HOCHSCHULE
DER MEDIEN

# Masterarbeit

im Studiengang Audiovisuelle Medien

Theoretische Ausarbeitung eines Programmtools zur Sprachverständlichkeitsanalyse von Sprachsignal-Audiodateien aus dem Broadcastumfeld

vorgelegt von Elias Thomas Weißenrieder

an der Hochschule der Medien Stuttgart am 26.09.2024

zur Erlangung des akademischen Grades eines Master of Engineering

Erstprüfer: Prof. Oliver Curdt

Zweitprüfer: Prof. Dr. Frank Melchior

# Ehrenwörtliche Erklärung

Hiermit versichere ich, Elias Thomas Weißenrieder, ehrenwörtlich, dass ich die vorliegende Masterarbeit mit dem Titel: “Theoretische Ausarbeitung eines Programmtools zur Sprachverständlichkeitsanalyse von Sprachsignal-Audiodateien aus dem Broadcastumfeld“ selbstständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Ebenso sind alle Stellen, die mithilfe eines KI-basierten Schreibwerkzeugs erstellt oder überarbeitet wurden, kenntlich gemacht. Die Arbeit ist noch nicht veröffentlicht oder in anderer Form als Prüfungsleistung vorgelegt worden.

Ich habe die Bedeutung der ehrenwörtlichen Versicherung und die prüfungsrechtlichen Folgen (§ 23 Abs. 2 Master-SPO (Vollzeit)) einer unrichtigen oder unvollständigen ehrenwörtlichen Versicherung zur Kenntnis genommen.

Wadern, 25.09.2024, E. Weißenrieder

# Kurzfassung

In dieser Masterarbeit wird die theoretische Ausarbeitung eines Programmtools zur Messung der Sprachverständlichkeit von Audiodateien aus dem Broadcastumfeld beschrieben. Es wird erforscht, welches Sprachverständlichkeitsmessverfahren für ein solches Programm am besten geeignet ist, und wie dieses die Ergebnisse der Analysen – möglichst verständlich – kategorisieren und darstellen kann. Für die Auswahl des geeigneten Messverfahrens werden zunächst eine Reihe an bekannten Verfahren aufgeführt, welche definiert und erklärt werden. Im Anschluss daran folgt die Einzelbetrachtung jedes Messverfahrens bezüglich seiner Vor- und Nachteile sowie seines möglichen Anwendungsgebietes. Um aus der dadurch entstehenden Vorauswahl von drei Messalgorithmen den passenden für das Programm auszuwählen, wird ein Versuch ausgeführt. Bei diesem werden die Ergebnisse der Algorithmen mit dem Ergebnis eines Hörtests verglichen. Im Anschluss an diesen Versuch werden unter anderem durch die Antworten eines Nutzerinterviews die Funktionen des Programmtools beschrieben, die notwendig sind, damit es Übersichtlichkeit aufweist und somit anwenderfreundlich ist.

# Abstract

In this master thesis, the theoretical development of a software tool for measuring the speech intelligibility of broadcast audio files is described. The research focuses on determining which speech intelligibility measurement method is best suited for such a tool and how the analysis results can be categorized and presented in the most comprehensible way. To select the appropriate measurement method, a number of well-known methods are first presented, defined, and explained. This is followed by an individual examination of each method with regard to its advantages, disadvantages, and potential application areas. To select the most suitable method from the resulting shortlist of three measurement algorithms, an experiment is conducted. In this experiment, the results of the algorithms are compared with the outcome of a listening test. Following this experiment, the necessary functions of the program are described, which are essential for ensuring clarity and user-friendliness. This is supported in part by responses from an user interview.

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# Formelverzeichnis

# Abkürzungsverzeichnis

| | |
|---|---|
| **AI** | **Artikulationsindex** |
| **AIcons** | **Articulation loss of consonants** |
| **ANSI** | **American National Standards Institute** |
| **ARD** | **Arbeitsgemeinschaft der öffentlich-rechtlichen Rundfunk-anstalten der Bundesrepublik Deutschland** |
| **Atmo** | **Atmosphäre** |
| **AutoMOS** | **Automatic Mean-Opinion-Score** |
| **CCP** | **Commentary-Control-Panel** |
| **CSV** | **Comma-Separated Values** |
| **DAW** | **Digital-Audio-Workstation** |
| **FFT** | **Fast Fourier Transform** |
| **(G)UI** | **(Graphical) User Interface** |
| **HD** | **High Definition** |
| **Hz** | **Hertz** |
| **IIS** | **Fraunhofer-Institut für Integrierte Schaltungen** |
| **ITU** | **International Telecommunication Union** |
| **KI** | **Künstliche Intelligenz** |
| **LU(FS)** | **Loudness Units (relative to Full Scale)** |
| **MOS** | **Mean-Opinion-Score** |
| **MOV** | **Modell Output Variables** |
| **ms** | **Millisekunden** |

| | |
|---|---|
| **NISQA** | **Non-Intrusive Speech Quality Assessment** |
| **ODG** | **Objective Difference Grade** |
| **ORF** | **Österreichischer Rundfunk** |
| **ORTF** | **Office de Radiodiffusion Télévision Française** |
| **PEAQ** | **Perceptual Evaluation of Audio Quality** |
| **PESQ** | **Perceptual Evaluation of Speech Quality** |
| **POLQA** | **Perceptual Objective Listening Quality Analysis** |
| **RASTI** | **Rapid Speech Transmission Index** |
| **SII** | **Speech Intelligibility Index** |
| **STI** | **Speech Transmission Index** |
| **STI-PA** | **Speech Transmission Index for Public Address Systems** |
| **STOI** | **Short-Time Objective Intelligibility** |
| **TV** | **Television** |
| **USB** | **Universal Serial Bus** |
| **WAV** | **Waveform Audio File Format** |
| **ZDF** | **Zweites Deutsches Fernsehen** |

# 1 Einleitung

Innerhalb seiner Signalkette durchläuft mittlerweile nahezu jedes Audiosignal eine digitale Verarbeitung. Die Wahrscheinlichkeit, eine ausschließlich analoge Audiosignalkette vorzufinden, ist demnach sehr gering. Meist wird das Signal im Verlauf der Verarbeitung mindestens einmal Analog-Digital bzw. Digital-Analog gewandelt. Dies führt dazu, dass die moderne Audiotechnik sich mit den Gegebenheiten der Digitaltechnik auseinandersetzen muss. Ein besonderes Augenmerk liegt dabei auf der digitalen Übertragung von Audiosignalen. Speziell die drahtlose Übertragungsart ist meist auf niedrige Datenraten der Signale angewiesen.

In vielen Fällen ist gesprochene Sprache ein inhaltlicher Hauptbestandteil der Audiosignale, welche drahtlos übertragen werden sollen. Hier seien beispielsweise die Anwendung von Betriebsfunk oder anderer Kommunikationsmöglichkeiten erwähnt. Zweifelsohne ist der wichtigste Aspekt bei der Übertragung von gesprochener Sprache die Verständlichkeit am Ende der Übertragungskette. Diese Verständlichkeit eines Signals hängt unter anderem von seiner Audioqualität ab. So hat beispielsweise eine Audiodatei mit einer hohen Qualität eine bessere Sprachverständlichkeit als eine Datei mit einer niedrigen Audioqualität. Die Audioqualität eines Signals wiederum wird zum Großteil durch seine Datenrate bestimmt. Vereinfacht gesagt gilt: Je höher die Audioqualität sein soll, desto höher muss die Datenrate der Datei sein.

Dies führt dazu, dass bei der drahtlosen Übertragung von Sprachsignalen meist ein Kompromiss eingegangen werden muss. Denn eine maximal mögliche Verständlichkeit fordert eine hohe Audioqualität und demnach eine große Datenrate. Eine hohe Datenrate technisch zu realisieren ist jedoch nicht sinnvoll, daher ist sie meist begrenzt. Um trotz geringer Datenrate eine möglichst gute Sprachverständlichkeit bei der Übertragung zu gewähren, werden unter anderem Audiokodierungsverfahren angewendet. Die Forschung nach der Frage, wie diese Verfahren eine möglichst perfekte Verständlichkeit bieten können, ist von großer Wichtigkeit und in der heutigen Zeit noch immer aktuell.

Unter anderem wäre für diese Forschungen ein Programm von Nutzen, welches automatisiert Audiodateien hinsichtlich ihrer Sprachverständlichkeit analysieren kann und die Ergebnisse dieser Analyse anschaulich darstellt und kategorisiert. Die theoretische Ausarbeitung der nötigen Funktionen eines solchen Programms sowie die Auswahl des geeigneten Analysealgorithmus werden Inhalt dieser Arbeit sein.

## 1.1 Bedeutung für die Fachwelt

Die konzeptionelle Ausarbeitung der Programmfunktionen in dieser Arbeit könnte der Fachwelt das Gerüst eines zuverlässigen Messinstruments für Sprachverständlichkeit liefern. Dies hätte den Vorteil, dass zukünftig schnell und automatisiert Messungen zur Verständlichkeit durchgeführt werden können. Die Möglichkeit, dies mittels eines automatisierten Tools und nicht mehr beispielsweise mittels Hörtests zu tun, könnte sich positiv auf die Geschwindigkeit der Entwicklung der Sprachverständlichkeitsforschung auswirken. Ebenso würde das Programm den Vergleich von mehreren Audiodateien hinsichtlich ihrer Verständlichkeit erleichtern. Somit könnten die bereits durch Hörtests gesammelten Ergebnisse der Forschung auch mittels des Programms bestätigt und in vergleichbaren Zahlen dargestellt werden.

Des Weiteren kann durch die Ausarbeitung der verschiedenen Algorithmen zur Sprachverständlichkeitsanalyse (siehe Kapitel 3 (Grundlagen und Definitionen)) eine Übersichtlichkeit dargestellt werden. Zusätzlich bietet Kapitel 4 (Vergleich der Sprachverständlichkeitsmessverfahren) die Möglichkeit, die Vor- und Nachteile der einzelnen Algorithmen anschaulich darzustellen und zu vergleichen. Infolgedessen kann dieser Teil der Arbeit als Ratgeber hinsichtlich des Verwendungszwecks des jeweiligen Analysealgorithmus verstanden werden.

## 1.2 Gesellschaftliche Relevanz der Forschung

Wie bereits in Kapitel 1.1 (Bedeutung für die Fachwelt) erwähnt, könnte das Ergebnis der Arbeit dazu beitragen, schneller Audiodateien hinsichtlich ihrer Sprachverständlichkeit zu untersuchen. Dadurch besteht die Möglichkeit, dass die Entwicklung besserer Sprachkodierungsverfahren beschleunigt wird und sich dies im Endeffekt positiv auf die Verständlichkeit von aktuellen Sprachsignalen auswirkt.

## 1.3 Zielstellung

Ziel der Arbeit ist es, die Funktionen eines Programms, welches eine Sprachverständlichkeitsanalyse ausführen kann, auszuarbeiten und einen geeigneten Analysealgorithmus für das Programm vorzuschlagen. Es wird darauf geachtet, dass auch die Nutzerfreundlichkeit des ausgearbeiteten Tools nicht vernachlässigt wird. Das Hauptanwen-

dungsgebiet des endgültigen Programms soll dabei im TV-Broadcastumfeld liegen. Um dieses Ziel zu erreichen, sollen zunächst die in Kapitel 3 (Grundlagen und Definitionen) aufgelisteten vorhandenen Algorithmen zur Sprachverständlichkeitsanalyse theoretisch betrachtet und erklärt werden. Der Vergleich dieser Messverfahren soll dann in Kapitel 4 (Vergleich der Sprachverständlichkeitsmessverfahren) ausgeführt und für jeden Algorithmus eine Empfehlung bezüglich seines Anwendungsbereichs abgegeben werden. Des Weiteren soll die Frage nach dem geeigneten Analysealgorithmus für das Programm unter anderem mittels eines Versuchs in Kapitel 5 (Versuch zur Auswahl des geeigneten Analysealgorithmus) beantwortet werden. Die endgültige Auswahl des Algorithmus wird dann in Kapitel 6 (Theoretische Ausarbeitung des Programms) erläutert. Des Weiteren soll in diesem Teil der Arbeit das restliche Programmgerüst ausgearbeitet werden. Hierzu wird aufgezeigt, wie die Analysedaten verarbeitet, kategorisiert und dargestellt werden. Außerdem wird in diesem Abschnitt auf die Nutzeroberfläche und deren Aussehen eingegangen. Diese Ausarbeitungen sind dabei unter anderem begründet und evaluiert durch die Ergebnisse eines durchgeführten Nutzerinterviews.

# 2 Darstellung der Hypothesen

Aus der Zielstellung ergeben sich unter anderem zwei Forschungsfragen: (1) Welches Sprachverständlichkeitsmessverfahren liefert als Implementierung in dem Programmtool die präzisesten Analyseergebnisse? (2) Wie können die Analyseergebnisse des Tools dem Anwender bzw. der Anwenderin übersichtlich dargestellt werden? Die in diesem Kapitel ausgeführten Hypothesen antizipieren die Antworten auf diese Fragen.

## 2.1 Präzisester Analysealgorithmus für das Programm

Die erste Forschungsfrage, welcher Analysealgorithmus die größte Genauigkeit bietet, lässt sich mit folgender Hypothese beantworten. Das Short-Time Objective Intelligibility-Verfahren scheint die exaktesten Ergebnisse der Verständlichkeitsanalyse auszugeben. Dies liegt zum einen daran, dass es als etablierter Messstandard für Sprachverständlichkeit gilt, und zum anderen hat es gegenüber den anderen Messverfahren den Vorteil, dass es die Audiosignale blockweise, also in Segmenten, hinsichtlich ihrer Sprachverständlichkeit analysiert werden. Deswegen ist zu erwarten, dass es auch genauere Ergebnisse der Verständlichkeitsbewertung generiert.

## 2.2 Darstellungsform der Analyseergebnisse des Programms

Die zweite Forschungsfrage, wie die Ergebnisse der Sprachverständlichkeitsbewertung dem Anwender bzw. der Anwenderin übersichtlich dargestellt werden können, lässt sich mit folgender Hypothese beantworten. Um die Analyseergebnisse für den Nutzer bzw. die Nutzerin anschaulich zu präsentieren und gleichzeitig Übersichtlichkeit zu garantieren, muss das Tool eine einfache grafische Darstellung der Ergebnisse in Form eines Balkendiagramms besitzen. Die einzelnen Balken des Diagramms sollen dabei die Sprachverständlichkeitswerte der einzelnen Audiosignale angeben. Die Auswahl dieser Darstellungsform liegt darin begründet, dass so die Verständlichkeitswerte mehrerer Dateien auf einen Blick durch den Anwender bzw. die Anwenderin verglichen werden können.

# 3 Grundlagen und Definitionen

In diesem Kapitel werden die wichtigsten Begriffe, die für das Verständnis und die Umsetzung der Zielstellung der Arbeit wichtig sind, definiert und erklärt. Außerdem enthält Kapitel 3.3 (Darstellung verschiedener Sprachverständlichkeitsmessverfahren) die Vorstellung bekannter Messverfahren der Verständlichkeit bzw. Sprachqualität.

## 3.1 Sprachverständlichkeit

Die Möglichkeit eines Hörers, Silben, Wörter und Sätze zu verstehen, wird als Sprachverständlichkeit bezeichnet.[1] Zwar sei sie auch abhängig vom Sichtkontakt zu einem Sprecher bzw. einer Sprecherin, allerdings ist dies nicht messtechnisch erfassbar.[2] Deswegen wird diesem Aspekt in dieser Arbeit keine weitere Bedeutung beigemessen werden. Sprachverständlichkeit ist ein grundlegendes Kriterium für die Beurteilung der tontechnischen Übertragungsqualität.[3] Sie lässt sich unterteilen in die Silbenverständlichkeit, die Wortverständlichkeit und die Satzverständlichkeit. Die Wortverständlichkeit liegt dabei stets über der Silbenverständlichkeit und die Satzverständlichkeit über der Wortverständlichkeit.[4] Somit ergibt sich, dass eine 100%ige Satzverständlichkeit bereits bei einer 80%igen Silbenverständlichkeit gegeben ist.[5] Dies rührt daher, dass nicht verstandene Silben und Wörter im Satzzusammenhang, also aus dem Kontext, ergänzt werden können.[6] In diesem Falle ist die Angabe in Prozent als Verhältnis zwischen richtig verstandenen und insgesamt gesendeten Silben, Wörtern oder Sätzen zu verstehen.[7] Das Sprachspektrum beschränkt sich in seinen wesentlichen Komponenten hauptsächlich auf den Frequenzbereich zwischen 100 Hz und etwa 8 kHz.[8] Innerhalb dieses Bereichs spielt das Oktavband um 2 kHz eine besonders dominante Rolle, denn ein Drittel der Sprachinformationen ist hier zu verorten.[9] In den Oktavbändern 1 kHz, 2 kHz sowie 4 kHz sind drei Viertel der Sprachinformationen und unterhalb 1 kHz nur noch ein Viertel zu lokalisieren.[10] Innerhalb einer Räumlichkeit ist die Sprachverständlichkeit abhängig von der Nachhallzeit, den frühen Reflexionen, im Bereich unter 1 ms sowie den späten

[1] Vgl. Friesecke (2014, S. 869).
[2] Vgl. Friesecke (2014, S. 869).
[3] Vgl. Dickreiter, Dittel, Hoeg und Wöhr (2014, S. 72).
[4] Vgl. Dickreiter et al. (2014, S. 72).
[5] Vgl. Dickreiter et al. (2014, S. 72).
[6] Vgl. Dickreiter et al. (2014, S. 72).
[7] Vgl. Dickreiter et al. (2014, S. 72).
[8] Vgl. Dickreiter et al. (2014, S. 1248).
[9] Vgl. Friesecke (2014, S. 101).
[10] Vgl. Friesecke (2014, S. 101).

Reflexionen ab 100 ms, dem Verhältnis zwischen Direktschall und Diffus Schall, dem Ruhegeräusch des Raumes und möglichen Kammfiltereffekten durch mehrere Schallquellen.[11] Auch bedingt der Störabstand zwischen dem Nutzsignal und den Hintergrund- bzw. Störgeräuschen die Wortverständlichkeit und dadurch auch die Sprachverständlichkeit.[12]

## 3.2 Betriebsarten der Signalübertragung

Prinzipiell wird in der Signal- und Kommunikationstechnik zwischen den drei Betriebsarten: Simplex-, Halbduplex- und Vollduplexbetrieb unterschieden.[13] Zur besseren Erklärung der unterschiedlichen Betriebsweisen wird in diesem Kapitel das einfachste Kommunikationsmodell, bestehend aus nur zwei Kommunikationsparteien (bezeichnet mit „A“ und „B“), verwendet. Der Simplexbetrieb, auch als Richtungsbetrieb[14] bezeichnet, lässt Kommunikation nur in eine Richtung, also nur von „A“ nach „B“ zu. Ein Beispiel, bei dem diese Betriebsart angewendet wird, ist der Rundfunk. Als Halbduplex- oder auch Wechselbetrieb[15] wird ein System bezeichnet, bei dem „A“ – anders als im Simplexbetrieb – auch an „B“ zurücksenden kann. Allerdings ist stets nur eine Übermittlung in eine Signalrichtung möglich, also entweder „A“ sendet und „B“ empfängt, oder „B“ sendet und „A“ empfängt. Als Beispiel für dieses System lässt sich normaler Betriebsfunk anführen. Eine Betriebsweise, bei der diese Einschränkung nicht besteht, wird als Vollduplex- oder Gegenbetrieb[16] bezeichnet. Es ist also bei diesem System möglich, dass „A“ simultan mit „B“ sendet und auch simultan mit „B“ empfängt. Ein Beispiel für diese Betriebsweise ist die klassische Telefonie.

## 3.3 Darstellung verschiedener Sprachverständlichkeitsmessverfahren

In den folgenden Unterkapiteln sind gängige Messverfahren zur Sprachverständlichkeit ausgeführt sowie deren Anwendungsgebiete definiert. Die Auflistung erhebt jedoch keinen Anspruch auf Vollständigkeit, sondern soll lediglich als Überblick dienen. Grund-

[11] Vgl. Friesecke (2014, S. 101).
[12] Vgl. Dickreiter et al. (2014, S. 72-73).
[13] Vgl. Nocker (2005, S. 79).
[14] Nocker (2005, S. 79).
[15] Nocker (2005, S. 79).
[16] Nocker (2005, S. 79).

sätzlich lassen sich Verständlichkeitstests in subjektive und objektive Verfahren einteilen. Dabei werden bei subjektiven Messverfahren (einer Gruppe) Hörern und Hörerinnen Hörbeispiele vorgespielt, die diese dann hinsichtlich der Sprachverständlichkeit bewerten sollen. Bei objektiven Verfahren hingegen erfolgt die Ermittlung stets ohne eine menschliche Einschätzung der Verständlichkeit, sondern mittels Software.

### 3.3.1 Mean-Opinion-Score

Der Mean-Opinion-Score, auch MOS abgekürzt, ist ein von der International Telecommunication Union (ITU) definierter Standard, der unter der Nummerierung P.800 geführt wird. Es handelt sich um ein subjektives Messverfahren von Sprachproben bezüglich ihrer „Sprachsignal-Wiedergabe-Qualität“.[17] Vielen repräsentativ ausgewählten Versuchspersonen werden diese Sprachproben vorgespielt.[18] Im Anschluss daran beurteilt jede Person die Verständlichkeit der Probe innerhalb einer fünfstufigen Skala.[19] Dabei spiegelt 5 eine exzellente Sprachsignal-Qualität mit nicht wahrnehmbaren Störeffekten wider. Stufe 4 beschreibt eine gute Qualität mit kaum wahrnehmbaren Störeffekten. Sind Störgeräusche wahrnehmbar und leicht störend, ist der Wert 3 auf der Skala zu wählen. Bei einer schlechten Sprachsignal-Qualität und einigen Störeffekten ist der Wert 2, und bei sehr vielen Störeffekten, die auch als unangenehm wahrgenommen werden, ist die unterste Stufe der Skala auszuwählen. Die durchschnittliche Bewertung aller Testpersonen ergibt den Wert des Mean-Opinion-Score.[20] Des Weiteren wird die Bewertungsskala des MOS auch verwendet, um das Ergebnis von objektiven Sprachverständlichkeitsverfahren zu kategorisieren (siehe Kapitel 3.3.9 (Perceptual Evaluation of Speech Quality)). Außerdem wird die Skala auch bei Verfahren angewandt, die die Sprachqualität mittels netzgestützter Algorithmen bewerten.[21] Diese drei aufgezählten Messvarianten, also subjektiv, objektiv und objektiv mittels netzgestützter Algorithmen, können wiederum jeweils im Schmalband oder im Breitband durchgeführt werden. Die Einteilung bezieht sich auf den Frequenzumfang der Audioanalyse und ist meist bei Schmalband auf 300 Hz bis 3,4 kHz und bei Breitband auf 50 Hz bis 7 kHz definiert.[22] Des Weiteren wird der MOS-Wert noch bezüglich der Messmethode unterschieden. So definiert die ITU hier drei Optionen, wie gemessen werden kann: (1) „conversational quality“, (2) „listening quality“ und (3) „talking quality“, auch bekannt als Interview- bzw.

[17] Vgl. Nocker (2005, S. 87).
[18] Vgl. Nocker (2005, S. 87).
[19] Vgl. Nocker (2005, S. 88).
[20] Nocker (2005, S. 87).
[21] Vgl. ITU-T P.800.1, S. 4.
[22] Vgl. ITU-T P.800.1, S. 5.

Befragungsoption.[23] Diese Unterteilung bezieht sich unter anderem auf den Umfang der Tests. So soll bei der (1) Option so gut wie möglich die Realität während des Audiotests simuliert werden. Als Beispiel sei hier ein Wählton erwähnt, welcher bei einem Hörtest, der ein Telefongespräch beinhaltet, vorangestellt wird.[24] Bei (2) Variante hingegen soll ein geringer Standard an Simulation angewendet werden.[25] Die (3) Option ist die aufwendigste Variante und soll unter anderem über die eigentliche Skala hinaus Befragungsbögen beinhalten.[26] Diese umfassen beispielsweise Fragen bezüglich der Begründung der Sprachverständlichkeit und müssen von den Probanden und Probandinnen zusätzlich ausgefüllt werden.[27] Um die Unterscheidung zwischen den einzelnen Messvarianten des MOS deutlich zu machen, empfiehlt die ITU Kennbuchstaben, die im Falle der Anwendung die Abkürzung MOS als Suffix begleiten. Die Reihenfolge, in der die einzelnen Suffixe angefügt werden, entspricht der Reihenfolge, in der sie im Folgenden ausgeführt werden. Wurde mittels „listening"-Option gemessen, ist der Kennbuchstaben „LQ"[28] anzustellen. Bei Messungen, die mittels „conversational"-Methode durchgeführt wurden, ist ein „CQ" und bei „talking"-Methoden ein „TQ" hinten anzustellen.[29] Ob die Messungen im Schmalband oder im Breitband durchgeführt wurden, ist mittels „N" für „narrowband" (Schmalband) und „W" für „wideband" (Breitband) zu kennzeichnen.[30] Wurde der MOS-Wert subjektiv gemessen, so soll dieser mit einem „S"[31] gekennzeichnet werden. Ein mittels netzgestützter Algorithmen gemessener MOS soll mit einem „E" (für „estimated") und ein objektiv evaluierter Wert mit einem „O" kenntlich gemacht werden.[32] Beispielsweise hätte ein subjektiv, mittels der „listening"-Methode im Schmalband evaluierter MOS-Wert die Suffixe „LQSN".

### 3.3.2 Articulation loss of consonants

Der Articulation loss of consonants, also der Artikulationsverlust von Konsonanten, wird auch als Alcons abgekürzt und bietet eine recht zuverlässige Messmethode für die Messung von Verständlichkeit.[33] Abermals spielt bei dieser Art der Messung der Frequenzbereich um 2 kHz eine entscheidende Rolle. Etwas weniger populär ist die Messung

[23] Vgl. ITU-T P.800.1, S. 4.
[24] Vgl. ITU-T P.800, S. 3.
[25] Vgl. ITU-T P.800, S. 3.
[26] Vgl. ITU-T P.800, S. 5.
[27] Vgl. ITU-T P.82, S. 3.
[28] ITU-T P.800.1, S. 4.
[29] Vgl. ITU-T P.800.1, S. 4.
[30] Vgl. ITU-T P.800.1, S. 4.
[31] ITU-T P.800.1, S. 4.
[32] Vgl. ITU-T P.800.1, S. 4.
[33] Vgl. Friesecke (2014, S. 870).

bezogen auf 1 kHz.[34] In einem dieser beiden Frequenzbereiche wird das Ausklingverhalten des Raumes gemessen.[35] Eine frequenzabhängige Angabe wird gewissermaßen nicht ausgeführt.[36] Die Berechnung des Alcons-Werts erfolgt mittels des Raumvolumens, der Nachhallzeit (T), des Abstandes zur Schallquelle ($r_{QH}$) und des Hallradius ($r_H$) (siehe Formel 1).[37]

$$Al_{cons} \approx 0{,}625 \left(\frac{r_{QH}}{r_H}\right)^2 * T$$

*Formel 1: Berechnung des Alcons-Werts[38].*

Der errechnete Alcons-Wert lässt sich zudem mittels einer Formel in einen RASTI-Wert (siehe Kapitel 3.3.5.1 (Rapid Speech Transmission Index)) umrechnen (siehe Formel 2).[39]

$$RASTI = 0{,}9482 - 0{,}1845 ln(AL_{cons})$$

*Formel 2: Umrechnung des Alcons-Werts in den RASTI-Wert[40].*

Auch das Alcons-Verfahren besitzt eine subjektive Bewertungsvariante ähnlich dem Mean-Opinion-Score (siehe 3.3.1 (Mean-Opinion-Score)). So ist es möglich, den Alcons-Wert zu erhalten, indem von geübten Sprechern eine Wortliste vorgelesen wird und diese im Anschluss durch Probanden und Probandinnen hinsichtlich ihrer Verständlichkeit bewertet wird.[41] Der Alcons-Wert ist stets in Prozent angegeben. Da er den Wert des Artikulationsverlustes beschreibt, entspricht ein Alcons-Wert von 0 % einer sehr guten Sprachverständlichkeit. Ab einem Wert von 33 % definiert sich eine eher unbrauchbare Verständlichkeit.[42]

### 3.3.3 Artikulationsindex

Der Artikulationsindex, auch AI abgekürzt, ist ein weiteres Maß für Sprachverständlichkeit. Er lässt sich mittels Formel (siehe Formel 3) aus dem Sprecherpegel ($L_{S(A)}$) und

[34] Vgl. Weinzierl (2008, S. 193).
[35] Vgl. Friesecke (2014, S. 870).
[36] Weinzierl (2008, S. 193).
[37] Vgl. Weinzierl (2008, S. 193).
[38] Weinzierl (2008, S. 193).
[39] Vgl. Houtgast und van Wijngaarden (2002, S. 81).
[40] Houtgast und van Wijngaarden (2002, S. 81).
[41] Vgl. Friesecke (2014, S. 870).
[42] Vgl. Friesecke (2014, S. 869).

dem Ruhegeräuschpegel ($L_{N(A)}$), jeweils in A-bewerteter Form, berechnen. Da mittels A-bewerteten Schallpegeln gemessen wird, ist eine frequenzabhängige Wichtung bereits im Ergebnis des Sprecher- und Ruhegeräuschpegels enthalten. Somit müssen nicht separat für jede Frequenz die Werte des Artikulationsindex ausgerechnet und aufsummiert werden.

$$AI = \frac{L_{S(A)} - L_{N(A)} + 12}{30dB}$$

*Formel 3: Berechnung des Artikulationsindex[43].*

Dabei entsteht als Ergebnis immer ein Wert zwischen 0 und 1. Ist dieser 0, bedeutet dies, dass keinerlei Sprachinformationen verfügbar, also hörbar bzw. nutzbar sind.[44] Ein AI-Wert von 1 bedeutet, dass alle übertragenen Sprachinformationen verfügbar sind.[45] Es ist möglich, aus den errechneten Werten eine Einschätzung der Sprachverständlichkeit abzuleiten. So besitzen ganze Sätze bei einem Artikulationsindex von 0,2 bis 0,4 eine Sprachverständlichkeit von etwa 90 %.[46] Werden jedoch Silben oder einzelne Worte gemessen, muss der AI bei über 0,7 liegen, damit 90 % Verständlichkeit vorherrscht.[47] Bei der Berechnung des Artikulationsindex ist jedoch zu beachten, dass er nur für einen Signal-Geräuschabstand von -18 dB bis 12 dB sowie für einen Dynamikbereich der Sprache von 30 dB definiert ist.[48]

### 3.3.4 Speech Intelligibility Index

Der Speech Intelligibility Index (SII) ist eine Weiterentwicklung des Artikulationsindexes und wurde 1997 vorgestellt.[49] Mithilfe verschiedener Studien von Pavlovic und Studebaker wurden die genaueren Bedingungen des Artikulationsindexes überprüft, um daraus Verbesserungen für den SII vorzuschlagen.[50] Diese wurden dann schließlich in die „American National Standards Institute“ (ANSI)-Norm S3.5 1997 eingepflegt.[51] So werden bei der Berechnung des Speech Intelligibility Indexes beispielsweise die Ruhehörschwelle und die Pegeldynamik der Sprache stärker miteinbezogen.[52] Außerdem

[43] Vgl. Lazarus, Sust, Steckel, Rita, Kulka, Marko und Kurtz Patrick (2007, S. 262).
[44] Vgl. Hornsby (2004, S. 10).
[45] Vgl. Taghavi, Mohammadkhani und Jalilvand (2022, S. 150).
[46] Vgl. Friesecke (2014, S. 869).
[47] Vgl. Friesecke (2014, S. 869).
[48] Vgl. Lazarus et al. (2007, S. 249).
[49] Vgl. Lazarus et al. (2007, S. 263).
[50] Vgl. Lazarus et al. (2007, S. 263).
[51] Vgl. Houtgast und van Wijngaarden (2002, S. 46).
[52] Vgl. Lazarus et al. (2007, S. 263).

wurde der Einfluss, den die Weitabverdeckung[53] – also einer speziellen Form der Maskierung – auf den SII nimmt, einfacher formuliert.[54] Der SII arbeitet also mit anderen Wichtungen der Frequenzbereiche als der Artikulationsindex. Um den SII-Wert zu berechnen, wird zunächst der Frequenzbereich zwischen 100 Hz und 9,5 kHz in „Sprachbänder“[55] eingeteilt. Die Abstände bzw. Größe dieser Bänder liegen im Oktav- oder Terzbereich. Außerdem ist es auch möglich, eine Einteilung in „kritische Frequenzbänder“[56] vorzunehmen. Die Anzahl der Bänder (n) hängt dementsprechend von der Auswahl der Länge der Frequenzbänder ab. Wird die Oktave als Größe gewählt, gibt es sechs Bänder; bei der Auswahl der „kritischen Frequenzbänder“ sind es 21.[57] Es ist jedoch zu beachten, dass der errechnete Wert meist genauer wird, wenn kleinere und damit mehr Bänder gewählt werden.[58] Im Anschluss daran werden die Faktoren Sprachspektrum, Störgeräuschspektrum und Hörschwelle, entsprechend den jeweiligen Gegebenheiten, bewertet.[59] Diese spielen eine wichtige Rolle bei der Berechnung der „Frequenz-Wichtigkeitsfunktion“[60] ($I_i$), also der Gewichtung des gegebenen Frequenzbandes bezüglich Sprachverständlichkeit und der „Hörbarkeitsfunktion“[61] ($A_i$), quasi dem hörbaren Anteil der Sprachsignale im Frequenzbereich. Die Summierung der einzelnen Produkte dieser Funktionen bezogen auf jedes Sprachband ($_i$) ergibt den SII-Wert (siehe Formel 4).

$$SII = \sum_{i=1}^{n} I_i A_i$$

*Formel 4: Berechnung des Speech Intelligibility Index[62].*

Analog zum Wertebereich des Artikulationsindex (siehe Kapitel 3.3.3 (Artikulationsindex)) liegt der errechnete SII-Wert stets zwischen 0 und 1.[63] Im Gegensatz zu der Einschränkung des Artikulationsindex darf der Speech Intelligibility Index bei einem Signal-Geräuschabstand zwischen -15 dB und 15 dB angewendet werden.[64]

[53] Lazarus et al. (2007, S. 24).
[54] Vgl. Lazarus et al. (2007, S. 263).
[55] Stiles (2019, S. 2).
[56] Stiles (2019, S. 2).
[57] Vgl. Hornsby (2004, S. 12).
[58] Vgl. Stiles (2019, S. 3).
[59] Vgl. Stiles (2019, S. 3).
[60] Stiles (2019, S. 2).
[61] Hornsby (2004, S. 12).
[62] Stiles (2019, S. 2).
[63] Vgl. Stiles (2019, S. 2).
[64] Institute of Electrical and Electronics Engineers (2001, S. 263); vgl. Lazarus et al. (2007, S. 263).

