Handbuch für Versuchsdesign in der Psychoakustik

Planung, methodische Konzeption, Durchführung und Analyse von Hörversuchen

Diplomarbeit

Silvie Yvonne Müller

Institut für Signalverarbeitung und Sprachkommunikation

Leitung: Univ.-Prof. DI Dr. Gernot Kubin

Begutachtung: DI Dr. Martin Hagmüller

In Kooperation mit JOANNEUM RESEARCH – Institut DIGITAL

Betreuung: DI[in] Maria Fellner, MBA

Graz, November 2013

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am …… ……..

(Unterschrift)

Danksagung

Allen voran möchte ich Dir danken, Klaus. Du bist mein bester Freund und warst für mich während meines gesamten Studiums eine Riesenunterstützung, hast mich herausgefordert, kritisiert, zu mir gehalten, mich aufgebaut, an mich geglaubt und mit mir gelacht. Ich liebe Dich sehr!

Liebe Mum, lieber Dad, danke dass ihr immer für mich da ward, ihr seid die besten Eltern auf der ganzen Welt! Ihr habt mir so viele Chancen ermöglicht, so viele verschiedene Möglichkeiten, lernen zu können. Euch verdanke ich mit, wie und wer ich heute bin und sein kann, wie ich denke und die Welt sehe. Und danke für die Freundschaft, die erholsamen Wochenenden bei und mit Euch, die Köstlichkeiten und den Raum für Ruhe und Erholung.

Liebe Yvette, lieber Michel, Euch möchte ich für das einander Zuhören danken und dafür, dass wir viel diskutiert und gelacht haben. Ihr habt immer versucht, meinen Weg zu verstehen und ich auch den euren. Ich liebe meine große Schwester und meinen kleinen Bruder von ganzem Herzen und hoffe, wir werden weiter so lieb füreinander da sein!

Omi, danke für die schöne gemeinsame Zeit, die Geschichten von früher, für das Spielen und für die berühmten Rezepte. Bitte schau weiter gut auf dich! Ich hab dich sehr lieb!

Liebe Marianne, danke für die schönen, erholsamen Stunden in Salzburg, für die wunderbaren Köstlichkeiten und für die vielen intensiven Gespräche.

Und dir, lieber Rainer, möchte ich für die lustige Zeit danken, das Lachen und für die Geschichten.

An alle meine lieben Freunde – Danke für euer Verständnis, wenn ich keine Zeit hatte oder grantig war, oder frustriert oder müde… Und danke für die lustigen Stunden wenn ich munter und übermütig war!

Ich möchte mich ganz herzlich bei Maria Fellner für die liebevolle Betreuung und die Möglichkeit bedanken, mich mit einem Thema auseinander zu setzen, das mich begeistert und interessiert. Martin Hagmüller möchte ich für die gewissenhafte und konstruktive Kritik danken, für seine freundliche Art und sein Interesse an meiner Arbeit. Ganz besonders lieben Dank an Michaela Dvorzak für die spannenden Diskussionen über die statistischen Inhalte, für den mathematisch wissenschaftlichen Standpunkt zum Thema und die gewissenhaften Korrekturen. Ebenso möchte ich mich bei Harald Rainer für Denkanstöße und Korrekturen der akustisch relevanten Teile meiner Arbeit und den schönen Arbeitsplatz bedanken.

Und dann, danke an die liebe Akustikrunde, Su und Bernhard, Flo und Moritz, Franz, Sandra, Johannes und nicht zuletzt auch Martina. Die Zeit bei euch war sehr schön für mich, wegen jedem Einzelnen von Euch. Danke für die Diskussionen und Ideen, die Fragen und Antworten, danke für alles Neue. Ein bisschen durfte ich euch kennenlernen, das war sehr schön. Es war fein mit euch zu lachen und mit zu fiebern und auch das Zuhören habe ich genossen. Ich hoffe, wenn ihr an mich denkt, müsst ihr lächeln.

Ein Danke an jene, die gelehrt haben, manche von Ihnen mit weit größerem Engagement als andere. Und ebenso danke ich deren Assistentinnen und Vertretungen.

Ich möchte zudem noch den freundlichen Sekretärinnen an den Instituten danken.

Und zum Schluss noch ein Dankeschön an das freundliche Personal und alle guten Geister der TU, KUG und bei JR Digital.

Zusammenfassung

Hörversuche stellen in Forschung und Entwicklung ebenso wie in Industrie und Wirtschaft nach wie vor ein unverzichtbares Instrumentarium zur Erlangung von Bewertungen psychoakustischer Parameter durch den Menschen dar. Die Kombination der Verwendung eines gemäß ITU-R BS 1116-1 geplanten Multimediaraums gemeinsam mit einem Expert Listening Panel repräsentiert für viele Anwendungen von Hörtests optimierte Rahmenbedingungen.

Zur Implementierung standardisierter Verfahren aus dem Sprach- und Audiobereich in ein bestehendes, forschungsnahes Testinstitut, wurden Anleitungen für Planung, Durchführung und Analyse dieser Verfahren erstellt.

Um die Aussagekraft von Hörversuchen zu steigern, ist es bereits in der Planungsphase essentiell, den Fokus auf die Fragestellung zu legen und danach das optimale Versuchsdesign auszurichten. Dieses beinhaltet neben der korrekten Parameterauswahl die Skalen- und Attributfindung, sowie das Experiment Design zur Fehlerminimierung. Der Einsatz naiver Hörer und jener von Experten wird in Bezug auf Allgemeingültigkeit, Kostenaufwand und Verfügbarkeit kritisch hinterfragt und die Vorteile des Screenings von Versuchspersonen werden aufgezeigt. Betrachtungen aus statistischer und sozialwissenschaftlicher Sicht weisen auf Eckpunkte in der nachfolgenden Datendarstellung, -analyse und Ergebnisinterpretation hin und werden anhand von Beispielen verdeutlicht. Für den Forschungsbereich werden überdies Restriktionen, unter welchen Pilotversuche reliabel durchgeführt werden können, erläutert. Attributtabellen, eine Kostenaufwandsschätzung und eine Berichtvorlage ergänzen das Handbuch.

Abstract

Listening tests still represent indispensable instruments for Research and Development as well as for industry and economics concerning the achievement of subjective assessments of psychoacoustic parameters by man. The combination of the usage of a multi media room planned according to ITU-R BS 1116-1 together with an Expert Listening Panel incorporates optimized conditions for many applications of listening tests.

To implement standardised methods from the field of speech and audio into an existing institute with close proximity to research, instructions for the planning, performance and analysis of these procedures were developed.

To increase the significance of listening tests, it is yet essential during the planning phase to focus on the enquiry and to align the optimal test design with this question. The test design implies the correct elicitation of parameters, scales and attributes as well as the experiment design process to minimize bias. The use of naive listeners is opposed to the use of expert listeners with respect to generality, cost and availability and the advantages of screening of listeners are depicted. Considerations of data representation, analysis and the interpretation of results are shown from a statistical and a socio-scientific point of view and furthermore are illustrated with examples. In addition, for research, restrictions to conduct reliable pilot tests were added. Tables of attributes, an estimation of expense and a report form complete the handbook.

Formales Vorwort

Die Arbeit wurde unter folgenden, formalen Gesichtspunkten erstellt:

- Englische Fachbegriffe und Abkürzungen

  Um den Leser nicht unnötig mit Fachbegriffen aus dem deutschen zu verwirren, die vielerorts nicht übereinkommend existieren, wird gerade in Bezug auf internationale Standards zumeist der englische Fachbegriff beim erstmaligen Auftreten im Text in Klammern und Kursiv eingeführt. Dabei wird auf den zusätzlichen Hinweis „engl.:“ bewusst verzichtet.

  Sofern eine Abkürzung eines Wortes im Text nützlich erscheint oder eine Abkürzung für die Bezeichnung einer Methode allgemeinem Habitus entspricht, wird die Abkürzung ebenso in runde Klammern und kursiv gestellt bei erstmaligem Auftreten eingeführt.

- Gendern

  Die Anforderung die ich an mich als Frau stelle ist jene, ehrliche Gleichberechtigung zu leben. Das bedeutet für mich im Kontext meiner Diplomarbeit, nicht dem Versuch zu erliegen, sämtlichen Wortschatz der deutschen Sprache zwanghaft zu verweiblichen. Ebenso wenig ist es aber zielführend, lediglich die männliche Form zu nennen. Die verschiedenen Formen des /-innen, also der simultanen Berücksichtigung beider Geschlechter erscheint mir überdies für den Textfluss nicht sinnvoll und darüber hinaus eben auch ein wenig künstlich.

  Mein Ansatz sieht daher den Versuch vor, innerhalb dieser Arbeit die männliche und weibliche Form, sofern von Individuen die Rede ist, gleichermaßen häufig im Fließtext zu verwenden.

- Beispiele

  Beispiele im Text erscheinen kursiv. Es sei angemerkt, dass diese durchaus bewusst konstruiert erscheinen, um den Fokus klar auf die zu erklärende Problemstellung zu lenken.

Inhaltsverzeichnis

Abbildungsverzeichnis

Einleitung

„Wir produzieren Klänge für Menschen, und nicht für technische Messgeräte. Die finale Instanz muss daher immer das Gehör sein. Je mehr wir über das Gehör wissen und berücksichtigen, desto wahrscheinlicher wird es, dass wir das produzieren, was wir produzieren wollen.“ nach Prof. Dr. Ing Thomas Sporer

Die Psychoakustik ist ein Teilgebiet der Psychophysik (die wiederum einen Teil der experimentellen Psychologie darstellt) das versucht, die sensorische Wahrnehmung von akustischen Reizen durch den Menschen qualitativ oder quantitativ zu erfassen und zu beschreiben [Gre88].

Entscheidend für diese sensorische Wahrnehmung ist der Faktor Mensch. Der Mensch als Individuum ist nach wie vor nicht durch ein Modell repräsentierbar. Zu viele Einflussfaktoren wie Gemütslage, Prägung, Erwartungshaltung, Konzentrationsfähigkeit, Vorwissen u.a. machen es unumgänglich, die Person selbst in eine Bewertung von Sprach- oder Audioqualität nach bestimmten Attributen (das sind vorerst allgemein betrachtet Eigenschaften) einzubeziehen. Die Versuchsteilnehmerinnen können Menschen aus dem alltäglichen Leben ohne spezielle Vorkenntnisse oder aber auch eigens für eine bestimmte Anwendung trainierte Experten sein.

Ein Fokus der Psychophysik liegt jedenfalls auf dem Versuchsdesign, der Erstellung von Verfahren, Abläufen und psychometrischen Methoden, die mit der Verwendung von Versuchspersonen als komplexe Individuen valide und wiederholbare Antworten bzw. quantitativ erfassbare Ergebnisse liefern. Unerwünschte Einflüsse auf das Experiment, soweit diese bekannt sind, können eliminiert oder überwacht und damit konstant gehalten werden; potentiellen Fehlerquellen wird versucht, beispielsweise durch die Verwendung von standardisierten Verfahren, vorzubeugen. Die Mittel der Statistik werden dazu verwendet, Datenmengen aus einem subjektiven Hörtest auf ihre Validität und Objektivität hin zu prüfen, Hypothesen zu festigen oder zu verwerfen und damit Aussagen über eine Grundgesamtheit zu treffen.

Die Ergebnisse von subjektiven Hörversuchen finden einen großen Anwendungsbereich in Forschung und Wirtschaft. Im Forschungsbereich steht etwa das Experiment bei neuem Produktdesign, zur robusten Prozessentwicklung und -optimierung im Vordergrund. Produkte (Signale, Stimuli) können auf ihre Funktionalität hin genauso getestet werden, wie auf ihre Annehmlichkeit, Benutzerfreundlichkeit oder Kundenzufriedenheit. Die subjektive Beurteilung von akustischen Faktoren im Bereich des Benchmarking, beispielsweise in der Automobilindustrie, ist in ihrer Wertigkeit als kaufentscheidender Faktor bereits seit Jahren

unumstritten [Pfe11, Kap F,G]. Mit akustischen Faktoren kann das Audiosignal als Gesamtkonzept (beispielsweise der angenehme Klang der Soundanlage) genauso gemeint sein, wie ein Teilaspekt davon, z.B. die Klangfarbe, der wiederum als Optimierungsindikator des Prototyping fungiert.

Motivation

Den Ausgangspunkt bildet der Multimediaraum am JOANNEUM RESEARCH – Institut DIGITAL, welcher im Rahmen vom AAP - Projekt gemäß ITU-R BS.1116-1 als Referenzraum für Hörversuche geplant und errichtet wurde. Die zugehörige Rekrutierung eines Expert Listening Panels als objektive Prüfstelle sowie die softwaretechnische Umsetzung erfolgten dabei in Kooperation mit dem Institut für Signalverarbeitung und Sprachkommunikation der Technischen Universität Graz und dem Institut für elektronische Musik der Universität für Musik und darstellende Kunst Graz.

Das Potential der vorhandenen Ressourcen soll als Basis für die Umsetzung sowohl innovativer und forschungsnaher Aufgabenstellungen als auch wirtschaftsnaher Problemstellungen fungieren. Dabei steht die praktische Umsetzung psychoakustischer Verfahrensabläufe als interdisziplinäres Gebiet im Vordergrund mit dem Ziel, Synergien der vorhandenen Expertise von Ingenieuren, Statistikern, Medizinern und Audiotechnikern zu optimieren.

Das „Handbuch für Versuchsdesign in der Psychoakustik“ soll die Leserin während der Erarbeitung eines Hörversuchs von der Projektakquise bis zum Endbericht begleiten und als Entscheidungshilfe unterstützen. Ebenso soll für die Versuchsleiterin mit Erfahrung auf dem Gebiet die Möglichkeit bestehen, einzelne Kapitel nachzuschlagen.

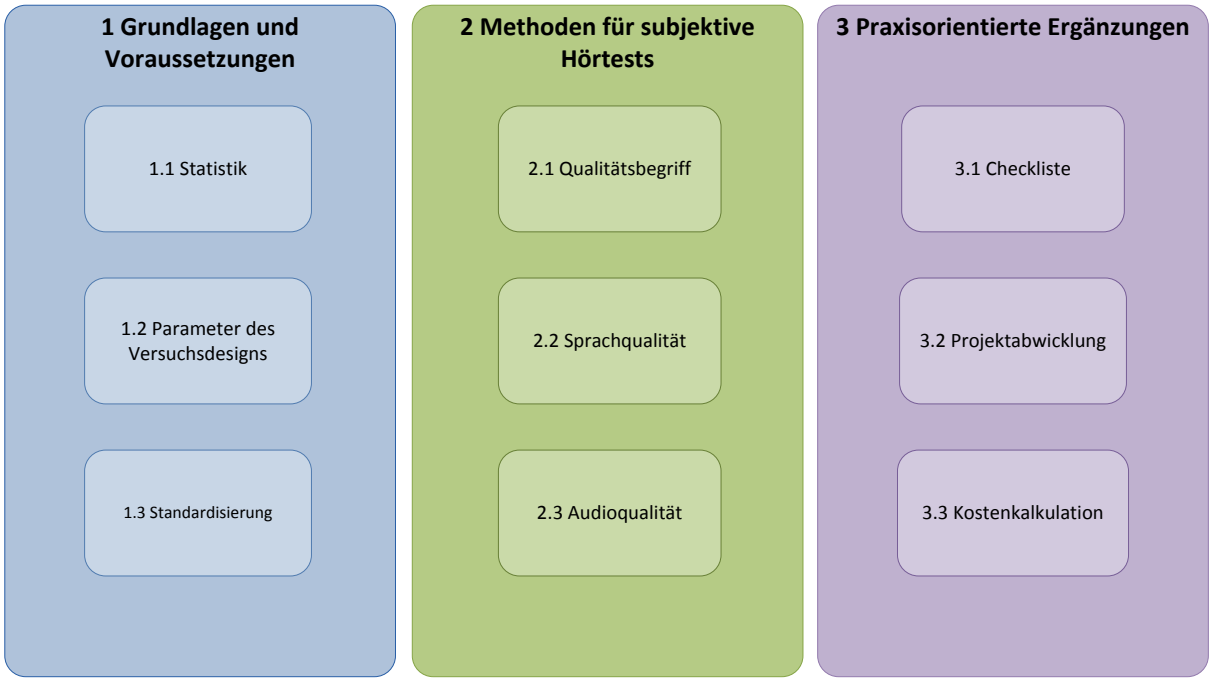

Abb. 1: Inhaltliche Gliederung des Handbuchs

Die Arbeit ist in drei Teile gegliedert, vgl. Abb. 1. Teil 1 behandelt allgemeine Grundlagen und Voraussetzungen zur Durchführung von Hörversuchen. Zu diesen gehören Teilbereiche der Statistik, Parameter des Versuchsdesigns wie die Spezifizierung von Versuchspersonen, Skalierungsmethoden oder Rahmenbedingungen und eine Erläuterung der zugrundeliegenden Standards. Abb. 2 gibt einen detaillierteren Überblick zu den Inhalten aus Teil 1. Der zweite Teil der Arbeit beschäftigt sich mit standardisierten Methoden zur Evaluierung von Sprach- und Audioqualität, ihrem Anwendungsbereich, sowie mit Besonderheiten, Vor- und Nachteilen des jeweiligen Verfahrens. Zu Beginn des Kapitels erfolgt eine Begriffsdefinition von Qualität. Der dritte Teil der Arbeit liefert praxisnahe Ergänzungen zur Durchführung. Eine Checkliste dient als Entscheidungshilfe in der Planung, weiterführende Gedanken zu Akquise, Kostenkalkulation, Projektabwicklung und Berichterstellung sollen den Ablauf erleichtern.

1 Grundlagen und Voraussetzungen

Die inhaltliche Struktur des ersten Teils der Arbeit, welcher die Grundlagen behandelt, soll der Leserin ohne Vorkenntnisse auf dem Gebiet des Versuchsdesigns entgegenkommen. Dementsprechend sind in Abschnitt 1.1 statistische Grundbegriffe zu finden. Es wird neben den Kennwerten der deskriptiven Statistik eine Einführung in die Möglichkeiten der Datendarstellung und der Datenanalyse gegeben und das Regressionsmodell wird vorgestellt. Zudem wird eine Möglichkeit für den Umgang mit Ausreißern im Datensatz gezeigt. Abschnitt 1.2 liefert grundlegende Parameter als Werkzeug für die Versuchsplanung und –durchführung. Dabei werden die verschiedenen Arten von Versuchspersonen ebenso definiert, wie der Informationsgehalt eines Skalenniveaus oder potentielle Fehlerquellen. Ansätze des Experiment Design werden ebenso beschrieben. Der dritte Abschnitt 1.3 betrachtet die Rolle der Standardisierung von Hörversuchen und erklärt die im Rahmen des Handbuchs relevanten Standards.

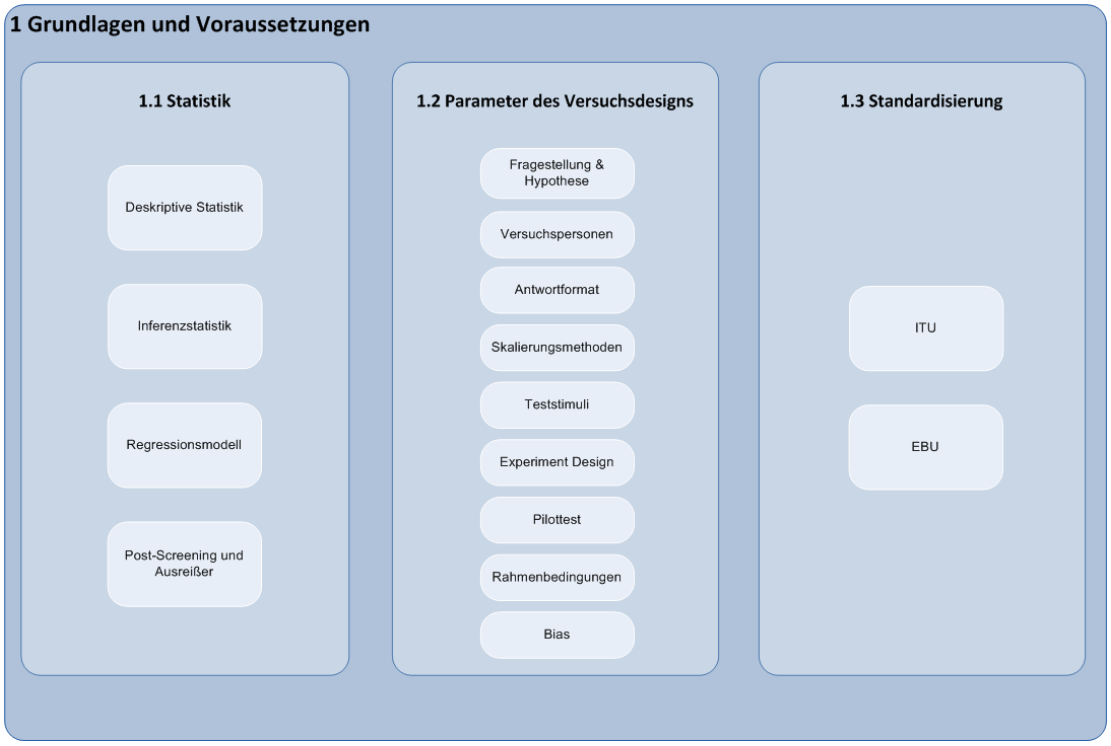

Abb. 2: Gliederung des ersten Teils des Handbuchs

1.1 Statistik

Dieses Kapitel beschäftigt sich mit den im Rahmen dieser Arbeit relevanten Grundbegriffen der Statistik und mit wichtigen Hintergrundinformationen, die zu einem optimierten Testdesign beitragen können. Dazu zählen beispielsweise eine Abschätzung der Anzahl an benötigten Versuchspersonen, die Herstellung eines Bezugs zum Informationsgehalt der Skalenniveaus und letztendlich auch die Erklärung von einigen Werkzeugen zur Datenanalyse. Außerdem wird auf die Überprüfung der Beschaffenheit/Qualität der Daten und deren Verteilungen eingegangen. Keinesfalls wird hier der Anspruch erhoben, den Statistiker zu ersetzen, lediglich Zusammenarbeit und Diskussion über die Thematik sollen für beide Parteien erleichtert werden.

