OPEN ACCESS

Originalarbeit

Untersuchung einer synthetischen Stimme für den Freiburger Einsilbertest

Examination of a synthetic voice for the Freiburg Monosyllabic Speech Test

Thomas Schwarz[1]
Marlitt Frenz[1]
Alina Bockelmann[1]
Hendrik Husstedt[1]

1 Deutsches Hörgeräte Institut GmbH, Lübeck, Deutschland

Abstract

The Freiburg Speech Test is a commonly used speech test in German-speaking countries. The test corpus was recorded in 1969 and reference curves for performing the Freiburg monosyllabic speech test (FET) in quiet are defined in DIN 45621-1. In the context of this work, test words generated by a synthetic voice are compared with the original speech material with regard to speech intelligibility in quiet. For this purpose, the synthetic speech material was generated by using a commercial text-to-speech system (TTS). The motivation for using a synthetic voice is that an update or extension of the speech material with the same voice is also possible in future. In addition, this would avoid costly and time-consuming new recordings. On the basis of measurements with 40 normal-hearing subjects, psychometric functions for the FET with the original and synthetic test material and speech intelligibility values for the single words and lists were determined. The test was performed in free field in quiet in an appropriate audiological test room.
When comparing the determined psychometric functions of the FET in the original-voice-condition with the FET in the synthetic-voice-condition, there is no significant difference in the mean SRT or the mean slope. Looking at the single-word comprehension, there are isolated words that were understood significantly worse by the test subjects due to the generation of the TTS system compared to the original recordings. When listening to these words in synthetic condition, an unnaturalness in pronunciation is noticeable, which can be attributed to different reasons. The results of this study show, that the creation and use of the FET with a synthetic voice seems to be feasible and reasonable.

Keywords: Freiburg monosyllabic speech test, synthetic voice, speech intelligibility, psychometric function

Zusammenfassung

Der Freiburger Sprachtest ist der im deutschsprachigen Raum am häufigsten verwendete Sprachtest.
Die Aufnahmen der Testwörter stammen aus dem Jahr 1969 und Sprachverständlichkeits- Bezugskurven für Messungen mit dem Freiburger Einsilbertest (FET) in Ruhe sind in der DIN 45621-1 definiert. Im Rahmen dieser Arbeit wurden mittels synthetischer Stimme einsilbige Testwörter erzeugt und mit dem originalen Sprachmaterial im Hinblick auf die Sprachverständlichkeit in Ruhe verglichen. Dafür wurde das synthetische Sprachmaterial des FET über ein kommerzielles Text-to-Speech (TTS)-System erzeugt. Die Entwicklung eines Sprachtests mit synthetischer Stimme findet vor dem Hintergrund statt, eine langfristige Lösung für einen um Sprachbestandteile austauschbaren und erweiterbaren Sprachtest zu finden. So ließen sich kosten- und zeitaufwendige Neuaufnahmen vermeiden und es bestände die Möglichkeit einer stetigen Weiterentwicklung des Sprachtests. Auf Grundlage einer Probandenmessung mit 40 normalhörenden Probanden wurden psychometri-

sche Funktionen für den FET mit originalem und synthetischem Testmaterial und Sprachverständlichkeitswerte für die Einzelwörter und Listen ermittelt. Bei der Probandenmessung wurde der FET im Freifeld in Ruhe in einer geeigneten Messkabine durchgeführt. Der Vergleich zwischen ermittelten psychometrischen Funktionen des FET mit originaler und synthetischer Stimme für den gesamten Test zeigt weder im mittleren SRT noch in der mittleren Steigung einen signifikanten Unterschied. Bei der Untersuchung zum Einzelwortverstehen gibt es einzelne Wörter, die durch die Erzeugung vom TTS-System im Vergleich mit den originalen Aufnahmen von den Probanden schlechter verstanden wurden. Beim Anhören dieser Wörter fällt eine durch das Synthesesystem erzeugte Unnatürlichkeit in der Aussprache auf, die auf unterschiedliche Ursachen zurückgeführt werden kann. Für die Zukunft wäre nach den Ergebnissen dieser Studie die Erstellung und Durchführung eines mit synthetischer Stimme erstellten FET mit einer angepassten Synthesestimme sinnvoll möglich.

Schlüsselwörter: Freiburger Einsilbertest, synthetische Stimme, Sprachverständlichkeit, psychometrische Funktion

Einleitung

Ein zentrales Ziel der Hörsystemversorgung ist ein verbessertes Sprachverstehen des Hörsystemträgers [1]. Um dies zu ermitteln, können verschiedene Sprachtestverfahren genutzt werden, mit denen eine erfolgreiche Hörsystemanpassung nachvollziehbar überprüft werden kann [2]. Im deutschsprachigen Raum wird zu diesem Zweck am häufigsten der Freiburger Sprachtest (FST) nach DIN 45621-1 verwendet [3]. Das Ergebnis des FST ist nach Heil- und Hilfsmittelrichtlinie § 21/22 zusammen mit dem Tonaudiogrammergebnis für die Indikation einer Hörgeräteversorgung und damit für die Kostenübernahme der Krankenkassen entscheidend [2].

In der Vergangenheit wurde der FST, insbesondere der Freiburger Einsilbertest (FET), in verschiedenen Aspekten kritisiert und diskutiert. Ein zentraler Aspekt der Kritiker ist, dass die Wortlisten phonemisch unausgeglichen sind [4]. Aber auch psychische Hemmnisse, einzelne Worte an den Prüfer gerichtet nachzusprechen (z.B.: Sau, Schwein, Sarg etc.), regionale Sprachbesonderheiten, Kontextbezug und eine unrealistische Überartikulation wurden unter anderem durch Bangert [4], Alich [5] und von Wedel [6] bemängelt. Ebenso stellten Winkler und Holube in einer Untersuchungsreihe fest, dass die verschiedenen Digitalisierungen des FET nicht alle Mindestanforderungen eines Sprachtests nach DIN EN ISO 8253-3 [7] erfüllen [8]. Differenzen zeigten sich unter anderem im mittleren Sprachpegel und der unnatürlichen Artikulation des Sprechers.

Steffens stellte 2016 in seinen Recherchen zur Verwendungshäufigkeit der Freiburger Einsilber in der Gegenwartssprache zudem fest, dass etwa 45% der Einsilber in der Alltagssprache praktisch nicht mehr verwendet werden, was auf ein veraltetes Testinventar schließen lässt. Zudem vermutete er, dass auch die höhere Verwendungshäufigkeit der Einsilber in der Schriftsprache zu positiven Messabweichungen bei belesenen Probanden führen kann [9].

Winkler et al. untersuchten mit einer weiteren Studie neben dem Einfluss der Wortfrequenz (Verwendungshäufigkeit der Einsilber in Schriftkorpora) auf die mit dem FET gemessene Sprachverständlichkeit auch die Nachbarschaftsdichte, die die 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 haben [10].

Den FET gemäß dieser Kritikpunkte zu überarbeiten, ohne dabei, wie es beispielsweise im Rahmen der Bachelorarbeit von Felix Hahn praktiziert wurde [11], den Testkorpus um Wörter reduzieren zu müssen, ist ohne eine Neuaufnahme des Tests nur bedingt möglich.