### 3.3.5 Speech Transmission Index

Der Speech Transmission Index (STI) ist ein weiteres Verfahren zur Messung der Sprachverständlichkeit.[65] Dabei beschreibt sein Wert die Qualität der Übertragung von Sprachinformationen und liegt stets zwischen 0 und 1.[66] Ein STI-Wert von 0 beschreibt dabei eine eher unbrauchbare[67] Sprachverständlichkeit, und bei einem Wert zwischen 0,3 und 0,45 lässt sich die gemessene Verständlichkeit als schlecht[68] klassifizieren. Ist das Messergebnis zwischen 0,45 und 0,6, liegt eine akzeptable Sprachverständlichkeit vor; ab einem Wert von 0,6 wird von einer guten und ab 0,75 von einer exzellenten Verständlichkeit gesprochen.[69] Die Messungen können auf zwei Arten ausgeführt werden:

1) Modulationsübertragungsfunktionen werden in verschiedenen Oktavbändern gemessen.[70] Die Modulationsübertragungsfunktion gibt an, wie ein amplitudenmoduliertes Signal nach der Übertragung beeinflusst wird.[71] Diese Beeinflussung können beispielsweise die Übertragung über eine Beschallungsanlage oder auch die Raumakustik sein.[72] Beispielsweise wird ein 100 % amplitudenmoduliertes Signal, bei dem also die Amplitude von 0 bis zum Maximum und wieder zurückgeht, gesendet.[73] Wenn nun das empfangene Signal aufgrund eines hohen Grundgeräusches oder eines starken Nachhalls nur noch zu 60 % moduliert ist, liegt eine Modulationsübertragungsfunktion von 0,6 vor. Die Modulationsfrequenz beschreibt die Geschwindigkeit, in der moduliert wird, und soll der menschlichen Sprache entsprechen.[74] Da diese jedoch nicht immer gleichmäßig lauter und leiser wird, kommen 14 verschiedene Modulationsfrequenzen, die von 0,63 Hz bis 12,5 Hz reichen, zum Einsatz.[75] Wie bereits in Kapitel 3.1 (Sprachverständlichkeit) beschrieben, liegt der wichtigste Teil, in dem gesprochene Worte im Frequenzspektrum auftreten, ungefähr zwischen 100 Hz und 8 kHz. Deswegen wird bei dieser Art von Messung in 7 Oktavbändern von 125 Hz bis 8 kHz gemessen.[76] Diese 7 Oktavbänder, multipliziert mit den erwähnten 14 Modulationsfrequenzen, ergeben insgesamt 98 einzelne Messergebnisse. Mit-

[65] Vgl. Friesecke (2014, S. 869).
[66] Vgl. Probst und Böhm (57).
[67] Friesecke (2014, S. 869).
[68] Friesecke (2014, S. 869).
[69] Houtgast und van Wijngaarden (2002, S. 27).
[70] Vgl. Weinzierl (2008, S. 194).
[71] Vgl. Weinzierl (2008, S. 194).
[72] Vgl. Friesecke (2014, S. 869).
[73] Vgl. Friesecke (2014, S. 869).
[74] Vgl. Friesecke (2014, S. 869).
[75] Vgl. Probst und Böhm (58).
[76] Vgl. Probst und Böhm (57).

tels gewichteten Mittelwertes der 98 Modulationsübertragungsfunktionen wird schließlich der STI-Wert berechnet.[77]

2) Des Weiteren ist es mittlerweile auch möglich, den STI-Wert eines Raumes anhand seiner Impulsantwort zu berechnen.[78] Häufig wird heutzutage diese Art der Messung verfolgt, da sie dank moderner computergestützter Messverfahren leicht umgesetzt werden kann.[79] Dazu muss die Impulsantwort in hinreichender Länge in Bezug auf die Nachhallzeit des Raumes erfasst werden und außerdem eine ausreichende Bandbreite herrschen. Dies bedeutet, dass die Länge der Impulsantwort größer oder gleich der Nachhallzeit sein muss und eine Abtastrate von mindestens 48 kHz vorliegen sollte.[80] Ein guter Wert liegt hierbei bei der doppelten Länge der Nachhallzeit.[81] Des Weiteren ist für die Messung der Impulsantworten die Wahl eines Sweeps als Anregungssignal wichtig.[82] Durch die Gleichung (siehe Formel 5) werden mittels der Impulsantwort h(t) die Modulationsübertragungsfunktionen m(f) berechnet.

$$m(f) = \frac{\int_0^\infty h^2(t) e^{-j2\pi ft}\, dt}{\int_0^\infty h^2(t)\, dt}$$

*Formel 5: Berechnung der Modulationsübertragungsfunktion[83].*

Nachdem der Störabstand und die pegelabhängigen Effekte der Maskierung und der absoluten Hörschwelle ausgemacht und aus den 98 Modulationsübertragungsfunktionen herausgerechnet wurden, kann der STI-Wert errechnet werden.[84]

#### 3.3.5.1 Rapid Speech Transmission Index

Eine Unterform des Speech Transmission Index ist der Rapid Speech Transmission Index, auch als RASTI abgekürzt. Er wurde definiert, da die 98 Einzelmessungen des STI in der Vergangenheit sehr viel Zeit in Anspruch nahmen. Wie in Kapitel 3.1 (Sprachverständlichkeit) erwähnt, ist der Bereich um 2 kHz im Frequenzspektrum von Sprache essenziell für die Verständlichkeit. Deswegen werden beim RASTI nur noch

[77] Vgl. Friesecke (2014, S. 869).
[78] Vgl. Weinzierl (2008, S. 194).
[79] Vgl. Weinzierl (2008, S. 194).
[80] Vgl. Weinzierl (2008, S. 538).
[81] Vgl. Weinzierl (2008, S. 539).
[82] Vgl. Weinzierl (2008, S. 539).
[83] Weinzierl (2008, S. 538).
[84] Vgl. Weinzierl (2008, S. 540).

insgesamt neun Modulationsübertragungsfunktionen – davon vier im Oktavband 500 Hz und fünf im Bereich von 2 kHz – berechnet.[85] Zwar lässt sich die Messung dadurch über zehnmal schneller realisieren, kann allerdings zu falschen Ergebnissen führen und Fehler im Beschallungssystem können nicht richtig erkannt werden.[86] Deswegen wird dieses Messverfahren zunehmend weniger verwendet[87]. Die Ergebnisse der Messung mittels Rapid Speech Transmission Index liegen dabei in der gleichen Messskala wie die des STI.[88]

#### 3.3.5.2 Speech Transmission Index for Public Address Systems

Der Speech Transmission Index for Public Address Systems (STI-PA) ist speziell für die Messung von Beschallungsanlagen entwickelt worden. Er verwendet zur Messung ausschließlich eine Anregung mit moduliertem Rauschen und keine Berechnung anhand der Impulsantwort.[89] Anders als beim Rapid Speech Transmission Index werden beim STI-PA wieder alle sieben Oktavbänder für die Messung verwendet.[90] Allerdings werden aus den 14 Modulationsfrequenzen jeweils zwei ausgewählt und pro Oktavband angewendet.[91] Somit ergeben sich 14 Modulationsübertragungsfunktionen.[92] Die Beschränkung auf diese Anzahl ist in diesem Fall möglich, da diese Kombinationen die entscheidenden Punkte, die bei Beschallungsanlagen kritisch sind, abdecken.[93] Logisch ist infolgedessen, dass dieses Messverfahren hauptsächlich zur Bewertung der Beeinflussung der Sprachverständlichkeit durch Beschallungsanlagen dienen kann.[94] Ein Vorteil des Verfahrens ist jedoch, dass es sehr anwenderfreundlich ist, da auch eine fachfremde Person mittels eines USB-Sticks und eines einfachen portablen Empfängergerätes die Messung an der Beschallungsanlage durchführen kann.[95]

### 3.3.6 Short-Time Objective Intelligibility

Mit dem Short-Time Objective Intelligibility (STOI) Messverfahren wiesen C. H. Taal, R. C. Hendriks, R. Heusdens und J. Jensen im Jahr 2010 darauf hin, dass Sprachverständ-

[85] Vgl. Weinzierl (2008, S. 195).
[86] Vgl. Friesecke (2014, S. 870).
[87] Weinzierl (2008, S. 195).
[88] Vgl. Friesecke (2014, S. 870).
[89] Vgl. Weinzierl (2008, S. 195).
[90] Vgl. Houtgast und van Wijngaarden (2002, S. 91).
[91] Vgl. Houtgast und van Wijngaarden (2002, S. 91).
[92] Weinzierl (2008, S. 537).
[93] Vgl. Houtgast und van Wijngaarden (2002, S. 13).
[94] Vgl. Houtgast und van Wijngaarden (2002, S. 13).
[95] Vgl. Weinzierl (2008, S. 195).

lichkeitsmessungen exakter durchgeführt werden können, wenn die Bewertung nicht über das gesamte Signal, sondern über kleinere Abschnitte ausgeführt wird.[96] Anders als beispielsweise der AI oder STI agiert deswegen der STOI in kurzen Segmenten.[97] Die Gefahr bei der nicht segmentierten Messung eines langen Signals ist, dass kurze Stellen mit großem Signalausschlag das Gesamtmessergebnis stark beeinflussen können.[98] Deswegen wird beim STOI-Algorithmus eine Zerteilung des zu analysierenden Signals vorgenommen. Diese Segmente sollen sich überlappen und im Optimalfall eine Länge von ca. 382 ms aufweisen.[99] Der Messalgorithmus misst die Korrelation der Hüllkurven zwischen einem einwandfreien Referenzsignal und einem zu analysierenden Signal (im Folgenden Testsignal genannt).[100] Dabei ist der Ausgabewert des Endergebnisses ein Skalarwert und steht in direkter Relation zur Sprachverständlichkeit.[101] Die Messung basiert auf einer einfachen diskreten Fourier-Transformation und ist eine von Referenz- und Testsignal abhängige Funktion.[102] Für das Referenzsignal wird eine möglichst perfekte Aufnahme benötigt, also frei von Rausch- und Störgeräuschen. Der Inhalt dieser Aufnahme muss äquivalent zum Inhalt des Testsignals sein, da sonst das Ergebnis der Messung des Korrelationskoeffizienten verfälscht ist. Für die Fensterung der Frequenzanalyse wird ein Hann-Fenster verwendet.[103] Zunächst werden jedoch stille Stellen in den Signalen entfernt, da diese die Verständlichkeitsberechnung beeinflussen könnten. Dazu wird der größte Schallleistungspegel im Referenzsignal ausgemacht und alle Stellen entfernt, bei denen der Leistungspegel weniger als 40dB (bezogen auf das zuvor ausgemachte Maximum) beträgt.[104] Wichtig ist, dass das Referenzsignal und das Testsignal die gleiche Abtastrate und gleiche Länge besitzen. Des Weiteren ist es zusätzlich essenziell, dass beide Signale zeitlich synchronisiert sind.[105] Die eigentliche Segmentierung findet erst nach dem Schritt der Fourier-Transformation statt. Außerdem wird das Testsignal vor der Berechnung des Korrelationskoeffizienten noch normalisiert und „geclippt“[106], also begrenzt. Hintergrund bezüglich der Normalisierung ist, dass globale Pegelunterschiede ausgeglichen werden, da diese sonst einen Einfluss auf die Sprachverständlichkeit haben, beispielsweise weil das Referenzsignal und das Testsignal unterschiedliche Wiedergabepegel aufweisen.[107] Mittels der Begrenzung wird sicher-

[96] Vgl. Kondo (2012, S. 63).
[97] Vgl. Kondo (2012, S. 63).
[98] Vgl. C. H. Taal, Hendriks, Heusdens und Jensen (2011, S. 2126).
[99] Vgl. C. H. Taal et al. (2011, S. 2135).
[100] Vgl. C. H. Taal et al. (2011, S. 2126).
[101] Vgl. C. H. Taal et al. (2011, S. 2127).
[102] Vgl. C. H. Taal, Hendriks, Heusdens und Jensen (2010, S. 4214).
[103] Vgl. C. H. Taal et al. (2011, S. 2126).
[104] Vgl. C. H. Taal et al. (2011, S. 2126).
[105] Vgl. C. H. Taal et al. (2010, S. 4214).
[106] Vgl. C. H. Taal et al. (2010, S. 4215).
[107] C. H. Taal et al. (2011, S. 2127).

gestellt, dass ein einzelnes stark verrauschtes Segment im Testsignal das Gesamtmessergebnis nicht zu sehr beeinflusst.[108] Dies ist notwendig, da ansonsten eine weitere Verschlechterung eines bereits vollkommen unverständlichen Segments möglich wäre, und dies die Verständlichkeitsvorhersage der Messung negativ beeinflussen würde.[109] Nachdem die Messergebnisse begrenzt, normalisiert und die Korrelationskoeffizienten berechnet wurden, wird aus den Koeffizienten ein skalarer Wert gemittelt, der im direkten Zusammenhang mit der Sprachverständlichkeit steht.[110] Da mit dem Short-Time Objective Intelligibility Messverfahren im Endeffekt ein Korrelationskoeffizient gemessen wird, liegt der Wert des Ergebnisses stets zwischen 0 und 1. Der erwähnte Zusammenhang zwischen Korrelationskoeffizient und Sprachverständlichkeit äußert sich wie folgt: Bei einem Messwert von 1 herrscht volle Sprachverständlichkeit im Testsignal.[111] Analog dazu bedeutet ein Wert von 0, dass das Testsignal vollkommen unverständlich ist. Die genaue Bewertung des Korrelationswertes bezüglich der Sprachverständlichkeit lässt sich zur besseren Einschätzung wie folgt vereinfacht darstellen:[112]

- Ein Korrelationswert von 0 bis 0,5 bedeutet, dass das Testsignal nahezu keine Verständlichkeit besitzt.
- Liegt der Messwert zwischen 0,5 und 0,75 liegt eine eher moderate Sprachverständlichkeit vor.
- Ab einem Wert von 0,75 lässt sich von einer hohen Verständlichkeit sprechen.

### 3.3.7 Non-Intrusive Speech Quality Assessment

Unter Non-Intrusive Speech Quality Assessment (NISQA) ist kein bestimmtes Sprachverständlichkeitsmessverfahren, sondern eine Herangehensweise, bzw. Art der Verständlichkeitsmessung zu verstehen. Wie bereits in Kapitel 3.3 (Darstellung verschiedener Sprachverständlichkeitsmessverfahren) erwähnt, wird zwischen subjektiven und objektiven Bewertungsverfahren unterschieden. Zwar liefern subjektive Verfahren das genaueste Messergebnis, allerdings sind sie sehr zeit- sowie kostenintensiv und nicht in Echtzeit umsetzbar.[113] Unter anderem aus diesen Gründen wurden Verfahren entwickelt, die ohne die Anwendung von Hörtests die Sprachverständlichkeit eines Signals analysieren können.[114] Diese Verfahren werden in „Intrusive Speech Quality

[108] C. H. Taal et al. (2011, S. 2127).
[109] C. H. Taal et al. (2010, S. 4215).
[110] Vgl. C. H. Taal et al. (2011, S. 2127).
[111] Vgl. C. H. Taal et al. (2011, S. 2131).
[112] Vgl. C. H. Taal et al. (2011, S. 2132).
[113] Vgl. Dubey und Kumar (2013, S. 89).
[114] Vgl. Shen, Yan, Hu und Ye (2024, S. 1).

Assessment“ und „Non-Intrusive Speech Quality Assessment“ unterteilt.[115] „Intrusive Speech Quality Assessment“ Messungen benötigen zur Beurteilung der Verständlichkeit immer ein Referenzsignal.[116] Dieses muss den gleichen Inhalt sowie die gleiche Länge und Abtastrate wie das zu messende Sprachsignal aufweisen. Als Beispiel ist an dieser Stelle das Short-Time Objective Intelligibility Messverfahren (siehe Kapitel 3.3.6 (Short-Time Objective Intelligibility)) zu nennen. Zwar sind diese Messungen sehr genau, jedoch auch niemals durchführbar, ohne dass das Referenzsignal vorliegt.[117] Da es jedoch Fälle gibt, in denen kein Referenzsignal für die Sprachverständlichkeitsanalyse vorliegt, wurde bezüglich „Non-Intrusive Speech Quality Assessment“ Messverfahren geforscht.[118] Diese benötigen zur Bewertung der Verständlichkeit eines Signals lediglich das eigentliche Signal und kein Referenzsignal.[119] Daher lassen sich diese Messungen nahezu in Echtzeit und automatisiert durchführen. Außerdem sind die Ergebnisse schneller verfügbar als beim „Intrusive Speech Quality Assessment“-Verfahren. Als ein Beispiel für ein solches Messverfahren ist der ITU-T P.563-Standard zu nennen. Er ist zur Verständlichkeitsanalyse von Schmalband-Sprachsignalen entwickelt worden und arbeitet in drei Schritten: Vorverarbeitung, Verzerrungsabschätzung und Wahrnehmungsabbildung.[120]

### 3.3.8 Perceptual Evaluation of Audio Quality

Ebenso wie das Short-Time Objective Intelligibility Messverfahren (siehe Kapitel 3.3.6 (Short-Time Objective Intelligibility)) basiert auch das Perceptual Evaluation of Audio Quality (PEAQ) auf dem „Intrusive Speech Quality Assessment“-Ansatz (siehe Kapitel 3.3.7 (Non-Intrusive Speech Quality Assessment)).[121] Auch PEAQ benötigt zur Bewertung der Audioqualität eines zu analysierenden Signals ein Referenzsignal. Nach den üblichen Vorverarbeitungsschritten, wie Normalisierung und Synchronisierung der beiden Signale, werden diese mittels eines perzeptuellen Modells des menschlichen Gehörs analysiert.[122] Daraus resultiert auch der Name dieses wahrnehmungsbasierten Analyseverfahrens. PEAQ ist in zwei Versionen definiert. Die Basisversion legt dabei den Fokus auf eine schnelle Analyse, während die fortgeschrittene Version das Augen-

[115] Vgl. Shen et al. (2023, S. 3377).
[116] Vgl. Shen et al. (2023, S. 3377).
[117] Vgl. Shen et al. (2024, S. 1).
[118] Vgl. Shen et al. (2024, S. 1).
[119] Vgl. Dubey und Kumar (2013, S. 90).
[120] Vgl. Shen et al. (2024, S. 4).
[121] Vgl. Torcoli, Kastner und Herre (2021, S. 1531).
[122] Vgl. Thiede et al. (S. 4).

merk auf ein möglichst genaues Analyseergebnis richtet.[123] Dazu nutzt die Basisversion ausschließlich ein FFT-basiertes Wahrnehmungsmodell, um die Maskierungsschwelle und den wahrnehmungsbasierten Vergleich anzuwenden.[124] Dabei handelt es sich bei der Maskierungsschwelle um die Wahrnehmbarkeitsschwelle, ab der die Differenz zwischen dem Testsignal und dem Referenzsignal hörbar ist. Der wahrnehmungsbasierte Vergleich entspricht dem Vergleich der beiden Signale, nachdem simuliert wurde, wie diese durch das menschliche Ohr wahrgenommen wurden. Die fortgeschrittene Version des PEAQ verwendet nur zur Anwendung der Maskierungsschwelle ein FFT-basiertes Wahrnehmungsmodell, denn der wahrnehmungsbasierte Vergleich wird mittels eines filterbank-basierendem Wahrnehmungsmodells angewendet.[125] Im Anschluss an diese Verfahren werden sogenannte Modell Output Variables (MOV) aus dem Test- und dem Referenzsignal extrahiert. Dabei handelt es sich um perzeptuelle Merkmale wie bspw. Lautstärke, spektrale Verteilung oder Verzerrungen.[126] Die Basisversion nutzt insgesamt elf MOV, die fortgeschrittene Version hingegen nur fünf.[127] Allerdings ist bestätigt, dass die wenigeren MOV der fortgeschrittenen Version aufgrund des insgesamt komplexeren Analyseverfahrens, trotzdem alle relevanten Daten erfassen, um ein valides Ergebnis zu liefern.[128] Dies wurde evaluiert, indem zu Testzwecken die Analyse mit der fortgeschrittenen Version und allen elf MOV der Basisversion durchgeführt wurde. Das Ergebnis unterschied sich nicht von dem der standardisierten fortgeschrittenen Version mit nur fünf MOV. Die Modell Output Variables werden dann durch ein mehrschichtiges neuronales Netzwerk mit entweder drei (in der Basisversion) oder fünf Einheiten (in der fortgeschritteneren Version) in einer einzigen verborgenen Schicht zu einem Qualitätsindex, dem Objective Difference Grade (ODG), berechnet.[129] Dieser ODG bildet das Endergebnis des Messverfahrens. Das neuronale Netzwerk ist unter anderem trainiert auf Basis der verfügbaren Daten von MPEG90 und ITU93.[130]

### 3.3.9 Perceptual Evaluation of Speech Quality

Bei dem Perceptual Evaluation of Speech Quality (PESQ)-Verfahren handelt es sich ebenfalls um ein intrusives Messverfahren (siehe Kapitel 3.3.7 (Non-Intrusive Speech

[123] Vgl. Torcoli et al. (2021, S. 1531).
[124] Vgl. You, Reiter, Hannuksela, Gabbouj und Perkis (2010, S. 486).
[125] Vgl. You et al. (2010, S. 486).
[126] Vgl. Torcoli et al. (2021, S. 1531).
[127] Vgl. You et al. (2010, S. 486).
[128] Vgl. Thiede et al. (S. 10).
[129] Vgl. Thiede et al. (S. 10).
[130] Vgl. You et al. (2010, S. 486).

Quality Assessment)).[131] Es ist standardisiert in der ITU-T Recommendation P.862 und eines der meistgenutzten Verfahren zur Messung von Audioqualität, gefolgt von seiner Weiterentwicklung Perceptual Objective Listening Quality Analysis (POLQA) (siehe Kapitel 3.3.10 (Perceptual Objective Listening Quality Analysis)).[132] PESQ wurde entwickelt, um die Einflüsse auf die Audioqualität durch Komprimierung von Sprachsignalen, beispielsweise Audiokodierungsverfahren bei Telefonsystemen[133], zu messen.[134] Daher ist es entsprechend weniger präzise bezüglich einer Bewertung des Einflusses von Rauschen und Nachhall auf die Audioqualität.[135] Da es sich bei Telefonsystemen um Schmalbandsysteme[136] handelt, analysiert PESQ (im Standardmodus, auch Schmalbandmodus genannt) auch nur im Frequenzbereich von ca. 100 Hz bis 3,5 kHz.[137] Allerdings gibt es seit 2005 auch einen Breitbandmodus, welcher von 50 Hz bis 7 kHz operiert.[138] Ziel der Entwicklung des Perceptual Evaluation of Speech Quality war es, ein Messverfahren zu erstellen, das objektiv, also ohne Hilfe von Hörtests, den MOS-Wert (siehe Kapitel 3.3.1 (Mean-Opinion-Score)) eines Audiosignals ausgeben kann.[139] Auch bei PESQ werden zunächst die beiden Signale, also das Referenz- und das Testsignal, bezüglich ihrer Lautstärke angeglichen.[140] Danach werden sie mittels einer Fast Fourier Transformation (FFT) gefiltert, um einen Standard-Telefonhörer zu imitieren.[141] Nach der FFT werden beide transformierten Signale zeitlich synchronisiert und einem psychoakustischen Modell unterzogen. Dieses bildet sie unter anderem durch Anwendung einer weiteren FFT in einer Darstellung der wahrgenommenen Lautstärke in Zeit und Frequenz ab.[142] Der messbare absolute Unterschied zwischen dem Testsignal und dem Referenzsignal gibt eine Einschätzung über die Audioqualität des Testsignals. Daher integriert PESQ die Abweichungen mittels einer Methode, die optimal auf die Verteilung von Fehlern in Zeit sowie Amplitude abgestimmt ist, und ermittelt daraus einen MOS-Wert.[143] Bei einer schlechten Audioqualität liegt dieser Wert bei 1, bei einer guten Qualität ohne hörbare Verzerrungen bei 4,5. Allerdings kann in Einzelfällen, bei sehr starken Verzerrungen, der MOS-Wert (welcher durch den PESQ errechnet wurde) auf eine unterste Grenze von bis zu -0,5 fallen.[144]

[131] Vgl. Shen et al. (2024, S. 1).
[132] Vgl. Avila et al. (2019, S. 1).
[133] Vgl. Torcoli et al. (2021, S. 1530).
[134] Vgl. Dubey und Kumar (2013, S. 90).
[135] Vgl. Avila et al. (2019, S. 1).
[136] Vgl. Torcoli et al. (2021, S. 1531).
[137] Vgl. Beerends et al. (2013, S. 385).
[138] Beerends et al. (2013, S. 385).
[139] Vgl. Kondo (2012, S. 65).
[140] Vgl. Rix, Beerends, Hollier und Hekstra (2001, S. 749).
[141] Rix et al. (2001, S. 749).
[142] Vgl. Rix et al. (2001, S. 749-750).
[143] Vgl. Rix et al. (2001, S. 750-751).
[144] Vgl. Rix et al. (2001, S. 751).

### 3.3.10 Perceptual Objective Listening Quality Analysis

Wie bereits in Kapitel 3.3.9 (Perceptual Evaluation of Speech Quality) erwähnt, ist das Perceptual Objective Listening Quality Analysis (POLQA)-Verfahren eine Weiterentwicklung des PESQ-Verfahrens. Dementsprechend handelt es sich auch bei POLQA um ein intrusives Messverfahren.[145] Während PESQ hauptsächlich zur Analyse von Schmalbandaudiosystemen realisiert wurde, lässt sich bei POLQA auswählen, ob es im Schmalband-, im Breitband- oder im Super-Breitbandanalysemodus operieren soll.[146] Der Super-Breitbandmodus umfasst dabei eine Audiobandbreite von 14 kHz und der Schmalbandmodus eine Bandbreite von 2 kHz.[147] POLQA überwindet Schwächen des PESQ-Standards wie zum Beispiel die unzureichende Bewertung der Auswirkung von linearen Verzerrungen, Nachhall sowie Jitter, wie sie in Voice-over-IP auftreten.[148] Außerdem lässt sich mittels POLQA die Auswirkung der Wiedergabelautstärke auf die Sprachqualität messen. Dies gelingt, da POLQA ein anderes perzeptuelles Modell verwendet, um das Testsignal hinsichtlich der vorherrschenden Sprachqualität zu analysieren.[149] Des Weiteren idealisiert das Verfahren das Referenzsignal bezüglich leiser Hintergrundgeräuschen und der Klangfarbe, die die Messung beeinflussen können.[150] Außerdem beträgt die obere Grenze der Bewertungsskala in der Breitbandanalyse des Perceptual Objective Listening Quality Analysis-Verfahren 3,75 anstatt 4,5.[151] Es lässt sich festhalten, dass POLQA, im Gegensatz zu PESQ, in nahezu jedem Anwendungsbereich genauere Ergebnisse bei der Messung der Sprachqualität liefert. Zudem ist es gut geeignet für die Analyse von Voice-over-IP und Mobilfunksystemen.[152]

### 3.3.11 Automatic Mean-Opinion-Score

Eine interessante Weiterentwicklung des Mean-Opinion-Score ist der 2016 vorgestellte Automatic Mean-Opinion-Score (AutoMOS)[153]. Dieser ist ebenfalls kein subjektives Bewertungsverfahren, sondern zählt zu den objektiven Verfahren. Er unterscheidet sich jedoch von den anderen objektiven Bewertungsverfahren, die auf dem MOS beruhen, da er ein nicht-intrusives Messverfahren ist und demnach kein Referenzsignal für die Ana-

[145] Vgl. Shen et al. (2024, S. 1).
[146] Vgl. Torcoli et al. (2021, S. 1531).
[147] Vgl. Beerends et al. (2013, S. 386).
[148] Vgl. Beerends et al. (2013, S. 385-386).
[149] Vgl. Beerends et al. (2013, S. 387).
[150] Vgl. Beerends et al. (2013, S. 387).
[151] Vgl. Beerends et al. (2013, S. 396).
[152] Vgl. Beerends et al. (2013, S. 401).
[153] Patton, Agiomyrgiannakis, Terry, Wilson und Saurous (2016, S. 1).

lyse benötigt.[154] Der Automatic Mean-Opinion-Score wurde unter anderem entwickelt, da ein geeignetes Bewertungstool für die Verständlichkeit von Sprachsynthesealgorithmen benötigt wurde, also für Software, die aus einer Texteingabe gesprochenes Wort erstellt. In diesem Anwendungsfall ist es nicht möglich, die Analyse der Sprachverständlichkeit durch einen Vergleich zu einem Referenzsignal durchzuführen, da es kein Referenzsignal gibt. Daher ist es nur möglich, ein nicht-intrusives Messverfahren anzuwenden.[155] AutoMOS analysiert also die Verständlichkeit eines Audiosignals allein anhand der Eigenschaften des Signals.[156] Für die Umsetzung dieser Analyse bedient sich AutoMOS eines neuronalen Netzwerks.[157] Der Vorteil an diesem – auf Künstliche Intelligenz (KI)-gestütztem – System ist, dass es trainiert werden und sich somit stets selbst verbessern kann.[158]

[154] Vgl. Patton et al. (2016, S. 1).
[155] Vgl. Patton et al. (2016, S. 1).
[156] Vgl. Patton et al. (2016, S. 2).
[157] Vgl. Zhou et al. (2024, S. 876).
[158] Vgl. Patton et al. (2016, S. 1).

# 4 Vergleich der Sprachverständlichkeitsmessverfahren

In diesem Kapitel sollen die im vorherigen Kapitel 3 (Grundlagen und Definitionen) ausgeführten Sprachverständlichkeitsmessverfahren kategorisiert und deren Vor- und Nachteile einzeln aufgelistet werden. Dies soll zum einen der Übersicht und zum anderen bei der Auswahl des geeigneten Sprachverständlichkeitsmessverfahrens als Grundlage des zu entwickelnden Tools dienen.

## 4.1 Kategorisierung der Verfahren

Die in Kapitel 3 (Grundlagen und Definitionen) vorgestellten Sprachverständlichkeitsmessverfahren lassen sich wie folgt übersichtlich unterteilen. Zunächst einmal werden die verschiedenen Verfahren unterschieden zwischen einer subjektiven und einer objektiven Herangehensweise. Die Anwendung eines subjektiven Messverfahrens gelingt durch Hörtests. Diese werden einer möglichst großen Hörergruppe vorgespielt und im Anschluss hinsichtlich der Sprachqualität bzw. Sprachverständlichkeit bewertet. Objektive Messverfahren hingegen arbeiten im eigentlichen Auswertungsschritt ohne menschliche Bewertung, sondern meist mittels Software in Form von diversen Analysealgorithmen. Allerdings gibt es auch Verfahren, bei denen die Berechnung mittels Messung von verschiedenen Schallpegeln, bspw. dem Ruhegeräuschpegel eines Raumes im Vergleich zum Sprecherpegel, ausgeführt wird. Die rein softwarebasierten objektiven Analyseverfahren werden des Weiteren unterschieden nach intrusiven und nicht-intrusiven Analysealgorithmen. Die Benamung „full reference“ und „no reference“ sowie „double-ended“ und „single-ended“ impliziert die gleiche Unterteilung, wählt jedoch differente Begrifflichkeiten. Intrusive Verfahren analysieren die Qualität eines Sprachsignals mithilfe des Originalreferenzsignals. Dieses enthält die Sprachinformation absolut störungsfrei. Der Vergleich des Referenzsignals und des Testsignals bildet die Analysegrundlage dieser Methoden. Dabei ist unter anderem wichtig, dass die gesprochenen Worte in beiden Signalen absolut identisch sind, und die Signale die gleiche Länge besitzen. Nicht-intrusive Herangehensweisen hingegen benötigen kein Referenzsignal für die Bewertung der Sprachverständlichkeit, sondern bestimmen diese nur durch Analyse des Testsignals.

Zwar bieten subjektive Bewertungsverfahren wegen der Evaluation durch reale Testhörer und Testhörerinnen eine höhere Genauigkeit des Endergebnisses, sie sind allerdings durch die Notwendigkeit einer großen Testpersonengruppe auch kosten- und zeitintensiv. Da jedoch zum Beispiel im Entwicklungsprozess eines neuen Audiocodec stets neue Messungen bezüglich der Sprachqualität gefordert werden, um ein schnelles Feedback für neue Versionen zu erhalten, wurde nach schnelleren Verfahren gesucht. Unter anderem wurden aus diesen Gründen objektive Bewertungsverfahren entwickelt.

Auf das Endergebnis bezogen sind intrusive Messmethoden im Vergleich zu Non-Intrusiven aktuell meist noch präziser. Ein naheliegender Grund dafür ist, dass ein nicht-intrusiver Analysealgorithmus das ursprüngliche Signal nicht kennt. Dies bedeutet, dass sich die Analyse der Sprachverständlichkeit des Testsignals nur auf die Auswertung bezüglich bekannter Merkmale von Sprachqualität in einem Audiosignal stützen kann. Intrusive Verfahren hingegen müssen nur die beiden Signale miteinander vergleichen, um eine solide Aussage über die Sprachverständlichkeit des Testsignals treffen zu können. Dennoch sind nicht-intrusive Analysemethoden nicht obsolet. So gibt es zum Beispiel Szenarien, in denen es nicht möglich ist, eine Messung mit einem intrusiven Verfahren durchzuführen, da das Referenzsignal nicht zur Verfügung steht.

## 4.2 Tabellarische Auflistung der Verfahren

Im Folgenden sind die in Kapitel 3 (Grundlagen und Definitionen) aufgelisteten Methoden zur besseren Übersichtlichkeit tabellarisch dargestellt.

| Name des Verfahrens | Messverfahren Methode |
|---|---|
| Mean-Opinion-Score | Subjektives Verfahren |
| Articulation loss of consonants | Objektives Verfahren |
| Artikulationsindex | Objektives Verfahren |
| Speech Intelligibility Index | Objektives Verfahren |
| Speech Transmission Index (inkl. RSTI und STIP) | Objektives Verfahren |
| Short-Time Objective Intelligibility | Objektiv-Intrusives Verfahren |
| Perceptual Evaluation of Audio Quality | Objektiv-Intrusives Verfahren |

| | |
|---|---|
| Perceptual Evaluation of Speech Quality | Objektiv-Intrusives Verfahren |
| Perceptual Objective Listening Quality Analysis | Objektiv-Intrusives Verfahren |
| Automatic Mean-Opinion-Score | Objektiv-nicht-intrusives-Verfahren |

*Tabelle 1: Sprachverständlichkeitsmessverfahren, Quelle: der Verfasser*

## 4.3 Vorteile, Nachteile und passende Anwendungsbereiche der einzelnen Verfahren

In diesem Unterkapitel sollen die Vor- und Nachteile der einzelnen Verfahren ausgearbeitet und dargestellt werden. Weiterhin soll im Anschluss daran für jedes Verfahren ein möglicher Anwendungsbereich definiert werden.

### 4.3.1 Mean-Opinion-Score

Zunächst einmal ist positiv zu erwähnen, dass der Mean-Opinion-Score in der ITU-T Recommendation P.800 sehr umfangreich definiert ist. Insbesondere die Definitionen, wie die Hörtests zu gestalten sind, also welche Testumgebungen herrschen sollen, sind präzise gestaltet. Als Beispiel wären hier die Bedingungen des Aufnahmeraums, in dem die Sprachproben aufgezeichnet werden, zu nennen. Dieser muss für die „listening“-Option (siehe Kapitel 3.3.1 (Mean-Opinion-Score)), ein Raumvolumen von 30 bis 120 m³, eine Nachhallzeit von weniger als 500 ms (im Optimalfall zwischen 200 und 300 ms) sowie ein Ruhegeräusch von weniger als 30 dBA mit keinen aus dem Frequenzspektrum herausstechenden Frequenzen besitzen.[159] Ein weiterer Vorteil des MOS ist, dass der Einfluss von vielen Faktoren auf die Sprachverständlichkeit berücksichtigt wird, beispielsweise laute Hintergrundgeräusche, Jitter oder andere Verzerrungen des Sprachsignals. Dies ergibt sich aus der Art des Verfahrens, denn die subjektive Bewertung von Hörproben bietet, wie bereits in Kapitel 4.1 (Kategorisierung der Verfahren) ausgeführt, eine hohe Genauigkeit, da sie auf der Einschätzung der Sprachverständlichkeit durch menschliche Tester und Testerinnen beruht. Ein weiterer Vorteil ergibt sich daraus, dass der MOS bereits seit geraumer Zeit geläufig ist; deswegen und dank seiner

[159] Vgl. ITU-T P.800, S. 14.

großen Verbreitung gibt es viele Verständlichkeitsmessungen, die mittels des MOS gemacht wurden. Dies führt dazu, dass eine große Anzahl an Testdaten verfügbar ist.