Die Statistik wird in der Literatur zumeist in zwei Teilgebiete untergliedert, die jeweils auf einen spezifischen Bereich im Zuge des Testdesigns fokussieren. Dieses Kapitel gibt einen Überblick über die deskriptive und induktive Statistik und führt relevante Aspekte innerhalb der Gebiete näher aus. Sowohl die beschreibende als auch die schließende Statistik setzen sich mit den innerhalb vom Versuch erhobenen Daten näher auseinander.

Abb. 3 veranschaulicht den Prozess Versuchsdesign parallel aus Sicht der Statistik und soll einen Überblick über die Aufgabengebiete beider Disziplinen geben.

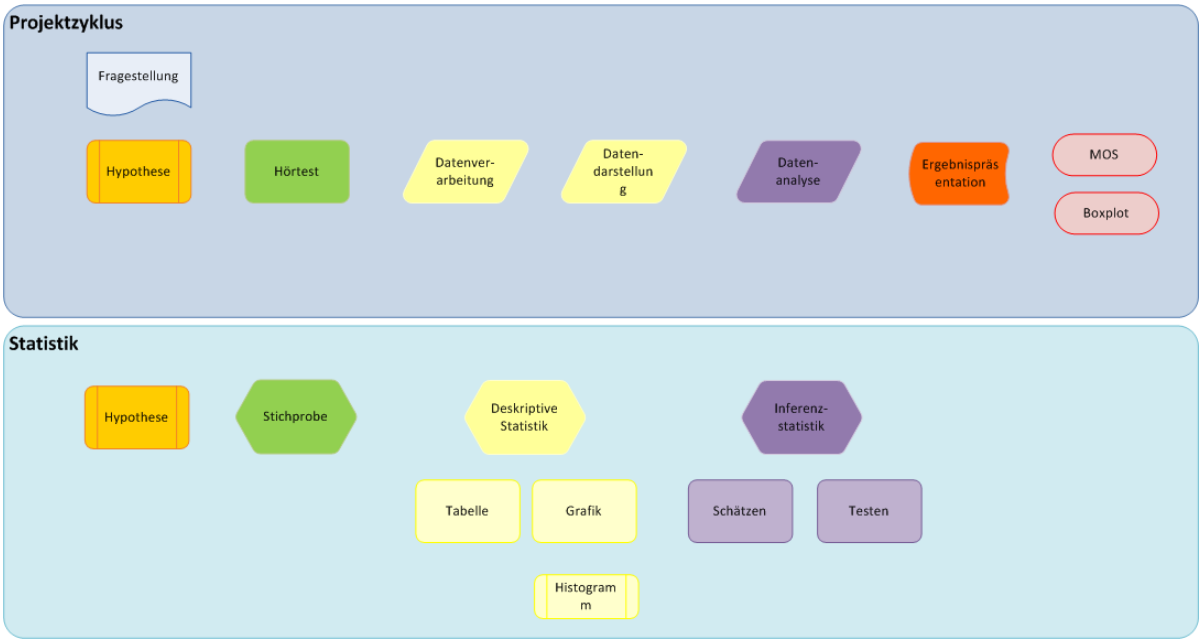

Abb. 3: Versuchsdesign aus statistischer Sicht

1.1.1 Deskriptive Statistik

Die deskriptive Statistik enthält Informationen über die empirisch gesammelten Daten. Sie beschreibt theoretische Hintergründe, Darstellungsformen, Arten von Skalen, und versucht, Datenmengen zu ordnen oder übersichtlicher zu gestalten. Die beschreibende Statistik macht keine Aussagen zur Grundgesamtheit. Damit einher geht die Tatsache, dass auch noch kein Wissen über die Aussagekraft des Datenmaterials bezüglich einer bestimmten Fragestellung vorhanden ist und dementsprechend auch Fehlschlüsse noch nicht kalkulierbar sind.

1.1.1.1 Kennwerte der deskriptiven Statistik entsprechend der Skalenarten

Grundsätzlich wird zwischen Kennwerten der Stichprobe - also den empirisch ermittelten Daten - und Kennwerten, die die Grundgesamtheit beschreiben unterschieden. Im Allgemeinen ist beispielsweise lediglich der Mittelwert $\bar{x}$ aus der Zufallsstichprobe bekannt, nicht aber der Mittelwert μ der Grundgesamtheit (ansonsten wäre die Prozedur eines Versuchs ad Absurdum geführt). Die deskriptive Statistik bedient sich der Kennwerte der Stichproben, wohingegen sich die induktive Statistik mit den Kennwerten der Grundgesamtheit befasst. Diese Kennwerte der Grundgesamtheit werden zumeist mit griechischen Buchstaben indiziert.

Für sämtliche nachfolgende Kennwerte ist das Skalenniveau der Daten entscheidend. Nominalskalierte Daten werden für die nachfolgenden Kennwerte nicht berücksichtigt, da mit ihnen lediglich einfachste Operationen durchführbar sind (vgl.

Tab. 7).

Lokalisationsmaße

Lokalisationsmaße, auch Maße der zentralen Tendenz oder Lagemaße genannt, repräsentieren Eigenschaften aller Datenpunkte der Menge. Dabei gilt es zu überlegen, welcher Kennwert sich bei gegebener Datenmenge am besten zur Veranschaulichung eignet.

- Median: Der Zentralwert ist robust gegenüber Ausreißern und zeigt die Datenmitte an, das bedeutet, er teilt die Datenmenge in zwei gleich große Hälften. Der Median kann im Gegensatz zum arithmetischen Mittel auch bei ordinal skalierten Daten verwendet werden und ist zudem robust gegen Ausreißer. Der Median ist auch das 0,5-Quantil einer Datenmenge.

$$\tilde{x}_{0,5} = \begin{cases} x_{((n+1)/2)} & , \textit{für n ungerade} \\ \frac{1}{2}(x_{(n/2)} + x_{((n+2)/2)}) & , \textit{für n gerade} \end{cases} \quad (1)$$

Nach [Har05, Kap1.4] wird der Median für eine gerade Anzahl n über das arithmetische Mittel der beiden mittigen Werte entsprechend (1) gebildet.

- Arithmetisches Mittel: Der Mittelwert setzt mindestens Intervallskalenniveau voraus und ist robust bei internen Werteverschiebungen, jedoch nicht gegen Ausreißer.

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right) = \frac{\sum_{i=1}^{n} x_n}{n} \qquad (2)$$

- winsorisiertes Mittel: Der Mittelwert, benannt nach Winsor Charles, kommt zur Anwendung wenn man Ausreißer im Datensatz erkennt.

  Dabei werden die Daten nach aufsteigender Größe sortiert und alsdann die Ausreißer auf beiden Seiten der Datenmenge durch ihren nächstgelegenen Datennachbarn ersetzt.

  Beispiel: Es sei eine Datenmenge x_1, x_2, …, x_{n-1}, x_n. x_1 und x_n werden als Ausreißer erkannt und durch ihre Nachbarn ersetzt. Das winsorisierte Mittel wird wie folgt gebildet:

$$\bar{x}_{win} = \frac{1}{n}\left(x_2 + x_2 + x_3 + x_4 + \ldots + x_{n-1} + x_{n-1}\right) \qquad (3)$$

  Der Vorteil gegenüber dem einfachen Wegschneiden der Daten besteht in der gleichbleibenden Anzahl *n* der Datenmenge (dies ist für die schließende Statistik von Bedeutung, vgl. Abs. 1.1.2.2, multiple Vergleiche). Grundsätzlich kommt die Verdichtung der Daten einem Informationsverlust gleich, der für den Fall Ausreißer gewünscht ist.

Dispersionsmaße (Streuungsmaße)

Da Lagemaße allein lediglich das Zentrum der Datenmenge beschreiben, nicht aber die Abweichung eines Merkmals von diesem, wird zusätzlich ein Streuungsmaß mit angegeben.

- Variationsbreite

  Die Variationsbreite oder Spannweite gibt den Wertebereich an, in dem die Datenmenge streut und wird über die Differenz zwischen dem Maximal- und Minimalwert gebildet.

$$VB = x_{\max} - x_{\min} \qquad (4)$$

Zur Bildung von VB wird lediglich Ordinalskalenniveau vorausgesetzt. Dabei kann mit der Kenntnis von VB allein noch keine Aussage über die Verteilung der Werte getroffen werden, jedoch ist erkennbar, dass die Variationsbreite stark von Extremwerten, also Ausreißern in den Daten abhängt.

Beispiel zur Verteilung der Werte: Ein einziger Teilnehmer bewertet den dargebotenen Stimulus mit 95, alle anderen Teilnehmer bewerten diesen in einem Bereich von 35 bis 42. Die Variationsbreite ergibt sich somit zu: $VB = 95 - 35 = 60$.

- Perzentil oder Quartilsabstand

  Perzentile betrachten restriktivere Teilbereiche einer Datenmenge, innerhalb derer sich entsprechend P_x ein gewisser Prozentsatz x [%] aller Daten befindet. Dies ist besonders dann sinnvoll, wenn Ausreißer im Datensatz potentiell vorhanden sind. Am häufigsten zur Anwendung kommen dabei Dezile (P_{10}, …, P_{90}) und Quartile (P_{25}, P_{50}, P_{75}). Somit ist der Interquartilbereich (bzw. der Interdezilbereich) jener Bereich, welcher mit den Grenzen P_{25} und P_{75} die mittleren 50 % an Streuung einer Datenmenge abbildet. Die Darstellung erfolgt beispielsweise in einem Boxplot, siehe Abs. 1.1.1.2, Datendarstellung.

- Varianz

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \qquad (5)$$

  Die Varianz ermittelt die Abweichung jedes einzelnen Wertes vom arithmetischen Mittel in quadratischer Form bezogen auf die Anzahl der Gesamtheit der Werte. Da das Quadrat in die Einheit eingeht, erweist sich die Standardabweichung als Streuungsmaß als repräsentativer. Die Varianz setzt intervallskalierte Daten voraus.

- Standardabweichung

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \qquad (6)$$

  Die Streuung oder Standardabweichung entspricht der Wurzel der Varianz und hat somit die gleiche Einheit wie der arithmetische Mittelwert. „Sie ist ein Maß für den *mittleren zu erwartenden Fehler der Einzelmessung*…“ [Hof04, Kap7].

  Sowohl Varianz als auch Standardabweichung referenzieren auf das arithmetische Mittel, das bedeutet, eine Angabe von s oder s^2 ohne $\bar{x}$ mit anzugeben ist nicht aussagekräftig. Selbige Aussage hat vice versa Gültigkeit.

  Bsp.: Die Standardabweichung ist mit s=20 gegeben. Je nachdem ob $\bar{x} = 5$ *oder* $\bar{x} = 5000$ *ist, wird die Streuung um den Mittelwert also groß oder klein sein.*

Hinweis: Eine Vergrößerung des Stichprobenumfangs führt nicht notwendigerweise zu einer Verringerung der Standardabweichung, da damit der Fehler einer einzelnen Vpn unbeeinflusst bleibt, nach [Hof04, Kap7]. Da trainierte Hörerinnen konsistentere Urteile bilden, die Streuung in ihrem Antwortverhalten im Allgemeinen also geringer ausfällt, ist die Versuchspersonenanzahl also im Vergleich zu naiven Hörern viel geringer.

- Variationskoeffizient

  Der Variationskoeffizient setzt arithmetisches Mittel und Standardabweichung direkt miteinander in Bezug, wie in nachfolgender Formel ersichtlich:

$$v = \frac{s}{\bar{x}} \tag{7}$$

  Er ist dimensionslos und macht einen Vergleich von Streuungen mehrerer Datenmengen, die verschiedene Mittelwerte haben, möglich.

zusätzliche Kennwerte für die unimodale Häufigkeitsverteilung (Formmaße)

Die beiden Kennwerte Schiefe und Exzess können lediglich für eingipflige Häufigkeitsverteilungen sinnvoll bestimmt werden.

- Schiefe

  Die Schiefe (engl.:*skewness*) gibt Auskunft über die Richtung, also links- oder rechtsschief, und das Ausmaß der Schiefe einer Häufigkeitsverteilung.

$$g_1 = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^3}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3 \tag{8}$$

$$g_1 > 0 \quad rechtsschief$$
$$g_1 < 0 \quad linksschief$$
$$g_1 = 0 \quad symmetrsich$$

  Für gruppierte Daten ist lediglich der Wert x_i in Gleichung (8) durch das arithmetische Mittel der Gruppierung zu ersetzen.

- Exzess

  Der Exzess (engl.:*curtosis*), auch Wölbung oder Kurtosis genannt, beschreibt die Breite des Unimodus, also die Breitgipfligkeit der Häufigkeitsverteilung. Dabei wird das absolute Maximum mit jenem der Normalverteilung (*NV*) verglichen [Har05, Kap1.5].

$$g_2 = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4\right)}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4 \quad (9)$$

$$g_2 > 0 \quad abs.Max. > NV$$
$$g_2 < 0 \quad abs.Max. < NV$$
$$g_2 = 0 \quad abs.Max. = NV$$

1.1.1.2 Datendarstellung

Man unterscheidet nach Bortz Variablen, Merkmalsausprägungen, Daten, unter Anderem nach der Anzahl ihrer Ausprägungen dahingehend, ob sie diskret (endlich viele Ausprägungen) oder stetig (beliebige reelle Werte) sind. Diskrete Daten, in die ein untersuchtes Merkmal anhand von den genannten Skalentypen unterteilt wird, können in einem Stabdiagramm dargestellt werden [Har05, Kap1].

Stabdiagramm

Das Säulen- oder Stabdiagramm (Barplot) eignet sich gut zur Visualisierung der Häufigkeitsverteilung einer oder mehrere diskreter Variablen. Im Gegensatz zum Histogramm ist hier lediglich die Höhe des Stabes aussagekräftig. Das Balkendiagramm, eine horizontale Darstellung der Häufigkeit von Daten, ist dem Säulendiagramm ähnlich.

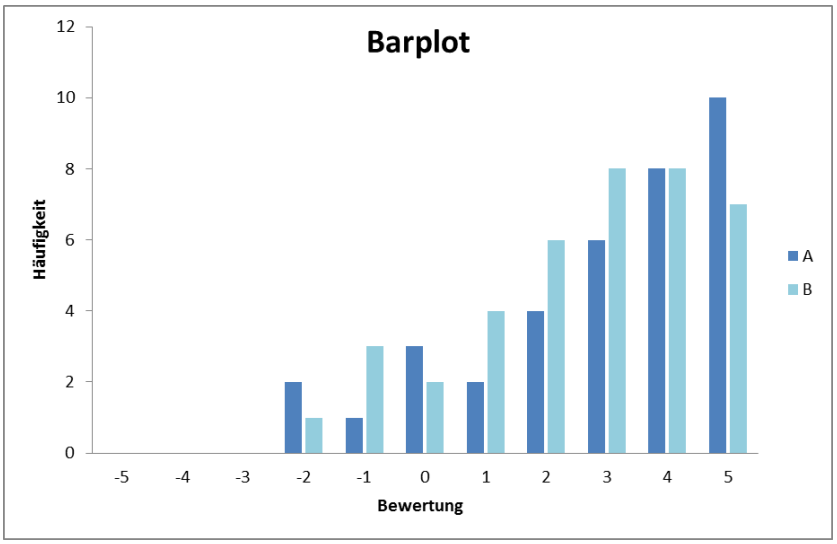

Abb. 4: Barplot von zwei Stichproben

In Abb. 4 sind die Vor- und Nachteile dieser Darstellungsform gut ersichtlich. Einzelne Kategorien sind bezüglich ihrer verschiedenen Häufigkeiten in den zwei Bewertungen A und B gut vergleichbar, die Übersichtlichkeit bezüglich einer einzelnen Gesamtbewertung, beispielsweise A, geht jedoch verloren (man stelle sich einen Vergleich von Bewertungen mit einer Anzahl >2 vor).

Boxplot

Der Boxplot vereint wichtige Parameter zur prägnanten Charakterisierung und Visualisierung einer Datenmenge. Zu diesen gehören folgende Lokalisations- und Streuungsmaße:

- Minimal- und Maximalwert (potentielle Ausreißer)
- Median
- Variationsbreite
- Unteres und oberes Quartil

In der Darstellung sind Ausreißer besonders gut zu erkennen. Mehrere Datensätze nebeneinander je als Boxplot dargestellt ermöglichen einen prägnanten Überblick beispielsweise über das beste System im Test, etc.

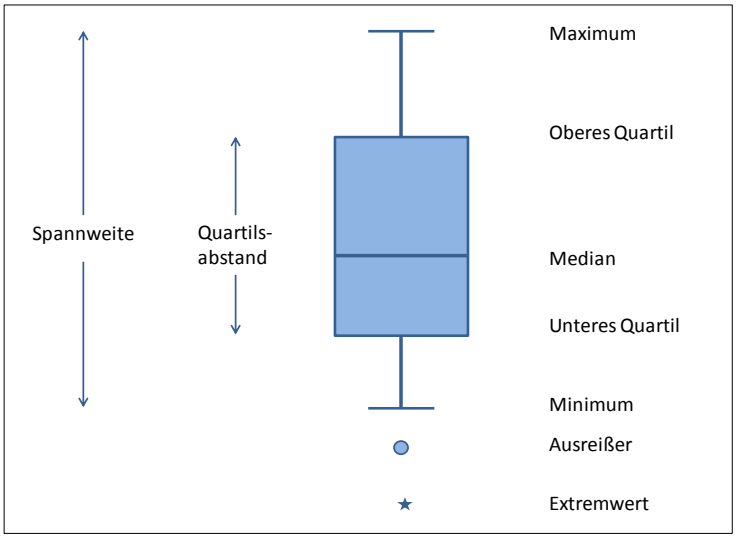

Abb. 5: Box-Plot

Netzdiagramm

Als letzte Darstellungsform sei speziell zum Zweck der Visualisierung der Abhängigkeit eines Parameters (beispielsweise Gesamteindruck) von verschiedenen Attributen das Netz- oder Spinnendiagramm erwähnt. Dabei wird der Hauptparameter, dessen Abhängigkeiten von Interesse sind, auf der y-Achse aufgetragen, die Attribute entsprechend ihrer Anzahl als Hauptnetzlinien (oder Hauptachsen). Die Achsenunterteilung ist dabei durch die Bewertung der Gesamtqualität, die beispielsweise auf einer MOS-Skala erfolgt, vorgegeben.

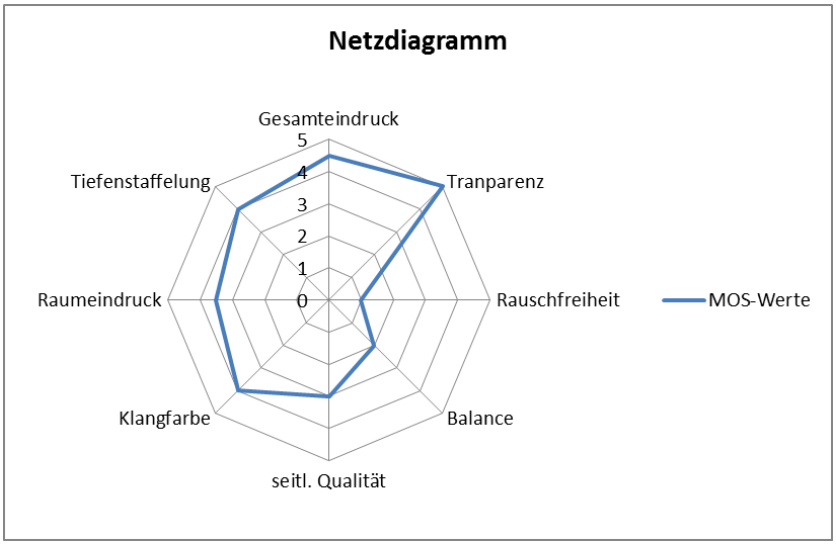

Abb. 6: Darstellung der Hauptparameter für Multikanalwiedergabe, [3286]

Die Darstellungsform eignet sich besonders zur Ergebnispräsentation von Versuchen, in denen verschiedene Aspekte der Klangqualität betrachtet wurden, wie zum Beispiel in Abb. 6 dargestellt.

1.1.2 Inferenzstatistik

Die Inferenzstatistik, auch induktive oder schließende Statistik genannt, prüft, ob und in welcher Weise von einer Stichprobe auf die Grundgesamtheit geschlossen werden kann. Das bedeutet gleichermaßen dass die aufgestellte Hypothese mit den Mitteln der Stochastik auf ihren allgemeinen Gültigkeitsbereich überprüft wird. Man beachte hier den umgekehrten Ansatz in der Aussagendefinition: ausgehend von der Empirie (Stichprobe, Daten aus dem Hörtest) werden Rückschlüsse auf die Theorie (Grundgesamtheit, allgemeine Aussage) getroffen. Dabei bedient sich die Inferenzstatistik der Werkzeuge der Wahrscheinlichkeitstheorie und Stochastik.

Eine Einteilung der Methoden innerhalb dieses Kapitels wurde anhand der Verteilung und des Skalenniveaus vorgenommen.

Innerhalb der inferenzstatistischen Betrachtungen werden die Kennwerte nicht abermals mathematisch ausgeführt oder erläutert, da sie abgesehen von der griechischen

Nomenklatur und dem Bezug auf Grundgesamtheit bzw. Stichprobe dieselbe Aussage haben wie jene Werte der Deskriptive.