Schon bei der Entwicklung des FST versuchte Karl-Heinz Hahlbrock 1952, die Bedeutung der Verwendung einer einheitlichen Stimme und Vortragsweise bei der Testwiedergabe für eine verlässliche Reproduzierbarkeit der Ergebnisse zu berücksichtigen. Es wurde mit einer einheitlichen Tonbandaufnahme des Testmaterials gearbeitet, die von einem ausgebildeten Sprecher aufgenommen wurde [12]. Die heute verwendete Aufnahme des Testmaterials wurde vom Sprecher Claus Wunderlich im Jahr 1969 aufgesprochen. Mit Hilfe des dadurch entstandenen Testmaterials wurden durch Messungen der Physikalisch-Technischen Bundesanstalt (PTB) unter definierten Bedingungen Bezugskurven für Normalhörende erstellt, die in der Norm DIN 45621-1 zu finden sind [13]. Der Sprecher dieser heute noch verwendeten Aufnahme des FET ist inzwischen verstorben, sodass Neuaufnahmen für den FET mit seiner Stimme nicht möglich sind.

Neue Wörter für den FET mit einem neuen Sprecher zu produzieren, wie es beispielsweise in den Dissertationen von Mahfoud [14] und Qualen [15] in Würzburg durchgeführt wurde, ist wiederum mit einem großen Aufwand und hohen Kosten, zum Beispiel für professionelle Studioaufnahmen, verbunden. Durch den stetigen Wandel von Sprache [16] werden in der Gegenwart häufig genutzte Einsilber in 10 Jahren vielleicht seltener benutzt, sodass

es sich bei diesem Ansatz lediglich um eine temporäre Lösung handeln würde.
Ein langfristiger Ansatz könnte ein mittels Text-to-Speech-System (TTS-System) erstellter FET sein. Im Verhältnis zu wiederkehrenden Neuaufnahmen stellt dieser eine kostengünstigere Alternative dar. Wörter, die nicht mehr in der Alltagssprache verwendet werden, könnten durch aktuell bekannte Wörter unter definierten Randbedingungen ersetzt werden. Somit wären Verbesserungen am Sprachkorpus oder auch eine Erweiterung des Tests um weitere Testlisten unkomplizierter.
Die Sprachverständlichkeit von synthetischen Stimmen wurde bereits in verschiedenen Arbeiten untersucht [17], [18], [19]. Einen deutschen Sprachtest mit synthetischer Stimme unter audiologischen Gesichtspunkten durchzuführen, war dagegen bisher ausschließlich Bestandteil von Untersuchungen zum Oldenburger Satztest (OLSA) aus 2019 von Nuesse et al. [20]. Die Ergebnisse dieser Studie zeigen, dass eine Sprachaudiometrie mit synthetischer Stimme, die an Originalaufnahmen angepasst war, beim OLSA zu ähnlichen Ergebnissen führte, wie die Durchführung mit originalen Aufnahmen. Im Rahmen dieser Studie wurde nun überprüft, inwiefern bei einer Durchführung des FET mit synthetischer Stimme Unterschiede zu einer klassischen Durchführung mit den Originalaufnahmen des FET bei der Sprachverständlichkeit in Ruhe entstehen. Dafür wurde in einem ersten Schritt ein TTS-System mit einer Stimme, die Ähnlichkeiten zur Klangfarbe des Sprechers der aktuellen Aufnahme hat, ausgewählt. Zusätzlich dazu wurde die Lautstärke und Sprechgeschwindigkeit der Stimme des TTS-Systems möglichst nahe an die Sprecheigenschaften der Stimme von Claus Wunderlich angeglichen, um mögliche Ursachen für ein unterschiedliches Sprachverstehen zu vermeiden. Neben der Lautstärke können nämlich auch Unterschiede bei der Grundfrequenz und Sprechgeschwindigkeit der synthetischen Stimme zu Unterschieden im Sprachverstehen beim FET führen [21]. Der FET wurde dann mit den Originalaufnahmen und dem durch das TTS-System erzeugten Sprachmaterial mit insgesamt 40 normalhörenden Probanden durchgeführt. Durch das Erstellen von psychometrischen Funktionen für die beiden Testkorpora sowie das Ermitteln von Sprachverständlichkeitswerten pro Wort im Allgemeinen und Liste je Schallpegel wurde ein Vergleich über Veränderungen der Sprachverständlichkeit gezogen. Auf diese Weise soll die Forschungsfrage beantwortet werden, inwieweit die Durchführung des FET mit synthetisch erstelltem Testmaterial zu signifikanten Abweichungen der Sprachverständlichkeit im Ergebnis führt.

Material und Methoden

Auswahl der synthetischen Stimme

Bei der Auswahl der synthetischen Stimme und des Anbieters wurde auf die wahrgenommene Natürlichkeit der produzierten Sprache und auf die subjektiv empfundene Ähnlichkeit zur Stimme des FET-Sprechers Claus Wunderlich in seiner tiefen Grundfrequenz geachtet. Die Auswahl fiel auf die Stimme „Klaus“ der Acapela Group, einem Unternehmen für Sprachtechnologie mit Hauptsitz in Belgien, die für das Projekt daher erworben wurde.
Die Ansteuerung der Stimme mit Matlab erfolgte über eine Anwendungsprogramm-Schnittstelle (API) in Python, die Zugriff auf die Cloud der Acapela Group hatte. Zur Erstellung des Testmaterials konnte über die Schnittstelle – also vor der Erstellung einer Audiodatei (.wav) – die Sprechgeschwindigkeit, spektrale Veränderungen und die Lautstärke festgelegt werden. Für jeden Einsilber wurde eine Datei mit einer Abtastrate von 44.100 Hz erstellt.

Anpassung des synthetischen Sprachmaterials

Um den FET mit synthetischer Stimme mit dem originalen FET vergleichbar zu halten und um den Normbestimmungen der Pegelverteilung über die Wörter aus DIN 45626-1 [22] möglichst zu entsprechen, wurden die mit synthetischer Stimme erzeugten Wörter in den Aspekten Sprechgeschwindigkeit pro Wort und mittlerer Leistungspegel pro Wort angepasst. Auf eine exakte Anpassung der Grundfrequenz wurde verzichtet, da nach den Ergebnissen der Studien von Williamson und Harmon-Smith sowie von Bradlow et al. die Sprachverständlichkeit durch die Verwendung von Sprechern mit unterschiedlichen, gemittelten Grundfrequenzen nicht signifikant beeinflusst wird [23], [24].
Für die Anpassungen wurde für alle Audiodateien eine Root Mean Square (RMS)-Hüllkurve mit einer Fensterbreite von 0,04 s über die Signale berechnet. Mithilfe dieser Einhüllenden und einem festgelegten Schwellenwert wurde das vor- und nachlaufende Aufnahmerauschen bzw. die „Stille“ weggeschnitten, sodass die eigentliche Länge und der mittlere Schallpegel des gesprochenen Wortes ermittelt werden konnten. Bei einigen Worten, wie beispielsweise Worten mit langem stimmlosem Auslaut (z.B. Biss, Fels, Schiff), wurde die Begrenzung nochmal manuell korrigiert, da sich die Definition der Wortlänge über den Schwellwert für diese Worte als zu grob herausstellte.
Um die synthetisch produzierten Wörter an die Sprechgeschwindigkeit der Wörter des Originalsprachtestmaterials anzupassen, wurde jedes Wort einmal in der Standardgeschwindigkeit des Synthesesystems, Stufe 100 (entspricht 100%), erzeugt. Dann wurde die Länge des synthetischen und des originalen Wortes wie beschrieben im Millisekunden Bereich ermittelt. Das synthetische Signal wurde in der entsprechend angepassten Sprechgeschwindigkeit neu erzeugt und dabei auch an den Schallpegel des Originalwortes angepasst.
Zu bemerken ist, dass das Schneiden der Wörter lediglich zur Bestimmung der Wortlänge und Berechnung des Schallpegels durchgeführt wurde. Für die Nutzung im