Allerdings ergeben sich aus der klassischen subjektiven Testmethode mittels Hörtests auch Nachteile. Wie bereits in Kapitel 4.1 (Kategorisierung der Verfahren) erwähnt, sind solche Tests kosten- und zeitintensiv. Von dieser Beeinträchtigung bleibt auch der subjektive Mean-Opinion-Score nicht verschont. Ein spezieller Nachteil des MOS-Verfahrens ist, dass der Faktor der Übertragungsverzögerung, also der Verzögerung, die bei einer digitalen Audioübertragung eintritt, bei der Messung des MOS keinen Einfluss auf die Verständlichkeit hat. Aus diesem Grund bezeichnet Nocker den MOS auch als einen Wert, der nur „die Qualität einer Sprachsignal-Wiedergabe“[160] angibt. Weiterhin kann der MOS-Wert den Einfluss eines sogenannten Sprecherechos, also ein Echo, welches der Sprecher oder die Sprecherin wahrnimmt, nicht messen. Dieser Effekt tritt häufig bei Telefonanlagen auf und kann den Sprecher bzw. die Sprecherin irritieren. Dies wiederum kann dazu führen, dass die gesamte Gesprächs- und somit auch Sprachqualität leidet. Diese beiden Effekte sind bei Kommunikationssystemen, die im Vollduplexbetrieb arbeiten, ein nicht zu unterschätzender Faktor.[161] Als weiterer Nachteil kann die diskrete Bewertungsskala, die nur fünf Auswahlmöglichkeiten besitzt, gesehen werden. Diese hat zum Beispiel den Makel, dass bei hohen Qualitätsdifferenzen zwischen den einzelnen Sprachproben das Endergebnis verzerrt dargestellt werden kann, da diese fünf Stufen nicht unbedingt ausreichen, um die Qualitätsunterschiede adäquat darzustellen. Des Weiteren kann dies dazu führen, dass die Versuchspersonen in ihrer Urteilsfähigkeit eingeschränkt sind. Daher empfiehlt es sich, Zwischenwerte, beispielsweise im Wertebereich von 0,2 zu ergänzen, um eine weniger diskrete und somit kontinuierlichere Skala zu kreieren. Ein weiterer Kritikpunkt am MOS-Verfahren ist, dass die Hörtests keine Referenzsignale beinhalten, und damit die Einschätzung der Hörproben stets relativ ist. Somit können genau genommen keine zwei unterschiedlichen MOS-Tests, geschweige denn deren Ergebnisse, valide miteinander verglichen werden. Als letzter Nachteil ist auszuführen, dass bei den objektiven Bewertungsverfahren, die auf dem Mean-Opinion-Score basieren, nur reine Sprachsignale zuverlässig untersucht werden können. Die Bewertung eines Audiosignals, welches zusätzlich zur Sprache noch Musikanteile enthält bzw. nur aus Musik besteht, ist dementsprechend nicht ohne Weiteres zu empfehlen.[162]

[160] Nocker (2005, S. 87).
[161] Vgl. Nocker (2005, S. 87-88).
[162] Vgl. ITU-T P.800.1, S. 4.

Ein wichtiger Punkt bei der Einschätzung eines möglichen Anwendungsgebietes des Mean-Opinion-Score ist der Fakt, dass ein subjektiv gemessener MOS-Wert die Einflüsse von Übertragungsverzögerungen auf die Sprachverständlichkeit nicht wahrnimmt. Daraus ergibt sich, dass nur Sprachsignale valide getestet werden können, die nicht in einem Vollduplexbetrieb zum Einsatz kommen. Dementsprechend ist davon abzuraten, neue Audiokodierungsverfahren, deren letztendlicher Hauptanwendungsbereich in vollduplexbetriebenen Kommunikationssystemen liegt, nur mittels dem subjektiven MOS-Verfahren zu evaluieren. Bei halbduplexbetriebenen Systemen hingegen ist dieser Effekt weniger relevant. Jedoch kann es trotzdem in manchen Systemen zu Sprecherechos kommen, die Einfluss auf die Sprachverständlichkeit haben können. Wie bereits erwähnt, berücksichtigt das MOS-Verfahren diesen Einfluss nicht. Daher sollte bei der Verständlichkeitsmessung von Sprachübertragungsverfahren, die theoretisch gefährdet sind, Sprecherechos zuzulassen, beachtet werden, dass MOS-Bewertungen hier unkorrekte Ergebnisse liefern können. Dieser Effekt tritt selbstverständlich auch bei vollduplexbetriebenen Systemen auf. Allgemein betrachtet lässt sich jedoch bei simplexbetriebenen Systemen eine Sprachverständlichkeitsbewertung mittels MOS-Verfahren sehr valide durchführen. Voraussetzung der Verständlichkeitsmessungen ist allerdings, dass die diskrete fünf Punkte Bewertungsskala die Testpersonen nicht einschränkt. Zusätzlich ist bei subjektiven MOS-Bewertungen stets zu beachten, dass die Ergebnisse aufgrund der fehlenden Referenzsignale in der Befragung nicht mit Ergebnissen anderer MOS-Tests verglichen werden können.

### 4.3.2 Articulation loss of consonants

Eingangs ist zu erwähnen, dass der Articulation loss of consonants hauptsächlich für die Bewertung der Sprachverständlichkeit von Räumen entwickelt wurde. Dies resultiert daraus, dass zur Zeit der Alcons-Forschungen von Peutz[163] und Klein[164] im Jahr 1971[165] die Bedeutung von Audiokodierungsverfahren im Vergleich zu heute noch relativ gering war. Ein Vorteil des Alcons ist, dass schnell eine gute Einschätzung der Sprachverständlichkeit eines Raumes erhalten werden kann, ohne dass Messungen im Raum durchgeführt werden müssen. Voraussetzung ist, dass die Variablen der Formel zur Berechnung des Alcons (siehe Formel 1) bekannt sind. Bei diesen Variablen handelt es sich jedoch lediglich um den Abstand zwischen der Position des Hörers bzw. der Hörerin und der

[163] Peutz (1972, S. 915-919).
[164] Klein (1972, S. 920-922).
[165] Houtgast und van Wijngaarden (2002, S. 81).

Schallquelle, der Nachhallzeit sowie dem Hallradius des Raumes. Beispielsweise sind Konzertsälen sind diese Faktoren eines Raumes meist sowieso bekannt.

Allerdings ist, wie bereits anfänglich, erwähnt der Fakt, dass Alcons sich nur auf Sprachverständlichkeitsmessungen von Räumen anwenden lässt, ein Nachteil dieses Verfahrens. Heutzutage richtet sich das Augenmerk der Verständlichkeitsforschung hauptsächlich auf die Entwicklung von Messmethoden, die Audiokodierungs- oder im Allgemeinen Sprachübertragungsverfahren analysieren können. Dies ist mit Alcons offensichtlich nicht möglich. Ein weiterer Nachteil bezogen auf die subjektive Bewertungsmethode des Alcons ist, dass für die Ermittlung des Ergebnisses mittels Wortliste trainierte Sprecher und Sprecherinnen sowie aufmerksame Hörer und Hörerinnen benötigt werden.[166] Dies macht das ohnehin schon kostspielige subjektive Bewertungsverfahren Alcons noch teurer.

Als Empfehlung eines Anwendungsgebietes für das Articulation loss of consonants-Verfahren lassen sich wie bereits erwähnt nur die Sprachverständlichkeitsmessungen von Räumlichkeiten anführen. Im Besonderen ist jedoch anzuführen, dass Alcons gerade dann effizient und zielführend sein kann, wenn die Daten Nachhallzeit, Hallradius und Abstand zwischen Hörposition und Schallquelle eines Raumes gegeben sind. In diesem Fall führt das Ausführen der Formel des Alcons zu einem schnellen und validen Ergebnis der Sprachverständlichkeit eines Raumes.

### 4.3.3 Artikulationsindex

Ein großer Vorteil des Artikulationsindexes ist, dass der Wert schnell und einfach berechnet werden kann. So werden nur der Pegel des Sprechers bzw. der Sprecherin und der Ruhegeräuschpegel für die Berechnung benötigt. Beispielsweise bei Sprachverständlichkeitsberechnungen von Räumen lässt sich zum Beispiel anführen, dass häufig der Ruhegeräuschpegel bekannt ist. Den Pegel des Sprechers bzw. der Sprecherin direkt in A-bewerteter Form zu messen, stellt dank der heutigen Schallpegelmessgeräte auch keine Hürde mehr dar. Somit kann schnell ein brauchbares Ergebnis bezüglich der Sprachverständlichkeit eines Raumes gefunden werden.

Analog zu den Nachteilen des Alcons (siehe Kapitel 4.3.2 (Articulation loss of consonants)) lässt sich jedoch auch für den Artikulationsindex anführen, dass er eher weniger gut zur Evaluierung von neuen Sprachkodierungsverfahren genutzt werden

[166] Vgl. Bistafa und Bradley (2000, S. 531).

kann. Grund dafür ist, dass er sich hauptsächlich auf den Einfluss von Störgeräuschen (in Form von Hintergrundgeräuschen) auf die Sprachverständlichkeit bezieht. Dies ist jedoch bei der Erforschung von modernen Audiokodierungsverfahren längst nicht mehr der einzige Faktor, der Einfluss auf die Verständlichkeit nimmt. Des Weiteren wird durch den konventionellen Artikulationsindex nur die zentrale, also im Frequenzspektrum unmittelbar benachbarte, Verdeckung des Sprachsignals durch das Störgeräusch berücksichtigt.[167] Zwar gibt es eine Berechnungsmethode, auch die sogenannten Weitabverdeckungen in die Kalkulation einfließen zu lassen; allerdings wird diese in der Praxis nicht benutzt, da sie mit erheblich mehr technischem Aufwand verbunden ist.[168] Jedoch kamen Studien von Kryter[169] aus dem Jahr 1962 zu dem Schluss, dass die Einbeziehung dieser Verdeckungen bei schmalbandigen Geräuschen für ein valides Messergebnis notwendig sei.[170] Außerdem ist der Artikulationsindex nur definiert bei einem Signal-Geräuschabstand zwischen -12 dB und 18 dB. Diese Einschränkung ist im Vorhinein zu beachten, da ansonsten kein valides Ergebnis aus der Messung entnommen werden kann.

Aufgrund dieser Vor- und Nachteile lässt sich festhalten, dass auch der Artikulationsindex sich gut für die Anwendung von Verständlichkeitsmessungen von Räumen eignet. Da er jedoch nur Störgeräusche in Form des Ruhegeräuschpegels in die Messung miteinbezieht, ist es nicht ratsam, ihn für die Sprachverständlichkeitsmessungen von Audiokodierungsverfahren oder auch von Beschallungsanlagen zu wählen.

### 4.3.4 Speech Intelligibility Index

Zunächst ist noch einmal zu erwähnen, dass es sich bei dem Speech Intelligibility Index um eine Weiterentwicklung des Artikulationsindex handelt. Dementsprechend wurden einige Nachteile des AI erkannt und bei der Entwicklung des SII mitberücksichtigt. Hierbei ist beispielsweise als Vorteil des SII der größere Signal-Geräuschabstand von -15 dB bis 15 dB, verglichen zum Artikulationsindex, zu erwähnen.[171] Des Weiteren bezieht der SII im Gegensatz zum AI auch Weitabverdeckungen in die Analyse der Sprachverständlichkeit mit ein.[172] Wie bereits in Kapitel 4.3.3 (Artikulationsindex) erwähnt, haben diese besonders bei schmalbandigen Geräuschen einen nicht zu vernachlässigenden Einfluss

[167] Vgl. Lazarus et al. (2007, S. 262).
[168] Vgl. Lazarus et al. (2007, S. 263).
[169] Kryter (1962, S. 1698-1702).
[170] Vgl. Lazarus et al. (2007, S. 263).
[171] Vgl. Lazarus et al. (2007, S. 266).
[172] Vgl. Lazarus et al. (2007, S. 263).

auf die Verständlichkeit eines Sprachaudiosignals. Unter anderem auch daher ist der SII effektiv, wenn ein Sprachsignal stationäres additives Rauschen enthält oder Filter, die die Sprachbandbreite einschränken.[173] Abgesehen von diesen Faktoren sei noch erwähnt, dass es sich bei dem Speech Intelligibility Index um ein objektives Verfahren handelt, welches sich auf akustische und psychoakustische Grundlagen stützt. Ein weiterer Vorteil des SII ist, dass die Daten der Ruhehörschwelle vor der Berechnung leicht angepasst werden können, denn dafür muss nur die Hörbarkeitsfunktion assimiliert werden. Diese Flexibilität führt dazu, dass die Berechnung des Speech Intelligibility Index auch für die Wahrnehmung von hörgeschädigten Personen durchgeführt werden kann und dabei dementsprechend aussagekräftige Ergebnisse liefert.[174] Voraussetzung für diese Variation des SII ist, dass die „grundlegenden physischen und wahrnehmungsbedingten Annahmen des Verfahrens“[175] weiterhin beachtet werden. Abgesehen von diesen Sonderfällen wurde die normale Hörschwelle im Vergleich zum Artikulationsindex in verbesserter Form in die Berechnung miteinbezogen und der Frequenzbereich erweitert.[176]

Allerdings besitzt das Speech Intelligibility Index Verfahren auch Nachteile. So berücksichtigt es beispielsweise nur stationäres Rauschen und ist bei fluktuierendem Rauschen ungenau.[177] Außerdem ist die gesamte Berechnung des SII komplexer als die des Artikulationsindexes. Dementsprechend ist es nicht möglich, nur mittels der Messergebnisse zweier A-bewerteter Schalldruckpegel schnell ein erstes Ergebnis, welches als Einschätzung der Sprachverständlichkeit dient, zu erhalten. Die Berechnung des SII kann dementsprechend weniger spontan und nur mit Hilfsmitteln durchgeführt werden, während für die Kalkulation des AI lediglich ein Schalldruckpegelmessgerät, welches eine A-Bewertung durchführen kann, benötigt wird. Diesem Argument ist entgegen zu setzten, dass der Mehraufwand auch mit einem genaueren Messergebnis und damit einem präziseren Wert der Sprachverständlichkeit einhergeht. Ein weiterer Nachteil des SII ist, dass das Verfahren den Faktor der zeitlichen Verzerrungen wie Hall oder Amplitudenkompression nicht korrekt zu berücksichtigen scheint.[178] Demnach bezieht sich der SII nur auf einen Aspekt der Verständlichkeit und schließt genannte Faktoren bei der Berechnung nicht mit ein.

Interessant sind die möglichen Anwendungsgebiete des SII. Hierbei sind drei Beispiele aus hörmedizinischen Untersuchungen anzuführen. Da Frequenz-Wichtigkeits-

[173] Kates und Arehart (2005, S. 2224).
[174] Vgl. Kates und Arehart (2005, S. 2225).
[175] Kates und Arehart (2005, S. 2225).
[176] Lazarus et al. (2007, S. 263).
[177] Kates und Arehart (2005, S. 2225).
[178] Vgl. Houtgast und van Wijngaarden (2002, S. 72).

funktionen für klinische Sprachtests in Datenbanken vorhanden sind, kann die Sprachwahrnehmung vorhergesagt werden. Dies bedeutet, dass Mediziner und Medizinerinnen dadurch die tatsächliche Hörleistung mit der vorhergesagten Leistung des Patienten oder der Patientin, basierend auf dem SII, vergleichen können.[179] Außerdem kann der SII hilfreich bei der Auswahl geeigneter Hörgerätverstärkungen sein, denn ein Hörgerät, welches einen besseren unterstützenden Speech Intelligibility Index bietet, kann dem Patienten bzw. der Patientin eine größere Hilfe sein.[180] Als Letztes Beispiel aus der Hörmedizin kann angeführt werden, dass der SII als gutes Maß für die Anpassung von Hörgeräten verwendet werden kann, denn ein Vergleich von zwei SII-Tests, einmal mit Hörgerät und einmal ohne, kann evaluieren, ob eine tatsächliche Verbesserung der Sprachverständlichkeit für den Patienten bzw. die Patientin vorherrscht oder nicht.[181]

### 4.3.5 Speech Transmission Index

Zwar wurde innerhalb des Kapitels 3.3.5 (Speech Transmission Index) noch eine Unterscheidung zwischen dem konventionellen Speech Transmission Index, dem Rapid Speech Transmission Index und dem Speech Transmission Index for Public Address Systems durchgeführt, jedoch sollen diese drei Varianten des STI in diesem Kapitel gemeinsam behandelt werden.

Als Vorteil des Speech Transmission Index ist anzuführen, dass bei diesem Verfahren alle wesentlichen Einflüsse, die ein Raum auf die Sprachverständlichkeit nehmen kann, bei der Berechnung des Ergebnisses berücksichtigt werden. Außerdem wird der Einfluss von Hintergrundgeräuschen sowie die „Verringerung der Modulationstiefe durch die Hörschwelle bei niedrigen und durch Maskierungseffekte bei hohen Signalpegeln“[182] beachtet. Nahezu alle Einflüsse auf die Sprachverständlichkeit, die im Vorhinein bei einer akustischen Raumplanung beeinflusst werden können, sind somit durch den STI kalkulierbar.[183] Außerdem lässt sich der STI aufgrund seiner Berechnungsmethode gut mit Akustiksoftwaretools simulieren.[184] Des Weiteren ist der Speech Transmission Index im Gegensatz zu dem Speech Intelligibility Index (siehe Kapitel 4.3.4 (Speech Intelligibility Index)) auch darauf ausgelegt, eine Vorhersage der Sprachverständlichkeit unter den Bedingungen von zeitlichen Verzerrungen wie Nachhall oder Amplitudenkompression zu

[179] Vgl. Stiles (2019, S. 4).
[180] Vgl. Stiles (2019, S. 4).
[181] Vgl. Stiles (2019, S. 4-5).
[182] Probst und Böhm (57).
[183] Vgl. Probst und Böhm (65).
[184] Vgl. Probst und Böhm (63–64).

treffen.[185] Wie bereits in Kapitel 3.3.5.2 (Speech Transmission Index for Public Address Systems) erwähnt, ist ein Vorteil von STI-PA, dass eine Messung mittels dieses Verfahrens sehr einfach anzuwenden ist und keiner besonderen Fachkenntnisse bedarf.[186] Um das Jahr 1990[187] hatte der ungenauere Rapid Speech Transmission Index den Vorteil, dass er auch in Echtzeit ausgeführt werden konnte. Allerdings ist mittlerweile die Rechenleistung von modernen Computern so hoch, dass auch die Grundform des STI schnell und somit auch in Echtzeit durchgeführt werden kann.

Da das RASTI-Verfahren ohnehin nur ein Kompromiss zwischen einer möglichst schnellen und einer präzisen Messung war und dementsprechend für heutige Standards nicht mehr exakt genug ist, wird es nur noch sehr selten eingesetzt.[188] Ein weiterer Nachteil des konventionellen Speech Transmission Index ist, dass er bei Anwendung bei dynamischen Amplitudenkompressionen, wie sie beispielsweise in der Hörgerätetechnik Verwendung finden, keine genauen Messergebnisse liefert.[189] Diese Abweichungen können sich in beide Richtungen auswirken. Dies meint, dass entweder der STI-Wert geringer ist als die wirkliche Sprachverständlichkeit eines Signals, oder aber der eigentliche Verständlichkeitswert unterhalb des mittels des STI gemessenen Wertes liegt. Ein Beispiel, bei dem der erste der beiden Fälle eintritt, ist, wenn bei einer STI-Messung ein digitales Signal mit einer falschen Abtastrate abgespielt wird, bzw. ein analoges Signal mit einer falschen Geschwindigkeit. Bei einer Veränderung der Abtastrate um ein Prozent verringert sich bereits der STI-Wert von 1 auf 0,74.[190] Dies entspricht zwar in der STI-Skala nur einer Verschlechterung um eine der fünf Einteilungen; bemerkenswert ist jedoch, dass sich bei dieser Änderung nahezu keine Veränderung der realen Sprachverständlichkeit ergibt.[191] Der umgekehrte Fall tritt ein, wenn das Signal regelmäßige Aussetzer (Drop-outs) besitzt. Bei digitalen Systemen kann es dabei zu Verzerrungen an den Stellen der Drop-outs kommen. So wurde beispielsweise bei einem Testsignal, welches in regelmäßigen Abständen (alle 100 ms und 250 ms) Aussetzer aufweist, noch ein STI-Wert zwischen 0,85 und 0,94 gemessen, was einer exzellenten Verständlichkeit entspricht.[192] Natürlich war jedoch die wirkliche Sprachverständlichkeit quasi nicht vorhanden. Eine weitere Schwäche des STI-Verfahrens ist Center-Clipping. Diese Form der Verzerrung tritt auf, wenn niedrige Pegelanteile eines Audiosignals nicht korrekt übertragen, oder komplett stummgeschaltet werden. Dies kann eine erhebliche Auswirkung auf

[185] Houtgast und van Wijngaarden (2002, S. 72).
[186] Vgl. Weinzierl (2008, S. 195).
[187] Weinzierl (2008, S. 195).
[188] Vgl. Weinzierl (2008, S. 195).
[189] Vgl. Houtgast und van Wijngaarden (2002, S. 61).
[190] Houtgast und van Wijngaarden (2002, S. 62).
[191] Vgl. Houtgast und van Wijngaarden (2002, S. 62).
[192] Houtgast und van Wijngaarden (2002, S. 62).

die Verständlichkeit haben. Jedoch unterschätzt eine Messung mittels des Speech Transmission Index die Auswirkung auf die Sprachverständlichkeit, wie folgendes Beispiel zeigt: Um dies darzustellen, wurden STI-Tests durchgeführt, bei denen die Clipping-Pegel 10 % und 20 % unter dem maximalen Signalpegel lagen.[193] Der STI-Wert lag bei 0,69 und bei 0,61, diese beiden Werte entsprechen einer Einschätzung der Verständlichkeit als gut.[194] 10 % Center-Clipping führt zwar zu stark verzerrter, aber verständlicher Sprache, jedoch ist bei 20 % Clipping gar keine Sprachverständlichkeit mehr gegeben.[195] Als letzter Nachteil des STI sei erwähnt, dass die Messung nur mit speziellem technischem Gerät durchgeführt werden kann wie einem besonderen Schallpegelmessgerät.

Wie bereits aus den Vorteilen des STI zu erkennen ist, bietet es sich an, dieses Verfahren für akustische Planungen von Räumen heranzuziehen. Im Besonderen, wenn bei der Planung bereits feststeht, welche Sprachverständlichkeit in dem endgültigen Raum vorherrschen soll. Weiterhin lässt sich der Speech Transmission Index aber auch bei der Messung der Verständlichkeit von schon existierenden Räumen gut und präzise einsetzen. Aufgrund seiner Einfachheit und Schnelligkeit ist heutzutage der Speech Transmission Index for Public Address eine sehr gute Methode. Dies zeigt sich bei Verständlichkeitsmessungen von Beschallungsanlagen, auch in Verbindung mit der Raumakustik des Raumes, in der die Anlagen aufgebaut sind. Außerdem wird der STI heutzutage weiterhin zur Messung der Sprachverständlichkeit von Notfallansagesystemen benutzt.

### 4.3.6 Short-Time Objective Intelligibility

Wie bereits erwähnt, zeichnet sich das Short-Time Objective Intelligibility-Verfahren dadurch aus, dass es das Audiosignal in kleinen Segmenten analysiert. Im Vergleich zu anderen Methoden wird das Messergebnis dadurch jedoch erheblich präziser und kann somit die real existierende Verständlichkeit besser erfassen.[196] Außerdem ist ein weiterer Vorteil von STOI, dass die eigentliche Analyse der Sprachverständlichkeit unkompliziert umgesetzt wird und somit verhältnismäßig leicht anzuwenden ist. Dies macht das Verfahren transparent, denn der STOI arbeitet als objektiv-intrusives Bewertungsverfahren mittels eines Vergleichs zwischen dem Testsignal und dem entsprechenden Referenz-

[193] Vgl. Houtgast und van Wijngaarden (2002, S. 62).
[194] Vgl. Houtgast und van Wijngaarden (2002, S. 62).
[195] Vgl. Houtgast und van Wijngaarden (2002, S. 62).
[196] Vgl. C. H. Taal et al. (2011, S. 2135).

signal.[197] Die Entwickler Taal, Hendriks und Heusdens gaben in mehreren Zeitschriftenartikeln sogar frei verfügbare Matlab-Implementierungen zur Berechnung des STOI an.[198] Außerdem ist die Verarbeitung von kurzen Segmenten auch weniger rechenintensiv. Ein weiterer Vorteil von STOI ist, dass es im Gegensatz zu anderen Methoden nicht nur bei einer hohen Verständlichkeit angemessene Messwerte ausgibt, sondern auch, wenn diese bei einem Audiosignal relativ gering ausfallen. Tests haben gezeigt, dass die Ergebnisse von STOI über den gesamten Bereich der Sprachverständlichkeit die reale Verständlichkeit adäquat abbilden.[199] Die gleichen Tests von Taal, Hendriks und Heusdens haben auch visualisiert, dass ihr STOI auch für verrauschte Sprachsignale eine gute Reliabilität aufweist.[200]

Als Nachteil des Short-Time Objective Intelligibility-Verfahren sei zunächst noch einmal erwähnt, dass es wegen seiner intrusiven Bewertungsmethode immer ein Referenzsignal benötigt, um eine Analyse durchzuführen. Des Weiteren scheint STOI die Auswirkungen von Pausen im Sprachsignal nicht korrekt zu bewerten. Wenn diese Pausen beispielsweise durch Aussetzer in der Übertragung auftreten, können sie sich sehr störend auf die Sprachverständlichkeit des Audiosignals auswirken. Hintergrund ist, dass die Pausen Auswirkungen auf den Sprachrhythmus haben können. Diese Beeinflussung nimmt STOI jedoch nicht korrekt wahr und bewertet infolgedessen die Verständlichkeit von Signalen mit solchen Pausen höher als den realen Wert.[201] Daher muss auch davon ausgegangen werden, dass bei einem Sprachsignal, welches in erheblichem Maße Zeitverschiebungen und Zeitkompressionen enthält, die Sprachverständlichkeit mit STOI nicht korrekt gemessen werden kann.[202] Außerdem neigt das Short-Time Objective Intelligibility-Verfahren dazu, den Einfluss von stationärem und schwankendem Rauschen verzerrt wahrzunehmen.[203] So gibt STOI einen zu schlechten Verständlichkeitswert an, wenn das Testsignal moduliertes Rauschen enthält.[204] Im Gegensatz dazu wird der Wert der realen Sprachverständlichkeit überschätzt, wenn das Signal sprachförmiges Rauschen enthält.[205] Dieses Rauschen enthält spektrale Eigenschaften, die denen von reinen Sprachsignalen ähneln. Es imitiert sozusagen ein Störgeräusch, welches durch andere Sprecher bzw. Sprecherinnen verursacht wird.

[197] Vgl. Kondo (2012, S. 63).
[198] C. H. Taal et al. (2010, S. 4217).
[199] C. H. Taal et al. (2010, S. 4216).
[200] Vgl. C. H. Taal et al. (2010, S. 4216-4217).
[201] Vgl. Tang und Cooke (2011, S. 347).
[202] Tang und Cooke (2011, S. 347).
[203] Vgl. Tang und Cooke (2011, S. 347).
[204] Vgl. Tang und Cooke (2011, S. 347).
[205] Vgl. Tang und Cooke (2011, S. 347).

Die möglichen Anwendungsgebiete von STOI liegen im Besonderen bei Telefonsystemen, im Mobilfunk, bei verrauschten Sprachsignalen sowie bei der Bewertung der Verständlichkeit und Leistung von Hörgeräten und Hörhilfen wie Cochlea-Implantaten.[206] Des Weiteren ist Short-Time Objective Intelligibility generell ein weit verbreitetes Maß zur Sprachverständlichkeitsmessung von neuen Sprachverarbeitungsalgorithmen.

### 4.3.7 Non-Intrusive Speech Quality Assessment

Natürlich ist Non-Intrusive Speech Quality Assessment, wie bereits in Kapitel 3.3.7 (Non-Intrusive Speech Quality Assessment) erwähnt, kein explizites Verfahren, sondern beschreibt im Allgemeinen Messverfahren, die eine nicht-intrusive Analysemethode anwenden. Diese benötigen dabei kein Referenzsignal für die Analyse der Sprachverständlichkeit. Dadurch besitzen sie den großen Vorteil, dass sie schneller ein Ergebnis ausgeben können. Theoretisch ist dadurch auch eine Echtzeitanwendung möglich. Ein möglicher Anwendungsfall für einen Echtzeit-Verständlichkeitsanalysealgorithmus ist der Einsatz als Qualitätskontrolle an verschiedenen Knotenpunkten in einem Telekommunikationsnetz.[207] Damit kann die Sprachverständlichkeit an den einzelnen Knoten überwacht und somit garantiert werden, dass jeder Kommunikationsteilnehmer das Signal auch verständlich empfängt. Des Weiteren sind NISQA-Verfahren kostengünstiger anzuwenden und besser in bestehende Systeme zu integrieren als intrusive Verfahren. Außerdem sind Non-Intrusive Speech Quality Assessment-Algorithmen flexibler und skalierbarer, da sie beispielsweise einfach an verschieden große Kommunikationssysteme angepasst werden können. Ein weiterer Vorteil speziell der KI-gestützten NISQA-Verfahren ist, dass sie durch die fortschreitende Verbesserung von Systemen mit neuronalen Netzen mittlerweile auch sehr brauchbare Ergebnisse bei komplexeren Sprachsignalen generieren.[208] Zuletzt sei noch erwähnt, dass dieses Verfahren sehr bedienungsfreundlich ist. Dies ist sehr vorteilhaft, da der Nutzer bzw. die Nutzerin nur das zu analysierende Signal einladen muss, und die Verständlichkeit jedes Sprachsignals gemessen werden kann.

Allerdings weisen nicht-intrusive Bewertungsverfahren auch Nachteile auf. So sind die Ergebnisse der meisten Algorithmen, die kein neuronales Netzwerk, bzw. keine KI einsetzen, im Vergleich zu den Ergebnissen von intrusiven Verfahren ungenauer. Als Beispiel sei hier eines der ersten NISQA-Verfahren genannt, welches 1998 von Au und

[206] Vgl. Kolbæk, Tan und Jensen (2019, S. 283-284).
[207] Vgl. Dubey und Kumar (2013, S. 90).
[208] Vgl. Shen et al. (2024, S. 10).

Lam vorgestellt wurde: Dieser Algorithmus analysiert die Sprachverständlichkeit nur anhand der visuellen Eigenschaften des Spektrogramms von verzerrter Sprache.[209] Des Weiteren sei noch erwähnt, dass selbst KI-gestützte nicht-intrusive Sprachverständlichkeitsalgorithmen auf gute Trainingsdaten angewiesen sind. Außerdem sind sie sehr rechenintensiv, da sie nicht nur zwei Signale miteinander abgleichen, sondern komplexe neuronale Netze verwenden.

Es gibt viele Fälle, in denen eine Verständlichkeitsmessung gefragt oder sogar benötigt wird, jedoch kein Referenzsignal vorherrscht, und es auch nicht unkompliziert ist, ein reines Sprachsignal zu erstellen. Dann besteht meist keine andere Möglichkeit, als eine objektive nicht-intrusive Messung durchzuführen, und es muss ein NISQA-Verfahren eingesetzt werden. Ein Anwendungsbeispiel hierfür wären, wie bereits erwähnt, Echtzeitanwendungen wie Live-Übertragungen von Sprache, Voice-over-IP oder Mobilfunknetze. Abgesehen davon werden nicht-intrusive Bewertungsverfahren auch bei der Verständlichkeitsbewertung von Sprachsynthesealgorithmen wie Sprachassistenzsystemen verwendet. Als Letztes sei noch die Audio-Postproduktion von Filmen erwähnt, denn meist liegt bei Originalton-Aufnahmen kein Referenzsignal vor, und die Sprachqualität kann nur anhand des fertigen Audiosignals bewertet werden.

### 4.3.8 Perceptual Evaluation of Audio Quality

Zunächst ist als Vorteil des Perceptual Evaluation of Audio Quality-Verfahren zu erwähnen, dass es über 20 Jahre nach seiner Entwicklung immer noch ein sehr gutes Wahrnehmungsmodell beinhaltet. Dies führt dazu, dass es in der heutigen Zeit weiterhin als Grundlage für viele Verständlichkeitsmessungen verwendet wird und auch teilweise die Basis von neuen Bewertungsverfahren bildet.[210] Des Weiteren berücksichtigt der Algorithmus bei der Messung viele Faktoren, welche Einfluss auf die Sprachverständlichkeit nehmen können. So werden nichtlineare sowie lineare Verzerrungen, Maskierungen, Veränderungen in der Modulation und Einflüsse auf die harmonische Struktur beachtet.[211] Ein weiterer Vorteil des Messverfahrens ist, dass es frei verfügbar ist. So existieren beispielsweise Open-Source Matlab-Implementierungen, in denen der PEAQ-Algorithmus genutzt wird. Außerdem ist ein weiterer Vorteil von PEAQ, dass er ebenso auf Musikaufnahmen trainiert wurde und somit auch die Verständlichkeit dieser bewerten kann.

[209] Vgl. Dubey und Kumar (2013, S. 90-91).
[210] Vgl. Torcoli et al. (2021, S. 1539).
[211] Vgl. Thiede et al. (S. 22).

Allerdings weist das Verfahren auch Nachteile auf. Zunächst ist zu erwähnen, dass die Grundversion des Algorithmus nur Stereo-Dateien unterstützt. Eine Verständlichkeitsanalyse von Dateien mit mehr als zwei Audiokanälen ist daher nicht möglich.[212] Außerdem zeigte sich folgende Schwäche des Verfahrens: Der Algorithmus liefert unpräzisere Ergebnisse, wenn Sprachsignale analysiert werden, welche in ihrer Art dem Verfahren unbekannt erscheinen. Dies liegt daran, dass das neuronale Netz des Algorithmus mit speziellen Trainingsdaten angelernt wurde und daher nur bedingt auf andersartige Signale reagieren kann. Allerdings wurde versucht, möglichst unterschiedliche Audiosignale zu trainieren, sodass dieser Effekt zwar zu beachten ist, jedoch auch weniger häufig auftritt. Des Weiteren sind die Messungen von PEAQ bei Signalen mit Audiokodierungsverfahren mit niedrigen Bitraten unzuverlässig.[213] Es lässt sich festhalten, dass Perceptual Evaluation of Audio Quality hauptsächlich dafür ausgelegt ist, Sprachsignale, die eine hohe Verständlichkeit aufweisen, adäquat einzuschätzen.[214] Hingegen sollten Messergebnisse von Sprache mit geringer Sprachverständlichkeit kritisch begutachtet werden.

Erstaunlich ist, dass das Perceptual Evaluation of Audio Quality-Verfahren unter anderem auch geeignet ist für die Verständlichkeitsmessung von Lautsprecheranlagen. So fanden Fischer, Feneberg und Krump heraus, dass der Algorithmus für diese Anwendung eine hohe Genauigkeit aufweist und somit auch dafür herangezogen werden kann.[215] Einzige Voraussetzung ist die Ergänzung eines Spektralkriteriums. Über diesen Spezialfall hinaus wird PEAQ jedoch auch wie andere wahrnehmungsbasierte Bewertungsverfahren bei der Entwicklung und Forschung von Audiokodierungsverfahren sowie bei Streamingdiensten oder in Broadcastumgebungen eingesetzt.