1.1.2.1 Stichprobe/n und Grundgesamtheit

Um überhaupt Aussagen über die Grundgesamtheit treffen zu können und zu wissen, welches statistische Testverfahren zur weiteren Analyse anwendbar ist, benötigt man einerseits Kenntnis über die Verteilung der Stichprobe (d.h. es wird überprüft, ob die Stichprobe groß genug ist, um eine Aussage über die Grundgesamtheit machen zu können) , andererseits ist das Skalenniveau (vgl. Abs. 1.2.4.2) interessant. Die aus dem Hörversuch gewonnenen Daten (die Stichprobe) werden also vorerst auf Normalverteilung hin überprüft. Mit Hilfe von Anpassungstests kann also die Wahrscheinlichkeitsverteilung einer Zufallsvariablen, demnach der Typ einer Verteilung, geprüft werden. Signifikanztests überprüfen demgegenüber einen Parameter der Verteilung, untersuchen also die Aussagekraft eines Parameters [Har05, KapIII].

Sind die Daten normalverteilt, können Testverfahren aus der parametrischen Statistik (vgl. Abs. 1.1.2.2) angewendet werden. Hat die Datenmenge hingegen keine Normalverteilung, stehen Tests der parameterfreien Statistik (nicht parametrische Statistik, vgl. Abs. 1.1.2.3) zur Verfügung. Grundsätzlich wird eine Normalverteilung der Daten bereits in der Planungsphase des Versuchs angestrebt, da mit parametrischen Verfahren genauere Aussagen zur Grundgesamtheit getroffen werden können.

zentraler Grenzwertsatz, Normalverteilung des Mittelwerts der Stichprobe

Nachfolgende Erläuterung bezieht sich idealerweise auf eine große Stichprobe und gleichzeitig eine intervallskalierte Datenmenge. Selbst wenn für kleine Stichproben[1] ebenso wie für ordinalskalierte Datenmengen[2] Restriktionen vorzunehmen sind, ist diese Vorgehensweise plakativ.

Es sei X ein Merkmal aus der Grundgesamtheit. Dieses Merkmal ist in der Grundgesamtheit verteilt mit einem unbekannten Mittelwert μ, dem Erwartungswert und einer unbekannten Varianz σ^2. Um μ zu bestimmen, wird eine Zufallsstichprobe (mit einer Anzahl n) aus der Grundgesamtheit gezogen, und der Mittelwert $\bar{x}$ aus der Zufallsvariable $\overline{X}$ berechnet. Je größer nun die Zufallsstichprobe ist, desto mehr nähert sich dieser berechnete Mittelwert dem wahren Erwartungswert der Grundgesamtheit an. Um zu wissen wie weit $\bar{x}$ von μ abweicht, ist es notwendig die Standardabweichung $\sigma_{\bar{x}}$ von $\bar{x}$ zu kennen. Die Standardabweichung $\sigma_{\bar{x}}$ des Mittelwertes $\bar{x}$ ist proportional zu σ der Grundgesamtheit, dieses σ ist aber nicht bekannt und muss daher geschätzt werden.

[1] Kleine Stichproben: es kann nicht mehr von einer Normalverteilung der Stichprobe ausgegangen werden. Die Verteilung wird nach Student-t betrachtet und ist abhängig von den Freiheitsgraden bei Annahme einer normalverteilten Grundgesamtheit. Eine Stichprobe ist klein, wenn n<30. [Bor04]

[2] Die Betrachtungen für ordinalskalierte Datenmengen beschränken sich lediglich auf den Median.

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} \qquad (10)$$

$\sigma_{\bar{x}} \ldots Standardabweichung\ von\ \bar{x}$

$\sigma^2 \ldots Varianz\ der\ Grundgesamtheit, unbekannt$

$$\hat{\sigma}^2 = \underbrace{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}_{s^2} \cdot \underbrace{\frac{n}{n-1}}_{Schätzfaktor} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad (11)$$

$\hat{\sigma}^2 \ldots geschätzte\ Varianz\ der\ Grundgesamtheit$

Der Schätzfaktor (vgl. Abs. 1.1.2.2, t-Test, Freiheitsgrad *df*) kompensiert die Tatsache, dass die Stichprobenvarianz die Populationsvarianz nicht erwartungstreu schätzt:

$$E(S^2) = \sigma^2 - \sigma_{\bar{x}}^2 = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2 \cdot \frac{n-1}{n} \qquad (12)$$

$E(S^2) \ldots Erwartungswert\ der\ Stichproben\ varianz\ s^2$

Aus der geschätzten Varianz der Grundgesamtheit $\hat{\sigma}^2$ lässt sich nun die geschätzte Standardabweichung des Mittelwertes berechnen und somit die Mittelwertverteilung charakterisieren.

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{\hat{\sigma}^2}{n}} \qquad (13)$$

Daraus folgt nun, dass Stichprobenmittelwerte $\bar{x}$ nach dem zentralen Grenzwertsatz normalverteilt sind um μ, bzw. in folgenden Intervallen mit genannter Wahrscheinlichkeit liegen:

$$\begin{aligned} &\mu \pm \hat{\sigma}_{\bar{x}} \quad 68\% \\ &\mu \pm 2\hat{\sigma}_{\bar{x}} \quad 95{,}5\% \end{aligned} \qquad (14)$$

nach [Bor04, Kap5].

Anpassungstests

Als Anpassungstests an die Normalverteilung, die in diesem Fall vorrangig interessiert, können der Chi^2-Anpassungstest oder der Anpassungstest nach Kolmogoroff-Smirnov verwendet werden.

- Chi^2-Anpassungstest: zwei mögliche Hypothesen H_0 können bei festgelegtem Signifikanzniveau α und bekannten μ_0 sowie σ_0^2 getestet werden:
    - H_0: Die Grundgesamtheit ist $N(\mu_0, \sigma_0^2)$-verteilt.
    - H_0: Die Stichprobe ist $N(\mu_0, \sigma_0^2)$-verteilt.
- Kolmogoroff-Smirnov: Der Test überprüft, ob die tatsächliche Verteilungsfunktion der Grundgesamtheit mit einer gewünschten übereinstimmt und wird aufgrund der approximativen Funktionsweise bei kleinen Stichprobenumfängen angewendet.

Nähere Informationen zu Anspassungstests findet man zum Beispiel bei Hartung [Har05, KapIV].

Hypothesenprüfung und Signifikanzniveaus

In den inferenzstatistischen Betrachtungen können verschiedene Fragestellungen untersucht werden. Zu diesen gehören beispielsweise:

- H_0: Stimmen die Erwartungswerte /Varianzen zweier Stichproben überein? Zugrunde liegende Frage: Unterscheiden sich die Erwartungswerte/Varianzen zweier Stichproben signifikant?
- H_0: Stimmen Erwartungswerte/Varianzen mehrerer Stichproben überein? Zugrunde liegende Frage: Unterschieden sich die Erwartungswerte/Varianzen mehrerer Stichproben signifikant?

Ist die Hypothese aufgestellt, wird man sich aufgrund der vorhandenen Stichprobe für die Null- oder Alternativhypothese entscheiden. Dabei ist es in beiden Fällen möglich, sich für die falsche Hypothese entschieden zu haben (was man im Allgemeinen nicht weiß, es besteht lediglich eine Wahrscheinlichkeit dafür). Ein Fehler 1. Art entsteht, wenn man sich für die H_1 entscheidet, obwohl die H_0 gültig ist, umgekehrtes trifft für den Fehler 2.Art zu. Entscheidend ist nun folgendes zu wissen:

- je nach Fragestellung kann das tatsächliche Eintreten des α- oder des β-Fehlers den gravierenderen Ausgang des Versuchs darstellen. (man stelle sich vor, das bessere zweier Hörgeräte aufgrund einer Fehlentscheidung für das schlechtere zu halten und deswegen nicht weiter zu entwickeln, vermarkten…)
- Man kann festlegen, dass mit einer Wahrscheinlichkeit α ein Versuch ausgewertet wird, ohne dabei eine Aussage über β zu treffen. Dabei möchte man β ebenfalls möglichst gering halten.

- übliche Werte für das Signifikanzniveau α sind:
    - α=5% (signifikant)
    - α=1% (höchst signifikant).

Nachfolgende Tabelle gibt einen Überblick über die beiden Fehlerarten.

Entscheidung für	**Es liegt vor**	
	H_0	**H_1**
H_0	Richtig	**Fehler 2. Art (β-Fehler)**
H_1	**Fehler 1. Art (α-Fehler)**	richtig

Tab. 1: Fehlerarten beim Testen von Hypothesen [Har05, Kap4]

Stichproben abhängig/unabhängig

Liefert ein Hörtest mehrere Stichproben, besteht er also aus mehreren Durchläufen und testet dabei verschiedene Attribute eines Systems, können diese Stichproben voneinander abhängig oder unabhängig sein.

Bsp.: Ein Lautsprecher soll nicht alleinig hinsichtlich seiner Gesamtqualität (Basic Audio Quality) beurteilt werden, zusätzlich wird die Klangfarbe untersucht. Der gewünschte Versuch soll zudem möglichst kosteneffizient durchgeführt werden. Dazu beschließt die Versuchsleiterin, die Durchführung auf eine Versuchspersonengruppe zu reduzieren. Als Vpn wird ein ELP ausgewählt, da die Anzahl der Vpn geringer ist, gleichzeitig die Antworten aufgrund der geschulten Hörfähigkeit reliabler sind und überdies bei Audioanwendungen EL naiven Hörern zu bevorzugen sind. Den Vpn werden zwei Durchläufe präsentiert, einmal zur Bewertung der Gesamtqualität , der andere Durchlauf dient der Beurteilung der Klangfarbe. Die beiden Durchläufe führen zu zwei Stichproben, die voneinander abhängig sind. Die Abhängigkeit wird hier durch dieselben Teilnehmer in beiden Durchläufen hergestellt. Ein Versuch bei dem zwei Versuchsteilnehmergruppen je ein Attribut behandeln würden, würde zu zwei voneinander unabhängigen Stichproben führen.

- Abhängige Stichproben (gepaarte Stichproben): Dieselbe Versuchsperson bewertet Systeme in zwei Durchläufen hintereinander hinsichtlich je eines Attributs. (für jeden Durchlauf entsteht als Summe aller Vpn eine Stichprobe → zwei Stichproben) Die beiden Stichproben stehen über die Versuchsperson miteinander in Zusammenhang (*ein typisches Beispiel hierfür repräsentiert der vollständige Paarvergleich*).

- Unabhängige Stichproben: Experten bewerten ein System hinsichtlich seiner Qualität (eine Stichprobe) und naive Hörer (oder eine zweite Gruppe von Experten) bewerten dasselbe System nach denselben Kriterien (eine Stichprobe). Die Stichproben sind voneinander unabhängig.

1.1.2.2 Parametrische Statistik

Voraussetzung für die Anwendung parametrischer Testverfahren ist eine Normalverteilung der Daten.

Eine normalverteilte Datenmenge wird meist über den arithmetischen Mittelwert beziehungsweise den Erwartungswert μ und die Standardabweichung σ der zugehörigen Verteilungsfunktion charakterisiert.

Normalverteilung

Im Versuchsdesign entsprechend standardisierter Verfahren wird oftmals von quasi intervallskalierten Daten ausgegangen, die zudem Normalverteilung aufweisen.

Die Normalverteilung $N(\mu,\sigma^2)$ ist eine unimodale, symmetrische Häufigkeitsverteilung mit glockenförmigem Verlauf, die durch die Funktion $f_x(x)$, die Wahrscheinlichkeitsdichte oder Dichte einer normalverteilten Zufallsvariable X, näherungsweise beschrieben wird:

$$f_x(x) = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad (15)$$

$$mit \quad \begin{matrix} \mu & \ldots Erwartungswert \\ \sigma^2 & \ldots Varianz \end{matrix}$$

Das bedeutet idealerweise, dass Modalwert, Median und Erwartungswert zu einem Punkt zusammenfallen. Außerdem, dass Schiefe und Exzess idealerweise Null sind.

Zudem gibt die Standardabweichung in folgender Art Auskunft über die Häufigkeitsverteilung der Daten:

$$\begin{matrix} \overline{X} \pm \sigma & \sim 68\% \, aller \, Werte \, in \, diesem \, Bereich \\ \overline{X} \pm 2\sigma & \sim 95\% \, aller \, Werte \, in \, diesem \, Bereich \end{matrix} \qquad (16)$$

Die zu $f_x(x)$ gehörige Verteilungsfunktion $F_x(x)$:

$$F_x(x) = P(X \le x) \qquad (17)$$

wird zur vereinfachten Lösung linear transformiert (z-Transformation):

$$Y = \frac{(X - \mu)}{\sigma} \qquad (18)$$

und somit in Standardnormalform gebracht, nach [Har05, KapII.7].

Standardnormalverteilung

Sie bildet eine besondere Form unter den Normalverteilungen und wird mit N(0,1) abgekürzt, das bedeutet ihr Erwartungswert ist μ=0 und ihre Standardabweichung ist σ=1. Ihre Verteilungsfunktion ϕ der Dichte ist tabelliert und macht somit eine Wahrscheinlichkeitsbestimmung möglich.

Die Dichte φ(y) ist dabei gegeben mit:

$$\varphi(y) = \sigma \cdot f_x\left(\left(y + \frac{\mu}{\sigma}\right)\sigma\right) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot y^2} \qquad (19)$$

wobei $f_x(x)$ als Wahrscheinlichkeitsdichte der Normalverteilung (Gauß-) gegeben ist, nach [Har05, KapII.7].

Damit folgt für die Wahrscheinlichkeit:

$$P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \phi\left(\frac{x-\mu}{\sigma}\right) = F_x(x) \qquad (20)$$

Mit ϕ(x) als Verteilungsfunktion der Standardnormalverteilung, nach [Har05, Kap4.1].

Normalisierung von Ergebnissen

Eine Datenmenge wird nach nachfolgender Formel normalisiert, sobald auf verbale Deskriptoren in der Skala verzichtet wird (vgl. Abs. 1.2.4.3, Direkte Antwortskalierung).

Für die i-te Versuchsperson gilt:

$$z_i = \frac{x_i - \bar{x}_i}{s_i} \cdot s + \bar{x} \qquad (21)$$

$z_i \ldots z - Wert\ der\ i - ten\ Vpn$
$x_i \ldots Beurteilung\ der\ i - ten\ Vpn$
$\bar{x}_i \ldots Mittelwert\ der\ Vpn\ im\ Durchlauf$
$s_i \ldots S\,\mathrm{tan}\,dardabweichung$
$s \ldots S\,\mathrm{tan}\,dardabweichung\ aller\ Vpn$
$\bar{x} \ldots Mittelwert\ aller\ Vpn$

nach [1284, 4.1].

t-Test

Die Student-t-Verteilung[3] gehört zu den stetigen Verteilungen, ist schmalgipflig und geht mit zunehmendem Freiheitsgrad df→∞ in die Standardnormalverteilung über.

Anwendung

Der t-Test findet grundsätzlich Anwendung bei intervallskalierten Daten aus kleinen Populationen, bei denen die Datenpunkte nicht ausreichen, um einen gültigen Mittelwert und eine Standardabweichung entsprechend einer Normalverteilung zu generieren.

Überprüft werden können Hypothesen, die

- den Mittelwert einer Stichprobe $\bar{x}$ mit dem Erwartungswert μ einer Grundgesamtheit vergleichen.
- zwei Stichprobenmittelwerte $\bar{x}_1, \bar{x}_2$ aus unabhängigen Stichproben miteinander vergleichen.
- zwei Stichprobenmittelwerte $\bar{x}_1, \bar{x}_2$ aus abhängigen Stichproben miteinander vergleichen.

Voraussetzungen:

Nachdem für die Stichproben selbst nicht von einer Normalverteilung ausgegangen werden kann, weil die Anzahl an Versuchspersonen zu gering ist, geht man von einer Normalverteilung der jeweiligen Grundgesamtheit, aus der die Stichprobe gezogen wird, aus. (*Beispiel: naive Hörer repräsentieren eine Stichprobe; die naive Bevölkerung repräsentiert die normalverteilte Grundgesamtheit*)

Die theoretische t-Verteilung ist um einen Freiheitsgrad reduziert, d.h.:

$$df = n - 1 \qquad (22)$$

$df \ldots Freiheitdg\,rad\,(\deg ree\ of\ \ freedom)$
$n \ldots Stichprobengröße$

[3] t-Verteilung: 1908 von Gosset unter dem Pseudonym „Student“ entwickelt. [Bor04, Kap2.5]

t-Test: Vergleich von $\bar{x}$ mit bekanntem μ

Die gerichteten Hypothesen für den Erwartungswert der Grundgesamtheit lauten:

$$H_0 : \mu \leq a$$
$$H_1 : \mu > a \qquad (23)$$

Man beachte, dass sich beide Hypothesen auf die Grundgesamtheit beziehen. Sodann ermittelt man den Stichprobenmittelwert $\bar{x}$ und die zugehörige Streuung $\sigma_{\bar{x}}$ aus (10), wenn σ bekannt ist, sonst aus (13) aus den Beurteilungen der Hörer.

Ist die Stichprobe klein (n<30) und σ unbekannt, wird der t-Wert für die Teststatistik wie folgt berechnet:

$$t_{obs} = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}} \qquad (24)$$

$$t_{obs} \ldots beobachtete\,\mathrm{Pr}\ddot{u}fgröße\,(observed)$$

(Selbige Formel wird im Fall einer großen Stichprobe angewendet. Lediglich für den Fall der zusätzlich bekannten Standardabweichung der Grundgesamtheit σ, kann die geschätzte Streuung durch die empirische Streuung $\sigma_{\bar{x}}$ ersetzt werden und die Prüfgröße wäre somit nicht mehr mit n-1 Freiheitsgraden t-verteilt, sondern standardnormalverteilt.)

Dieser empirische Wahrscheinlichkeitswert wird nun mit dem kritischen t-Wert aus der Tabelle verglichen, welcher abhängig von Signifikanzniveau α und Freiheitsgrad gebildet wird.

$$t_{1crit} = t(df)_{1-\alpha}$$
$$|t_{2crit}| = t(df)_{1-\frac{\alpha}{2}} \qquad (25)$$

$$t_{1crit} \ldots krit.t-Wert\ für\ 1-seitige\ H_0$$
$$|t_{2crit}| \ldots krit.t-Wert\ für\ 2-seitige\ H_0$$

Erfolgt eine einseitige Hypothesenprüfung, wird für α=5% für den jeweiligen Freiheitsgrad die Spalte mit 0,95 gewählt. Bei zweiseitiger Hypothesenprüfung wird das Vertrauensintervall von 5% auf beide Seiten (und somit je 2,5%) der Verteilung aufgeteilt, weswegen die Spalte mit 0,975 gewählt würde (vgl. (25)).

$$t_{obs} \overset{?}{=} t_{crit} \qquad (26)$$

Beispiel: Es wurde ein Pilottest durchgeführt, in dem zehn Hörer die Qualität eines Aufnahmegeräts auf einer kontinuierlichen Skala von 0-11 bewerteten. Die gerichteten Hypothesen lauten beispielsweise wie folgt:

$$H_0 : \mu \leq 6$$
$$H_1 : \mu > 6 \qquad (27)$$

Die übrigen Berechnungen ergeben sich aufgrund der kleinen Stichprobe und einseitiger Hypothesenprüfung zu:

$$n = 10; \mu = 6; \bar{x} = 7; \hat{\sigma}^2 = 1{,}6$$
$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{1{,}6}{10}} = 0{,}4$$
$$t_{obs} = \frac{7-6}{0{,}4} = 2{,}5 \qquad (28)$$
$$t_{crit} = t(9)_{0{,}95} = 1{,}833$$
$$t_{obs} > t_{crit} \Rightarrow H_0 \; verworfen$$

Die Alternativhypothese, welche besagt, dass der Erwartungswert der Grundgesamtheit über dem Wert 6 liegt, wird also angenommen. Es kann demzufolge angenommen werden, dass die Grundgesamtheit das Aufnahmegerät mit 95%iger Wahrscheinlichkeit für überdurchschnittlich hinsichtlich seiner Gesamtqualität bewerten würde.

nach Bortz [Bor04, Kap5].

t-Test mit unabhängigen Stichproben

Es erfolgt hier der Vergleich zweier Mittelwerte aus zwei unabhängigen Stichproben.

Die ungerichtete H_0 postuliert, dass die Differenz der beiden Erwartungswerte Null ist:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0 \qquad (29)$$

Die mit der Tabelle zu vergleichende Prüfgröße berechnet sich zu:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{(\bar{x}_1 - \bar{x}_2)}} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{(\bar{x}_1 - \bar{x}_2)}} \qquad (30)$$

Dabei wurde die Varianz der Grundgesamtheit der beiden Stichproben wiederum geschätzt nach (11) und die Differenz der Streuung berechnet sich zu

$$\hat{\sigma}_{(\bar{x}_1-\bar{x}_2)} = \sqrt{\frac{(n_1-1)\cdot\hat{\sigma}_1^2+(n_2-1)\cdot\hat{\sigma}_2^2}{(n_1-1)\cdot(n_2-1)}}\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \quad (31)$$

Der Freiheitsgrad für kleine Stichproben (n<30) ist hier

$$df = n_1 + n_2 - 2 \quad (32)$$

Damit der t-Test robust ist, sollten die beiden zu vergleichenden Stichprobenumfänge gleich groß und deren Grundgesamtheiten normalverteilt sein. Ist dies nicht der Fall, sollten beiden Varianzen gleich sein, um Fehlentscheidungen durch den Test zu vermeiden [Bor04, Kap5].

Um zu wissen, ob die Voraussetzung erfüllt ist, kann hier der F-Test verwendet werden. Der F-Test prüft die Varianz zweier Stichproben dahingehend, ob die Stichproben derselben Grundgesamtheit entstammen und somit die Streuungen zufällig durch die Stichprobe verursacht sind.

