode zur Aktualisierung und grundlegenden Optimierung von Sprachtests zu entwickeln, um die Effizienz zu steigern und die Objektivität der Ergebnisse zu gewährleisten. Eine solche Methodik kann dazu beitragen, die Ressourcen, die für die manuelle Anpassung und Bewertung benötigt werden, zu reduzieren und gleichzeitig eine konsistentere und reproduzierbare Qualität der Testergebnisse sicherzustellen. Dadurch soll die methodische Validität des FET gewahrt und gleichzeitig eine leicht anpassbare, zunächst perzeptiv ausgewogene und zeitgemäße Lösung entwickelt werden.

## 2 Grundlagen

### 2.1 Aufbau des Freiburger Einsilbertests

Der FET wurde von Karl-Heinz Hahlbrock bereits im Jahr 1953 entwickelt. Die heutzutage bekannte Audioaufnahme erfolgte erst 1969 und wurde mit dem Nachrichtensprecher Claus Wunderlich aufgenommen. Der Sprachtest besteht aus zwei Sprachmaterialien, dem Freiburger Mehrsilbertest (FMT) und dem FET. Mit dem FMT wird die Sprachverständlichkeitsschwelle in Dezibel und mit dem FET wird der Diskriminationsverlust für Sprache in Prozent bei jeweils unterschiedlichen Prüfpegeln gemessen. Beim FMT besteht das Testmaterial aus zehn Listen mit jeweils zehn mehrsilbigen Zahlwörtern von 13 bis 99. Der FET enthält 20 Listen mit 400 einsilbigen Substantiven in 20er-Listen angeordnet, wobei das Wort Schrift sowohl in Liste 2 als auch in Liste 14 vertreten ist. Auch Homophone, akustisch identisch ausgesprochene Wörter, wie z.B. Rat (r a: t) und Rad (r a: t) sind im original FET enthalten.

Die Listeneinteilung mit der standardisierten Reihenfolge der Wörter ist in DIN 45621-1:1995 [1] sowohl für den original FMT als auch für den FET abgebildet. Diese festgelegt Reihenfolge der Einsilber in den 20 FET Listen kann bei wiederholten Messungen Lerneffekt ergeben. Auch Wortassoziationen und Nachbarschaftseffekte zu einem darauffolgenden Wort, sollten möglichst vermieden werden [5], wie zum Beispiel bei den Wörtern Arm und Neid, die in diesem Fall beim FET in derselben Liste direkt hintereinander sind. Aufgrund der Digitalisierung des FET ist eine Randomisierung bei den meisten Messanlagen möglichen und Kontexteffekte und Lerneffekte deutlich unwahrscheinlicher, aber nicht ausgeschlossen. Die von der Physikalisch Technische Bundesanstalt Braunschweig festgelegte Referenzkurve nach DIN 45621-1:1995-08 für Sprachverständlichkeit bei monauralem Hören kann für binaurale Messungen (z.B. über Kopfhörer) angepasst werden, indem sie um 3 dB zu niedrigeren Pegeln verschoben wird, um den Lautstärkeanstieg beim binauralem Hören zu berücksichtigen [1]. Für die Messung des Diskriminationsverlusts einsilbige Substantive zu verwenden, lag die Überlegung durch Hahlbrock zugrunde, dass für das Verstehen von Sprache ein möglichst einheitlicher Stimulus zur Gewährleistung einer Test übergreifenden Vergleichbarkeit der Listen gewählt werden muss. Die Alternative statt einsilbiger Substantive Logatome zu verwenden wurde durch Hahlbrock ausgeschlossen, da laut seiner Aussage der Mensch dazu neigt dem Gehörten stets eine sinnvolle Bedeutung geben zu wollen. Außerdem begründet er, dass ein Test mit bedeutungslosen Logatomen den Probanden schneller ermüden lassen würde [9].

## **2.2 Optimierungsbedarf des Freiburger Einsilbertests**

### **2.2.1 Aktualität der Einsilber**

Bei der Erstellung des FET im Jahr 1953 legte Karl-Heinz Hahlbrock besonderen Wert darauf, dass die im Test verwendeten Wörter in der gesamten Bevölkerung, über alle Altersstufen, sozialen Schichten, Bildungsniveaus und Regionen hinweg, bekannt waren [9]. Aufgrund der großen interindividuellen Unterschiede der Zielgruppe für Sprachtests lassen sich solche Kriterien jedoch nicht vollständig erfüllen. Eine Untersuchung der im FET verwendeten Einsilber auf die Verwendungshäufigkeit in der aktuelleren Sprache durch Steffens, lies für mehrere Korpora für gesprochene und geschriebene Sprache große Differenzen in der Verwendungshäufigkeit der Freiburger Wörter erkennen [4]. Eine Stichprobe zur Wortfrequenz der FET Sammlung in den verwendeten Korpora bestätigte diese große Schwankungsbreite ebenfalls. In der Studie von Winkler et al. wird als Maß für die Bekanntheit der Wörter die Wortfrequenz angenommen. Sie untersuchten neben dem Einfluss der Wortfrequenz auf die mit dem FET gemessene Sprachverständlichkeit auch die Auswirkung der Nachbarschaftsdichte, welche die lexikalische Ähnlichkeit zu anderen Wörtern beschreibt. Mit ansteigender Hörbarkeit eines Wortes steigt auch der Einfluss Verwendungshäufigkeit, im Folgenden einheitlich Wortfrequenz genannt und Nachbarschaftsdichte für die Sprachverständlichkeit der Wörter an. Es stellte sich heraus, dass beide Parameter und damit auch die Auswahl der Testlisten Einfluss auf die Ergebnisse des FET hatten [5].

### **2.2.2 Perzeptive Äquivalenz**

Für die Beurteilung der perzeptiven Äquivalenz des Freiburger Einsilbertests (FET) ist vor allem die Äquivalenz zwischen den Testlisten von Bedeutung. Bisherige Untersuchungen von A. Winkler et al. [13] zeigen, dass der original FET große Unterschiede beim Sprachverstehen zwischen den einzelnen Listen aufweist. Dies gilt für Messungen im Störgeräusch und in Ruhe mit dem Unterschied, dass im Störgeräusch andere Listen die größten Abweichungen aufweisen als in Ruhe, siehe Abbildung 2. Die perzeptive Äquivalenz ist ausschlaggebend in der Testlistenerstellung, da sie direkt die Reproduzierbarkeit der Messergebnisse des Sprachverständlichkeit beeinflusst und bei großen Differenzen zwischen den Listen keine Vergleichbarkeit mehr besteht. Die Norm DIN EN ISO 8253-3:2022-11 [12] legt fest, wie die perzeptive Äquivalenz der Testlisten zu ermitteln ist, um sicherzustellen, dass die Ergebnisse nicht von der Auswahl spezifischer Listen abhängen.

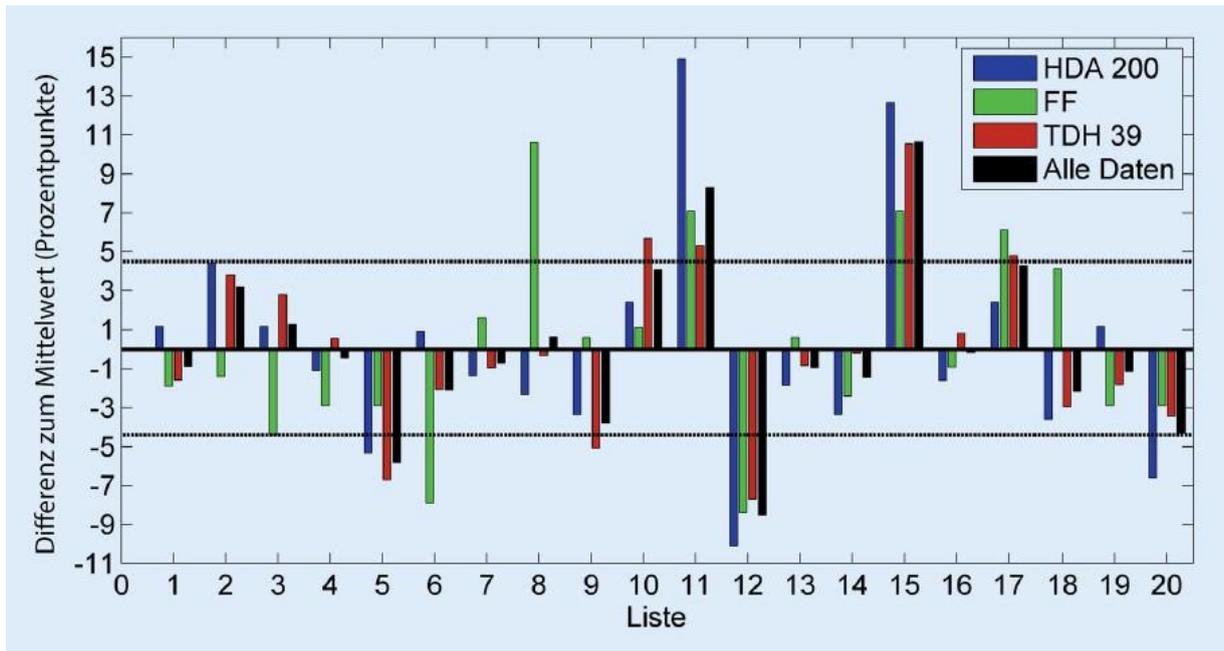


Abbildung 1: Differenz zwischen dem erwarteten Sprachverstehen (SV) von 50 % und dem erreichten SV der Listen in Ruhe mit zwei verschiedenen Kopfhörern und im Freifeld. Nachdruck von: I. Baljić et al. aus der HNO-Zeitschrift, 64:572–583, 2016 [14].

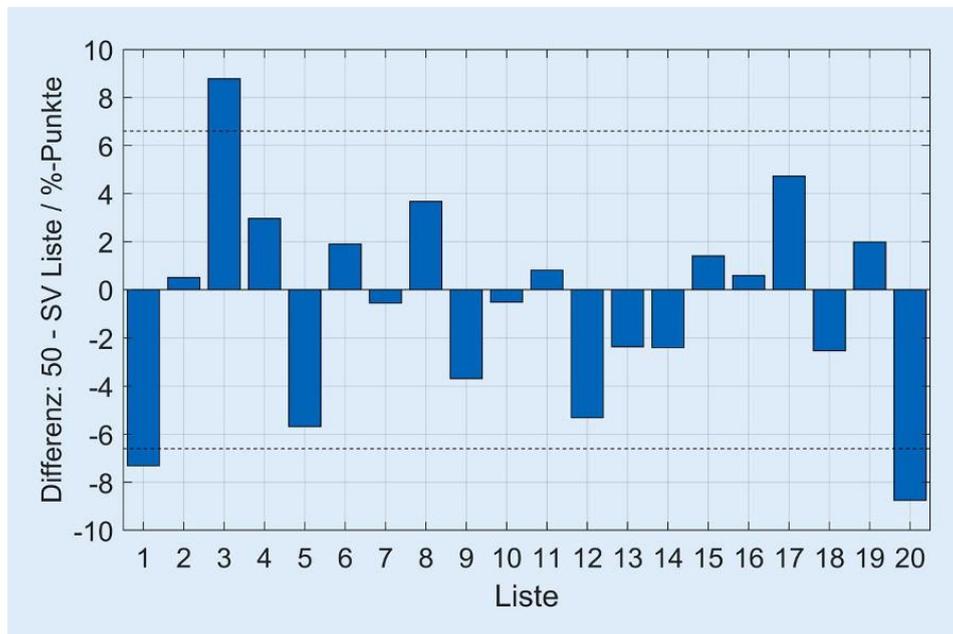


Abbildung 2: Differenz zwischen dem erwarteten Sprachverstehen (SV) von 50 % und dem erreichten SV der Listen im Störgeräusch. Die maximale Differenz von  $\pm 6,6$  % ist als gestrichelte Linie dargestellt. Nachdruck von: A. Winkler et al. aus der HNO-Zeitschrift, 68:14–24, 2020 [13].

Sie schreibt vor, dass für Listen, die auf ein optimales Sprachverständnis abzielen, die Gleichwertigkeit durch die Bestimmung des mittleren 95 Prozent-Konfidenzintervalls für die Sprachverständlichkeitswerte geprüft werden soll. Dieses Verfahren erfordert den Einsatz unterschiedlicher Testlisten bei denselben Probanden. Das 95-Prozent-Konfidenzintervall, das eine statistische Maßzahl für die Präzision der Messergebnisse darstellt, lässt sich mittels der Formel

$$KI(p) = 1.96 \cdot \sqrt{\frac{p \cdot (1 - p)}{n}} \quad (1)$$

berechnen [12]. Durch die Einhaltung dieser Richtlinien kann die Validität der Sprachtests erhöht und eine reproduzierbare und präzise Messung des Sprachverstehens sichergestellt werden.

### 2.2.3 Phonemische Äquivalenz

In jeder Liste des FET sind 73 Phoneme enthalten, die sich wie in Abbildung 4 dargestellt, auf die einsilbigen Substantive verteilen. Es gibt pro Liste jeweils ein Wort mit zwei Phonemen, sieben Wörter mit drei Phonemen, zehn Wörter mit vier Phonemen und zwei Wörter mit fünf Phonemen. Bei den Wörtern wird zusätzlich zur Phonemanzahl auch die Stellung des Vokals bzw. Diphthongs des Einsilbers beachtet. Phoneme sind die kleinsten bedeutungsunterscheidenden Lauteinheiten in einer Sprache, während Allophone die konkreten Lautvarianten sind, die ein Phonem haben kann [15]. Dies ist später wichtig für die Entscheidung der Phonemkategorisierung und -zählung. Jede Gruppe sollte phonemisch ähnlich zusammengesetzt sein. Ziel ist also eine phonemische Äquivalenz, so dass jeder Laut innerhalb einer Gruppe gleich häufig vorkommt. Abhängig von der Testart sollte bei einem standardisierten Sprachtest zusätzlich die Häufigkeit eines jeden Lautes in den Gruppen sein prozentuales Vorkommen in der alltäglichen deutschen Sprache widerspiegeln. Es hat sich herausgestellt, dass das Initialphonem eine besonders wichtige Bedeutung zur Verständlichkeit von Einsilbern liefert. Daher wurde bei der Zusammenstellung der FET Testlisten auf eine gleichmäßige Verteilung der initiiierenden Phonemanzahlgruppe geachtet, siehe Abbildung 4 [9]. So besitzt beispielsweise jede Liste des FETs ein Wort mit Vokal bzw. Diphthong als Initialphonem.

Das Verhältnis der Anzahl an Konsonanten zu Vokalen von 72,7 % zu 27,3 % im Test entspricht durch die ausschließliche Verwendung von einsilbigen Substantiven als Testmaterial nicht der deutschen Sprache. Diese weist ein Verhältnis der Konsonanten zu Vokalen von 61,3 % zu 38,2 % auf [9]. In Abbildung 3 werden die prozentualen Unterschiede zwischen der deutschen Sprachstatistik von Kohler und dem FET beispielhaft für einzelne Konsonanten in SAMPA aufgezeigt [16]. Es ist kaum möglich, eine phonemische Äquivalenz zwischen den statistischen Werten der deutschen Sprache und einem Test, der ausschließlich aus Einsilbern besteht, vollständig ausbalanciert herzustellen. Um die nach DIN EN ISO 8253-3:2022-11 geforderte phonemische Äquivalenz in den Testlisten bestmöglich herstellen zu können, wurde zunächst eine Zielverteilung der Phoneme benötigt. Außerdem wurde beachtet, dass die Listenstruktur mit 73 Phonemen über alle Listen, wie in Abbildung 4 dargestellt, entspricht. In der deutschen Sprache tragen Artikel, Pronomen und Mehrsilber, die eine deutlich häufigeres Vorkommen haben, zu einer Verschiebung des Konsonant- Vokalverhältnisses zugunsten der Vokale bei [9].

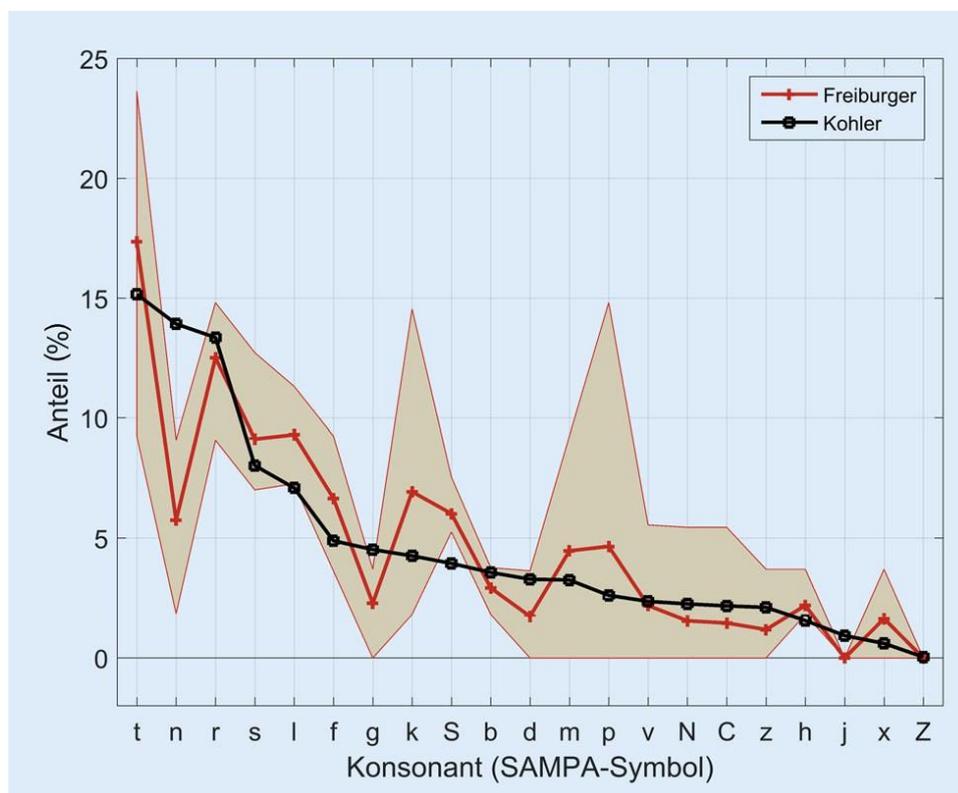


Abbildung 3: Prozentualer Anteil der Konsonanten in SAMPA im FET (rote Kreuze) über alle 20 Listen sowie nach der Literaturstatistik von Kohler (schwarze Kreise) [17] als Mittelwerte und FET-Streubereich (rot umrahmt). Nachdruck von: M. Exter et al. aus der HNO-Zeitschrift, 64:557–563, 2016 [16].

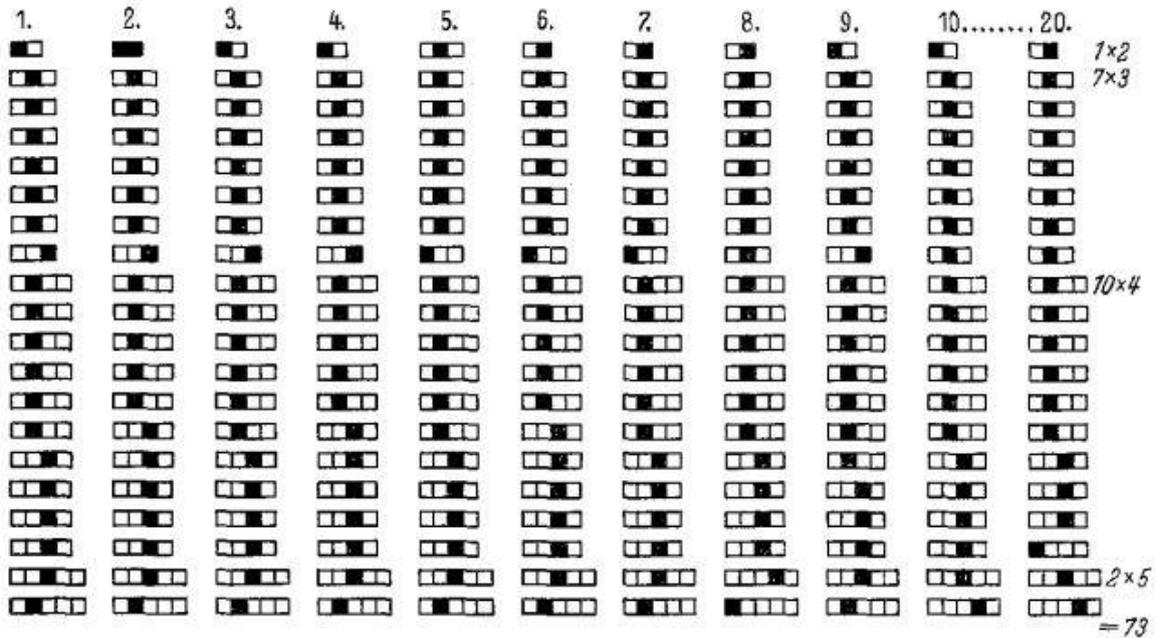


Abbildung 4: Jede Liste des FET besteht aus 73 Phonemen mit Darstellung der Wortlängen und Vokalstellung in den einzelnen, phonemisch gleichen Gruppen des Verständnistests. Jedes Quadrat stellt ein Phonem dar, die Vokale sind schwarz markiert [9].

### 2.3 Psychometrische Funktion in der Sprachaudiometrie

Die psychometrische Funktion beschreibt den Zusammenhang zwischen der Stärke des akustischen Reizes und der kognitiven Leistung des Probanden, ausgedrückt durch eine logistische Funktion.

Bei der Formel

$$f_{\text{sig}}(c_{50}, \beta) = \frac{1}{1 + e^{-\beta(\text{Stimulus-Level} - c_{50})}} \quad (2)$$

steht  $c_{50}$  für die Reizintensität bei 50 % Detektionswahrscheinlichkeit und  $\beta$  für die Steigung der Funktion.

Für Sprachtests spezifisch wird die Funktion oft so formuliert,

$$R = \frac{100}{1 + e^{0.04 \cdot s(L_{\text{SRT}} - L)}} \quad (3)$$

wobei  $L$  den Sprachpegel in Dezibel darstellt,  $s$  die Steigung der Sprachverständlichkeitskurve in Prozent pro Dezibel und  $L_{\text{SRT}}$  den Schalldruckpegel, bei dem 50 % Verständlichkeit erreicht wird [12].

## **2.4 Aktuelle deutsche Sprachkorpora**

### **2.4.1 Dudenkorpus**

Das Dudenkorpus ist eine umfangreiche Sammlung elektronischer Texte aus unterschiedlichen Quellen, wie Zeitungsartikeln, Romanen, Fachtexten aus Bereichen wie Medizin, Technik und Wirtschaft, sowie weiteren Genres. Dieses Korpus, das rund sechs Milliarden Wortformen umfasst, dient der Dudenredaktion als Basis für die Analyse und Dokumentation des aktuellen Sprachgebrauchs im Deutschen. Die Texte im Korpus fungieren als Belege bei der täglichen Arbeit der Duden Redaktion, beispielsweise zur Klärung von Fragen wie der korrekten Artikelverwendung, der Schreibweise, regionalen Sprachvarianten oder der Präferenz für bestimmte grammatische Konstruktionen [18].

Des Weiteren ermöglicht das Dudenkorpus es, die Häufigkeit bestimmter Wörter oder grammatikalischer Strukturen zu ermitteln und unterstützt die Redaktion bei der Identifikation und möglichen Aufnahme neuer Wörter in die Duden-Wörterbücher. Die Worthäufigkeitsangaben des Dudenkorpus sind aus einer digitalen Volltextsammlung mit über 5 Milliarden Wortformen aus verschiedenen Textsorten der letzten 25 Jahre und werden in fünf grafischen dargestellten Strich-Kategorien eingeteilt. Diese Informationen sind nur pro Wort einzeln über die Duden-Homepage abrufbar.

Im Gegensatz zum Digitalen Wörterbuch der deutschen Sprache (DWDS), das öffentlich über eine API und verschiedene Schnittstellen zugänglich ist und Nutzern direkten Zugriff auf Textdaten und Suchfunktionen bietet, ist das Dudenkorpus nicht öffentlich über eine API zugänglich. Die Textdaten des Dudenkorpus sind ausschließlich für die interne Nutzung durch die Dudenredaktion vorgesehen und für externe Nutzer weder über eine API noch in Textdateiform verfügbar. Die Nutzung dieser Daten ist daher nur indirekt möglich, über die Homepage oder gedruckte Bücher, ohne die Möglichkeit zur Automatisierung oder Integration in externe Systeme.

### **2.4.2 Datenbank für gesprochenes Deutsch (DSG)**

Die DGD wird vom Leibniz-Institut für Deutsche Sprache Mannheim betrieben und stellt die größte Datenbank für Korpora mit gesprochener deutscher Sprache dar. Die Datenbank enthält verschiedene Korpora aus real gesprochener Sprache. Die in der Datenbank abrufbaren Korpora lassen sich in Variationskorpora und Gesprächskorpora aufteilen. Mit den Variationskorpora wird versucht die dialektale Vielfalt von Variationen des gesprochenen Deutsch abzubilden. Mit den Gesprächskorpora wird hingegen versucht Sprache eher als empirisches Untersuchungsobjekt abzubilden.

Als Referenzkorpus der DGD kann der Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) angesehen werden. Der FOLK enthält Aufnahmen von spontanen Gesprächen und Interviews, die von verschiedenen Regionen und sozialen Gruppen stammen. Der Referenzkorpus wurde an der Universität Duisburg-Essen erstellt und dient der Untersuchung von Phänomenen des gesprochenen Deutsch in unterschiedlichen Kontexten. Zugriff auf die Daten erfolgt über das Online-Portal der DGD [19].

### **2.4.3 Leipzig Corpora Collection (LCC)**

Das LCC-Korpus ist eine Sammlung von Sprachkorpora, die an der Universität Leipzig erstellt und gepflegt wird [20]. Diese Korpora enthalten große Mengen an Texten in verschiedenen Sprachen und werden für linguistische Forschungszwecke verwendet. Die Korpora sind in mehreren Sprachen verfügbar, darunter Englisch, Deutsch, Französisch, Spanisch, Portugiesisch, Italienisch, Russisch, Niederländisch, Schwedisch und Polnisch. Sie werden von Forschern weltweit genutzt, um sprachliche Phänomene zu untersuchen und zu analysieren. Das Korpus wird durch die Universität Leipzig online über ein Portal sowie teilweise zum Download zur Verfügung gestellt. Das Projekt wurde durch Quasthoff im Jahr 1998 gestartet. Der freie Zugriff auf die Korpora ist seit dem Jahr 2006 möglich. Die Korpora enthalten Informationen zur Wortfrequenz, allerdings nicht in lemmatisierter Form. Dies führt dazu, dass neben der Wortgrundform auch Nebenformen, Rechtschreibfehler, Plurale und bei Verben Konjugationen im Korpus inkl. Wortfrequenz enthalten sind. Quellen, die für die Zusammenstellung der Korpora genutzt wurden, sind überwiegend Zeitungen, Wikipedia, öffentliche Internetauftritte und Nachrichtenfeeds. In weiteren Verarbeitungsschritten findet eine Auflösung der html-Formatierungen im Text statt, sowie eine Identifikation der richtigen Sprache, eine Segmentierung von Sätzen sowie eine Bereinigung der Daten. Bereinigt werden die Daten, indem alle Zeichenfolgen, die nicht in einer Satzstruktur stehen, nicht der zugeordneten Sprache entsprechen gelöscht werden. Außerdem findet ein Abgleich der Artikel mit bereits gesammelten Daten statt, sodass Artikel, die beispielsweise auf mehreren Seiten publiziert werden, nicht doppelt in die Datenbank aufgenommen werden.

### **2.4.4 Deutsche Referenz Wortlisten (DeReWo)**

Die Deutsche Referenz Wortformenliste (DeReWo) sind häufigkeitsbasierte Grund- und Wortformenlisten, die durch verschiedene Verarbeitungen des DeReKo generiert wurden [21]. Zusammengestellt wurden die Listen vom Leibniz-Institut für die deutsche Sprache. Das DeReKo besteht aus über drei Milliarden Wörtern, die aus Texten der Belletristik, Wissenschaft, Populärwissenschaft, Zeitungsartikeln und weiteren Textarten. Die Datierung der Texte geht zurück bis zum Jahr 1964. Bei den zusammengestellten Wortlisten handelt es sich um Wortformenlisten und Grundformenlisten. Die Grundformenlisten entsprechen der lemmatisierten Form. In den Ausführungen der Benutzerdokumentation für die DeReWo ist explizit erwähnt, dass ein Vergleich unterschiedlicher Wörterlemmastrecken oft zu unterschiedlichen Ergebnissen führen wird. Dies kann teilweise an unterschiedlichen Prinzipien bei der Zusammenstellung der Wortlisten begründet liegen, wie beispielsweise die Berücksichtigung von Fremdwörtern, Fachbegriffen, veralteten und verschiedenen Schreibweisen, unselbstständigen Morphemen, Einzelbuchstaben und Akronymen, Eigennamen, Wortreihen oder Kurzwörtern.

### **2.4.5 Digitales Wörterbuch Deutscher Sprache DWDS**

Das Digitale Wörterbuch der deutschen Sprache (DWDS) ist ein umfassendes online verfügbares Wörterbuch, das von der Berlin-Brandenburgischen Akademie der Wissenschaften herausgegeben wird. Es umfasst eine Sammlung von Wörtern und Begriffen aus der deutschen Sprache, einschließlich ihrer Bedeutungen, Statistiken, Aussprache, Etymologie und ihrer Verwendung in verschiedenen Kontex-

ten. Der zugrundeliegende Korpus beinhaltet aufsummiert 35 271 873 584 Tokens und ist damit der größte seiner Art [22]. Tokens sind die Anzahl der Vorkommen eines einzelnen Wortes oder eines bedeutungsunterscheidenden Zeichens (Phonem) in einer Korpus-Sammlung. Das DWDS verwendet vier verschiedene Korpora, den DWDS-Kernkorpus, Metakorpus WebXL, DWDS-Zeitungskorpus und den Wikipedia-Korpus.