### 4.3.9 Perceptual Evaluation of Speech Quality

Eingangs ist zu erwähnen, dass analog zum PEAQ-Verfahren (siehe Kapitel 4.3.8 (Perceptual Evaluation of Audio Quality)) auch das Wahrnehmungsmodell des Perceptual Evaluation of Speech Quality bereits über 20 Jahre alt ist. Wie bereits erwähnt, ist es trotzdem bis heute noch sehr aktuell und „akkurat“[216]. Ein weiterer Vorteil von PESQ ist folgender: Das Verfahren wurde speziell für die Verständlichkeitsbewertung von Signalen entwickelt, die Verzerrungen durch Sprachkodierungs- und Sprach-

[212] Vgl. You et al. (2010, S. 483).
[213] Vgl. You et al. (2010, S. 483).
[214] Vgl. You et al. (2010, S. 487).
[215] Vgl. Fischer und Feneberg, Gregor, Kump, Gerhard (2011, S. 598).
[216] Torcoli et al. (2021, S. 1539).

kompressionsalgorithmen enthalten.[217] Dementsprechend liefert es bei der Analyse dieser Art von Audiosignalen präzise Messergebnisse, ist jedoch nicht spezialisiert auf eine besondere Kodierungsart. Außerdem ist nachgewiesen, dass PESQ im Schmalbandanalysemodus zu sehr realistischen Ergebnissen der Sprachverständlichkeitsanalyse führt.[218] Des Weiteren ist die einfache Zugänglichkeit von PESQ positiv anzuführen. So sind einige PESQ-Anwendungen frei verfügbar: jeder und jede mit entsprechendem Basiswissen, zum Beispiel über Python-Implementierungen, kann seine bzw. ihre eigene Verständlichkeitsmessung konstruieren.

Jedoch hat das Perceptual Evaluation of Speech Quality-Verfahren auch Nachteile. So sind die Ergebnisse der Sprachverständlichkeitsanalyse ungenau, wenn die Testsignale Nachhall oder Rauschen enthalten. Auch wird der Einfluss von Algorithmen, die die Verständlichkeit verbessern, nicht korrekt von PESQ wahrgenommen.[219] Außerdem ist die Genauigkeit des Verfahrens, wenn es im Breitbandmodus operiert, im Vergleich zu anderen Algorithmen nur „akzeptabel“[220] und nicht sehr hoch. Des Weiteren hat sich gezeigt, dass lineare Frequenzgangverzerrungen sowie Zeitdehnungen bzw. Zeitkompressionen, wie sie in Voice-over-IP-Anwendungen auftreten können, einen Einfluss auf die Sprachverständlichkeit nehmen, der von PESQ nicht korrekt erfasst werden kann.[221] Ebenso verhält es sich mit bestimmten Arten von Verzerrungen von Audiokodierung und Nachhall.[222] Die Forschergruppe (bestehend aus Rix, Beerends, Hollier und Hekstra) fand überdies heraus, dass das Perceptual Evaluation of Speech Quality, wie alle wahrnehmungsbasierten Bewertungsverfahren, Probleme hat, die Verständlichkeit bei Sprachsignalen mit kurzzeitiger Stille richtig einzuschätzen.[223] PESQ überbewertet den Einfluss auf die Sprachverständlichkeit von kurzzeitigem Clipping (im Bereich von 50 ms) am Anfang oder Ende des Signals.[224] Im Gegensatz dazu unterschätzt der Algorithmus den Einbruch der Verständlichkeit eines Signals, wenn während der Sprachinformation Stille eintritt.[225] Dieses Phänomen tritt unter anderem bei Paketverlust in Audio-over-IP-Anwendung auf. Ein weiterer Nachteil des Verfahrens ist, dass es nur eine maximale Abtastrate von 16 kHz[226] unterstützt. Dies ist in der heutigen Zeit insofern ungünstig, da die meisten Audiosignale eine weitaus höhere Abtastrate aufweisen. Daher müsste dieses dann zunächst konvertiert werden. Außerdem kann der Algorithmus auf-

[217] Vgl. Avila et al. (2019, S. 1).
[218] Beerends et al. (2013, S. 385).
[219] Vgl. Avila et al. (2019, S. 1).
[220] Beerends et al. (2013, S. 385).
[221] Vgl. Beerends et al. (2013, S. 386).
[222] Beerends et al. (2013, S. 386).
[223] Vgl. Rix et al. (2001, S. 752).
[224] Vgl. Rix et al. (2001, S. 752).
[225] Vgl. Rix et al. (2001, S. 752).
[226] Beerends et al. (2013, S. 396).

grund seiner Analyseart weder den Einfluss von Sprecherechos auf die Sprachverständlichkeit noch den von Gesprächsverzögerungen bewerten.[227] Der Begriff der Gesprächsverzögerung besitzt in diesem Fall die gleiche Bedeutung wie der in Kapitel 4.3.1 (Mean-Opinion-Score) genannte und erklärte Effekt der Übertragungsverzögerung. Abgesehen von diesen Umständen ist das Verfahren bei großen Datenmengen relativ rechenintensiv und benötigt dementsprechend einen hohen Rechenaufwand. Zwar ist dies eine Schwäche, die durch die im Allgemeinen steigende Rechenleistung von modernen Computern weniger relevant ist, jedoch gilt es, dies bei der Implementierung in einer mobilen kleinen Recheneinheit zu beachten. Als letzter hier angeführter Nachteil sei noch die Inflexibilität des Verfahrens gegenüber unterschiedlichen Wiedergabepegeln erwähnt. So setzt PESQ voraus, dass die gesamte Wiedergabe auf „demselben optimalen Wiedergabepegel erfolgt“[228].

Trotz der zahlreich aufgezählten Nachteile des Perceptual Evaluation of Speech Quality ist das Verfahren vielseitig einsetzbar und gehört noch heute zu den weltweit etablierten Industrie-Standards der objektiven Messverfahren für Sprachqualität.[229] So wird es unter anderem nach wie vor für die Evaluierung von Telefonnetzen und Telefonsystemen sowie Sprachkodierungsverfahren verwendet.[230] Allerdings muss aufgrund der Schwäche des Algorithmus – bezüglich des Einflusses von Nachhall auf die Verständlichkeit – davon ausgegangen werden, dass er eher ungeeignet für Sprachverständlichkeitsmessungen im raumakustischen Kontext ist.

### 4.3.10 Perceptual Objective Listening Quality Analysis

Wie bereits erwähnt, handelt es sich bei dem Perceptual Objective Listening Quality Analysis-Verfahren um eine Weiterentwicklung von Perceptual Evaluation of Speech Quality. Daher wurden einige Beeinträchtigungen von PESQ bei der Ausarbeitung von POLQA berücksichtigt. So ist zunächst zu erwähnen, dass POLQA im Gegensatz zu PESQ zusätzlich zu den beiden Analysemodi Schmalband und Breitband auch einen sogenannten Super-Breitbandmodus[231] besitzt. Dieser operiert mit einer Audiofrequenzbandbreite von 14 kHz[232] und kann somit auch Audiosignale mit einem größeren Frequenzspektrum, wie zum Beispiel high definition (HD)-Voice-over-IP-Signale, auswerten.

[227] Vgl. Rix et al. (2001, S. 752).
[228] Beerends et al. (2013, S. 391).
[229] Beerends et al. (2013, S. 400).
[230] Vgl. Rix et al. (2001, S. 752).
[231] Shen et al. (2024, S. 2).
[232] Beerends et al. (2013, S. 386).

Dieser neue Modus erweitert also das Spektrum der Audiosignale, die mittels POLQA getestet werden können. Aber auch durch seinen verbesserten Analysealgorithmus ist das Verfahren in der Lage, die Einflüsse von Paketverlusten, Amplitudenclipping und unterschiedlichen Hintergrundgeräuschen auf die Sprachverständlichkeit angemessen zu beurteilen.[233] Des Weiteren kann POLQA im Vergleich zu PESQ die Auswirkungen auf die Verständlichkeit von folgenden Effekten besser erfassen. Zum einen sind dies lineare Frequenzgangverzerrungen, Zeitdehnung und Zeitkompression, wie sie beispielsweise in Audio-over-IP-Anwendungen auftreten, zum anderen aber auch Verzerrungen durch Kodierungsverfahren und Nachhall.[234] Aus diesem Grund und weil POLQA auf insgesamt acht unterschiedliche Sprachen trainiert wurde, ist das Verfahren für die Verständlichkeitsmessung einer Vielzahl an unterschiedlichen Audiokodierungsverfahren geeignet.[235] Abgesehen davon erlaubt Perceptual Objective Listening Quality Analysis im Gegensatz zu vielen anderen Algorithmen, die Auswirkungen des Wiedergabepegels auf die Sprachverständlichkeit zu beurteilen.[236] Dies liegt daran, dass POLQA im Gegensatz zu seinem Vorgängerverfahren PESQ nicht voraussetzt, dass die Lautstärkepegel des Referenzsignals und des Testsignals gleich laut analysiert werden, sondern das Referenzsignal auf einen festgelegten Schallpegel von ca. 73 dB[237] setzt und den Lautstärkepegel des Testsignals dementsprechend anpasst. Dadurch kann das Testsignal im Vergleich zum festgelegten Schalldruckpegel eine Differenz zwischen -20 dB und +6 dB aufweisen.[238] Somit ergibt sich als Bedingung für das zu analysierende Signal eine Schallpegel-Toleranz zwischen 53 dB und 78 dB (jeweils A-bewertet).[239] Wie bereits in Kapitel 3.3.10 (Perceptual Objective Listening Quality Analysis) erwähnt, ist ein weiterer Vorteil von POLQA, dass es in den Bewertungsmodi, die auch PESQ besitzt, also Schmalband und Breitband, genauere Messergebnisse liefert als der Vorgänger.[240] Somit kann, wenn die Möglichkeit besteht, POLQA nahezu immer anstatt PESQ für die Sprachverständlichkeitsanalyse eingesetzt werden. Des Weiteren ist POLQA; bezüglich kommender Entwicklungen in der Sprachverständlichkeitsforschung, sehr anpassungsfähig gestaltet. So ist es ohne größere Probleme möglich, den Algorithmus aufgrund von neuen Erkenntnissen an den Stand der Technik anzupassen. Dadurch ist er noch flexibler und seine Zukunftssicherheit ist gewährleistet. Der letzte hier aufgeführte Vorteil von Perceptual Objective Listening Quality Analysis ist

[233] Vgl. Shen et al. (2024, S. 2).
[234] Vgl. Beerends et al. (2013, S. 401).
[235] Vgl. Shen et al. (2024, S. 2).
[236] Beerends et al. (2013, S. 386).
[237] Beerends et al. (2013, S. 390).
[238] Vgl. Beerends et al. (2013, S. 391).
[239] Vgl. Beerends et al. (2013, S. 401).
[240] Vgl. Beerends et al. (2013, S. 396).

die Bearbeitung des Referenzsignals. POLQA entfernt, bevor es mit der Messung beginnt, leises Hintergrundrauschen im eingeladenen Referenzsignal.[241] Dies führt dazu, dass das eigentliche Analyseergebnis präziser wird, da nur das Nutzsignal des Referenzsignals berücksichtigt wird. Hintergrund für dieses Vorgehen ist unter anderem, dass eine Aufnahme eines komplett rauschfreien Sprachsignals aufgrund von technischen Gegebenheiten nur selten möglich ist. Die meisten Signale besitzen daher ein, wenn auch nur leichtes, Hintergrundrauschen, welches sich jedoch negativ auf das Ergebnis der Sprachverständlichkeitsmessung auswirken kann. Allerdings darf dieser Vorteil von POLQA nicht missverstanden werden. Ein Trugschluss wäre es, anzunehmen, dass die Qualität des Referenzsignals dadurch nicht mehr hoch sein muss. Denn ein zu stark verrauschtes Referenzsignal wirkt sich im Endeffekt sehr negativ auf die Verständlichkeitsanalyse von POLQA aus und führt somit zu falschen Messergebnissen.

Dies lässt sich auch als ersten Nachteil des Verfahrens aufzählen, denn durch diesen Effekt liefert POLQA nur dann realistische Messergebnisse, wenn eine gewisse Qualität des ausgewählten Referenzsignals gewährleistet ist.[242] Dies muss der Nutzer bzw. die Nutzerin des Algorithmus vor Auswahl des Referenzsignals also weiterhin beachten. Außerdem besitzt POLQA den Nachteil, dass es ungeeignet für die Verständlichkeitsanalyse von Musikaufnahmen ist. Weiterhin ist es ungenau, wenn Messungen an Kodierungsverfahren durchgeführt werden, mit denen POLQA nicht trainiert wurde, bzw. die es nicht kennt. Abgesehen von den technischen Nachteilen besitzt POLQA den Makel, dass es sich bei ihm um einen lizenzrechtlich geschützten Standard handelt. Dementsprechend ist er aktuell nicht frei verfügbar und es gibt bisher keine Implementierungen für Programmierumgebungen. Bei der Entscheidung für POLQA muss also bedacht werden, dass die Benutzung mit Lizenzkosten verbunden ist. Dies macht das Verfahren im Vorhinein schon kostspieliger als einige bereits erwähnte Messalgorithmen, da von diesen zum Teil komplett kostenlose Implementierungen vorhanden sind. Auch führt dieser Umstand dazu, dass es nicht mühelos erscheint, den Algorithmus in ein Programmgerüst einzupflegen. Als letzten Nachteil von POLQA ist auch bei diesem Verfahren zu erwähnen, dass es rechenintensiv ist; insbesondere bei großen Datenmengen kann dies zu Verzögerungen bei der Auswertung führen. Unter anderem auch deswegen ist es de facto weniger gut für Echtzeit-Berechnungen geeignet.

[241] Beerends et al. (2013, S. 387).
[242] Vgl. Beerends et al. (2013, S. 401).

Als Letztes seien die möglichen Anwendungsbeispiele des Perceptual Objective Listening Quality Analysis-Verfahrens erwähnt. Dadurch, dass der Algorithmus für die Verständlichkeitsmessung von einigen Audiokodierungsverfahren ausgelegt wurde, ist er entsprechend universell bei der Evaluation und Entwicklung von neuen Kodierungsverfahren einsetzbar. Aufgrund der drei unterschiedlichen Bewertungsmodi (Schmalband, Breitband und Super-Breitband) ist der Standard beispielsweise für Festnetztelefonsysteme und 2G-Mobilfunk, aber auch für Anwendungen geeignet, bei denen Audiosignale mit einem größeren Frequenzspektrum übertragen werden wie HD-Audio-over-IP. Des Weiteren wird er bei der Hardware-Entwicklung von beispielsweise Telefonen oder Headsets sowie Freisprecheinrichtungen verwendet. Außerdem ist POLQA infolge seiner Ähnlichkeit mit PESQ für nahezu alle Einsatzmöglichkeiten, in denen auch PESQ eine Anwendung findet, prädestiniert.[243] Weiterhin findet POLQA Einsatz auch in militärischen Systemen, bei denen die Sprachverständlichkeit der übertragenen Audiosignale sichergestellt werden muss. Von den aufgezählten Anwendungsgebieten abgesehen ist es denkbar, POLQA wegen seiner Fähigkeit, den Einfluss von Nachhall auf die Sprachverständlichkeit zu berücksichtigen, auch bei Verständlichkeitsmessungen von Räumen zu verwenden.

### 4.3.11 Automatic Mean-Opinion-Score

Als erster Vorteil des Automatic Mean-Opinion-Score ist zu nennen, dass von dem Bewertungsverfahren frei verfügbare Python Implementierungen zur Verfügung stehen. Dadurch ist das Verfahren leicht zugänglich und kann theoretisch ohne Komplikationen in einem vorhandenen Programmgerüst installiert werden. Außerdem handelt es sich bei dem Algorithmus um ein KI-gestütztes Messverfahren. Daraus resultiert der Vorteil, dass es stetig und unkompliziert verbessert werden kann. Dies gelingt unter anderem, indem es mit besseren Trainingsdaten geschult wird. Da diese Verfahren problemlos Verbesserungen erfahren können, sind sie zukunftssicher und flexibel. Nicht zuletzt sei die Möglichkeit erwähnt, dass KI-basierte Bewertungsverfahren sehr genau die Fehlerquellen, bzw. Schwachstellen in einem Audiosignal angeben können. Dies ermöglicht es, eventuell bei zukünftigen Algorithmen nicht nur ein Ergebnis der MOS-Bewertung zu erhalten, sondern auch direkt eine genaue Begründung bzw. einen Verweis auf die genaue Stelle im Sprachsignal, die den endgültigen MOS-Wert am stärksten beeinflusst hat.

[243] Vgl. Beerends et al. (2013, S. 401).

Allerdings besitzen KI-gestützte-Analyseverfahren auch Nachteile. So ist zum Beispiel ein großer Makel dieser Systeme, dass ihre Leistung stark davon abhängig ist, mit welchen Trainingsdaten sie angelernt worden sind. Im konkreten Fall der Sprachverständlichkeitanalyse bedeutet dies, dass sie – um zu korrekten Messergebnissen zu führen – explizit auf die Anwendungsgebiete geschult sein müssen. Sind hingegen die Trainingsdaten sehr unähnlich den eigentlichen Daten, mit denen die Analyse durchgeführt wird, führt das Verfahren zu einem sehr ungenauen MOS-Wert.[244] Dies zeigte sich auch bei den Ergebnissen der „The VoiceMOS Challenge 2023“, einer Veranstaltung, bei der es um die Bewertungsleistung von Messverfahren geht, welche versuchen, den Mean-Opinion-Score ohne subjektiven Test, also objektiv vorherzusagen. Hier zeigte sich, dass keines der Verfahren eine konsistente Bewertungsleistung bei allen Tests mit demselben Modell und Trainingsdaten erreichte.[245] Dementsprechend kamen die Teilnehmer zu dem Schluss, dass diesbezüglich die Forschung noch zu besseren Ergebnissen kommen müsse. Ein weiterer Nachteil von KI-Messverfahren ist, ähnlich einem subjektiven MOS-Test (siehe Kapitel 4.3.1 (Mean-Opinion-Score)), die fehlende Referenz. Dieser Umstand ist problematisch, da es schwierig ist, den Verfahren das Trainieren der MOS-Bewertung von Signalen beizubringen, wenn bereits bei subjektiven MOS-Tests keine klare Referenz für die Bewertungsergebnisse vorhanden ist, da diese meist unbewusst von den Testpersonen gewählt wird.[246] Diese Auswahl der Referenz wird außerdem sehr durch die Verfassung, in der sich die Testhörer und Testhörerinnen befinden, und welche Erfahrungen sie besitzen, geprägt. Des Weiteren spielt bei den Verfahren zur automatischen Berechnung des MOS-Wertes die Größe des Testdatensatzes eine entscheidende Rolle. Denn ist dieser zu klein, kann die Verständlichkeitsanalyse zu ungenau ausfallen und somit falsche bzw. unrealistische MOS-Werte als Ergebnisse liefern.[247] Des Weiteren dürfen die Trainingsdaten jedoch auch nicht zu groß sein. In diesem Fall könnte nämlich die Leistungsfähigkeit aufgrund der entstehenden hohen Latenz zu sehr eingeschränkt sein und das Verfahren sich somit beispielsweise nicht mehr für eine realistische Echtzeitanwendung anbieten.[248] Zuletzt sei noch als Nachteil aufgeführt, dass die Funktionsweise der KI-basierten Bewertungsverfahren kompliziert und undurchsichtig erscheinen kann. Aus diesem Grund ist es nicht immer eindeutig nachvollziehbar, warum der Algorithmus zu einer bestimmten Entscheidung gelangt ist. Als Beispiel sei hier eine hohe Einschätzung der Verständlichkeit einer eher unverständlichen

[244] Vgl. Zhou et al. (2024, S. 876).
[245] Vgl. Cooper et al. (2023, S. 6).
[246] Vgl. Hu, Yasuda und Toda (S. 546).
[247] Vgl. Hu et al. (S. 546).
[248] Vgl. Tseng, Kao und Lee (S. 1).

Audiodatei erwähnt. Dieser Nachteil führt dazu, dass die Entwicklungen und Verbesserungen solcher Verfahren insbesondere in der Anfangsphase komplex sind.

Die Anwendungsgebiete, in denen der Automatic Mean-Opinion-Score und auch andere KI-basierte Bewertungsverfahren verwendet werden, sind – trotz der aufgezählten Nachteile – vielfältig. Natürlich werden sie unter anderem dann eingesetzt, wenn kein Referenzsignal existiert und es auch nicht möglich ist, ein entsprechendes Vergleichssignal für die Analyse zu erstellen. Dies ist beispielsweise bei der Sprachverständlichkeitsmessung von Algorithmen, die Text-zu-Sprache kreieren, der Fall. Aber auch bei Echtzeitanwendungen, bei denen in einem bestehenden und laufenden System Verständlichkeitsprognosen gemacht werden müssen, sind KI-basierte Verfahren in Anwendung.

# 5 Versuch zur Auswahl des geeigneten Analysealgorithmus

Um das geeignetste Sprachverständlichkeitsmessverfahren für das Programm begründet auszuwählen, wurde ein Versuch ausgeführt. Dafür wurden drei Verfahren ausgewählt und an einem realen Audiosignal angewendet. Parallel dazu wurde die Verständlichkeit des gleichen Signals mittels eines minimalistischen Hörtests mit Testpersonen untersucht. Die Ergebnisse des Hörtests wurden mit denen der drei Sprachverständlichkeitsmessverfahren verglichen. Dieser Vergleich diente der Bewertung der Urteilsfähigkeit der einzelnen Verfahren, da die Hörtestergebnisse als eine Referenz für die real wahrgenommene Sprachverständlichkeit stehen. Die umfangreiche Beschreibung und Durchführung des Versuchs sowie die resultierenden Ergebnisse sind Inhalt dieses Kapitels.

## 5.1 Auswahl der drei Algorithmen für den Versuch

Ein Versuch, welcher alle aufgelisteten Algorithmen aus Kapitel 3.3 (Darstellung verschiedener Sprachverständlichkeitsmessverfahren) sinnvoll miteinander vergleicht, ist technisch unter anderem aufgrund der verschiedenen Analysearten der Verfahren nicht umsetzbar. Des Weiteren wäre er auch ohne großen wissenschaftlichen Mehrwert, da die Analyseverfahren auf unterschiedliche Anwendungsgebiete spezialisiert wurden. Beispielsweise ist es selbsterklärend, dass mit dem Articulation loss of consonants-Verfahren keine valide Verständlichkeitsbewertung eines Sprachkodierungsverfahrens abgegeben werden kann. Deswegen wurde für den Versuch eine Vorauswahl, bestehend aus drei der vorgestellten Sprachverständlichkeitsmessverfahren, getroffen.

Die drei Verfahren sind zum einen Short-Time Objective Intelligibility (STOI) und Perceptual Evaluation of Speech Quality (PESQ) als zwei Vertreter der intrusiven Messverfahren, und zum anderen eine Variante des Non-Intrusive Speech Quality Assessment (NISQA) Ansatzes als nicht-intrusives Bewertungsverfahren. Die Hintergründe der Wahl werden im Folgenden erläutert.

Zunächst ist zu erwähnen, dass einige Sprachverständlichkeitsmessverfahren aufgrund ihrer Methodik nicht als Analysealgorithmus für das zu entwickelnde Programmtool zur Auswahl standen. Dies liegt daran, dass das Tool hauptsächlich zur objektiven Verständ-

lichkeitsanalyse von bereits bestehenden Audiodateien angedacht ist. Dadurch ist es nicht möglich, ein Verfahren, welches eine subjektive Sprachverständlichkeitsmessungen durchführt, für das Programmtool auszuwählen. Ebenso sind auch Verfahren, die für die raumakustische Bewertung der Verständlichkeit ausgelegt sind, ungeeignet für den Einsatz im Programm, zum einen aufgrund der Inkompatibilität der zu erwartenden Messergebnisse, aber auch zum anderen – wie bereits erwähnt – wegen ihrer Messmethodik, beispielsweise, weil sie für die Bewertung der Sprachverständlichkeit den Ruhegeräuschpegel eines Raumes als Eingabeparameter benötigen. Aus diesen Gründen schieden die Verfahren: Articulation loss of consonants (Alcons), Artikulationsindex (AI), Speech Intelligibility Index (SII), Speech Transmission Index (STI), inklusive Rapid Speech Transmission Index (RASTI) und Speech Transmission Index for Public Address Systems (STI-PA) sowie der Mean-Opinion-Score (MOS) in seiner Grundform – als subjektives Bewertungsverfahren – aus.

Ein weiterer Punkt, der bei der Auswahl der drei Messverfahren eine nicht unerhebliche Relevanz hatte, ist die Lizenzfrage der einzelnen Algorithmen. Denn von nicht allen Sprachverständlichkeitsmessverfahren existiert eine kostenlose Open-Source-Software wie beispielsweise die Python-Implementierungen von STOI, PESQ und NISQA. Dies ist auch der Hauptgrund, weswegen keine Tests mit POLQA durchgeführt werden konnten. Denn POLQA ist – zumindest zum Stand dieser Ausführungen – noch ein nicht kostenlos verfügbarer Standard. Die Lizenz für die Anwendung besitzen zum Beispiel in Deutschland das Unternehmen Opticom GmbH und das Fraunhofer-Institut für Integrierte Schaltungen (IIS). Deswegen sind keine legalen frei zugänglichen Implementierungen von POLQA vorhanden. Kapitel 4.3.10 (Perceptual Objective Listening Quality Analysis) zeigte jedoch, dass die Messergebnisse von POLQA im Vergleich zu denen des Vorgängerverfahrens (PESQ) in allen Anwendungsgebieten präziser sind.

Als weitere Begründung bezüglich der Wahl der drei erwähnten Messverfahren sei die Varianz der Messmethoden erwähnt. Zwar wäre es im Gegensatz zu POLQA möglich gewesen, eine Implementierung des Perceptual Evaluation of Audio Quality (PEAQ)-Verfahren zu realisieren, allerdings wären dann ausschließlich drei intrusiv agierende Algorithmen Versuchsgegenstand gewesen. Zusätzlich ist zu erwähnen, dass solche Vergleiche bereits in einer Vielzahl in der Vergangenheit durchgeführt wurden. Abgesehen davon wäre das Ergebnis vermutlich auch vorhersehbar gewesen, da der Algorithmus, der die neuesten Entwicklungen in der Sprachverständlichkeitsforschung berücksichtigt, vermutlich die beste Verständlichkeitsbewertung ergeben hätte. Aus diesen Gründen wurden zwei etablierte intrusive Bewertungsverfahren und ein nicht-intrusives

Verfahren (NISQA) ausgewählt. Außerdem sei noch erwähnt, dass diese Auswahl noch einen weiteren Hintergrund hatte. Denn sollte der Versuch das Ergebnis liefern, dass das NISQA-Bewertungsverfahren der geeignetste Algorithmus für das Programm sei, könnten weitere Funktionen in das Programmgerüst mit eingefügt werden. Als Beispiel für diese erweiterten Funktionen des Programms sei die Möglichkeit von Echtzeitanalysen der Sprachverständlichkeit erwähnt. Diese sind nämlich aufgrund des fehlenden Referenzsignals nur mit nicht-intrusiven Verfahren sinnvoll durchführbar.

Die letzte hier ausgeführte Begründung für die Auswahl der drei Bewertungsverfahren beantwortet die Frage, warum eine Form des NISQA-Verfahrens und nicht der Automatic Mean-Opinion-Score (AutoMOS) ausgewählt wurde. Die im Versuch angewendete NISQA-Implementierung stammt von den Forschern Mittag, Naderi, Chehadi und Möller und wurde im Jahr 2021 vorgestellt.[249] Sie verwendet zum einen neueste Erkenntnisse der Sprachverständlichkeitsforschung in ihrer Analyse und ist zum anderen verglichen mit anderen nicht-intrusiven Verfahren sehr präzise bezüglich ihrer Messergebnisse.[250] Da mittlerweile nahezu alle zukunftsorientierten nicht-intrusiven Bewertungsverfahren eine Künstliche Intelligenz beinhalten, hängt ihr Urteilsvermögen in erheblichem Maße von der Leistungsfähigkeit ihrer KI ab. Es scheint offensichtlich, dass durch Forschungsarbeiten die Anzahl und Qualität der Fähigkeiten der Künstlichen Intelligenz in den letzten Jahren rapide gestiegen sind.[251] Daher ist es besonders interessant, zu untersuchen, wie ein aktuelles KI-basiertes nicht-intrusives Bewertungsverfahren im Vergleich zu zwei etablierten intrusiven Verfahren abschneidet. Da die Entwicklung des AutoMOS schon 2016 erfolgte, ist das Verfahren von Mittag et al. von 2021 fünf Jahre jünger. Mit Blick auf die Weiterentwicklung der KI in dieser Zeit ist deswegen davon auszugehen, dass AutoMOS ein weniger genaues neuronales Netzwerk als NISQA von Mittag et al. besitzt. Dementsprechend lässt sich ableiten, dass AutoMOS auch unpräzisere Ergebnisse der Verständlichkeitsmessung generieren wird, als das Verfahren von Mittag et al.

## 5.2 Implementierungen der drei Algorithmen

Wie bereits in Kapitel 5.1 (Auswahl der drei Algorithmen für den Versuch) erwähnt, werden im Versuch die Algorithmen STOI, PESQ und NISQA verwendet. In diesem Kapitel sollen die drei Software-Implementierungen, die die drei unterschiedlichen Bewertungsverfahren repräsentieren, vorgestellt werden. Es ist nochmals zu erwähnen, dass jede

[249] Mittag, Naderi, Chehadi und Möller (2021, S. 2127-2131).
[250] Vgl. Mittag et al. (2021, S. 2130).
[251] Vgl. Bolkart (S. 6).

der Implementierungen auf der Website github.com kostenlos zur Verfügung steht und somit ohne den Kauf einer Lizenz verwendbar ist. Demnach ist es mittels folgender Beschreibung möglich, die kompletten Implementierungen nachzubilden.

Als eine Version des Short-Time Objective Intelligibility-Verfahren wurde die Python-Implementierung „pystoi“[252] von Manuel Pariente ausgewählt. Es handelt sich dabei um eine direkte Adaption der Matlab-Implementierung „STOI – Short-Time Objective Intelligibility Measure –“ von Cees Taal.[253] Diese wurde von Taal als Open-Source-Software zur Verfügung gestellt. Zwar hätte auch direkt die Matlab-Implementierung von Taal für den Versuch ausgeführt werden können, allerdings sollten aus Gründen der Einheitlichkeit alle drei Bewertungsverfahren mittels eines Python-Codes angewendet werden. Da der vorgegebene Code zur Ausführung der pystoi-Implementierung (siehe Abbildung 1) keine Berücksichtigung der Abtastfrequenz unternimmt, musste dieser noch entsprechend angepasst werden.

```
1 import soundfile as sf
2 from pystoi import stoi
3
4 clean, fs = sf.read('path/to/clean/audio')
5 denoised, fs = sf.read('path/to/denoised/audio')
6
7 # Clean and den should have the same length, and be 1D
8 d = stoi(clean, denoised, fs, extended=False)
```

*Abbildung 1: Vorlage des Codes zur Ausführung der pystoi-Implementierung, Quelle: (Pariente).*

Zur besseren Zugänglichkeit und einfacheren Fehlersuche wurde der Code in der integrierten Entwicklungsumgebung „Thonny“[254] geschrieben. Wie in Abbildung 2 zu sehen ist, wurde der Programmcode – zur Ausführung der Implementierung – angepasst. Um eine bessere Übersichtlichkeit zu gestalten, wurde eine Funktion, welche die Implementierung ausführt, definiert und im Anschluss aufgerufen. Außerdem musste – wie bereits erwähnt – auch die Variable der Abtastrate berücksichtigt werden. Deswegen wurde diese mit dem Namen „sr“ für „Samplingrate“ definiert. Im Anschluss an diese Definition erfolgt der Sicherheitstest, ob es sich bei den eingeladenen Dateien um Mono- oder Stereodateien handelt. Sollte es sich um Stereosignale handeln, werden diese durch Anwendung der Numpy-Funktion „mean“ auf Monosignale konvertiert. Danach wird noch eine Sicherheitsabfrage gestellt, ob versehentlich Dateien mit drei oder mehr Spuren eingeladen wurden. Wenn diese Abfrage positiv ist, wird ein sogenannter „ValueError“ ausgelöst, der dies dem Nutzer bzw. der Nutzerin mitteilt. Des Weiteren wurde noch eine

[252] Pariente.
[253] C. Taal.
[254] Thonny.

einfache Ausgabe des berechneten STOI-Wertes mittels der „print“-Funktion ergänzt. Zuletzt wurde der Dateipfad, in der sich die zu analysierenden Audiodateien befanden, entsprechend angepasst. Die eigentliche Implementierung, also die Funktion „stoi“ (siehe Quellcode (Abbildung 2) Zeile 24), wurde jedoch nicht verändert.

```
1  import pystoi
2  import soundfile as sf
3  import numpy as np
4
5  def calculate_stoi(reference_file, distorted_file):
6      # Audiodateien einlesen
7      reference_audio, _ = sf.read(reference_file)
8      distorted_audio, _ = sf.read(distorted_file)
9
10     # Sampling Rate (muss für beide Dateien gleich sein)
11     sr = 16000  # Beispielwert. Wert stets auf Samplingrate der Dateien einstellen!!!
12
13     # Konvertiere zu Mono, falls die Dateien Stereo sind
14     if reference_audio.ndim == 2:
15         reference_audio = np.mean(reference_audio, axis=1)
16     if distorted_audio.ndim == 2:
17         distorted_audio = np.mean(distorted_audio, axis=1)
18     if reference_audio.ndim >= 3:
19         raise ValueError("Die Referenzdatei darf maximal eine Stereo-Datei sein.")
20     if distorted_audio.ndim >= 3:
21         raise ValueError("Die Testdatei darf maximal eine Stereo-Datei sein.")
22
23     # STOI-Berechnung durchführen
24     stoi_score = pystoi.stoi(reference_audio, distorted_audio, sr)
25     return stoi_score
26
27 reference_file = 'audio/ref.wav'  # Pfad zur Referenzdatei
28 distorted_file = 'audio/dist.wav'  # Pfad zur degradierten Datei
29
30 score = calculate_stoi(reference_file, distorted_file)
31 print(f"STOI-Wert: {score}")
```

*Abbildung 2: angepasster Code zur Ausführung der pystoi-Implementierung, Quelle: der Verfasser.*

Als Perceptual Evaluation of Speech Quality-Verfahren kam die „PESQ“ Python-Implementierung von Miao Wang, Christoph Boeddeker, Rafael G. Dantas und Ananda Seelan aus dem Jahr 2022 zum Einsatz.[255] Diese bietet entsprechend der Definition des PESQ-Verfahrens (siehe Kapitel 3.3.9 (Perceptual Evaluation of Speech Quality)) zwei Analysearten an: Zum einen den Narrowband-, also Schmalbandmodus und zum anderen den Wideband, also Breitbandmodus. Es gilt zu beachten, dass die beiden Audiosignale für die Schmalbandanalyse eine Abtastrate von 8 kHz und für die Breitbandanalyse eine Abtastrate von 8 kHz oder 16 kHz aufweisen müssen. Ähnlich der Anpassung des ausführenden Programmcodes von pystoi wurden auch bei dem von PESQ kleine Überarbeitungen in der Thonny Umgebung durchgeführt, hauptsächlich, um die Ausführung der Implementierung stabiler zu gestalten. Die vorgeschlagene Version des Codes importiert die notwendigen Module, lädt die beiden Audiodateien ein und ruft innerhalb des Aufrufs der „print“-Funktion die eigentliche PESQ-Funktion auf (siehe Abbildung 3). Dies ist zwar ein sehr effizienter Code, allerdings nicht sehr robust. Das Programm versucht, die Funktion auch auszuführen, wenn die Abtastraten der Dateien

[255] Vgl. Wang, Boeddeker, Dantas und seelan (2022).

nicht übereinstimmen. Des Weiteren kann die Funktion nur Monodateien verarbeiten. Dies wird aber in dem vorgeschlagenen Funktionsaufruf nicht im Vorhinein überprüft.

```
1 from scipy.io import wavfile
2 from pesq import pesq
3
4 rate, ref = wavfile.read("./audio/speech.wav")
5 rate, deg = wavfile.read("./audio/speech_bab_0dB.wav")
6
7 print(pesq(rate, ref, deg, 'wb'))
8 print(pesq(rate, ref, deg, 'nb'))
```

*Abbildung 3: Vorlage des Codes zur Ausführung der PESQ-Implementierung, Quelle: (Wang, Boeddeker, Dantas & seelan, 2022).*

Deswegen wurde der Code, der die PESQ-Implementierung aufruft, entsprechend angepasst (siehe Abbildung 4).

```
1  from scipy.io import wavfile
2  from pesq import pesq
3  import numpy as np
4
5  # Lade die Referenz- und verzerrten Sprachdateien
6  reference_rate, reference_audio = wavfile.read('audio/ref.wav')
7  distorted_rate, distorted_audio = wavfile.read('audio/dist.wav')
8
9  # Überprüfe, ob die Abtastraten übereinstimmen
10 if reference_rate != distorted_rate:
11     raise ValueError("Die Abtastraten der beiden Dateien stimmen nicht überein.")
12
13 # Konvertiere zu Mono, falls die Dateien Stereo sind
14 if reference_audio.ndim == 2:
15     reference_audio = np.mean(reference_audio, axis=1)
16 if distorted_audio.ndim == 2:
17     distorted_audio = np.mean(distorted_audio, axis=1)
18 if reference_audio.ndim >= 3:
19     raise ValueError("Die Referenzdatei darf maximal eine Stereo-Datei sein.")
20 if distorted_audio.ndim >= 3:
21     raise ValueError("Die Testdatei darf maximal eine Stereo-Datei sein.")
22
23 # Berechne den PESQ-Wert für Breitband
24 score = pesq(reference_rate, reference_audio, distorted_audio, 'wb')
25
26 print(f'PESQ-Wert (Breitband): {score}')
```

*Abbildung 4: angepasster Code zur Ausführung der PESQ-Breitbandanalyse-Implementierung, Quelle: der Verfasser.*

Zum einen wurden wieder die Dateipfade auf die Speicherorte der beiden Audiodateien für die Analyse angepasst, zum anderen erfolgt zunächst in Zeile 10 die Abfrage, ob die beiden Abtastraten der Dateien identisch sind. Im Anschluss daran erfolgt erneut, analog zu dem ausführenden Code von pystoi, die Abfrage, ob Stereo-Dateien oder Dateien mit mehr als drei Kanälen eingeladen wurden. Sollte dies nicht der Fall sein, wird ein „ValueError“ ausgelöst. Erst nach diesen Sicherheitsabfragen erfolgt die Übergabe an die Analysefunktion zur Berechnung des Breitband PESQ-Wertes. Aus Gründen der Vollständigkeit sei noch erwähnt, wie der Aufruf der PESQ-Funktion aussehen muss, um

eine Schmalbandanalyse zu aktivieren. Hierbei sind die Zeilen 1 bis 22 des in Abbildung 4 präsentierten Codes zu übernehmen, nur die letzten vier Zeilen sind entsprechend Abbildung 5 anzupassen.

```
23 # Berechne den PESQ-Wert für Schmalband
24 score = pesq(reference_rate, reference_audio, distorted_audio, 'nb')
25
26 print(f'PESQ-Wert (Schmalband): {score}')
```

*Abbildung 5: Unterschied zu Ausführung der PESQ-Schmalbandanalyse, Quelle: der Verfasser.*

Der Vorteil der PESQ-Implementierung von Miao Wang, Christoph Boeddeker, Rafael G. Dantas und Ananda Seelan ist, dass anhand mitgelieferter Audiodateien geprüft werden kann, ob alle Schritte der Implementierung erfolgreich waren. So sind im downloadbaren Archiv der Implementierung zwei Audiodateien enthalten, deren Analyseergebnisse bereits bekannt sind. Die PESQ-Werte dieser Dateien liegen bei einer Breitbandmessung bei 1.0832337141036987 und bei einer Analyse im Schmalbandmodus bei 1.6072081327438354.[256] Die für den Versuch ausgeführte Implementierung der PESQ-Funktion berechnete – für die beiden mitgelieferten Audiodateien – die gleichen Werte. Deswegen kann davon ausgegangen werden, dass die Implementierung erfolgreich war und die Funktion im Sinne der Entwickler läuft.