Trifft die Normalverteilungsvoraussetzung nicht zu, kann statt des t-Tests der Test nach Mann Whitney angewendet werden (vgl. Abs. 1.1.2.3, Mann Whitney).

t-Test mit abhängigen Stichproben

Um zwei Mittelwerte aus abhängigen Stichproben miteinander vergleichen zu können, wie dies beispielsweise bei einem vollständigen Paarvergleich der Fall ist, ist die Abhängigkeit der Varianzen der Stichproben zueinander zu berücksichtigen und dementsprechend die Vorgehensweise auf die Differenz der Beurteilungspaare zu adaptieren. Das bedeutet für die ungerichtete H_0, dass der Erwartungswert μ_d der Differenz zweier Beurteilungen Null ist:

$$H_0 : \mu_d = 0 \quad (33)$$

Demnach werden die Mittelwerte für die Differenzen der Beurteilungspaare d_i gebildet,

$$\bar{x}_d = \sum_{i=1}^{n} d_i \quad (34)$$

$$mit\, d_i = x_{i1} - x_{i2}$$

und deren Streuung geschätzt nach:

$$\hat{\sigma}_{\bar{x}_d} = \frac{\hat{\sigma}_d}{\sqrt{n}} \quad (35)$$

mit

$$\hat{\sigma}_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{x}_d)^2}{n-1}} \quad (36)$$

$\hat{\sigma}_d \ldots geschätzte\ Streuung\ der\ Differenzen$

Die Prüfgröße ergibt sich dann zu

$$t = \frac{\bar{x}_d - \mu_d}{\hat{\sigma}_{\bar{x}_d}} = \frac{\bar{x}_d}{\hat{\sigma}_{\bar{x}_d}} \quad (37)$$

Der Freiheitsgrad für die Stichprobenpaare ist hier:

$$df = n_a - 1 \quad (38)$$

$n_a \ldots Anzahl\ der\ Beurteilungspaare$

Damit der t-Test für abhängige, kleine Stichproben robust ist, wird davon ausgegangen, dass die Differenzen in der Grundgesamtheit, aus der beide Stichproben stammen, normalverteilt sind [Bor04, Kap5].

Trifft diese Normalverteilungsvoraussetzung hingegen nicht zu, kann stattdessen der Test nach Wilcoxon angewendet werden (vgl. Abs. 1.1.2.3, Wilcoxon-Test).

Varianzanalyse - ANOVA

Die Varianzanalyse (*analysis of variance, ANOVA*) dient als Auswerteverfahren für Auswirkungen von mehreren, unabhängige Variablen auf eine abhängige Variable. Dementsprechend ist sie dann anzuwenden, wenn für einen Versuch eine Vielzahl an t-Tests durchzuführen wäre. Grundsätzlich werden drei verschiedene Modelle voneinander unterschieden. Modell I betrachtet feste Effekte (*fixed effect models*), Modell II zufällige Effekte (*random effect models*) und Modelle der Gruppe III mischen feste und zufällige Effekte (*mixed effect models*) als Betrachtungsgrundlage. Anschließend wird die Vorgehensweise zur zwei-Faktor (*two way*) Vairanzanalyse erläutert, die als Basis für sämtliche, komplexere Aufgabenstellungen dient. Die ANOVA ist in gängiger Statistiksoftware (SPSS, R, Matlab) implementiert. Weiterführend sei vor Allem auf Ott [Ott10] verwiesen, es können aber auch Bortz [Bor04, Kap7] und Bech [Bec06, Kap6] verwendet werden.

Vorgehensweise

Ausgehend von Beurteilungen nach einem Hörversuch, in dem p Systeme miteinander verglichen wurden, betrachten wir die Bewertungen des i-ten Systems (i=1,…,p). Es wird angenommen dass die Beurteilungen $Y_{i1},\ldots,Y_{ini}$ aus einer $N(\mu,\sigma^2)$-verteilten Grundgesamtheit stammen. Somit kann ein Vergleich der Mittelwerte $\mu_1,\ldots, \mu_p$ über entsprechende Schätzer mit dem F-Test angewendet werden. Kann man hingegen nicht von einer normalverteilten Grundgesamtheit ausgehen, ist stattdessen der Test von Kruskal und Wallis anzuwenden.

Die zu verwerfende Nullhypothese postuliert, die Mittelwerte aller Systeme wären gleich. Demnach gäbe es keinen signifikanten Unterschied zwischen den Systemen.

- Gesamtanzahl aller Beurteilungen aller Systeme

$$N = \sum_{i=1}^{p} n_i > p \qquad (39)$$

$N \ldots Anzahl\ aller\ Beurteilungen\ des\ Versuchs$
$n_i \ldots Beurteilungen\ je\ System / Gruppe$
$p \ldots Anzahl\ an\ Systemen\ im\ Versuch$

- Mittelwerte berechnen

$$\overline{Y_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \qquad (40)$$

$\overline{Y_i} \ldots Mittelwert\ einer\ Gruppe / Datenreihe / Systems$
$Y_{ij} \ldots j-te\ Beurteilung / Datenpunkt\ d.i-ten\ Systems$

Der in Gleichung (40) angeführte Mittelwert repräsentiert den Schätzer für den wahren Mittelwert μ_i der Grundgesamtheit.

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij} \qquad (41)$$

$\overline{Y} \ldots Gesamtmittelwert\ aller\ Beurteilungen$
$Y_{ij} \ldots j-te\ Beurteilung / Datenpunkt\ d.i-ten\ Systems$

- Berechnung der Variation zwischen den Gruppen

$$SS_T = \sum_{i=1}^{p} n_i (\overline{Y_i} - \overline{Y})^2 \qquad (42)$$

$SS_T \ldots Sum\ of\ squares\ for\ treatments$

- Berechnung der Variation innerhalb einer Gruppe

$$SS_E = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2 \quad (43)$$

$SS_E \ldots Sum of\ squares\ for\ errors$

- Die Gesamtvariation setzt sich zusammen aus:

$$SS_G = SS_T + SS_E \quad (44)$$

$SS_G \ldots Grand\ sum of\ squares$

- Definition der mittleren Quadratsummen/Varianzen aus den beiden Variationen:

$$mS_T = \frac{SS_T}{p-1} \quad (45)$$

$mS_T \ldots mean\ of\ squares\ for\ treatments$
$df = p-1 \quad Freiheitsgrade$

Die Varianz zwischen den Gruppen (auch mit *between groups* assoziiert) soll möglichst groß sein, um zu gewährleisten, dass ein signifikanter Unterschied zwischen den Gruppen besteht beziehungsweise erkennbar ist.

$$mS_E = \frac{SS_E}{N-p} \quad (46)$$

$mS_E \ldots mean\ of\ squares\ for\ errors$
$df = N-p \quad Freiheitsgrade$

Die Fehlerquadratsumme (mS_E, auch mit *within groups* assoziiert) repräsentiert einen erwartungstreuen Schätzer für die Varianz σ^2. Die Varianz innerhalb einer Gruppe soll demnach möglichst klein sein, um zu zeigen, dass die Variabilität im Antwortverhalten der Vpn gering ist.

Beide Varianzen geben somit Auskunft über das Beurteilungsverhalten der Versuchspersonen bzw. des Panels.

- Für die Überprüfung der Hypothese mit Hilfe des F-Tests gilt nun:

$$F = \frac{mS_T}{mS_E} \qquad (47)$$

F ist folglich abhängig von den Freiheitsgraden im Zähler und Nenner. Dieser beobachtete F-Wert wird nun mit den tabellierten Quantilen zur Hypothesenprüfung verglichen und liefert Signifikanz zum Niveau γ für:

$$F > F_{p-1,N-p;1-\gamma} \qquad (48)$$

$$F_{p-1,N-p;1-\gamma} \ldots vertafeltes\, Quantil\, der\, F-Verteilung\ \ zum\, Niveau\, \gamma$$

Sämtliche angeführte Werte sind, gerade wenn die ANOVA softwaremäßig kalkuliert wurde, in einer Tabelle zu finden, die beispielsweise folgendermaßen aussieht:

Streuungsursache	Sum of Squares	Degrees of freedom	Mean Square	F	Signifi-cance
Zwischen Gruppen	SS_T	p-1	mS_T	F	0.00
Innerhalb d. Gruppe	SS_E	N-p	mS_E	-	-
Gesamt	SS_G	N-1	-	-	-

Tab. 2: ANOVA-Tabelle zum Modell I

nach [Har05, KapXI].

Die ANOVA prüft letztendlich also die H_0, dass kein signifikanter Unterschied in den Mittelwerten der Beurteilungen des Systems besteht. Die F-Statistik liefert demnach bestenfalls einen signifikanten Wert, um die H_0 zu verwerfen.

Multiple Vergleiche

Für den Fall dass ein signifikanter Unterschied in den Mittelwerten besteht, ist es beispielsweise interessant zu wissen, welche dieser Mittelwerte aus den p Systemen paarweise verschieden sind. Dazu verwendet man die multiplen Vergleiche nach Scheffé oder Tukey.

Es wird darauf hingewiesen, dass es eine Vielzahl an post hoc Tests gibt, für die unterschiedliche Voraussetzungen gelten und die unterschiedliche Fragen abklären. Als Beispiel sei noch Dunett genannt, der es ermöglicht, eine Kontrollgruppe mit einer Vielzahl an Gruppen zu vergleichen, ohne dabei die beurteilenden Gruppen miteinander zu vergleichen (*beispielweise sinnvoll, wenn per Internet zig Gruppen wiederum bestehend aus zig Vpn weltweit ein System beurteilen*).

Scheffé und Tukey

Beide Tests setzen wiederum eine normalverteilte Grundgesamtheit voraus. Der Unterschied der beiden Tests liegt im Wesentlichen im Anwendungsbereich. Scheffé wendet man bei verschieden großen Gruppen (≙versch. lange Messreihe) an, Tukey bei gleich großen Gruppen.

Scheffé gilt als konservativer Test und kann auch verwendet werden, wenn eine verschiedene Anzahl an Beurteilungen zum jeweiligen Mittelwert geführt hat. Demnach wird der Test beispielsweise eingesetzt, wenn aus einer Datenmenge Ausreißer entfernt wurden.

Tukey ist der in den Standards vorgeschlagene Post-Hoc-Test und setzt voraus, dass für jeden Mittelwert gleich viele Beurteilungen vorhanden sind.

Nachdem die beiden Tests die Grenzdifferenz auf verschiedene Weise kalkulieren, führen sie auch zu unterschiedlichen Ergebnissen, weswegen durchaus vorgeschlagen wird, nicht ausschließlich Scheffé einzusetzen.

Nach [Har05, KapXI].

1.1.2.3 Nicht Parametrische Statistik

Die nicht-parametrische Statistik ist jenes Gebiet innerhalb dessen Verfahren zur Hypothesenprüfung eingesetzt werden, die keine Annahme über die Verteilung einer Menge oder ihrer Grundgesamtheit machen. Demnach können sie angewendet werden, wenn die Stichproben zu klein für die Annahme einer Normalverteilung sind. Nachfolgend werden zwei häufig verwendete Verfahren, die Einsatz bei ordinalskalierten Daten finden, erläutert.

Mann Whitney

Der Test nach Mann Whitney wird für ordinalskalierte Datenmengen zur Überprüfung der zentralen Tendenz zweier unabhängiger Stichproben verwendet. Dabei erfolgten die Beurteilungen auf einer Skala mit k Kategorien. Den Beurteilungen beider Stichproben, jede Datenmenge ist durch eine Spalte repräsentiert, werden entsprechend einer gemeinsam aufsteigenden Rangfolge Zahlenränge zugeordnet. Für gleiche Beurteilungen erfolgt eine Bindungskorrektur (*Beispiel: Beurteilung 1 kommt in beiden Systemen je einmal vor, würde also die Rangplätze 1 und 2 erhalten. Daraus das Mittel gebildet ergibt für den Rang 1,5.*). Nachfolgendes Beispiel der Beurteilung zweier Systeme auf einer ACR Skala, soll die Vorgehensweise verdeutlichen:

System 1	System 2
4	2
3	5
3	4
2	2
1	1
	4

Tab. 3: Beurteilung von zwei Systemen anhand von 5 Kategorien

Die Anzahl an Beurteilungen muss also nicht zwangsläufig gleich sein. *In diesem Beispiel ist* $n_1=5$, $n_2=6$.

System 1	Rang	System 2	Rang
4	9	2	4
3	6,5	5	11
3	6,5	4	9
2	4	2	4
1	1,5	1	1,5
		4	9

Tab. 4: Zuweisung von Rangzahlen zu den Beurteilungen

Aus diesen Rangzahlen kann nun für jede Stichprobe die Rangsumme r und der Durchschnitt $\bar{r}$ gebildet werden.

Die beiden Rangsummen sind dabei über folgende Beziehung miteinander verknüpft:

$$r_1 + r_2 = \frac{n \cdot (n+1)}{2} \qquad (49)$$

$$mit\ n = n_1 + n_2$$

$$r \ldots Rangsumme$$

$$n \ldots Anzahl\ aller\ Beurteilungen$$

Der U-Wert wird nach folgender Beziehung ermittelt:

$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - r_1 \qquad (50)$$

Selbe Formel gilt analog unter Verwendung vertauschter Indizes für das zweite System.

Der U-Wert setzt sich zusammen aus den Vergleichen des Rangs für einen Wert mit der Anzahl aus höheren Rängen aus dem Vergleichssystem. (*Beispiel: System 1, 1. Beurteilung hat den Wert 4 und dieser den Rang 9 → Vergleich mit System 2: wie viele Ränge sind dort höher als 9? usw. Man könnte den U-wert also auch auszählen.*)

$$\begin{aligned} &n_1 = 5 \quad n_2 = 6 \\ &r_1 = 27{,}5 \quad r_2 = 38{,}5 \\ &\bar{r}_1 = 5{,}5 \quad \bar{r}_2 = 6{,}417 \\ &U_1 = 17{,}5 \quad U_2 = 12{,}5 \\ &mit \quad U_1 + U_2 = n_1 \cdot n_2 \end{aligned} \qquad (51)$$

Die Nullhypothese behauptet, dass sich die beiden Systeme nicht voneinander unterscheiden. Demnach würden wir mit einem Erwartungswert für U und einer Streuung wie folgt rechnen :

$$\begin{aligned} \mu_U &= \frac{n_1 \cdot n_2}{2} \\ \sigma_U &= \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}} \end{aligned} \qquad (52)$$

Die Prüfgröße lautet dann wie folgt:

$$z = \frac{U - \mu_U}{\sigma_U} \qquad (53)$$

Da U_1 und U_2 symmetrisch zu μ_U liegen, kann einer der beiden Werte wahlweise für U eingesetzt werden.

$$\mu_U = 15$$
$$\sigma_U = 30 \qquad (54)$$
$$z = 0{,}083$$

In der Tabelle ergibt sich für U=12 eine Irrtumswahrscheinlichkeit von 33,1 %.

Je nach der Anzahl an Versuchspersonen ist nun die Überprüfung der Prüfgröße z zu unterscheiden: bis n_1 bzw n_2 von 8 sind die exakten Irrtumswahrscheinlichkeiten vertafelt, für $1 \le n_1 \le 20$ und $9 \le n_2 \le 20$ sind kritische U-Werte für die einseitige und zweiseitige Prüfung tabelliert [Bor04, TabF]. Zudem kann für eine Anzahl an Beurteilungen >10 von einer annähernden Normalverteilung ausgegangen werden und der z-Wert auch in der Standardnormalverteilungstabelle eingesehen werden.

Nach [Bor04, Kap5.2].

Wilcoxon-Test

Der Test nach Wilcoxon überprüft als verteilungsfreies Verfahren ebenfalls ordinalskalierte Datenmengen hinsichtlich ihrer zentralen Tendenz. Dabei werden im Gegensatz zu Mann Whitney abhängige Stichproben betrachtet, die wiederum in k Kategorien beurteilt wurden.

Analog zum t-Test für abhängige Stichproben werden für die Beurteilungspaare Differenzen gebildet, diese dürfen auch negative sein. Die Differenzen werden dann wieder in eine Rangfolge gebracht, bei der das Vorzeichen zusätzlich berücksichtigt wird. Nachfolgendes Beispiel soll dies verdeutlichen.

Experten bewerten ein System 1 in zwei Durchläufen:

System 1 erster Run	System 1 zweiter Run	Differenz	Rangfolge
4	2	2	3,5
3	5	-2	3,5(-)
3	4	-1	1,5(-)
3	2	1	1,5
1	1	0	
5	1	4	5

Tab. 5: Beispiel für einen Wilcoxon-Test

Wie in Tab. 5 ersichtlich bleiben Paare mit übereinstimmenden Beurteilungen unberücksichtigt. Die Anzahl n der Paare wird in diesem Fall um eins reduziert. Gibt es viele Übereinstimmungen zwischen den Beurteilungen, ist die Nullhypothese bestätigt. Diese besagt, dass es keinen Unterschied in der zentralen Tendenz der beiden Stichproben gibt. Zudem wird eines der beiden Vorzeichen gekennzeichnet. Die Rangsummen werden für beide Vorzeichen getrennt gebildet:

$$r_{neg} = 5; \quad r_{pos} = 10; \quad n = 5$$
$$r_{neg} + r_{pos} = \frac{n \cdot (n+1)}{2} \qquad (55)$$

$$n \ldots Anzahl\ der\ Differenzpaare$$

Nach Bortz ist die H_0 umso unwahrscheinliche je weiter die beiden Rangsummen voneinander abweichen.

Der Erwartungswert für r ergibt sich zu:

$$\mu_r = \frac{n \cdot (n+1)}{4} \qquad (56)$$

Dieser Erwartungswert wird nun mit der Rangsumme r verglichen, deren Vorzeichen in der Tabelle seltener aufschien. Je deutlicher die beiden Werte dann voneinander abweichen, desto unwahrscheinlicher ist die H_0.

Für das Beispiel bedeutet das:

$$\mu_r = 7{,}5$$
$$r_{neg} = 5$$
$$mit\ n = 5,\ \alpha = 5\% \rightarrow Tabelle \qquad (57)$$
$$r_{krit} = 0$$

Da die empirische Rangsumme größer als die kritische ist, darf die H_0 nicht verworfen werden. Dazu müsste der empirische Wert mindestens so stark vom Erwartungswert abweichen wie der kritische.

1.1.3 Regressionsmodell

Die statistische Analyse zur Auswertung von Daten aus einem Hörtest basiert auf dem Wunsch, den Zusammenhang oder die Auswirkung von unabhängigen Variablen auf eine abhängige Variable darzustellen und zu untersuchen. Dieser Zusammenhang wird nachfolgend anhand eines linearen, additiven Modells aus der Regressionsanalyse, erläutert.

$$Y_{t,i} = \mu + \alpha_t + \varepsilon_{t,i}$$
$$mit\ \mu_t = \mu + \alpha_t \qquad (58)$$

$Y_{t,i} \ldots i-te\ Bewertung\ der\ Vpn\ für\ t-ten\ Stimulus$
$\mu_t \ldots \mathrm{Re}aktion\ der\ Vpn\ auf\ i-te\ Wiederholung\ des\ t-ten\ Stimlus,$
$die\ durch\ die\ kontrollierten\ Variablen\ im\ Experiment\ festge\mathrm{lg}t\ wird$
$\mu \ldots arithmetischer\ Mittelwert\ aller\ Bewertungen\ im\ Experiment\ (idealisiert)$
$\alpha_t \ldots Verarbeitungseffekt, E\mathrm{inf}luss\ der\ kontrollierten\ Variablen\ verursacht$
$durch\ den\ t-ten\ Stimulus\ wenn\ für\ den\ Gesamtdurchschnitt\ korrigiert$
$\varepsilon_{t,i} \ldots Effekt, verursacht\ durch\ zufälligen\ \mathrm{exp}erimentellen\ Fehler$

Zur Versuchsplanung gehört die Auflistung von Variablen, die in mehrere Klassen unterteilt werden [Bec06, Kap5]:

- Unabhängige Variable α_t: ist die zu prüfende Variable, das Merkmal, dessen Auswirkung auf die abhängige Variable getestet wird.
- Abhängige Variable $Y_{t,i}$: ist die Variable, deren Auswirkung auf die unabhängige Variable bewertet wird, idealerweise ohne Störeinflüsse.
- Kontrollvariable: auch Moderatorvariable genannt, wird miterhoben, um im Nachhinein Aussagen über deren Einfluss/Kontrollfunktion auf die abhängige Variable machen zu können. (z.B. Raum, Lautsprecherposition, etc) Sie sind aber streng genommen nicht Teil des Experiments.
- Unkontrollierte Variablen: diese unterteilt man weiter in
    - nicht berücksichtigte Variablen: sind zwar unter Umständen für den Versuch von Bedeutung, werden aber entweder nicht berücksichtigt (da ansonsten der Aufwand immens würde) oder aber auch nicht erkannt bzw. vergessen.
    - Störvariablen: der Versuchsleiter ist bemüht, diese Variablen im Zuge des Versuchsdesigns in Variablen mit zufälligem Einfluss zu konvertieren, um den Störfaktor zwischen unabhängiger und abhängiger Variable zu minimieren (und sie somit im weitesten Sinn zu „kontrollieren“).

- Zufällige Variablen: ihr Effekt auf die abhängigen Variablen ist von rein zufälliger Natur. Sie sind vergleichbar mit dem zufälligen Fehler in der Messtechnik.[4]

Im Allgemeinen kann man sagen, dass unbekannte Variablen für die Varianz des Fehlers verantwortlich sind. Die nicht berücksichtigten Variablen und auch die Störvariablen des angeführten Modells können die Ursache für einen systematischen Fehler darstellen. Dieser streut nicht wie der zufällige Fehler um einen Mittelwert, sondern manifestiert sich als konstante Abweichung, die oftmals unerkannt bleibt und somit als „normale“ Bewertung fehlinterpretiert wird. Die Schwierigkeit dieser Fehlerart liegt aufgrund der Fehlinterpretation auch in mangelnden Möglichkeiten im Bereich des Postprocessing. Umso wichtiger ist es, sich der Fehlerursachen im Designprozess bewusst zu sein und diese möglichst zu vermeiden, vgl. Abs. 1.2.9.