Sprachtest wurden sowohl originale wie auch synthetische Worte ungeschnitten genutzt.
Beim Anhören des synthetischen Testmaterials fiel bei einigen Wörtern eine unnatürliche Aussprache auf. Da die genutzte API auch eine Worteingabe über Phoneme anbietet, wurden 8 auffallende Wörter durch die ebenfalls vom System unterstützte Eingabe von Phonemen neu erzeugt, um eine natürlichere Aussprache zu erzielen. Anschließend wurden die so erzeugten Wörter ebenfalls nach dem beschriebenen Verfahren in Länge und Schallpegel angepasst.

Messaufbau

Die Probandenmessungen fanden in einem audiologischen Testraum der Deutschen Hörgeräte Institut GmbH statt. Der Hintergrundschallpegel im Raum liegt unter dem in DIN EN ISO 8253-2 [25] für tonaudiometrische Messungen im freien Schallfeld vorgegebenen Grenzwert. Damit ist er auch für die Aufnahme von sprachaudiometrischen Ruhebezugskurven geeignet, da es sich bei der Sprachaudiometrie um eine überschwellige Messung handelt. Die Durchführung des FET sowie die Datenauswertung erfolgten mithilfe eines selbstgeschriebenen Matlab-Scriptes unter der Version MATLAB R2020a.
Der Lautsprecher vom Typ Genelec 8351A wurde in einem audiologischen Testraum in 1 m Entfernung frontal zur Mitte des Probandenkopfes aufgestellt. Die Höhe des Schallaustritts des Lautsprechers befand sich auf der Medianebene des Probanden. Zur Vermeidung von unerwünschten Stehwellen im Raum wurde der Messaufbau schräg ausgerichtet. Der Bildschirm des Prüfers wurde für die Messung des FET vor den Blicken des Probanden abgeschirmt.
Um sicherzustellen, dass das Quantisierungsrauschen der Soundkarte und das Eigenrauschen des Lautsprechers die Messungen nicht beeinflussen, wurde die Verstärkung zunächst am Lautsprecher und dann an der Soundkarte reduziert. Die Soundkarte wurde mit 16-bit betrieben und der Fullscale-Wert entsprach 80 dB SPL. Die Ruhebezugskurve für binaurales Sprachverstehen nach DIN 45626-1 liegt im Bereich von 10 dB bis 45 dB SPL [22], sodass der Dynamikbereich für die Messungen ausreichend war.
Das genutzte Messsystem wurde mit einem kalibrierten Mikrofon vom Typ Brüel & Kjær 4190 über alle Terzbänder im Frequenzbereich von 125 bis 8.000 Hz unter Berücksichtigung der Raumimpulsantwort entzerrt. Bei der Entzerrung wurde entsprechend der DIN EN ISO 8253-3 eine Abweichung von maximal 2 dB toleriert [7].
Um die Korrektheit der Pegelabgabe zu gewährleisten, wurde das CCITT-Rauschen der Siemens-CD für den FET bei 65 dB vor jede Probandenmessung über das entzerrte Messsystem wiedergegeben und nach DIN 45626-1 an der Position des Referenzmikrofons mit einem Pegelmesser der Klasse 1 der äquivalente, impulshaft gewertete Schalldruckpegel des CCITT-Rauschens überprüft [22].

Probandenkollektiv

Die Probandenmessungen fanden in zwei Messblöcken mit jeweils 20 Probanden in einem zeitlichen Abstand von 2 Monaten statt. Das Durchschnittsalter aller Probanden betrug 25 Jahre. Die Altersspanne erstreckte sich von 18 bis 34 Jahre. 24 der Probanden waren weiblich und 16 männlich. Im ersten Probandenkollektiv besaßen 16 Probanden bereits Vorerfahrung mit dem FET. Im zweiten Probandenkollektiv wurde darauf geachtet, dass die Anforderungen zur Aufnahme von Bezugskurven für einen Sprachtest nach DIN EN ISO 8253-3 (Proband normalhörend und Alter 18–25 Jahre) erfüllt waren.
Weitere Einschlusskriterien für die Studienteilnahme waren: Deutsch als Muttersprache, keine Erkrankungen oder vergangene Operationen an den Ohren, die mit einem Tonschwellenaudiogramm nachgewiesene Normalhörigkeit, kein Tinnitus und keine akute Erkältung bzw. Erkältungssymptome oder akutes Allergieleiden. Zu Beginn eines jeden Probandentermins wurde ein Anamnesegespräch geführt, in dem die Einschlusskriterien abgefragt wurden. Im weiteren Verlauf wurde mittels Otoskop das äußere Ohr wie auch das Trommelfell auf pathologische Auffälligkeiten untersucht. Daran anschließend wurde die Normalhörigkeit der Probanden mit einer Aufnahme eines Tonschwellenaudiogramms über Kopfhörer überprüft. Als Definition der Normalhörigkeit wurden die in der DIN EN 8253-3 formulierten Empfehlungen, ein Hörverlust im Frequenzbereich von 250 Hz bis 8.000 Hz von höchstens 15 dB in zwei gemessenen Frequenzen, herangezogen [7]. Für die Messungen gab es eine Aufwandsentschädigung von 7€ pro Stunde sowie eine Anfahrtspauschale von 7€ für die Probanden. Die Messungen waren für eine Dauer von 2 Stunden ausgelegt.

Messablauf

Die Messungen des FET erfolgten alle im Freifeld aus 0°. Die Wörter wurden bei den vier Schallpegeln 20 dB, 27 dB, 34 dB und 41 dB gemessen. Die Auswahl dieser Schallpegel richtete sich danach, die Sprachverständlichkeits-Bezugskurve für Einsilber aus der DIN 45626-1 [22] möglichst gut abzutasten. Diese wurde für binaurale Messungen der DIN 45626-1 entsprechend um 3 dB zu niedrigeren Schallpegeln verschoben [22]. In einer Vormessung zeigten die gewählten Schallpegel eine gute Abtastung der Bezugskurve.
Der Proband wurde in einer Einweisung auf das zu prüfende Ohr, die Art der Testelemente und die von ihm geforderte Antwort hingewiesen.
Der Prüfer befand sich während der Messung im selben Raum wie der Proband, um die Antworten des Probanden über direkten Weg zu hören und in das Programm aufnehmen zu können. Über eine Benutzeroberfläche hatte der Prüfer Einsicht auf den momentanen Prüfpegel, die Art des Wortmaterials (synthetisch oder original), die aktuelle Listennummer, die Anzahl getesteter Wörter aus der aktuellen Liste und auf das Verständnis in Prozent der aktuellen Liste sowie der bereits überprüften Listen.