Wie bereits in Kapitel 5.1 (Auswahl der drei Algorithmen für den Versuch) erwähnt, stammt die dritte Python-Implementierung von Gabriel Mittag, Babak Naderi, Assmaa Chehadi und Sebastian Möller und ist aus dem Jahr 2021.[257] Da dieses Verfahren mit einem relativ komplexen neuronalen Netzwerk arbeitet, war es nötig, die Open-Source-Distribution „Anaconda“[258] zu installieren. Nach der Installation konnten im sogenannten „Anaconda Prompt“ – einer Befehlszeilen-Schnittstelle – die folgenden Befehle[259] eingegeben werden:

```
(1) conda env create -f env.yml
(2) conda activate nisqa
```

Der erste Befehl erstellt dabei die Umgebung und installiert alle benötigten Module und Bibliotheken, während der zweite im Anschluss die neue Umgebung aktiviert. Danach kann auch schon mit der Berechnung der Sprachverständlichkeit eines Audiosignals begonnen werden. Dabei gibt es drei Möglichkeiten, mittels NISQA den MOS-Wert zu berechnen. So können (1) eine einzelne Wavedatei, (2) alle Dateien in einem bestimmten

[256] Vgl. Wang et al. (2022).
[257] Vgl. Mittag et al. (2021, S. 2127).
[258] Continuum Analytics.
[259] Mittag.

Ordner oder (3) alle Wavedateien in einer Tabelle mit dem Dateiformat „CSV“ analysiert werden. Nachfolgend sind die im Anaconda Prompt einzugebenden Befehle[260] aufgelistet:

(1) `python run_predict.py --mode predict_file --pretrained_model weights/nisqa_mos_only --deg /path/to/wav/file.wav --output_dir /path/to/dir/with/results`

(2) `python run_predict.py --mode predict_dir --pretrained_model weights/nisqa_mos_only --data_dir /path/to/folder/with/wavs --num_workers 0 --bs 10 --output_dir /path/to/dir/with/results`

(3) `python run_predict.py --mode predict_csv --pretrained_model weights/nisqa_mos_only --csv_file files.csv --csv_deg column_name_of_filepaths --num_workers 0 --bs 10 --output_dir /path/to/dir/with/results`

Es ist zu erwähnen, dass jeweils die letzten Teile der Befehle durch die entsprechenden Speicherorte zu ersetzen sind, und diese entsprechend für den Versuch angepasst wurden. Die Ergebnisse der Analyse werden zum einen im Anaconda Prompt angezeigt, zum anderen aber auch in einer neu angelegten CSV-Datei in Form einer Tabelle gespeichert. Des Weiteren ist erwähnenswert, dass an den Modellen der NISQA-Implementierung sogenanntes „Finetuning“[261] vorgenommen werden kann. Dies bedeutet, dass die Modelle an ein anderes Datenset angepasst werden können. Des Weiteren ist es sogar möglich, NISQA auf ein komplett neues Modell umzustellen und zu trainieren. Allerdings konnte dies aufgrund des damit verbundenen erheblichen Mehraufwandes nicht für die Durchführung des Versuchs realisiert werden. Außerdem würde ein vorher speziell auf die zu analysierende Audiodatei trainiertes NISQA-Verfahren ein besseres Analyseergebnis liefern. Somit wäre das Versuchsergebnis stark verzerrt, und der Versuchsaufbau entspräche weniger den realen Bedingungen, unter denen ein Sprachverständlichkeitsmessverfahren ein Audiosignal bewerten muss.

## 5.3 Versuchsvorbereitung: Auswahl des Testsignals

Bei der Auswahl des Hörbeispiels für den Versuch wurde sich für ein Audiosignal entschieden, welches in dieser Art häufig in der realen TV-Broadcast-, und zum Teil auch in der Hörfunk-Umgebung gesendet wird. So wurde festgelegt, dass es sich bei dem Testsignal um einen Ausschnitt aus dem Sendeton einer Live-Fußballübertragung handeln soll.

[260] Mittag.
[261] Mittag.

Bei Fernseh-Live-Übertragungen wird ein Geschehen oft durch die Erläuterungen eines Kommentators oder einer Kommentatorin begleitet. Der Sendeton lässt sich in diesem Fall meist in zwei Arten von Signalen unterteilen. Zum einen in das Sprachsignal, welches den Kommentar enthält und zum anderen in ein Atmosphären-Audiosignal (Atmo-Signal). Dieses besteht aus Hintergrundgeräuschen, wie zum Beispiel im Fall der Fußballübertragung den Fan-Gesängen, sowie den Ball-Trittgeräuschen der Spieler bzw. Spielerinnen. Zweifelsfrei ist die Verständlichkeit des Sprachsignals in diesem Sendeformat essenziell. Jedoch folgen nicht alle Live-Übertragungen diesem Aufbau. So gibt es beispielsweise auch Sendungen, die musikalische Darbietungen als Sendeschwerpunkt haben. Ein weiteres Beispiel sind Fernsehshows, bei denen ein Moderator oder eine Moderatorin anstatt eines Kommentator bzw. einer Kommentatorin auftritt. Auch bei dieser Art von Show ist die Sprachverständlichkeit des Sendetons von elementarer Bedeutung. Allerdings ist diese zum Teil anders zu bewerten. Dies liegt daran, dass eine Moderation im fernsehtechnischen Kontext eine Position beschreibt, die hauptsächlich vor der Kamera agiert. Somit ist für den Zuschauer bzw. die Zuschauerin die Gestik und im Besonderen die Mimik – wie beispielsweise die Lippenbewegungen – dieser Person zu sehen, während Sprachsignale gesendet werden. Wie bereits zu Beginn des Kapitels 3.1 (Sprachverständlichkeit) ausgeführt, ist die Sprachverständlichkeit auch durch diesen visuellen Kontakt zum Sprecher bzw. der Sprecherin beeinflusst. Diese Beeinflussung kann jedoch bisher nicht durch ein gängiges Verständlichkeitsmessverfahren berücksichtigt werden. Dementsprechend darf sie bei der Bewertung der Sprachverständlichkeit durch die Testpersonen in diesem Versuch auch keine Rolle spielen. Um diesen Effekt zu sondieren, wurden auch keine Sprachsignale von Moderatoren oder Moderatorinnen für die Versuchsdurchführung verwendet, sondern Sprachsignale von Kommentatoren. Denn das Präsenz-Verhältnis ihrer Sprache ist im Vergleich zu ihrem Erscheinungsbild im Sendesignal sehr unausgeglichen. So sind sie, wenn überhaupt, meist nur wenige Sekunden zu sehen, während sie über mehrere Sendeminuten zu hören sind. Unter anderem aus diesen Gründen bietet sich für den Versuch die Imitation eines Audiosignals einer Fußballübertragung an, die auditiv ein Atmo-Signal und ein Kommentatoren-Signal enthält.

Es ist zu erwähnen, dass diese Art von Testsignal sehr speziell ist. Denn ein Versuch, der die Sprachverständlichkeit eines beispielsweise verrauschten Sprachsignals misst, würde universellere Messergebnisse liefern. Jedoch ist das endgültige Programmtool, wie bereits erwähnt, hauptsächlich für die Analyse von TV-Broadcast-Audiosignalen gedacht. Außerdem ist ein Test mit dem erwähnten Testsignal durchaus praxisnäher als ein

Test mit Rauschen. Aus diesen Gründen ist die Auswahl des Audiosignals für den Versuch trotzdem gerechtfertigt.

Es konnten reale Aufnahmen, die während einer Fußball-Live-Übertragung aufgezeichnet wurden, für den Versuch verwendet werden. Diese wurden im Kontext des Tests eines neuen 3D-Audio-Mikrofons der Marke „Audio-Technica“ durchgeführt. Nachfolgend werden jedoch nur Gegebenheiten dieses Mikrofontests erwähnt, welche für die Versuchsvorbereitung des Versuchs zur Auswahl des geeigneten Analysealgorithmus von Bedeutung sind. So stammen die Aufnahmen des Mikrofontests vom 21.10.2023 und wurden während der Begegnung des „Sport-Club Freiburg e. V.“ und des „Verein für Leibesübungen Bochum 1848 Fußballgemeinschaft e. V“ im Europa-Park-Stadion in Freiburg im Breisgau erstellt. Es wurde das Sprachsignal des Kommentators sowie ein Atmo-Signal aufgezeichnet. Der Kommentator war wie gewöhnlich mit einem Headset der Marke „beyerdynamic“ mit Modellnamen DT 797 PV (siehe Abbildung 6) mikrofoniert.

*Abbildung 6: „beyerdynamic“ DT 797 PV, Quelle: (Beyerdynamic, 2024a).*

Die Kopfhörerbauform dieses Headsets ist geschlossen und als Mikrofon ist eine Kondensatorkapsel mit Nierencharakteristik verbaut. Wie bereits erwähnt ist es zusammen mit wenigen anderen Modellen die Standardwahl als Headset im Kommentatoren-Bereich des TV-Broadcasts. Diese Headsets werden nahezu immer an sogenannten Kommentator-Einheiten betrieben. In diesem Fall kam eine Einheit der Marke „RIEDEL“, die den Namen „Commentary-Control-Panel (CCP)-1116“ trägt, zum Einsatz (siehe Abbildung 7). Diese Einheiten ermöglichen dem Kommentator oder der Kommentatorin zum einen, sich selbst stummzuschalten sowie die Abhörlautstärken individuell einzustellen, zum anderen aber auch beispielsweise mit der Redaktion oder dem Regisseur zu kommunizieren, ohne dass diese Kommunikation gesendet wird.

*Abbildung 7: “RIEDEL” Commentary-Control-Panel (CCP)-1116, Quelle: (RIEDEL, 2024).*

Das Atmo-Signal wurde mittels eines 3D-Audio-Mikrofons aufgezeichnet. Dafür wurde das „ORTF-3D“ der Marke „SCHOEPS“ verwendet. Es ist ein acht-kanaliges Mikrofon mit vier Mikrofonkapselpärchen, die in einer ORTF-Anordnung in einem Windkorb angebracht sind. Es besteht aus vier CCM41 und vier CCM41V Mikrofonkapseln. Alle acht Kapseln besitzen eine Supernieren-Charakteristik. Es ist zu erwähnen, dass die Version des „ORTF-Surround“ von „SCHOEPS“ ebenfalls vier CCM41 Kapseln besitzt und als Hauptmikrofon in den meisten deutschen Fußballstadien, in denen Spiele der 1. Bundesliga ausgetragen werden, standardmäßig fest verbaut ist. Auch aus diesem Grund wurde das dem Standard sehr ähnliche „ORTF-3D“ für die Erstellung der Testdateien verwendet. Allerdings wurde das Atmo-Mikrofon für die Aufzeichnung nicht, wie im fest verbauten Zustand, an das Stadiondach montiert, sondern zwecks der besseren Handhabung auf der Pressetribüne auf einem großen Stativ aufgebaut. Den genauen Aufbau des Mikrofons zeigt die Abbildung 8. Die eigentliche Aufzeichnung der beiden Audiosignale wurde mittels der Digital-Audio-Workstation (DAW) „Reaper“ sowie einem Audiointerface der Marke RME-Audio und der Modellbezeichnung „MADIface USB“ realisiert. Dabei wurden die Signale direkt als „Direct Out“ Ausgänge aus der Input-Sektion des LAWO-Broadcastmischpults geroutet und erfuhren demnach keine weiteren audiotechnischen Anpassungen vor der Aufzeichnung.

*Abbildung 8: Aufbau des "ORTF-3D" (linksseitig hängend auf dem Stativ), Quelle: der Verfasser.*

## 5.4 Versuchsvorbereitung: Bearbeitung der Testsignale

Die beiden Audiodateien, also das Sprachsignal des Kommentators und das Atmo-Signal, benötigten zur Anwendung der Verständlichkeitsmessung durch die Messverfahren und den Hörtest einige Schritte der Bearbeitung. Um die genauen Versuchsbedingungen nachvollziehen zu können, sind diese Bearbeitungen in diesem Kapitel beschrieben.

Da die Analyse der Sprachverständlichkeit nicht im 3D-Audio-Kontext, sondern lediglich im Stereobereich durchgeführt werden sollte, musste zunächst das Atmo-Signal des „SCHOEPS ORTF-3D“ entsprechend bearbeitet werden. Von den acht Kanälen wurden entsprechend der Kanalbelegung (siehe Abbildung 9) die beiden Signale für vorne-unten-links und vorne-unten-rechts ausgewählt. Dazu wurde die achtkanalige Mehrspurdatei in die frei verfügbare Software Audacity in der Version 3.6.1 geladen. Im Anschluss daran wurden die ersten beiden Spuren als linke und rechte Spur der entstehenden Stereodatei definiert und entsprechend als wav-Datei mit den gleichen Exporteigenschaften wie die Originaldatei exportiert. Die Kommentator-Audiodatei musste diesbezüglich nicht angepasst werden, da es sich bei ihr um eine Monodatei handelte. Nach diesem Schritt wurden beide Audiodateien in die DAW „Nuendo 12“ in der Version 12.0.70 eingeladen. Da beide Aufzeichnungen die exakt gleiche Spiellänge besaßen, musste kein Time-Alignement durchgeführt werden. Zur besseren Übersichtlichkeit wurden die zwei Spuren des Atmo-Signals bei dem Dateiimport aufgeteilt. Daher musste noch das entsprechende Stereo-Panning dieser beiden Spuren durchgeführt werden.

**CHANNEL ALLOCATIONS OF THE ORTF-3D SETUP**

| Channel # | Labelling on the cable outlet and the Multicore breakout cable | **Channel routing** only valid if small side is front and windshield is hanging |
|---|---|---|
| 1 | Bottom layer: Ch 1 (L) – Yellow | L (LEFT) |
| 2 | Bottom layer: Ch 2 (R) – Red | R (RIGHT) |
| 3 | Bottom layer: Ch 3 (LS) – Blue | LS (LEFT SURROUND) |
| 4 | Bottom layer: Ch 4 (RS) – Green | RS (RIGHT SURROUND) |
| 5 | Top layer: Ch 1 (L) – Yellow | LTF (LEFT TOP FRONT) |
| 6 | Top layer: Ch 2 (R) – Red | RTF (RIGHT TOP FRONT) |
| 7 | Top layer: Ch 3 (LS) – Blue | LTR (LEFT TOP REAR) |
| 8 | Top layer: Ch 4 (RS) – Green | RTR (RIGHT TOP REAR) |

*Abbildung 9: Kanalbelegung des ORTF-3D von „SCHOEPS“, Quelle: (Schoeps GmbH, 2024, S. 8).*

Im Anschluss daran erfolgte die Auswahl möglicher Stellen, die sich für eine Sprachverständlichkeitsmessung eignen. Alle Stellen, in denen die Kommentar-Spur kein Sprachsignal enthielt, konnten aus den 139 Minuten langen Signalen gekürzt werden. Im Anschluss wurde nach Passagen gesucht, in denen das Atmo-Signal relativ laute und diffuse Publikumsgeräusche wie beispielsweise Torjubel oder Pfiffe enthielt. Dies wurde deswegen durchgeführt, da während dieser Phasen das Atmo-Signal und das Kommentator-Signal den geringsten Pegelunterschied besitzen, und dementsprechend davon auszugehen ist, dass die Sprachverständlichkeit nicht bei 100 % liegt. Im Hinblick auf die Endanwendung des zu entwickelnden Programmtools wären dies die Momente, bei denen eine objektive Auswertung der real-vorhandenen Verständlichkeit sehr von Nutzen wäre. Ein weiterer Kritikpunkt bei der manuellen Auswahl der passenden Stelle für das Testsignal des Versuchs war, dass der Kommentator in diesen Momenten möglichst keine Spielernamen bzw. generell keine Eigennamen nennen sollte. Denn dies könnte die Verständlichkeitsmessung sowohl bei den Hörtests als auch im Besonderen bei dem nicht-intrusiven Messalgorithmus verzerren, da diese für die Hörer und Hörerinnen und höchstwahrscheinlich auch den Messverfahren unbekannt sein könnten. Außerdem wurde darauf geachtet, dass der Kommentator einen ganzen zusammenhängenden Satz in diesem Moment sprach. Nach dieser Vorauswahl blieben vier Stellen übrig, bei denen die genannten Bedingungen zutreffen. Bei zwei dieser Passagen war das Atmo-Signal zu Beginn des Satzes des Kommentators noch leise und wurde im Verlauf der

Ausführungen lauter. Dies schien für das Testsignal ungeeignet, da es vielleicht dazu geführt hätte, dass sich beispielsweise die Testhörer schon einen kurzen Moment lang an die Stimme gewöhnen konnten. Von den beiden verbleibenden Stellen hatte die erste eine Länge von ca. 18 Sekunden, die zweite jedoch nur eine Länge von 5 Sekunden. 5 Sekunden erschienen für den Versuch zu kurz, daher wurde die 18-Sekunden-Passage für das Testsignal ausgewählt.

Um Clipping am Anfang und Ende der drei Spuren zu vermeiden, wurden diese mit kurzen Ein- bzw. Ausblendungen innerhalb der DAW-Umgebung versehen. Im Anschluss daran mussten die für die Sprachverständlichkeitsmessverfahren notwendigen Bedingungen erfüllt werden. Da es sich bekanntlich bei STOI und PESQ um zwei intrusive-Bewertungsverfahren handelt, benötigen beide für die Analyse der Verständlichkeit ein möglichst sauberes – also mit wenigen Störgeräuschen versehenes – Referenzsignal. Dieses lag durch die Audiospur des Kommentators vor. Zwar war es nicht komplett störgeräuschfrei, jedoch ist es, wie bereits erwähnt, in der Realität auch nahezu unmöglich, ein entsprechend perfektes Referenzsignal für die Analyse bereitzustellen. Dementsprechend wurde das Kommentator-Signal als Referenzsignal für die beiden intrusiven Bewertungsalgorithmen verwendet. Da im fernsehtechnischen Kontext stets ein Durchschnitts-Lautheitswert von -23 LUFS gefordert ist, sollten beide Signale auch diesen Wert aufweisen. Daher wurden für die Erstellung des Referenzsignal zunächst die beiden Spuren des Atmo-Signals gemutet und eine Lautheitsmessung durchgeführt. Nach dieser Messung wurde der Gesamtausgabepegel so angepasst, dass das exportierte Referenzsignals eine Lautheit von -23 LUFS aufwies. Bei den Exporteinstellungen war zu beachten, dass die Abtastrate der fertigen Datei durch die Begrenzung von PESQ (siehe 5.1 (Auswahl der drei Algorithmen für den Versuch)) auf 16 kHz festgelegt werden musste. Nachdem der Export des Referenzsignals aus der DAW erfolgt war, konnte nun die Erstellung des Testsignals ausgeführt werden. Dazu wurde die Mute-Einstellung der beiden Spuren des Atmo-Signals deaktiviert und im Anschluss der Lautheitspegel des entstehenden Signals so abgesenkt, dass -23 LUFS als Lautheitswert eingestellt war. Danach konnte auch diese Datei exportiert werden. Zum Schluss wurden die beiden Dateien in die angegebenen Ordner der einzelnen Implementierungen der Messalgorithmen verschoben. Zuletzt mussten noch die Dateinamen entsprechend den ausführenden Programmcodes von pystoi und PESQ (siehe Abbildung 2 und Abbildung 4) auf „dist“ für das Testsignal und „ref“ für das Referenzsignal angepasst werden.

## 5.5 Versuchsvorbereitung: Ausarbeitung der Messskala der Hörtests

Zwei der drei Sprachverständlichkeitsmessverfahren, die im Versuch angewendet werden, basieren auf der MOS-Bewertung, welche in einer Skala zwischen eins und fünf definiert ist. Zwar weicht die Bewertung durch den PESQ-Algorithmus davon leicht ab, da die obere Grenze der Skala auf 4,5 und die theoretische untere Grenze auf -0,5 definiert ist, jedoch entspricht die zugrunde liegende Skala trotzdem dem MOS-Wert. Abgesehen davon umfasst der Wertebereich die gleiche Ausdehnung, sodass der PESQ-Wert gegebenenfalls mühelos in einen korrekten MOS-Wert umgerechnet werden kann. Aus diesen Gründen wurde bei der Auswahl der Messskala für den Hörtest die Skala des Mean-Opinion-Score-Verfahrens festgelegt, bei der der Wert 1 einer äußerst schlechten Sprachverständlichkeit und der Wert 5 einer tadellosen Verständlichkeit entspricht. Die genauen Umrechnungen der Ergebnisse der Sprachverständlichkeitsmessverfahren werden jedoch erst in den beiden Kapiteln der Ergebnispräsentation der Algorithmen definiert, in Kapitel 5.9 (Messergebnis des STOI-Algorithmus) und Kapitel 5.10 (Messergebnis des PESQ-Algorithmus). Um den Probanden und Probandinnen des Hörtests jedoch eine Hilfestellung bezüglich der Verständlichkeitseinschätzung der Hörprobe zu geben, sollte die Skala nicht nur Zahlenwerte zwischen eins und fünf, sondern auch Adjektive als Beschreibung enthalten. Diese Definitionen orientieren sich an denen des MOS-Wertes (siehe Kapitel 3.3.1 (Mean-Opinion-Score)). Wie nachfolgend in aufsteigender Reihenfolge ausgeführt, wurden die Beschreibungen gewählt: „sehr schlecht“, „schlecht“, „passabel“, „gut“ und „ausgezeichnet“. Der komplette Fragebogen, der den Testpersonen zu Beginn des Hörtests ausgehändigt wurde, ist in Anlage 1 dargestellt.

## 5.6 Versuchsvorbereitung: Hörtest mit Testpersonen

Für die Umsetzung des Hörversuchs wurde die Datei, welche das in Kapitel 5.4 (Versuchsvorbereitung: Bearbeitung der Testsignale) erwähnte Testsignal enthielt, bereitgestellt. Des Weiteren erfolgte noch die Ausarbeitung des Fragebogens (siehe Anlage 1). Dieser sollte zwar aus zwei Teilen bestehen, Ziel war jedoch, beide Teile auf einer DIN-A4-Seite darzustellen. Der erste Teil bestand aus der Abfrage der persönlichen Daten und der zweite aus der Qualitätseinschätzung der Sprachverständlichkeit des

Hörbeispiels. Die Abfrage der personenbezogenen Daten der Testpersonen bezog sich auf ihr Alter und Geschlecht sowie die Frage, ob sie ein Hörgerät während des Versuchs verwendeten. Die Ergebnisse dieser Fragestellung sollten der Kategorisierung der Versuchsergebnisse dienen und sind in Kapitel 5.12 (Aufschlüsselung der personenbezogenen Daten der Testpersonen) dargestellt.

## 5.7 Versuchsaufbau des Hörtests für Testpersonen

Der Versuchsaufbau des Hörtests bestand im Wesentlichen aus einem Kopfhörer, der mittels Audiokabel direkt an den Klinken-Audioausgang eines Computers angeschlossen war. Auf dem PC befand sich die Audiodatei mit dem Testsignal und die Mediaplayer-Software „VLC media player“ in der Version 3.0.20. Es handelte sich um einen Kopfhörer mit geschlossener Bauform der Marke „beyerdynamic“ mit dem Modellnamen „Custom Studio“. Die Wahl eines geschlossenen Kopfhörers lässt sich begründet, dass der Hörtest so wenig wie möglich durch Hintergrundgeräusche beeinflusst werden sollte. Des Weiteren besitzt das Modell Custom Studio Schieberegler an beiden Ohrmuscheln. Diese bieten die Möglichkeit, den Frequenzgang des Kopfhörers zu verändern. Es sind die vier Einstellungen „light bass“, „linear“, „bass boost“ und „vibrant bass“ auswählbar (siehe Abbildung 10). Da für den Hörtest eine möglichst unbeeinflusste Wiedergabe durch die Frequenz-Charakteristik des Kopfhörers gewünscht war, wurde die Einstellung linear gewählt. Wie in Abbildung 11 zu sehen, folgt bei dieser Einstellung die Wiedergabe möglichst genau dem Eingangssignal, das in den Kopfhörer gespielt wird.

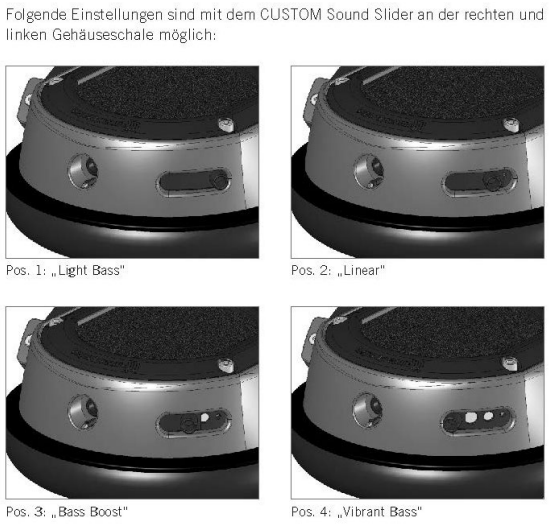


*Abbildung 10: Einstellungen des „beyerdynamic“ Custom Studio, Quelle: (Beyerdynamic, 2024b, S. 12).*

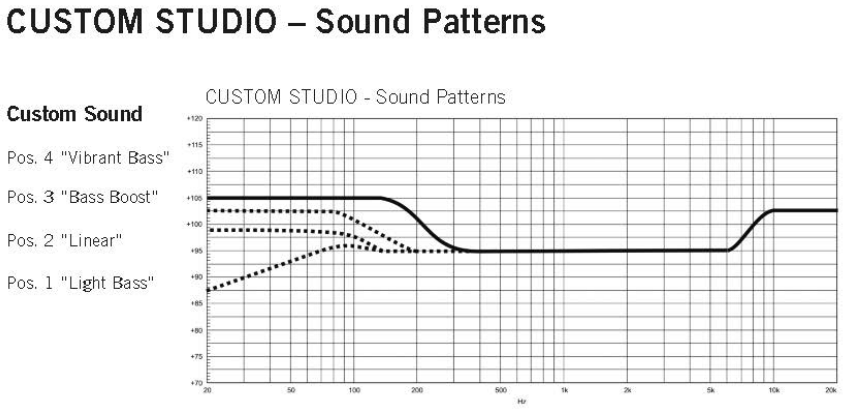


*Abbildung 11: Frequenzgang des „beyerdynamic“ Custom Studio, Quelle: (Beyerdynamic, 2024b, S. 13).*

Der Kopfhörer war mittels des angeschlossenen Wendelkabels an den Audioausgang des Computers angeschlossen. Die Windows-Lautstärkeeinstellungen sowie die Lautstärke in der Abspielsoftware VLC media player wurde auf 100 % eingestellt. Diese Einstellungen wurden während der gesamten Versuchsdurchführung nicht verändert. Die Hörtests fanden in einem Raum mit einer Länge von 5,6 m, einer Breite von 4,9 m und einer Deckenhöhe von 2,4 m statt. Ein Stuhl, der relativ mittig im Raum positioniert war, diente den Probanden und Probandinnen während des Versuchs als Sitzgelegenheit.

## 5.8 Versuchsablauf mit den Testpersonen

Nachfolgend ist beschrieben, wie der Hörtest mit den Testpersonen durchgeführt wurde. Um bei dieser Beschreibung eine bessere Übersichtlichkeit zu gewähren, ist dies nachfolgend in Singularform ausgeführt. Allerdings wurde darauf geachtet, dass der Ablauf bei allen Testpersonen gleich war und ihnen die gleichen Informationen gegeben wurden.

Bevor die Testperson den Raum betrat, wurde der Versuchsaufbau gemäß Kapitel 5.7 (Versuchsaufbau des Hörtests für Testpersonen) entsprechend aufgebaut und alles für die Durchführung des Versuchs hergerichtet. Die Testperson wurde in den Raum gelassen und gebeten, auf dem Stuhl Platz zu nehmen. Sie wurde angewiesen, den oberen Teil des Fragebogens, also die „Angaben zur Kategorisierung der Ergebnisse“ auszufüllen und den restlichen Teil durchzulesen. Währenddessen passte der Versuchsleiter den Kopfhörer, ohne ihn aufzusetzen, ungefähr auf die Kopfgröße der Person an. Nachdem der Versuchsleiter wahrgenommen hatte, dass die Testperson die Aufgabe beendet hatte, machte er sie mit dem genauen Ablauf des Versuchs vertraut. Der Proband bzw. die Probandin wusste im Vorhinein nur wenig über den Versuch und ihm bzw. ihr wurde erst jetzt erklärt, dass für den Versuch ein Hörbeispiel vorgespielt wird und dieses be-

züglich seiner Sprachverständlichkeit bewertet werden soll. Nachfolgend definierte der Versuchsleiter noch die auf dem Fragebogen abgedruckte Begrifflichkeit der „akustischen Sprachverständlichkeit“. Er konkretisierte, dass hierbei nicht gemeint sei, ob, sondern wie gut bzw. wie leicht das Hörbeispiel verständlich sei. Des Weiteren erwähnte er den Begriff der Sprachsignal-Qualität zur besseren Anschaulichkeit des Bewertungskriteriums. Diese Erklärung musste gegeben werden, da diese Formulierung erst nachdem das Layout des Fragebogens ausgearbeitet worden war, ausgewählt wurde. Da jedoch im Vorhinein feststand, dass der Testperson die Begrifflichkeit der Sprachverständlichkeit sowieso erklärt werden müsste, wirkte sich dieser Umstand nicht negativ auf die Durchführung des Versuchs aus. Nachdem der Versuchsleiter die Rückfrage gestellt hatte, ob dies klar sei und ob noch Fragen zum Ablauf offen wären, setzte er der Person den Kopfhörer auf. Nachdem die finale Größe des Kopfhörers eingestellt und somit die korrekte Position sichergestellt war, begab sich der Versuchsleiter an den Computer, auf dem sich die Datei mit dem Hörbeispiel befand. Dann wurde die Testperson noch gebeten, während des Abspielens der Audiodatei die Augen geschlossen zu halten. Erst nach Ablauf der Datei sollten die Augen geöffnet werden und eine Einschätzung in Form von einer Ankreuzung auf dem Fragebogen abgegeben werden. Im Anschluss wurde die Testperson gefragt, ob sie bereit sei, das Hörbeispiel zu hören. Nachdem diese Frage bejaht wurde, stellte der Versuchsleiter sicher, dass keine Umgebungsgeräusche vorhanden waren und zählte laut von drei auf null, bevor er die Wiedergabe der Hörbeispieldatei startete. Nachdem das Hörbeispiel abgespielt war, konnte die Person den Kopfhörer absetzen und ihre Einschätzung der Verständlichkeit ankreuzen. Abschließend ist noch zu erwähnen, dass während der Durchführung des Versuchs nur die Testperson und der Versuchsleiter im Raum anwesend waren.

## 5.9 Messergebnis des STOI-Algorithmus

Für die Messung der Sprachverständlichkeit des Testsignals wurde zunächst innerhalb der Software Thonny der in Abbildung 2 zu sehende Programmcode aufgerufen und die Variable „sr“, welche die Abtastrate der zu analysierenden Signale angibt, auf 16000 Hz eingestellt. Da die Dateien des Referenzsignals und Testsignals bereits im entsprechenden Projektordner abgespeichert waren, konnte direkt mit der Ausführung des Programmcodes begonnen werden. Der Algorithmus lieferte bei der erstmaligen Anwendung nach ca. 3,75 Sekunden das Messergebnis. Aus Sicherheitsgründen wurde die Implementierung auch ein zweites Mal ausgeführt. Bei der zweiten Ausführung dauerte die Berechnung nur ca. 1,49 Sekunden. Beide Messungen kamen zum gleichen Ergebnis:

einem STOI-Wert von 0,7235219427725607. Nach der Berechnung meldete die Thonny Entwicklungsumgebung, dass der Programmcode zwei Warnungen provozierte. Diese beliefen sich jedoch beide nur auf den Umstand, dass im Code in Zeile 27 und 28 jeweils eine globale Variable definiert wird, die innerhalb des Aufrufs der Funktion durch eine lokale Variable überschrieben wird. Da es sich jedoch um die Variablen „reference_file“ und „distorted_file“ handelt, ist dieser Effekt gewollt. Hintergrund ist, dass so wenig wie mögliche Variablen bei der Erstellung des Codes verwendet werden sollten, um den Programmcode so einfach wie möglich zu gestalten. Daher können die beiden Warnungen der Thonny Umgebung bewusst ignoriert werden.