1.1.4 Post – Screening und Ausreißer

Statistische Betrachtung von Ausreißern

Als Ausreißer werden Daten aus den Beurteilungen des Hörers angesehen, die im Vergleich zur Datenmenge der übrigen Hörer extrem abweichen oder stark schwanken. Die Entscheidung zu treffen, einen vom Normverhalten abweichenden Wert als Ausreißer und nicht als Wissenszuwachs zu postulieren, ist auch mit statistischen Mitteln oftmals nicht trivial. Aus diesem Grund ist die Ursache für das Zustandekommen eines Extremwerts grundsätzlich vorerst kritisch zu hinterfragen (Warum ist es zu diesem Datenpunkt gekommen?).

Anmerkung: Die erhobene Datenmenge wird üblicherweise nach Versuchsende und vor der statistischen Analyse mit Hilfe deskriptiver Techniken veranschaulicht und auf ihre logische Konsistenz hin überprüft. Hatte die Messskala lediglich das Niveau einer Nominal- oder Ordinalskala, sind die statistisch ermittelbaren Werte gegenüber Ausreißern relativ robust (beispielsweise Median). Erst wenn eine Intervallskala vorliegt und Mittelwert und Standardabweichung berechenbar sind, gewinnt die Betrachtung von potentiellen Ausreißern an entscheidender Bedeutung.

Eine Möglichkeit, besonders bei großen Stichprobenumfängen n≥25 abzuschätzen (beispielsweise beim Einsatz naiver Hörer), ob es sich um einen Ausreißer handelt, ist die allgemeine Regel des „4-Sigma-Bereichs“. Diese besagt, dass ein Wert als Ausreißer betrachtet und somit auch von der Analyse ausgeschlossen werden kann, wenn er außerhalb des nachfolgenden Wertebereichs liegt:

$\mu \pm 4\sigma$ (Mittelwert ± 4 mal Standardabweichung)

[4] *Zufälliger Fehler*: bei wiederholter Messung streut das Ergebnis um einen Mittelwert (hebt sich bei ∞ Wiederholung auf). Der z.F. kann in mindestens einem der Merkmale Amplitude, Vorzeichen oder Zeitpunkt seines Auftretens nicht vorhergesagt werden.[Hof04, Kap7]

Die Regel wird allerdings auch gerne bei Umfängen ab n=10 zur Abschätzung herangezogen.

Die Statistik bedient sich für verschiedene Stichprobenumfänge (*entspricht z.B. der Anzahl an Versuchspersonen*), jedoch unter der Voraussetzung normalverteilter Daten, sogenannten Ausreißer Tests. Gerade für kleine Stichproben n≤25 und einseitige Hypothesenprüfung ist folgender Test nach Dixon hilfreich zur Ausreißerdetektion:

Die gewonnenen Datenpunkte werden ihrer Größe nach so angeordnet, dass der potentielle Ausreißer x_1 jedenfalls den Beginn der Rangfolge bildet. Je nachdem, ob der Wert nun der kleinste oder größte im Datenset ist, ergeben sich somit zwei Möglichkeiten:

$$x_1 < x_2 < x_3 < ... < x_n$$
$$x_1 > x_2 > x_3 > ... > x_n$$

Nun wird die Prüfgröße *M* entsprechend gebildet:

$M_1 = \left| \frac{x_1 - x_2}{x_1 - x_n} \right|$ für 3≤n≤7; $M_2 = \left| \frac{x_1 - x_2}{x_1 - x_{n-1}} \right|$ für 8≤n≤10;

$M_3 = \left| \frac{x_1 - x_3}{x_1 - x_{n-1}} \right|$ für 11≤n≤13; $M_4 = \left| \frac{x_1 - x_3}{x_1 - x_{n-2}} \right|$ für 14≤n≤25

Die Prüfgröße *M* wird mit dem kritischen Wert für das jeweilige Signifikanzniveau aus der Tabelle verglichen. Ist *M* größer als jener kritische Wert, stellt x_1 einen Ausreißer dar [Sac02, Kap36].

Teststatistik		Signifikanzniveau α			Teststatistik		Signifikanzniveau α		
	n	α=0,1	α=0,5	α=0,01		n	α=0,1	α=0,5	α=0,01
	3	0,886	0,941	0,988		14	0,492	0,546	0,641
	4	0,679	0,765	0,889		15	0,472	0,525	0,616
M1	5	0,557	0,642	0,780		16	0,454	0,507	0,595
	6	0,482	0,560	0,698		17	0,438	0,490	0,577
	7	0,434	0,507	0,637		18	0,424	0,475	0,561
	8	0,479	0,554	0,683		19	0,412	0,462	0,547
M2	9	0,441	0,512	0,635	M4	20	0,401	0,450	0,535
	10	0,409	0,477	0,597		21	0,391	0,440	0,524
	11	0,517	0,576	0,679		22	0,382	0,430	0,514
M3	12	0,490	0,546	0,642		23	0,374	0,421	0,505
	13	0,467	0,521	0,615		24	0,367	0,413	0,497
						25	0,360	0,406	0,489

Tab. 6: kritische Werte nach Dixon [Sac02, Kap38]

Beispiel: An einem Hörversuch nehmen 12 Expert Listener teil. Diese beurteilen die Signale anhand einer kontinuierlichen Skala mit Schrittweite 0,1. Die gewonnenen Datenpunkte werden graphisch dargestellt. Auffälligkeiten innerhalb der Versuchsabläufe können so schneller detektiert werden und für diese abweichenden Datenpunkte kann eine Überprüfung mittels Ausreißertest nach Dixon durchgeführt werden. Der Ausreißertest wird hier anhand eines beurteilten Stimulus gezeigt:

Die Beurteilungen für Stimulus x seien wie folgt:

3,7 3,9 3,7 4,0 3,2 3,6 3,8 3,8 3,8 3,0 3,6 3,7

Frage: Ist der Datenpunkt mit dem Wert x_1=3,0 ein Ausreißer? Beantwortung über Dixon:

- *Rangfolge bilden mit Startpunkt bei x_1=2,8:*

  3,0 3,2 3,6 3,6 3,7 3,7 3,7 3,8 3,8 3,8 3,9 4,0

- *Berechnung von $\hat{M}$ für n=12:*

$$\hat{M} = \left| \frac{3,0 - 3,6}{3,0 - 3,9} \right| = 0,\dot{6}$$

- *Vergleich mit kritischem Wert aus der Tabelle für α=0,05:*

  $0{,}\dot{6} > 0{,}546$

  Bei einem Signifikanzniveau von α=0,05 kann x_1 als Ausreißer angesehen werden. Das bedeutet jedoch nicht zwangsläufig, dass bei einem Signifikanzniveau von α=0,01 der Datenpunkt x_1=3,0 auch als Ausreißer anzusehen ist.

Bezug zum Versuchsdesign

Versteckte Referenzen und Anker werden aus Sicht der Vpn wie jeder übrige Stimulus im Test bewertet. In der Regel wurde die Hörerin vor Beginn des Versuchs über das Vorhandensein einer versteckten Referenz in Kenntnis gesetzt.

Die Intention dabei ist, die Versuchsperson dazu zu veranlassen, die vermeintliche Referenz als das Optimum (das sie ja darstellt) innerhalb des jeweiligen Trials[5] zu beurteilen. Die Beurteilung der Referenz kann demnach als Indikator für Ausreißer innerhalb eines Datensets herangezogen werden.

Bsp: Die Skala geht von 1, also schlecht, bis 5, demnach sehr gut (das Beispiel ist bewusst allgemein gehalten mit „die Skala" und die Bewertung überdies dem amerikanischen angepasst). Außerdem besteht der Test aus 30 Trials. Innerhalb eines Trials werden mehrere Stimuli, darunter auch eine hidden reference, miteinander verglichen. Die versteckte Referenz soll idealerweise in jedem Trial mit 5 bewertet werden.

Um zuverlässige Ergebnisse zu erhalten, kann man nun Bedingungen festlegen, die, sollten sie nicht erfüllt werden, zu einem Ausschluss der Daten für die statistische Analyse führen. Folgende zwei Möglichkeiten werden vorgeschlagen:

- *Grenzwert: wird die Referenz innerhalb eines Trials schlechter als 4,5 bewertet, also <4,5, wird dieses Trial der VP x nicht ausgewertet; das Trial stellt also einen Ausreißer im Datensatz der VP x dar.*

- *VP als Ausreißer: beurteilt die Testperson innerhalb des Versuchs in a=3 oder mehr Trials die Referenz nicht mit dem Optimum (also hier mit 5), geht der vollständige Datensatz der Versuchsperson nicht in die statistische Analyse ein.*

Es wird darauf hingewiesen, dass bei Anwendung von Ausschlussbedingungen zur Ausreißerdetektion diese den Versuchspersonen vor Testbeginn bekannt sein müssen. Überdies ist diese Vorgehensweise statistisch betrachtet mit Vorsicht zu genießen, da sie zu einer Verfälschung der Daten führen kann und sollte demnach mit einem Ausreißertest oder der 4-Sigma-Bereichs-Regel überprüft werden (siehe vorheriger Absatz zur Statistischen Betrachtung).

[5] Trial (engl.; Probe): ein Trial besteht aus dem Abspielen von Samples plus zugehöriger Beantwortung durch die Vp. Mehrere Trials formen einen run (Durchlauf).

1.2 Parameter des Versuchsdesigns

Um die Relevanz der in diesem Abschnitt beschriebenen Parameter zu verdeutlichen wird der Ablauf eines typischen Projekts aus Sicht des den Hörversuch durchführenden Instituts erklärt. Den zentralen Ausgangspunkt der Darstellung (vgl. Abb. 7) bildet der Projektablauf. Parallel zu diesem wird der benötigte Input dargestellt, der als Voraussetzung für einen reibungslosen Ablauf anzusehen ist. Der Output zeigt neben den vom Auftraggeber geforderten Leistungen relevante Dokumente und Daten. Eine Möglichkeit der Verantwortungsverteilung zeigt der Ansatz über das DEMI-Modell. Folgende vier Verantwortliche werden definiert:

- Versuchsleiter: der Versuchsleiter hat die meiste Expertise auf dem Gebiet der Akquise, Planung und Durchführung von Hörversuchen. Er trifft grundsätzlich die Entscheidungen und leitet das Projekt.

- Senior Scientist: Die Mitarbeiterin hat bereits in diversen Projekten mitgearbeitet und ist mit Hörversuchen grundsätzlich vertraut. Sie ist zumeist in der Lage, Daten aus einem Pilottest richtig zu interpretieren und geeignete Adaptionsmaßnahmen vorzuschlagen. Unterstützend ist sie mit der Durchführung und Mitarbeit im Projekt betraut.

- Junior Scientist: Der Forscher verfügt über wenig oder gar keine Erfahrung auf dem Gebiet und arbeitet im Projekt mit. Er sammelt Erfahrung in der Versuchsdurchführung und assistiert in verschiedenen Bereichen des Ablaufs.

- Hersteller/Kunde: Zur vereinfachten Darstellung werden Hersteller und Kunde zu einem Verantwortlichen zusammengefasst. Die Position kennzeichnet die Verknüpfung zu einem Vertrags- oder Kooperationspartner.

Zu Versuchsbeginn stellt der Kunde relevante Informationen das zu beurteilende Produkt betreffend zur Verfügung. Dem Kundenwunsch entsprechend wird eine für den Versuch evaluierbare Fragestellung und auf dieser aufbauend eine Hypothese gebildet. Nach dem Auswahlprozess der zu testenden Attribute, welcher bereits protokolliert werden soll, wird eine geeignete Versuchsmethode gewählt, die den Versuchsaufbau und –ablauf, sowie Besonderheiten in der Generierung der Stimuli enthält. Dem Anwendungszweck entsprechend werden die Versuchspersonen ausgewählt und mit einer kleinen Anzahl dieser wird ein erster Pilottest durchgeführt. Liefert der Vorversuch zufriedenstellende Daten, kann mit dem eigentlichen Hörtest fortgefahren werden. Andernfalls sind Adaptionsmaßnahmen vorzunehmen welche üblicherweise erneut in einem Pilottest erprobt werden. Der Hörversuch liefert eine Datenmenge, welche anschließend mit statistischen Mitteln beschrieben und analysiert wird. Die Ergebnisse der Statistik werden gemeinsam mit der Beschreibung des Projektablaufs in einem Bericht erfasst und dieser nach eventuellen Korrekturen dem Kunden übergeben.

Hier wird lediglich der Fall betrachtet, dass ein Hörversuch aufgrund seiner erwarteten, sinnvollen Ergebnisse zu neuen Erkenntnissen führen kann. Da von einem Wissenszuwachs nicht zwangsläufig ausgegangen werden kann, obliegt es dem Versuchsleiter vor Beginn der Planungsphase und Durchführung eingehend zu prüfen, ob die Durchführung eines Hörversuchs sinnvoll ist und unter welchen Voraussetzungen. Die Überlegungen schließen eine Literaturrecherche, gute Kenntnis des zu untersuchenden Systems (Messobjekt) und statistische Überlegungen, beispielsweise die Anzahl an benötigten Versuchspersonen um ein bestimmtes Signifikanzniveau zu erreichen, mit ein. Gerade bei geringen, hochsensitiven Aufgabenstellungen an die Versuchspersonen gewinnen diese Überlegungen an Bedeutung in Bezug auf die Realisierbarkeit und den Aufwand. Um der Vollständigkeit Genüge zu tun wird auch darauf hingewiesen, dass es umgekehrt zu Fragestellungen kommen kann deren offensichtliche perzeptive Unterschiede den Aufwand eines Hörversuchs nicht rechtfertigen.

In der Bestimmung von Qualität (vgl. Abs. 2.1) ist der Mensch als bewertender Faktor maßgeblich beteiligt. Um Qualität von Signalen im hörbaren Frequenzbereich messbar oder erfassbar zu machen, werden Hörversuche durchgeführt, die entstandenen subjektiven Ergebnisse analysiert und über Rückschlüsse auf die Grundgesamtheit objektiviert und verallgemeinert.

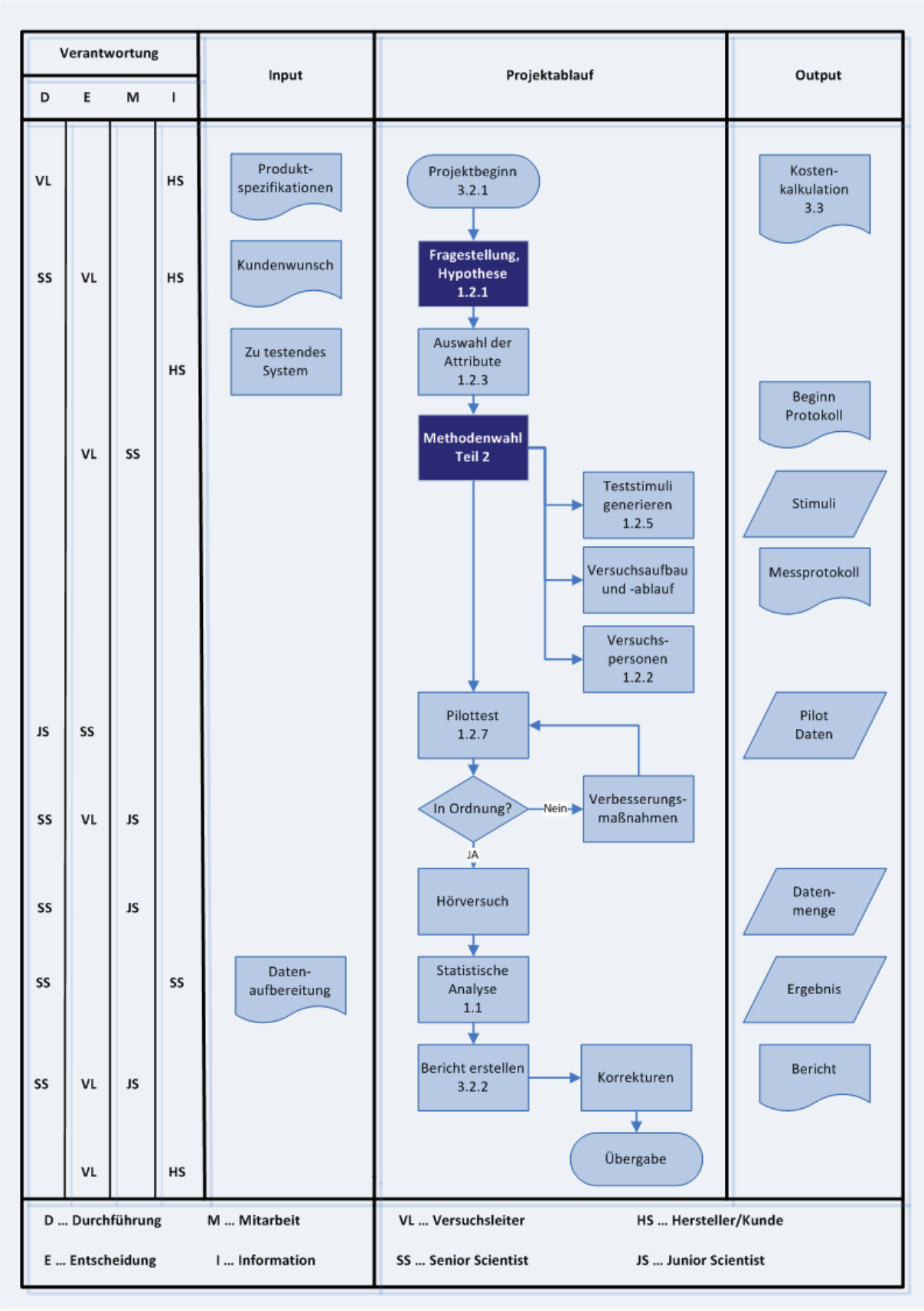

Abb. 7: Ablauf eines Projekts mit psychoakustischem Versuch

Nach Osgood hat eine zufriedenstellende Messmethode, hier bezogen auf einen Hörversuch, folgende sechs Charakteristika [Cot11, Kap2]:

- Objektivität: die Empfindung ist von anderen Hörerinnen reproduzierbar (vgl. Abs. 1.1.2.2)
- Reliabilität (Zuverlässigkeit, reliability): die Empfindung ist von demselben Hörer reproduzierbar (vgl. Abs. 1.1.2.2). Es wird also der Aspekt betrachtet, wie genau die zu beurteilende Eigenschaft bewertet wird. Reliabilität gilt als Voraussetzung für Validität.
- Validität: Die Methode liefert zuverlässige Ergebnisse, sie ist demnach reliabel, und die Ergebnisse lassen sich verallgemeinern. Es kann demnach eine Schlussfolgerung gebildet werden.
- Sensitivität: Die Diskriminierungsfähigkeit der Versuchspersonen stimmt mit der Empfindlichkeit der Methode überein.
- Komparabilität: Vergleichbarkeit der Methode und ihrer Ergebnisse mit anderen Versuchsgruppen
- Brauchbarkeit: die Ergebnisse haben einen sinnvollen Anwendungsbereich

Neben diesen Aspekten zu den Messmethoden ist eine gute Kenntnis des Messobjekts (im Weiteren wird allgemein der Begriff System verwendet) eine Grundvoraussetzung für die richtige Wahl einer geeigneten Methode. Informationen über das zu testende System können vom Hersteller bereitgestellt werden und sind über eine Literaturrecherche oder fachkundige Forscherinnen zu ergänzen. Im weiteren Verlauf der Arbeit wird zudem davon ausgegangen, dass psychoakustische Versuche von Versuchsleitern mit entsprechenden Kenntnissen auf dem Gebiet der Akustik und Psychoakustik, sowie anderen relevanten Disziplinen geplant werden und diese ihre Mitarbeiter in der Durchführung betreuen.

1.2.1 Fragestellung und Hypothese

Ohne zu wissen, wonach man sucht, ist es schwer, eine zufriedenstellende Antwort zu finden. Oder nach Möller [Moe10, Kap2]: *„Vor der Planung eines Versuchs sollte zunächst das **Ziel der Messung** genau festgelegt werden. Dabei reicht es nicht aus, dass man allgemein die „Qualität eines Systems messen möchte“*.“ Die Fragestellung bildet daher den zentralen Ausgangspunkt eines Experiments, auf der die Methodik, Datengenerierung und Datenanalyse aufsetzen. Der Ablauf sieht demnach vor, herauszufinden, was die Intention des Versuchs sein soll (z.B. durch Kundenwunsch repräsentiert), auf dieser aufbauend eine Hypothese zu formulieren (vgl. Abs. 1.2.1) und mit dieser Grundlage die Fragestellung auszuarbeiten. Es sei erwähnt, dass ein wesentlicher Unterschied zwischen der Hypothese, die im Allgemeinen den Ausgangspunkt für das Versuchsdesign darstellt (obgleich sie im Hintergrund steht), und der Fragestellung an den Versuchsteilnehmer, die auf der Hypothese aufsetzt, besteht.