Jedem Probanden wurden in der Messung alle 20 Listen des originalen und des synthetischen FET vorgespielt. Durch eine Randomisierung der Listen mittels lateinischen Quadrats, der Reihenfolge der Stimmparameter (original/synthetisch) pro Probanden, der Schallpegel innerhalb von 4er Blöcken und der Wörter innerhalb der Liste, wurde sichergestellt, dass dieselben Listen innerhalb eines Probandendurchgangs den maximalen Abstand von 19 Listen hatten und Lerneffekte sowie andere Störparameter minimiert wurden.
Die Studie und das Vorgehen innerhalb dieser wurde durch die Ethikkommission der Technischen Hochschule Lübeck mit Schreiben vom 11.09.2020 genehmigt.

Ergebnisse

Bestimmung des Probandenkollektivs

Anhand der Messergebnisse aus der Probandenstudie wurde für jeden Probanden in allen vier Prüfpegeln das durchschnittliche Sprachverstehen berechnet. An diese Werte wurde dann pro Proband eine psychometrische Funktion in Form einer logistischen Funktion für die kleinste quadratische Abweichung angepasst (Abbildung 1 und Abbildung 2 graue Kurven). In der vorliegenden Studie wurden zwei Probandenkollektive mit jeweils 20 Probanden mit verschiedener Altersstruktur und Vorerfahrung mit dem FET in zwei Messzeiträumen untersucht. Für beide Probandenkollektive wurden die Mediane der SRTs und der Steigungen der individuell pro Proband angepassten psychometrischen Funktionen getrennt berechnet (siehe Tabelle 1). Eine Überprüfung auf normalverteilte Daten mit dem Shapiro-Wilk Test zeigte, dass die Ergebnisse der SRTs und der Steigungen nur teilweise als normalverteilt angesehen werden können. Daher wurde bei den nachfolgenden statistischen Untersuchungen auf nicht-parametrische Tests zurückgegriffen.
Zunächst wurden die SRTs und Steigungen zwischen den Probandenkollektiven mit dem Mann-Whitney-U Test verglichen. Hier zeigte sich nur bei der Steigung des synthetischen Sprachmaterials ein statistisch signifikanter Unterschied mit p=0,008**. Die Mediane beider Steigungen liegen mit 5,97%/dB beim 1. Kollektiv und 5,37%/dB beim 2. Kollektiv aber sehr nahe beieinander. Insgesamt werden die beiden Datensätze als ausreichend ähnlich angesehen, um diese im Folgenden zusammen zu betrachten. Diese Entscheidung wurde insbesondere auch vor dem Hintergrund getroffen, dass die nachfolgenden Untersuchungen auf den relativen Vergleich zwischen dem originalen und synthetisch erzeugten Sprachmaterial und nicht auf die Erstellung von Bezugskurven abzielen.

Psychometrische Funktionen

Beim Vergleich der psychometrischen Funktionen des originalen und synthetisch erzeugten Sprachmaterials für das gesamte Probandenkollektiv mit dem Wilcoxon-Vorzeichen-Rang-Test zeigte sich kein signifikanter Unterschied (p=0,129) für die Sprachverständlichkeitsschwellen (SRT), jedoch für die Steigungen mit p=0,029*. Zusätzlich zu den Medianen über die individuell angepassten psychometrischen Funktionen wurden für beide Sprachmateriale auch psychometrische Funktionen entsprechend DIN EN ISO 8253-3:2012 angepasst. Dafür werden zunächst die Mediane bei den vier verwendeten Schallpegeln berechnet, welche dann für die Anpassung der psychometrischen Funktionen herangezogen werden. Dadurch ergeben sich mit $SRT_{orig.}$=28,80 dB, $SRT_{synth.}$=28,84 dB, $s_{orig.}$=5,38%/dB und $s_{synth.}$=5,67 %/dB teilweise leicht andere Werte als in der unteren Zeile von Tabelle 1 angegeben. Da auch Brinkmann [13] zur Erstellung der Sprachverständlichkeitsbezugskurve aus DIN 45626-1 [22] so vorgegangen ist, werden in Abbildung 1 (schwarze Kurve), Abbildung 2 (blaue Kurve) und im nachfolgenden Text bei der Betrachtung gemittelter psychometrischer Funktionen nur diese Werte betrachtet. Die zugehörigen Verteilungsparameter der Messdaten sind in Tabelle 2 dargestellt.
Die an die Mediane angepassten psychometrischen Funktionen der beiden Sprachmaterialien sind in Abbildung 3 gemeinsam abgebildet. Außerdem ist die Sprachverständlichkeits-Bezugskurve der DIN 45626-1:1995-08 [22], welche monaural über Kopfhörer bestimmt wurde, für binaurale Messungen um 3 dB zu geringeren Schallpegeln verschoben eingezeichnet. Um die psychometrischen Funktionen dieser Studie besser mit den Normwerten vergleichen zu können, wurde an die linearisierte Sprachverständlichkeits-Bezugskurve eine psychometrische Funktion in allen 10% Schritten von 0 bis 100% angepasst. Diese besitzt eine Sprachverständlichkeitsschwelle von $SRT_{DIN45626-1}$=26,85 dB und eine Steigung von $s_{DIN45626\,1}$=4,03 %/dB.

Wortverständlichkeit

Um die Verständlichkeit der Einsilber pro Wort gut miteinander vergleichen zu können, wurden die Antworten (richtig=1 und falsch=0) pro Einsilber über alle Probanden und über die vier Messschallpegel gemittelt. In Abbildung 4 sind die Werte im Streudiagramm zwischen originalem und synthetischem Testmaterial dargestellt. Durch die begrenzte Probanden- und Messschallpegelanzahl liegen einige Ergebnisse übereinander. Diese sind durch den Punktdurchmesser und die Farbgebung je nach Anzahl gekennzeichnet, wobei ein größerer Durchmesser sowie eine dunklere Tönung für eine höhere Anzahl an gleichen Ergebnissen steht. Auffällig sind die vermehrt auftretenden Ausreißer hin zu schlechter Sprachverständlichkeit beim synthetischen Sprachmaterial. Die in grau eingefärbte doppelte Standardabweichung der Binomialverteilung zeigt die zufällige Streuung der Verständlichkeitswerte pro Wort für 40 Probanden an.

Listenverständlichkeit

In Abbildung 5 ist das durchschnittliche Sprachverstehen je Liste des FET mit originaler Stimme in den vier Mess-

Tabelle 1: Mediane der Parameter der individuell für jeden Probanden angepassten psychometrischen Funktionen.