Abschließend muss bei dem Ergebnis des Short-Time Objective Intelligibility-Verfahren noch beachtet werden, dass der Wert nicht als ein MOS-Wert zu verstehen ist, sondern eine eigene Skala besitzt. Wie bereits in Kapitel 3.3.6 (Short-Time Objective Intelligibility) beschrieben, ist diese Skala von 0 bis 1 definiert. Das Ergebnis ließe sich nach den zugehörigen Beschreibungen der Verständlichkeit innerhalb der einzelnen Zahlenbereiche also als moderate Sprachverständlichkeit definieren, da es zwischen 0,5 und 0,75 liegt. Um jedoch das Messergebnis besser mit denen der anderen beiden Analysealgorithmen zu vergleichen, ist es in Prozent umzurechnen. Deswegen lässt sich abschließend die Aussage treffen, dass die Berechnung des STOI-Wertes der Testdatei ergab, dass diese eine 72,3522%ige Sprachverständlichkeit aufweist.

## 5.10 Messergebnis des PESQ-Algorithmus

Zunächst wurde der Programmcode zur Ausführung einer PESQ-Breitbandanalyse (siehe Abbildung 4) in der Entwicklungsumgebung Thonny geöffnet. Im Gegensatz zur Ausführung der STOI-Implementierung musste bei diesem Code keine Abtastrate eingestellt werden, da diese durch die Anwendung des Algorithmus im Breitbandmodus bereits auf 16 kHz festgelegt ist. Da sich auch in diesem Projektordner bereits die beiden Audiodateien befanden, wurde die Implementierung des intrusiven Bewertungsverfahrens PESQ im Anschluss ausgeführt. Die Zeit, die der Algorithmus für die erste der beiden Berechnungen benötigte, lag bei ca. 2,03 Sekunden. Bei der zweiten Analyse lag die Bearbeitungszeit bei ca. 1,10 Sekunden. Auch in diesem Fall führten beide Auswertungen zu dem gleichen Ergebnis, das bei 1,4726208448410034 lag. Zwar gab es nach der Ausführung des Programmcodes keine Warnungen durch die Thonny Umgebung selbst, jedoch wurden zwei von der PESQ-Implementierung ausgegeben. Die ausgegebenen Warnungen wurden als „WavFileWarning“ geführt und bezogen sich auf die bei-

den Audiodateien. Sie enthielten die Aussage, dass die zwei wav-Audiodateien Chunks enthielten, die nicht ausgelesen werden konnten und deswegen ignoriert bzw. übersprungen wurden. Die Recherche nach diesen Warnmeldungen ergab, dass sie durch die Bibliothek „Scipy“ ausgelöst wurden und meist daraus resultieren, dass diese die Metadaten der zu lesenden wav-Datei nicht richtig erfassen konnten. Da die Metadaten der beiden Audiodateien jedoch keinen bedeutenden Einfluss auf die Sprachverständlichkeitsbewertung haben, konnte diese Warnung ebenfalls ignoriert werden.

Genaugenommen gibt das Perceptual Evaluation of Speech Quality-Verfahren keinen gewöhnlichen MOS-Wert als Ergebnis aus. Die Skala des MOS-Verfahren ist – wie bereits erwähnt – von 1 bis 5 definiert; Die Bewertungsskala von PESQ hingegen von -0,5 bis 4,5. Deswegen musste, um einen validen Vergleich der Ergebnisse der drei Analysealgorithmen durchführen zu können, das Ergebnis des PESQ ebenfalls umgerechnet werden. Aus Gründen der Vergleichbarkeit wurde auch hier eine Umrechnung in eine prozentuale Angabe der Sprachverständlichkeit gewählt. Das Ergebnis dieser Umrechnung und damit der Verständlichkeitsbewertung durch den PESQ-Algorithmus betrug 39,4524 %.

## 5.11 Messergebnis des NISQA-Algorithmus

Da die NISQA-Implementierung nicht innerhalb der Thonny Umgebung realisiert worden, sondern mittels eines Anaconda Prompts auszuführen war, gestaltete sich die Vorgehensweise auch anders als bei den ersten beiden Algorithmen. Deswegen musste zur Ausführung des Analysealgorithmus zunächst der Anaconda Prompt geöffnet und dort die NISQA-Implementierung aktiviert werden. Dies geschah mittels des Befehls:

```
conda activate nisqa_env
```

Im Anschluss daran musste noch der Dateipfad ausgewählt werden, in dem der Ordner mit dem ausführenden Programmcode sowie die Audiodateien installiert waren. Dies erfolgte mit dem Befehl:

```
cd C:\Users\elias\OneDrive\Dokumente\nisqaUmgebung\nisqa
```

Nun konnte die Sprachverständlichkeitsanalyse des Testsignals gestartet werden. Dies geschah durch Eingabe des bereits in Kapitel 5.2 (Implementierungen der drei Algorithmen) erwähnten Befehls:

```
python run_predict.py --mode predict_dir --pretrained_model
weights/nisqa_mos_only.tar --data_dir
/Users/elias/OneDrive/Dokumente/nisqaUmgebung/nisqa/audio --num_workers 0 --bs
10 --output_dir /Users/elias/OneDrive/Dokumente/nisqaUmgebung/nisqa/output
```

Die Analyse des NISQA-Algorithmus dauerte bei der erstmaligen Ausführung ca. 13,81 Sekunden. Im Anschluss daran wurde die zweite Messung absolviert. Bei dieser lag das Ergebnis nach ca. 5,93 Sekunden vor. Bei beiden Analysen lag das Endergebnis bei 1,91069. Auch bei der Anwendung des NISQA-Verfahrens wurden danach Warnmeldungen im Anaconda Prompt ausgegeben. Zum einen eine „FutureWarning“ und zum anderen eine „UserWarning“. Die erste Warnung bezog sich auf die Verwendung eines Befehls aus der „PyTorch“ Bibliothek. Da bei dem Aufruf des Befehls ein Argument nicht definiert wurde, wurde er mit dem „default“-Wert, also dem Standardwert für dieses Argument, aufgerufen. Die Warnung gab an, dass dieser default-Wert jedoch bald in einer neueren Version der Bibliothek verändert werden würde, und dass dies zukünftig beachtet werden sollte. Dementsprechend konnte diese Warnmeldung ignoriert werden, da dieser Wert bei der Ausführung der Implementierung noch richtig war. Die zweite Warnung, also die „UserWarning“, wurde durch die „librosa“ Bibliothek erzeugt. Sie gab an, dass aufgrund der geringen Abtastrate von 16 kHz nicht alle Filter des neuronalen Netzwerks, welche die Frequenzwahrnehmung des menschlichen Ohrs nachahmen, angewendet werden konnten und daher einige Ergebnisse dieser Filter leer sind. Dies ist also eine Warnmeldung, die auf die geringe Abtastrate des Testsignals zurückzuführen ist. Dies bestätigte sich darin, dass bei der Probe-Ausführung des Algorithmus mit einer Datei mit 48 kHz Abtastrate diese Warnmeldung nicht auftrat. Zwar hätte für die NISQA-Analyse auch ein Testsignal mit einer entsprechenden Abtastrate exportiert werden können, jedoch wären dann die Ergebnisse der drei Analysealgorithmen nicht mehr valide vergleichbar gewesen. Deswegen konnte auch diese Warnung ignoriert werden.

Zwar gibt das NISQA-Verfahren im Gegensatz zum PESQ-Algorithmus eine MOS-Bewertung des Audiosignals ab, jedoch sollte diese ebenfalls aus Gründen der Einheitlichkeit und besseren Vergleichbarkeit mit den Ergebnissen des STOI-Verfahrens in eine Prozentangabe umgerechnet werden. Das Ergebnis des NISQA-Bewertungsalgorithmus bezüglich der Sprachverständlichkeit des Testsignals liegt demnach bei 22,7673 %.

Ergänzend sind noch zwei Dinge zu erwähnen, die zu Testzwecken ausgeführt wurden: So wurde eine Verständlichkeitsanalyse des Referenzsignals durch die NISQA-Implementierung durchgeführt. Der entsprechende Ablauf dieser Messung ist analog zu der beschriebenen Analyse des Testsignals. Der Algorithmus errechnete einen MOS-Wert

der Referenzdatei von 2,878098. Dies entspricht einer Sprachverständlichkeit von 46,9525 %. Des Weiteren wurde eine weitere Testsignal-Datei mit der gleichen Methodik, wie in Kapitel 5.4 (Versuchsvorbereitung: Bearbeitung der Testsignale) beschrieben, jedoch mit einer Abtastrate von 48 statt 16 kHz erstellt und mittels des NISQA-Algorithmus analysiert. Die erste Auswertung dieser Datei dauerte etwa 7,57 Sekunden und die zweite ca. 5,45 Sekunden. Auch bei diesen beiden Messungen waren die Ergebnisse identisch und beliefen sich auf einen MOS-Wert von 2,882135. Die prozentuale Sprachverständlichkeit liegt in diesem Fall bei 47,0534 %.

## 5.12 Aufschlüsselung der personenbezogenen Daten der Testpersonen

Der Versuch wurde mit insgesamt sechs Testpersonen durchgeführt und es gelang, eine paritätische Verteilung von weiblichen und männlichen Personen zu erreichen. Dabei trug kein Proband und keine Probandin während des Tests – bzw. generell im Alltag – ein Hörgerät, weswegen diesbezüglich auch kein Ergebnis gesondert zu betrachten ist. Die jüngste Probandin war 30 Jahre und der älteste Proband 73 Jahre alt. Da es sich um eine kleine Testpersonengruppe handelt, seien hier die Alter der restlichen vier Probanden bzw. Probandinnen aufgeführt. Diese lagen bei 34, 60, 63 und 66. Die Testpersonen waren also im Durchschnitt 54,33 Jahre alt. Damit liegt das Durchschnittsalter der Versuchsgruppe zwischen dem durchschnittlichen Alter der Bevölkerung in Deutschland im Jahre 2023 und dem Durchschnittsalter der Zuschauer des „Ersten Deutschen Fernsehens“ im Jahr 2022. Diese beiden Alter liegen nämlich laut Statista bei 44,6[262] bzw. bei 64[263] Jahren. Der Altersdurchschnitt der Testpersonengruppe entspricht damit in etwa dem durchschnittlichen Alter der Hörer und Hörerinnen der Audiodateien, die mit dem fertigen Tool analysiert werden.

## 5.13 Ergebnisse des Hörtests

Von den sechs Testpersonen wählten vier das Kästchen „ausgezeichnet (5)“ als Einschätzung der Sprachverständlichkeit des Hörbeispiels. Eine Probandin wählte die Ankreuzmöglichkeit „gut (4)“ und ein Proband „passabel (3)“. Damit liegt die durchschnittli-

[262] Statistisches Bundesamt (2024).
[263] DWDL.de (2022).

che Bewertung der Verständlichkeit des Hörbeispiels bei 4,5. Zur Übersichtlichkeit und besseren Einschätzung dieses Wertes sei er in Prozent umgerechnet. Diese Umrechnung ergibt eine Sprachverständlichkeit von 87,5 %.

Erstaunlicherweise wählten die beiden jüngeren Testpersonen schlechtere Verständlichkeitseinschätzungen als die vier Testpersonen mit einem Alter über 60 Jahren. So wählte die 30-jährige Probandin aus, dass das Hörbeispiel nur eine gute Verständlichkeit besäße, und der 34-jährige Proband schätzte die Sprachverständlichkeit als passabel ein. Hierzu soll die Anmerkung einer 60-jährigen Testperson erwähnt sein. Sie war eine der letzten Personen und fragte nach der Durchführung des Versuchs, wie die Einschätzung der anderen Probanden und Probandinnen ausgefallen sei. Der Versuchsleiter teilte ihr mit, dass die beiden jüngeren Testpersonen als einzige nicht die Einschätzung „ausgezeichnet“ wählten. Sie stellte die These auf, dass die Hörerfahrung der älteren Versuchspersonen noch geprägt sei von Röhrenfernsehern, analoger terrestrischer Fernsehübertragung und schlechten Einbau-Lautsprechern in den TV-Geräten. Daher sei ihre Wahrnehmung der Sprachverständlichkeit divergent, zumal es sich bei dem Hörbeispiel um einen TV-Beitrag handelte. Zwar ist diese These interessant, jedoch kann aufgrund der kleinen Größe der Testpersonengruppe diese Beeinflussung auf die Testdaten nicht aussagekräftig untersucht werden. Daher soll dies hier nur als Kommentar angeführt werden.

Ein anderes Feedback von einem Probanden, der die Verständlichkeit als ausgezeichnet einschätzte, beinhaltete, dass das Hörbeispiel laut genug gewesen sei, wenig störende Hintergrundgeräusche vorhanden wären und deswegen die Einschätzung auf den Wert 5 fiel. Dies ist insofern interessant, da der Signal-Rauschabstand zwischen dem Kommentatoren-Sprachsignal und dem Hintergrund Atmo-Signal im Hörbeispiel nur 3,5 LU betrug. Zur Einschätzung dieses Wertes sei hier die Empfehlung aus den technischen Produktionsrichtlinien von ARD, ZDF und ORF erwähnt. Diese sagt aus, dass der Signal-Rauschabstand von Sprache zu Musik oder Geräuschen zwischen 7 und 10 LU liegen soll.[264] Da jedoch während einer Live-Sendung im Broadcast der Effekt eintreten kann, dass der Signal-Rauschabstand kurzzeitig außerhalb dieses Wertebereichs liegt, wurde bewusst für das Hörbeispiel ein Wert kleiner als 7 LU gewählt.

Als Letztes sei noch der Kommentar eines anderen Probanden ausgeführt, der ebenfalls die Einschätzung wählte, dass die Sprachverständlichkeit des Hörbeispiels ausgezeichnet sei. Er relativierte seine Einschätzung damit, dass er zwar 100 % des gesprochenen

[264] Vgl. TPRF HDTV 2016, S. 16.

Wortes des Kommentators verstanden hätte, er sich jedoch vorstellen könnte, dass dies anderen Hörern und Hörerinnen nicht leichtfallen könnte. Er meinte damit, dass die Sprachverständlichkeit zwar vorhanden sei, jedoch eventuell nur mit einer gewissen Höranstrengung das komplette Sprachsignal verstanden werden könnte. Diese Aussage deutet darauf hin, dass – bei einer größeren Testgruppe – nicht alle Personen die Verständlichkeit als ausgezeichnet einschätzen würden. Dies spiegelte sich auch im Gesamtergebnis des Hörtests wider, da es zwei Abweichungen von der besten Einschätzung der Sprachverständlichkeit im Endergebnis gab. Daher ist davon auszugehen, dass trotz der kleinen Versuchsgruppe ein annähernd realistisches Umfrageergebnis erstellt wurde.

## 5.14 Vergleich der Ergebnisse

Zu besseren Übersichtlichkeit seien zu Beginn dieses Kapitels nochmal die Ergebnisse der drei Messalgorithmen und des Hörtests aufgelistet. Der STOI-Algorithmus sagte eine 72,3522%ige Sprachverständlichkeit voraus. Bei PESQ betrug die errechnete Verständlichkeit 39,4524 % und bei NISQA 22,7673 %. Der subjektive Hörtest ergab eine Sprachverständlichkeit von 87,5 %.

Wie bereits in Kapitel 4.1 (Kategorisierung der Verfahren) erwähnt, ist die subjektive Verständlichkeitsmessung mittels Testpersonen die genaueste Messmethode. Daher ist davon auszugehen, dass die tatsächliche Sprachverständlichkeit des Hörbeispiels in etwa bei 87,5 % liegt. Damit korreliert das Ergebnis des Short-Time Objective Intelligibility-Verfahrens mit 72,3522 % am höchsten mit dem tatsächlichen Ergebnis. Das Ergebnis des Perceptual Evaluation of Speech Quality-Verfahren war deutlich ungenauer und lag 48,0476 Prozentpunkte unterhalb der tatsächlichen Verständlichkeit. Am schlechtesten schnitt das Non-Intrusive Speech Quality Assessment-Verfahren ab. Dieses lag 64,7327 Prozentpunkte unterhalb des Ergebnisses des Hörtests.

Der große Unterschied zwischen dem Analysewert des NISQA-Verfahrens und dem des STOI-Algorithmus erscheint zunächst unerklärlich. Insbesondere, da die Verständlichkeit, die NISQA vorhersagt, sehr weit von dem tatsächlichen Wert der Sprachverständlichkeit des Testsignals entfernt ist. Allerdings ist zu erwähnen, dass, wie bereits – zu Testzwecken – in Kapitel 5.11 (Messergebnis des NISQA-Algorithmus) ausgeführt, NISQA die Verständlichkeit des Referenzsignals nur auf 47,0534 % einschätzt. Die Sprachverständlichkeitsanalysen des STOI- und des PESQ-Verfahrens basieren jedoch auf der Annahme, dass das Referenzsignal eine 100%ige Verständlichkeit aufweist, und

deswegen das Testsignal mit diesem verglichen wird. Dies ist einer der Gründe, warum NISQA schlechtere Analyseergebnisse ausgibt. Dieser Effekt lässt sich – theoretisch betrachtet – dadurch ausblenden, indem die 22,7673 % als ein Wert auf der Skala von 0 % bis 47,0534 % betrachtet und erneut in Prozent umgerechnet werden. Diese Umrechnung würde zu dem Ergebnis führen, dass NISQA das Testsignal mit einer Sprachverständlichkeit von 48,3861 % bewertet. Dies zeigt, dass trotz der Berücksichtigung der Beeinflussung des genannten Effekts auf das Ergebnis des NISQA-Verfahrens der STOI-Algorithmus ein genaueres Ergebnis ausgibt.

## 5.15 Endergebnis des Versuchs

Der in Kapitel 5.14 (Vergleich der Ergebnisse) ausgeführte Vergleich der Analyseergebnisse der drei Sprachverständlichkeitsmessverfahren mit dem Ergebnis des Hörtests zeigt, dass das STOI-Verfahren die Verständlichkeit am genauesten vorhergesagt hat. Dies ist sogar der Fall, wenn berücksichtigt wird, dass das NISQA-Verfahren das Referenzsignal eher als verhältnismäßig schlecht verständlich einschätzt, und dieser Effekt – wie im vorherigen Kapitel ausgeführt – bei den Analyseergebnissen theoretisch ausgeblendet wird. Von den drei Sprachverständlichkeitsmessverfahren, die im Versuch angewendet wurden, repräsentiert also der STOI-Algorithmus das exakteste Bewertungsverfahren. Deswegen lässt sich schlussfolgern, dass die Verständlichkeitsbewertung von Audiodateien mit ähnlichem Signal, wie das des Hörbeispiels, am genausten mit dem Short-Time Objective Intelligibility-Algorithmus gelingen kann. Mit dem Hörbeispiel bzw. Testsignal dieses Versuchs wurde außerdem ein Audiosignal gewählt, welches im täglichen Live-Broadcastumfeld regelmäßig produziert wird und damit Bestand im tatsächlichen Sendeprogramm hat. Infolgedessen ist der STOI-Algorithmus für das Programmtool zur automatischen Berechnung der Verständlichkeit von Sprachsignal-Audiodateien aus dem Broadcastumfeld das genaueste Sprachverständlichkeitsmessverfahren.

Bei der Bestimmung des zweit genausten Bewertungsverfahrens sind theoretisch zwei Ergebnisse möglich. Werden die unangepassten Ergebnisse betrachtet, ist das Perceptual Objective Listening Quality Analysis-Verfahren anzuführen. Allerdings ist das Non-Intrusive Speech Quality Assessment-Verfahren genauer, wenn die Umrechnung des Ergebnisses – wie bereits erklärt – durchgeführt wird. Jedoch ist generell zu erwähnen, dass auch das angepasste Ergebnis des NISQA-Algorithmus – wie auch das des PESQ-Algorithmus – verhältnismäßig ungenau ist, und die Verfahren daher in dem durchgeführten Versuch keine valide Vorhersage der Sprachverständlichkeit generierten.

# 6 Theoretische Ausarbeitung des Programms

In diesem Teil der Arbeit soll zum einen die aus dem vorherigen Kapitel hervorgehende Auswahl des geeigneten Analysealgorithmus nochmals kritisch betrachtet werden und zum anderen die weiteren Funktionen, die das Programmtool enthalten soll, definiert werden. Des Weiteren soll ein Blockschaltbild des fertigen Programmtools dargestellt und das Aussehen sowie die Funktionen der Nutzeroberfläche ausgearbeitet werden.

## 6.1 Nutzerinterview mit Broadcast-Toningenieur

Um die in den nachfolgenden Kapiteln untersuchten Ausarbeitungsschritte des Programmtools zu begründen sowie neue Sichtweisen auf die notwendigen Gegebenheiten des Tools zu bekommen, wurde ein Interview durchgeführt. Da dieses maßgeblich die Ausarbeitung beeinflusste, seien die wichtigsten Hintergrundinformationen in diesem Kapitel übersichtlich dargestellt. Das komplette Interview ist in Anlage 2 abgebildet. Bei dem Interview handelte es sich um ein Nutzerinterview. Die Methodik des Nutzerinterviews schien für die Ausarbeitung der Funktionen des Programmtools und im Besonderen für die Ausarbeitung der Programmoberfläche sowie die Sicherstellung der einfachen Bedienbarkeit am geeignetsten. Dies liegt unter anderem daran, dass die geplanten Funktionen des Tools direkt auf Nützlichkeit geprüft werden konnten, und das Programm nutzerfreundlich gestaltet werden konnte. Bei dem Nutzer handelte es sich um einen freiberuflichen Toningenieur, der hauptberuflich in der Fernsehbranche tätig ist. Er ist 36 Jahre alt und weist eine 13-jährige Broadcast-Berufserfahrung auf. Da er deswegen als Fachkraft qualifiziert ist, scheint er geeignet, eine Fachmeinung sowie eine Anwendermeinung bezüglich der Ideen der Ausarbeitung des Programmtools abzugeben. Der Nutzer bekam vor der Durchführung des Interviews bereits Hintergrundinformationen, bezogen auf das generelle Anwendungsgebiet und die Grundfunktionen, die das fertige Tool besitzen sollte. Des Weiteren wurde ihm versichert, dass die Tonaufnahmen, die während des Interviews gemacht wurden, nicht veröffentlicht und seine Aussagen anonym behandelt werden. Dies sollte dafür sorgen, dass der Nutzer in seiner Meinungsäußerung während des Interviews nicht eingeschränkt war. Allerdings ist bei einem Nutzerinterview zu beachten, dass es oft persönliche Meinungen des Interview-Gastes bzw. der Interview-Gästin enthält. Dahingehend positiv zu erwähnen ist, dass der Nutzer beispielsweise bei den Fragen nach der Ergebnisdarstellung nicht nur seine eigene Meinung vertreten hat, sondern auch Einschätzungen darüber abgab, was Arbeitskollegen

und -kolleginnen von der Darstellung der Ergebnisse erwarten würden. Trotzdem ist zu beachten, dass es nicht pauschal möglich ist, aus den Ergebnissen des Interviews Rückschlüsse auf die durchschnittliche Meinung der Fachwelt zu ziehen.

## 6.2 Ausarbeitung der Funktionen des Programms

In diesem Kapitel sollen alle Funktionen des Programmtools, welche – über die Berechnung der Sprachverständlichkeit hinaus – Anwendung im fertigen Tool finden sollen, ausgeführt, erklärt und begründet werden. Die Reihenfolge, in der diese hier genannt sind, orientiert sich aus Übersichtlichkeitsgründen an der Bearbeitungskette, die die Audiodateien durchlaufen.

Zunächst ist hierbei die Funktion zu erwähnen, dass die Anwender und Anwenderinnen die Möglichkeit haben sollen, die zu analysierenden Audiodateien innerhalb des Programmfensters einladen zu können. Dies ist essenziell, da die Nutzer und Nutzerinnen in der fertigen Anwendung auch mehrere Dateien auswählen können sollen. Wäre dann kein Einfügen der Audiodateien in das Programmfenster möglich, sowie es beispielsweise bei der Anwendung der drei Analysealgorithmen der Fall ist (siehe Kapitel 5.2 Implementierungen der drei Algorithmen), würde dies die Nutzerfreundlichkeit in erheblichem Maße einschränken. Die genaue Funktion des Einfügens der Dateien soll auf drei Arten möglich sein. So soll zum einen die klassische Variante über eine Schaltfläche mit dem Namen „Dateien einladen“ umgesetzt werden. Nach Betätigen dieses Knopfes in der Programmoberfläche soll sich ein neues Fenster öffnen, in dem der Speicherort der Dateien ausgewählt und diese dann selektiert werden können. Zum anderen sollen die Nutzer und Nutzerinnen mittels eines Doppelklicks in den leeren Datei-Einladungsbereich zum gleichen Fenster, wie zuvor beschrieben, gelangen. Als letzte Variante soll eine „Drag and Drop“-Einladung realisiert werden, mit der die Dateien direkt in das Programmfenster abgelegt und somit eingefügt werden können. Um die Nutzer und Nutzerinnen auf diese drei Möglichkeiten hinzuweisen, wird ein entsprechender Informationstext im leeren Programmfenster angezeigt, der ausgeblendet wird, sobald die erste Datei eingeladen wird. Die Realisierung der drei unterschiedlichen Einladungsarten soll dazu führen, dass die Anwender und Anwenderinnen ihr Nutzungsverhalten so wenig wie möglich an das Programmtool anpassen müssen und somit eine gewisse Barrierefreiheit erreicht wird. Dies bestätigte auch der Nutzer im Interview, da er auch diese Formen des Einladens der Dateien vorschlug.

Die Idee zu einer anderen bedeutenden Funktion, die das fertige Programm besitzen sollte, kam dem Nutzer, während er im Interview gefragt wurde, wie die eingeladenen Audiodateien vor der Berechnung dargestellt werden sollten. Er führte an, dass es der Übersichtlichkeit dienen würde, wenn nach dem Laden der Dateien nicht nur die Dateinamen angezeigt werden würden, sondern auch Metadaten. Diese sollten in Form von Titel, Interpret, Dateilänge und Abtastrate ausgelesen und, wie im nächsten Abschnitt beschrieben, ausgegeben werden. Der Nutzer sagte, dass dies im Besonderen bei der Analyse einer großen Anzahl an Audiodateien hilfreich sei. Des Weiteren käme es manchmal vor, dass Dateien ähnlich bzw. nicht eindeutig benamt sind, insbesondere, wenn die Sprachverständlichkeit von ähnlichen Audiodateien verglichen werden soll. In solchen Fällen würde die Auslesung und Darstellung der Metadaten der einzelnen Audiodateien schnell Eindeutigkeit generieren.

Als weitere Funktion sei die Darstellung der eingeladenen Dateien in dem Programmfenster erwähnt. Den Nutzern und Nutzerinnen sollte eine Tabelle präsentiert werden, die folgenden Aufbau besitzt: Die Zeilen repräsentieren die einzelnen eingeladenen Dateien. Die Spalten die Datei-Attribute, also Einlade-Reihenfolge, sowie die Metadaten: Titel, Interpret, Erstellungsdatum, Dateilänge, Abtastrate und Bitrate der einzelnen Dateien. Dabei sollen die Metadaten schon beim Laden der Audiodateien ausgelesen werden, sodass diese schon vor dem Ausführen der Analyse dem Nutzer bzw. der Nutzerin angezeigt werden können. Dies würde dazu führen, dass den Anwendern und Anwenderinnen eine gewisse Übersichtlichkeit über die eingeladenen Audiodateien gegeben wird. Auch der Nutzer sprach diesen Aspekt im Interview an und erwähnte das Beispiel, bei dem andernfalls auch unabsichtlich Dateien eingefügt werden könnten. Des Weiteren ist diese Art der Darstellung eine gute Möglichkeit, um den Anwendern und Anwenderinnen Rückmeldung zu geben, ob die Audiodateien korrekt eingefügt wurden und ob sie überhaupt durch das Programmtool auslesbar sind. Dementsprechend müsste die Funktion auch eine Fehlermeldung ausgeben, wenn beispielsweise versucht wird, Bilddateien anstelle von Audiodateien zu laden. Im Interview äußerte der Nutzer außerdem die Idee, dass diese Tabelle auch dynamisch sortierbar ausgeführt werden sollte. Damit meinte er, dass die Anwender und Anwenderinnen die Anordnung der Zeilen in der Tabelle nach den Spalten sortieren können sollen. Ein Beispiel dafür wäre die Sortierung nach der aufsteigenden oder abfallenden Einlade-Reihenfolge der Dateien, oder nach der Länge der Audiosignale. Diese Funktion würde ebenfalls die Übersichtlichkeit der Programmoberfläche erhöhen und stellt daher einen Mehrwert für die Anwenderfreundlichkeit des Tools dar. Deswegen sollte diese Einstellungsmöglichkeit per Mausklick auf die entsprechende Spalte, nach der die Zeilen angeordnet werden sollen, ausführbar

sein. Des Weiteren schlug der Nutzer vor, in diese Funktion eine Bildlaufleiste – auch als Scrollbalken bezeichnet – einzufügen. Dabei sollte diese sowohl in waagerechter als auch in senkrechter Ausrichtung, also bezogen auf die Anordnung der Spalten und Zeilen, ausgeführt sein. Auch diese Funktion soll bei der Ausarbeitung des Programmtools aufgegriffen werden. Als Letztes erwähnte der Nutzer noch die Nützlichkeit einer Detailansicht einer einzelnen Audiodatei aus der Tabelle. So sollte sich bei der Selektion einer geladenen Datei ein kleines Fenster öffnen, in dem die in den Spalten aufgeführten Attribute übersichtlich dargestellt sind. Dies verschaffe den Anwendern und Anwenderinnen einen detaillierten Blick auf die einzelne Audiodatei, ohne dass diese durch die Spalten scrollen müssen. Da dies positive Auswirkungen auf die Anwendbarkeit des fertigen Programmtools nehmen könnte, wurde es als erstrebenswert für die Ausarbeitung des Tools gewertet.

Eine weitere Anforderung an das Programmtool wurde durch den Nutzer initiiert. Im Interview fragte er, ob geplant sei, die Sprachverständlichkeit nur als Mittelwert über die gesamte Länge der einzelnen Audiodateien oder blockweise auszugeben. Er führte an, dass es von Vorteil sei, wenn die Verständlichkeit in kleineren Abschnitten innerhalb der Dateien berechnet werden würde. Hintergrund davon sei, dass so in einem langen Audiosignal Schwachstellen lokalisiert werden könnten, und außerdem die blockweisen Angaben die Sprachverständlichkeit genauer angeben würden, als es ein Mittelwert täte. Anfänglich schlug der Nutzer vor, sekündlich einen Verständlichkeitswert auszugeben. Dies ist allerdings aus Sichtweise einer Sprachverständlichkeitsmessung wenig sinnvoll, da diese nur über einen längeren Zeitraum ermittelt werden kann. In diesem kurzen Intervall wäre aus zeitlichen Gründen beispielsweise nur eine Messung der Wortverständlichkeit möglich. Nachdem dieser Umstand dem Nutzer erklärt wurde, schlug er vor, dass die Anwender und Anwenderinnen des Programmtools die Länge der Blöcke selbst einstellen können sollten. Allerdings sollte dies nur in einem sinnvollen Rahmen möglich sein. Des Weiteren sollte diese Funktion erst verfügbar sein, wenn die Anwender und Anwenderinnen ein Kästchen mit dem Namen „erweiterter Modus“ aktivierten. Dadurch soll sichergestellt werden, dass die Programmoberfläche nicht zu unübersichtlich für unerfahrenere Anwender und Anwenderinnen wird. Dieser Vorschlag ist ein Kompromiss zwischen Einfachheit der Benutzeroberfläche und Funktionsumfang des Programmtools, und sollte dementsprechend bei der Ausarbeitung des Tools beachtet werden. Die Wichtigkeit dieser Funktion begründete der Nutzer anhand eines persönlichen Beispiels, bei dem es um die effiziente Ton-Bearbeitung von Podcast-Audiodateien geht (siehe Anlage 2).

## 6.3 Exkurs: Echtzeitanalysefunktion des Programms

Die in diesem Kapitel beschriebene Funktion ist im Vergleich zu den in Kapitel 6.2 (Ausarbeitung der Funktionen des Programms) aufgeführten Programmfunktionen gesondert zu betrachten, unter anderem deswegen, da sie nicht trivial in das bestehende Programmkonzept implementiert werden kann. Daher ist ihre Ausarbeitung auch in einem separaten Kapitel behandelt.

Die Funktion beschreibt eine Echtzeitanalyse der Sprachverständlichkeit eines Audiosignals. Dies bedeutet, dass innerhalb des Tools ausgewählt werden können soll, aus welcher Quelle ein Signal simultan in das Programmtool eingespeist wird. Als Beispiel sei hier der analoge Audioeingang eines Audiointerfaces erwähnt. Dieses Interface ist dabei wiederum an einen Computer, auf dem das Softwaretool installiert ist, angeschlossen. So könnte mit einem gewissen zeitlichen Versatz das Live-Signal bezüglich seiner Sprachverständlichkeit analysiert werden. Der Nutzer wurde im Interview gefragt, ob er diese Funktion für sinnvoll halte. Dies bejahte er und führte folgendes Beispiel an: Bei internationalen Live-Übertragungen kommt es vor, dass der Toningenieur bzw. die Toningenieurin mehrere Kommentar-Signale verwalten muss. Da es ihm jedoch nicht möglich ist, alle Kommentatoren bzw. Kommentatorinnen gleichzeitig abzuhören, wäre ein Tool hilfreich, das die Verständlichkeit der Sprachsignale monitort und ihm diese Werte ausgibt. Somit hätte er einen Anhaltspunkt, dass die unterschiedlichen Sendewege während einer Live-Sendung keine groben audiotechnischen Makel enthalten, ohne ständig die Abhörquelle zu verändern. Dies sind im TV-Broadcastumfeld reale Umstände. Daher wäre es vorteilhaft, wenn das Programmtool diese Funktion enthalten würde. Während des Nutzerinterviews stellte sich heraus, dass zwei Analysewerte als Ergebnis für diese Echtzeitanalyse sinnvoll wären. Zum einen eine Langzeitanalyse, bei der kontinuierlich der Wert der Sprachverständlichkeit mit den vorherigen Werten verrechnet wird, und zum anderen eine Kurzzeitanalyse der Verständlichkeit in 10-sekündigen Abschnitten, die isoliert betrachtet werden. Der Langzeitanalysewert lässt sich dabei ähnlich beschreiben wie eine integrierte Lautheitsmessung und sollte auch dementsprechend interpretiert werden. Also als Gesamtwert der Sprachverständlichkeit über die komplette Dauer des Audiosignals. Die Unterteilung in diese beiden Analysewerte ist essenziell, da der integrierte Verständlichkeitswert bei längeren Signalen von beispielsweise einer Stunde einzelne Einbrüche der Sprachverständlichkeit nicht valide wiedergeben kann. Dies hat jedoch negative Konsequenzen, da die Verständlichkeit eines Kommentar-Signals bei einer TV-Übertragung stets gewährt sein sollte. Damit verbunden wäre ein Warnsystem, welches den Anwender bzw. die Anwenderin des Programmtools benachrichtigt, wenn

der Wert der Kurzzeitanalyse unterhalb eines gewissen Schwellwerts fällt, von großem Nutzen. Da davon auszugehen ist, dass der Anwender bzw. die Anwenderin die Ausgabe der Ergebnisse des Tools stets in seinem bzw. ihrem Blickfeld hat, ist davon auszugehen, dass ein visuelles Warnsignal für die Benachrichtigung ausreichen würde. Der Nutzer wurde im Interview auch gefragt, ob die Echtzeitanalysefunktion auch nützlich sei, wenn die Ergebnisse mit einer gewissen Verzögerung dem Anwender oder der Anwenderin vorliegen würden. Dies bejahte er, da es besser sei, diese Form des Monitorings zu haben, als keine Möglichkeit der Überwachung zu besitzen.