Das DWDS basiert auf einer umfangreichen Sprachkorpus-Analyse inkl. Frequenz. Es ermöglicht daher, die Verwendung von Wörtern in verschiedenen historischen und aktuellen Kontexten zu untersuchen. Es ist in der Lage, die Verwendung von Wörtern und Begriffen in der deutschen Sprache im Laufe der Zeit zu verfolgen und ihre Bedeutungsverschiebungen und Entwicklungen aufzuzeigen. Die Datenbasis für die Worthäufigkeit (Wortfrequenzbarometer) bilden gegenwartssprachliche Korpora. Diese Funktionen können nicht nur über die Homepage, sondern auch über eine API-Abfrage genutzt werden. Die API-Abfrage und der gesamte Service des DWDS ist kostenlos online verfügbar, so dass jeder darauf zugreifen und nachschlagen kann. Das Wortfrequenzbarometer, bildet eine siebenstufige, logarithmische Skala ab. Es wird die Frequenz für alle Flexionsformen eines Wortes ermittelt. Entscheidend für die Berechnung sind sowohl die absolute Häufigkeit (Frequenz) des jeweiligen Wortes als auch das Verhältnis dieser Zahl zur Gesamtgröße des Korpus. Diese Angabe erfolgt nur für Wörter, die insgesamt mindestens fünfmal in oben genannten vier Korpora vorkommen. Die Korpora im DWDS werden stetig aktualisiert und daher sind die Worthäufigkeitsangaben variabel und zeitabhängig. Die nachfolgende Abbildung 5 listet die sieben Häufigkeitsstufen mit ihren jeweiligen Intervallen und entsprechenden Beispielen auf [22].

Skalenwert	Worthäufigkeit		Beispiel		Skala
	von	bis	Wort	Frequenz	
0	5	1 115	Kontorsion	609	
1	1 116	11 153	schwurbeln	1 940	
2	11 154	111 539	gutgläubig	27 586	
3	111 540	1 115 394	Bildschirm	621 437	
4	1 115 395	11 153 945	Krieg	3 224 836	
5	11 153 946	111 539 457	gut	60 353 035	
6	111 539 458	35 271 873 584	sein	549 233 351	

Abbildung 5: DWDS Wortfrequenzbarometer mit sieben Häufigkeitsstufen (0 sehr selten bis 6 sehr häufig) mit ihren jeweiligen Intervallen, Tokenzahl und entsprechenden Beispielen.

### 2.4.6 CELEX

Das CELEX-Sprachkorpus ist eine umfangreiche Datenbank, die zur linguistischen Forschung und Sprachverarbeitung verwendet wird. Es ist ein elektronisches Wörterbuch, das Informationen zu verschiedenen Sprachen und deren Struktur enthält. Die Datenbank enthält Informationen zu mehreren europäischen Sprachen, darunter Englisch, Deutsch, Niederländisch und Französisch. Sie enthält morphologische und syntaktische Informationen, wie z.B. Wortformen, Wortklassen, Stammformen und grammatische Funktionen. Darüber hinaus enthält die Datenbank auch semantische Informationen, wie z.B. Bedeutungen und Konnotationen von Wörtern. Die Informationen in der Celex-Datenbank werden von verschiedenen Quellen gesammelt, darunter Wörterbücher, Textkorpora und sprachliche Experimente. Sie wird oft in der Sprachforschung und -verarbeitung verwendet, um Sprachmodelle zu entwickeln und zu verbessern, die maschinelles Lernen und künstliche Intelligenz verwenden [23].

### 2.4.7 Wikipedia Einsilbersammlung

Die Einsilbersammlung auf Wikipedia ist eine Zusammenstellung von deutschen Wörtern, die nur aus einer Silbe bestehen und aus dem Wikipedia-Korpus extrahiert wurden. Diese Sammlung ist Teil des Projekts Wortschatz auf Wikipedia, das sich zum Ziel gesetzt hat, den Wortschatz der deutschen Sprache zu dokumentieren und zugänglich zu machen. In der Einsilber-Sammlung finden sich über 1 100 Wörter, die nur aus einer Silbe bestehen, wie zum Beispiel Hund, Baum, Maus, Licht, Wind oder Brot. Es handelt sich dabei um eine alphabetisch geordnete Liste, die ständig erweitert und aktualisiert wird [24].

## 2.5 Phonemaufteilung

In der hochdeutschen Sprache<sup>1</sup> werden insgesamt 40 Phoneme in 9 Phonemgruppen wie folgt unterteilt: [15], [25]:

- Lange Vokale:  
a:, e, e:, i, i:, o, o:, ø, ø:, u, u:, ε:, y, y:
- Kurze Vokale:  
a, ɑ, ɐ, ε, ɪ, ɔ, ʊ, ʏ, œ, ə
- Standard-Diphthonge (sonstige Vokale):  
aʊ, aɪ, ɔɪ
- Diphthonge mit vokalischem r (sonstige Vokale):  
aɐ, ɛɐ, ɪɐ, ɔɐ, œɐ, ʊɐ, ʏɐ

<sup>1</sup>Die hier präsentierte Liste enthält regionale Phoneme. Die Kategorisierung und Anzahl der Phoneme kann in der literarischen Fachwelt variieren. Eine einheitliche Festlegung existiert nicht. Die 40 Standardphoneme werden im Kapitel zur Methodik detailliert definiert.

- Plosive stimmlose Konsonanten:  
k, t, p
- Plosive stimmhafte Konsonanten:  
b, d, g
- Frikative stimmlose Konsonanten:  
f, s, ʃ, x, ç, ç, h
- Frikative stimmhafte Konsonanten:  
v, z, ʒ
- Nasale:  
m, n, ŋ
- Sonstige Konsonanten:  
l, j, r
- Standard-Affrikaten (sonstige Konsonanten):  
ts, tʃ, pf

Die Kategorisierung der deutschen Phoneme in lange und kurze Vokale, Diphthonge sowie Affrikate zeigt die phonetische Vielfalt der deutschen Sprache. Affrikaten, definiert als Konsonanten, die aus einer Plosiv- und einer Frikativ-Komponente bestehen und phonemisch als ein einziges Phonem wahrgenommen werden, umfassen die drei Standardaffrikate ts, tʃ, und pf [15]. Aufgrund der Verwendung von Anglizismen müssen in dieser Arbeit jedoch auch weitere Affrikate und vokalisierte Diphthonge wie dz und œ berücksichtigt werden, siehe auch Unterabschnitt 3.2.

Die Einordnung von Diphthongen, insbesondere jener mit einem vokalisiertem r, sowie von Affrikaten variiert in der Literatur und ist kontextabhängig [26]. Diphthonge und Affrikate werden generell als ein einzelnes Phonem betrachtet, wenn sie phonemisch untrennbar sind und innerhalb einer Silbe liegen [18].

Regionale und stilistische Variationen beeinflussen zudem die Aussprache des r in Deutschland. Im Norden und Osten des Landes sowie bei schneller Sprechweise wird das r nach einem kurzen Vokal vokalisiert. Im Gegensatz dazu neigen Sprecher in der Schweiz und in Österreich, besonders bei langsamer Sprechgeschwindigkeit und auch im Hochdeutschen, zur konsonantischen Aussprache des r [27]. Solche Phoneme werden als frei variierende Allophone bezeichnet. Sie können in derselben Umgebung auftreten, ohne die Bedeutung zu verändern. Ein Beispiel dafür im Deutschen ist das oben genannte vokalische r, das am Wortende oder vor einem Konsonanten in verschiedenen Varianten auftritt, hier beispielsweise in den Wörtern Turm, Tour und Wurm beobachtet werden. Aufgrund der Anglizismen musste für die korrekte Phonemzählung das vokalische r bei einem Diphthong œ als ein Phonem berücksichtigt werden. Die Diskussion über die angemessensten Transkriptionsmethoden dieser Laute nach den IPA-Richtlinien hält weiterhin an [28].

## 2.6 Text-to-Speech Synthese und Anwendung

TTS-Systeme fungieren als Sprachsynthesysteme, die aus Text ein hörbares Sprachsignal erzeugen [29]. Die menschliche Sprachverarbeitung und die Generierung von Sprache durch Maschinen weisen grundlegende Unterschiede auf. Menschen interpretieren Sprache kontinuierlich und setzen sie in Bezug zu ihrer Umgebung. Das ist ein Prozess, der als Verständnis bezeichnet werden kann [29]. Im Gegensatz dazu benötigt eine Sprachsynthese kein solches Verständnisniveau. Für die natürliche Sprachsynthese ist die Kenntnis über die Identität und Sequenz der Wörter ausreichend [30].

Die Funktionsweise eines TTS-Systems gliedert sich in zwei Hauptebenen: Die segmentale Ebene, die Ausspracheregeln innerhalb eines Wortes oder Segments umfasst. Die suprasegmentale Ebene, die sich auf Satzakzente, Gruppierungen, Pausen und Variationen in der Tonhöhe bezieht.

Ein TTS-System führt in erster Linie eine syntaktische Analyse des Textes durch, ohne semantische Inhalte zu berücksichtigen, und kann so für viele Sätze korrekte Sprachsignale erzeugen [29].

Die Sprachsynthese umfasst eine Transkriptionsstufe für die linguistische Verarbeitung und eine phonoakustische Stufe für Prosodiesteuerung und Signalproduktion. Während der Transkriptionsstufe werden die Laute für die Sprachausgabe festgelegt. In der phonoakustischen Stufe wird die Prosodie der Laute bestimmt, indem Grundfrequenz, Intensität und Dauer für jeden Laut festgelegt werden [29]. Es gibt verschiedene Transkriptionsansätze, die hauptsächlich in direkte und linguistische Ansätze unterteilt sind. Direkte Ansätze konzentrieren sich ausschließlich auf die Erzeugung korrekter Lautfolgen, während linguistische Ansätze linguistische und logische Zusammenhänge nutzen, um die Anzahl der notwendigen Regeln zu reduzieren [29]. Der Verkettungsansatz, der Sprache aus einer vordefinierten Menge von Grundelementen zusammensetzt, ist der am häufigsten verwendete Ansatz in kommerziellen Sprachsynthesystemen [30]. Bei der Korpussynthese wird besonderer Wert auf die Kontinuität an den Verbindungsstellen der Grundelemente gelegt. Die Prosodiesteuerung umfasst außerlinguistische Funktionen, die Emotionen über Betonung und Intonation vermitteln, und linguistische Funktionen, die sich auf spezifische Wort- und Satzstrukturen beziehen. Moderne Systeme nutzen häufig neuronale Netze zur Steuerung von Lautdauer und Grundfrequenz, da lineare Anpassungen aufgrund der Vielzahl an Einflussfaktoren komplex sind. Die Lautstärkeanpassung individueller Phoneme hingegen wird oft vernachlässigt, um natürlichere Ergebnisse zu erzielen [29].

Für die im Folgenden verwendete Sprachsynthese wurde die Acapela Cloud genutzt. Sie setzt auf fortschrittliche Technologien, um natürliche und flüssige Sprachausgaben zu erzeugen. Während die oben beschriebene konkatenative Synthese auf der Verkettung vorab aufgenommener Sprachsegmente basiert. Es werden bei Acapela moderne Ansätze wie die parametrische Synthese sowie Techniken eingesetzt, die auf künstlicher Intelligenz, Deep Neural Network (DNN) und maschinellem Lernen basieren. Deep Learning, ein Teilbereich des maschinellen Lernens, nutzt neuronale Netzwerke mit vielen Schichten, um große Datenmengen effizient zu verarbeiten. In TTS-Systemen hat Deep Learning die Entwicklung revolutioniert, indem es ermöglicht, direkt von Text zu Sprache zu gelangen, ohne die Zwischenschritte traditioneller Methoden. Dies führt zu einer deutlich verbesserten Natürlichkeit und Flüssigkeit der Sprache [31]. Diese Methoden erlauben es, die Sprachsynthese flexibler und dynamischer zu gestalten, indem sie Sprache direkt aus Text unter Berücksichtigung von lin-

guistischen und akustischen Modellen generieren. Durch den Einsatz von künstlicher Intelligenz und maschinellem Lernen ist das System in der Lage, kontinuierliche Verbesserungen in der Natürlichkeit und Verständlichkeit der Sprachausgabe zu erzielen. Es lernt aus großen Mengen an online bereitgestellten Sprachdaten, erkennt Muster und passt die Synthese entsprechend an, um eine hohe Qualität und Natürlichkeit der erzeugten Sprache zu gewährleisten [32]. Diese Technologien ermöglichen es neue Stimmen und Sprachen effizienter zu entwickeln. Durch die Integration einer API bietet Acapela basierend auf der Technik von DNN eine einfache Einbindung die Sprachsynthese in verschiedene Anwendungen und Plattformen, einschließlich der direkten Ansteuerung durch Skriptsprachen wie Python und MATLAB. Entwickler und Forscher können so die Funktionen von Acapela nutzen, um individuell angepasste Sprachausgaben zu erstellen, die spezifischen Anforderungen gerecht werden. Die Acapela Cloud ermöglicht die Sprachsynthese durch eine Vielzahl von Stimmen, Einstellmöglichkeiten und Sprachen, darunter die Stimme des Acapela Sprechers Klaus für den deutschsprachigen Raum. Neben anderen Sprechern bietet Klaus nach eigener Analyse eine große Übereinstimmung in der gemittelten Grundfrequenz von ca. 113 Hz und der Aussprache mit dem ehemaligen Sprecher des FET Claus Wunderlich mit ca. 123 Hz. Zusätzlich ergaben Untersuchungen von Saskia Ibelings, dass das TTS-System von Acapela mit dem synthetischen Sprecher Klaus 22K-HQ (basierend auf DNN) der Studie mit dem Göttinger Satztest gute und teils bessere Ergebnisse bezüglich des Speech Reception Threshold (Sprachverständlichkeitsschwelle) (SRT) und der Steigung erzielte, wobei die SRT-Werte etwa 1,2 dB besser waren als die des Originalmaterials [33].

Ein spezifisches Feature der Acapela Cloud ist die Möglichkeit, zum Beispiel durch den Befehl `sel=altN`, alternative Aussprachen für nachfolgende Wörter zu wählen, was eine präzisere Steuerung der Sprachausgabe mit verschiedenen Atemgruppen ermöglicht [34]. Insgesamt geht Acapela somit über die Grenzen der konkatenativen Synthese hinaus und bietet eine flexible, leistungsstarke DNN-Lösung für die Erzeugung von synthetischer Sprache.

### 3 Methodik

#### 3.1 Übersicht zum Vorgehen der Testlistenenerstellung

Der Prozess zur Erstellung von Testlisten beginnt mit der Auswahl von sechs Sprachkorpora, aus denen automatisiert einsilbige Substantive basierend auf ihrer Wortfrequenz extrahiert werden. Dabei liegt der Fokus auf den im deutschen Sprachgebrauch am häufigsten verwendeten Einsilbern. Diese werden in eine Datenstruktur überführt, die das IPA und das SAMPA integriert, um eine Basis für die folgenden Analyse- und Syntheseschritte zu schaffen. Basierend auf der neu zusammengestellten Einsilberkollektion werden zufällige Testlisten zusammengestellt und anschließend durch TTS-Verfahren synthetisiert. Darauf folgt eine detaillierte phonemische Analyse der ausgewählten einsilbigen Substantive, die mit Statistiken der deutschen Literatur, verglichen wird. Die Evaluation dieser synthetischen Testlisten umfasst sowohl die Messung des Sprachverstehens mit Probanden als auch vorab eine interne Bewertung der Natürlichkeit der Synthese. Ein Schwerpunkt der Auswertung liegt auf der phonemischen und perzeptiven Äquivalenz. Die Ergebnisse dienen zunächst der iterativen perzeptiven Optimierung der Testlisten, sodass die Aktualisierungen gewährleisten, dass die synthetischen Testlisten in zukünftigen audiologischen Messungen dem aktuellen deutschen Wortschatz entsprechen und untereinander vergleichbarer sind als bisher.

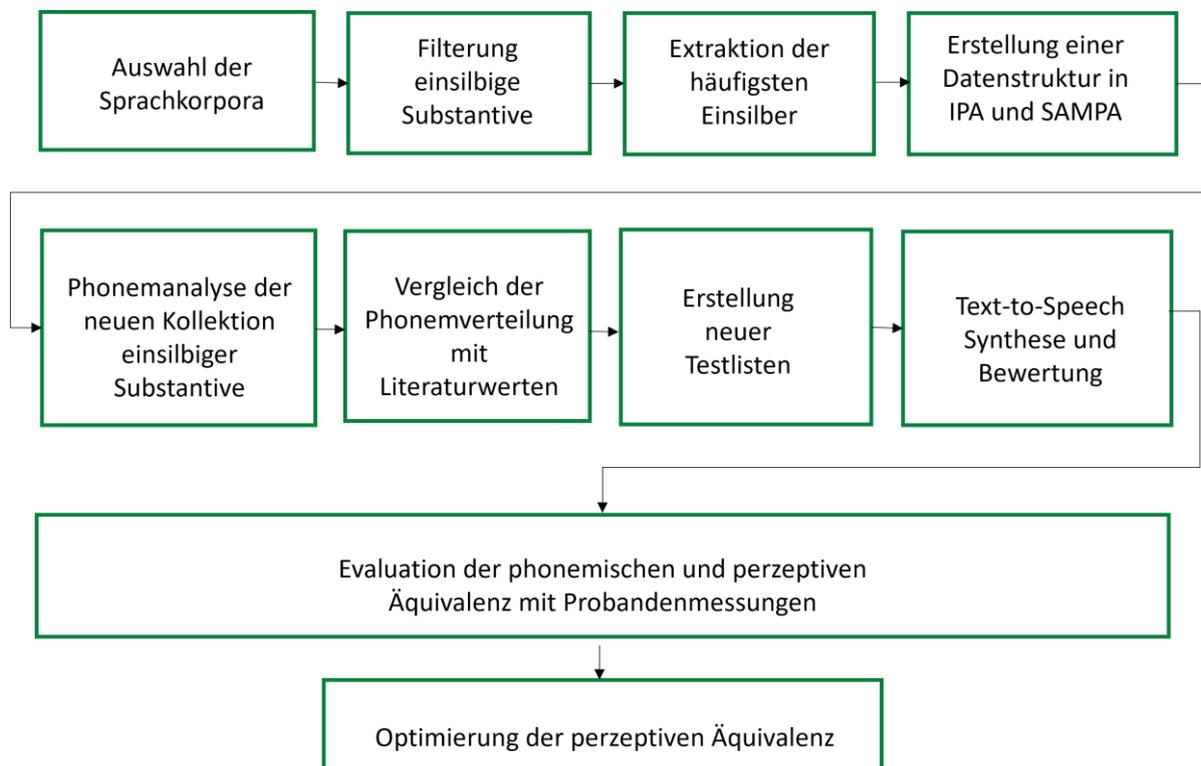


Abbildung 6: Prozess der Testlistenenerstellung und Evaluation.

### 3.2 Sichtung und Auswahl der Einsilber

Die Auswahl und spätere Transkription und Analyse der einsilbigen Substantive aus allen Sprachkorpora erfolgte automatisiert mit einem individuellem Skript in MATLAB R2022b. Zunächst wurden aus sechs verschiedenen Sprachkorpora alle Wörter ausgewählt, die mit einem Großbuchstaben beginnen und ausschließlich aus lateinischen Buchstaben bestehen. Wörter mit Leerzeichen, inneren Majuskeln (Großbuchstaben innerhalb des Wortes) und Sonderzeichen wurden entfernt. Nach diesem ersten Schritt erfolgt eine Filterung nach einsilbigen Substantiven, durch Prüfung auf einen Monophthong (einzelner Vokal, z.B. a, e, i, o, u), einen Diphthong (zwei unterschiedliche Vokale, z.B. eu, au, ei) oder zwei aufeinander folgende Monophthonge (z.B. aa, ee, ii, oo, uu) enthalten. Einsilber bestehen immer nur aus einem Vokal, Monophthong oder Diphthong und einem Konsonanten oder Affrikat. Die extrahierten Wörter wurden daraufhin abgeglichen, um Duplikate zu vermeiden und sicherzustellen, dass jedes Wort nur einmal vorkommt. Diese beiden Schritte sind wichtig, um die einsilbige Struktur und die automatische Substantivselektion zu gewährleisten, die Mehrsilber zu entfernen und damit auch die Analysezeit der API gesteuerten Wortfrequenzkategorisierung im dritten Schritt zu reduzieren. Die Wikipedia Sammlung wurde ergänzend genutzt, um nach der Einsilberextraktion aus fünf Korpora zu prüfen, ob diese vollständig ist oder noch deutsche Einsilber fehlen.

Bei der API-basierten Wortfrequenzfilterung wird die Worthäufigkeit für jedes einsilbige Substantiv online aus dem DWDS geladen. Nach einer ersten Sichtung der Häufigkeiten wurde ein Grenzwert von 3 bis 6 festgelegt, um Wörter als häufig zu kategorisieren, dargestellt in Abbildung 5. Diese Wörter werden dann für die Erstellung der neuen Testlisten verwendet. In Korpora mit gesprochener Sprache wurden teilweise ungewollt Füllwörter wie Ähm transkribiert. Diese Wörter erfüllen manchmal die Kriterien für die erste automatische Extraktion von einsilbigen Substantiven, werden aber anschließend vom API-Wortfrequenzfilter automatisch aussortiert. Das Gleiche gilt für Schimpfwörter, Regionalismen, Fachsprache und Fantasiewörter. Zahlen wurden ebenfalls automatisch aufgrund ihrer niedrigen Wortfrequenz gefiltert. Dies ist nützlich, da der verwandte FMT, der Freiburger Mehrsilber-Sprachtest, nur aus Zahlen besteht, was zu Irritationen führen könnte. Heutzutage sind auch Anglizismen in der deutschen Sprache häufig und wurden aufgrund ihres Vorkommens in den verschiedenen Korpora zusammen mit ihrer hohen Wortfrequenz automatisch selektiert. Wörter wie z.B. die Anglizismen fügen jedoch neue Phoneme zu den deutschen Sprachkorpora hinzu, z.B. das Phonem  $\text{ʊə}$  wie in Tour.

Im Allgemeinen ist es ratsam wegen der Phonemtrennung, Einsilber mit fremden Phoneme möglichst wenig zu verwenden, aber die Entscheidung, wie viele davon zunächst in die endgültige Auswahl aufgenommen wurden, ist ebenfalls automatisch über die Wortfrequenz erfolgt. Es wurden ausschließlich die häufigsten Wörter ausgewählt, definiert durch einen Skalenbereich von Wert drei (ab 111 540 Tokens) bis sechs (ab 11 540 458 bis zur maximalen Tokenanzahl von 35 271 873 584), siehe Abbildung 5. Diese Auswahlstrategie dient dazu, im Deutschen veraltete Wörter, Fremdwörter, seltene alte Rechtschreibvarianten sowie Abkürzungen und die meisten fehlerhaft transkribierten Wörter zu vermeiden. Zwar erscheint es fast unmöglich, das Wortmaterial eines Sprachtests so zu optimieren, dass es für alle Regionen, sozialen Klassen, Bildungsniveaus und Altersgruppen gleichermaßen geeignet ist, doch diese Problematik wird durch die Berücksichtigung der Wortfrequenzverteilung deutlich reduziert.

Nach Überprüfung aller vorab automatisiert gefilterten Wörter wurde beschlossen, konsequent die neue deutsche Rechtschreibung anzuwenden, um ungewollte Dopplungen zu vermeiden und den aktuellen phonetischen Standards zu entsprechen [35], [36]. Dazu gehören alle deutschen Affrikate ts, tʃ, pf sowie das englische Affrikat dʒ, beispielsweise verwendet in Jeans, und das vokalische Diphthong *ʊə*, wie es in Tour vorkommt. Diese Maßnahme stellt sicher, dass die Phonemzählung auch bei Anglizismen korrekt durchgeführt wird. Final mussten alte Deutsche Rechtschreibung und Eigennamen manuell entfernt werden.

### 3.3 Listenerstellung unter Berücksichtigung der Phonemstruktur

Die Erstellung der Listen mit einsilbigen Substantiven, basierend auf der Anzahl der Phoneme pro Wort, folgte der von Hahlbrock definierten Struktur der FET-Listen, wie bereits in den Grundlagen in der Abbildung 4 dargestellt. In jeder Liste des FETs sind im Mittel 73 Phoneme enthalten, die sich auf die einsilbigen Substantive verteilen. Es gibt pro Liste jeweils ein Wort mit zwei Phonemen, sieben Wörter mit drei Phonemen, zehn Wörter mit vier Phonemen und zwei Wörter mit fünf Phonemen [9]. Um die Phonemanzahl der Einsilber zu bestimmen und eine spätere Analyse sowie Optimierung der Testlisten in neun Phonemklassen sicherzustellen, war eine Transkription der gesammelten Einsilber von lateinischen Buchstaben in einzelne Phoneme gemäß dem IPA erforderlich. Für die Transkription wurde das Online-Tool Grapheme to Phoneme (G2P) des Bayerischen Archivs für Sprachsignale genutzt [37]. Eine vollständige Übersicht aller offiziellen IPA-Zeichen und weiterer Symbole mit dem auch G2P arbeitet, befindet sich im Anhang in Abbildung 34.

Für den späteren Vergleich des original FETs und aktualisierten Einsilbertests wurde auf die Phonemstatistik aus K. J. Kohlers Buch Einführung in die Phonetik des Deutschen, Grundlagen der Germanistik zurückgegriffen. Diese basiert auf dem Kiel Corpus of Read Speech mit einer Größe von 23 985 Wörtern gelesener deutscher Sprache. Besonders beachtet wurde von Kohler die Unterscheidung zwischen allgemeinen und akzentuierten Vokalen, da akzentuierte Vokale weniger häufiger in der gesprochenen Sprache vorkommen [17]. Diese differenzierte Betrachtung der Vokalhäufigkeiten wurde mit einer entsprechenden Gewichtung berücksichtigt.

Die gewichtete Häufigkeit von Vokalen, einschließlich akzentuierter Vokale, wird durch die folgende Formel gegeben:

$$\frac{(\text{Häufigkeit Vokal} \cdot \text{Gesamtanzahl Vokale}) + (\text{Häufigkeit akz. Vokals} \cdot \text{Gesamtanzahl akz. Vokale})}{\text{Gesamtanzahl Vokale} + \text{Gesamtanzahl akz. Vokale}} \quad (4)$$

In der Literaturstatistik von Kohler wurden 40 Phoneme in IPA unterschieden und gemäß DIN EN ISO 8253-3:2022-11 [12] in sieben Phonemklassen bzw. mit Berücksichtigung der sonstigen Vokale und Konsonanten, in neun Klassen unterteilt. Die genaue Zuordnung und prozentuale Verteilung der einzelnen Phonemklassen ist nachfolgend in Tabelle 1 aufgeführt.

Tabelle 1: Phonemverteilung in der deutschen Sprache nach K. J. Kohler.

Phonemklasse	IPA-Zeichen	Phonemverteilung Kohler
<b>Vokale</b>		
Lange Vokale	a:, e, i, o, ø, u, ε:, y	25,25 %
Kurze Vokale	a, ɑ, ɐ, ε, ɪ, ɔ, ʊ, ʏ, œ, ə	64,33 %
Sonstige Vokale	au, ai, ɔɪ	10,42 %
Summe		100%
<b>Konsonanten</b>		
Stimmhafte Plosive	d, g, b	10,84 %
Stimmlose Plosive	k, t, p	20,99 %
Stimmhafte Frikative	v, z, ʒ	4,3 %
Stimmlose Frikative	s, f, ʃ, h, x, ç	20,21 %
Nasale Laute	m, n, ŋ	18,51 %
Sonstige Konsonanten	l, j, r	25,15 %
Summe		100%

Die Einteilung der Phoneme in 18 Phonemklassen nach Hahlbrock wurde nicht umgesetzt. Es wurde die Empfehlung der Norm DIN EN ISO 8253-3:2022-11 bevorzugt, da es die Komplexität verringert und Exter et al. darauf hingewiesen haben, dass die damalige detaillierte Aufteilung von Hahlbrock aus heutiger Sicht oftmals als unbegründet angesehen wird [16].

Nicht alle der 40 Phoneme, die für die Analyse verwendet wurden, lassen sich den sieben Hauptklassen gemäß der Norm DIN EN ISO 8253-3:2022 [12] zuordnen. Dies führt zur Bildung zusätzlicher Klassen. Einerseits die Klasse der sonstigen Vokale bzw. Diphthonge (au, ai, ɔɪ), andererseits die Klasse der sonstigen Konsonanten (l, j, r). Der Glottisverschluss, ein stimmloser Laut, wurde in der Analyse nicht berücksichtigt, da er hauptsächlich die Artikulation betrifft und nicht direkt mit der phonemischen Struktur der Sprache in Verbindung steht. Er wird daher nicht als eigenständiges Phonem betrachtet [38]. Die abschließende Phonemanalyse der endgültigen Wortauswahl von 540 Wörtern erfolgte in IPA und wurde in neun Phonemklassen zusammengefasst, um eine Vergleichbarkeit mit der gewählten Literatur zu gewährleisten.

Im Prozess der Phonemumwandlung wurde dem IPA gegenüber SAMPA der Vorzug gegeben, insbesondere wegen der genaueren Übereinstimmung mit den phonemischen Statistiken von Kohler, der ebenfalls IPA für seine Analysen nutzte [17]. Nach der Transkription der Wörter in ihre Phoneme wurden die Einsilber gemäß ihrer Phonemlänge über ein individuelles MATLAB Programm kategorisiert, wie bereits in Abbildung 4 von Hahlbrock vorgegeben. Es gibt pro Liste jeweils ein Wort mit zwei Phonemen, sieben Wörter mit drei Phonemen, zehn Wörter mit vier Phonemen und zwei Wörter mit fünf Phonemen. Bei 278 Einsilbern, die aus vier Phonemen bestehen, erwies sich dies als limitierender Faktor für die Listenzusammenstellung, da jede Liste zehn Wörter mit vier Phonemen enthalten soll. Unter diesen Vorgaben war die Erstellung von mehr als 27 Listen aus dem gesammelten Material nicht möglich. Vier Wörter mit einer Länge von sechs Phonemen und ein Wort mit nur einem Phonem (Ei) wurden aufgrund ihrer fehlenden Übereinstimmung mit der Listenstruktur von Hahlbrock ausgeschlossen. Eine Überprüfung der 600 transkribierten Einsilber in das IPA erfolgte anhand von 100

Stichproben. Unter Beachtung der erläuterten Listenstruktur nach Hahlbrock wurden die Testlisten zufällig und in maximal möglicher Anzahl zusammengestellt.