Beispiel: Ein System soll ohne Vergleichswerte (kein Referenzstimulus oder Vergleichssystem) hinsichtlich seiner Gesamtqualität absolut beurteilt werden. Für die Generierung der Teststimuli ist bedeutend, welche Variable variiert wird, beispielsweise das Artefakt Rauschen, oder die Lautstärke. Die Hypothese richtet sich nach dem Grad der Variation der Variable bzw. des Artefakts. Die Fragestellung an die Versuchsperson hingegen bezieht sich auf den qualitativen Gesamteindruck, beispielsweise <Wie wird die Gesamtqualität des Systems beurteilt?>, oder <Beurteilen Sie die Gesamtqualität der nachfolgenden Stimuli>. Dementsprechend würden die Versuchspersonen dazu aufgefordert, auf der 5-Kategorienskala zu bewerten, wie Sie die Gesamtqualität nachfolgender Stimuli beurteilen würden. (vgl. Abs. 2.2.1.1., ACR)

In der Wissenschaftstheorie beschäftigt man sich neben Zielen und Voraussetzungen von Wissenschaft auch mit der Frage nach geeigneten Ansätzen um zu neuer wissenschaftlicher Erkenntnis zu gelangen (deswegen auch die Parallelen zur Epistemologie, Erkenntnistheorie). Dabei stehen einander Empirismus und Rationalismus - oder anders formuliert Induktion und Deduktion - als zwei grundlegende Prinzipien gegenüber.

Empirismus als induktives Prinzip schließt von der Einzelbeobachtung oder einem Experiment auf das Allgemeine. Das Problem dabei ist, dass auch aus vielen Beobachtungen nicht zwangsläufig die Verallgemeinerung für gültig erwiesen werden kann. *Beispiel: Person A hört gerne Volksmusik. Person B hört gerne Volksmusik. Person C hört gerne Volksmusik. Schlussfolgerung: Alle Menschen hören gerne Volksmusik. Es kann aus diesem Experiment genauso nicht der Schluss gezogen werden, dass alle Menschen gerne Musik hören*. Zudem fokussiert der Empirismus auf Gemeinsamkeiten von Eigenschaften und nicht auf deren Unterschied.

Der Rationalismus als deduktives Prinzip geht von der Allgemeingültigkeit[6] aus und grenzt diese theoretisch bis auf die Einzelbeobachtung ein. Diese Aussage oder Beobachtung wird dann als Hypothese innerhalb eines Experiments auf ihre Gültigkeit hin überprüft. Der deduktive Ansatz von der Theorie zur Empirie soll innerhalb dieser Arbeit als Ausgangspunkt für das nachfolgende Experiment weiter verfolgt werden. Es wird aber darauf hingewiesen, dass innerhalb der statistischen Analyse durchaus weiterführende Fragestellungen (nach dem induktiven Prinzip) beantwortet werden können (vgl. Abs. 1.1.2) und somit beide Prinzipien innerhalb eines Versuchs kombiniert werden können.

Die Formulierung der Nullhypothese bestimmt die Art ihrer Überprüfung, sowohl in der Wahl der geeigneten Methode zur Durchführung des Versuchs, als auch in der anschließenden inferenzstatistischen Analyse. Nach Bortz [Bor04, Kap4] wird grundsätzlich zwischen zwei Arten von Hypothesen und den damit einhergehenden inferenzstatistischen Betrachtungen unterschieden:

- Unterschiedshypothese: Diese Hypothesen werden über Häufigkeitsverteilungen oder Vergleich der Mittelwerte verifiziert und sie postulieren einen Unterschied zweier Variablen bzw. Systeme.

[6] Ausgangspunkt des deduktiven Prinzips ist eine nomologische Hypothese.

- Ungerichtet: Es gibt einen Unterschied in den abhängigen Variablen, aber die Richtung ist unbekannt. z.B.: System A ist besser oder schlechter als System B. Daraus folgt für die H_0: Es gibt keinen Qualitätsunterschied zwischen den Systemen A und B.

$$H_0 : \mu_0 = \mu_1 \quad H_1 : \mu_1 \neq \mu_0 \tag{59}$$

Ungerichtete Hypothesen werden mit zweiseitigen Verfahren überprüft.

- Gerichtet: Die Tendenz des Unterschieds wird in der Formulierung zusätzlich bekannt gegeben. z.B.: *System A ist <u>besser als</u> System B*.

$$\begin{aligned} H_0 : \mu_0 > \mu_1 \quad H_1 : \mu_1 \leq \mu_0 \\ H_0 : \mu_0 < \mu_1 \quad H_1 : \mu_1 \geq \mu_0 \end{aligned} \tag{60}$$

Gerichtete Hypothesen werden mit einseitigen Testverfahren geprüft.

- Zusammenhangshypothese: Diese Hypothesen beschreiben einen Aspekt, den zwei oder mehrere Variablen miteinander gemein haben und sie werden mit Korrelationsrechnungen überprüft.
  - Ungerichtet: *Bsp.: Die Lautstärke beeinflusst den Hörkomfort des Kopfhörers.*
  - Gerichtet: *Bsp.: Der Hörkomfort des Kopfhörers sinkt mit steigender Lautstärke.* Der Einfluss der unabhängigen (*Lautstärke*) auf die abhängige Variable (*Hörkomfort*) wird bereits in der Fragestellung postuliert.

Die gerichtete Hypothese erfordert zu Beginn des Versuchs ein höheres Maß an Vorkenntnissen, liefert jedoch in der Überprüfung gleichzeitig auch mehr Gehalt an Information.

Formulierung der Hypothese zu Versuchsbeginn

Ausgehend von der nomologischen Hypothese (gesetzgebend, allgemeingültig) gelangt man durch schrittweise Deduktion (Herabsetzen des Gültigkeitsbereichs) zu Beobachtungen oder Prämissen. Aus diesen Prämissen (Explanans) ergibt sich eine logische Schlussfolgerung (Explanandum), die innerhalb eines Experiments geprüft werden kann.

- Prämisse 1 (Nullhypothese H_0): ist z.B. eine probabilistische Hypothese
- Prämisse 2 (initial condition): beschreibt sämtliche Rahmenbedingungen oder Voraussetzungen innerhalb eines Setups
- Schlussfolgerung (testable statement)

Die Nullhypothese H_0 (Prämisse 1) soll so formuliert werden, dass sie verworfen werden muss, wenn die Schlussfolgerung falsch ist und die Prämisse 2 gültig ist. (=Prinzip der indirekten Deduktion, Falsifikation). Andernfalls wenn das Explanans gültig ist, also beide Prämissen stimmen, muss auch die Schlussfolgerung richtig sein.

Eine theoretische Aussage, die keine möglichen Falsifikatoren aufweist (also immer wahr ist), nennt man tautologisch oder logisch inkonsistent. Bei der Findung einer theoretischen Aussage sucht man jedoch nach einer Formulierung, die aufgrund möglichst vieler potentieller Falsifikatoren den Informationsgehalt der wissenschaftlichen Frage maximiert und somit in sich auch logisch konsistent ist. Um eine theoretische Aussage nun auch logisch konsistent und unmissverständlich zu formulieren, ist die Präzision der Begriffe bzw. deren möglichst eindeutige Erklärung relevant [Bor04, Kap0]. Die Hypothese bildet nun gemeinsam mit der initial condition und der deduktiven Argumentation die prüfbare Aussage (testable statement).

Der initial condition kommt höchste Bedeutung zu, da sie die Qualität des Experiments über die Definition der Rahmenbedingungen festlegt [Jac13],[Bec06, Kap2]. Die initial conditions entstehen gemeinsam mit dem treatment design (vgl. Abs. 1.2.6.1).

Beispiel:

H_0: Beide Active Noise Cancellation Algorithmen (ANC) führen im selben Ausmaß zu einer Verbesserung in der Bewertung der Sprachqualität.

Initial condition: Ein analoger und ein digitaler ANC werden im Rahmen eines Hörversuchs über Kopfhörer hinsichtlich ihrer Verbesserungen der Sprachverständlichkeit bei babble noise gegenüber der Situation ohne Noise Cancellation miteinander verglichen. Die Ergebnisse des Versuchs sind wahre Repräsentative für die wahrgenommene Sprachverständlichkeit beider ANCs. …(man kann hier noch den Raum, die Methode, etc. anführen)

Schlussfolgerung: Die wahrgenommene Sprachverständlichkeit wird mit beiden ANCs zu verbesserten/denselben Ergebnissen führen.

Die H_0 ist so formuliert, dass sie verworfen werden muss, da man davon ausgehen kann, dass zum derzeitigen Standpunkt der Entwicklung der analoge ANC nach wie vor besser ist als der ihm nachempfundene digitale oder umgekehrt.

Ziel des Hörversuchs kann nun sein: nachzuweisen, dass der analoge ANC tatsächlich besser ist (als besser empfunden wird) als der digitale und in welchem Maß (oder umgekehrt, die Richtung ist nur entscheidend, wenn sie bekannt ist). Zusätzlich können weitere Fragestellungen postuliert werden, welche die Teilaspekte von Sprache hinsichtlich ihrer Qualitätsbewertung genauer betrachten, z.B. Betrachtung von Verstehbarkeit von Plosivlauten etc.

Dabei kommt das Prinzip der verfeinerten Falsifikation zum Einsatz, welches es dem Versuchsleiter erlaubt, innerhalb der Pilotphase die Variablen zu adaptieren ohne sogleich die H_0 verwerfen zu müssen. [Jac13]

1.2.2 Versuchspersonen

Die Expertise von Versuchsteilnehmerinnen im auditiven Bereich bestimmt maßgeblich die Art der an sie gerichteten Fragestellung, den damit verbundenen Informationsgehalt und damit auch den Einsatzbereich bei Hörversuchen. Nachfolgende Begriffsdefinitionen innerhalb dieses Kapitels entstammen der Norm für das Vokabular zur sensorischen Analyse [5492].

„Prüfperson (sensory assessor): jede Person, die an einer sensorischen Prüfung teilnimmt. Anmerkung 1: Ein Laie ist eine Person, die kein bestimmtes Kriterium erfüllt. Anmerkung 2: Eine eingeführte Prüfperson hat bereits an einer sensorischen Prüfung teilgenommen.“ Die Prüfpersonen werden innerhalb dieses Dokuments Versuchspersonen (Vpn), Versuchsteilnehmer, Hörerinnen genannt.

„Panel, Prüfpersonengruppe: Gruppe von Prüfpersonen, die an einer sensorischen Prüfung teilnehmen.“

Im weiteren Verlauf werden lediglich Experten als Panel zusammengefasst, nämlich als Expert Listening Panel (*ELP*). Die Definition von naiven Hörern oder auch erfahrenen Hörern als Panel erscheint mir nicht sinnvoll, da sich die Zusammensetzung naiver Hörer von Versuch zu Versuch zu ändern hat. Demnach besteht nach meiner Meinung auch ein Unterschied zwischen einer Gruppe an Prüfpersonen und einem Panel bezüglich ihrer zeitlichen Konstanz.

1.2.2.1 Naiver Hörer

Der Begriff naiv wurde aus dem Englischen „naive assessor“ (für Laie) übernommen, da er die Art des Hörens ohne jegliche Erfahrung und Vorkenntnis, auf die es eben bei der Auswahl von Versuchspersonen ankommt, im Gegensatz zum Laien metaphorisch beschreibt.

Naive Hörer werden für Versuche ausgewählt, in denen ein Signal auf seinen Gesamteindruck hin getestet wird. Die Personengruppe soll Menschen aus dem Alltag möglichst realistisch repräsentieren. Dabei ist die selektive Wahrnehmung einzelner Attribute oft nicht gewünscht, ein Verständnis der zugrundeliegenden Signalgenerierung und -verarbeitung aufgrund von Expertise ebenso wenig. Dementsprechend erfolgt die Beschreibung des Antwortformats durchaus auch anhand von wertenden Attributen.

Verwendung und Einsatzbereich

Der naive Hörer hat der Theorie nach noch nie an einem Hörversuch teilgenommen. Er ist demnach auditiv und experimentell unbelastet und soll somit die Allgemeinheit repräsentieren. Hören ist diesem Kontext nach als Mittel zum Zweck zu verstehen und nicht als Passion oder grundlegende Berufung. Dieser Umstand ist besonders bei Bewertungen der Sprachqualität oftmals gewünscht. In der Praxis wird ein Hörer nach einer Pause von sechs Monaten wieder als naiv eingestuft, was allerdings nicht demselben Maß an Naivität

gleichkommt, da Rahmenbedingungen wie Räumlichkeiten, Ablauf, Software bereits bekannt sind und damit das Verhalten der Hörerin ein anderes ist.

Sobald jedoch irgendeine Form von Training vor dem Versuch, diese ist nicht zu verwechseln mit der Eingewöhnungsphase (vgl. Abs. 1.2.8.5, Versuchsablauf), verlangt oder notwendig wird, kann es sich bei der Personengruppe per Definition streng genommen nicht mehr um naive Hörer handeln. Bezüglich der Auswahl der Population liefert die Literatur kontradiktorische Aussagen. So schließt Quackenbush beispielsweise nicht aus, dass unter genauer Auswahl der Hörerinnen und Achtsamkeit bezüglich der Repräsentation der Population welche das Kommunikationssystem verwendet, durchaus repräsentativere Ergebnisse erzielt werden können, indem der Judgement Bias verringert wird [Qua88, Kap2.1].

Es wird angemerkt, dass im methodischen Teil 2 bei den Verfahren, welchen die ITU-T P.800 zugrundeliegt, naive Hörer empfohlen werden. Diese Entscheidung der ITU ist jedoch zweckgebunden, das bedeutet daher keinesfalls, dass beispielsweise für den Audiobereich, in dem ein sensitiver Umgang mit dem Stimulus gewünscht wird, nicht trotzdem Experten als Versuchspersonen in Kombination mit der beschriebenen Methode verwendet werden dürfen. Streng genommen, führt dies lediglich zu der Konsequenz, dass das Verfahren dann nicht mehr standardisiert genannt werden darf. Dabei wird keine Aussage darüber gemacht, dass diese Art der Durchführung zu keinen zuverlässigen Ergebnissen führt. Weitere Vergleiche zwischen naiven Hörern und Experten wie jene von Frank würden in diesem Bereich sicherlich Aufschluss geben [Fra12].

1.2.2.2 Experten

„Sensoriker (*expert sensory assessor*): Subst. ausgewählter Prüfer mit nachgewiesener sensorischer Empfindlichkeit und mit umfassender Schulung und Erfahrung hinsichtlich der sensorischen Prüfung, der in der Lage ist, verschiedene Prüfmaterialien widerspruchsfrei und wiederholbar sensorisch zu beurteilen.“

Anhand der Definition von Antwortattributen (vgl. Abs. 1.2.3.1) kann der sinnvolle Einsatz eines Expert Listening Panels (*ELP*) im Audiobereich gut veranschaulicht werden. Ein Listening Panel durchläuft nach ITU-R-BS.1116-1 einen dreistufigen Auswahlprozess, der die Eignung der Versuchsperson auf folgende Eigenschaften hin prüft [1116-1]:

- Fragebogen: allgemeine Fragen zur Person, (Alter, Geschlecht, Muttersprache, musikalische Interessen, Vorbildung,…) die deren Eignung als potentielle Kandidatin prüfen
- Sprachgewandtheit
- Auditorische Fähigkeit

Ist der Auswahlprozess abgeschlossen, erarbeitet das Panel gemeinsam eine Attributliste, die dann als Grundlage für die Auswahl der zu testenden Eigenschaften in einem Hörversuch dient und bei Bedarf erweitert werden kann. Der Auswahlprozess von Attributen ist als

kontinuierlicher Prozess anzusehen, gemeinsam mit dem regelmäßigen Training [Bec06, Kap4].

Trainingseinheit versus Trainingsphase (familiarization phase) direkt vor dem Versuch

Experten (*expert listener*, *EL*) haben den Vorteil, aufgrund ihrer regelmäßigen Übungseinheiten und auch wegen ihres guten Hörvermögens zu konsistenten und reliablen Urteilen in Hörversuchen zu kommen. Zusätzlich zur Sensitivitätserhaltung kann ein Training mit speziellem Augenmerk auf einen durchzuführenden Hörversuch designt werden. Der Zweck liegt in der zusätzlichen Sensibilisierung auf speziell zu testende, kleine Artefakte in Signalen, ganz abgesehen von der Verankerung mit den zu prüfenden Attributen der Signale.

Eine Trainingseinheit stellt der Hörerin eine Liste möglicher Stimuli zur Verfügung, welche in einem Zeitraum von mehreren Stunden beliebig häufig gehört und somit eingeprägt werden kann. Vorteil dieser Einheit ist, dass der EL sich die Stimuli beliebig oft und zu dem von ihm präferierten Zeitpunkt anhören kann. Dieser Benefit ist allerdings mit zusätzlichen Kosten an das Panel verbunden.

Als Alternative können auch kurze Trainingsphasen für ausreichend erachtet werden, die direkt vor dem jeweiligen Versuch stattfinden. Diese Variante wird bevorzugt zum Einsatz kommen, wenn der Versuchsleiter ähnliche Artefakte zum wiederholten Mal abfragt. Nachteilig wirkt sich jedenfalls die zusätzliche Dauer der wenn auch kurzen Trainingsphase auf Ermüdungseffekte aus.

Verwendung und Einsatzbereich

Im Audiobereich (vgl. Abs. 2.3) werden nahezu ausschließlich Expertinnen als Versuchspersonen verwendet. Ein bestehendes ELP ermöglicht in diesem Fall gezieltes Training der Versuchspersonen, sofern für den jeweiligen Anwendungsbereich erwünscht, bei gleichzeitigem Screening der Teilnehmerinnen. Darüber hinaus ist mit einer geringeren Anzahl an Vpn eine effiziente Versuchsdurchführung und auch Auswertung der Daten möglich. Es ist aber anzumerken, dass der Einsatz eines ELP bezüglich der Verallgemeinerung der zugrundeliegenden Grundgesamtheit (Population) kritisch zu hinterfragen ist.

1.2.3 Quantifizieren von Eindrücken; die Antwort

Abhängig von der Art der Fragestellung ist es wichtig, im Antwortverhalten der Versuchsperson zwei wesentliche Punkte zu beachten; die Definition des Antwortattributs, sowie die Definition des Antwortformats. Das Antwortattribut beschreibt in Hinblick auf die Fragestellung gezielt die zu beurteilende Eigenschaft des untersuchten Stimulus. Das Antwortformat legt die Art und Weise (Rahmenbedingung) der Beurteilung fest. Beide Aspekte tragen wesentlich zur Objektivierbarkeit der abhängigen Variable, der Antwort der Versuchsperson, bei.

Das Attribut, welches im Hörtest abgefragt wird, ist nicht zwangsläufig bekannt und kann über direkte oder indirekte Auswahlverfahren (direct and indirect elicitation methods) ermittelt werden. Dabei wird neben der Attributbestimmung zumeist auch ein dichotomes Eigenschaftspaar ermittelt, das in weiterer Folge für die Endpunkte der Skala als Beschriftung dient. Diese Verfahren stellen einen mehrstufigen, komplexen und zeitintensiven Prozess dar, weshalb es ratsam ist, Attributlisten für den speziellen Anwendungsfall zuvor zu recherchieren, mit der Expertengruppe zu validieren und einzustudieren.

1.2.3.1 Definition des Antwortattributs

Experimente können unterschiedliche Ziele verfolgen. Ein Ziel kann sein, ein spezifisches Attribut wie Lautstärke, Tonhöhe, etc. zu testen. Es kann aber auch interessant sein, den Gesamteindruck der Versuchsperson auf ein Signal hin, beispielsweise Akzeptanz, Annehmlichkeit, Belästigung, zu erfassen. Der Gesamteindruck kann aber auch lediglich mit hörbar/ nicht hörbar oder auch gut/ schlecht bewertet werden.

Die Art des zu untersuchenden Stimulus spielt bei der Auswahl von einer geeigneten Versuchsmethode und dementsprechend auch bei der Attributzuschreibung eine große Rolle. Ein Sinus beispielsweise wird auf ein spezifisches Attribut hin relativ einfach zu untersuchen sein und der Versuchsperson wird auch schnell eingängig sein, welche Eigenschaft getestet werden soll. Der zu testende Stimulus ist dann allerdings synthetisch und es stellt sich die Frage, in wie weit dies für die Hypothesenformulierung gewünscht ist.

Bei komplexeren Stimuli wie Musik oder Umgebungslärm hingegen wird es schwieriger sein, ein einzelnes, spezifisches Attribut zu extrahieren. Verschiedene Attribute werden angeregt wie Signaldauer, Tonhöhen, Timbre, Lautstärke und der Versuchsleiter hat die Aufgabe, das zu testende Attribut möglichst exakt zu definieren und abzufragen. Der Stimulus ist dementsprechend komplexer in der Handhabung, allerdings liefert er auch die realeren Testbedingungen [Bec06, Kap4].

Filtermodell nach Pederson & Fog

Das Filtermodell beschreibt die Sinneswahrnehmung eines physikalischen Stimulus über zwei menschliche Filter und die sich dadurch ergebende Unterteilung von Messungen der Wahrnehmung. Der erste Filter – die Sinneswahrnehmung – ist gekennzeichnet durch Sensitivität und Selektivität und lässt ein auditorisches Ereignis im Gedächtnis des Hörers entstehen. Dieses Ereignis, bestehend aus mehreren spezifischen Attributen, hat die analytische Messung jener Attribute zur Folge. Die Attributsdefinition und -erkennung nennt man auch perzeptive Messung. Der zweite Filter – die kognitiven Faktoren – steht für Emotion, Stimmungslage, Erwartung, Kontext und macht der Versuchsperson gemeinsam mit den relevanten Attributen die Entscheidung nach dem bevorzugten Stimulus möglich. Die Messung jenes Gesamteindrucks wird affective measurement genannt, nach [Bec06, Kap4].

Perceptive Measurement

Expert listeners haben neben der auditorischen Sensitivität den Vorteil, dass sie gemeinsam den Auswahlprozess verschiedener Attribute durchlaufen und sich somit einen einheitlichen Wortschatz zur Beschreibung von Hörwahrnehmungen aneignen.