Probandenkollektiv	SRT original	SRT synth.	Steigung original	Steigung synth.
1. Kollektiv 1–20	28,19 dB	28,93 dB	5,30%/dB	5,97%/dB
2. Kollektiv 21–40	29,31 dB	28,33 dB	5,47%/dB	5,37%/dB
gesamt 1–40	28,78 dB	28,84 dB	5,40%/dB	5,67%/dB

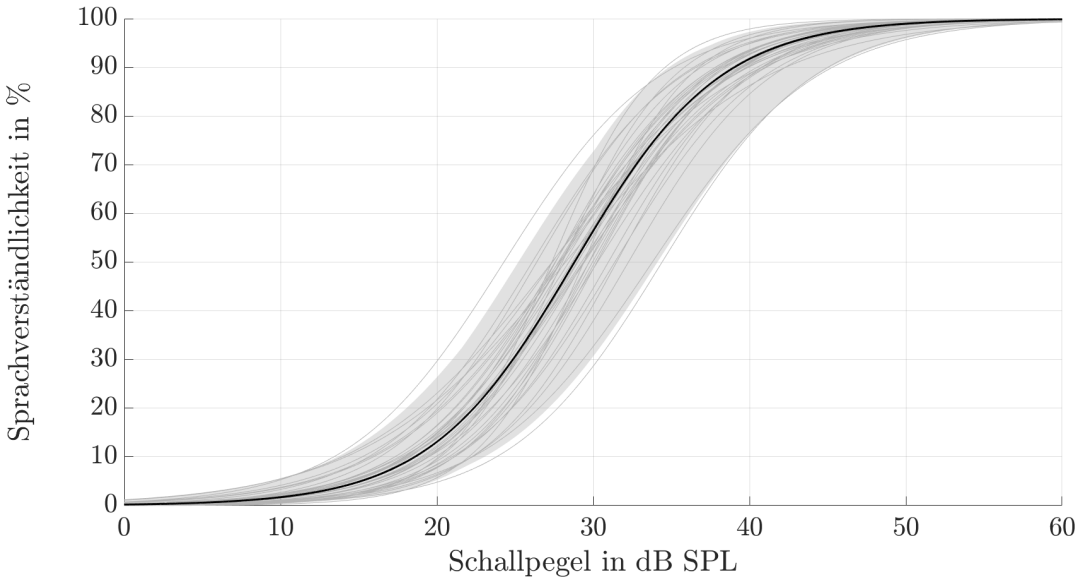

Abbildung 1: Ermittelte psychometrische Funktionen aller Probanden der Messung mit originaler Stimme; die schwarz hervorgehobene Kurve entspricht einer an den Median in den vier Messschallpegeln angepassten psychometrischen Funktion mit $SRT_{original}$=28,80 dB und $s_{original}$=5,38%/dB; grau hinterlegter Bereich zeigt das 95%-Konfidenzintervall über die Sprachverständlichkeit an.

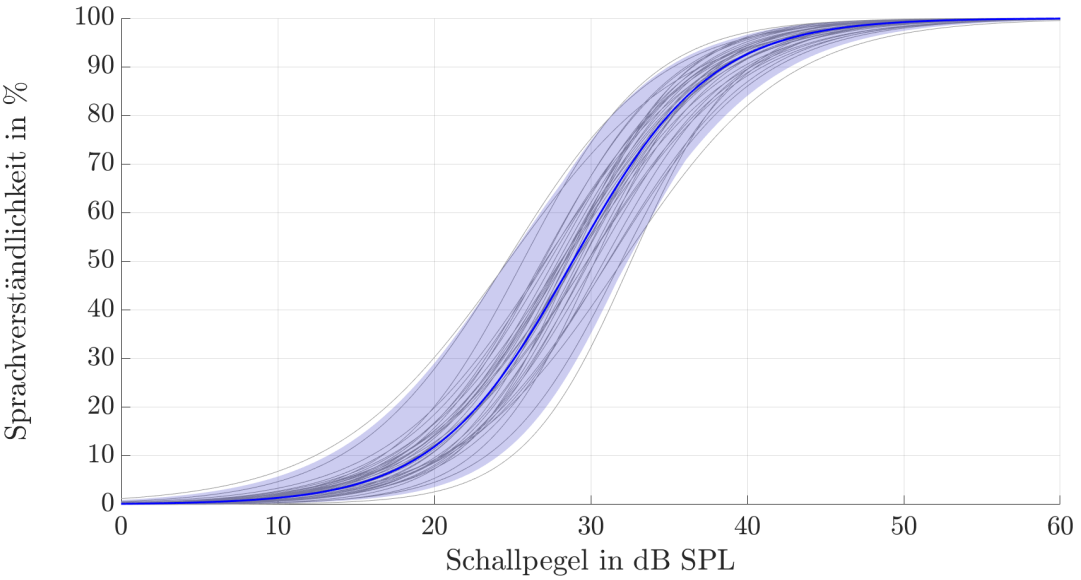

Abbildung 2: Ermittelte psychometrische Funktionen aller Probanden der Messung mit synthetischer Stimme; die blau hervorgehobene Kurve entspricht einer an den Median in den vier Messschallpegeln angepassten psychometrischen Funktion mit $SRT_{synth.}$=28,84 dB und $s_{synth.}$=5,67%/dB; blau hinterlegter Bereich zeigt das 95%-Konfidenzintervall über die Sprachverständlichkeit an.

Tabelle 2: Verteilungsparameter der Messdaten von 40 Probanden für den FET mit originaler Stimme und synthetischer Stimme.

Parameter	Originale Stimme				Synthetische Stimme			
Schallpegel	20 dB	27 dB	34 dB	41 dB	20 dB	27 dB	34 dB	41 dB
Median	9,5%	43,5%	74,0%	92,0%	9,0%	41,5%	76,5%	91,0%
Ø	10,7%	41,8%	73,2%	91,0%	9,2%	42,5%	75,5%	90,6%
SD	5,9	10,3	7,8	4,4	6,1	10,9	6,5	3,8
95%-KI	[5,3; 26,3]	[19,6; 59,0]	[54,4; 90,4]	[79,8; 98,0]	[3,6; 29,0]	[19,4; 61,0]	[65,2; 90,0]	[86,5; 97,4]

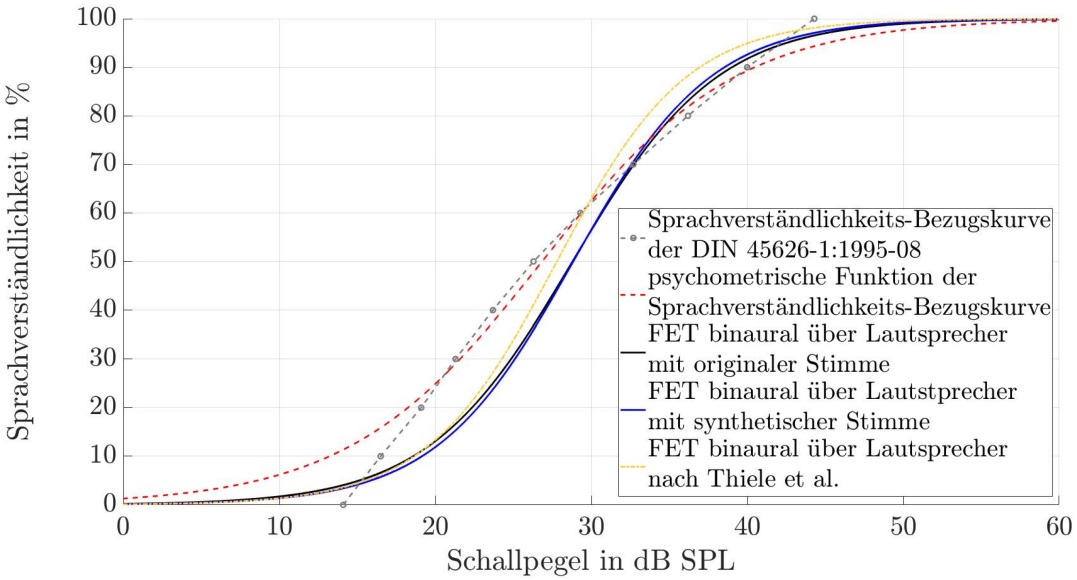

Abbildung 3: Vergleich der ermittelten psychometrischen Funktionen für den FET mit originalem und synthetischem Sprachmaterial mit der psychometrischen Funktion, die an die Sprachverständlichkeitsbezugskurve aus DIN 45626-1 [22] angepasst wurde und einer binaural über Lautsprecher aufgenommen psychometrischen Funktion aus einer Studie nach Thiele et al. [26].