### 3.4 Synthetisierung der Einsilberauswahl

Als Ersatz für die Stimme des inzwischen verstorbenen Sprechers Claus Wunderlich wurde die bereits existierende, männliche synthetische Stimme Klaus der Acapela Group genutzt, die eine Ähnlichkeit mit der Stimme von Claus Wunderlich aufweist. Der Versuch, in der Bachelorarbeit von Thomas Schwarz eine Stimme auf Basis des alten Tonmaterials über Acapela zu erstellen, scheiterte aufgrund unnatürlicher Aussprache und für den aktuellen Stand der Technik zu wenig Trainingsmaterial [10]. Die Stimme Klaus wurde mittels MATLAB und einer Python API gesteuert, die Zugriff auf die Acapela Group Cloud bietet. Vor der Erstellung der Audiodateien konnten Parameter wie Sprechgeschwindigkeit, spektrale Änderungen und Lautstärke über das API angepasst werden. Es wurden alle Werte bis auf die Sprechgeschwindigkeit im default gelassen. Die 100 Prozent Sprechgeschwindigkeit als default Einstellung von Acapela war subjektiv zu schnell und führte zu verfremdeten und unverständlichen Aussprache bei fast allen Einsilbern. Standardmäßig wurde deshalb für die TTS-Anwendung eine Geschwindigkeit von 85 Prozent in der Acapela Synthese eingestellt. Das Ergebnis der Bachelorarbeit von J. Karl ergab, dass eine schnelle Sprechgeschwindigkeit zu signifikant niedrigeren Sprachverständlichkeiten führt, während eine langsame Sprechgeschwindigkeit keinen Unterschied ergibt [39].

Die Audiodateien wurden mit einer Abtastrate von 44 100 Hz als WAV-Dateien gespeichert, wobei für jedes Wort eine separate Datei angelegt wurde. Die synthetischen Einsilber von Acapela enthielten Pausen vor und nach den Wörtern, die durch automatisierte Erkennung von Nullen am Anfang und Ende des Signals entfernt werden mussten. Anschließend wurden alle bereinigten 540 WAV-Dateien sequenziell angeordnet, um den RMS-Schallpegel zu ermitteln und an den des FETs in der Siemens Version anzugleichen [40]. Für den RMS-Pegel des original FETs, der als Referenzwert dient, beträgt der Wert -23,32 dBFS. Der C-bewertete Schalldruckpegel vom CCITT-Rauschen (Comité Consultatif Internationale Télégraphique et Téléphonique) beträgt -16,97 dBFS und um die gleiche Pegeldifferenz wie die des Freiburgers zu bekommen, muss das synthetische Sprachmaterial auf -23,32 dBFS gebracht und C-bewertet werden [41]. Bei der C-Bewertung handelt es sich um einen Frequenzbewertungsfilter, der dazu dient, die menschliche Schallwahrnehmung abhängig von der Frequenz genauer zu simulieren. Dieses Vorgehen ist wichtig, da beim Pegelabgleich des FETs in der verwendeten Siemens-Variante ebenfalls keine Pausen enthalten sind. Zusammengefasst gewährleistet dies, dass die Gesamtdifferenz der synthetischen Signale im Vergleich zum originalen FET vollständig berücksichtigt wird, was für den Vergleich mit dem bisherigen Sprachmaterial entscheidend ist.

### 3.5 Subjektive Bewertung der synthetischen Einsilber

Bei der Überprüfung des synthetischen Testmaterials wurde teilweise eine unnatürliche oder falsche Aussprache festgestellt. Der Versuch, diese Unstimmigkeiten mithilfe der API-Funktion zur Phonemeingabe mit SAMPA-Zeichen zu korrigieren, führte nur selten zu einer Lösung. Die Eingabe spezifischer Phonemsequenzen erwies sich als nicht ausreichend, um die gewünschten Korrekturen vorzunehmen. Daher wurde eine alternative Synthesemethode (Alternative 1) durch Acapela direkt mit lateinischen Buchstaben gewählt. Diese Methode bot die größte Automatisierung für eine korrekte Artikulation und Aussprache der Einsilber und stellte sich als Ansatz zur Generierung der korrekten Atemgruppe heraus. Eine Atemgruppe ist ein Segment gesprochener Sprache, das in einem Atemzug geäußert wird. Sie bildet eine natürliche Einheit der Sprechdynamik und ist entscheidend für den Rhythmus und die Betonung in der mündlichen Kommunikation. Die Aussprache einer Atemgruppe können je nach Sprechtempo und Intention des Sprechers variieren [34]. Für die Durchführung der Probandenstudie mussten nach der automatisierten TTS-Synthese über die Atemgruppe 1 trotzdem insgesamt 95 Wörter wegen eindeutiger Aussprachefehler (Artefakte, Unverständlichkeit, falsche Betonung und Sprechgeschwindigkeit) und 26 Wörter wegen schlechter Bewertung bezüglich der Natürlichkeit neu synthetisiert werden.

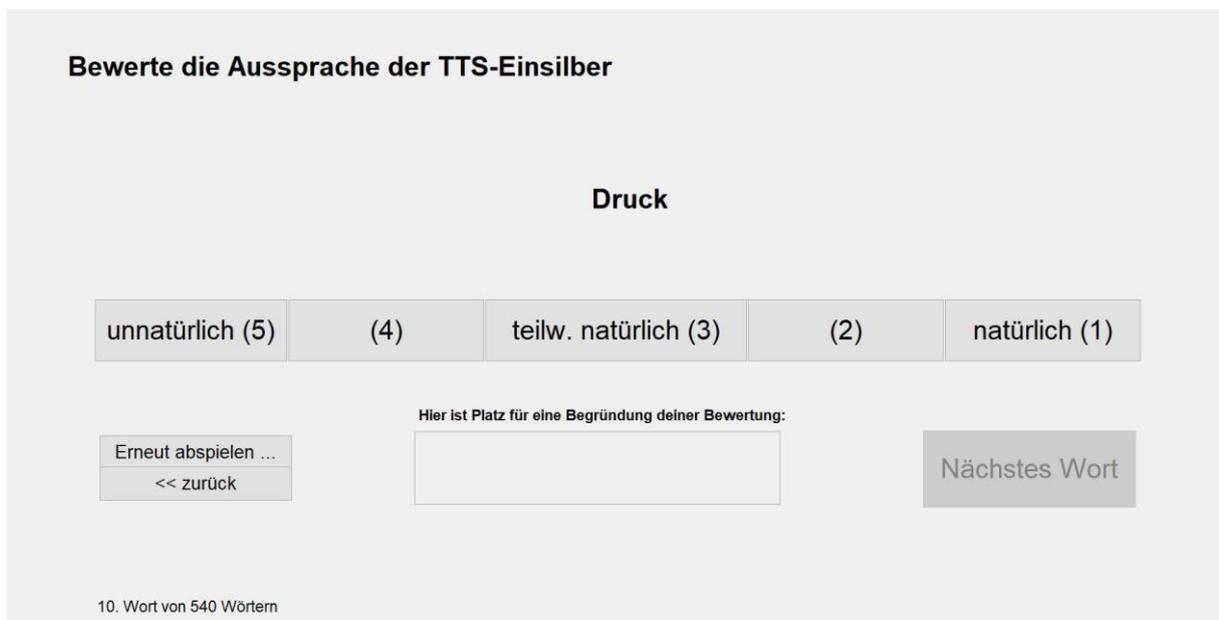


Abbildung 7: MATLAB GUI zur Bewertung der aktualisierten Wortsammlung für die 27 neuen Listen mit 540 Wörtern.

Sieben Mitarbeiter des Deutschen Hörgeräte Instituts bewerteten anschließend die Qualität jedes Wortes anhand eines Schulnoten Systems von eins bis fünf, wobei besonders auf die Korrektheit und Natürlichkeit der Aussprache, einschließlich der Prosodie, geachtet wurde [42]. Die Einweisung lautete Bewerte die Natürlichkeit der Text-To-Speech (TTS)-Einsilber in Schulnoten (1 - 5), berücksichtige dabei Sprechgeschwindigkeit, Deutlichkeit, Intonation, Lautübergänge, Dauer der Laute, Akzente und

Artefakte. Anschließend wurden die Wörter, die eine Bewertung schlechter oder gleich der Note 3 erhielten, erneut über die Alternativsynthese erzeugt. Für die Optimierung der Aussprache wurden die Alternativen 0 bis 5 verwendet, wie von Acapela empfohlen, da diese die häufigsten Varianten darstellen, die jeweils spezifischen Atemgruppen zugeordnet sind [34]. Mithilfe der Alternativen 2 und 3 konnten die meisten Aussprache- und Betonungsfehler behoben werden. Dies führte dazu, dass in der Bewertung der Aussprachequalität keine Ergebnisse schlechter als eine Bewertung von 3 verbleiben sollten. Diese Maßnahmen stellen sicher, dass die synthetisierten Sprachdaten eine hohe Qualität in der Artikulation und Betonung aufweisen, was für die Genauigkeit des Sprachtests entscheidend ist.

### 3.6 Evaluation der neuen Testlisten

#### 3.6.1 Messaufbau

Die Durchführung der Probandenmessungen fand in der schallisolierten Kabine G001 des Deutschen Hörgeräte Instituts statt, deren Grundfläche 11,7 m<sup>2</sup> beträgt und deren Hintergrundschallpegel unter den von DIN EN ISO 8253-2 für tonaudiometrische Messungen im Freifeld vorgeschriebenen Grenzwerten liegt [43]. Der Raum ist mit Absorbermaterialien und Teppich ausgestattet, um eine optimale Schallabsorption zu gewährleisten, siehe nachfolgend in Abbildung 8.

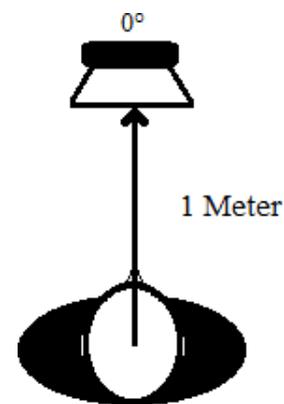


Abbildung 8: Messaufbau bestehend aus einem Lautsprecher in 1 Meter Entfernung bei 0 Grad.

Der Lautsprecher wurde im Messraum frontal in einer Entfernung von 1 Meter zur Mitte des Probandenkopfes positioniert, mit der Höhe des Schallaustritts auf der Medianebene des Probanden. Um unerwünschte Stehwellen im Raum zu minimieren, wurde der Versuchsaufbau innerhalb des Raumes schräg ausgerichtet. Der Monitor des Untersuchers war durch seine spezielle Anordnung und einen Blickschutzfilter für den Probanden nicht sichtbar. Als Soundkarte wurde die RME fireface 802 verwendet und die Steuerung der Messungen erfolgte mit MATLAB R2023b. In Tabelle 2 sind alle weiteren Geräte aufgelistet, die für den Messaufbau genutzt wurden.

Tabelle 2: Auflistung der verwendeten Geräte.

<b>Gerät</b>	<b>Zweck</b>
Soundkarte: RME fireface 802	Audiowandlung und Mix
Vorverstärker Brüel und Kjær Type 2669	Verstärkung der Signale
Freifeldmikrofonkapsel Brüel und Kjær Type 4190	Aufnahme der Signale
Power Supply	Phantomspannung des Mikrofons mit 48V
Pistonphon Type 4228	Kalibrierung des Mikrofons
Freifeldlautsprecher: Genelec 8351A	Abgabe der Signale für den FET
ACAM5 mit freifeldentzerrtem Kopfhörer DT48	Signalabgabe Tonaudiometrie

Die Kalibrierung des Referenzmikrofons erfolgte vor Beginn der Studie, um die Genauigkeit der Messungen sicherzustellen. Dazu wurde ein Pistonphon verwendet, das einen konstanten Sinuston von 250 Hz erzeugt. Nach der Kalibrierung wurde die Entzerrung des Frequenzgangs des Messsystems durch die Wiedergabe eines Exponentialweeps durchgeführt. Dieser Sweep wurde gleichzeitig von einem Referenzkondensatormikrofon aufgenommen, das in einer Entfernung von einem Meter und ausgerichtet auf den Lautsprecher an der Stelle des Stuhls bei 0 Grad positioniert war. Diese Methode ermöglichte eine präzise Anpassung des Systems an die akustischen Eigenschaften des Raumes. Das Ziel der Entzerrung war es, einen gleichmäßigen Frequenzgang ab 200 Hz mit einer Toleranz von  $\pm 2$  dB zu erreichen, um eine genaue Wiedergabe der Signale während der Messungen zu gewährleisten. Die Freifeldentzerrung des Lautsprechers wurde vor jedem Probanden kalibriert und überprüft.

### 3.6.2 Probandenkollektiv

An der Studie nahmen 27 Probanden<sup>2</sup> im Alter zwischen 18 und 25 Jahren teil, mit einem Durchschnittsalter von 22 Jahren. Unter den Teilnehmern waren 25 Frauen und 2 Männer. Allen Teilnehmenden war der FET unbekannt. Für die Teilnahme war Deutsch als Muttersprache Voraussetzung. Die Messungen mit den Probanden wurden innerhalb eines durchschnittlichen Zeitrahmens von 1,5 Stunden im November und Dezember 2023 durchgeführt. Laut den Ergebnissen der Tonschwellenaudiometrie, dargestellt in Abbildung 9, erfüllten alle Probanden die Vorgabe einer Normalhörigkeit. Als Definition der Normalhörigkeit wurden die in der DIN EN 8253-3:2012-08 formulierten Empfehlungen herangezogen. Die Probanden sollten eine Hörschwelle von 10 dB oder weniger bei möglichst allen der folgenden Frequenzen aufweisen: 0,125 kHz, 0,25 kHz, 0,5 kHz, 0,75 kHz, 1 kHz, 1,5 kHz, 2 kHz, 3 kHz, 4 kHz, 6 kHz und 8 kHz. Für bis zu zwei dieser Frequenzen durfte die Hörschwelle für das zu messende Ohr maximal 15 dB betragen [12].

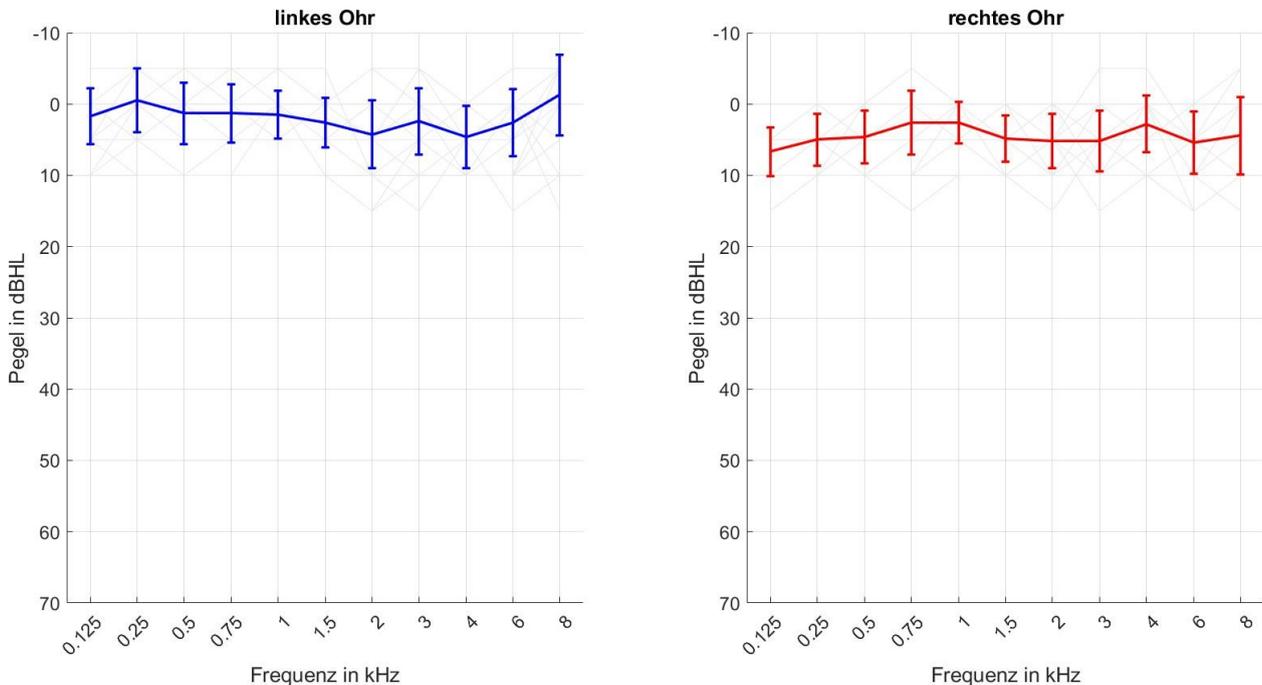


Abbildung 9: Durchschnittliches Reintonhörvermögen der 27 Studienteilnehmer mit Fehlerbalken, welche die Standardabweichung pro Messfrequenz angeben.

<sup>2</sup>Im Folgenden wird aus Gründen der besseren Lesbarkeit und Vereinfachung nur die männliche Form verwendet. Dies impliziert keine Benachteiligung des weiblichen Geschlechts, sondern soll im Sinne der sprachlichen Vereinfachung als geschlechtsneutral verstanden werden.

### 3.6.3 Studiendesign

In der Studie wurden zu Beginn die Teilnehmenden bezüglich ihres Alters, ihrer Muttersprache, früherer Ohrerkrankungen, des bevorzugten Ohrs, vorhandenem Tinnitus und akuten Erkältungen oder Allergien in einer Anamnese befragt. Anschließend erfolgte eine Otoskopie und eine Tonschwellenaudiometrie über Kopfhörer, basierend auf den Normalhörigkeitskriterien der DIN EN 8253-3:2012-08, mit Schwellen unter oder bei maximal 15 dB für die in der Norm hinterlegten Frequenzen [43]. Der aktualisierte synthetische Einsilbertests wurde binaural im Freifeld bei den drei Schallpegeln 21,5 dB, 27,5 dB und 33,5 dB dargeboten. Mithilfe einer Pilotstudie wurden die Schallpegel so gewählt, dass ein Sprachverstehen von ungefähr 20 %, 50 %, und 80 % erreicht wird. Die RMS-Pegel der Einsilber wurden gemäß DIN 45626-1 [44] unter Verwendung von Korrekturfaktoren für den C-bewerteten äquivalenten Schalldruckpegel und die Impulszeitbewertung angepasst, um die übliche Verwendung des Freiburger Einsilbertests in klinischen Anwendungen zu reproduzieren. Der Untersucher war im selben Raum, um Antworten direkt aufnehmen zu können. Dieses Studiendesign wurde so gewählt, um eine möglichst große Vergleichbarkeit zu Karl-Heinz Halbrock, Thomas Schwarz und den Normvorgaben aus der Norm DIN EN 8253-3 [12] zu gewährleisten. Dieser Aufbau hatte in der Praxis auch den Vorteil den Probanden direkt im Blick zu behalten, bezüglich der Aufmerksamkeit oder einer Positionsänderungen. Mittels einer MATLAB GUI wurden Parameter wie Schallpegel, Wortmaterialtyp, Listennummer und das Verstehen in Prozentpunkten angezeigt, wie in Abbildung 10 dargestellt.

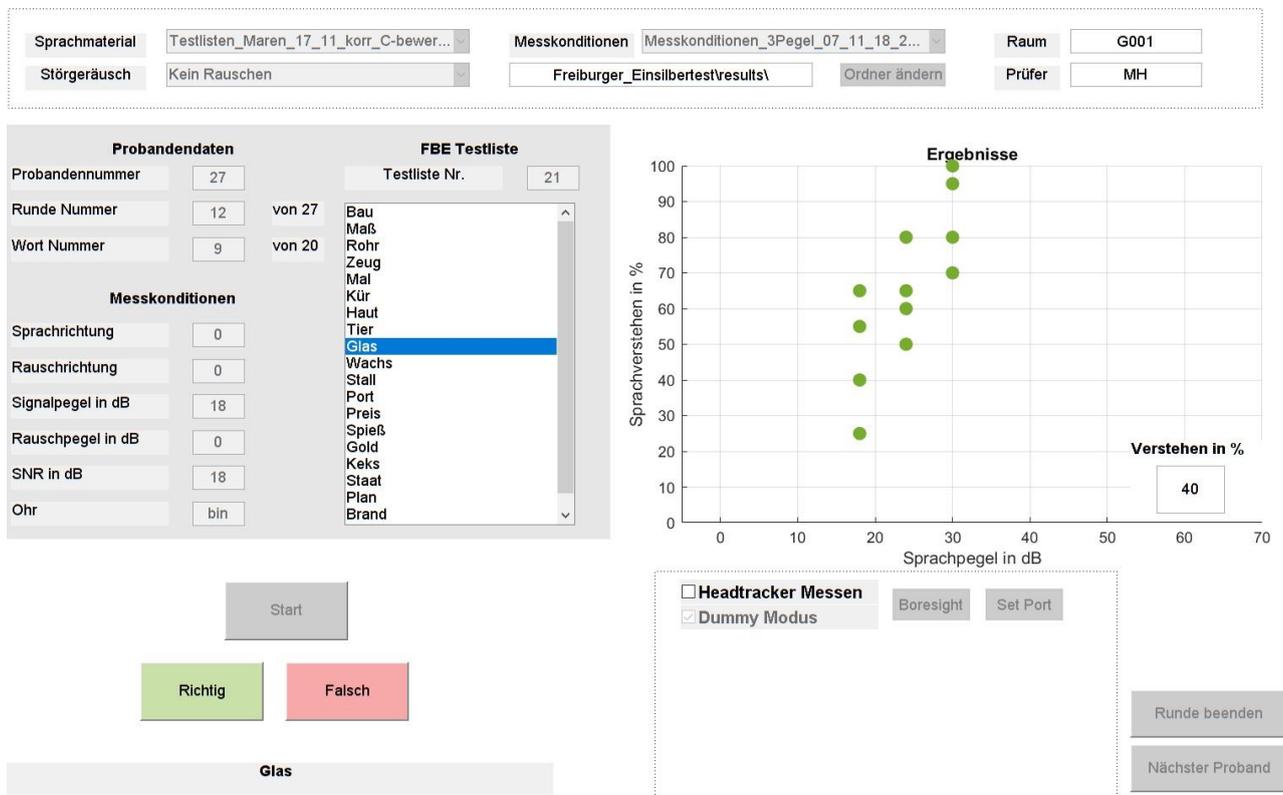


Abbildung 10: MATLAB GUI zur Messung des aktualisierten Freiburger Einsilbertests.

Ein Sprachtest kann als Bernoulli-Prozess modelliert werden, wobei das Sprachverstehen aus einer Testliste als Zufallsvariable betrachtet wird. Für eine Testliste mit 20 Elementen und einem Zielwert des Sprachverstehens von 50 % ist zur Erreichung einer Unsicherheit von 4,5 % mindestens eine Anzahl von 24 Versuchspersonen erforderlich, um die Ergebnisse im 95 %-Vertrauensintervall zu gewährleisten [12]. Da der aktuelle Test 27 Listen umfasst, wurde die Anzahl der Versuchspersonen auf 27 erhöht, wobei die Reihenfolge der Listen und die Schallpegel per Lateinischem Quadrat randomisiert wurden, um Reihenfolgeeffekte und Messdauer zu reduzieren. Jedem der 27 Probanden wurden alle 27 Listen des neu zusammengestellten synthetischen Tests präsentiert. Die neuen synthetischen Wörter wurden pro Probanden bei den genannten Schallpegeln abgespielt, wobei nicht jedes Wort mit jedem Schallpegel getestet wurde, sondern so verteilt, dass am Ende neun Messwerte pro Schallpegel und Wort vorlagen. Die Reihenfolge der Einsilber in einer Liste wurde ebenfalls randomisiert.

### 3.7 Optimierungsziele für den Freiburger Einsilbertest

Die phonemische und perzeptive Äquivalenz in Sprachtests stellt ein komplexes Optimierungsproblem dar, das mindestens zwei konkurrierende Bedingungen umfasst. Wenn ein Optimierungsproblem zwei oder drei solcher Bedingungen enthält, wird dies als mehrdimensionale Optimierung bezeichnet. Ziel bei solchen Problemen ist es, z.B. durch eine differenzierte Gewichtung der konkurrierenden Ziele einen optimalen Kompromiss zu finden.

**Phonemische Äquivalenz basierend auf literarischen Statistiken:** Es wird angestrebt, die Phonemverteilung so abzubilden, wie sie in statistischen Untersuchungen der deutschen Literatur dokumentiert ist, etwa nach den von Kohler beschriebenen Phonemverteilungen im Deutschen. In der Praxis sind diesem Ansatz jedoch durch den begrenzten Wortschatz in Testlisten Grenzen gesetzt.

**Phonemische Äquivalenz durch Gleichverteilung der Phonemklassen:** Um die Begrenzungen der direkten literarischen Anpassung zu überwinden, kann eine gleichmäßigere Phonemverteilung über alle Testlisten hinweg angestrebt werden, indem die Listen untereinander angeglichen werden.

**Perzeptive Äquivalenz:** Das Ziel ist, alle Testlisten an einen gemeinsamen Mittelwert des Sprachverständlichkeitsschwellenwertes (SRT) anzupassen, um eine konsistente Wahrnehmbarkeit und Verständlichkeit sicherzustellen. Dies ist entscheidend für die Vergleichbarkeit und Konsistenz der Testergebnisse. In MATLAB wurde zunächst mit einem iterativen Verfahren versucht nur eine Konditionen (perzeptive Äquivalenz) zu optimieren, um ein Minimum mit einem möglichst niedrigen RMS-Fehler zu finden. Die Ergebnisse dieser Optimierung werden im Abschnitt 4 dargestellt. Eine Gewichtung der Konditionen ist entscheidend, um eine optimale Balance zwischen phonemischer Ausgewogenheit und perzeptiver Äquivalenz zu erreichen. Dies erfordert eine iterative Vorgehensweise, bei der neue Daten zur Sprachverständlichkeit genutzt werden, um die Gewichtungen kontinuierlich anzupassen. So können ausbalancierte Listen für phonemische und perzeptive Äquivalenz für endgültige Studien erstellt werden.

## 4 Ergebnisse

### 4.1 Beschreibung der Einsilbersammlung

Das systematische Extrahieren aller Wörter mit einer Silbe und einem Großbuchstaben als ersten Filterschritt, ergab korpusabhängig unterschiedliche Ergebnismengen, wie in Tabelle 3 aufgelistet. Nach dem Abgleich von Dopplungen wurden insgesamt 7 851 verschiedene potentielle einsilbige Substantive aus den genutzten Sprachkorpora zusammengestellt.

Tabelle 3: Tabellarische Übersicht über die Anzahl an automatisch extrahierten einsilbigen Substantiven aus den verschiedenen Sprachkorpora sowie die Gesamtsumme aller Korpora.

<b>Sprachkorpus</b>	<b>Anzahl</b>
DGD	1 277
LCC	6 711
DeReKo	2 198
DWDS	1 861
CELEX	1 197
Wikipedia	1 172
<b>Summe mit Dopplungen</b>	<b>14 416</b>
<b>Summe ohne Dopplungen</b>	<b>7 851</b>

Die gesammelten 7 851 einsilbigen Substantive wurden nach ihrer Häufigkeit in der neuen Rechtschreibung mit Hilfe der DWDS-API in sieben Wortfrequenzskalen sortiert. Wie in Abbildung 11 dargestellt, kommen einsilbige Substantive nur in den Kategorien null (sehr selten) bis fünf (häufig) vor. Die automatisierte Wortfrequenzfilterung entfernte aus der ursprünglichen Menge von 7 851 einsilbigen Substantiven vorrangig bedeutungslose Zeichenfolgen, Abkürzungen, Regionalismen, Fremdwörter, Fachbegriffe sowie veraltete oder kaum noch genutzte Ausdrücke. Bei der Betrachtung der Verteilung aller einsilbigen Substantive, können die höchsten Anteile in den Kategorien null mit 30 %, eins mit 28 % und zwei mit 29 % erkannt werden. Dies weist darauf hin, dass in der Gesamtheit der sechs Korpora überproportional viele seltene Einsilber vorhanden sind. Wie bereits beschrieben enthält die LCC ohne Lemmatisierung zahlreiche falsch geschriebene Wörter, weshalb diese Sammlung zusätzlich ohne dieses Korpus bewertet wurde. Dies führte zu einem niedrigeren Anteil an Wörtern in den Kategorien null und eins. Im Zuge der Erstellung der neuen Einsilbersammlung wurden selten genutzte Wörter, die im Frequenzbereich von null bis zwei lagen, entfernt. Die anschließende Filterung unter Verwendung der DWDS-Datenbank führte zu 994 häufig genutzten einsilbigen Substantiven aus einer Gesamtmenge von 7 851. Diese reduzierte Kollektion repräsentiert die Wortfrequenzkategorien drei bis fünf und wurde zusätzlich auf alte Rechtschreibung überprüft, um versteckte Dopplungen auszuschließen. Ebenfalls entfernt wurden nicht erkannte Eigennamen. Von den ursprünglich 994 automatisch vorgefilterten Einsilbern wurden 600 beibehalten.