Zwar sind die Aspekte der Nützlichkeit dieser Funktion zweifelsfrei vorhanden, allerdings ist die Umsetzung wie bereits erwähnt nicht trivial. Dies liegt unter anderem daran, dass als Analysealgorithmen eines Tools zur Echtzeitanalyse der Verständlichkeit nur nicht-intrusive Sprachverständlichkeitsmessverfahren gewählt werden können, da diese als einzige kein Referenzsignal benötigen. Des Weiteren würde dieses Tool viel Rechenleistung benötigen, da beispielsweise ein KI-gestütztes Analyseverfahren permanent Abschnitte des Audiosignals laden und im Anschluss hinsichtlich seiner Sprachverständlichkeit analysieren müsste. Generell scheint für diese Verwendung nur ein KI-gestütztes Messverfahren geeignet zu sein, da das Tool im Gesamten sehr dynamisch und agil arbeiten muss, um die Ergebnisse adäquat zu erzeugen. Somit würde diese Funktion auch großen Einfluss auf die Auswahl des geeigneten Analysealgorithmus für das Programmtool nehmen.

## 6.4 Ausarbeitung der Darstellungsform der Ergebnisse

In diesem Kapitel wird ausgearbeitet, wie die Analyseergebnisse des Programmtools dem Anwender bzw. der Anwenderin dargestellt werden sollen. Unter anderem bestätigte die Meinung des Nutzers, dass eine gute und übersichtliche Darstellungsform die in Kapitel 6.2 (Ausarbeitung der Funktionen des Programms) beschriebene Tabelle – in der die geladenen Audiodateien angezeigt werden – verkörpern würde. Die Tabelle sollte zwei Spalten enthalten, die bis zum Ausführen der eigentlichen Sprachverständlichkeitsanalyse leer sind und danach erst mit Werten beschrieben werden. Die eine Spalte sollte dabei den Verständlichkeitswert angegeben, gemittelt über die gesamte Länge der Datei. Die zweite Spalte sollte den schlechtesten Sprachverständlichkeitswert der Audiodatei angeben, bezogen auf die vorher gewählte Blockgröße im erweiterten Modus (siehe Kapitel 6.2 (Ausarbeitung der Funktionen des Programms)). Wenn diese zuvor nicht durch den Anwender bzw. die Anwenderin festgelegt wurde, sollte sie als Standard-

wert bei 10 Sekunden liegen. Zusätzlich zu dem Verständlichkeitswert des schlechtesten Blocks sollte auch noch die Stelle, an der dieser Abschnitt in dem Audiosignal gemessen wurde, ausgegeben werden. Dabei sollte es den Anwendern und Anwenderinnen auch möglich sein, die Tabelle nach einer dieser beiden Spalten sortieren zu können.

Eine genauere Form der Ergebnispräsentation soll bei der Nutzung des erweiterten Modus möglich sein. Nachdem die Dateien, wie beschrieben, blockweise hinsichtlich ihrer Sprachverständlichkeit analysiert wurden, wäre es sinnvoll, die Einzelergebnisse der Audiosignale per Selektion der entsprechenden Datei in der Tabelle aufzurufen. Nach diesem Aufruf soll die Wellenform des Signals horizontal abgebildet werden. Innerhalb dieser sollten Abschnittseinteilungen erkennbar sein, die die einzelnen Analyseblöcke repräsentieren, und innerhalb derer jeweils der momentane Verständlichkeitswert angegeben wird. Dabei würde die zusätzliche Darstellung der Wellenform des Audiosignals zur besseren Orientierung innerhalb des Signals beitragen.

Aus dem Interview ging außerdem hervor, dass der Nutzer nicht mit dem MOS-Wert vertraut war und auch seiner Einschätzung nach dieses Maß im TV-Broadcastumfeld nicht gebräuchlich sei. Diese Beurteilung ist zwar subjektiv und auch nicht validiert, allerdings führt sie trotzdem zu der Überlegung, ob die Ausgabe der Sprachverständlichkeitsanalyse nur in Form des MOS-Wertes nicht doch zu fachspezifisch ist. Daher sollten die Ergebnisse der Analysen der Verständlichkeit primär in Prozent ausgegeben werden, um auf jeden Fall aussagekräftig zu sein. Allerdings sollte eine Umrechnungsfunktion in Form einer Umschaltung der Ergebnisse auf die MOS-Werte mit in das Programmtool implementiert werden.

## 6.5 Kategorisierung der erwarteten Audiodateien

Da das fertige Programmtool nach der Berechnung der Sprachverständlichkeit der Audiodateien diese auch automatisch hinsichtlich der Analyseergebnisse kategorisieren können soll, ist die Umsetzung dieser Funktion Inhalt dieses Kapitels. Um dem Anwender bzw. der Anwenderin eine Auswahl zu bieten, wie die Kategorisierung der Ergebnisse erfolgen soll, ist es erstrebenswert, diese auf zwei Arten umzusetzen. Die erste Möglichkeit sollte die Ausgabe der in der Programmoberfläche dargestellten und fertig ausgefüllten Tabelle in Form einer CSV-Datei sein. Dafür sollte – nach Abschluss der Analyse – der Anwender bzw. die Anwenderin einen lokalen Speicherort dieser Datei auswählen können.

Die zweite Art der Kategorisierung sollte die Audiodateien in einen neuen Ordner kopieren und dabei ihren Namen ändern. Der neue Dateiname sollte dabei dem folgenden Muster entsprechen: Im ersten Teil des Namens eine vierstellige Zahl, die den Rang der Datei, bezogen auf den durchschnittlichen Sprachverständlichkeitswert des enthaltenen Audiosignals, widerspiegelt. Dabei sollte in absteigender Reihenfolge nummeriert werden, also so, dass die Datei mit dem größten Verständlichkeitswert die kleinste Zahl erhält. Aus Gründen der Übersichtlichkeit sollten führende Nullen mit angegeben werden. Der zweite Teil des Dateinamens sollte den gemittelten Verständlichkeitswert des Signals in einer Prozentangabe – mit zwei Nachkommastellen – und dem Suffix „Sprachverständlichkeit“ enthalten. Da stets eine größtmögliche Kompatibilität des Programmtools angestrebt wird und Sonderzeichen wie beispielsweise das Prozentzeichen nicht bedenkenlos bei allen Betriebssystemen benutzt werden können, sollte es im Dateinamen nicht verwendet und stattdessen ausgeschrieben werden. Analog dazu sollte anstelle des Kommas bei der Angabe des Verständlichkeitswerts ein Unterstrich angeführt werden. Die beiden Teile müssen mittels eines Bindestriches voneinander separiert werden, um die Übersichtlichkeit zu gewähren. Ein Beispiel für eine Datei, die das verständlichste Audiosignal enthält und dieses dabei eine Sprachverständlichkeit von 80,42 % besitzt, sähe demnach wie folgt aus: „0001-80_42 Prozent Sprachverständlichkeit“.

Um die Kategorisierung der Audiodateien nach Abschluss der Analyse im Programmtool zu starten, empfiehlt es sich, eine Schaltfläche mit dem Namen „Kategorisierung der eingeladenen Dateien erstellen“ in die Benutzeroberfläche einzufügen. Damit der Anwender bzw. die Anwenderin auswählen kann, welche der beiden Varianten er bzw. sie als Ausgabe erhalten möchte, ist es sinnvoll, sogenannte Auswahlkästen, auch als „Checkbox“ bezeichnet, in den Dialog in der Programmoberfläche mit einzufügen. Dabei wäre es zweckmäßig, dass diese die Namen „als CSV-Datei“ und „in neuem Ordner ausgeben“ tragen würden. Der Nutzer sagte im Interview, dass er sich vorstellen könne, dass diese beiden Kategorisierungsarten für die Anwender und Anwenderinnen des fertigen Programmtools gute und übersichtliche Varianten der Darstellung wären.

## 6.6 Auswahl des Analysealgorithmus des Programms

Wie bereits in Kapitel 5.15 (Endergebnis des Versuchs) beschrieben, sagte die Messung mit der Implementierung des Short-Time Objective Intelligibility-Verfahrens die Verständlichkeit des Testsignals am genauesten vorher. In folgendem Kapitel soll untersucht wer-

den, ob sich deswegen dieser Algorithmus auch am besten als Sprachverständlichkeitsmessverfahren des fertigen Programmtools eignet.

### 6.6.1 Vorteile der Verwendung des STOI-Algorithmus im Programmtool

Ein Tool zur Messung der Sprachverständlichkeit von Sprachsignal-Audiodateien sollte valide und möglichst präzise Analyseergebnisse liefern. Darin läge der größte Vorteil der Implementierung des STOI-Verfahrens in das Programmtool. Denn wie der in Kapitel 5 (Versuch zur Auswahl des geeigneten Analysealgorithmus) ausgeführte Versuch zeigt, ist dieses Verfahren – aus den drei ausgewählten – mit Abstand das präziseste bezüglich der Bestimmung der Sprachverständlichkeit des realen Audiosignals aus dem Broadcastumfeld. Daher kann davon ausgegangen werden, dass das Tool mit dieser Implementierung überwiegend zu sehr validen Messergebnissen führen würde.

Des Weiteren benötigt die Berechnung des Verständlichkeitswertes mittels des STOI-Verfahrens wenig Rechenleistung. Dies spiegelt sich unter anderem auch in der kürzeren Bearbeitungszeit im Vergleich zu der des NISQA-Verfahrens wider (siehe Kapitel 5.9 (Messergebnis des STOI-Algorithmus)). Zum Vergleich: bei NISQA betrug diese 13,81 Sekunden, während das STOI-Verfahren nach 3,75 Sekunden ein Ergebnis ausgab.

### 6.6.2 Nachteile der Verwendung des STOI-Algorithmus im Programmtool

Allerdings würde die Verwendung des STOI-Verfahrens als Analysealgorithmus im fertigen Programmtool auch zu Nachteilen der Anwendung führen. Wie bereits in Kapitel 4.3.6 (Short-Time Objective Intelligibility) beschrieben, unterschätzt der STOI die Beeinflussung durch Pausen im Sprachsignal auf die Sprachverständlichkeit. Dies ist insofern kritisch, weil diese Aussetzer häufig bei Fehlern in digitaler Audioübertragung auftreten können. Da beispielsweise mittlerweile der Hauptteil der Signalverteilung innerhalb eines Übertragungswagens digital erfolgt, betrifft dieser Aspekt das TV-Broadcastumfeld unmittelbar.

Des Weiteren wird der Wert der Verständlichkeit wohl überschätzt, wenn das Nutzsignal mit einem Störgeräusch in Form von anderen Sprechern bzw. Sprecherinnen überlagert wird.[265] Dies ist insofern im Broadcastumfeld ebenfalls problematisch, da es beispiels-

[265] Vgl. Tang und Cooke (2011, S. 347).

weise durch Fehler im Signalrouting bei einer Live-Sendung vorkommen kann, dass der Moderator bzw. die Moderatorin, obwohl er bzw. sie eigentlich nicht mehr auf Sendung sein sollte, trotzdem noch zu hören ist. Die Sprachverständlichkeit in dieser Situation könnte jedoch von STOI falsch eingeschätzt werden und dadurch das Tool dem Anwender bzw. der Anwenderin falsche Analyseergebnisse ausgeben.

Allerdings ist der größte Nachteil des STOI-Algorithmus, dass es sich bei ihm um ein intrusives Bewertungsverfahren handelt. Wie bereits erwähnt, benötigt er also immer zur Analyse der Sprachverständlichkeit ein Referenzsignal. Dies wäre nicht nur bei der möglichen Umsetzung der in Kapitel 6.3 (Exkurs: Echtzeitanalysefunktion des Programms) erwähnten Echtzeitanalysefunktion ein unüberwindbares Hindernis, sondern würde den Anwender bzw. die Anwenderin bei der Nutzung des fertigen Programmtools auch in erheblichem Maße einschränken. Beispielsweise sei hier der Anwendungsfall der Analyse und Überprüfung von Podcast-Audiodateien – den der Nutzer als persönliches Beispiel erwähnte – angeführt. Im Vergleich dazu könnte ein Programmtool, welches mit der NISQA-Implementierung als Analysealgorithmus agiert, diese Anwendungsfälle abdecken.

### 6.6.3 Festlegung des Analysealgorithmus des Programmtools

Die Ausarbeitung der Funktionen des Programmtools in Kapitel 6.2 (Ausarbeitung der Funktionen des Programms) zeigten, dass ein Tool, welches eine hohe Bedienungsfreundlichkeit bietet und für viele Anwendungsfälle nutzbar sein soll, nur mit einem nichtintrusiven Analysealgorithmus wie dem NISQA-Verfahren umgesetzt werden kann. Demgegenüber steht die Tatsache, dass der in Kapitel 5 (Versuch zur Auswahl des geeigneten Analysealgorithmus) ausgeführte Versuch zeigte, dass für ein reales Beispiel aus dem TV-Broadcastumfeld das Short-Time Objective Intelligibility-Verfahren mit Abstand die genauesten Ergebnisse der Vorhersage der Sprachverständlichkeit lieferte. Demnach ist auch davon auszugehen, dass von den drei verglichenen Analysealgorithmen das STOI-Verfahren – für diese Art von Audiosignal – im Durchschnitt die präziseste Verständlichkeitsbewertung abgibt. Diese Umstände führen zu folgendem Dualismus: Bei der Entwicklung des finalen Programmtools muss im Hinblick auf den derzeitigen Erkenntnisstand zwangsläufig ein Kompromiss eingegangen werden. Mit den in Kapitel 5.1 (Auswahl der drei Algorithmen für den Versuch) zur Verfügung stehenden Mitteln sind so prinzipiell zwei unterschiedliche Programmtools umsetzbar. So legt eines den Fokus auf die genaueste Sprachverständlichkeitsbewertung, ist dabei aber für den Anwender bzw. die Anwenderin nur in speziellen Anwendungsfällen verwendbar. Demge-

genüber weist ein anderes Tool eine weniger präzise Verständlichkeitsanalyse auf, ist jedoch für nahezu alle Anwendungsfälle geeignet. Bei der finalen Ausarbeitung des Programmtools muss also entsprechend dem Entwicklungsziel des Tools eine Entscheidung zwischen diesen beiden Varianten gefällt werden. Die erste Option beschreibt dabei ein Programmtool, welches den STOI-Algorithmus als Implementierung verwendet, und die zweite eines, welche das NISQA-Verfahren als Analysealgorithmus beinhaltet.

## 6.7 Skizzierungen der Programmoberfläche

Dieses Kapitel beinhaltet drei Skizzen des „Graphical User Interface“ (GUI) des Tools. Sie sollen zum einen die in den Kapiteln 6.2 (Ausarbeitung der Funktionen des Programms) und 6.4 (Ausarbeitung der Darstellungsform der Ergebnisse) vorgestellten Funktionen des Programms visualisieren und zum anderen eine Übersicht über die generelle Nutzeroberfläche des geplanten Tools ermöglichen. Des Weiteren können sie bei der finalen Ausarbeitung der Programmoberfläche als ein möglicher Design-Vorschlag verstanden werden. Die drei Skizzen wurden mithilfe der Software „Balsamiq“[266] angefertigt. Zuletzt ist noch wichtig zu erwähnen, dass die nachfolgenden GUI-Skizzen sich nur auf ein Programmtool, welches das NISQA-Verfahren als Sprachverständlichkeitsmessverfahren benutzt, beziehen. Die GUI eines Tools, welches STOI als Analysealgorithmus verwendet, könnte jedoch sehr ähnlich gestaltet sein. Allerdings müsste aus Gründen der Übersichtlichkeit eine separate Einlade-Möglichkeit für die Referenzsignale umgesetzt werden.

Abbildung 12 zeigt die Grundform der Benutzeroberfläche des Programmtools nach dem Einladen der Audiodateien. Wie zu erkennen, sind die Verständlichkeitswerte in der Tabelle noch nicht ausgefüllt, da die Analyse noch nicht gestartet wurde. Per Auswahl von einer der eingeladenen Audiodateien erscheinen die bereits tabellarisch dargestellten Informationen noch einmal kompakt in einem kleinen Fenster in der Nähe des Mauszeigers.

[266] Balsamiq.

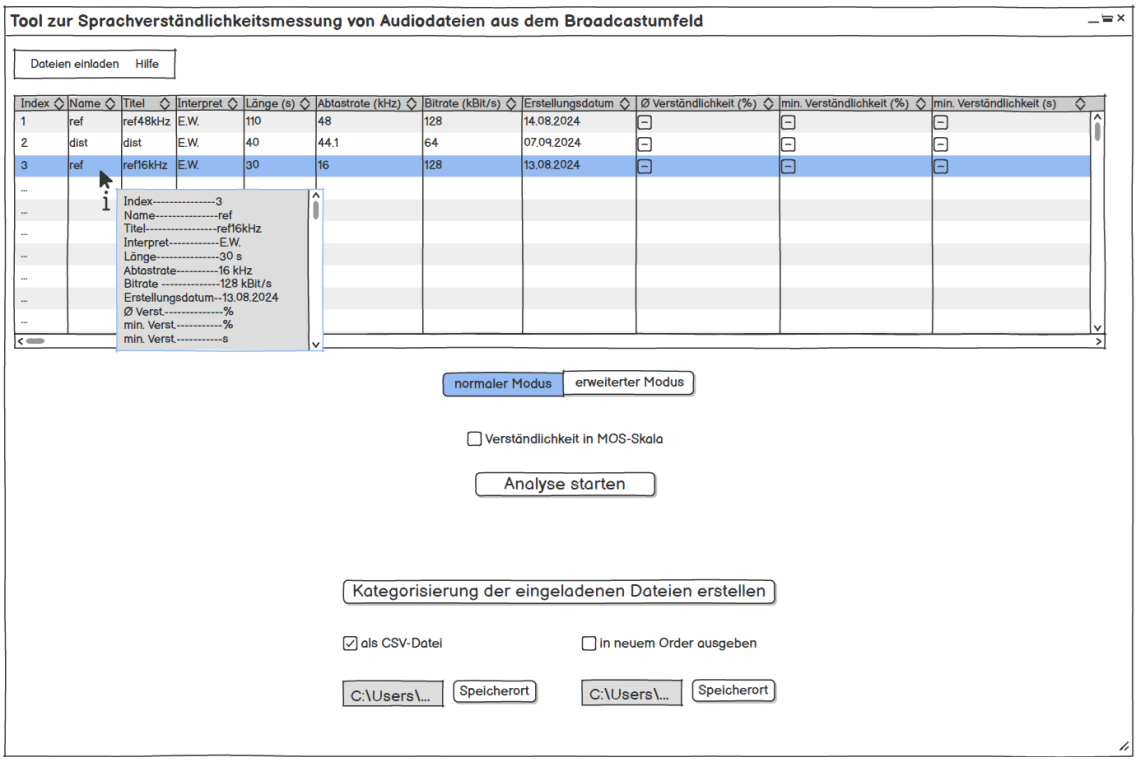


*Abbildung 12: Skizze der GUI im normalen Modus vor der Berechnung, Quelle: der Verfasser.*

Um in den erweiterten Modus zu gelangen, ist die entsprechende Schaltfläche umzuschalten. Daraufhin soll sich die Programmoberfläche – wie in Abbildung 13 dargestellt – anpassen. Die Einstellung der Abschnittsgröße der Verständlichkeitsanalyse ist nun möglich. Allerdings ist zunächst standardmäßig die Schaltfläche „Analyse starten“ ausgegraut. Dies liegt daran, dass – wie im Unterfenster des erweiterten Modus beschrieben – zunächst eine Datei für die Darstellung selektiert werden muss. Hat der Anwender bzw. die Anwenderin dies getan, kann er bzw. sie die Verständlichkeitsanalyse starten.

Die nach Ausführung der Analyse im erweiterten Modus entstehende Ergebnispräsentation ist in Abbildung 14 dargestellt. Dabei wurde entsprechend den Antworten des Nutzerinterviews die Darstellung so umgesetzt, dass die gewählte Blockgröße – innerhalb der das Audiosignal analysiert wurde – grafisch zu erkennen ist. Außerdem sind die jeweiligen Werte der Verständlichkeit sowie ein entsprechend großer Balken abgebildet. Des Weiteren zeigt diese Skizze auch die Präsentation der Ergebnisse im normalen Modus, da für diesen lediglich die letzten drei Spalten der Tabelle ausgefüllt werden.

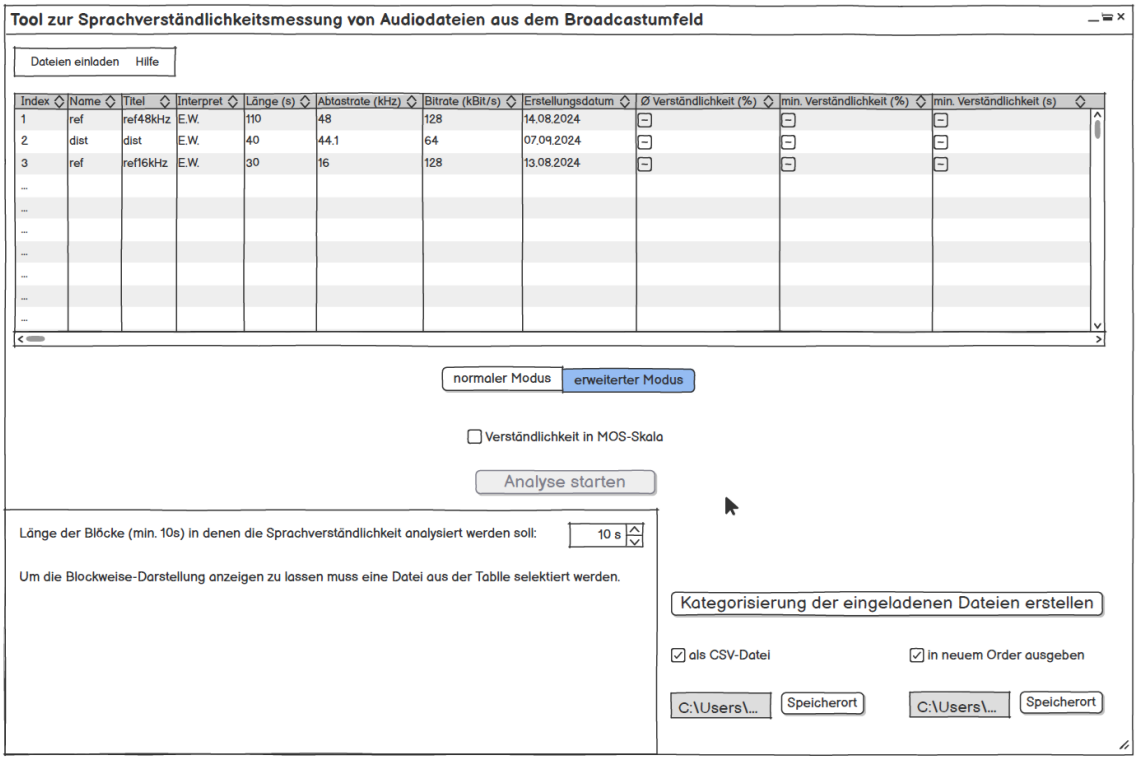


*Abbildung 13: Skizze der GUI im erweiterten Modus vor der Berechnung, Quelle: der Verfasser.*

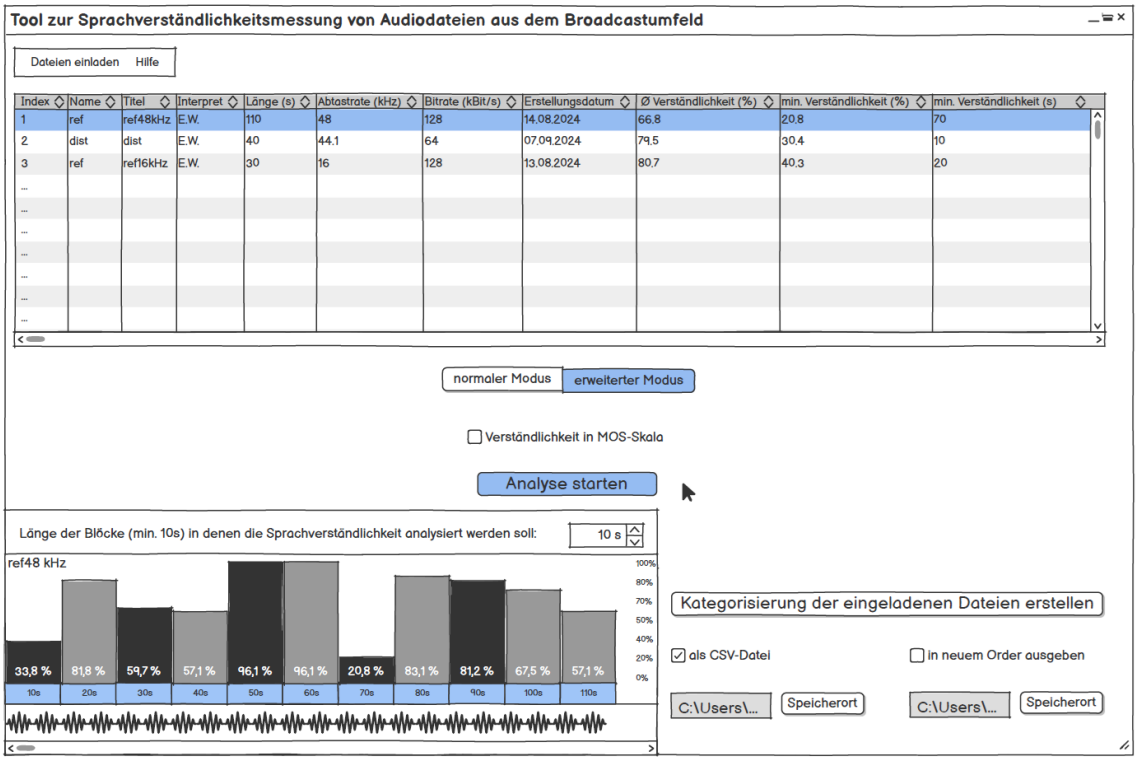


*Abbildung 14: Skizze der GUI im erweiterten Modus nach der Berechnung, Quelle: der Verfasser.*

## 6.8 Blockschaltbilder der beiden Programmvarianten

Um einen technischen Überblick über die beiden vorgeschlagenen Varianten des Programmtools zu erstellen, sind in diesem Kapitel zwei einfache Blockschaltbilder dargestellt. Dabei beschreibt das erste ein Programm, bei dem der STOI-Algorithmus als Analyseverfahren implementiert ist, und das zweite ein Tool, welches mittels der NISQA-Implementierung die Berechnung der Sprachverständlichkeit ausführt. Die Nummerierungen neben den Pfeilen in den beiden Blockschaltbildern symbolisieren jeweils die Reihenfolge, in der diese im Programmablauf ausgeführt werden.

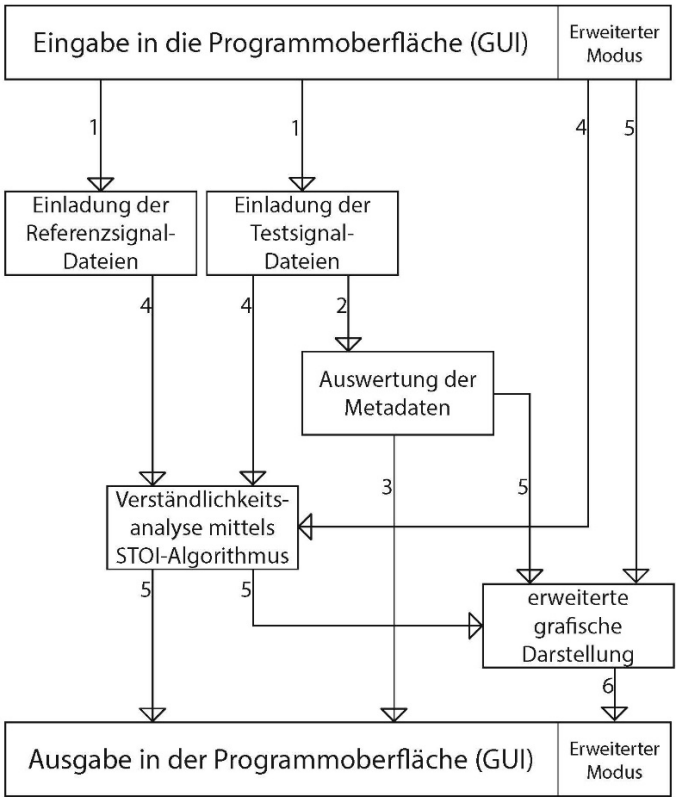


*Abbildung 15: Blockschaltbild des Programmtools mit Implementierung des STOI-Algorithmus als Sprachverständlichkeitsmessverfahren, Quelle: der Verfasser.*

Wie in Abbildung 15 zu sehen, werden zunächst die Referenzsignal-Dateien und Testsignal-Dateien über die Eingabe durch den Anwender bzw. die Anwenderin in der Programmoberfläche geladen. Im Anschluss werden die Metadaten der Dateien mit den Testsignalen ausgelesen, um die Tabelle in der Nutzeroberfläche entsprechend auszufüllen. Nachdem noch eventuelle Einstellungen – bezogen auf die Blockgröße der detaillierten Verständlichkeitsanalyse – im erweiterten Modus über die Programmoberfläche eingegeben wurden, wird die Analyse ausgeführt. Die Ergebnisse werden dann zum einen direkt in der Tabelle ausgegeben und zum anderen für die grafische block-

weise Darstellung der einzelnen Audiodateien im erweiterten Modus aufbereitet. Als letzter Schritt wird auch diese Aufbereitung in der Programmoberfläche ausgegeben.

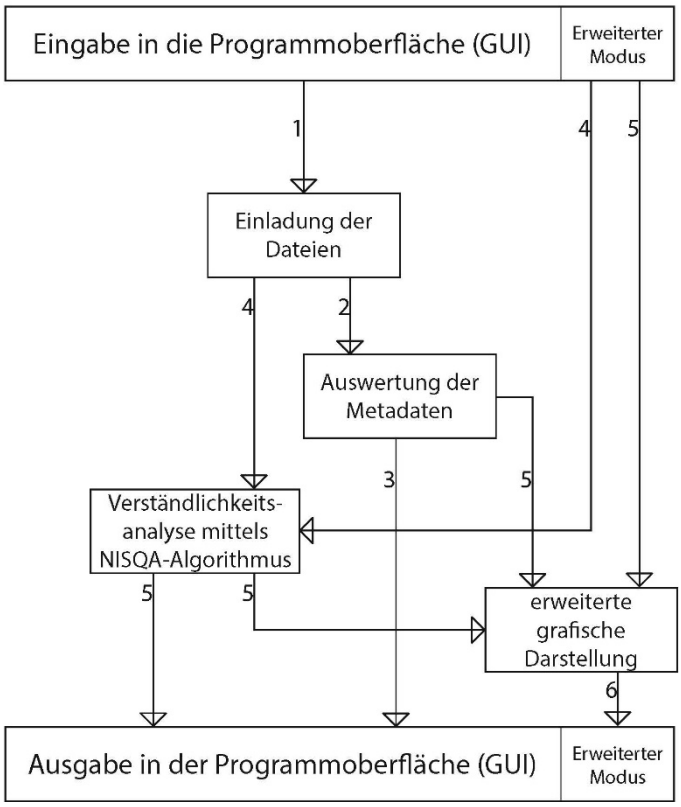


*Abbildung 16: Blockschaltbild des Programmtools mit Implementierung des NISQA-Algorithmus als Sprachverständlichkeitsmessverfahren, Quelle: der Verfasser.*

Wie in Abbildung 16 zu erkennen, ist der Aufbau der zweiten Variante des Programmtools ähnlich dem ersten Blockschaltbild. Allerdings fällt das Laden der Referenzsignal-Audiodateien weg, da nur die Dateien mit den Testsignalen für die Analyse benötigt werden. Deswegen ist in diesem Blockschaltbild dies nicht mit aufgeführt, und die Testsignal-Dateien sind allgemein als Dateien bezeichnet. Des Weiteren ist der Name des Blocks, der das eigentliche Verständlichkeitsmessverfahren beschreibt, auf NISQA angepasst.

# 7 Beantwortung der Hypothesen

Dieses Kapitel dient der Beantwortung der in Kapitel 2 (Darstellung der Hypothesen) aufgestellten Hypothesen. Die Beantwortung resultiert aus den Forschungsergebnissen dieser Arbeit.

## 7.1 STOI als der präziseste Analysealgorithmus für das Tool

Die erste aufgestellte Hypothese beinhaltete, dass das Short-Time Objective Intelligibility-Verfahren das genaueste Sprachverständlichkeitsmessverfahren für das Programmtool sei. Diese Hypothese kann nach den genauen Untersuchungen in Kapitel 5 (Versuch zur Auswahl des geeigneten Analysealgorithmus) verifiziert werden. Denn wie in Kapitel 5.14 (Vergleich der Ergebnisse) gezeigt, gab die STOI-Implementierung mit Abstand das genaueste Analyseergebnis aus, also jenes mit der höchsten Korrelation zu dem des Hörtests. Da in dem Versuch ein Audiosignal verwendet wurde, welches direkten Bezug zum TV-Broadcastumfeld hat, ist davon auszugehen, dass STOI für ein Programmtool, welches solche Signale bewerten soll, am genauesten agiert.

## 7.2 Balkendiagramm als Ergebnisdarstellung des Tools

Die zweite Hypothese bezog sich auf die geeignetste Darstellungsform der Sprachverständlichkeitswerte für das fertige Programmtool. Sie besagte, dass ein Balkendiagramm die übersichtlichste Form der Ergebnispräsentation sei. Im Hinblick auf die Meinung des Nutzers im Interview sowie die letztendlich in Kapitel 6.4 (Ausarbeitung der Darstellungsform der Ergebnisse) ausgearbeitete Darstellung in Form einer Tabelle lässt sich diese Hypothese falsifizieren. Dies liegt hauptsächlich daran, dass sich während der Ausarbeitung die Notwendigkeit von mehr als einem Verständlichkeitswert pro eingeladenen Audiosignal herausstellte. Des Weiteren muss zur besseren Identifizierung der einzelnen Dateien mehr als nur der Dateiname in der Ergebnispräsentation angegeben werden. Deswegen schien ein übersichtliches Balkendiagramm für die Darstellung nicht realisierbar.

# 8 Zusammenfassung und Ausblick

Die Zielstellung der Arbeit, die Funktionen eines Programmtools zur Sprachverständlichkeitsanalyse von Sprachsignal-Audiodateien aus dem Broadcastumfeld zu beschreiben, ist beantwortet worden. So ist ausgearbeitet worden, wie die Programmoberfläche aussehen sollte, um möglichst anwenderfreundlich zu sein und gleichzeitig alle relevanten Informationen abzubilden sowie die Ergebnisse anschaulich darzustellen. In diesem Kontext sind Skizzen der GUI ausgearbeitet worden, welche eine konkrete Darstellung der Nutzeroberfläche vorschlagen. Außerdem wurden vereinfachte Blockschaltbilder erstellt, die die Funktionsweise des Programmtools übersichtlich beschreiben.