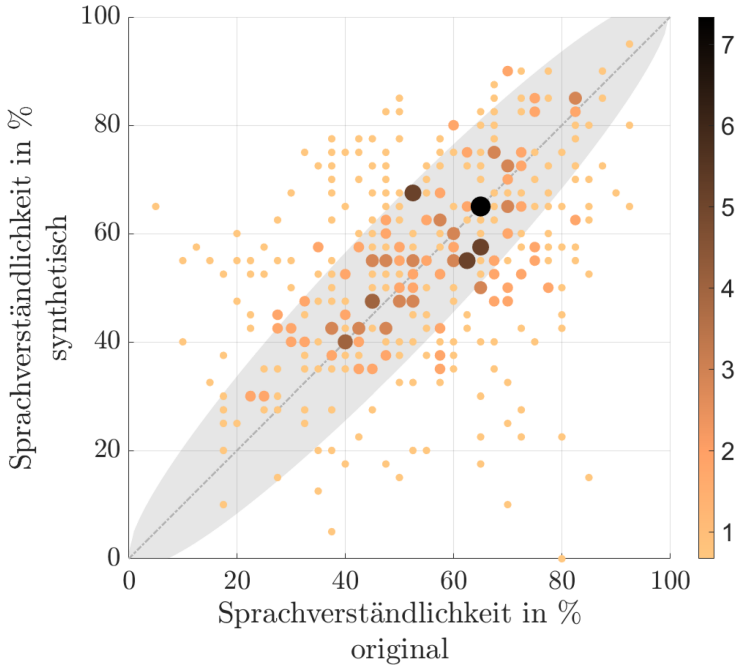

Abbildung 4: Vergleich der Sprachverständlichkeit pro Wort über die Schallpegel gemittelt im Streudiagramm; Farbskala und Punktgröße zeigen die Anzahl der aufeinanderliegenden Punkte des Diagramms an. Der grau eingefärbte Bereich stellt die zufällige zweifache Standardabweichung der Binomialverteilung dar.

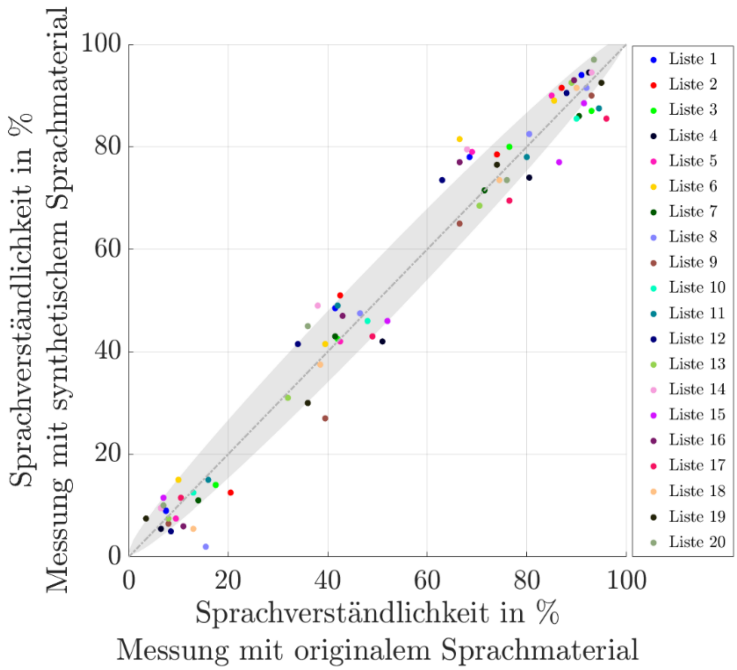

Abbildung 5: Vergleich der Sprachverständlichkeit resultierend aus den Listen im FET mit originaler und synthetischer Stimme im Streudiagramm über die vier Messpegel 20 dB, 27 dB, 34 dB und 41 dB. Der grau eingefärbte Bereich stellt die zufällige zweifache Standardabweichung der Binomialverteilung dar.

schallpegeln im Vergleich zum durchschnittlichen Sprachverstehen der Listen des FET mit synthetischer Stimme in den vier Messschallpegeln als Streudiagramm dargestellt. Der Vergleich der aus den Listen resultierenden Sprachverständlichkeitswerte für die verschiedenen Messschallpegel mit originaler und synthetischer Stimme zeigt mit einem Korrelationskoeffizienten nach Spearman von r_s=0,951 mit p<0,001 eine hochsignifikante, stark positive Korrelation an.

Für jeden der vier Messschallpegel ergeben sich im Diagramm Cluster aus den Verständlichkeitswerten der 20 Listen. Bei der statistischen Untersuchung wurden nichtparametrische Tests verwendet, da der Shapiro Wilk-Test bei einem Messschallpegel von 27 dB nicht normalverteilte Daten anzeigte (p=0,019). Signifikante Unterschiede im Median der Cluster im Vergleich pro Messschalldruckpegel lassen sich nicht ermitteln (p_{20dB}=0,330, p_{27dB}=0,371, p_{34dB}=0,184, p_{41dB}=0,684).

Die in grau eingefärbte doppelte Standardabweichung der Binomialverteilung zeigt die zufällige Streuung der Verständlichkeitswerte für die 20 Listen mit 20 Wörtern, die unterteilt in die vier Messchallpegel über 40 Probanden ermittelt wurden.

Diskussion

Psychometrische Funktionen

Zur Validierung der Messergebnisse wird die mittlere psychometrische Funktion für das originale Sprachmaterial mit der Sprachverständlichkeitsbezugskurve der DIN 45626-1 [22], wie in Abbildung 3 dargestellt, verglichen. Die psychometrische Funktion zum originalen Sprachmaterial verläuft im SRT zu höheren Pegeln verschoben und insgesamt steiler als die psychometrische Funktion der Sprachverständlichkeitsbezugskurve. Auch in einer Studie von Thiele et al. [26] wurden bereits bei der Überprüfung von Ruhebezugskurven im Freifeld im SRT um 1,5 dB zu höheren Schallpegeln verschobene und um 1,6%/dB steiler verlaufende psychometrische Funktionen beobachtet. Diese ähneln den in dieser Studie ermittelten Kurven in der Abweichung sehr. Die Differenzen lassen sich wahrscheinlich auf Abweichungen zwischen Freifeldmessungen und Messungen mit freifeldentzerrtem Kopfhörer zurückführen.

Die mit originaler und synthetischer Stimme ermittelten psychometrischen Funktionen dieser Studie unterscheiden sich in der Sprachverständlichkeitsschwelle nicht signifikant voneinander, jedoch in der Steigung. Die Ergebnisse sind vergleichbar gut mit bisherigen Ergebnissen von Sprachtests, die bei normalhörenden Probanden mit synthetischer Stimme durchgeführt wurden, wie dem OLSA mit weiblicher synthetischer Stimme, der von Nuesse et al. [20] beschrieben wurde. In der Studie wurden ebenfalls aus Probandenmessungen ermittelte psychometrische Funktionen für den Sprachtest untersucht. Über alle Probanden zeigte sich für das gesamte Testmaterial eine Verschiebung des gemittelten SRT-Werts in der Kondition mit synthetischer Stimme um 0,5 dB zu höherem SNR (Signal-to-Noise-Ratio). Die Steigung der psychometrischen Funktion für das synthetisch erzeugte Sprachmaterial verlief ähnlich wie auch beim FET mit synthetischer Stimme um 0,3%/dB steiler im Vergleich zur psychometrischen Funktion für das Originalmaterial [20].