Es war außerdem möglich, wenig repräsentierte Einsilber im FET zu identifizieren, die als Indiz für das Vorkommen veralteter Begriffe dienen. Die Analyse gemäß der neuen deutschen Rechtschreibung

ergab 80 seltene Einsilber, darunter solche aus Kategorie eins wie Grog oder Lump. Somit ist fast ein Viertel des Sprachmaterials als veraltet anzusehen, was durch den gelben Balken des FETs der Kategorien eins und zwei in Abbildung 11 dargestellt wird. Eine vollständige Aufstellung der 80 veralteten Einsilber des FET ist im Anhang, Unterabschnitt 7.3, zu finden.

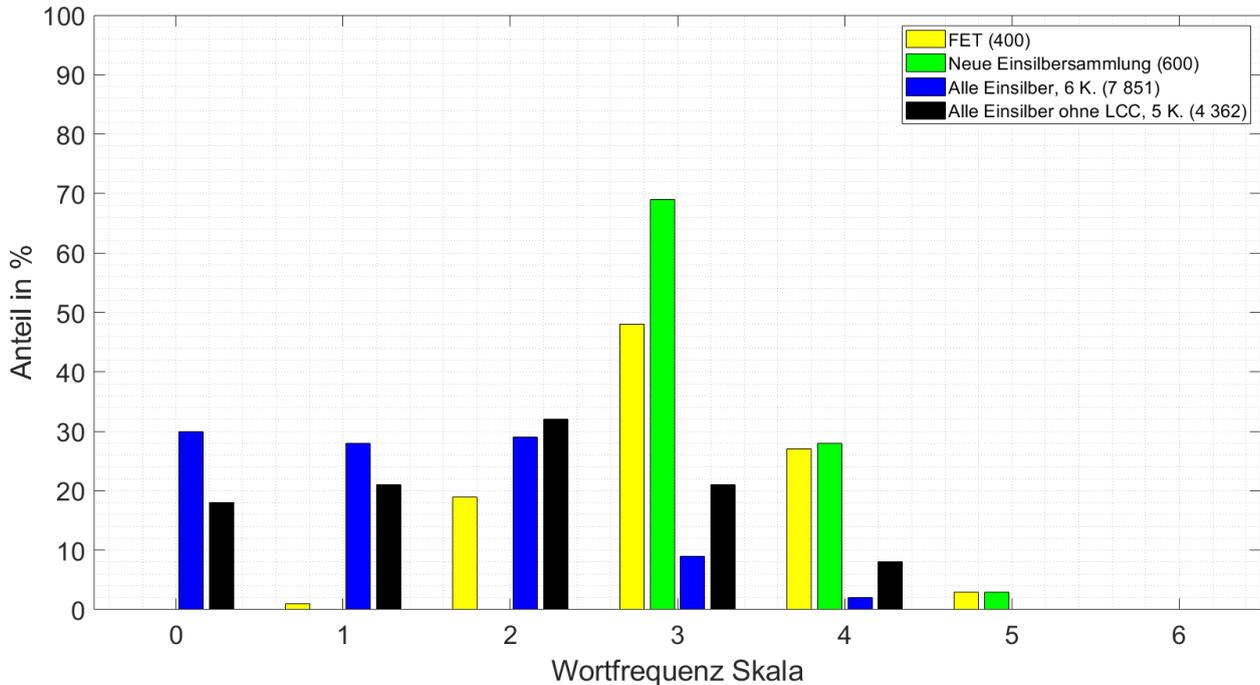


Abbildung 11: Wortfrequenzverteilung im Vergleich mit allen ungefilterten Einsilbern der sechs Korpora (7 851), den ungefilterten Einsilbern ohne LCC (4 362), dem FET (400) und der neuen Einsilbersammlung (600).

## 4.2 Subjektive Bewertung der synthetisch erzeugten Einsilber

Im Rahmen der Untersuchung wurden alle synthetisch erzeugten Wörter von sieben Mitarbeitern des Deutschen Hörgeräte Institut (DHI) mit Hilfe einer MATLAB GUI, siehe Methodik Abbildung 7, hinsichtlich ihrer Natürlichkeit auf einer Schulnotenskala von eins bis fünf bewertet. Abbildung 12 zeigt die durchschnittliche Bewertung jedes Wortes, basierend auf den subjektiven Einschätzungen. Die horizontale Linie markiert die Mitte der Bewertungsskala. Die Analyse der Wortbewertungen zeigt, dass die Mehrheit der Wörter eine Bewertung besser oder gleich drei erhielt. Lediglich 26 Wörter fielen mit einer Note schlechter als drei auf, wie in Abbildung 12 dargestellt.

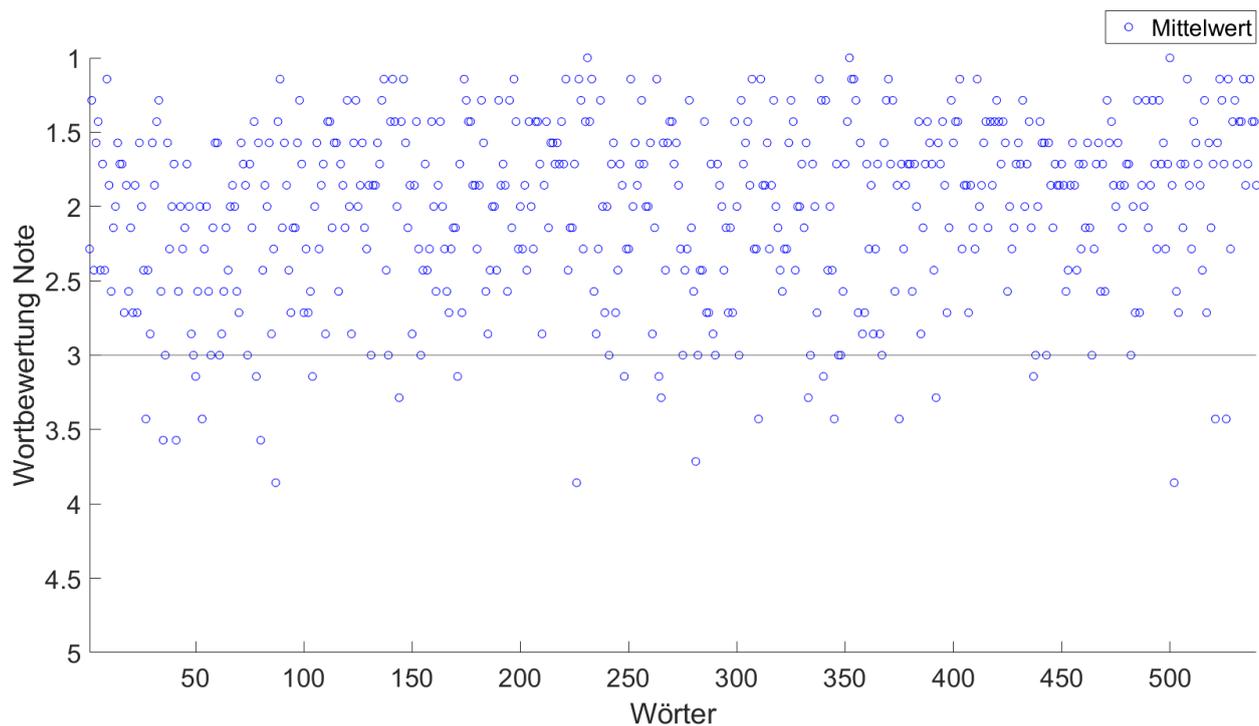


Abbildung 12: Mittelwerte der Bewertung der Synthesequalität von 540 Wörtern anhand der Schulnotenskala von 1 (sehr gut) bis 5 (mangelhaft) durch sieben Probanden.

Die 540 Wörter erzielten eine durchschnittliche Gesamtbewertung mit der Synthese der Atemgruppe Alternative 1 und der beschriebenen ersten manuellen Korrektur eine Note von 2,01. Nach Korrektur der Synthesequalität der 26 ausreichend oder mangelhaft ausgesprochenen Wörtern verbesserte sich der Wert der Gesamtbewertung aller Wörter auf eine Durchschnittsnote von 1,96. Die 26 schlecht bewerteten Wörter isoliert betrachtet, hatten zunächst eine Benotung von durchschnittlich 3,41 und verbesserten sich auf eine Durchschnittsnote von 2,37, siehe Tabelle 4.

Nach der Bewertung und erneuten Bearbeitung der Synthese blieb lediglich das Wort Bob in seiner Standardsynthese, da keine Alternative nach erneuter Bewertung innerhalb der Acapela Alternativen in den Atemgruppen null bis fünf für bessere Aussprache gefunden werden konnte.

Tabelle 4: Bewertung TTS-Synthese der 26 schlecht synthetisierten (Note > 3) Einsilber vor und nach der Korrektur der Synthesequalität.

<b>Wort</b>	<b>Note vor Korrektur</b>	<b>Note nach Korrektur</b>
Mai	3,71	2,57
Band	3,29	1,43
Schluck	3,14	2,57
Busch	3,29	2,43
Gleis	3,43	2,57
Milch	3,43	1,86
Wurm	3,14	2,83
Schmelz	3,57	2,29
Fan	3,86	1,86
Druck	3,14	2,57
Eck	3,57	2,14
Bug	3,43	1,86
Bob	3,14	3,14
Schmutz	3,14	1,86
Reim	3,43	2,29
Mail	3,86	3,00
Top	3,14	1,71
Öl	3,43	3,00
Trip	3,29	3,00
Müll	3,86	2,14
Nest	3,43	2,42
Start	3,14	2,71
Bruch	3,57	2,29
Watt	3,43	1,86
Hab	3,29	1,57
Bär	3,14	1,85
<b>Mittelwert</b>	<b>3,41</b>	<b>2,37</b>

### 4.3 Beschreibung der neuen Testlisten

Nach der Transkription der Wörter in ihre Phoneme mithilfe des G2P-Tools [37], wurden die Einsilber automatisch nach ihrer Anzahl an Phonemen gezählt und sortiert, wie in Tabelle 5 dargestellt. Bei der Zusammenstellung von Listen, die insgesamt 540 einsilbige Substantive enthalten, darunter 278 Einsilber mit vier Phonemen, muss beachtet werden, dass laut dem Hahlbrock-Schema (siehe Abschnitt Grundlagen, Abbildung 4) jede Liste zehn Wörter mit genau vier Phonemen umfassen muss. Unter diesen Vorgaben lassen sich aus dem vorhandenen Material maximal 27 Listen erstellen.

Tabelle 5: Tabellarische Übersicht von der finalen Phonemaufteilung, der neu extrahierten Einsilbersammlung (600) aus den verschiedenen Sprachkorpora.

<b>Anzahl der Phoneme</b>	<b>Anzahl an einsilbigen Substantiven</b>
1 Phonem	1 Wort
2 Phoneme	28 Wörter
3 Phoneme	225 Wörter
4 Phoneme	278 Wörter
5 Phoneme	64 Wörter
6 Phoneme	4 Wörter

Tabelle 6: Vergleich der Verteilung der Phonemanzahl in der neuen listenbezogenen Einsilbersammlung (540) mit den Vorgaben von Hahlbrock.

<b>Phonemanzahl</b>	<b>Anzahl Wörter</b>	<b>Anteil Sammlung</b>	<b>Anteil Hahlbrock</b>
2	27	4,7%	5%
3	186	34,4%	35%
4	265	49,1%	50%
5	62	11,5%	10%

Die Daten zeigen, dass die Zusammensetzung der neuen Einsilbersammlung weitgehend mit den Vorgaben von Hahlbrock übereinstimmt. Um die nach DIN EN ISO 8253-3:2022-11 [12] geforderte phonemische Äquivalenz für einen Vergleich der Testlisten miteinbeziehen zu können, wurde zunächst eine Zielverteilung der Phoneme benötigt. Hier wurde, wie unter Methodik in Abschnitt 3 beschrieben, die Statistik von Kohler zur Hilfe genommen [17]. Zur besseren Vergleichbarkeit mit der Literatur von Kohler wurde die Analyse von der endgültigen Wortauswahl für die 27 Listen von 540 Einsilbern und für die 20 Listen des FET von 400 Einsilbern in IPA transkribiert und in neun Phonemklassen zusammengefasst. Eine Einteilung der einzelnen Phoneme in die Phonemklassen ist in Abbildung 13 zu finden. Zusätzlich zu den sieben, die die Norm nennt, wurden nach Kohler die Phonemklassen sonstige Konsonanten und sonstige Vokale hinzugenommen, um eine eindeutige Gruppierung der Phoneme gewährleisten zu können. Sie sind bisher nicht in der Norm [12] beschrieben. Die sonstigen Vokale *aI*, *au*, *ɔI*, *œ* und Konsonanten *j*, *l*, *r*, *ʒ*, *ts*, *tʃ*, *pf*, *dʒ* waren relevant bei der Trennung und Zählung der einzelnen Phoneme pro Wort, um die Listenstruktur nach Hahlbrock korrekt umzusetzen und konnten daher nicht weggelassen werden. Die sonstigen Konsonanten der Affrikate *ts*, *tʃ*, *pf*, *dʒ* und sonstigen Vokale des vokalisiertem Diphthongs *œ* wurden in der Abbildung 15 nicht mit ausgewertet, da sie in der Literatur von Kohler keine Statistikreferenzen besitzen.

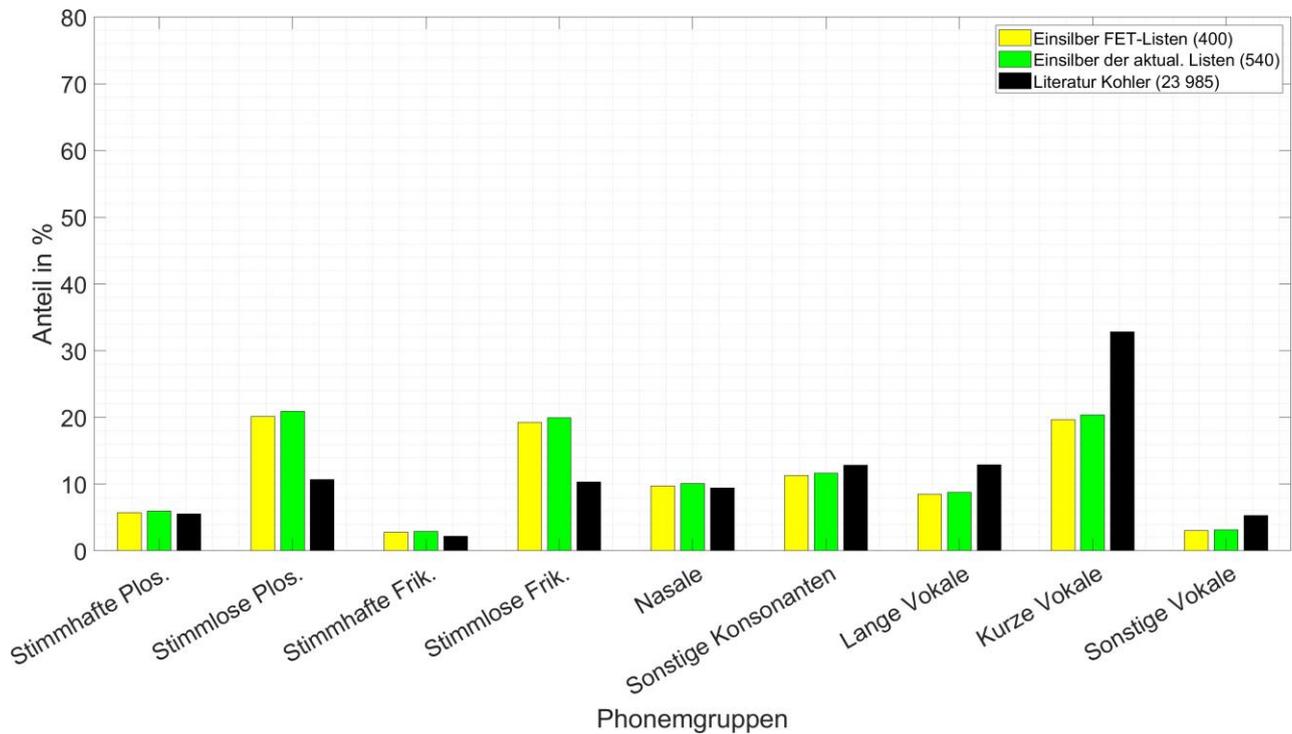


Abbildung 13: Phonemklassen mit ihrem Anteil an der Deutschen Sprache in Prozent nach K. J. Kohler [17] als Zielwerte für phonemisch ausgewogene Listen im Vergleich mit dem FET (400) und dem Korpus der aktualisierten Einsilberlisten (540).

In Abbildung 13 werden die neun Phonemklassen und ihr Anteil an der deutschen Sprache gezeigt, basierend auf Daten von Kohler im Vergleich zu den Phonemklassen des FET und aktualisierten Listen. Besonders auffällig ist, dass die stimmlosen Plosive und Frikative in den neuen Listen und dem FET etwa 10 % stärker vertreten sind als in den Zielwerten von Kohler und die kurzen Vokalen hingegen bei beiden etwa 10 % schwächer. Auch bei den langen Vokalen sind der FET und die neue Einsilber-Sammlung ca. 5 % geringer repräsentiert. Dieser Unterschied kann durch die typische Struktur von Einsilbern verursacht werden und zeigte sich unabhängig auch in größeren Einsilbersammlungen, in denen noch keine Reduzierung für die Listen vorgenommen wurde. Die anderen fünf Gruppen der stimmhaften Plosive und Frikative, Nasale, sonstige Konsonanten und Vokale sind vergleichbar in ihrer Verteilung zur Literatur und weichen im Mittel weniger als 0,5 Prozentpunkte ab. Die neun Phonemklassen mit dem zugehörigen genauen Zielprozentsatz nach Kohler sind in der Methodik Tabelle 1 aufgeführt.

In Abbildung 14 ist beispielhaft eine vollständige Liste mit den Wörtern, der phonemischen Repräsentation in IPA-Darstellung und der zugehörigen Übersicht der Worte mit zwei bis fünf Phonemen von klein nach groß dargestellt.

FET(10).word						
	1	2	3	4	5	6
1 Zoo		ts	o:			
2 Raub	r		aʊ	p		
3 Ring	r		ɪ	ŋ		
4 Wohl	v		o:	l		
5 Arm	a		r	m		
6 Kick	k		ɪ	k		
7 Knie	k		n	i:		
8 Lob	l		o:	p		
9 Mars	m		a	r	s	
10 Dank	d		a	ŋ	k	
11 Kampf	k		a	m	pf	
12 Blatt	b		l	a	t	
13 Plan	p		l	a:	n	
14 Fleck	f		l	ɛ	k	
15 Gramm	g		r	a	m	
16 Mast	m		a	s	t	
17 Mist	m		ɪ	s	t	
18 Block	b		l	ɔ	k	
19 Brand	b		r	a	n	t
20 Werft	v		ɛ	e	f	t

Abbildung 14: Beispielhaft Testliste 10 mit IPA-Phonemschriftdarstellung und Verteilung der Phone-  
me auf die jeweiligen Einsilber sortiert von klein (2-Phonemer) nach groß (5-Phonemer).

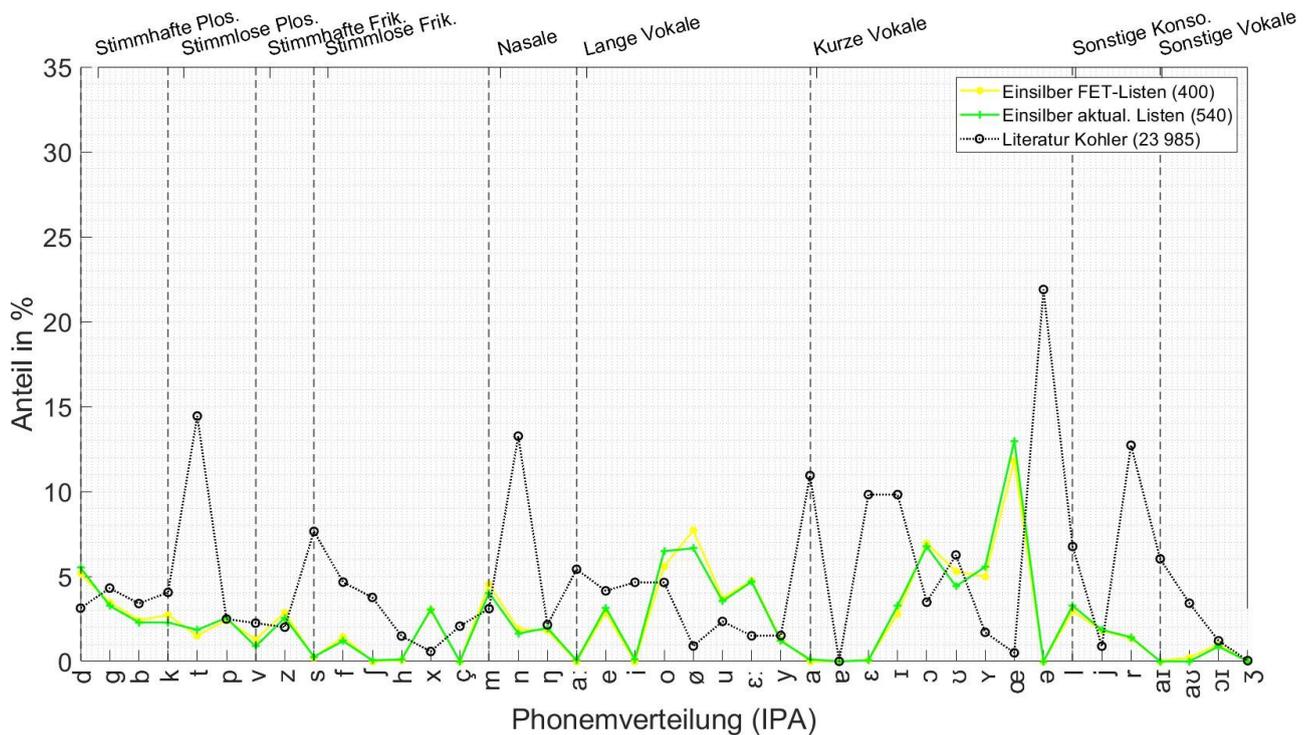


Abbildung 15: Phonemische Verteilung der Einsilber-Testlisten im Vergleich mit den Literatur nach Kohler transkribiert in 40 IPA Zeichen und sortiert in neun Phonemklassen.

Abbildung 15 veranschaulicht die Verteilung der einzelnen Phoneme, um spezifische Abweichungen deutlicher einzuordnen. Die Analyse zeigt, dass sowohl im FET als auch in der aktualisierten Sammlung, neben dem Schwa-Laut ə, der häufig an unbetonten Wortendungen auftritt (z.B. Lehrer), auch andere Phoneme wie t, n, a und r aus den Gruppen der stimmlosen Plosive, Nasale, kurzen Vokale und sonstigen Konsonanten in ihrer Häufigkeit um mehr als 10 % von den Literaturwerten nach Kohler abweichen. Bei den Phonemengruppen der stimmhaften Plosive und Frikative, langen Vokalen und sonstigen Vokalen besteht weitgehend Übereinstimmung mit den Literaturwerten. Diese Unterschiede spiegeln hauptsächlich die charakteristische Struktur von Mehrsilbern und Sätzen im Deutschen wider, die sich von der einfachen Struktur einsilbiger Substantive unterscheidet. Die Abbildung 13 und Abbildung 15 zeigen deutlich die große Ähnlichkeit zwischen dem FET und den aktualisierten Einsilberlisten und dokumentieren zugleich die Abweichungen zur Phonemverteilung in der deutschen Literatur nach Kohler.

### 4.3.1 Perzeptive Äquivalenz

Eine Analyse der in der Probandenstudie ermittelten Sprachverständlichkeitswerte für den aktualisierten synthetischen FET ergab unterschiedliche Sprachverständlichkeitsschwellen zwischen den Testlisten, wobei der mittlere SRT50 für die einzelnen Testlisten variierte. In Abbildung 16 sind die 27 psychometrischen Funktionen für jede Liste und eine gemittelte gestrichelte Funktion dargestellt. Testliste 4 zeigt mit einem Pegel von 29,4 dB den höchsten SRT50, während Testliste 8 mit 25,9 dB den niedrigsten Pegel aufweist. Der gemittelte SRT50 über alle Testlisten hinweg beträgt 27,9 dB und

die durchschnittliche Steigung am SRT50 beläuft sich auf 5,5 % pro dB. Diese Ergebnisse der psychometrischen Funktionen sind konsistent mit früheren Studien [11], [10], [45] und zeigen Ähnlichkeit zu den Bezugskurven [44]. Jedoch zeigen die Unterschiede am SRT50 von bis zu 3,5 dB zwischen den einzelnen Testlisten eine zu optimierende perzeptive Äquivalenz.

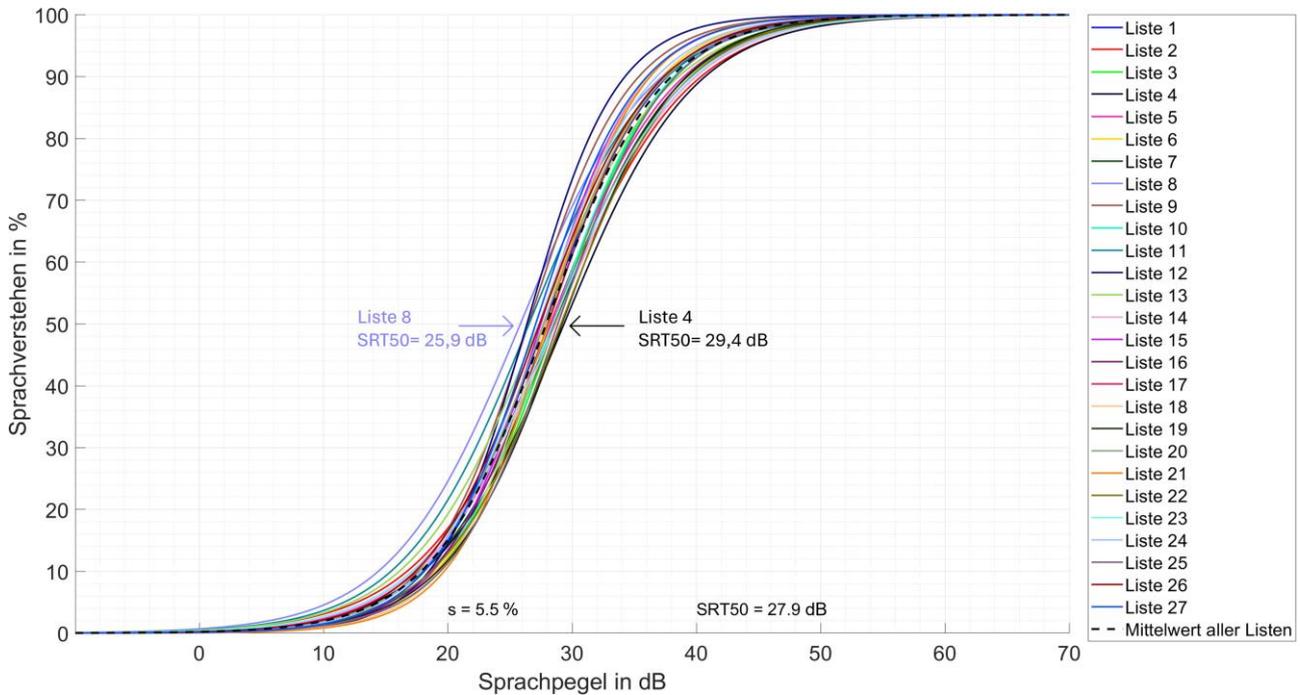


Abbildung 16: Psychometrische Funktionen der 27 Listen für die Schallpegel 21,5 dB, 27,5 dB und 33,5 dB mit Angabe der besten Liste 8 (lila) und der schlechtesten Liste 4 (schwarz).

Im Folgenden werden in Tabelle 7 und Abbildung 17 die Ergebnisse der SRT50 Werte und Steigungen für den aktualisierten synthetischen Freiburger Einsilbertest im Vergleich zu den original Bezugskurven aus der Norm DIN45626-1:1995-08 [44] und der weiteren Literatur dargestellt.

Tabelle 7: Vergleich der gemittelten Sprachverständlichkeitsschwelle und Steigung der psychometrischen Funktionen.

Messung	SRT (dB)	Steigung (%/dB)
Original FET monaural (DIN 45626-1:1995-08 [44])	29,3	4,0
Original FET binaural mod. (-2,5 dB) (DIN 45626-1:1995-08 [44])	26,8	4,0
Original FET binaural (Thiele et al., 2014 [45])	27,8	6,0
Original FET binaural (Baljic et al., 2016 [14])	29,2	4,6
Original FET binaural (Schwarz et al., 2022 [10])	28,8	5,4
Synthetischer FET binaural (Schwarz et al., 2022 [10])	28,8	5,7
Aktualisierter synthetischer FET binaural (Harries et al., 2024)	27,9	5,5

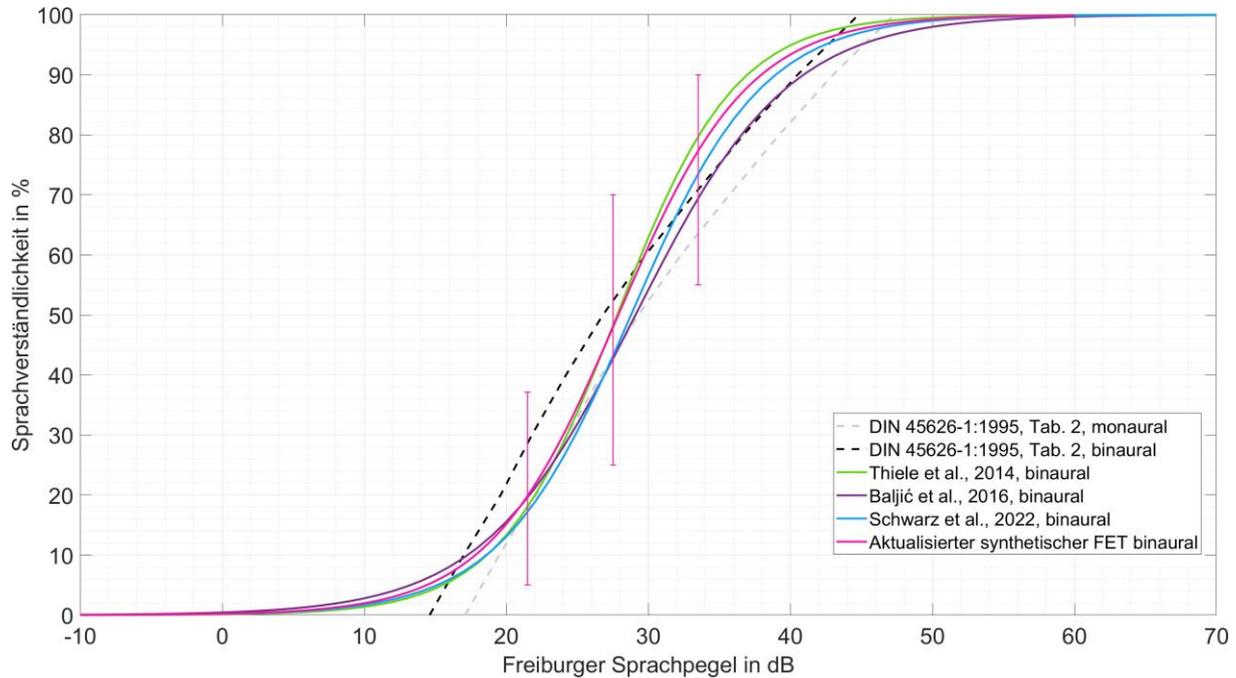


Abbildung 17: Psychometrische Funktionen der 27 Listen im Literaturvergleich. Für die psychometrische Funktion des aktualisierten synthetischen FET (pink) sind zusätzlich die 95 %-Konfidenzintervalle der Sprachverständlichkeit pro Messpegel gezeigt.