Dieser Wortschatz wird in einer Attributliste konkretisiert und macht es möglich, nicht nur den Gesamteindruck eines Signals zu erfassen, sondern nach spezifischen, perzeptiven Merkmalen oder Eigenschaften des Klangs zu fragen. Dabei ist die perzeptive Messung in diesem Sinn nicht mit jener der analytischen Betrachtung aus der Psychophysik, die sich mit auditory events auseinandersetzt, zu verwechseln. Vielmehr ist hier die Ebene zwei des Layer Modells gemeint, genauer gesagt, die perzeptive Analyse auraler Objekte (vgl.Abb. 10), die bereits einen integrativeren Zugang zum Hörereignis aufzeigt, aber dabei dennoch ein selektives Wahrnehmen von Klang zur Bewertung darstellt. (Klang meint hier ganz allgemein Stimuli im hörbaren Frequenzbereich, also auch Sprache.)

Verschiedene Techniken für den gemeinsamen Vokabelwortschatz (*consensus vocabulary techniques*) können angewendet werden, um die Liste von zu verwendenden Attributen (*semantic differential*) zu erstellen und sie bei Bedarf in ihrem Anwendungsbereich zu erweitern. Somit wird es mit einem Expert Listening Panel praktisch überflüssig, vor jedem einzelnen Hörversuch den Wortschatz und damit auch die Attributeigenschaften genau zu erläutern. Dennoch kann der kreative Prozess der Diskussion über die Eignung konkreter Bezeichnungen, gerade wenn es um die Findung von Antonymen geht, durchaus effizient sein.

Einführende Erläuterungen zu den zu beurteilenden Eigenschaften erleichtern dennoch den Einstieg in den Versuch und sind in der Eingewöhnungsphase durchaus ratsam.

Für naive Hörer stehen zur Attributfindung die *individual vocabulary techniques* zur Verfügung, bei denen jedes Individuum für sich den geeigneten Wortschatz zur Beurteilung von Stimuli im Allgemeinen beschreibt. Diese Vorgehensweise hat den Vorteil, dass Diskussionen über die Eignung von Worten zur Attributbeschreibung innerhalb der Gruppe wegfallen und auch kein Training der Versuchspersonen notwendig ist, was zu geringerem Kosten- und Zeitaufwand führt. Die Gemeinsamkeiten innerhalb der Gruppe werden letztendlich über statistische Methoden wie Flash Profile [Lor05] oder Repetory Grid Technique [Ber99] ermittelt. Anzumerken ist, dass die Zeitersparnis bei der Attributfindung nicht im Verhältnis zum wesentlich höheren zeitlichen Aufwand bei der Versuchsdurchführung steht. Ein Hörversuch, der mit untrainierten Hörern durchgeführt wird, erfordert nämlich eine wesentlich höhere Teilnehmerzahl (im Vergleich zu einem Versuch, der mit EL durchgeführt wird) um statistisch signifikante Aussagen machen zu können. Damit steigen aber auch der Kostenaufwand und jener der nachgehenden Analyse.

Beide erwähnte Techniken, jene, die den gemeinsamen (Anwendung EL) und jene die den individuellen Sprachwortschatz (Anwendung naive Hörer) ermitteln, gehören zu den direkten Auswahlmethoden. Diese direct elicitation methods verfolgen den Ansatz, dass ein direkter Zusammenhang zwischen auditivem Reiz und dem Gesagten besteht. Andere Methoden sind jedoch der Auffassung, dass verbale Fähigkeiten sehr eng an die Versuchsperson geknüpft sind und versuchen daher, den wahrgenommen Reiz und die

Verbalisierung zu trennen. Zu diesen indirekten Auswahlmethoden zählen beispielsweise multi dimensional scaling (*MDS*) oder perceptual structure analysis (*PSA*) [Cho05]. Es soll darauf hingewiesen werden, dass mit MDS anfangs lediglich eruiert werden kann, wie viele verschiedene, relevante Attribute zur Beschreibung eines mehrdimensionalen Raums benötigt werden, demnach aber noch nicht bekannt ist, welche Attribute dies sind. Zur einführenden Information sei auf die Arbeit von Wickelmaier verwiesen [Wic03].

Affective Measurement

Um den Gesamteindruck eines Hörereignisses in einem Hörversuch zu erfassen, gibt es unterschiedliche Möglichkeiten.

- *Präferenztest*

  Die Versuchsperson reiht zwei oder auch mehrere Stimuli nach ihrer Präferenz. Typisches Beispiel ist der Paarvergleich (AB-Vergleich). Die Bewertung mehrerer Stimuli anhand einer Skala kann entweder eine Entscheidung in eine Richtung erzwingen (es gibt keinen neutralen Punkt – geradzahlige Skalen) oder aber auch eine neutrale Bewertung erlauben [Bec06, Kap4].

  Die Ergebnisse von Präferenztests sind relativ zu betrachten. Demnach kann zwar Produkt B besser bewertet werden als Produkt A, allerdings erhält der Versuchsleiter keine Information darüber, in welchem Ausmaß B besser ist als A und auch nicht darüber, ob diese Aussage allgemein gültig ist.

- *Akzeptanztest*

  Die Aufgabenstellung richtet sich beim Akzeptanztest nach dem Grad von Akzeptanz oder Bevorzugung eines Produkts. Dabei soll die Versuchsperson bewerten, in welchem Maß sie das Hörereignis präferiert oder ablehnt, beispielsweise auf einer mehrstufigen hedonischen Skala von „mag sehr“ bis „mag gar nicht“ oder „zu wenig“ bis „zu viel“. Die Art der Fragestellung spielt auch in diesem Fall eine zentrale Rolle.

- *Eignungstest*

  In diesem Fall wird das Ausmaß an Gefallen an einem Produkt an eine bestimmte Ausgangsposition, einen Umstand geknüpft, um somit das Hörereignis unter bestimmten Bedingungen evaluierbar zu machen.

Für Messungen des Gesamteindrucks wird an sich ein naiver Hörer einem Experten vorgezogen, da dieser nicht auf spezifische Eigenschaften eines Signals trainiert wurde und so den Gesamteindruck des Signals (mit all seinen Attributen) inklusive Emotion, Erwartung etc. besser widerspiegeln kann.

Es wird in der Literatur aber auch nicht dezidiert von der Verwendung von EL abgeraten [Bec06]. Nach Frank [Fra12] gibt es bei der Messung der Sprachverständlichkeit keinen signifikanten Unterschied zwischen den beiden Gruppen untrained und trained listeners. Diese Aussage legt unter Berücksichtigung der Definition nach ISO 8586:2012 von naive listeners und expert assessors die Vermutung nahe, dass EL unter bestimmten

Voraussetzungen durchaus in der Lage sind, den Gesamteindruck eines Normalhörenden getreu widerzuspiegeln. Insbesondere bei der Durchführung von Versuchen, die als Teststimulus mitunter Sprache verwenden und unter der Voraussetzung, dass der trainierten Versuchsperson der verwendete Sprachkorpus nicht bekannt ist (beziehungsweise sie diesen im vergangenen Jahr nicht gehört hat) sollte es keine signifikanten Unterschiede zwischen den Gruppen geben, sofern die EL auch nicht auf Sprachverständlichkeit als solches trainiert wurden (das ist noch nachzuweisen). Überdies schreibt die ITU für die Bewertung von Audiosystemen Experten vor, die dann auch affektive Attribute, wie Basic Audio Quality beurteilen [1116-1]. Beim Einsatz des Listening Panels ist es jedenfalls ratsam, dem EL klare Vorgaben bezüglich der Beurteilung des Teststimulus zu geben und diese, wenn möglich, auch durch ein Hörbeispiel zu verdeutlichen. Beispielsweise soll die Versuchsperson nicht beurteilen, ob das Klangbeispiel zu laut oder scharf ist, sondern, als in welchem Maß störend sie den Klang in seiner Gesamtheit empfindet (es kann ein Klang ja durchaus laut sein, dabei aber schön klingen und deswegen nicht als störend empfunden werden.). Jedenfalls wird ein Vergleich der beiden Gruppen innerhalb eines Pilottests, auch zum Zweck der statistischen Analyse, (und Sammlung von solchen Versuchsdaten) empfohlen, um Zweifel bezüglich der Eignung auszuräumen.

1.2.3.2 Definition des Antwortformats

Hat man das Attribut, welches getestet werden soll, durch perzeptive Messung oder Messung des Gesamteindrucks bestimmt, oder ist dieses per se bekannt, stellt sich die Frage, in welcher Weise diese Eigenschaft von der Versuchsperson bewertet werden soll. Als Bewertungsgrundlage bei Hörversuchen dienen dazu beispielsweise Skalierungsmethoden, auch als Antwortskalen bezeichnet, die nach direkten und indirekten Verfahren unterschieden werden. Dabei verschwimmt der Skalenbegriff durchaus, als der Antwortskala eine Skala im ursprünglichen Sinn (Statistik) zugrunde liegt, die mit Begriffen als verbale Zuordnung einer Wertung des zu untersuchenden Attributs versehen ist (vgl. Abs. 1.2.4).

Auf die Beschriftung ist bei der Wahl der Skalierungsmethode jedenfalls zu achten. Diese stellt eine potentielle Fehlerquelle dar, da bei perzeptiven Messungen die dichotomen Endpunkte der Attribute (welche zur Beschreibung des Antwortformats in der Skalenbeschriftung verwendet werden) beschreibenden, jedoch keinesfalls Präferenzcharakter (wertend, z.B. gut schlecht) aufweisen sollen, wie das bei Messungen des Gesamteindrucks wiederum durchaus der Fall ist [Bec06, Kap4].

Im Allgemeinen ist die Art der zu verwendenden Skalierung des Antwortformats im jeweiligen Standard verankert und soll nach Möglichkeit auch nicht weiter verändert werden. Der Grund hierfür liegt in der nicht trivialen Lösung der Erarbeitung einer eigenständigen Skalierungstechnik, die verbale Qualifikatoren auf ihre Äquidistanz und kulturellen Bias hin prüfen. Der interessierte Leser möge sich beispielsweise Rohrmann [Roh78][Roh07] genauer ansehen.

Auch die Thurstone-, die Likert- und die Guttmanskala beschreiben Verfahren zur Skalenfindung. Die Likertskala, welche gerne mit der 5-Kategorien-Skala verwechselt wird, die wiederum schon eine Skalierungstechnik für sich darstellt [Mut06, Kap2], kommt nach

dem auditiven Teil eines Hörversuchs für die Beantwortung qualitativer Aussagen zum Einsatz (vgl. Abs. 1.2.8.5, Versuchsablauf).

Die unterschiedlichen Skalierungsmethoden bzw. -techniken sind in Abs. 1.2.4.3 Antwortskalen/ Skalierungstechniken näher beschrieben. Weiterführende Informationen liefert zudem Preston [Pre00].

1.2.4 Skalierungsmethoden

Es ist darauf zu achten, dass ein wesentlicher Unterschied zwischen Messskalen und Skalierungstechniken/Antwortskalen besteht, die oftmals gleichermaßen als Skala bezeichnet werden. Ansätze zur Unterscheidbarkeit dieser beiden Begriffe werden deshalb nachfolgend erläutert.

1.2.4.1 Skalenbegriff nach ÖNORM EN ISO 5492

- „***Skala***, Subst.: Begriff, der sich entweder auf eine Antwortskala oder eine Messskala bezieht.“

- „***Messskala***: formale Beziehung (z.B. bei Verwendung einer Ordinal-, Intervall- oder Verhältnisskala) zwischen einer Eigenschaft (z.B. der Intensität einer sensorischen Wahrnehmung) und der zur Darstellung der Werte dieser Eigenschaft verwendeten Zahlen (z.B. Zahlen, die von den Prüfpersonen notiert bzw. festgehalten oder aus den Antworten der Prüfpersonen abgeleitet werden)

  Anmerkung 1: Der Begriff „Skala“ wird vielfach gleichbedeutend mit dem der „Messskala“ verwendet.“

- „***Antwortskala***: Hilfsmittel (z.B. numerisch, verbal oder piktographisch) mit dem eine Prüfperson eine qualitative Antwort notiert oder auf andere Weise festhält.

  Anmerkung 1: Im Falle sensorischer Analysen handelt es sich dabei um ein Gerät oder Werkzeug, mit dessen Hilfe die Reaktion einer Prüfperson auf eine oder mehrere Eigenschaften so erfasst werden kann, dass sie sich in Zahlen umwandeln lässt.

  Anmerkung 2: Der Begriff „Skala“ wird vielfach gleichbedeutend mit dem der „Antwortskala“ verwendet.“

1.2.4.2 Messskala

Messskalen werden in der Statistik (in der Wissenschaft im Allgemeinen) dazu verwendet, um dem Informationsgehalt und Skalenniveau entsprechende Operationen und Analysen von den erfassten Daten durchführen zu können. In diesem Zusammenhang dienen sie der quantitativen Erfassung der Bewertungen von sensorischen Ereignissen durch Versuchspersonen.

Voraussetzung für die Anwendung von statistischen Verfahren und Analysen ist das Vorhandensein von quantitativen Daten und Merkmalsausprägungen [Bor04, Kap1]. Dementsprechend wird für eine statistische Weiterverarbeitung von qualitativen Aussagen versucht, diese Aussagen zahlenmäßig zu erfassen. Merkmalsausprägungen und Ereignisse aus dem empirischen Relativ[7] - Antworten der Versuchspersonen - sollen homomorph in ein numerisches Relativ abgebildet werden.

Als Skalen bezeichnet man nach Bortz homomorphe Abbildungsfunktionen, die Ereignissen (Objekten aus dem empirischen Relativ) Zahlen (aus dem numerischen Relativ) zuordnen. [Bor04, Kap1]

Stanley Smith Stevens, der bereits 1938 das Sone zur subjektiven Bewertung von Lautheit definierte, definiert Skalen folgendermaßen: „Measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement.”[Ste46]

Skalenarten nach Stevens

Nach Stevens [Ste46] unterscheidet man nachfolgende vier verschiedene Skalenarten. Dabei kumuliert der Informationsgehalt hierarchisch mit den restriktiven Attributen der Skala:

- *Nominalskala:* es gibt keine Hierarchie, die Merkmalsausprägungen unterscheiden sich nur hinsichtlich eines Parameters.

- *Ordinalskala:* eine Reihung ist möglich, allerdings wird keine Aussage über die Intervallgröße zwischen zwei benachbarten Ausprägungen getroffen. Die Skala wird auch Rangskala genannt [Bor04, Kap1].

- *Intervallskala:* die Abstände zwischen benachbarten Ausprägungen sind konstant. Im Rahmen des Versuchdesigns ist dies die am häufigsten angewendete Skala, da die wichtigsten statistischen Verfahren bereits angewendet werden können und somit auch viel genauere Aussagen getroffen werden können, als bei den beiden vorherigen Skalen.

- *Verhältnisskala/Ratioskala:* es existiert ein echter Nullpunkt. Die Angabe von gemessenen Ereignissen als Verhältnis macht die Darstellung in der dB Skala möglich. Nach Bortz sind Verhältnisskalen aufgrund ihrer geringen humanwissenschaftlichen Bedeutung statistisch ebenfalls eher uninteressant. Sie werden daher und ob ihrer genaueren Messungen (als Intervallskalen) mit den Intervallskalen zusammengefasst und als metrische Skalen bezeichnet [Bor04, Kap1].

[7] Relativ = eine Menge von Objekten und Relationen (in dem Fall eben Ereignissen, Merkmalsausprägungen), die den Zusammenhang der Objekte untereinander beschreiben

Skalenniveau	Empirische Operationen	Numerische Operationen	Invariante Statistik
Nominal	Gleichheit/Unterschied →Datenklassifizierung	Zählen	Anz. von Ereignissen Modalwert Kontingenzkorrelation
Ordinal	Größer/Kleiner Mehr/Weniger	Zählen Ordnen	Median Perzentile
Intervall	Gleichheit von Intervallen oder Differenzen	Zählen Ordnen Differenz bilden	Mittelwert Standardabweichung Rangordnungskorrelation Korrelationskoeffizient
Ratio	Gleichheit von Verhältnissen	Zählen Ordnen Differenz bilden Quotient bilden	Variationskoeffizient

Tab. 7: Skalenarten und deren Möglichkeiten [Bor04]

Wie aus Tab. 7 ersichtlich steigt die Genauigkeit bzw. der Informationsgehalt mit dem Skalenniveau. Für das Versuchsdesign ist es folglich sinnvoll, die jeweils höchste in Frage kommende Stufe zu wählen. Ein Herabsetzen des Skalenniveaus ist im Nachhinein möglich (beispielsweise, wenn Intervalle tatsächlich nicht konstant wären, Herabsetzen auf die Ordinalskala), die umgekehrte Prozedur jedoch nicht (das würde einem Wissenszuwachs im Nachhinein gleichkommen, was schlichtweg nicht realisierbar ist).

1.2.4.3 Antwortskalen/ Skalierungstechniken

Es wird nach direkten und indirekten Skalierungstechniken unterschieden. Direkte Skalierung verlangt von der Versuchsperson eine Zuweisung des Reizes zu einem entsprechenden Wert auf der Skala infolge einer Bewertung und wird vermehrt für perzeptive Messung eingesetzt (bei der Bewertung von Sprachqualität auch zur Messung des Gesamteindrucks). Indirekte Verfahren erfassen die Sinneswahrnehmung einer Versuchsperson über ihre Fähigkeit, zwei Stimuli zu diskriminieren und finden bei der Messung des Gesamteindrucks Anwendung [Bec06, Kap4].

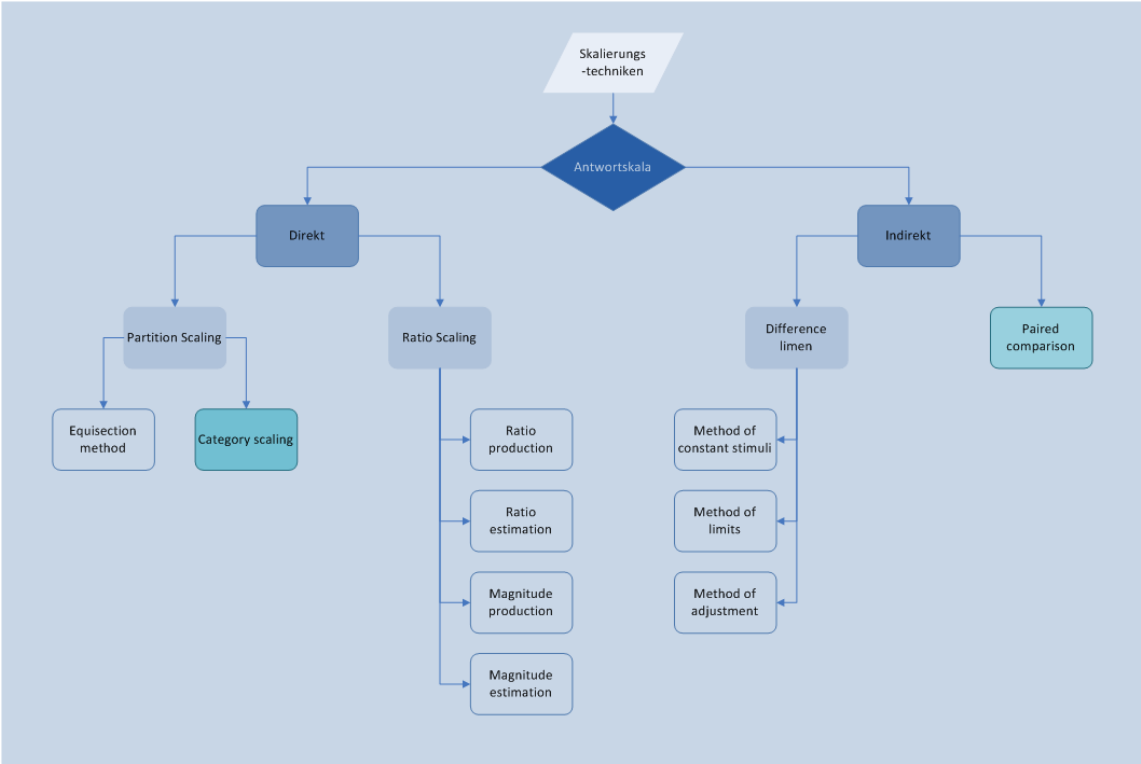

Abb. 8: Gliederung der verschiedenen Skalierungstechniken [Bec06]

Direkte Antwortskalierung

Nach Stevens werden die direkten Skalierungstechniken in Partition Scaling (Unterteilungsskalierung) und Ratio Scaling (Verhältnisskalierung) unterteilt, vgl. Abb. 8.

Die Kategorienzuweisung (Category Scaling) erfolgt dabei in diskrete oder kontinuierliche Stufen und ist im Rahmen der standardisierten Verfahren die am meisten verwendete Skalierungstechnik. Ab einer Unterteilung von 11 Stufen (damit sind nicht nur Kategorien, sondern auch etwaige Dezimalstellen als Feinabstufung der Skala gemeint) kann man die Skala als kontinuierlich betrachten und demnach auch statistische Mittel für intervallskalierte Daten anwenden [Bec06, Kap6].

Es ist wichtig anzumerken, dass um kulturellen Bias (vgl. Abs. 1.2.9 Bias) zu minimieren, es auch im Rahmen einer standardisierten Methode zulässig ist, auf die verbalen Deskriptoren zu verzichten, solange die Richtung der Wertigkeit der numerischen Skala eindeutig identifizierbar bleibt oder gekennzeichnet wird und eine Normalisierung der Ergebnisse erfolgt (vgl. (21)). Als Indikatoren können grafische Ankerpunkte wie -, +, !, verwendet werden, wie es von der EBU vorgeschlagen wird [Hoe97].