Diese gute Übereinstimmung des Verlaufs von den psychometrischen Funktionen zwischen originalem und synthetischem Sprachmaterial erscheint erstmal überraschend, da bis auf die Auswahl einer ähnlichen Stimme sowie die Anpassung der Wortlänge und des mittleren Wortschallpegels keine weiteren Maßnahmen zum Abgleich des Testmaterials vorgenommen wurden. So hätten beispielsweise Parameter wie der Frequenzumfang der Stimme sowie Vokalabstände, wie in [24] beschrieben, oder auch weitere spektral-zeitliche Modulationen und Feinstrukturen von Sprache Einfluss auf die Sprachverständlichkeit haben können.

Wortverständlichkeit

Die Ergebnisse der Einzelwortverständlichkeit können als das kleinste gemessene Maß Unterschiede für die Sprachverständlichkeit zwischen den beiden Sprachmaterialien im Detail aufzeigen. Die in Abbildung 4 eingetragene doppelte Standardabweichung der Binomialverteilung gibt die Streuung für die über 40 Probanden gemittelten Wortverständlichkeitswerte an. Wortverständlichkeiten, die außerhalb des grauen Bereiches liegen, sind daher eher auf die Unterschiede zwischen originalem und synthetischem Sprachmaterial als auf statistische Schwankungen zurückzuführen. Die sich über den Bereich der Sprachverständlichkeit abbildende unsymmetrische Streuung der einsilbigen Worte deutet auf einen grundsätzlichen Einfluss der synthetischen Stimme auf die Sprachverständlichkeit hin, da die Streuung im Bereich der schlechter verstandenen Wörter mit synthetischer Stimme größer ist. Ein Reihenfolgeeffekt lässt sich an dieser Stelle aufgrund der durchgeführten Randomisierung ausschließen. Für den gesamten Test betrachtet gleichen sich die Unterschiede in der Sprachverständlichkeit gut aus, sodass sich die psychometrischen Funktionen für beide Sprachmaterialien stark ähneln. Auffällig sind im Streudiagramm Abweichungen von einzelnen Einsilbern des synthetischen Sprachmaterials in Richtung schlechterer Sprachverständlichkeit von bis zu 80 Prozentpunkten („Hemd“, Liste 6; „Hecht“, Liste 7 und „Rind“, Liste 9). Bei genauerer Untersuchung der jeweiligen Teststimuli musste festgestellt werden, dass es sich bei den Wörtern um diejenigen handelte, die zwar phonetisch korrekt synthetisiert wurden, jedoch aus subjektiver Sicht unnatürlich und teils verzerrt klangen. Für eine Durchführung des Sprachtests in der Praxis wären einzelne Wörter des synthetisch erzeugten Sprachkorpus demzufolge nicht einzusetzen. Für die schlechte Synthesequalität dieser einzelnen Wörter kommen unterschiedliche Gründe in Frage. Eine mögliche Erklärung ist hier die vermutlich hohe Anzahl an Freiheitsgraden im Synthesesystem, speziell bei der Synthese von einsilbigen Worten im Unterschied zu ganzen Sätzen oder zusammenhängenden Texten. Hier könnte ein Erzeugen des Wortes im ganzen Satz die Natürlichkeit in der Aussprache des Synthesesystems positiv beeinflussen. Außerdem könnte auch eine zu starke Verlangsamung, die zur Anpassung an das originale Sprachmaterial teilweise durchgeführt werden musste, Grund für eine Minderung der Verständlichkeit eines synthetisierten Wortes sein. Gegenüber den gehäuften Abweichungen in negative Richtung steht mit einer Abweichung in positive Richtung von 60 Prozentpunkten nur ein Wort („Draht“, Liste 6).

Listenverständlichkeit

Eine ausgeglichene Verständlichkeit der Listen des FETs ist für die Vergleichbarkeit der Testlisten von großer Bedeutung. So haben diese Werte in der Praxis eines Sprachtests unmittelbaren Einfluss auf dessen Genauigkeit. Die verschiedenen Listen dürfen bei demselben Schallpegel gemessen zu keinen bedeutenden Unterschieden bei der Sprachverständlichkeit führen. Beim Vergleich der Sprachverständlichkeitswerte der Listen bei den vier gemessenen Schallpegeln haben sich zwischen dem synthetischen und dem originalen Sprachmaterial keine signifikanten Unterschiede in der Lage der Mittelwerte ergeben. Sprachverständlichkeitswerte in Abbildung 5, die außerhalb der in grau eingetragenen zufälligen Streuung der Binomialverteilung liegen, können als überzufällige Abweichungen betrachtet werden. Die in der Sprachverständlichkeit abweichenden Listen variieren jedoch je nach Pegel, sodass sich nur schwierig Rückschlüsse auf bestimmte Listen ziehen lassen, die durch Wortmaterial oder Synthesequalität signifikant schlechter oder besser verstanden werden.

Fazit

In dieser Studie wurde der FET mit originalem und synthetischem Sprachmaterial vergleichend bei 40 Probanden zur Ermittlung von Sprachverständlichkeitsschwellen in Ruhe durchgeführt. Die Einsilber des synthetischen Sprachmaterials wurden dafür in Sprechgeschwindigkeit und mittlerem Schalldruckpegel pro Wort an das originale Material angepasst. Die für die Sprachkorpora über den gesamten Test ermittelten psychometrischen Funktionen zeigen sehr ähnliche Verläufe im SRT und der Steigung, wobei die psychometrische Funktion für das synthetische Sprachmaterial mit $s_{synth.}$=5,67%/dB gegenüber $s_{original}$=5,38%/dB beim originalen Sprachmaterial signifikant steiler verläuft. Die praktische Bedeutung dieses Unterschieds wird jedoch als gering bewertet (insbesondere beim Vergleich in Abbildung 3). Im Hinblick auf Steigung und Sprachverständlichkeitswerte ergaben sich ähnliche Ergebnisse wie in einer Studie über den OLSA mit synthetischer Stimme von Nuesse et al. [20].
Bei den Sprachverständlichkeitswerten treten für einzelne Worte des synthetischen Sprachkorpus Ausreißer zu negativerer Sprachverständlichkeit auf. Im Mittel gleichen sich die Abweichungen jedoch über die Listen und den gesamten Test hinweg wieder aus, sodass die Ergebnisse bei Durchführung des FET mit synthetischer Stimme vergleichbar zu Ergebnissen bei Durchführung des FET mit originalem Sprachmaterial liegen. Vor einem Einsatz in

der Praxis sollten diese Negativausreißer zunächst korrigiert werden.
Es bleibt die Frage offen, inwieweit sich die mit Normalhörenden gesammelten Resultate in Bezug auf die Sprachverständlichkeit der synthetischen Stimme auch auf Menschen mit verringertem Hörvermögen übertragen lassen.
Die Studie hat gezeigt, dass eine Entwicklung des Freiburger Einsilbertests mit synthetischer Stimme sinnvoll möglich ist und zu vergleichbaren Ergebnissen führt wie eine Durchführung des Sprachtests mit dem originalen Sprachmaterial. Bei der Produktion der Wörter muss jedoch auf die Natürlichkeit und grundsätzliche Verständlichkeit der Wörter geachtet werden, sodass Negativausreißer im Test vermieden werden können. Insgesamt deuten die Ergebnisse darauf hin, dass sich Streuungen der Sprachverständlichkeitsschwellen von Wörtern über eine Liste hinweg sehr gut ausgleichen, sodass eine Durchführung von Sprachtests mit synthetisch erzeugtem Sprachmaterial in Zukunft als eine sinnvolle Möglichkeit erscheint.