Für eine verbesserte Vergleichbarkeit wurde die Bezugskurve der Norm (monaural mit Kopfhörern gemessen) gemäß DIN45626-1:1995 [44], die eine Steigung von 4,03 %/dB aufweist, um 2,5 dB zu niedrigeren Pegeln verschoben, um eine Annäherung an binaurales Hören im Freifeld zu simulieren. In der Literatur ist bekannt, dass binaurales Hören eine um etwa 3 dB verbesserte Wahrnehmung gegenüber monauralem Hören über Kopfhörer bietet, während der Gewinn im Freifeld bei etwa 2,5 dB liegt [46]. Die psychometrische Funktion der aktualisierten synthetischen Listen zeigt mit einer Steigung von 5,5 %/dB am SRT50 einen steileren Kurvenverlauf als die Bezugskurven der Norm. Die psychometrische Funktion des aktualisierten synthetischen FET kreuzt die binaurale Bezugskurve bei 60 % und liegt daher am SRT50 bei einem 1,1 dB höheren Pegel als die Norm.

In den grafischen Darstellungen wurden Fehlerbalken aus dem 2,5. und 97,5. Perzentil eingefügt, die die Probandenstreuung verdeutlichen. Diese repräsentieren das Konfidenzintervall, in dem 95 % der Daten liegen, und bieten eine Darstellung der interindividuellen Streuung der Probandenergebnisse. Im Vergleich zu der Norm DIN45626-1:1995 [44] repräsentieren die Kurven von Thiele, Schwarz und Baljić neuere binaurale Messungen im Freifeld aus der Literatur [11], [10], [45], [14], die durchweg steilere Anstiege und niedrigere SRT-Werte aufweisen. Ein zweiseitiger T-Test zeigt signifikante Unterschiede zwischen der neuen gemittelten psychometrischen Funktion der aktualisierten Testlisten und den Funktionen der Norm sowie von Baljić, sowohl beim SRT50 als auch bei der Steigung. Im Gegensatz dazu weicht die gemittelte psychometrische Funktion des neuen synthetischen FET nicht

signifikant unterscheiden ( $t(26) = 1,4, p = 0,17$ ) von den Messungen von Schwarz et al. [11] bezüglich der Steigung und von Thiele et al. ( $t(26) = -0,27, p = 0,79$ ) hinsichtlich des SRT50 ab [45]. Diese Übereinstimmungen verdeutlichen, dass die neuen Messergebnisse teilweise mit den Daten aus der Literatur übereinstimmen. Diese Ergebnisse bestätigen teilweise die Vergleichbarkeit der Messungen des noch nicht perceptiv und phonemisch optimierten Testmaterials unter den Bedingungen des binauralen Hörens und betonen die Bedeutung der Fehlerbalken für das Verständnis der Datenvarianz bei der eingesetzten Probandenanzahl von 27.

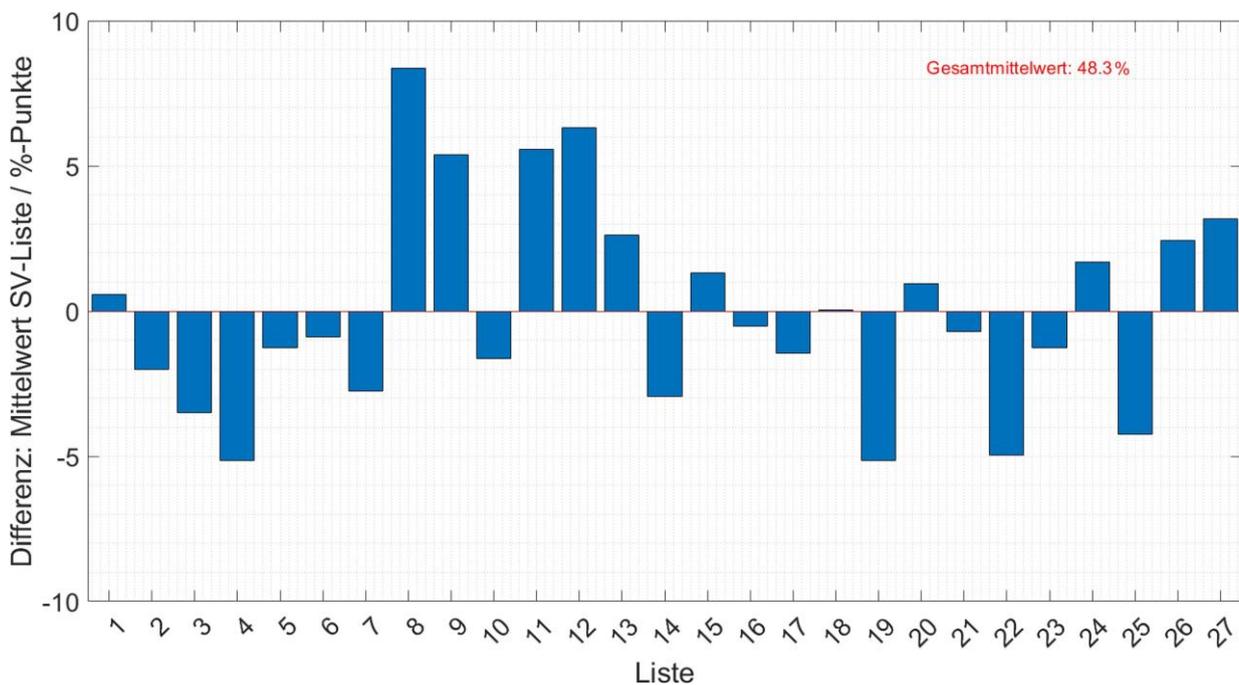


Abbildung 18: Abweichungen der Listenverständlichkeit vom Mittelwert (in Prozentpunkten).

Die Abbildung 18 illustriert die Differenzen der gemittelten Sprachverständlichkeit der einzelnen Listen über die Schallpegel 21,5 dB, 27,5 dB und 33,5 dB im Vergleich zum Gesamtmittelwert aller 27 Testlisten, der bei 48,3 % liegt. Die Balken zeigen, wie jede einzelne Testliste relativ zu diesem Durchschnittswert abschneidet. Diese Balkendiagramm Darstellung ermöglicht eine direkte Übersicht in die Differenzen der Listenverständlichkeit.

Testlisten mit überdurchschnittlicher Verständlichkeit sind durch positive Abweichungen gekennzeichnet, wobei Liste 8 die höchste positive Abweichung von gerundet +8 Prozentpunkten aufweist. Im Gegensatz dazu zeigen negative Abweichungen Listen mit unterdurchschnittlicher Verständlichkeit, mit Liste 4 als deutlichstes Beispiel, die eine negative Abweichung von gerundet -5 Prozentpunkten verzeichnet. Weitere Listen, die Abweichungen vom Gesamtmittelwert von mehr als 5 Prozentpunkten aufweisen, sind Liste 9, Liste 11, Liste 12 und Liste 19. Die mittlere absolute Abweichung der Verständlichkeitswerte aller Listen liegt bei 2,8 Prozentpunkten, während der RMS-Wert aufgrund der Ausreißer etwas höher bei 3,5 Prozentpunkten liegt. Diese Ergebnisse weisen auf eine Variabilität in

der Verständlichkeit über die 27 Listen hin und zeigen damit, noch deutlicher als die psychometrischen Funktionen, eine fehlende perzeptive Äquivalenz und ein umfassendes Optimierungspotential auf.

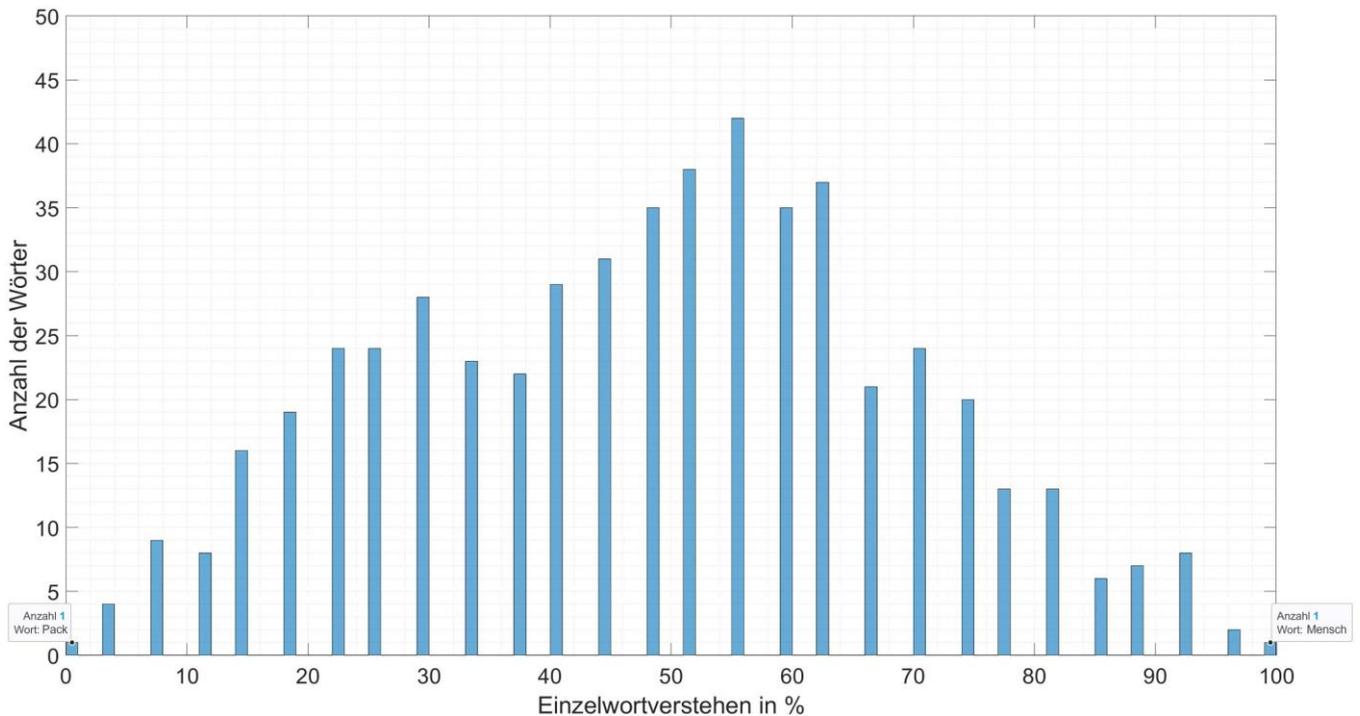


Abbildung 19: Häufigkeitsverteilung des prozentualen Einzelwortverstehens der 540 Einsilber aus den 27 Listen mit N=27 Probanden.

In Abbildung 19 ist die Häufigkeitsverteilung des prozentualen Einzelwortverstehens der 540 Einsilber aus den 27 Listen mit N=27 Probanden dargestellt. Der Shapiro-Wilk-Test ergab einen W-Wert von 0,9 und einen p-Wert ( $p < 0,001$ ), was auf eine nicht-normalverteilte Datenverteilung hinweist. Die Datenanalyse der Ergebnisse des Einzelwortverstehens, gemittelt aus oben genannten drei Schallpegeln in Abbildung 19 zeigt, dass die Häufigkeitsverteilung des Einzelwortverstehens aufgrund der überwiegend symmetrischen Verteilung (Schiefe = -0,01) um den Mittelwert trotzdem Ähnlichkeit mit einer Normalverteilung aufweist. Die größte Anzahl an Wörtern findet sich in der Mitte der Häufigkeitsverteilung. Dieser Bereich weist eine hohe Dichte auf, was bedeutet, dass ein Großteil der Einsilber im Mittel zu 50 % verstanden wurde, während die Häufigkeit des Einzelwortverstehens an den Rändern bei niedrigem und hohem prozentualen Sprachverstehen abfällt. Die Kurtosis (Wölbung) von 2,4 zeigt, dass die Verteilung platykurtisch ist, also flacher und mit dünneren Enden als eine Normalverteilung. Dies bedeutet, dass Extremwerte seltener auftreten. Der Wert unter 3 weist darauf hin, dass die Verteilung weniger extreme Ausreißer hat. Beispielhaft sind die wenigen Extremwerte der 540 Einsilber das Wort Mensch als einziges mit einem Sprachverstehen von 100 % vertreten und des Wortes Pack ebenfalls nur einmal mit einem Sprachverstehen von 0 %. Das heißt, dass diese beiden Wörter pegelunabhängig von allen Probanden immer bzw. nie verstanden wurden.

### 4.3.2 Korrelationsanalyse von Bewertung und Sprachverstehen

In der Abbildung 20 zur Korrelation wird der Zusammenhang zwischen der durchschnittlichen Note jedes einzelnen Wortes und dem Prozentsatz des gemittelten Sprachverstehen aller Probanden dargestellt. Die Korrelationsanalyse zwischen dem Einzelwortverstehen und der Wortbewertung zeigt eine negative Korrelation ( $r(538) = -0,25$ ). Dies bedeutet, dass mit einer schlechteren Wortbewertung (steigender Schulnote) das Sprachverstehen sinkt. Der ermittelte p-Wert ( $p < 0,001$ ) bestätigt zusätzlich, dass die Korrelation der Wortbewertung mit dem Einzelwortverstehen statistisch hoch signifikant ist. Abgesehen von einem Ausreißer (Bob), der trotz Überarbeitung eine Bewertung oberhalb der Zielnote drei blieb, erfüllten alle Wörter das Kriterium einer Schulnote von drei oder kleiner.

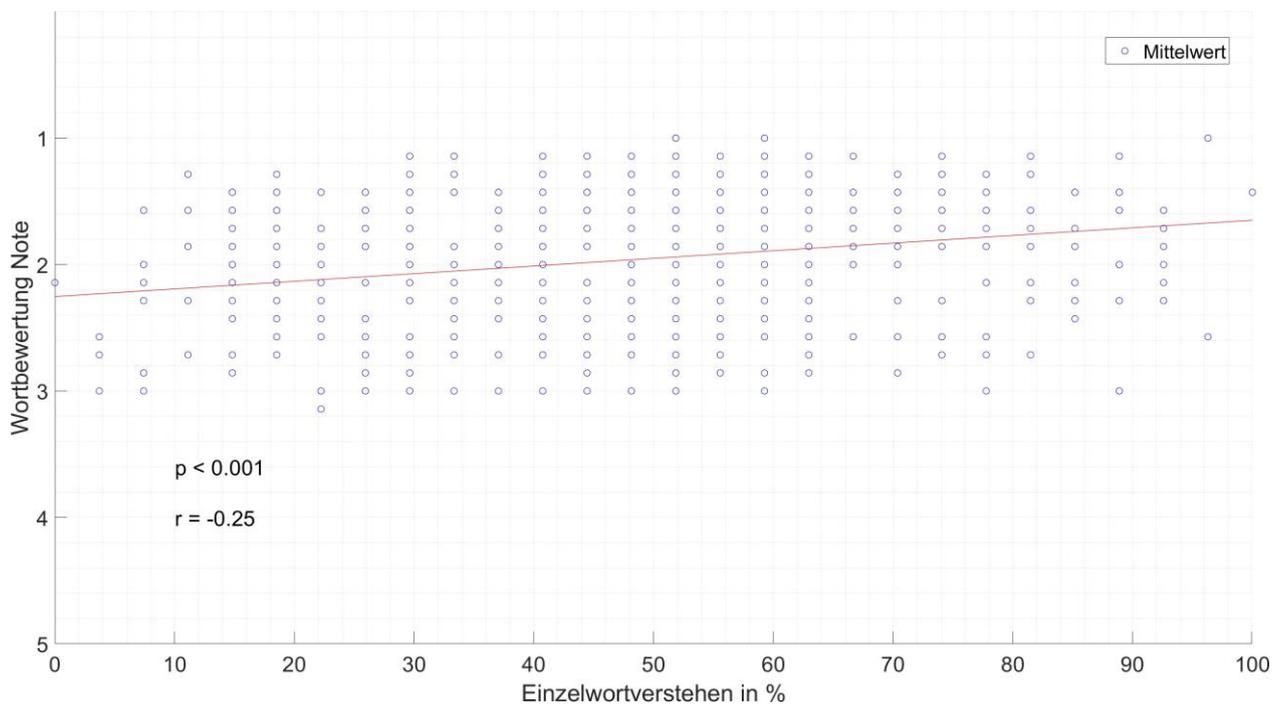


Abbildung 20: Korrelation von Wortbewertung der TTS-Synthese und Sprachverstehen der 540 ein-silbigen Substantive.

## 4.4 Optimierung

Im Optimierungsprozess wurde die perzeptive Äquivalenz der Testlisten durch ein iteratives Verfahren verbessert, indem die Differenzen zwischen den einzelnen Listen und ihrem jeweiligen Mittelwert systematisch reduziert wurden. Die phonemische Äquivalenz wurde außer Acht gelassen. Die in Abbildung 21 dargestellten Daten zeigen eine deutliche Verbesserung der perzeptiven Äquivalenz in allen 27 Listen. Im Rahmen der Optimierung wurde festgestellt, dass die initiale mittlere Differenz in der Sprachverständlichkeit, welche bei 3,53 Prozentpunkten lag, sehr stark reduziert werden konnte. Durch die Anpassung der Gewichtung der phonemischen Äquivalenz auf null, also durch deren bewusste Vernachlässigung, sank diese Differenz auf 0,07 Prozentpunkte. Die durchgeführte statistische Analyse mit einem zweiseitigen t-Test zeigt eine hoch signifikante  $p < 0,001$  Verbesserung nach der Optimierung.

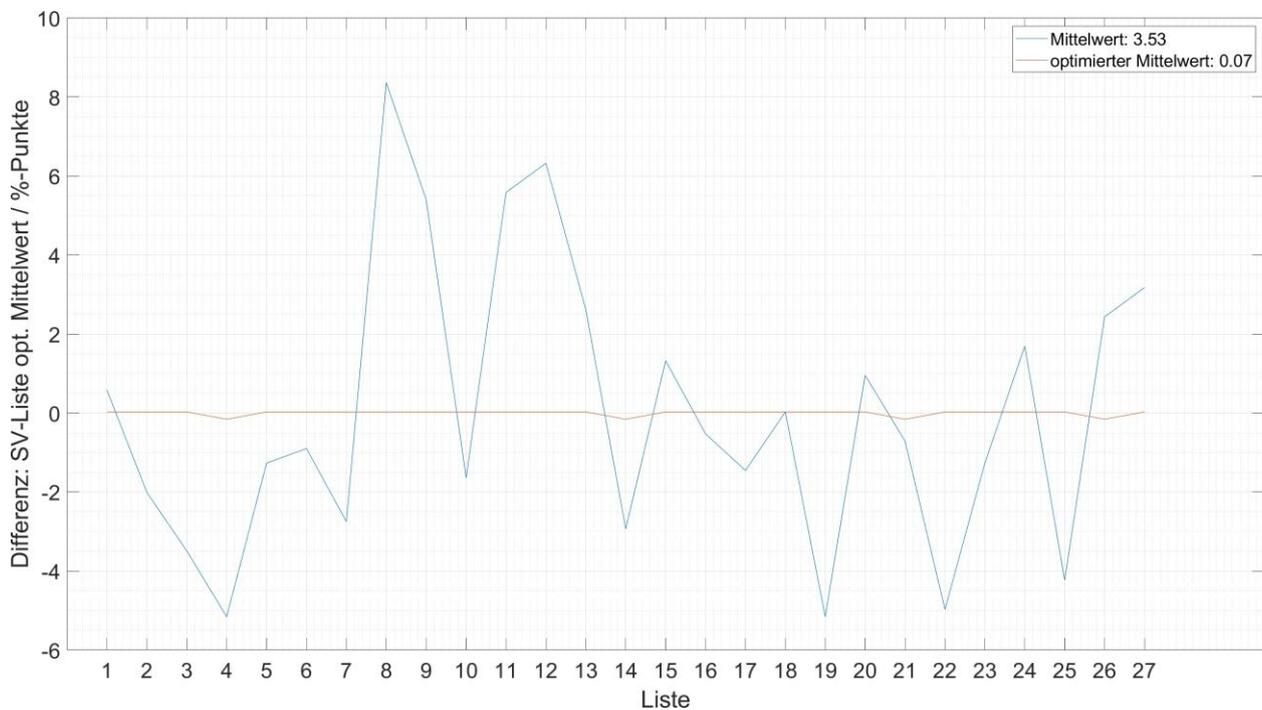


Abbildung 21: Perzeptive Äquivalenz pro Liste vor (blau) und nach (rot) der Optimierung.

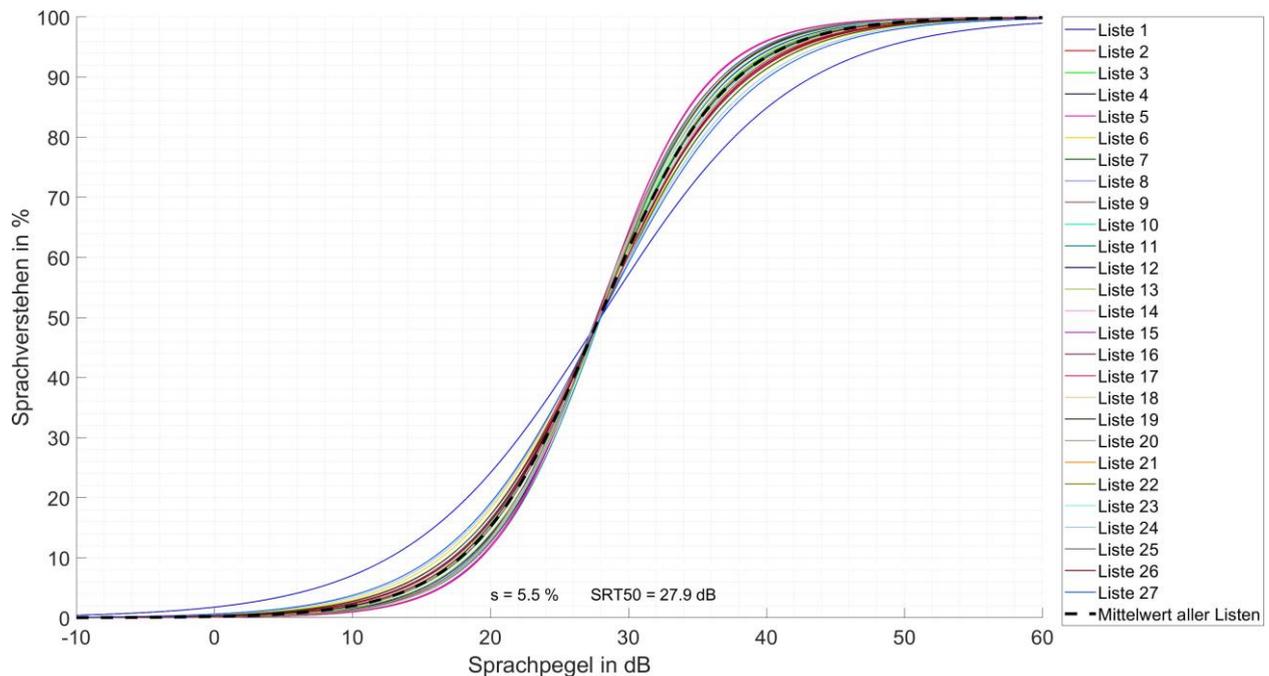


Abbildung 22: Psychometrische Funktionen der 27 Listen nach Optimierung der perzeptiven Äquivalenz.

Nach der Optimierung der perzeptiven Äquivalenz ist in Abbildung 22 zu beobachten, dass die einzelnen psychometrischen Funktionen für die Testlisten enger beieinander liegen, was auf eine verbesserte Konsistenz in der Sprachverständlichkeit hinweist. Die Ausnahme bildet die Liste 1, die eine flachere Steigung hat und außerhalb des SRT50 etwas weiter von den anderen psychometrischen Funktionen entfernt ist. Der durchschnittliche Sprachverständlichkeitsschwellenwert (SRT50) und die Steigung der Funktionen haben sich durch die Optimierung aufgrund der Mittelwertberechnung aller Werte nicht verändert, was bedeutet, dass die mittlere Verständlichkeit der Funktionen über alle Listen hinweg gleich geblieben ist.

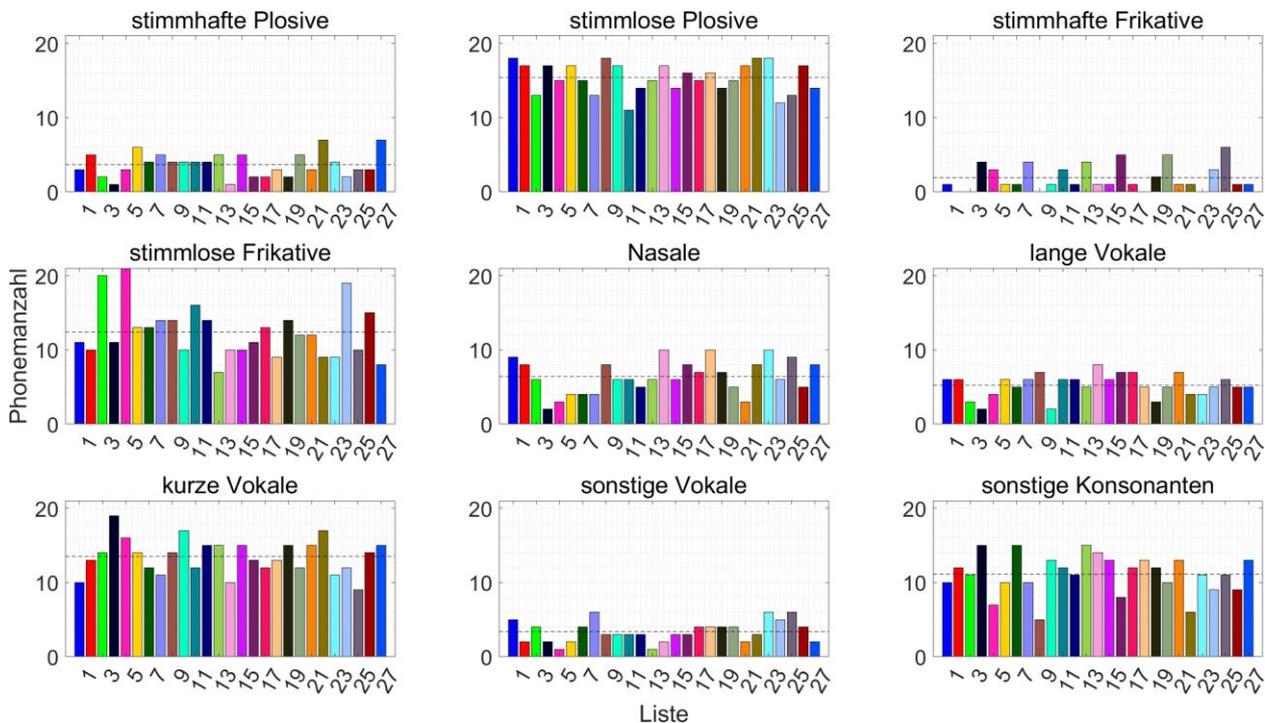


Abbildung 23: Darstellung der phonemischen Verteilung inkl. Mittelwertlinie der 27 Listen für jede der neun Phonemklassen.

In Abbildung 23 sind die einzelnen perzeptiv optimierten Testlisten in neun Phonemklassen aufgeteilt und mit ihrem jeweiligen Sprachverstehen als Prozentwerte dargestellt. Jede Phonemklasse ist durch eine Mittelwertlinie ergänzt, welche die durchschnittliche Verteilung der Phoneme in dieser Gruppe über alle 27 Listen hinweg kennzeichnet. Dies ermöglicht eine visuelle Auswertung der Schwankungen zwischen den einzelnen Listen sowie der Abweichungen jeder Liste vom Gesamtmittelwert ihrer Phonemklasse. Es ist zu sehen, dass die perzeptiv optimierte Zusammenstellung der 27 Testlisten über die neun Phonemklassen hinweg phonemisch nicht äquivalent sind. Besonders die stimmhaften und stimmlosen Frikative weisen die größten Schwankungen zwischen den Listen auf. Die Gruppen der Nasale, langen und kurzen Vokale sowie der sonstigen Konsonanten variieren ebenfalls, aber in einem geringeren Ausmaß. Nur die drei Gruppen der stimmhaften und stimmlosen Plosive und der sonstigen Vokale sind ausgeglichener. Diese Unterschiede in der phonemischen Verteilung bilden die Grundlage für weitere Analysen und der zukünftig zusätzlichen Optimierung der phonemischen Äquivalenz in Kombination mit der perzeptiven Äquivalenz.