Diese Ausarbeitung wurde dabei gestützt durch die Antworten eines Toningenieurs – einem potenziellen Nutzer des fertigen Programmtools – im Rahmen eines Nutzerinterviews. Allerdings beruhen die Ausführungen der Programmoberfläche und der Funktionen des Tools nicht ausschließlich auf den Antworten des Nutzers. Vielmehr wurden seine Einschätzungen als Anreize verstanden und stets kritisch begutachtet. Dies sollte gewährleisten, dass die GUI und die Programmfunktionen nicht nur den persönlichen Vorlieben einer Person, sondern im Endeffekt einer größeren Gruppe von Anwendern und Anwenderinnen nutzerfreundlich erscheinen. Dieser Effekt wurde ebenfalls dadurch begünstigt, dass der Nutzer – wie in Anlage 2 zu erkennen – des Öfteren versuchte, Einschätzungen, welche über seine Meinung hinausgehen, bspw. bezüglich einer ergonomischen Nutzeroberfläche abzugeben.

Weiterhin konnten anhand des ausgeführten Versuchs zwei Favoriten aus den in Kapitel 3.3 (Darstellung verschiedener Sprachverständlichkeitsmessverfahren) vorgestellten Verfahren für die Auswahl des Analysealgorithmus des fertigen Programmtools bestimmt werden. Es stellte sich heraus, dass die Wahl des zu implementierenden Algorithmus maßgeblich davon abhängt, welche Anforderungen oder Ziele im finalen Programm im Vordergrund stehen. Der STOI-Algorithmus ist geeignet, wenn Genauigkeit in der Verständlichkeitsbewertung Priorität hat und Anwenderfreundlichkeit eine geringere Rolle spielt. NISQA hingegen ist für vielfältigere Anwendungsfälle vorteilhaft, besitzt jedoch eine weniger präzise Sprachverständlichkeitsbewertung. Daraus resultiert die Erkenntnis, dass sich die Zielstellung der Arbeit nicht eindimensional beantworten lässt. Vielmehr muss im Vorhinein differenziert werden, wo der Schwerpunkt des Anwendungsszenarios des Programmtools liegt.

Als Kritik an diesem Teil der Forschungsmethode ist anzuführen, dass die Bewertung der Genauigkeit der drei getesteten Analysealgorithmen nur auf dem in Kapitel 5 (Versuch zur Auswahl des geeigneten Analysealgorithmus) beschriebenen Versuch basiert. Es ist jedoch wichtig zu betonen, dass das Hörbeispiel im Versuch ein reales Audiosignal aus dem Broadcastumfeld darstellt. Daher eignet es sich gut zur Evaluierung der Funktionalität des Programmtools.

Dass die in Kapitel 5.11 (Messergebnis des NISQA-Algorithmus) dargelegten Ergebnisse des NISQA-Verfahrens eine geringe Korrelation mit den Ergebnissen des Hörtests zeigten, ist auf den ersten Blick erstaunlich. Als mögliche Erklärung lässt sich jedoch anführen, dass die Implementierung aus dem Jahr 2021 stammt. Damit ist sie zwar im Vergleich zu den restlichen Sprachverständlichkeitsmessverfahren relativ aktuell, allerdings sind seitdem wiederum große Fortschritte in der Entwicklung von KI gemacht worden. Diese spiegeln sich darin wider, dass die Leistungsfähigkeit solcher Systeme in hohem Maße gestiegen ist (siehe Kapitel 5.1 (Auswahl der drei Algorithmen für den Versuch)). Daher ist davon auszugehen, dass auch die Leistungsfähigkeit der NISQA-Implementierung in naher Zukunft noch zunimmt. Dieses Potential ist auch der Hauptgrund, warum sie trotz der geringen Genauigkeit einer der Vorschläge für den Analysealgorithmus des Programmtools ist. In künftigen Forschungen wäre es deswegen interessant, zu untersuchen, wie die Sprachverständlichkeitsbewertung von NISQA sich weiterentwickelt und ob sich die These, dass der Algorithmus noch präziser wird, verifizieren lässt.

Im Allgemeinen ist die Weiterentwicklung – gerade von nicht-intrusiven – Sprachverständlichkeitsmessverfahren, die mit einer KI-gestützten-Analyse agieren, interessant für weitere Forschungen und im Besonderen auch von Bedeutung für die Entwicklung des finalen Programmtools zur Verständlichkeitsbewertung. So ist davon auszugehen, dass diese Arten der Sprachverständlichkeitsanalyse aufgrund ihrer Fähigkeiten zu lernen, in Zukunft deutliche Vorteile gegenüber nicht-KI-gestützten Algorithmen bieten werden.

Theoretisch gäbe es noch eine weitere Art, sich die Entwicklungssprünge KI-basierter-Systeme zunutze zu machen. So könnte ein externes Tool verwendet werden, welches mittels KI aus dem Testsignal das Sprachsignal extrahiert und entsprechend so anpasst, dass es sich als Referenzsignal für die Analyse der Sprachverständlichkeit eignet. Die Entwicklung ähnlicher Tools ist bekanntlich ebenfalls weit fortgeschritten. Diese externe Anwendung müsste dann in Form einer Vorverarbeitung in das Programmtool der Verständlichkeitsanalyse implementiert werden. Dadurch könnte letzt-

endlich die STOI-Implementierung als Analysealgorithmus verwendet werden, ohne dass die Notwendigkeit bestünde, Referenzsignale einzuladen. Allerdings müsste ein entsprechendes Programmtool, welches sich gewissermaßen selbst die erforderlichen Referenzsignale kreiert, ausgiebig getestet werden. Insbesondere die Erstellung der Referenzsignal-Dateien im Vorverarbeitungssystem muss valide und zuverlässig funktionieren, da ansonsten der eigentliche Analysealgorithmus erhebliche Fehler in der Verständlichkeitsbewertung erzeugt. Trotz dieser Gefahr wäre dies ein sehr interessantes Forschungsfeld für zukünftige Arbeiten in dem Fachgebiet der Sprachverständlichkeitsanalyse.

Des Weiteren wäre es interessant, wie das Perceptual Objective Listening Quality Analysis-Verfahren im Vergleich zu den Verfahren STOI und NISQA abschneiden würde. Da es sich bei POLQA bekanntlich um eine Weiterentwicklung des PESQ-Verfahrens handelt, ist davon auszugehen, dass es präzisere Verständlichkeitsbewertungen ausgeben würde. Wie bereits erwähnt, ist jedoch POLQA im Gegensatz zu PESQ nicht frei verfügbar, weswegen es kein Untersuchungsgegenstand des Versuchs dieser Arbeit sein konnte. Bei entsprechenden Forschungen müssten entweder Nutzungslizenzen von POLQA erworben werden, oder die Untersuchungen müssten in Kooperation mit Unternehmen bzw. Forschungseinrichtungen, welche bereits eine Lizenz besitzen, durchgeführt werden.

Als letzter interessanter Ansatzpunkt für weitere Untersuchungen auf dem Gebiet der Sprachverständlichkeitsforschung ist das Themengebiet „3D-Audio“ zu erwähnen. Der Hörtest dieser Arbeit wurde bekanntlich nur im Stereo-Kontext umgesetzt. Forschungen nach dem Einfluss von beispielsweise binauraler Audiowiedergabe auf die Sprachverständlichkeitseinschätzung von Testpersonen und im Besonderen, wie diese Effekte bei der Berechnung der Verständlichkeit durch Analysealgorithmen beachtet werden könnten, wären sehr erstrebenswert.

Zusammenfassend lässt sich sagen, dass eine umfassendere Forschung durch zusätzliche Ressourcen möglich wäre. Diese wären einerseits die vielversprechenden Forschungsergebnisse der zuvor erwähnten Ausblicke hinsichtlich zukünftiger KI-gestützter Analysealgorithmen. Andererseits wären erweiterte Forschungsmittel, Zugang zu kostenpflichtigen Nutzungslizenzen, mehr Zeit sowie eine größere Anzahl an Versuchen mit mehreren Probanden und Probandinnen vorteilhaft. Diese Aspekte würden im Endeffekt eine optimierte Ausarbeitung des finalen Programmtools implizieren.

# Literaturverzeichnis

Avila, A. R., Gamper, H., Reddy, C., Cutler, R., Tashev, I. & Gehrke, J. (2019, 16. März). *Non-intrusive speech quality assessment using neural networks*.

Balsamiq. Balsamiq Company Info [Computer software]. Verfügbar unter: https://balsamiq.com/company/

Beerends, J., Schmidmer, C. [Chris], Berger, J., Obermann, M., Ullmann, R., Pomy, J. et al. (2013). Perceptual Objective Listening Quality Assessment (POLQA). The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part II-Perceptual Model. *AES: Journal of the Audio Engineering Society*, *61*, 385–402.

Beyerdynamic. (2024, 13. Augusta). *beyerdynamic DT 797 PV: geschlossenes Headset mit Kondensatormikrofon*. Verfügbar unter: https://www.beyerdynamic.de/dt-797.html

Beyerdynamic. (2024, 15. Augustb). *Bedienungsanleitung Custom Studio*. Verfügbar unter: https://www.beyerdynamic.de/service/downloads?cid=footer

Bistafa, S. R. & Bradley, J. S. (2000). Revisiting algorithms for predicting the articulation loss of consonants AL(cons). *Journal of the Audio Engineering Society*, (48), 531–544.

Bolkart, J.. Künstliche Intelligenz (KI). Statista DossierPlus über die Bedeutung von KI. *Statista DossierPlus*.

Continuum Analytics. Anaconda. The Operating System for AI [Computer software]. Verfügbar unter: https://www.anaconda.com/

Cooper, E., Huang, W. -C., Tsao, Y., Wang, H. -M., Toda, T. [T.] & Yamagishi, J. (2023). The Voicemos Challenge 2023: Zero-Shot Subjective Speech Quality Prediction for Multiple Domains. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (S. 1–7).

Dickreiter, M., Dittel, V., Hoeg, W. & Wöhr, M. (Hrsg.). (2014). *Handbuch der Tonstudiotechnik*. München: Saur.

Dubey, R. K. & Kumar, A. (2013). Non-intrusive speech quality assessment using several combinations of auditory features. *International Journal of Speech Technology*, *16*(1), 89–101.

DWDL.de. (2022). *Durchschnittsalter der Zuschauer der einzelnen Fernsehsender in Deutschland im ersten Halbjahr 2022*. Verfügbar unter: https://de.statista.com/statistik/daten/studie/183279/umfrage/durchschnittsalter-der-fernsehzuschauer-nach-sender/

Fischer, J. & Feneberg, Gregor, Kump, Gerhard. (2011). Vergleich des Messverfahrens PEAQ mit Hörversuchen in Produkttests, 597–598.

Friesecke, A. (2014). *Die Audio-Enzyklopädie. Ein Nachschlagewerk für Tontechniker*. Berlin: De Gruyter Saur.

Hornsby, B. W. Y. (2004). The Speech Intelligibility Index: What is it and what's it good for? *The Hearing Journal*, (57), 10–17.

Houtgast, T. & van Wijngaarden, S. J. (2002). *Past, present and future of the speech transmission index*. Soesterberg: TNO Human Factors.

Hu, C.-H., Yasuda, Y. & Toda, T. [Tomoki].. Preference-based training framework for automatic speech quality assessment using deep neural network, 546–550.

TPRF HDTV 2016 (11.2016). *Technische Produktionsrichtlinien*.

Institute of Electrical and Electronics Engineers (Ed.). (2001). *Proceedings / 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. 7 - 11 May, 2001, Salt Palace Convention Center, Salt Lake City, Utah, USA*. Piscataway, NJ: IEEE Operations Center.

ITU-T, P.82 (1993). *Telephone Transmission Quality Subjective Opinion Tests*.

ITU-T, P.800 (1996). *Telephone Transmission Quality Subjective Opinion Tests*.

ITU-T, P.800.1 (2016). *Terminals and Subjective and Objective Assessment Methods*.

Kates, J. M. & Arehart, K. H. (2005). Coherence and the speech intelligibility index. *The Journal of the Acoustical Society of America*, *117*(4), 2224–2237.

Klein, W. (1972). Articulation loss of consonants as a basis for the design and judgment of sound reinforcement systems. *Journal of the Audio Engineering Society*, *March 1972*, 920–922.

Kolbæk, M., Tan, Z. -H. & Jensen, J. [J.]. (2019). On the Relationship Between Short-Time Objective Intelligibility and Short-Time Spectral-Amplitude Mean-Square Error for Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(2), 283–295.

Kondo, K. (2012). Estimation of speech intelligibility using objective measures. *Applied Acoustics*, *74*(1), 63–70.

Kryter, K. D. (1962). Validation of the Articulation Index. *The Journal of the Acoustical Society of America*, *34*(11), 1698–1702.

Lazarus, H., Sust, C. A., Steckel, Rita, Kulka, Marko & Kurtz Patrick. (2007). *Akustische Grundlagen sprachlicher Kommunikation*. Berlin, Heidelberg: Springer.

Mittag, G. NISQA. Non-Intrusive Speech Quality and TTS Naturalness Assessment [Computer software]. Verfügbar unter: https://github.com/gabrielmittag/NISQA

Mittag, G., Naderi, B., Chehadi, A. & Möller, S. (2021). NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets, 2127–2131.

Nocker, R. (2005). *Digitale Kommunikationssysteme 2. Grundlagen der Vermittlungstechnik*. Wiesbaden: Vieweg.

Pariente, M. Pystoi. Python implementation of the Short Term Objective Intelligibility measure [Computer software]. Verfügbar unter: https://github.com/mpariente/pystoi

Patton, B., Agiomyrgiannakis, Y., Terry, M., Wilson, K. & Saurous, R. (2016). AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech, 1–5.

Peutz, V. (1972). Articulation loss of consonants as a criterion for speech transmission in rooms. *Journal of the Audio Engineering Society*, *M21*, 915–919.

Probst, W. & Böhm, M.. Sprachverständlichkeit. Die Anwendung des STI zur Beurteilung von Sprachgeräuschen. *Lärmbekämpfung*, *2017*, S.57-65.

RIEDEL. (2024, 13. August). *RIEDEL » Kommentator*. Verfügbar unter: https://www.riedel.net/de/produkte-loesungen/intercom/sprechstellen/kommentator

Rix, A. W., Beerends, J. G. [J. G.], Hollier, M. P. & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of

telephone networks and codecs. In Institute of Electrical and Electronics Engineers (ed.), *Proceedings / 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. 7 - 11 May, 2001, Salt Palace Convention Center, Salt Lake City, Utah, USA* (S. 749–752). Piscataway, NJ: IEEE Operations Center.

Schoeps GmbH. (2024). SCHOEPS ORTF-3D Windshield Setup Guide, 1–14.

Shen, K., Yan, D., Hu, J. & Ye, Z. (2024). Non-intrusive speech quality assessment: A survey. *Neurocomputing*, *580*, 1–14.

Shen, K., Yan, D., Ye, Z., Xu, X., Gao, J., Dong, L. et al. (2023). Non-intrusive speech quality assessment with attention-based ResNet-BiLSTM. *Signal, image and video processing*, (17), 3377–3385.

Statistisches Bundesamt. (2024). *Durchschnittsalter der Bevölkerung in Deutschland von 2011 bis 2023 (in Jahren)*. Verfügbar unter: https://de.statista.com/statistik/daten/studie/1084430/umfrage/durchschnittsalter-der-bevoelkerung-in-deutschland/

Stiles, D. J. (2019). *The SAGE Encyclopedia of Human Communication Sciences and Disorders. Speech Intelligibility Index (SII)* (1st ed.). Thousand Oaks: SAGE Publications.

Taal, C. STOI – Short-Time Objective Intelligibility Measure – [Computer software]. Verfügbar unter: https://ceestaal.nl/code/

Taal, C. H., Hendriks, R. C., Heusdens, R. & Jensen, J. [Jesper]. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4214–4217.

Taal, C. H., Hendriks, R. C., Heusdens, R. & Jensen, J. [Jesper]. (2011). An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 2125–2136.

Taghavi, S. M. R., Mohammadkhani, G. & Jalilvand, H. (2022). Speech Intelligibility Index: A Literature Review. *Auditory and Vestibular Research*, 148–157.

Tang, Y. & Cooke, M. (2011). Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. *Proceedings of the*

*Annual Conference of the International Speech Communication Association, INTERSPEECH*, 345–348.

Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C. [Christian], Sporer, T., Beerends, J. G. et al.. PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality. *Journal of the Audio Engineering Society*, *2000*(48), 3–29.

Thonny. Thonny, Python IDE for beginners [Computer software]. Verfügbar unter: https://thonny.org/

Torcoli, M., Kastner, T. & Herre, J. (2021). Objective Measures of Perceptual Audio Quality Reviewed: An Evaluation of Their Application Domain Dependence. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 1530–1541.

Tseng, W.-C., Kao, W.-T. & Lee, H.. DDOS: A MOS Prediction Framework utilizing Domain Adaptive Pre-training and Distribution of Opinion Scores.

Wang, M., Boeddeker, C., Dantas, R. G. & seelan, a. (2022). ludlows/python-pesq. supporting for multiprocessing features [Computer software]: Zenodo.

Weinzierl, S. (Hrsg.). (2008). *Handbuch der Audiotechnik* (VDI-Buch). Berlin: Springer.

You, J., Reiter, U., Hannuksela, M. M., Gabbouj, M. & Perkis, A. (2010). Perceptual-based quality assessment for audio–visual services: A survey. *Signal Processing: Image Communication*, *25*(7), 482–501.

Zhou, W., Yang, Z., Chu, C., Li, S., Dabre, R., Zhao, Y. et al. (2024). MOS-FAD. Improving Fake Audio Detection Via Automatic Mean Opinion Score Prediction. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (S. 876–880).

# Anlagenverzeichnis

# Anlage 1: Fragebogen des Hörtests

Versuch zur Auswahl
des geeigneten
Analyse-Algorithmus
für das Programmtool

Angaben zur Kategorisierung der Ergebnisse. Bitte zutreffendes ankreuzen, oder eintragen.

Alter: ____ (in ganzen Jahren)

Geschlecht: ☐ weiblich ☐ männlich ☐ _____

Tragen Sie ein Hörgerät: ☐ ja ☐ nein

---

Bitte kreuzen Sie in Bezug auf die akustische Sprachverständlichkeit des Hörbeispiels die zutreffende Antwort an.

☐ sehr schlecht (1) ☐ schlecht (2) ☐ passabel (3) ☐ gut (4) ☐ ausgezeichnet (5)

# Anlage 2: Nutzerinterview mit Toningenieur

Moderator: Also die erste Frage wäre: Bei dem Tool (…) Also wir reden nochmal darüber, das ist ein Programmtool. Quasi du hast ein Programmfenster vor dir und kannst da über „Datei öffnen“ oder „Datei einladen“ mehrere Dateien einladen. Und im Endeffekt wird da ein Ergebnis ausgegeben, wie sprachverständlich die einzelnen Dateien sind. Was, glaubst du, wäre gut, um jetzt die Ergebnisse darzustellen? Also wäre es eher gut, wenn das eine grafische Darstellung in Form von zum Beispiel einem Balkendiagramm wäre? Sagen wir, du lädst fünf Dateien ein, wäre es da gut, wenn du dann das Ergebnis als ein Balkendiagramm vorliegen hast mit: Datei eins hat einen Wert von 50 % Sprachverständlichkeit. Zweite Datei, nächster Balken (…) Oder wäre es gut, die Ergebnisse nur mit den

genauen Zahlen zu hinterlegen? Oder wäre es sogar gut, dass auf beide Weisen zu machen? Das man die genauen Zahlenwerte hat und eine grafische Darstellung davon?

Nutzer: Also für mich wäre es am besten, wenn es eine tabellarische Ansicht wäre, die ich sortieren kann nach den Werten. Damit ich, wenn ich auf der Suche nach dem besten Wert bin (…) Also ich habe Dinge einfach gerne sortiert nach Qualität.

Moderator: Okay, also weniger grafisch?

Nutzer: Das zusätzlich gerne auch. Aber mehr Wert würde ich persönlich legen auf die Tabelle. Also quasi zwei Spalten und die Anzahl der Zeilen, entsprechend der Anzahl der Dateien.

Moderator: Okay und dann so, dass du auswählen kannst: Ich lasse nach Einlade-Reihenfolge sortieren, oder nach Sprachverständlichkeit. Das heißt, wir stellen uns eine Tabelle vor, in der die einzelnen Ergebnisse angezeigt werden. Wie würdest du dir das genau vorstellen?

Nutzer: Na ja, es gibt Zeilen, jede Zeile ist eine Datei. Und als Spalten nimmst du bspw. Titel und Interpret vielleicht zusammen, vielleicht auch getrennt, wann es eingeladen wurde, Dateilänge (…).

Moderator: Okay, also Metadaten möchtest du auch gerne mit in der Tabelle aufgeführt sehen?

Nutzer: Ja, denn es geht ja darum, dass ich eine große Anzahl an Dateien habe und dann beschließen muss, welche ich nehmen möchte. Wenn ich die Dinge sehe, bspw. wie lang die Datei ist, wann ich sie eingeladen habe (…). Wie soll ich sagen (…), dann habe ich bessere Übersichtlichkeit bei einer großen Anzahl an Dateien.

Moderator: Okay, perfekt. Also man müsste die Metadaten auslesen und in die Tabelle mit einfügen und dann eben die Werte der Sprachverständlichkeit in die Tabelle geben und sehr wichtig, man müsste sortieren können, zwischen Datei mit der höchsten Verständlichkeit, oder der Einlade-Reihenfolge (…). Dann nächste Frage, die Kategorisierung; fändest du es sinnvoll, wenn das Programm parallel zur Ausgabe im Programmfenster auch eine CSV-Datei generiert?

Nutzer: Na ja die Möglichkeit das zu tun, wäre sicherlich nicht verkehrt.

Moderator: Glaubst du, es hätte einen Mehrwert, wenn das Tool das hätte?

Nutzer: Für irgendwen hat es das bestimmt.

Moderator: Was natürlich auch eine Alternative wäre, wenn du einen Ordner angibst und das Tool sagt dir, ich habe hier deine Dateien, die beispielsweise inhaltlich alle das gleiche Signal enthalten, aber mit unterschiedlichen Audiokodierungsverfahren kodiert wurden und du gibst einen neuen Ordner an und das Tool schreibt die Dateien dort alle nochmal hinein, aber ändert die Dateinamen auf zum Beispiel: „0001 – 50 % Sprachverständlichkeit“.

Nutzer: Ja, ich verstehe schon. Das ist auch eine gute Idee.

Moderator: Glaubst du das hätte einen Mehrwert, oder denkst du eher an die eben erwähnte CSV-Datei?

Nutzer: Na ja, ich denke es gibt für vieles einen Anwendungszweck und genauso wird es Leute geben, die es gerne nach Dateinamen sortiert hätten und für die Leute ist es dann natürlich praktisch, wenn es mit dem Tool geht und nicht noch mit etwas anderem durchgeführt werden muss.

Moderator: Genau, man müsste dann die Daten aus der Tabelle nochmal händisch in die neuen Dateinamen einfügen. Aber glaubst du es wäre gut, wenn das Tool das automatisch machen könnte? Vielleicht auswählbar (…)

Nutzer: Na ja, Menschen, die sich dieses Tool dann im Endeffekt zulegen würden und damit arbeiten würden, denen würde das gelegen kommen.

Moderator: Okay, nächste Frage: Wie stellst du dir das Anwendungsfenster vor? Wie würde denn für dich eine gute GUI aussehen? Also, wir haben ein Fenster, wo wir oben auf „Datei einladen“ klicken können. Dann können wir da die Dateien auswählen bzw. Datei oder Dateien auswählen und diese (…) erscheinen dann irgendwo in Form von einer Liste, bevor du auf Analyse klickst, dass du siehst, dass sie auch korrekt eingeladen wurden. Dann kannst du auf Analyse klicken, kannst noch auf Ergebnisausgabe das einstellen, was wir gerade erwähnt haben. Und dann bekommst du vielleicht in einem neuen Fenster (…) Ich denke jetzt eher so ein bisschen an so Matlab orientierte Situationen, wo für einen neuen Plot dann ein neues Fenster aufgeht (…) mit den Analyseergebnissen. In dem neuen Fenster wird dir dann die erwähnte Tabelle angezeigt. Was, glaubst du denn, wäre noch brauchbar in der GUI? Oder wie stellst du dir eine übersichtliche GUI vor, so wie ich sie beschrieben habe? Oder fehlt da noch was? Oder ist da noch was zu viel?

Nutzer: Prinzipiell bräuchte ich ja kein Fenster, das sich neu öffnet für die Ergebnisse, sondern wie sehr oft in Software, wo du ein Stück leeres Fenster hast in der GUI, wo du doppelt reinklickst oder du ziehst da Dateien einfach rein. Das da praktisch alle Dateien aufgelistet werden, zur Not mit Scrollbalken, wenn es zu viele sind und die ganzen Metadaten direkt angezeigt werden, nur das die Spalte Verständlichkeitsbewertung noch nicht ausgefüllt ist und diese dann erst ausgefüllt wird, wenn du die Analyse wirklich startest (…).

Moderator: Okay, also du würdest vor der Analyse der Sprachverständlichkeit die Metadaten auslesen?

Nutzer: Ja, damit das schonmal gemacht ist. Denn es kann auch sein, dass du dadurch erkennst, dass Dateien reingeraten sind, die du eigentlich überhaupt nicht analysieren möchtest. Denn Dateiname ist ja schön und gut, aber wenn du von einem Signal unterschiedliche Versionen hast und die ähnlich oder gleich benannt sind, dann siehst du Abtastrate nicht am Dateinamen. Und ansonsten, wäre es noch gut, wenn man eine Datei markiert in dieser Liste, dass man dann in der zweiten Hälfte des GUIs, oder wo noch Platz ist, detaillierte Angaben erhält, übersichtlich.

Moderator: Für die einzelne Datei?

Nutzer: Genau. Also wenn jetzt die Tabelle und die Spalten zu lang werden, weil du viele Metadaten ausliest, was ich gerne in Software sehe, dann musst du viel scrollen. Da wäre es gut, wenn du dann eine Datei anklicken könntest und dann in einem Bereich alles in einem neuen kleinen Fenster angezeigt wird, damit du alles in einem Blick siehst. Quasi alle Informationen.

Moderator: Okay. Super!

Nutzer: Wird die Sprachverständlichkeit eigentlich nur im Gesamten beurteilt, oder ist da eine Art Timeline angedacht, wo du siehst, wo es besser ist in der Datei und wo schlechter?

Moderator: Eigentlich war bisher nur geplant, dass sie in der gesamten Datei analysiert wird und gemittelt wird. Aber du meinst es wäre praktisch, wenn man eine genauere Analyse der Datei hätte? Beispielsweise: Wie sehen die ersten 10 Sekunden der Datei aus und wie die nächsten 10 Sekunden?

Nutzer: Na ja, es kann dir ja auch passieren, dass du eine Datei hast, die eine halbe Stunde lang ist und die eigentlich eine gute Verständlichkeit hat bis auf 10 Sekunden, in

denen nichts zu verstehen ist (…) Dies würde dazu führen, dass das Tool die gesamte Sprachverständlichkeit schlechter einschätzt. Abgesehen davon, wäre es auch praktisch, um absolute „k.o.-Stellen“ herauszufinden. Weil du dann die schlechten Stellen im Audiosignal lokalisieren kannst.

Moderator: Okay also du sagst, es wäre gut, wenn das Tool es dir auch ermöglichen würde, zwar nur eine Datei, dafür aber eine genauere Analyse einer einzelnen Audiodatei zu unternehmen? Also quasi blockweise. Das heißt du bekommst Ergebnisse einer Datei in Form von: 0-5 Sekunden Sprachverständlichkeit x, 5-10 Sekunden Sprachverständlichkeit y (…), dann das ganze vielleicht in einer Tabelle ausgegeben, und vielleicht sortierbar, wo ist denn die verständlichste Stelle und wo ist die unverständlichste Stelle (…) Bleiben wir mal bei 5 Sekunden als Beispiel. Das meintest du, oder?

Nutzer: Ja, also das Bild, dass ich im Kopf habe, ist, dass die UI genauso ist, wie zuvor besprochen und in der Tabelle dann der Analysemittelwert untergebracht ist und aber auch der Wert der schlechtesten und der besten Stelle (…) mit Timecode dazu. Die Detailansicht-Seite sollte bezüglich dem grafischen Vergleich der unterschiedlichen Dateien zueinander eine Grafik anbieten, bei der du die komplette Länge der Datei angezeigt bekommst, mit eins bis fünf pro jeder Sekunde. Und du damit genau siehst, an welcher Stelle ist die Verständlichkeit wie hoch.

Moderator: Okay. Also dann pro Sekunde ein Verständlichkeitswert über der Wellenform der einzelnen Datei? Das wird schwierig (…), da die Sprachverständlichkeit ja nicht sekündlich gemessen werden kann, da sie sich ja auf zusammenhängende Worte bzw. Sätze bezieht.

Nutzer: Dann 10 Sekunden (…) oder gut wäre, wenn der Anwender diesen Bereich einstellen könnte, also in einem gewissen Rahmen.

Moderator: Okay, aber du willst die Ergebnisse nicht aufteilen, sondern die GUI prinzipiell wie beschrieben darstellen (…) nur zusätzlich noch einfügen, dass bei Selektion einer Datei die Sprachverständlichkeit dieser Datei dann blockweise angezeigt wird.

Nutzer: Ja, genau.

Moderator: Denn das Tool soll am Ende auch nutzerfreundlich und übersichtlich sein.

Nutzer: Na das schreit doch nach einem Advanced-Modus (…)

Moderator: Ja, sehr guter Punkt!

Nutzer: Ja, das bietet sich vielleicht an. Damit erweiterst du den potenziellen Anwenderbereich, da sich dadurch das Tool auch für Podcaster anbietet. Meine Frau – beispielsweise – nimmt Podcasts auf und ich muss es dann hinterher bearbeiten. Ich habe aber ganz oft keine Zeit, oder aber auch keine Lust mir da den 1,5 Stunden Podcast anzuhören. Ich höre mir natürlich die Schnitte an und bearbeite die, aber wenn zwischendurch etwas kaputt ist, dann bekomme ich das natürlich nicht mit. Dafür wäre das Tool auch gut.

Moderator: Ja, guter Punkt (…). Die nächste Frage bezieht sich auf die Skala der Sprachverständlichkeit (…). Möglichkeit 1: Ausgabe der Ergebnisse als MOS-Wert, also 1 bis 5. Dabei wäre dann 1 eine enorm schlechte Verständlichkeit und 5 eine enorm leicht zu verstehende und perfekte Sprachverständlichkeit. Möglichkeit 2 wäre eine prozentuale Darstellung.

Nutzer: Du kannst natürlich beides implementieren, da ja beides ineinander umrechenbar ist. Ich persönlich kann mehr mit einem Wert von 0 bis 100 etwas anfangen, weil mir dieser MOS-Wert unbekannt ist. Aber in einem anderen Rahmen könnte das schon interessant sein. Deswegen vielleicht einfach beides anzeigen lassen. Entweder auswählbar oder (…)

Moderator: Okay, andere Frage: Würdest du davon ausgehen, dass der MOS-Wert im Broadcastumfeld bekannt ist?

Nutzer: Na ja, also der normalsterbliche Broadcast-Toningenieur wird das wahrscheinlich nicht als Maß kennen.

Moderator: Okay also Prozente und MOS-Wert. Als Standardausgabe die Prozente und als Zusatzoption den MOS-Wert.

Nutzer: Ja genau.

Moderator: Okay, nächste Frage: Das ist ein etwas großer Punkt, der noch ein wenig in der Schwebe ist. Ich will eigentlich auch nur wissen, ob es hilfreich ist oder nicht. Wir reden noch mal vom Broadcastumfeld, also wir reden ja die ganze Zeit vom Broadcastumfeld, aber jetzt nochmal speziell bei der Frage bitte beachten. Glaubst du, es wäre praktisch, wenn dieses Tool auch Echtzeitanalyse könnte? Das heißt, du schickst einen Stream rein (…) lassen wir das mal offen, was das für ein Stream ist. In welcher Form (…)

Nutzer: Wie eine Lautheitsmessung, quasi?

Moderator: Sozusagen wie eine Lautheitsmessung (…) ja genau. Nur im Sinne einer Sprachverständlichkeitsmessung. Echtzeitanalyse meint jetzt in dem Fall (…) Also du kannst das wahrscheinlich nicht so schnell betreiben, wie eine Lautheitsanalyse (…)

Nutzer: Da kommt es also auf die Intervalle – die man vorher definiert – an, also sagen wir beispielsweise 10-Sekunden Blöcke.

Moderator: Genau richtig.

Nutzer: Und daraus einen fortlaufenden Schnitt auswählen.

Moderator: Ja genau, glaubst du, das hätte Anwendungspotential?

Nutzer: Ja (…) auf jeden Fall! Zum Beispiel dahingehend, wenn du in einer Sendung bist, in der du zum einen deinen eigenen Kommentator hast und zum anderen zusätzlich noch zwei, oder drei anderssprachige Kommentatoren. Die man ja nicht die ganze Zeit mit abhören kann.

Moderator: Ja (…) sehr guter Punkt.

Nutzer: Dass du da dann eine Automatik hast, wie eine Lautheitsmessung, die den Schnitt anzeigt. Also Short-Term, Middle-Term und den Integrated Wert, dass du so ungefähr weißt, was da so vor sich geht bezüglich der Sprachverständlichkeit.

Moderator: Okay.

Nutzer: Sowas wäre gut.

Moderator: Und du würdest es auch tatsächlich sogar in den drei Abständen Short-Term, Long-Term und Integrated machen? Oder würdest du es ungenauer machen?

Nutzer: Na ja (…).

Moderator: Die große Frage ist, brauchst du wirklich eine Short-Term Sprachverständlichkeitsmessung. Eigentlich benötigst du ja vielleicht nur Integrated und nur Long-Term, denn man sollte beachten (…) die Sprachverständlichkeit liegt immer über der Satzverständlichkeit und die Satzverständlichkeit liegt immer über der Wortverständlichkeit. Das ist ja logisch. Aufgrund des Satzkontext. Deswegen ist immer der große Punkt bei dieser Frage, ob denn eine kurze, also unter 10-sekündige, Sprachverständlichkeitsanalyse wirklich Sinn ergibt.

Nutzer: Darüber lässt sich jetzt bestimmt streiten. Es würde sich wahrscheinlich im Laufe der Nutzung des Tools ergeben. Aber gerade im Hinblick auf dieses Kommentatoren-Beispiel, wo es so viele sind, dass du nicht alle abhören kannst, wäre es natürlich auch gut, wenn kurzzeitige Einbrüche in der Verständlichkeit auch in die Bewertung mit eingehen.

Moderator: Ja, vollkommen richtig.

Nutzer: Dass man das eben mitbekommt. Da hilft nämlich eine Langzeitanalyse von einer viertel Stunde wenig, da ich das dann nicht mitbekomme. Schön wäre es, wenn ich bei Auftreten einer unverständlichen Stelle innerhalb eines 10-Sekundenblocks, benachrichtigt werden würde und dann darauf reagieren kann.

Moderator: Ja, sehr gut.

Nutzer: Also, wenn du Dateien beurteilst, ist das ja nicht zeitkritisch, aber wenn du bei einer Live-Sendung bist und das Kommentator-Signal war schon eine Viertelstunde sehr unverständlich, das ist dann schlecht.

Moderator: Okay. Großer Punkt ist natürlich das Delay, also man müsste davon ausgehen, dass – nur als Größenordnung – diese 10-Sekundenblöcke vielleicht erst nach 30 Sekunden ausgewertet sind.

Nutzer: Okay.

Moderator: Glaubst du, auch dann hätte die Echtzeitanalyse einen Nutzen?

Nutzer: Na ja. Sicherlich! Denn es ist ja natürlich besser als nichts.

Moderator: Okay. Ich habe auch nur noch eine Frage bezüglich Kritik bzw. Anregungen Also hast du noch irgendeinen Vorschlag, oder irgendwas, was dir noch wichtig erscheint, wo du sagst, das muss auf jeden Fall so in das Tool miteinfließen, ansonsten wird das nicht funktionieren, oder es ist „un-anwenderfreundlich“?

Nutzer: Nein (…), im Moment fällt mir nichts dazu ein.

Moderator: Okay, dann bedanke ich mich sehr für deine Antworten!