Anmerkungen

Interessenkonflikte

Die Autoren erklären, dass sie keine Interessenkonflikte im Zusammenhang mit diesem Artikel haben.

Literatur

1. Steffens T. 25 Hörgeräteversorgung. In: Strutz J, Mann W, editors. Praxis der HNO-Heilkunde, Kopf- und Halschirurgie. Stuttgart: Thieme; 2009.
2. G-BA. Richtlinie des Gemeinsamen Bundesausschusses über die Verordnung von Hilfsmitteln in der vertragsärztlichen Versorgung. 2020.
3. Deutsches Institut für Normung e.V. DIN 45621-1:1995-08, Sprache für Gehörprüfung – Teil 1: Ein- und mehrsilbige Wörter. Berlin: Beuth; 1995.
4. Bangert H. Probleme bei der Ermittlung des Diskriminationsverlustes nach dem Freiburger Sprachtest. Audiologische Akustik. 1980;19:166-70.
5. Alich G. Anmerkungen zum Freiburger Sprachverständnistest (FST). Sprache, Stimme, Gehör. 1985; 9:1-6.
6. von Wedel H. Untersuchungen zum Freiburger Sprachtest – Vergleichbarkeit der Gruppen im Hinblick auf Diagnose und Rehabilation (Hörgeräteanpassung und Hörtraining). Audiologische Akustik. 1986;25:60-73.
7. Deutsches Institut für Normung e.V. DIN EN ISO 8253-3:2012-08, Akustik –Audiometrische Prüfverfahren – Teil 3: Sprachaudiometrie (ISO 8253-3:2012), Deutsche Fassung EN ISO 8253-3:2012. Berlin: Beuth; 2012.
8. Winkler A, Holube I. Der Freiburger Einsilbertest und die Norm DIN EN ISO 8253-3: Technische Analyse. Z Audiol. 2016;55(3):106-13.
9. Steffens T. Verwendungshäufigkeit der Freiburger Einsilber in der Gegenwartssprache: Aktualität der Testwörter. HNO. 2016 Aug;64(8):549-56. DOI: 10.1007/s00106-016-0163-5
10. Winkler A, Carroll R, Holube I. Impact of Lexical Parameters and Audibility on the Recognition of the Freiburg Monosyllabic Speech Test. Ear Hear. 2020 Jan/Feb;41(1):136-42. DOI: 10.1097/AUD.0000000000000737
11. Hahn F. Freiburger reloaded [Bachelorarbeit]. Aaalen: Hochschule Aalen; 2014.
12. Hahlbrock KH. Sprachaudiometrie: Grundlagen und praktische Anwendung einer Sprachaudiometrie für das deutsche Sprachgebiet. Stuttgart: Thieme; 1957.
13. Brinkmann K. Die Neuaufnahme der „Wörter für Gehörprüfung mit Sprache“. Zeitschrift für Hörgeräteakustik. 1974;13: 14-40.
14. Mahfoud M. Neuaufsprache und Evaluation des Einsilber-Sprachverständnistests [Dissertation]. Würzburg: Julius-Maximilians-Universität Würzburg; 2009.
15. Qualen JF. Evaluation des Einsilber-Sprachmaterials M-2007 [Dissertation]. Würzburg: Julius-Maximilians-Universität Würzburg; 2010. DOI: 10.28937/1000107838
16. Bechmann S. Sprachwandel – Bedeutungswandel. Stuttgart: UTB GmbH; 2016. DOI: 10.36198/9783838545363
17. Cohn M, Zellou G. Perception of concatenative vs. neural text-to-speech (TTS): Differences in intelligibility in noise and language attitudes. Proc. Interspeech. 2020:1733-7. DOI: 10.21437/Interspeech.2020-1336
18. Benoit C, Grice M, Hazan V. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. Speech Commun. Jun 1996;18(4):381-92. DOI: 10.1016/0167-6393(96)00026-X
19. Valentini-Botinhao V, Toman M, Pucher M, Schabus D, Yamagishi J. Intelligibility of time-compressed synthetic speech: Compression method and speaking style. Speech Commun. 2015 Nov;74:52-64. DOI: 10.1016/j.specom.2015.09.002
20. Nuesse T, Wiercinski B, Brand T, Holube I. Measuring Speech Recognition With a Matrix Test Using Synthetic Speech. Trends Hear. 2019 Jan-Dec;23:2331216519862982. DOI: 10.1177/2331216519862982
21. Karl J. Investigation of the influence of pitch and speed for synthetic speech on intelligibility of the Freiburg monosyllabic speech test. Lübeck: Technische Hochschule Lübeck; 2021.
22. Deutsches Institut für Normung e.V. DIN 45626-1:1995-08, Tonträger mit Sprache für Gehörprüfung – Teil 1: Tonträger mit Wörtern nach DIN 45621-1 (Aufnahme 1969). Berlin: Beuth; 1995.
23. Williamson DG, Harmon-Smith A. Der Einfluß der Grundfrequenz auf die Sprachverständlichkeit. Zeitschrift für Audiologie. 1980;19(6):236-40.
24. Bradlow AR, Torretta GM, Pisoni DB. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. Speech Commun. 1996 Dec;20(3-4):255-72. DOI: 10.1016/S0167-6393(96)00063-5
25. Deutsches Institut für Normung e.V. DIN EN ISO 8253-2:2010-07, Akustik- Audiometrische Prüfverfahren – Teil 2: Schallfeld-Audiometrie mit reinen Tönen und schmalbandigen Prüfsignalen (ISO 8253-2:2009) Deutsche Fassung EN ISO 8253-2:2009. Berlin: Beuth; 2010.
26. Thiele C, Wardenga N, Lenarz T, Büchner A. Überprüfung der Vergleichbarkeit von Freifeld- und HDA200-Kopfhörermessungen für den Freiburger. HNO. 2014 Feb;62(2):115-20. DOI: 10.1007/s00106-013-2789-x

Korrespondenzadresse:
Thomas Schwarz
Deutsches Hörgeräte Institut GmbH, Anschützstraße 1, 23562 Lübeck, Deutschland
thschwarz11@gmail.com

Bitte zitieren als
Schwarz T, Frenz M, Bockelmann A, Husstedt H. Untersuchung einer synthetischen Stimme für den Freiburger Einsilbertest. GMS Z Audiol (Audiol Acoust). 2022;4:Doc04.
DOI: 10.3205/zaud000022, URN: urn:nbn:de:0183-zaud0000224

Artikel online frei zugänglich unter
https://doi.org/10.3205/zaud000022

Veröffentlicht: *08.06.2022*