## 5 Diskussion

### 5.1 Auswahl der Einsilber und Vergleich zum Freiburger Einsilbertest

Die Aktualisierung des FET erfolgte durch die Zusammenstellung aller einsilbigen Substantive aus sechs Sprachkorpora, ohne anfangs eine Filterung nach Wortfrequenz oder Tokenanzahl vorzunehmen. Eine systematische Analyse führte zur Identifikation von Wörtern, die durch Schreibfehler oder veraltete Rechtschreibvarianten beeinflusst waren. Nach dieser ersten Sichtung wurde mithilfe einer API-basierten Frequenzanalyse die Wortauswahl auf 600 häufige Substantive eingeschränkt, indem seltene Wörter mit einer Frequenz von weniger als drei ausgesondert wurden. Das LCC-Korpus wies durch Schreibfehler und die fehlende Lemmatisierung eine hohe Zahl an Mehrfachzählungen auf, was in der finalen Auswahl berücksichtigt wurde. Die Anpassung der Wortfrequenzgrenze auf drei (entsprach im November 2022 einer Tokenanzahl von mindestens 111 540) bis fünf (von 11 153 946 bis 111 539 457) beeinflusste die Menge der einbezogenen Substantive erheblich. Die Wortfrequenz sechs enthielt keine einsilbigen Substantive. Allerdings blieb diese Filterung wirkungslos bei Wörtern, die durch orthographische Fehler wie alte Rechtschreibung, Namen oder falsche Groß- und Kleinschreibung verzerrt waren. Die erforderliche Nachbearbeitung durch eine Person, die Wörter nach festgelegten Kriterien manuell auszusortieren, sollte künftig möglichst vermieden werden. Diese Phase ist anfällig für Fehler aufgrund subjektiver Einschätzungen bei der Verwendung von Wörtern und Namen. Beispielsweise wurde das seltene Wort Lee manuell entfernt, obwohl es eine Wortfrequenz von drei aufweist. Dies geschah, da Lee nicht nur ein Fachbegriff aus dem Wassersport ist, sondern auch als Familien-, Vor- und Markenname verwendet wird und daher die Wortfrequenz von drei überhöht erscheint. Die unterschiedliche Wortfrequenz ein und derselben Wörter in alter und neuer Rechtschreibung bereitete ebenfalls Selektionsprobleme in der Automatisierung. Teilweise wurden beide Varianten eines Wortes selektiert, dann musste die veraltete Variante händisch entfernt werden. In wenigen Fällen wurde durch die Wortfrequenz die neue Rechtschreibung aufgrund einer niedrigeren Frequenz verworfen und die alte Schreibweise wie z.B. bei Schloß selektiert. In diesem Fall wurde die Auswahl der Wörter mit hoher Wortfrequenz beibehalten, weil davon auszugehen ist, dass sie durch die Rechtschreibreform unzureichend von der API gezählt wurden. Diese Annahme bestätigt sich nach Prüfung über die Homepage des DWDS. Hier wird automatisch immer die höhere Tokenzahl berücksichtigt und beiden Orthografien zugeordnet.

Ziel sollte es sein, diesen Prozessschritt von persönlichen Entscheidungen zu entkoppeln und zu automatisieren. Der hier beschriebene Ansatz der Wortfrequenzfilterung, die direkten Zugriff auf das DWDS Online-Wörterbuch mit Wortfrequenzangaben bietet, hat viele der genannten Probleme behoben. Allerdings stößt auch dieser Ansatz bei Eigennamen, insbesondere bei solchen mit Doppelbedeutungen, und alter Rechtschreibung an seine Grenzen. Es bleibt daher zu überlegen, ob der fehlerbehaftete LCC durch andere, idealerweise lemmatisierte und auf Rechtschreibung geprüfte Korpora, ersetzt werden sollte. Ob eine finale manuelle Kontrolle aller automatisch extrahierten Wörter in Zukunft entfallen kann, hängt von der weiteren Entwicklung des DWDS API-Wortfrequenzfilters und der Auswahl der Korpora ab.

## 5.2 Listenzusammenstellung

Bei der initialen Zusammenstellung der Listen mit den aktualisierten einsilbigen Substantiven wurden, abgesehen von der Phonemstruktur einzelner Listen gemäß den Vorgaben von Hahlbrock zur Phonemgruppenanzahl [9], keine weiteren Merkmale für die Randomisierung berücksichtigt. Mit dem gewählten Ansatz, die Listen für den FET nach dem von Hahlbrock gewähltem Aufbau in Bezug auf die Wortlänge und Phonemverteilung zusammenzustellen, können mit den aus den sechs Korpora extrahierten 600 einsilbigen Substantiven maximal 27 Listen mit insgesamt 540 Wörtern zusammengestellt werden. Die Anzahl ist bisher überwiegend begrenzt durch die Einsilber mit zwei Phonemen, die in jeder Liste jeweils einmal vorkommen müssen, aber sehr selten sind und bei den neuen aktualisierten Listen durch die Einsilber mit vier Phonemen, die nach Hahlbrock in jeder Liste jeweils zehnmal vorkommen müssen. Eine Anpassung der Listenstruktur, etwa durch die Integration des einsilbigen Substantivs *Ei* mit nur einem Phonem und ggf. das Ersetzen eines fünfphonemigen Einsilbers durch einen sechsphonemigen wie *Strand*, könnte die Erstellung einer zusätzlichen Liste ermöglichen. Das ist in ähnlicher Weise auch bei der Erstellung des FET durch Hahlbrock in Liste 2 geschehen, siehe Abbildung 4. Die Summe pro Liste von 73 Phonemen ändert sich bei einem solchen Tausch nicht, jedoch erschwert dieses Vorgehen den Prozess der Abstimmung der automatisierten Listenzusammensetzung und führt zusätzlich wegen der in dieser Auswertung ebenfalls fehlenden vierphonemigen Einsilbern in dieser Studie zu keinen weiteren Listen. Die Flexibilisierung der Hahlbrock-Kriterien, die eine feste Anzahl von 73 Phonemen pro Liste vorsehen, wie sie bereits von Hahlbrock selbst in Liste 2, 5 und 15 praktiziert wurde, könnte für zukünftige Erweiterungen oder für die Optimierung der Listenverständlichkeit eine Option bieten. Eine solche Anpassung würde mehr Spielraum in der Zusammensetzung der Listen erlauben, was insbesondere hilfreich sein kann, wenn das Ziel darin besteht, eine größere Anzahl an Listen zu erstellen oder die phonemische Äquivalenz zwischen den Listen zu verbessern.

Aufgeteilt in Gruppen nach Phonemanzahl kann man in Tabelle 6 anhand der Anteile an der Gesamtheit der Wörter gut die von Hahlbrock in seiner Struktur angestrebten Proportionen und deren Übereinstimmung zwischen den Einsilberlängen erkennen. Kleinere Abweichungen in der Phonemanzahlverteilung sind aufgrund des vokalischen Diphthongs bei acht Wörtern (*Burg*, *Durst*, *Furcht*, *Kurs*, *Sturm*, *Sturz*, *Turm*, *Wurst*) und des Weglassens des Minimalpaars (Wort mit einem Phonem) *Ei* vorhanden. Bei diesen Wörtern ist ein Phonem zu wenig enthalten, weil der vokalische Diphthong als ein Phonem gezählt wurde. Die Vorgehensweise Diphthonge und Affrikate als ein Phonem zu zählen, da sie in der Regel untrennbar in einer Silbe liegen, ist grundsätzlich korrekt aber uneinheitlich geregelt. Hahlbrock zählte z.B. nur das *ts* als einzelnes Phonem bzw. als Affrikat. Die Transkription einer nicht hochdeutschen Variante durch das Bayerische Archiv für Sprachsignale (BAS), insbesondere bei dem konsonantischen *r*, führte zu leichten Abweichungen. Da Hahlbrock das Kriterium der 73 Phoneme pro Liste selbst flexibel auslegte, stellt diese Diskrepanz keine schwerwiegende Problematik dar. Ebenso erschwert die Inklusion von Anglizismen, die oft nichtdeutsche Affrikate und Diphthonge enthalten, eine einheitliche Gruppierung und Klassifizierung der Phoneme. Um eine korrekte Annäherung an das hochdeutsche Phoneminventar zu gewährleisten, könnte eine Anpassung der Transkriptionsstan-

dards oder des Transkriptionsprogramms notwendig sein. Die Transkription von Anglizismen und die Anwendung eines auf bayrische Dialektmerkmale ausgerichteten Transkription des Bavarian Archive for Speech Signals (BAS) haben dazu geführt, dass das hochdeutsche Phoneminventar für die acht genannten Wörter nicht adäquat abgebildet wurde. Diese Problematik wurde durch die geringe Stichprobengröße von 100 Transkriptionsüberprüfungen nicht offensichtlich. Für zukünftige Arbeiten sollte erwogen werden, ein anderes Transkriptionsprogramm zu nutzen und eine umfassendere Prüfung der Transkriptionen durchzuführen, bevor Auswertungen vorgenommen werden. Dies könnte auch durch den Einsatz mehrerer Transkriptionsprogramme und deren vergleichende Analyse mittels automatischer Filterung erfolgen, um konsistente und standardisierte Daten zu gewährleisten.

Hahlbrock zog neben der Ausgewogenheit der Phonemgruppenanzahl auch eine gleichmäßige Verteilung der Initialphoneme in Betracht, ähnlich dem Ansatz des englischen Einsilbertests, der sich auf Konsonant-Vokal-Konsonant (KVK) Strukturen konzentriert. Bei einer kleineren Anzahl von Listen, die jedoch mehr Wörter enthalten als der FET, ist es einfacher, eine phonemische Äquivalenz zu erreichen [9]. Dieses Kriterium könnte für zukünftige Verbesserungen des FET eine Möglichkeit bieten, um die Äquivalenz unter Reduzierung der Listenanzahl weiter zu optimieren. Die Auswertung der neun Phonemklassen aus Abbildung 13 zeigte, dass die aktualisierte Einsilbersammlung eine unterschiedliche phonemische Verteilung gegenüber der deutschen Literatur aufweist [17], die mehrheitlich auf die Eingrenzung auf aktuelle einsilbige Substantive und der Berücksichtigung von Anglizismen zurückzuführen ist. In der detaillierten Analyse der Phonemverteilung in Abbildung 15 wird deutlich, dass sowohl im FET als auch in der aktualisierten Sammlung bestimmte Phoneme, darunter der Schwa-Laut ə, sowie die Phoneme t, n, a und r größere Abweichungen von mehr als 10 % im Vergleich zu den in der Literatur verzeichneten Häufigkeiten aufweisen. Diese Phoneme umfassen die Gruppen der stimmlosen Plosive, Nasale, kurzen Vokale und sonstigen Konsonanten. Andererseits zeigen die Phonemgruppen der stimmhaften Plosive und Frikative sowie der langen Vokale und sonstigen Vokale eine weitgehende Übereinstimmung mit den literarischen Werten von Kohler [17]. Diese Beobachtungen spiegeln die strukturellen Unterschiede zwischen Mehrsilbern und Sätzen in der deutschen Sprache und der vereinfachten Struktur einsilbiger Substantive wider. Laut Hahlbrock liegt es in der Natur der Sache, dass die Lautverteilung in Einsilbertests, insbesondere bei Vokalen wie dem Schwa-Laut ə, nicht die gesamte Bandbreite der deutschen Sprache abdecken kann, die überwiegend aus Mehrsilbern besteht. Eine möglichst naher Angleich kann als Ziel gesetzt werden, wenn die Repräsentativität des Test in Bezug auf die deutsche Sprache angestrebt wird. Um eine äquivalenterer phonemische Repräsentation zu erreichen, könnte zukünftig ein alternativer Vorschlag aus der Norm DIN EN ISO 8253-3:2022-11 [12] berücksichtigt werden. Dieser empfiehlt, zwischen den Phonemklassen eine mittelwertbasierte Äquivalenz aller Testlisten innerhalb einer Klasse anzustreben, ohne die Verteilung des deutschen Sprachgebrauchs zu berücksichtigen. Trotz der Diskrepanzen gegenüber der Statistik von Kohler bei den stimmlosen Plosiven und Frikativen, kurzen Vokalen, Nasalen und sonstigen Konsonanten stimmen die Phonemverteilungen der Phonemklassen beider Einsilbertests weitgehend überein (Abbildung 13). Diese hohe Übereinstimmung, mit einer mittleren Differenz von weniger als 0,5 Prozentpunkten, weist darauf hin, dass der originale und der aktualisierte Test auf einer sehr ähnlichen phonemischen Grundlage basieren und in nahezu gleichem Maße von Kohlers Verteilung abweichen.

### 5.3 Bewertung der Synthese des aktualisierten Einsilbertests

Für die Durchführung der Probandenstudie mussten nach der automatisierten TTS-Synthese über die Atemgruppe 1 insgesamt 95 Wörter wegen eindeutiger Aussprachefehler (Artefakte, Unverständlichkeit, falsche Betonung und Sprechgeschwindigkeit) und 26 Wörter wegen schlechter Bewertung bezüglich der Natürlichkeit neu synthetisiert werden. Die Synthese über SAMPA erwies sich im Vergleich zu verschiedenen Atemgruppenalternativen bei der automatisierten Korrektur nur bedingt als effektiv, stellt jedoch zukünftig eine mögliche Option dar, um einzelne Wörter unter Beachtung einer korrekten und einheitlichen SAMPA-Transkription zu synthetisieren. Dies wurde in dieser Arbeit nicht angewandt, da die BAS-Transkription nicht immer die hochdeutsche Variante verwendete, insbesondere bei der Unterscheidung zwischen einem konsonantischen und einem vokalischen r. Insgesamt mussten 22 % der synthetisch erzeugten Einsilber aufgrund von Unregelmäßigkeiten in der Aussprache bearbeitet werden. Die Durchschnittsbewertung aller 540 einsilbigen Substantive nach Bearbeitung entspricht aufgerundet der Schulnote 2 (Note 1,96). Trotz der guten Durchschnittsnote und ausgewogenen Verständlichkeitswerte wäre es wünschenswert, dass die TTS von Acapela noch natürlicher wird, um effizienter und objektiver, ohne manuelle Korrekturen eingesetzt werden zu können.

### 5.4 Psychometrische Funktionen der neuen Testlisten

Die psychometrischen Funktionen der 27 Listen, dargestellt in Abbildung 16, zeigten unterschiedliche SRT50-Werte. Besonders auffällig waren Testliste 4, die mit einem SRT50 von 29,4 dB den höchsten Schallpegel aufwies, und Testliste 8, die mit 25,9 dB den niedrigsten verzeichnete. Die Analyse dieser Funktionen deutete auf eine enge Übereinstimmung mit den Messungen von Schwarz et al. [10] hin. Der durchschnittliche SRT50 aller Testlisten von 27,9 dB und die gemittelte Steigung am SRT50 von 5,5 Prozent pro dB bestätigten, dass die neuen Listen eine vergleichbare Sprachverständlichkeitsschwelle erreichten, wie sie bereits in Tabelle 7 im Literaturvergleich aufgezeigt wurde. Dies legte nahe, dass trotz der individuellen Unterschiede zwischen den Testlisten, die Gesamtqualität des aktualisierten synthetischen FET mit den Erwartungen konsistent war.

Der Vergleich zwischen den psychometrischen Funktionen des FET mit synthetischem Testmaterial zeigte in bisherigen Untersuchungen bei dem SRT50 von Thiele et al. [45] und bei der Steigung von Schwarz et al. [10] keinen signifikanten Unterschied gegenüber dem neuen, aktualisierten und synthetischen Material. Die vorhandenen, wenn auch geringfügigen Abweichungen des SRT und der Steigung zu den originalen Bezugskurven und den SRT50-Ergebnissen von Schwarz et al. könnten auf mehrere Faktoren zurückzuführen sein. Zum Beispiel könnte die Synthesequalität die Ergebnisse beeinflusst haben, indem die künstlich erzeugte Sprache als weniger natürlich empfunden wurde. Dies könnte schlechtere Sprachverständlichkeitswerte zur Folge gehabt und Deckeneffekte sowie ein ähnliches Einzelwortverstehen begünstigt haben, was wiederum eine höhere Steigung verursachte. Die Verwendung bekannterer Wörter mit einem sehr hohen Sprachverstehen von 95 % bis 100 % wie z.B. Mensch,

Tisch, Stadt, führte hingegen zu einer Verbesserung des SRT50 und könnte somit diese Effekte wieder ausgeglichen haben. Individuelle Unterschiede zwischen den Listen könnten durch eine übermäßige Verwendung von Anglizismen oder schlecht bewerteten Wörtern entstanden sein, da diese bei

der Verteilung in der Listenerstellung nicht gleichmäßig berücksichtigt wurden. Auch wurde die Verständlichkeit von Anglizismen in den Testlisten qualitativ untersucht. Von insgesamt 24 Anglizismen wurden nur sechs besser als 50 % verstanden. Unter diesen zeichnete sich das Wort Couch mit einer Verständlichkeit von 90 % besonders positiv ab, gefolgt von Sound mit 70 %. Im Gegensatz dazu wurden 18 Wörter schlechter als 50 % verstanden, wobei ein großer Anteil sogar unter 30 % lag. Das Wort Bob, welches zusätzlich die schlechteste nicht verbesserte subjektive Bewertung in der Sprachsynthese erhielt, war mit einer Verständlichkeit von 22 % Teil dieser Gruppe. Das schlechteste Ergebnis unter den Anglizismen erzielte Beat mit lediglich 3 % Sprachverstehen im Mittel. Diese qualitative Betrachtung deutet darauf hin, dass Anglizismen tendenziell schlechter verstanden werden. Bei einer zufälligen Zusammenstellung von Listen, die Anglizismen enthalten, könnte die Verständlichkeit der Testlisten daher negativ beeinflusst werden. Jedoch ist diese Problematik aufgrund der angestrebten perzeptiven Optimierung ohne weiteres ausgleichbar.

Ein weiterer relevanter Aspekt war die Tatsache, dass die Bezugskurven des Original-FET auf monauralen Kopfhörmessungen beruhten, wohingegen die aktuellen Messungen sowie jene von Schwarz et al. binaural im Freifeld durchgeführt wurden. Diese methodische Diskrepanz wurde mit einem Korrekturwert von 2,5 dB berücksichtigt. Die limitierte Anzahl von Probanden in dieser Studie, lediglich 27 im Vergleich zu 97 in den Bezugskurvenmessungen [46], könnte zu einer geringeren Repräsentativität geführt und zusätzlich zu statistischen Abweichungen beigetragen haben. Dies beeinträchtigte möglicherweise die Genauigkeit der abgeleiteten psychometrischen Funktionen. Die insgesamt geringen Unterschiede des SRT50 zwischen den aktualisierten Testlisten und den vorangegangenen Untersuchungen reflektierten daher sowohl die natürlichen Schwankungen innerhalb einer solchen Testreihe als auch die potenziellen oben genannten Einflüsse. Zukünftig müssen umfangreichere Probandenstudien durchgeführt werden, um zu überprüfen, ob sich diese Ergebnisse bestätigen lassen und ob sie als Grundlage für die Aktualisierung der Bezugskurven geeignet sind.

## 5.5 Optimierung der phonemischen und perzeptiven Äquivalenz

Im Anschluss der Evaluation erfolgte die Optimierung der perzeptiven Äquivalenz. Diese zielt darauf ab, vergleichbare Ergebnisse beim Sprachverstehen über alle Testlisten hinweg zu erzielen. Dies stellt eine Basis dar, um sicherzustellen, dass der Schwierigkeitsgrad, die Einsilber zu verstehen, über die verschiedenen Testlisten konsistent bleibt, was für die Zuverlässigkeit, Vergleichbarkeit und Validität der Sprachaudiometrie entscheidend ist.

In Vorbereitung auf die Optimierung der perzeptiven Äquivalenz wurden das Einzelwortverstehen über alle 540 Wörter gemittelt (48,3 %) und die Differenzen der Listenverständlichkeit in Prozentpunkten berechnet, siehe Abbildung 18. Der größte SRT-Unterschied ergab sich in den Testlisten vier, gerundet -5 Prozentpunkte, und acht, gerundet +8 Prozentpunkte. Trotz der zufälligen Listenzusammenstellung unter Einhaltung des Hahlbrock-Schemas ergaben sich nur geringe Unterschiede in den Differenzwerten des mittleren Sprachverstehens zur Literatur, siehe Grundlagen Abbildung 1 und Abbildung 2. Der Shapiro-Wilk-Test ergab eine Normalverteilung bei den Differenzwerten ( $W = 0,95$ ,  $p = 0,18$ ). Auf dieser Grundlage wurde ein zweiseitiger t-Test durchgeführt. Die Ergebnisse des t-Tests zeigten

keinen signifikanten Unterschied zwischen den Mittelwerten der neuen Daten und den Ergebnissen aus Abbildung 1 ( $t(26) = -0,362$ ,  $p = 0,719$ ) sowie Abbildung 2 ( $t(26) = 0,428$ ,  $p = 0,671$ ). Die Befunde deuten darauf hin, dass die neuen Daten größtenteils mit den Differenzwerten der Literatur des ursprünglichen FET übereinstimmen und keine großen Abweichungen in der perceptiven Differenz aufweisen. Das bedeutet, dass die in der Studie genutzte perceptiv und phonemisch zufällige Zusammenstellung mindestens gleich gute Ergebnisse liefert, was das Potenzial einer noch folgenden Optimierung bestätigt. Eine perceptiv ausgewogene Liste über alle Listen ist jedoch über die zunächst zufällige Listenerstellung nur mit Berücksichtigung des Hahlbrockschemas wie erwartet nicht zu erreichen. Diese Ergebnisse zeigen, ähnlich wie beim ursprünglichen FET, eine Variabilität in der Verständlichkeit über die 27 Listen hinweg. Dies deutet auf eine fehlende perceptiv Äquivalenz und ein umfassendes Optimierungspotential hin. Bei der Analyse von 540 Wörtern fiel auf, dass das Wort Mensch durchgängig richtig verstanden wurde, während das Wort Pack von keinem der Teilnehmer erkannt wurde. Die geringe Anzahl von Ausreißern und die nahezu normalverteilte Streuung der Wortverständlichkeitswerte legen nahe, dass die synthetische Version des Tests mindestens vergleichbare Ergebnisse in der listen-spezifischen Sprachverständlichkeit liefert. Diese Beobachtung wird durch die Studie von Mallinger [47] unterstützt, die aufzeigt, dass im ursprünglichen FET mit 400 Wörtern, trotz der Aufnahme durch Claus Wunderlich, sieben Wörter nicht verstanden wurden. Ähnlich wie bei der Synthese gibt es auch Kritik an der menschlichen Sprachaufnahme von Claus Wunderlich, insbesondere bezüglich der unnatürlich lang gezogenen Vokale und unterschiedlich starken Betonungen einhergehend mit schwankender Lautstärke [48]. Bei der synthetischen Sprache ist bei Acapela in der Standardsynthese aufgefallen, dass der Hauptvokal zu kurz und das Wortende zu lang betont wurde. Größtenteils konnte dies durch die Alternativsynthese 1 behoben werden.

Wie in Abbildung 21 dargestellt, zeigt sich nach der Optimierung der perceptiven Äquivalenz, dass die durchschnittliche Differenz zum Mittelwert des Sprachverstehens über alle Listen auf nur noch 0,065 Prozentpunkte reduziert werden konnte, was einer fast 100 prozentigen Verbesserung gegenüber den vorherigen Abweichungen entspricht. Diese Ergebnisse verdeutlichen, dass es trotz der Herausforderungen, eine natürliche und neutrale Aussprache zu realisieren, möglich ist, eine hohe perceptiv Äquivalenz zu erreichen, und dass innerhalb der 540 Einsilber mit verhältnismäßig geringem Aufwand durch den Tausch einzelner Wörter. Die von Baljić et al. [14] durchgeführte Studie unterstreicht zudem, dass die durchschnittliche Abweichung der Verständlichkeitswerte aller Listen des originalen FET bei 4,5 Prozentpunkten liegt, während die Abweichung bei der aktuellen zufälligen Testlistenzusammenstellung nur 2,8 Prozentpunkte beträgt. Dies zeigt eine geringere Variabilität in der aktuellen Zusammenstellung im Vergleich zum ursprünglichen FET. Diese Befunde bestätigen das erhebliche Potenzial für weitere Optimierungen und Verbesserungen in der Entwicklung von Sprachtests.

Abbildung 23 zeigt die Phonemverteilung der verschiedenen Phonemklassen über die 27 Testlisten. Es ist zu sehen, dass allein die perceptiv Äquivalenz zu optimieren für einen aktualisierten ausgewogeneren Sprachtest nicht reicht und auch wie zu erwarten keine gleichmäßige Verteilung innerhalb der Phonemklassen gewährleistet. Auffällig ist die hohe Anzahl an kurzen Vokalen, stimmhaften Frikativen und sonstigen Konsonanten in den Listen mit den höchsten SRT50-Werten. Dies könnte darauf hindeuten, dass diese Phonemgruppen besonders schwierig zu verstehen sind, was zu einer erhöh-

ten Höranstrengung und schlechteren SRT50-Werten bei ungleicher Verteilung innerhalb der Listen führt. Untersuchungen zum Freiburger Sprachverständnistest haben gezeigt, dass die phonemische Verteilung einen erheblichen Einfluss auf die Verständlichkeit hat. Listen mit mehr stimmhaften Frikativen und kurzen Vokalen haben ein tendenziell schlechteres Sprachverstehen [49]. Zukünftig sollte bei der Listenerstellung versucht werden, neben der perzeptiven Äquivalenz gleichzeitig die phonemische Äquivalenz miteinzubeziehen. Eine perfekte phonemische Äquivalenz, die sich an der deutschen Sprachstatistik orientiert, ist aufgrund der spezifischen Vorgaben zur Phonemanzahl in den Listen, wie sie von Hahlbrock mit jeweils 73 Phonemen festgelegt wurden, sowie der durch die Einsilber begrenzten Wortauswahl, kaum erreichbar.

Trotz dieser Einschränkungen stellt die hoch signifikante Verringerung der Listendifferenzen in der Verständlichkeit gegenüber dem Mittelwert bei der perzeptiven Äquivalenz einen ersten wesentlichen Teilschritt in der Optimierung dar. Es deutet sich an, dass aufgrund der Einschränkung im Wortmaterial, die perzeptiv Äquivalenz leichter zu erreichen ist als eine phonemische Äquivalenz bezogen auf Literaturwerte, siehe Abschnitt 4. Eine perzeptiv Optimierung verbessert nicht die phonemische Äquivalenz, siehe Abbildung 23. Die Anzahl der Phoneme variiert innerhalb der neun Phonemklassen bei jeder Liste im Vergleich zum durchschnittlichen Mittelwert jeder Phonemklasse. Entscheidend ist dabei, dass eine gleichmäßige Ausgewogenheit über alle Listen hinweg angestrebt wird, um Konsistenz zu gewährleisten. Diese gleichmäßige Verteilung könnte später als zusätzliches Optimierungskriterium herangezogen werden, insbesondere im Vergleich zur phonemischen Ausgewogenheit gemäß der deutschen Sprachstatistik. Da eine vollständige Äquivalenz zur deutschen Sprachstatistik nicht vollständig erreichbar ist, sollte eine Äquivalenz zwischen den Listen bezogen auf den Mittelwert aller Listen der neun Phonemgruppen ergänzend oder als Alternative zur Statistik angestrebt werden. Weiterhin ist zu bedenken, dass die ausschließliche Fokussierung auf prozentuale Abweichungen vom Mittelwert des Sprachverstehens eine vereinfachte Analyse ist. Die Interaktionen der Einsilber durch die Randomisierung innerhalb der Testlisten zwischen verschiedenen phonemischen Elementen, Nachbarschaftsdichte, Reihenfolge und deren Einfluss auf die Verständlichkeit sind in dieser Analyse nicht abgebildet. Winkler et al. [13] untersuchten mit einer Studie neben dem Einfluss der Wortfrequenz auf die mit dem FET gemessene Sprachverständlichkeit auch die Nachbarschaftsdichte, die eine lexikalische Ähnlichkeit zu anderen Wörtern beschreibt. Es stellte sich heraus, dass beide Parameter und somit auch die Auswahl der Testlisten Einfluss auf die Ergebnisse des FET hatten [5]. Eine Auswertung der Wortfrequenz in Verbindung mit der Tokenanzahl und dem Sprachverstehen zeigte auch in dieser Studie eine schwache aber hoch signifikante Korrelation ( $r = 0,14$ ,  $p < 0,001$ ). Je bekannter die Wörter, desto besser wurden sie teilweise verstanden, siehe Anhang Abbildung 25. Auch darf die Berücksichtigung der Listendifferenzen zum Mittelwert bei Messungen in Ruhe nicht mit Ergebnissen unter Störgeräuschbedingungen gleichgesetzt werden. Zukünftig sollten für die gegenwärtig zusammengestellten Listen auch Messungen im Störschall durchgeführt werden. Dies ist für die Optimierung der perzeptiven Äquivalenz von hoher Relevanz, da der Freiburger Einsilbertest (FET) gemäß den aktuellen Hilfsmittelrichtlinien sowohl in Ruhe als auch im Störgeräusch verwendet wird. Nur durch solche Messungen kann die Eignung der Listen für alle Hörbedingungen in der Hörsystemversorgung sichergestellt werden.

Im Zuge der Datenanalyse wurde festgestellt, dass die Randomisierung der Einsilber innerhalb der

Testlisten unvollständig war, was zur Folge hatte, dass die Listen überwiegend mit Wörtern starteten, die aus nur zwei Phonemen bestehen, siehe Anhang Abbildung 27 bis Abbildung 33. Dies könnte die Verständlichkeit in dieser spezifischen Phonemanzahlgruppe beeinträchtigen und unerwünschte Reihenfolgeeffekte nach sich ziehen. Eine Korrelationsanalyse mit den aktualisierten Listen ergab eine signifikante aber sehr schwache Korrelation ( $r = 0,09$ ,  $p = 0,04$ ), dass mit sinkender Phonemanzahl das Sprachverstehen abnimmt, siehe Anhang Abbildung 26. Laut Hahlbrock verbessert sich das Wortverständnis bei steigender Phonemanzahl grundsätzlich [9]. Dieser Zusammenhang ist bekannt und kann nicht direkt der unvollständigen Randomisierung zugewiesen werden. Obwohl die große Übereinstimmung des SRT50, der Steigung und der geringen aber signifikanten Korrelation zwischen Sprachverstehen und Phonemanzahl darauf hindeutet, dass der Einfluss solcher Verzerrungen begrenzt sind, ist ein Effekt nicht vollständig auszuschließen.

Die Komplexität der Sprache und die begrenzte Auswahl an einsilbigen Substantiven stellen eine Herausforderung für die Erstellung von Testlisten dar, die sowohl phonemisch als auch perceptiv äquivalent sind. Die perzeptive Äquivalenz in Ruhe und im Störschall sowie die phonemische Äquivalenz, basierend auf Literaturwerten oder durchschnittlichen Phonemklassenwerten, sind die entscheidende Kriterien, die für die weitere Optimierung berücksichtigt und entsprechend gewichtet werden müssen. Lösungen für eine optimale und aktualisierte Listenzusammenstellung auf Basis des FET können beispielsweise durch einen mehrdimensionalen Optimierungsansatz, iterative Verfahren und den Einsatz von Expertenwissen erreicht werden. Dieser komplexe Ansatz kann eine umfassendere Berücksichtigung verschiedener Aspekte der Sprachverarbeitung ermöglichen und verbessert die Chancen, Testlisten zu entwickeln, die den realen Anforderungen des Sprachverstehens in unterschiedlichen Prüfsituationen gerecht werden und vergleichbar sind.

## 6 Fazit und Ausblick

Das in dieser Arbeit beschriebene Verfahren dient als Vorbereitung für eine phonemisch und perzeptiv äquivalente Aktualisierung des Freiburger Sprachmaterials. Durch das Erstellen einer neuen Sammlung von häufig gebrauchten einsilbigen Substantiven im Deutschen mit einer automatischen Wortfrequenzfilterung wurden zeitgemäß aktualisierte Testlisten erstellt. Zunächst konnten unter Vernachlässigung der phonemischen und perzeptiven Äquivalenz mit dem beschriebenen Verfahren zur systematischen Erstellung von neuen FET Listen, Sprachverständlichkeitsmessungen und psychometrische Funktionen erstellt werden. Kombiniert mit dem Hahlbrock-Listenschema ergaben sich maximal 27 neue Listen, die allen grundlegenden Anforderungen des originalen FETs entsprachen. Für die Untersuchung der perzeptiven Äquivalenz und der psychometrischen Funktionen der einzelnen Listen wurde in dieser Arbeit eine Studie mit 27 normalhörenden Teilnehmenden durchgeführt. Die neu entwickelten Testlisten zeigen einen ähnlichen SRT50 und keine signifikanten Unterschiede in der Steigung, wie in der Studie von Schwarz et al. [10] dokumentiert. Weiterhin zeigt die Studie innerhalb der vier von Hahlbrock definierten Phonemanzahlgruppen, dass die Verteilung des Einzelwortverstehens ein hohes Optimierungspotenzial für die perzeptive Äquivalenz aufweist. Die Möglichkeiten zur Optimierung der phonemischen Äquivalenz sind jedoch aufgrund der Eigenschaften der deutschen Sprache und der unterschiedlichen phonemischen Verteilung zwischen ein- und mehrsilbigen Wörtern begrenzt. Dies zeigt die damit einhergehende Schwierigkeit auf, alle typisch deutschen Phonemanteile im richtigen Verhältnis in standardisierte Sprachtests zu integrieren. Aufgrund dieser Beschränkung besteht eine Möglichkeit darin, die phonemische Äquivalenz der Listen untereinander durch die Optimierung anhand des Mittelwertes ihrer einzelnen Phonemklassen zu verbessern, anstatt sich auf Literaturwerte zu beziehen. Nach Empfehlung der Norm DIN EN 8253-3:2022-11 bedeutet das konkret, wenn eine vollständige phonemische Gleichwertigkeit zwischen Testlisten nicht erreichbar ist, sollen die Testlisten zumindest innerhalb spezifischer Phonemklassen gleichwertig gestaltet werden. Solche Phonemklassen umfassen beispielsweise stimmhafte und stimmlose Plosive sowie Frikative, Nasallaute, sowie lange und kurze Vokale und sonstige Vokale und Konsonanten. Dies stellt sicher, dass die Testlisten in Bezug auf die phonemische Verteilung der Sprache ausgewogen sind und relevante phonetische Merkmalsgruppen angemessen repräsentieren. Es ist noch nicht abschließend geklärt, wie viele sinnvolle Listen unter Berücksichtigung der oben genannten Norm [12] erstellt werden können, um eine phonemisch und perzeptiv optimal ausgewogene Kombination der neu selektierten Einsilber zu erhalten. Um praktisch anwendbare neue Testlisten mit synthetischem Sprachmaterial für einen aktualisierten FET sowie entsprechende neue Bezugskurven zu erstellen, sind noch einige abschließende Schritte erforderlich. Insbesondere bedarf es weiterer Forschungsarbeiten und einer umfangreicheren Probandenstudie in Ruhe und im Störgeräusch. Die kontinuierliche Anpassung des FET an linguistische Veränderungen sollte durch automatisierte Verfahren in Zukunft angestrebt werden. Die Forschungsergebnisse unterstützen die Annahme, dass durch eine mehrdimensionale Optimierung ein besseres Gleichgewicht zwischen phonemischer und perzeptiver Äquivalenz erreicht werden kann, was die Entwicklung eines aktualisierten Sprachmaterials für den FET in absehbarer Zeit ermöglichen wird.

## Literatur

- [1] German Institute for Standardization e.V.: *DIN 45621-1:1995-08, Sprache für Gehörprüfung - Teil 1: Ein- und mehrsilbige Wörter*, 1995.
- [2] L. Jäger, I. Holube, A. Winkler, F. Denk, T. Sankowsky-Rothe, M. Blau und H. Husstedt: *Verwendung von Sprachtests im Freifeld in Deutschland*. Deutsches Hörgeräte Institut GmbH, Lübeck, 2024.
- [3] S. Hoth: *Der Freiburger Sprachtest. Eine Säule der Sprachaudiometrie*. HNO, 64:540–548, 2016. Online publiziert: 3. Juni 2016.
- [4] T. Steffens: *Verwendungshäufigkeit der Freiburger Einsilber in der Gegenwartssprache*. HNO, 64:549–556, 2016.
- [5] A. Winkler, I. Holube und R. Carroll: *Impact of Lexical Parameters and Audibility on the Recognition of the Freiburg Monosyllabic Speech Test*. Ear Hear, 2019.
- [6] F. Hahn: *Freiburger reloaded*. Bachelorarbeit, Hochschule Aalen, 2014.
- [7] M. Mahfoud: *Neuaufsprache und Evaluation des Einsilber-Sprachverständnistests*. Dissertation, Julius-Maximilians-Universität Würzburg, 2009.
- [8] J. F. Qualen: *Evaluation des Einsilber-Sprachmaterials M-2007 und Entwurf einer Methodik für die Zusammenstellung gleichwertiger Listen*. Dissertation, Julius-Maximilians-Universität Würzburg, 2010.
- [9] K. H. Hahlbrock: *Sprachaudiometrie*. Thieme Verlag, Stuttgart, 1970.
- [10] T. Schwarz und M. Frenz und A. Bockelmann und H. Husstedt: *Examination of a synthetic voice for the Freiburg Monosyllabic Speech*. GMS Zeitschrift für Audiologie - Audiological Acoustics, 4:321–334, 2022.
- [11] T. Schwarz: *Entwicklung eines systematischen Prozesses für die Erstellung neuer Listen für den Freiburger Einsilbertest*. Praktikumsbericht, Universität zu Lübeck, 2022.
- [12] International Organization for Standardization. ISO 8253-3:2022-11. Akustik - Audiometrische Prüfverfahren - Teil 3: Sprachaudiometrie (ISO 8253-3:2022); Deutsche Fassung EN ISO 8253-3:2022, November 2022. Ausgabe 2022-11.
- [13] A. Winkler, I. Holube und H. Husstedt: *Der Freiburger Einsilbertest im Störgeräusch*. HNO, 68:14–24, 2020.
- [14] I. Baljić, A. Winkler, T. Schmidt und I. Holube: *Untersuchungen zur perceptiven Äquivalenz der Testlisten im Freiburger Einsilbertest*. HNO, 64:572–583, 2016. Publiziert: 14. Juli 2016.
- [15] T. Becker: *Einführung in die Phonetik und Phonologie des Deutschen*. Phonologie, 1998.

- [16] M. Exter, A. Winkler und I. Holube: *Phonemische Ausgewogenheit des Freiburger Einsilbertests*. HNO, 64(8):557–563, 2016.
- [17] K. J. Kohler: *Einführung in die Phonetik des Deutschen*, Band 2, Seiten 220–224. E. Schmidt, Berlin, 1995.
- [18] Dudenredaktion: *Über Duden: Partner*. [https://www.duden.de/ueber\\_duden/Partner](https://www.duden.de/ueber_duden/Partner), 2023. Zugriff am: 18. April 2023.
- [19] IDS: *FOLK*. <http://dgd.ids-mannheim.de>, 2022. Zugriff am: 14. November 2022.
- [20] *Leipzig Corpora Collection*. <https://corpora.uni-leipzig.de?corpusId=deunewscrawl2011,2011>. Zugriff am: 20. Januar 2023.
- [21] Leibniz-Institut für Deutsche Sprache: *DeReWo - Korpusbasierte Grund-/Wortformenlisten*. <https://www.ids-mannheim.de/digspra/kl/projekte/methoden/derewo/>. Zugriff am: 20. Januar 2023.
- [22] DWDS: *Digitales Wörterbuch der deutschen Sprache*. <http://www.dwds.de>, 2022. Zugriff am: 21. November 2022.
- [23] *Lexical databases of German 2.0*. <https://catalog.ldc.upenn.edu/topten>. Zugriff am: 20. Januar 2023.
- [24] Wikipedia: *Wikipedia. Deutsche Einsilber*. <https://www.wiktionary.org/wiki/Category:German1syllablewords>. Zugriff am: 10. Januar 2023.
- [25] Learnattack: *Diphthong*. <https://learnattack.de/schuelerlexikon/latein/diphthong>, 2018. Zugriff am: 20. März 2018.
- [26] H. H. Wängler: *Grundriß einer Phonetik des Deutschen. Mit einer allgemeinen Einführung in die Phonetik*, Seite 90. Helmut Buske Verlag, 1983, ISBN 9783770807536.
- [27] Cornelsen Verlag: *Aussprache – Der R-Laut: Konsonantisches und vokalisches R*. <https://de.scribd.com/document/684676862/lektion13-lehrer-aussprache>, 2006. Zugriff am: 21.02.2023.
- [28] *Deutsch - Lautschrift*. <https://de.wiktionary.org/wiki/Wiktionary>, Deutsche Lautschrift, 2024. Zugriff am: 22.03.2024.
- [29] B. Pfister und T. Kaufmann: *Sprachverarbeitung: Grundlagen und Methoden der Sprachsynthese und Spracherkennung*. Springer Berlin Springer Vieweg, Berlin, Germany, 2017.
- [30] P. Taylor: *Text-to-speech synthesis*. Cambridge University Press, Cambridge, UK, 2009.
- [31] H. Schulz und S. Behnke: *Deep Learning: Layer-Wise Learning of Feature Hierarchies*. KI - Künstliche Intelligenz, 26(4):357–363, 2012, ISSN 0933-1875.

- 
- [32] Acapela Group: *Acapela DNN Technology*. <https://www.acapela-group.com/voices/acapela-dnn-technology/>, 2023. Zugriff am: 22.03.2024.
- [33] S. Ibelings, T. Brand und I. Holube: *Göttinger Satztest mit synthetischer Sprache*. 24. Jahrestagung der Deutschen Gesellschaft für Audiologie, 2022.
- [34] Acapela Group: *Acapela Cloud API Documentation*. [https://www.acapela-cloud.com/docs\\_tags/](https://www.acapela-cloud.com/docs_tags/), 2023. Zugriff am 21.03.2024.
- [35] Wiktionary: *Affrikate*. <https://de.wiktionary.org/wiki/Affrikate>, 2023. Zugriff am: 07.03.2023.
- [36] A. Busch und O. Stenschke: *Germanistische Linguistik. Eine Einführung*, Kapitel 3, Seite 45. Narr Francke Attempto Verlag, 2007.
- [37] U. D. Reichel: *PermA and Balloon: Tools for string alignment and text processing*. In: *Proc. Interspeech*, Seite 4, Portland, Oregon, 2012.
- [38] R. Wiese: *The Phonology of German*, Seiten 58–61. Oxford University Press, Oxford, 2000.
- [39] J. Karl: *Investigation of the influence of pitch and speed of synthetic speech in intelligibility of the Freiburg monosyllabic speech test*. Masterarbeit, Technische Hochschule Lübeck, 2021.
- [40] I. Holube und H. Husstedt: *Original Recording of Freiburg Words for Testing Hearing with Speech*, (*Hahlbrock, 1953*). <https://zenodo.org/records/10082491>. Zugriff am: 4. Januar 2024.
- [41] International Electrotechnical Commission, Geneva, Switzerland: *IEC 61672-1, Electroacoustics - Sound level meters - Part 1: Specifications*, 2. Auflage, 2013.
- [42] C. Guenther: *Prosodie und Sprachproduktion*. Niemeyer, Tübingen, 1999.
- [43] German Institute for Standardization e.V.: *DIN 8253-2:2010-07, Akustik - Audiometrische Prüfverfahren - Teil 2: Schallfeld-Audiometrie mit reinen Tönen und schmalbandigen Prüfsignalen*, 2010.
- [44] *DIN 45626-1: Tonträger mit Sprache für Gehörprüfung - Teil 1: Tonträger mit Wörtern nach DIN 45621-1 (Aufnahme 1969)*. Deutsches Institut für Normung, August 1995.
- [45] C. Thiele, N. Wardenga, T. Lenarz und A. Büchner: *Überprüfung der Vergleichbarkeit von Freifeld- und Freiburger Sprachtest*. *HNO*, 62:115–120, 2014. Online publiziert: 14. Februar 2014.
- [46] K. Brinkmann und U. Richter: *Ensuring reliability and comparability of speech audiometry in Germany*. In: M. Martin (Herausgeber): *Speech Audiometry*, Seiten 106–130. Whurr Publishers Ltd, Chichester, 2. Auflage, 1997.
- [47] E. Mallinger: *Trainingseffekte und Listenäquivalenz des Freiburger Einsilbertests im Störschall*. Inaugural-Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2011.

- 
- [48] J. Kiessling: *Moderne Verfahren der Sprachaudiometrie*. *Laryngo-Rhino-Otologie*, 79:633–635, 2000.
- [49] H. v. Wedel: *Untersuchungen zum Freiburger Sprachtest - Vergleichbarkeit der Gruppen im Hinblick auf Diagnose und Rehabilitation*. Unbekannt, Deutschland, 1986. Phonemische Ausgewogenheit des Freiburger Sprachverständnistests und dessen Einfluss auf die Verständlichkeit.
- [50] International Phonetic Association: *IPA-Kiel-Tabelle*, 2020. [https://www.internationalphoneticassociation.org/IPAcharts/IPA\\_chart\\_trans/pdfs/IPA\\_Kiel\\_2020\\_full\\_deu.pdf](https://www.internationalphoneticassociation.org/IPAcharts/IPA_chart_trans/pdfs/IPA_Kiel_2020_full_deu.pdf), Zugriff am 02. Mai 2024.

## 7 Anhang

### 7.1 Psychometrische Funktionen aller Listen aus der Pilotstudie

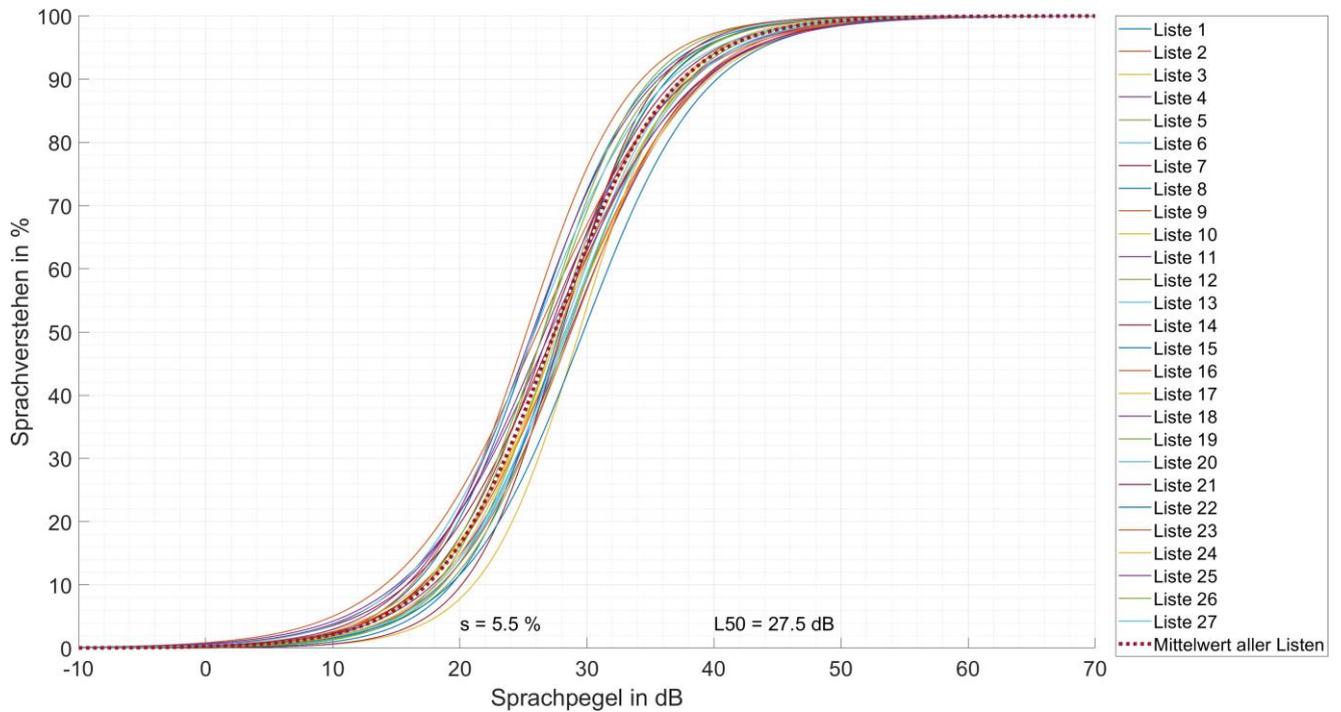


Abbildung 24: Psychometrische Funktionen der 27 Listen (Pegel 28,5 dB und 35,5 dB) mit acht Probanden, Pilotstudienresultate für Pegelermittlung 20 % (21,5 dB), 50 % (27,5 dB), 80 % (33,5 dB).

### 7.2 Korrelation des Einzelwortverstehens mit der Tokenanzahl und der Phonemanzahl

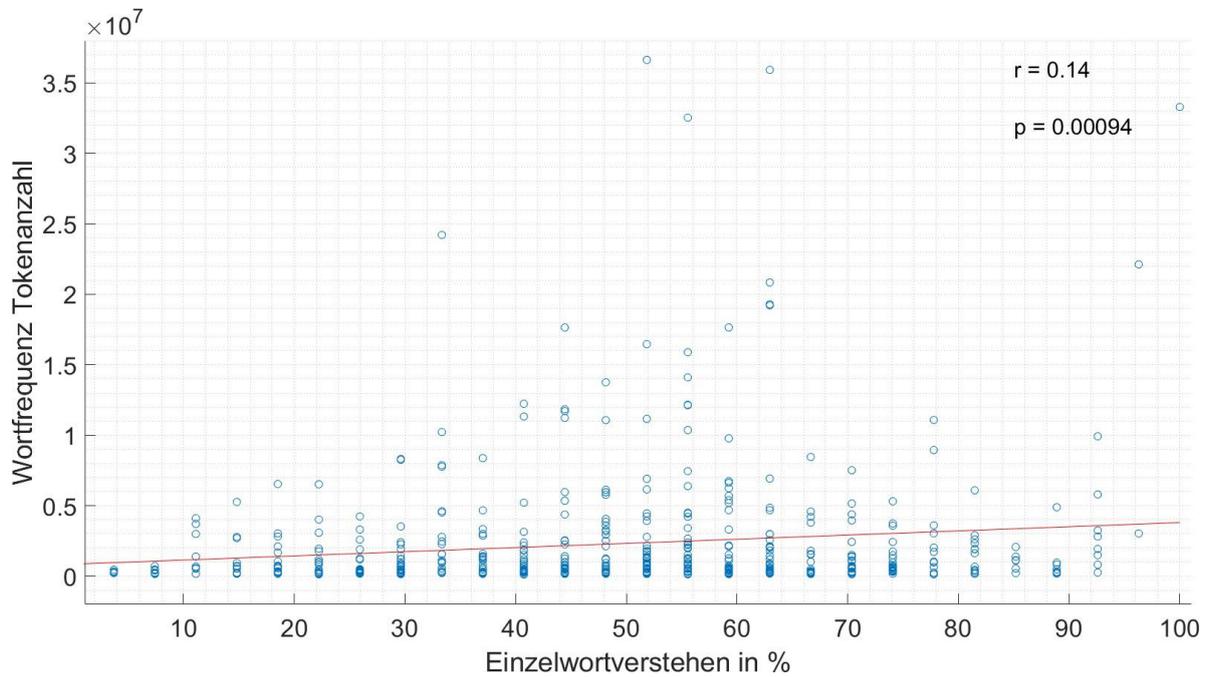


Abbildung 25: Korrelation zwischen Tokenanzahl und Einzelwortverstehen, Korrelationskoeffizient  $r = 0,14$  zeigt eine sehr schwache hoch signifikante positive Korrelation ( $p < 0,001$ ).

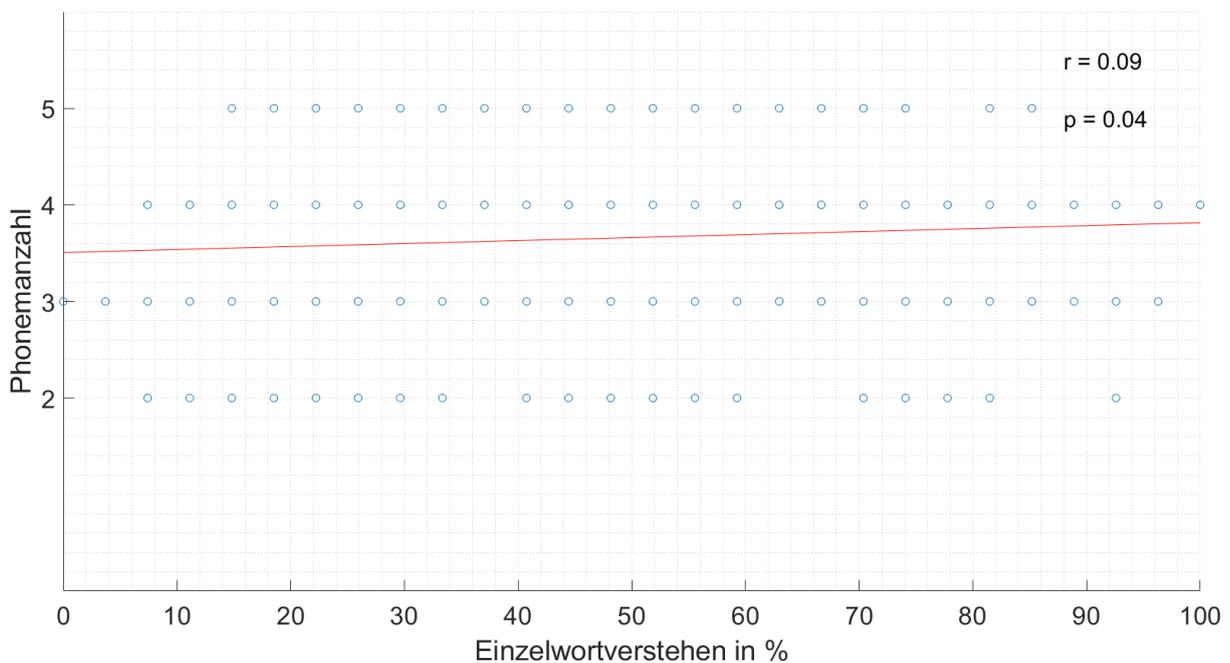


Abbildung 26: Korrelation zwischen Phonemanzahl und Einzelwortverstehen, Korrelationskoeffizient  $r = 0,09$  zeigt eine sehr schwache signifikante positive Korrelation ( $p = 0,04$ ).

### 7.3 Wortfrequenzfilterung der seltenen Einsilber im FET

Tabelle 8: Auflistung der 80 veralteten Einsilber des FET der Wortfrequenz 1 und 2, sortiert nach Tokenmenge.

Teil 1			Teil 2		
Token	Wortfrequenz	Wort	Token	Wortfrequenz	Wort
7320	1	Lump	86298	2	Schlitz
8402	1	Grog	90160	2	Ochs
14636	1	Biss	92857	2	Pult
16067	2	Fass	93048	2	Lehm
16872	2	Zank	97794	2	Reif
20187	2	Dung	100132	2	Fracht
20481	2	Molch	103583	2	Kinn
21467	2	Fraß	103619	2	Saum
22023	2	Pflock	105626	2	Pelz
22841	2	Docht	105843	2	Hohn
29109	2	Spind	105847	2	Axt
30965	2	Teer	109827	2	Schmied
31130	2	Grieß	110006	2	Kalk
33169	2	Huf	110089	2	Hanf
39007	2	Torf	110302	2	Stoß
39367	2	Dolch	112284	2	Pfand
41859	2	Napf	113374	2	Zopf
43464	2	Aas	115215	2	Schall
48085	2	Wuchs	115699	2	Trab
50023	2	Schilf	116200	2	Pest
50990	2	Laus	118135	2	Vieh
51933	2	Schmalz	120294	2	Blei
54698	2	Schleim	131432	2	Heu
55097	2	Narr	132047	2	Schreck
56417	2	Schopf	132289	2	Spott
63271	2	Roß	132786	2	Sieb
64551	2	Kork	134668	2	Kamm
64891	2	Pfau	137034	2	Klee
65993	2	Tau	137043	2	Stiel
66194	2	Glut	139572	2	Krach
66202	2	Erz	139811	2	Kies
67178	2	Floß	142781	2	Knecht
68318	2	Groll	146327	2	Keil
69138	2	Pfahl	147039	2	Reim
69312	2	Spalt	147881	2	Speer
69908	2	Floh	152160	2	Rast
78834	2	Dachs	156000	2	Stroh
80806	2	Gips	156036	2	Pfiff
85412	2	Leim			

### 7.4 Übersicht zur Wortverständlichkeit der aktualisierten Testlisten

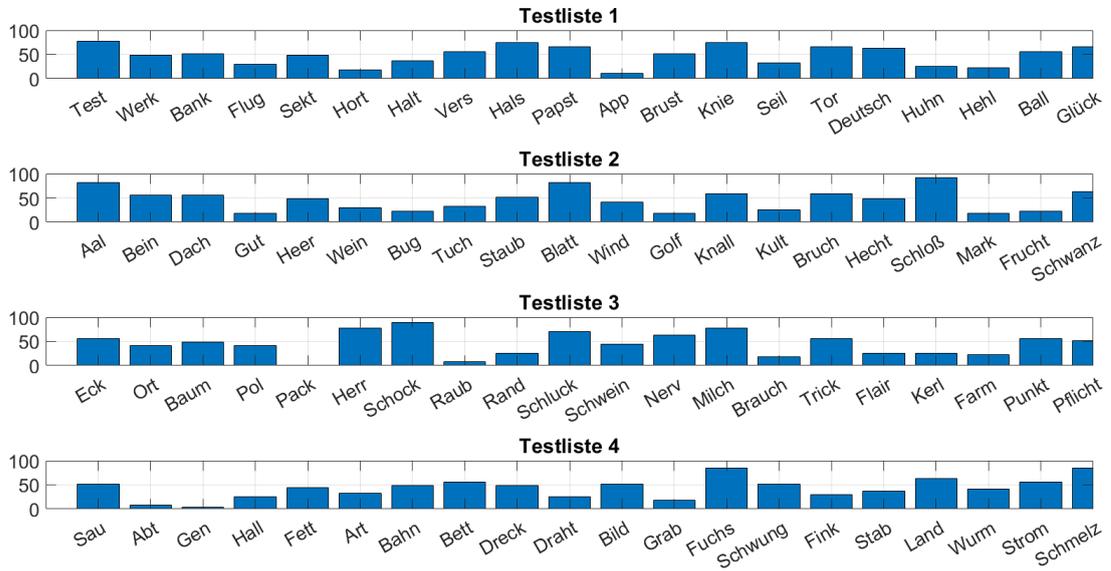


Abbildung 27: Balkendiagramm zum Einzelwortverstehen mit Wortnamen in Prozent sortiert nach Testlisten, Liste 1 bis 4.

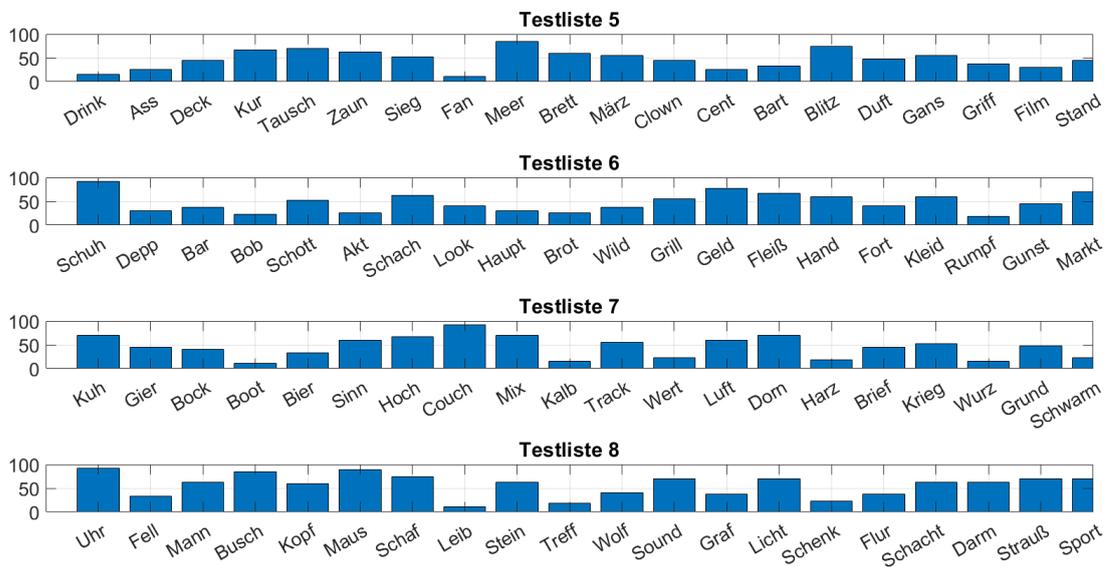


Abbildung 28: Balkendiagramm zum Einzelwortverstehen mit Wortnamen in Prozent sortiert nach Testlisten, Liste 5 bis 8.

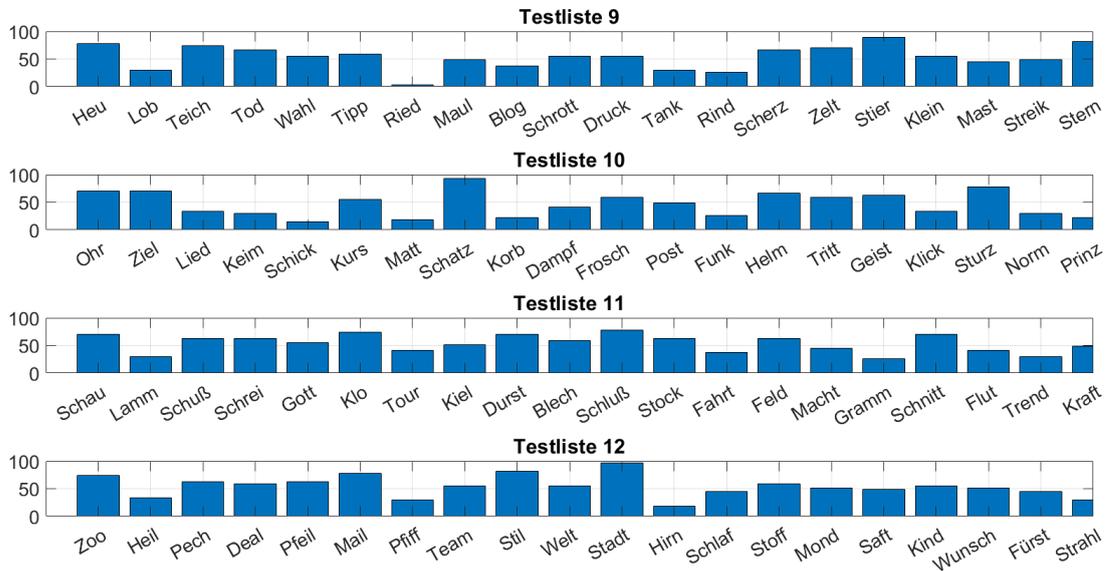


Abbildung 29: Balkendiagramm zum Einzelwortverstehen mit Wortnamen in Prozent sortiert nach Testlisten, Liste 9 bis 12.

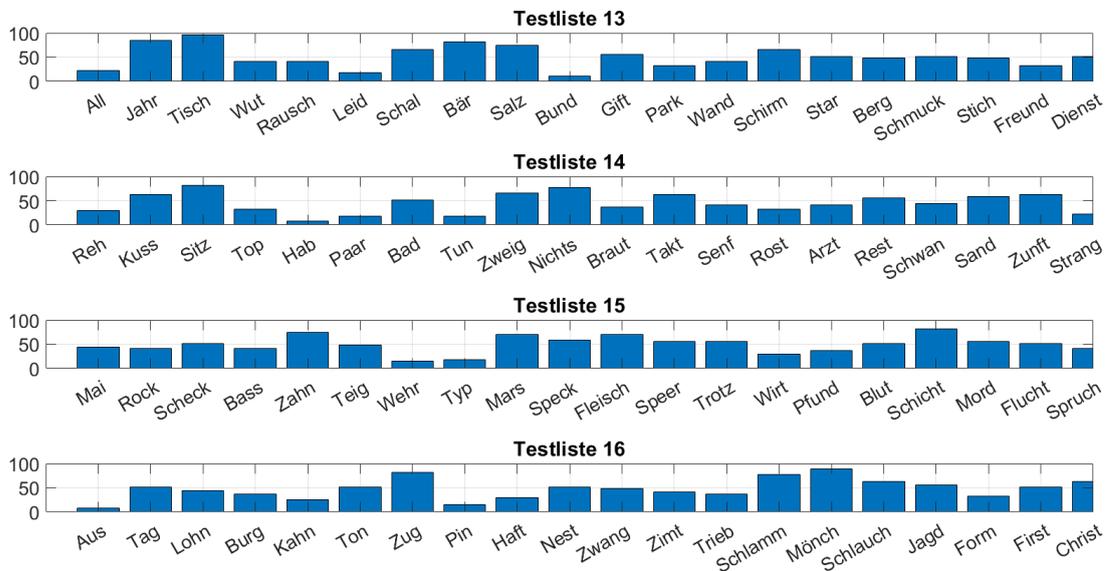


Abbildung 30: Balkendiagramm zum Einzelwortverstehen mit Wortnamen in Prozent sortiert nach Testlisten, Liste 13 bis 16.

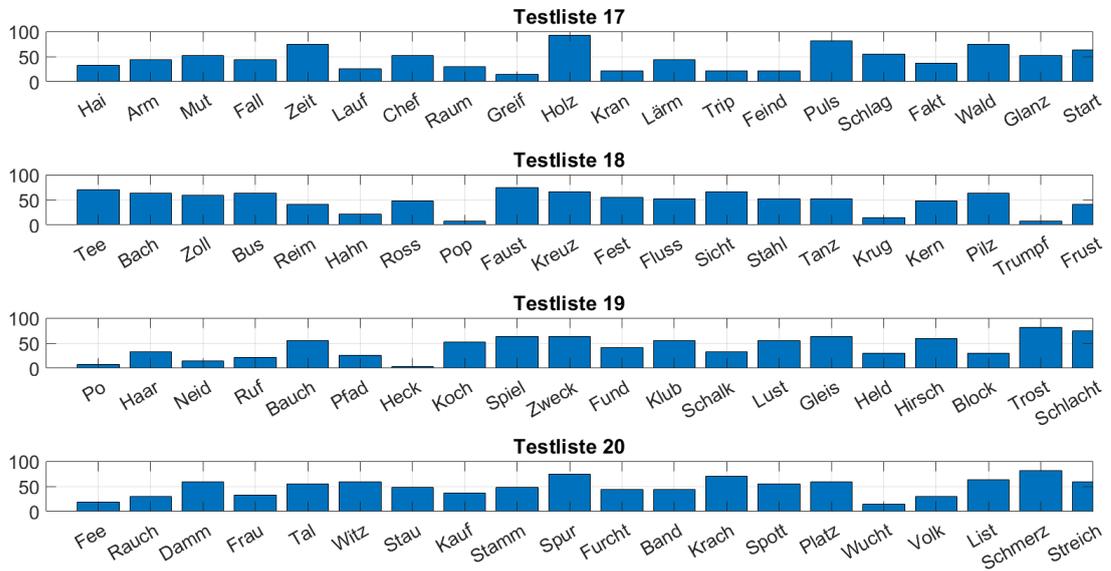


Abbildung 31: Balkendiagramm zum Einzelwortverstehen mit Wortnamen in Prozent sortiert nach Testlisten, Liste 17 bis 20.

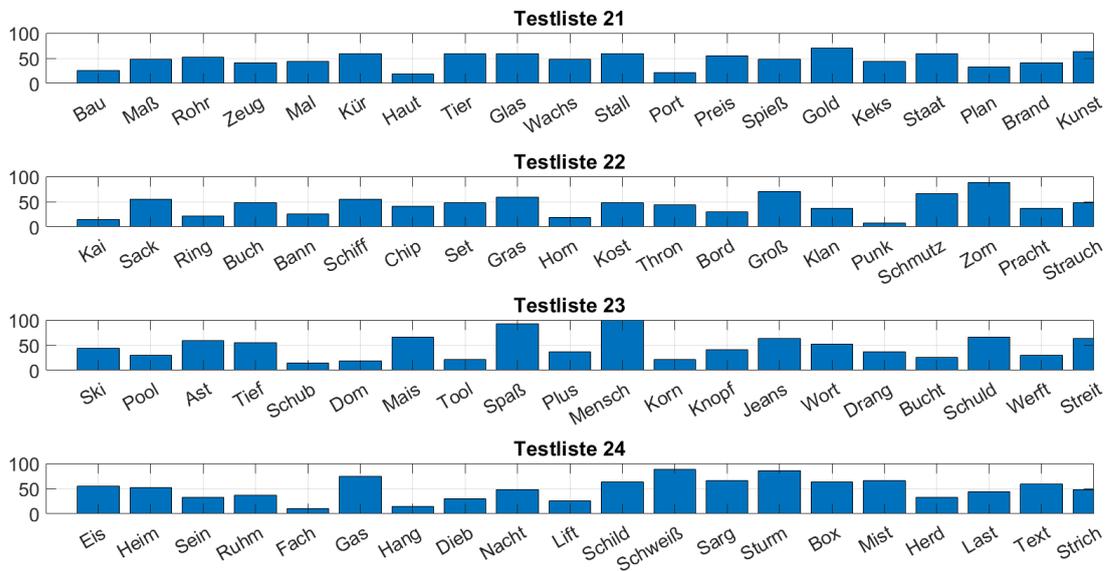


Abbildung 32: Balkendiagramm zum Einzelwortverstehen mit Wortnamen in Prozent sortiert nach Testlisten, Liste 21 bis 24.

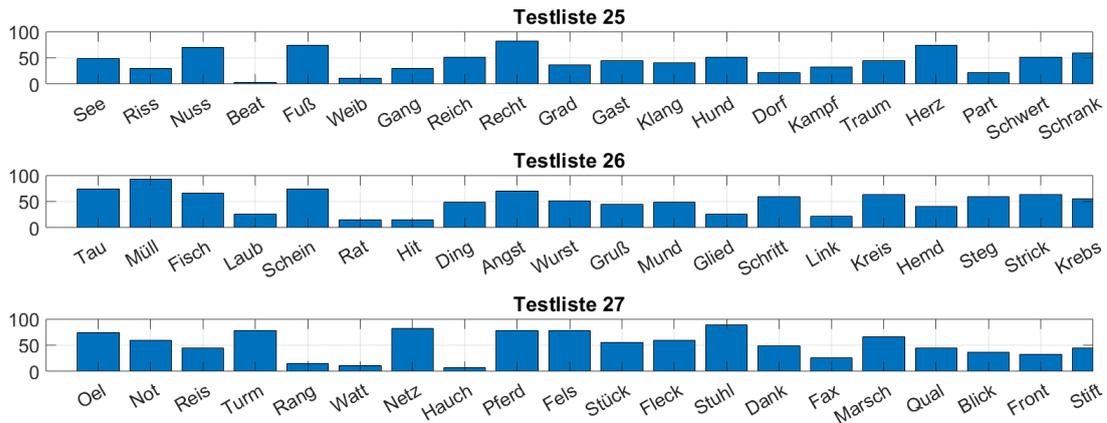


Abbildung 33: Balkendiagramm zum Einzelwortverstehen mit Wortnamen in Prozent sortiert nach Testlisten, Liste 25 bis 27.

**7.5 Auflistung der erstellten Testlisten**

Tabelle 9: 1. Version der Testlisten: Liste 1 bis Liste 9

Liste 1	Liste 2	Liste 3	Liste 4	Liste 5	Liste 6	Liste 7	Liste 8	Liste 9
'Test'	'Aal'	'Eck'	'Sau'	'Drink'	'Schuh'	'Kuh'	'Uhr'	'Heu'
'Bank'	'Bein'	'Baum'	'Abt'	'Ass'	'Akt'	'Bier'	'Busch'	'Lob'
'Flug'	'Bug'	'Herr'	'Art'	'Deck'	'Bar'	'Bock'	'Fell'	'Maul'
'Halt'	'Dach'	'Ort'	'Bahn'	'Fan'	'Bob'	'Boot'	'Kopf'	'Ried'
'Hort'	'Gut'	'Pack'	'Bett'	'Kur'	'Depp'	'Couch'	'Leib'	'Teich'
'Sekt'	'Heer'	'Pol'	'Fett'	'Sieg'	'Look'	'Gier'	'Mann'	'Tipp'
'Vers'	'Tuch'	'Raub'	'Gen'	'Tausch'	'Schach'	'Hoch'	'Maus'	'Tod'
'Werk'	'Wein'	'Schock'	'Hall'	'Zaun'	'Schott'	'Sinn'	'Schaf'	'Wahl'
'App'	'Blatt'	'Brauch'	'Bild'	'Bart'	'Brot'	'Brief'	'Darm'	'Blog'
'Brust'	'Bruch'	'Farm'	'Draht'	'Blitz'	'Fleiß'	'Dorn'	'Flur'	'Druck'
'Deutsch'	'Golf'	'Flair'	'Dreck'	'Brett'	'Fort'	'Harz'	'Graf'	'Klein'
'Hals'	'Hecht'	'Kerl'	'Fink'	'Cent'	'Geld'	'Kalb'	'Licht'	'Mast'
'Hehl'	'Knall'	'Milch'	'Fuchs'	'Clown'	'Grill'	'Krieg'	'Schacht'	'Rind'
'Huhn'	'Kult'	'Nerv'	'Grab'	'Duft'	'Hand'	'Luft'	'Schenk'	'Scherz'
'Knie'	'Mark'	'Rand'	'Land'	'Gans'	'Haupt'	'Mix'	'Sound'	'Schrott'
'Papst'	'Schloß'	'Schluck'	'Schwung'	'Griff'	'Kleid'	'Track'	'Stein'	'Stier'
'Seil'	'Staub'	'Schwein'	'Stab'	'März'	'Rumpf'	'Wert'	'Treff'	'Tank'
'Tor'	'Wind'	'Trick'	'Wurm'	'Meer'	'Wild'	'Wurz'	'Wolf'	'Zelt'
'Ball'	'Frucht'	'Pflicht'	'Schmelz'	'Film'	'Gunst'	'Grund'	'Sport'	'Stern'
'Glück'	'Schwanz'	'Punkt'	'Strom'	'Stand'	'Markt'	'Schwarm'	'Strauß'	'Streik'

Tabelle 10: 1. Version der Testlisten: Liste 10 bis Liste 18

Liste 10	Liste 11	Liste 12	Liste 13	Liste 14	Liste 15	Liste 16	Liste 17	Liste 18
'Ohr'	'Schau'	'Zoo'	'All'	'Reh'	'Mai'	'Aus'	'Hai'	'Tee'
'Keim'	'Gott'	'Deal'	'Bär'	'Bad'	'Bass'	'Burg'	'Arm'	'Bach'
'Kurs'	'Schrei'	'Heil'	'Jahr'	'Hab'	'Rock'	'Kahn'	'Mut'	'Reim'
'Lied'	'Klo'	'Mail'	'Leid'	'Kuss'	'Scheck'	'Lohn'	'Fall'	'Hahn'
'Matt'	'Tour'	'Pech'	'Rausch'	'Paar'	'Teig'	'Pin'	'Zeit'	'Zoll'
'Schatz'	'Lamm'	'Pfeil'	'Schal'	'Sitz'	'Typ'	'Tag'	'Lauf'	'Pop'
'Schick'	'Kiel'	'Pfiff'	'Tisch'	'Top'	'Wehr'	'Ton'	'Chef'	'Bus'
'Ziel'	'Schuß'	'Team'	'Wut'	'Tun'	'Zahn'	'Zug'	'Raum'	'Ross'
'Dampf'	'Blech'	'Hirn'	'Berg'	'Arzt'	'Blut'	'Form'	'Fakt'	'Faust'
'Frosch'	'Durst'	'Kind'	'Bund'	'Braut'	'Fleisch'	'Haft'	'Feind'	'Fest'
'Funk'	'Fahrt'	'Mond'	'Gift'	'Nichts'	'Mars'	'Jagd'	'Greif'	'Fluss'
'Geist'	'Feld'	'Saft'	'Park'	'Rest'	'Mord'	'Mönch'	'Holz'	'Kern'
'Helm'	'Flut'	'Schlaf'	'Salz'	'Rost'	'Pfund'	'Nest'	'Kran'	'Kreuz'
'Klick'	'Gramm'	'Stadt'	'Schirm'	'Sand'	'Schicht'	'Schlamm'	'Lärm'	'Krug'
'Korb'	'Macht'	'Stil'	'Schmuck'	'Schwan'	'Speck'	'Schlauch'	'Puls'	'Pilz'
'Post'	'Schluß'	'Stoff'	'Star'	'Senf'	'Speer'	'Trieb'	'Schlag'	'Sicht'
'Sturz'	'Schnitt'	'Welt'	'Stich'	'Takt'	'Trotz'	'Zimt'	'Trip'	'Stahl'
'Tritt'	'Stock'	'Wunsch'	'Wand'	'Zweig'	'Wirt'	'Zwang'	'Wald'	'Tanz'
'Norm'	'Kraft'	'Fürst'	'Dienst'	'Strang'	'Flucht'	'Christ'	'Glanz'	'Frust'
'Prinz'	'Trend'	'Strahl'	'Freund'	'Zunft'	'Spruch'	'First'	'Start'	'Trumpf'

Tabelle 11: 1. Version der Testlisten: Liste 19 bis Liste 27

Liste 19	Liste 20	Liste 21	Liste 22	Liste 23	Liste 24	Liste 25	Liste 26	Liste 27
'Po'	'Fee'	'Bau'	'Kai'	'Ski'	'Eis'	'See'	'Tau'	'Öl'
'Bauch'	'Damm'	'Haut'	'Bann'	'Ast'	'Dieb'	'Beat'	'Ding'	'Hauch'
'Neid'	'Stau'	'Kür'	'Sack'	'Schub'	'Fach'	'Reich'	'Fisch'	'Netz'
'Heck'	'Kauf'	'Mal'	'Chip'	'Mais'	'Gas'	'Gang'	'Hit'	'Not'
'Ruf'	'Witz'	'Maß'	'Set'	'Tool'	'Hang'	'Weib'	'Laub'	'Rang'
'Koch'	'Rauch'	'Rohr'	'Ring'	'Pool'	'Heim'	'Nuss'	'Müll'	'Reis'
'Haar'	'Frau'	'Tier'	'Buch'	'Dom'	'Ruhm'	'Fuß'	'Rat'	'Turm'
'Pfad'	'Tal'	'Zeug'	'Schiff'	'Tief'	'Sein'	'Riss'	'Schein'	'Watt'
'Block'	'Band'	'Glas'	'Bord'	'Bucht'	'Box'	'Dorf'	'Angst'	'Blick'
'Fund'	'Furcht'	'Gold'	'Gras'	'Drang'	'Herd'	'Gast'	'Glied'	'Dank'
'Gleis'	'Krach'	'Keks'	'Groß'	'Jeans'	'Last'	'Grad'	'Gruß'	'Fax'
'Held'	'List'	'Plan'	'Horn'	'Knopf'	'Lift'	'Herz'	'Hemd'	'Fels'
'Hirsch'	'Platz'	'Port'	'Klan'	'Korn'	'Mist'	'Hund'	'Kreis'	'Fleck'
'Klub'	'Spott'	'Preis'	'Kost'	'Mensch'	'Nacht'	'Kampf'	'Link'	'Marsch'
'Lust'	'Spur'	'Spieß'	'Punk'	'Plus'	'Sarg'	'Klang'	'Mund'	'Pferd'
'Schalk'	'Stamm'	'Staat'	'Schmutz'	'Schuld'	'Schild'	'Part'	'Schritt'	'Qual'
'Spiel'	'Volk'	'Stall'	'Thron'	'Spaß'	'Schweiß'	'Recht'	'Steg'	'Stück'
'Zweck'	'Wucht'	'Wachs'	'Zorn'	'Wort'	'Sturm'	'Traum'	'Wurst'	'Stuhl'
'Schlacht'	'Schmerz'	'Brand'	'Pracht'	'Streit'	'Strich'	'Schränk'	'Krebs'	Front'
'Trost'	'Streich'	'Kunst'	'Strauch'	'Werft'	'Text'	'Schwert'	'Strick'	'Stift'

## 7.6 Auflistung der perzeptiv optimierten Testlisten

Tabelle 12: 2. Version: Testlisten perzeptiv optimiert: Liste 1 bis Liste 9

Liste 1	Liste 2	Liste 3	Liste 4	Liste 5	Liste 6	Liste 7	Liste 8	Liste 9
'Aal'	'Uhr'	'Schau'	'Eis'	'Bau'	'Schuh'	'Ohr'	'App'	'Heu'
'Beat'	'Bauch'	'Bad'	'Bass'	'Bock'	'Akt'	'Heil'	'Dach'	'Ast'
'Gen'	'Boot'	'Chef'	'Couch'	'Fan'	'Bar'	'Hoch'	'Deutsch'	'Bahn'
'Heck'	'Haar'	'Hauch'	'Fall'	'Fisch'	'Bob'	'Müll'	'Leib'	'Dieb'
'Maus'	'Hab'	'Huhn'	'Kopf'	'Fuß'	'Herr'	'Not'	'Meer'	'Nuss'
'Ring'	'Paar'	'Lamm'	'Pfad'	'Schiff'	'Lob'	'Pol'	'Schaf'	'Ried'
'Schrei'	'Tisch'	'Pech'	'Rock'	'Tipp'	'Reim'	'Reis'	'Schuß'	'Tief'
'Turm'	'Zoll'	'Schein'	'Wehr'	'Zug'	'Schott'	'Top'	'Tuch'	'Tod'
'Brauch'	'Blitz'	'Duft'	'Fort'	'Bord'	'Arzt'	'Brief'	'Braut'	'Dorf'
'Mönch'	'Clown'	'Feld'	'Kerl'	'Flug'	'Bart'	'Bruch'	'Gleis'	'Draht'
'Punk'	'Darm'	'Haft'	'Puls'	'Lift'	'Brot'	'Durst'	'Hort'	'Horn'
'Schalk'	'Fahrt'	'Harz'	'Qual'	'Nest'	'Fest'	'Fax'	'Jeans'	'Klein'
'Spiel'	'Fink'	'Kampf'	'Rest'	'Plus'	'Gans'	'Feind'	'Pfund'	'Kost'
'Stadt'	'Fund'	'Krach'	'Speck'	'Schloß'	'Grab'	'Grill'	'Schlauch'	'Mast'
'Stück'	'Rost'	'Plan'	'Tritt'	'Spieß'	'Mars'	'Kalb'	'Wert'	'Sound'
'Stuhl'	'Schmutz'	'Rumpf'	'Volk'	'Stoff'	'Sand'	'Schluck'	'Wind'	'Stamm'
'Tank'	'Schnitt'	'Schicht'	'Wand'	'Wachs'	'Stab'	'Track'	'Wolf'	'Trieb'
'Wurz'	'Takt'	'Schuld'	'Wucht'	'Zweck'	'Test'	'Wild'	'Zweig'	'Vers'
'Krebs'	'Drink'	'Sport'	'Pflicht'	'Schwarm'	'Fürst'	'Prinz'	'Brust'	'Kraft'
'Markt'	'Punkt'	'Streich'	'Schmelz'	'Stift'	'Strauß'	'Strich'	'Papst'	'Stern'

Tabelle 13: 2. Version: Testlisten perzeptiv optimiert: Liste 10 bis Liste 18

Liste 10	Liste 11	Liste 12	Liste 13	Liste 14	Liste 15	Liste 16	Liste 17	Liste 18
'Tee'	'Sau'	'Tau'	'All'	'Reh'	'Ski'	'Po'	'Hai'	'Kuh'
'Ball'	'Hehl'	'Bach'	'Busch'	'Art'	'Gut'	'Burg'	'Buch'	'Abt'
'Keim'	'Klo'	'Bier'	'Chip'	'Bus'	'Rang'	'Fett'	'Lau'	'Damm'
'Matt'	'Riss'	'Deal'	'Schal'	'Mut'	'Rausch'	'Heim'	'Koch'	'Hahn'
'Schatz'	'Schach'	'Gang'	'Sitz'	'Netz'	'Stau'	'Jahr'	'Ton'	'Kuss'
'Schick'	'Scheck'	'Leid'	'Tour'	'Raub'	'Wut'	'Kahn'	'Pfiff'	'Look'
'Sein'	'Seil'	'Pfeil'	'Typ'	'Ross'	'Zahn'	'Pin'	'Mal'	'Maul'
'Tausch'	'Team'	'Pop'	'Wahl'	'Tun'	'Zeug'	'Sieg'	'Zeit'	'Zaun'
'Berg'	'Angst'	'Fluss'	'Bund'	'Jagd'	'Bank'	'Form'	'Druck'	'Bild'
'Blatt'	'Flut'	'Hirn'	'Hals'	'Klan'	'Dorn'	'Halt'	'Farm'	'Geist'
'Bucht'	'Glied'	'Klub'	'Drang'	'Knall'	'Flur'	'Hund'	'Kleid'	'Herz'
'Fels'	'Golf'	'Recht'	'Gift'	'Knopf'	'Glas'	'Lust'	'Lärm'	'Krieg'
'Klang'	'Graf'	'Schlaf'	'Sekt'	'Mist'	'Gold'	'Pferd'	'Link'	'Krug'
'Klick'	'Groß'	'Schluß'	'Kern'	'Nacht'	'Licht'	'Schweiß'	'Mark'	'Marsch'
'Milch'	'Hand'	'Speer'	'Kult'	'Pilz'	'Mord'	'Stock'	'Schenk'	'Mond'
'Platz'	'Land'	'Staat'	'Stahl'	'Schlag'	'Rind'	'Welt'	'Spaß'	'Mund'
'Port'	'Macht'	'Stich'	'Tanz'	'Schwan'	'Trick'	'Wirt'	'Spur'	'Preis'
'Stall'	'Schlamm'	'Wort'	'Sarg'	'Sturz'	'Trotz'	'Zimt'	'Stil'	'Scherz'
'Strahl'	'Strauch'	'Norm'	'Freund'	'Frucht'	'First'	'Dienst'	'Schwert'	'Frust'
'Strang'	'Werft'	'Stand'	'Glanz'	'Kunst'	'Spruch'	'Schwanz'	'Strick'	'Trumpf'

Tabelle 14: 2. Version: Testlisten perzeptiv optimiert: Liste 19 bis Liste 27

Liste 19	Liste 20	Liste 21	Liste 22	Liste 23	Liste 24	Liste 25	Liste 26	Liste 27
'Aus'	'Fee'	'Ass'	'Kai'	'Mai'	'Eck'	'See'	'Öl'	'Zoo'
'Bug'	'Kauf'	'Gier'	'Bann'	'Bett'	'Fach'	'Bein'	'Hit'	'Arm'
'Frau'	'Lied'	'Haut'	'Baum'	'Dom'	'Fell'	'Depp'	'Kur'	'Bär'
'Kiel'	'Ort'	'Kür'	'Deck'	'Heer'	'Gas'	'Knie'	'Laub'	'Hall'
'Raum'	'Rauch'	'Lohn'	'Ding'	'Mais'	'Hang'	'Tag'	'Maß'	'Mail'
'Ruf'	'Set'	'Rohr'	'Gott'	'Mann'	'Pool'	'Weib'	'Rat'	'Pack'
'Sack'	'Sinn'	'Teig'	'Kurs'	'Neid'	'Reich'	'Wein'	'Schock'	'Schub'
'Tal'	'Witz'	'Tier'	'Tor'	'Tool'	'Ruhm'	'Ziel'	'Teich'	'Watt'
'Box'	'Band'	'Gruß'	'Funk'	'Blick'	'Cent'	'Fleck'	'Blech'	'Blog'
'Film'	'Block'	'Hecht'	'Grad'	'Blut'	'Fleiß'	'Fuchs'	'Brett'	'Dampf'
'Helm'	'Dank'	'Hirsch'	'Held'	'Fakt'	'Frosch'	'Griff'	'Flair'	'Dreck'
'Kind'	'Fleisch'	'Keks'	'Last'	'Haupt'	'Gramm'	'Korn'	'Furcht'	'Faust'
'Kreis'	'Gras'	'Luft'	'Mix'	'Hemd'	'Herd'	'Kran'	'Gast'	'Geld'
'List'	'Greif'	'Schritt'	'Post'	'Holz'	'Saft'	'Salz'	'Korb'	'Glück'
'Nerv'	'Schacht'	'Schrott'	'Schirm'	'Kreuz'	'Sicht'	'Schwein'	'Schmuck'	'März'
'Park'	'Thron'	'Trip'	'Steg'	'Mensch'	'Stier'	'Star'	'Schwung'	'Nichts'
'Senf'	'Wald'	'Zelt'	'Werk'	'Part'	'Sturm'	'Traum'	'Stein'	'Rand'
'Spott'	'Wunsch'	'Zwang'	'Zorn'	'Schild'	'Wurst'	'Wurm'	'Treff'	'Staub'
'Schlacht'	'Text'	'Brand'	'Grund'	'Streit'	'Schrank'	'Flucht'	'Christ'	'Front'
'Schmerz'	'Trost'	'Start'	'Pracht'	'Trend'	'Streik'	'Zunft'	'Strom'	'Gunst'