Indirekte Antwortskalierung

Die indirekten Skalierungsverfahren beruhen auf Thurstone und werden in Methoden zur Bestimmung der Unterschiedsschwelle und Grenzwerte (*difference limen, thresholds*) und in den Paarvergleich unterteilt, siehe auch Abb. 8. Bei diesen Verfahren erfolgt die Zuweisung des numerischen Relativs nicht direkt über die Beurteilung des Stimulus durch die

Versuchsperson auf einer Skala, sondern beispielsweise über Ja/Nein - Entscheidungen, die in iterativen Zyklen gemeinsam mit statistischen Analysen zu einem Ergebnis führen. Der Paarvergleich (vgl. Abs. 2.2.2.2, CCR) wiederum ermöglicht durch eine Vielanzahl von Vergleichen zweier Stimuli eine Reihung der dadurch entstehenden Bewertungen a posteriori. Die Analyse erfolgt beispielsweise über eine Dominanzmatrix nach dem Modell von Bradley-Terry-Luce [Col80], [Bra52], [Bra55], [Wic04] oder das Modell V von Thurstone [Bra07].

Der entscheidende Vorteil der indirekten Antwortverfahren liegt im Verzicht der Zuweisung einer Beurteilung durch die Versuchsperson auf einer Skala. Das direkte Mapping der perzeptiven auf die objektive Ebene wird also umgangen. Nachteilig an diesen Verfahren ist meist ein zusätzlicher Zeitaufwand.

Skalenniveau der Skalierungstechniken

Bis dato wurde noch keine dezidierte Aussage über das jeweilige Skalenniveau und den damit verbundenen Informationsgehalt getroffen (vgl. Abs. 1.2.4.2). Die folgende Abbildung soll einen Überblick über den Informationsgehalt der jeweiligen Methode und ihre Anwendungsmöglichkeiten geben.

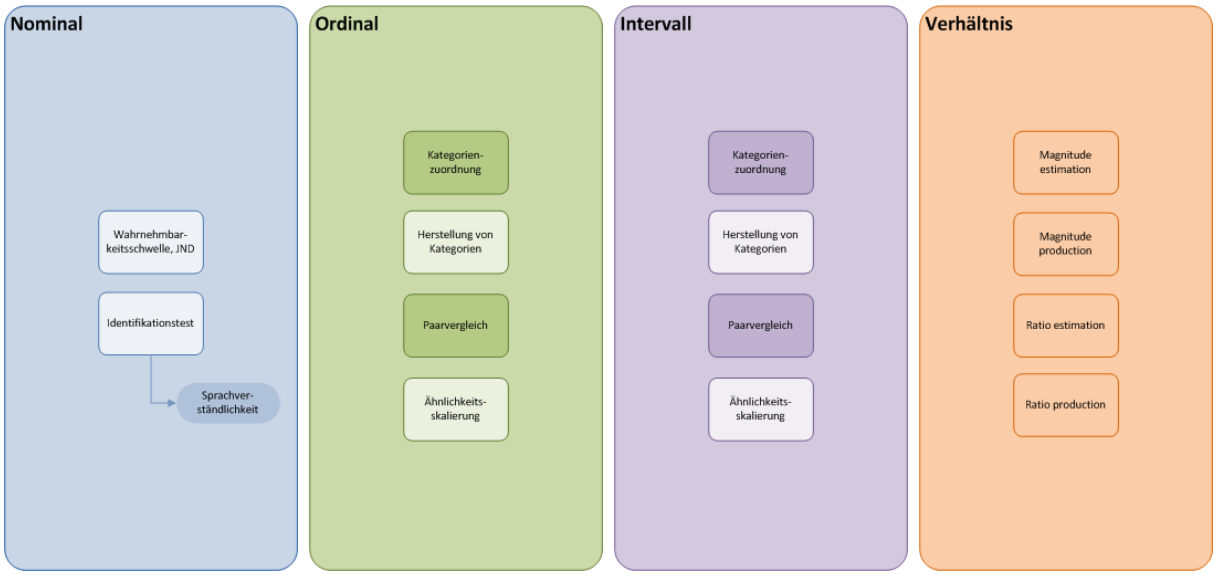

Abb. 9: Skalenniveau der Antwortskalen/Skalierungstechniken

Wie schon bei den Messskalen (vgl. Abs. 1.2.4.2) erwähnt, bestimmt das Skalenniveau den Informationsgehalt. Anders formuliert, legt das Niveau einer Skala die erlaubten statistischen Werkzeuge fest (vgl. Abs. 1.1.1.1, deskriptive Statistik). Ginge man für die Intervall- und Verhältnisskalen beispielsweise von quantitativen Daten und einer Normalverteilung aus, könnte man sich der Analyse durch die ANOVA bedienen. Wiederum wenn die Ergebnisse eines Experiments dichotome, also nominal skalierte Daten liefern würden, oder auch ordinal skaliert wären (wie es die meisten kategorialen Daten sind),

würde man sich anderer statistischer Werkzeuge bedienen. Jedenfalls ist die Verteilung der Daten immer auf eine Normalverteilung hin zu überprüfen, (vgl. Abs. 1.1.2.1, Stichprobe und Grundgesamtheit).

In Abb. 9 ist eine Übersicht der gängigen Methoden bzw. Skalierungstechniken und deren Skalenniveau gegeben. Die Kategorienzuweisung ist in diesem Fall sowohl der ordinalen als auch der Intervallebene zugeordnet. Das liegt daran, dass streng genommen bei kategorialen Daten nicht davon ausgegangen werden kann, dass die vordefinierten Abstände der Kategorien auch tatsächlich als konstante Abstände angesehen werden können (intervallskaliert). Ursache hierfür sind die verbalen Deskriptoren, deren Äquidistanz durch einfache Übersetzung noch nicht belegt ist. Jekosch [Jek10, Kap7.2] betrachtet die Anwendung von Verfahren aus der parametrischen Statistik als Analyse standardisierter, für intervallskaliert befundener Kategorienskalen ohne Prüfung der Varianz kritisch und liefert in Bezug auf verbale Deskriptoren weiterführende Literaturhinweise. Die ITU argumentiert aufgrund der langen Expertise auf dem Gebiet, dass man von quasi-intervallskalierten Daten ausgehen könne, solange keine großen Abweichungen von den Annahmen innerhalb der statistischen Analyse durch die ANOVA erkennbar sind [1116-1, 9].

1.2.5 Teststimuli

Ein beliebiges auditives Ereignis kann grundsätzlich als Teststimulus dienen. Dazu zählt man beispielsweise Rauschen, Musik, Sprache, aufgenommene Klangbeispiele, geräuschhafte Ereignisse, Sinustöne oder synthetische Klangereignisse.

Der sinnvolle Anwendungsbereich legt die zu verwendenden Testsignale zumeist fest. Anforderungen für den jeweiligen Anwendungsfall sind der einschlägigen Fachliteratur bzw. den der jeweiligen Methode zugrundeliegenden Standards und Papers zu entnehmen (vgl. Abs. 1.3, Standardisierung). Inputs bezüglich des richtigen Aufnahmemodus Mono oder Stereo liefert beispielsweise Toole [Too82].

Um dem Versuchsleiter jedoch den zumeist zeitintensiven und nicht immer trivialen Prozess der Auswahl geeigneten Materials zu erleichtern, sind der Arbeit Audiofiles in unkomprimiertem Format von der EBU beigefügt, die als Beispielstimuli für einige Anwendungsbereiche fungieren. Eine kurze Erläuterung zum Verwendungszweck sowie ein Datenblatt sind ebenfalls vorhanden.

Zur Testsignaldefinition können Signale vorerst in natürliche und synthetische Signale kategorisiert werden. Innerhalb der natürlichen Signale unterscheidet man weiter nach:

- Sprache,
- Audio,
- Störgeräusch (*noise*) und
- Geräuschen (z.B. das Geräusch eines Rasenmähers).

Synthetische Signale werden überwiegend zur Auswahl auditorischer Eigenschaften wie der Attributfindung verwendet, da ihre physikalischen Eigenschaften gut definierbar (messbar) sind [Bec06, Kap5].

Bei Beurteilungen der Sprachverständlichkeit (speech intelligibility) ist es möglich, einen bestehenden Korpus als Set von Testsignalen zu verwenden. Dies ist besonders bei den Wörter- und Satztests der Fall, ähnliches gilt für Bewertungen der Sprachverstehbarkeit (speech comprehensitivity).

Müssen die Teststimuli hingegen generiert, also aufgenommen, werden, gilt es, ein paar Dinge zu beachten (für Details sei auf die zugrundeliegenden Standards der jeweilige Methode in Teil 2 der Arbeit verwiesen; es werden allgemeine Punkte gelistet, Überlegungen, die aus anderen Quellen stammen, werden explizit angeführt):

- Rahmenbedingungen sollen während der Signalgenerierung dokumentiert und konstant gehalten werden. Dazu gehören der Aufnahmeraum mit Nachhallzeit, SNR, Aufnahmegeräte, Mikrofon, Abtastrate, etc. Der Prozess der Herstellung soll damit nachvollziehbar werden und ein entstandenes Stimuliset kann so bei Bedarf um weitere Signalproben erweitert werden.

- Testsignale sollen je nach Anwendungsbereich eine festgelegte Zeitdauer nicht überschreiten:

    - Zur Bestimmung der Gesamtqualität von Codern: Sprachmaterial (speech samples[8]) bestehend aus kurzen Sätzen mit einer Dauer von 2 - 3 s. (vgl. Abs. 2.2.1.1, Stimuli)

    - Sprachmaterial langer Zeitdauer: ein sample besteht aus vier verschiedenen Sätzen mit einer Dauer von 16 s. (die Extrema liegen bei 9-20 s aktiver Sprache im sample und gelten als Empfehlung (!), dürfen demnach unterschritten werden) Das sample soll eine realistische Verteilung der Degradierung des Signals durch das System darstellen (z.B. zeitvariante Entartungen wie packet loss im Internet, [Itu11]).

    - Sprachmaterial sehr langer Zeitdauer: ein sample dauert zwischen 45 s und 3 min. Das Sprachmaterial ist geeignet um Entartungseigenschaften sehr langer Zeitdauer realistisch darstellen zu können [Itu11, Kap5.1].

    - Sequenzen von maximal 20 s bei Vergleich von Audiosystemen mit mittleren/hohen Qualitätsverlusten (vgl. Abs. 2.3.2.3, MUSHRA)

    - Abschnitte von 10-25 s für den Vergleich von Audiosystemen mit geringen Qualitätseinbußen

[8] Sample: (engl. Muster, Probe) ein verarbeiteter Stimulus, der in einem Hörversuch beurteilt wird. Mehrere samples formen gemeinsam mit der Beurteilung ein trial.

- Zur Erzeugung von Sprachkorpora (Gesamtqualität) sollen je mindestens zwei Frauen und Männer verwendet werden, um Variabilitäten in der Bewertung aufgrund der Sprecherquelle zu minimieren.

  Bei Codern ist es auch möglich, den Einfluss der Sprecherstimme als unabhängige Variable und somit als Qualitätskriterium zu testen. In diesem Fall wird ein Korpus benötigt, der sich aus:

  - 8 Frauen
  - 8 Männern
  - 8 Kindern

    zusammensetzt [830, Abs.8].

Es ist durchaus sinnvoll, zu Beginn einen relativ großen Grundkorpus aufzunehmen und diesen dann als Basis für ACR-Tests zu verwenden, was eine gute Komparabilität aufgrund konstanter Kontrollparameter ermöglicht. Eine genaue Protokollierung der Rahmenbedingungen verringert zudem den Aufwand einer nachträglichen Erweiterung des Korpus durch zusätzliche Sprecher, wie beispielsweise Kinder, vgl. [800, AnnexB.1].

Bsp: Sprachkorpus Generierung zur Bestimmung der Gesamtqualität:

Verwendet werden dieselbe Anzahl an weiblichen und männlichen native speakern der gewünschten Sprachen. Diese können von Sprachinstituten oder Nachhilfeorganisationen rekrutiert werden und dürfen unterschiedlichen Alters sein. Wichtig ist ein Screening der Personen um zu überprüfen, ob diese gleichbleibend „monoton", jedoch mit korrekter Melodik und ohne Sprachfehler sich zu artikulieren in der Lage sind.

Als Sätze werden Passagen aus hochwertigen Zeitungen und anderen Zeitschriften gesucht, die jeweils eine Dauer von 2-3 s nicht überschreiten (so wird sowohl die häufige Bedingung für ACR erreicht, als auch die Möglichkeit offengehalten, Sprachmaterial längerer Dauer zu produzieren).

Als Aufnahmeraum steht der Multimediaraum nach ITU-R BS 1116-1 von JOANNEUM RESEARCH ebenso zur Verfügung, wie ein Kondensatormikrofon plus Verstärker mit niedrigem Eigenrauschen und geeignetes Aufnahmeequipment (Linearen Frequenzgang bis zur höchsten vom System zu übertragenden Frequenz gewährleisten!), das im Protokoll zum Zweck der Reproduzierbarkeit und Erweiterbarkeit des Korpus entsprechend spezifiziert wird (z.B.: A/D-Wandler: 16 bit, 2^{er} Komplement, fs=48 kHz). Der Abstand der Sprecherin zum Mikrofon beträgt 14 bis 20 cm und wird während der Aufnahmen möglichst konstant beibehalten.

Es können entweder mehrere Kanäle parallel verwendet werden um die Sprache als Schmalband-, Breitband- und Superbreitbandsignal aufzuzeichnen, oder aber das Sprachsignal wird mit voller Bandbreite bei fs=48 kHz aufgezeichnet und bei Bedarf auf die gewünschte Bandbreite reduziert.

Jedes Sample, welches aus (zwei bis) fünf Sätzen besteht, beginnt und endet mit einer halben Sekunde Ruhepause. Die Sätze werden jeweils mit einer halben Sekunde Pause aufgenommen. (Das Verkürzen von Sprachsamples ist in der Regel einfacher zu realisieren, als das Verlängern). Zusätzlich wird ein Kalibriersignal aufgezeichnet (z.B. 1 kHz Sinus, 30 s, -26 dBov.

Anschließend wird der active speech level beispielsweise nach dem Modell von ITU-T P56 [56] ermittelt. Jeder einzelne Satz soll mit dem active speech level bei -20 bis -30 dBov liegen.

Jedenfalls ist eine genaue Kenntnis der gewonnenen Sprachproben inklusive der zugehörigen Testbedingungen in Form eines Protokolls bei der Auswahl der Teststimuli und der anschließenden Interpretation der Daten enorm hilfreich [Itu11, Kap5.3].

Anmerkung: Die Generierung eines Sprachkorpus wird in diesem Fall sehr vereinfacht dargestellt. Auf linguistische, phonemische und phonetische Aspekte der Sprache wurde nicht weiter eingegangen. Diese Eigenschaften sind jedoch bei Bewertungen von Sprachverständlichkeit und zu einem bestimmten Anteil auch bei Bewertungen der Sprachverstehbarkeit von enormer Bedeutung. Damit steigen der Aufwand und die vorausgehende Kenntnis um das zu verwendende Material aber auch beträchtlich. Mit aus diesem Grund sind Testverfahren, die aus der Audiometrie kommen den Korpus betreffend copyright geschützt. Weiterführende Informationen zu akustischen und linguistischen Eigenschaften von Sprache unter dem Aspekt der Kommunikation liefert beispielsweise [Laz07].

1.2.5.1 Referenzstimulus

Das Referenzsignal dient als Vergleichsprobe zur Bewertung eines Attributs. Es entspricht dem unbearbeiteten Originalsignal voller Bandbreite [1534, S.5]. Vielfach ist die Definition beziehungsweise Generierung eines Referenzsignals ein Vorgang mit beträchtlichem Zeitaufwand, oder schlichtweg nicht realisierbar, und es ist darüber hinaus nicht immer offensichtlich, wie das Optimum zu definieren ist. Aus diesem Grund wird oftmals ein hochqualitatives Vergleichssystem als Referenzsystem verwendet, beispielsweise ein Lautsprecher oder Codec und die Bewertung erfolgt über Qualitätsdifferenzen [Bla12]. Nachteilig ist anzumerken, dass die resultierende relative Bewertung eine Vergleichbarkeit mit anderen Versuchsergebnissen oft nicht ermöglicht.

Der Einsatz einer Referenz kann auf zwei Arten erfolgen:

- das Referenzsignal steht der Versuchsperson während dem Test als Vergleichsprobe zur Verfügung. (*Präferenztest, siehe affective measurement*) In diesem Fall wird das Referenzsignal nicht auf der Skala beurteilt.

- *„Hidden reference“*: das Referenzsignal dient in einem Set als Teststimulus zur Überprüfung der Diskriminierungsfähigkeit einer Versuchsperson, siehe auch Post-Screening Abs. 1.1.4. Die Versuchsperson bewertet die versteckte Referenz nach denselben Kriterien, die sie für die übrigen Stimuli anwendet. Wird eine versteckte Referenz innerhalb eines Versuchsaufbaus verwendet, ist es grundsätzlich wichtig, die Versuchsperson vorher darüber in Kenntnis zu setzen (vgl. Abs. 2.3.2.2,

Versuchsablauf). (*eine Ausnahme stellt der Fall dar, dass man die Versuchspersonen hinsichtlich ihrer Performance und auch persönlichen Präferenzen miteinander vergleichen möchte und bewusst auf die Vorabinformation verzichtet. In diesem Fall erfüllt der Hörversuch nicht mehr seinen eigentlichen Zweck, eine Eigenschaft eines Signals zu beurteilen.*)

Referenzrauschen

Das Referenzrauschen kann auch als spezielle, standardisierte Form der Degradierung eines Referenzstimulus gedeutet werden. Dabei wird ein Referenzsignal mit einer standardisierten Rauschquelle (modulated noise reference unit, MNRU [810]) verrauscht und der sich ergebende Signal-Rausch-Abstand wird für die Normalisierung des arithmetischen Mittelwerts (*mean opinion score, MOS*) als sogenannter Q-Wert [entspricht SNR, in dB] herangezogen. Der MOS wird somit für Testergebnisse anderer Institute vergleichbar. Die Literatur liefert hier allerdings auch Nachweise, die entgegen der ITU zeigen, dass sich selbst der MOS als international konsistent anzusehender Parameter nur sehr eingeschränkt vergleichen lässt. Zudem ist dafür Sorge zu tragen, dass sich das Rauschen perzeptiv zum Vergleich mit den Störungen des jeweiligen Testsystems eignet, was in modernen Sprachübertragungssystemen nicht unbedingt der Fall ist [Moe10, Kap5.9] [Cot11, Kap 2.2].

1.2.5.2 Anker

Der Anker als Begriff hat mehrere Bedeutungen. Er bezieht sich entweder auf die verwendete Skala, auf die Eingewöhnungsphase oder auf das verwendete Klangmaterial.

Bei Skalen versteht man unter Ankerpunkten oder einem Anker:

- die Beschriftung der gesamten Skala
- die Beschriftung der Endpunkte oder lediglich des Mittelpunkts der Skala. Der Kontext des Endpunktes hat unterstützende Funktion.
- eine sichtbare Unterteilung der Skala.

Werden der Versuchsperson vor dem eigentlichen Experiment Stimuli vorgespielt, die bei der Orientierung im Hörtest helfen sollen, nennt man diesen Prozess verankern. Dazu werden meist Stimuli verwendet, die das zu sensibilisierende Artefakt oder die Verzerrung deutlich hörbar repräsentieren, um die Eigenschaften des abzufragenden Attributs zu verdeutlichen; dieser Aspekt betrifft vor Allem die Trainingsphase von Experten!

In Hörversuchen selbst kommen Anker (engl. Anchor) zumeist als versteckter Anker (*hidden anchor*) vor. Dabei wird ein Referenzstimulus hinsichtlich einer Eigenschaft, deren Auswirkung in Bezug auf die Klangqualität bekannt ist, verändert. Als Anker fungieren bei Tests von Audiomaterial beispielsweise folgende Abänderungen von Referenzsignalen, die nach ihrer Ähnlichkeit zu den zu testenden Qualitätskriterien angewendet werden [1534, S.6]:

- Tiefpassfilter und Begrenzung der Bandbreite auf 3,5 kHz (kommt aus Schmalband),

- Bandbreitenbegrenzung auf 7 kHz oder 10 kHz (Breitband, bzw. Superbreitband)
- Verringertes Stereoabbild
- Zusätzliches Rauschen
- Signalausfall (*drop outs*)
- Verlust von Datenpaketen (*packet loss*)

Die aus Sicht der Versuchsperson versteckten Ankersignale können dem Versuchsleiter zusätzliche Informationen über das zu untersuchende Audiosystem liefern. Tendenzen in der Ähnlichkeit von Auswirkungen der Artefakte und jenen Auswirkungen der bekannten Artefakte des Ankers auf das Antwortverhalten der Versuchspersonen, können Hinweise auf die Vergleichbarkeit bekannter Qualitätskriterien mit dem Testsystem liefern und somit zur Ergebnisinterpretation unterstützend eingesetzt werden.

Über die Verwendung von Hidden Anchors wird die Versuchsperson in der Regel nicht in Kenntnis gesetzt.

Anker erfüllen aber meist eine Kontrollfunktion innerhalb eines Versuchsablaufs für den Versuchsleiter.

Bsp. Es sollen sechs verschiedene Audiosysteme miteinander verglichen werden, die zu testende Signale mit verschiedenartigem Rauschen und unterschiedlichem Rauschanteil liefern. Der Versuchsleiter beschließt, auch aufgrund der Einfachheit der Umsetzung, einen Anker zu verwenden. Dazu hinterlegt er das hochqualitative, unverrauschte Originalsignal (= Referenzsignal) mit einem ihm bekannten Rauschen (Rosa Rauschen, Weisses Rauschen, etc.), das jenem der Testsignale am meisten ähnelt. Innerhalb des Versuchsablaufs werden die Teststimuli und der Anker in randomisierter Reihenfolge bewertet. Liefert der Anker nun ähnliche Ergebnisse wie die zum Audiosystem zwei gehörenden Teststimuli, ist zu vermuten, dass dieses System ähnliche Rauschanteile liefert wie jene, die für den Anker gewählt wurden. Der Anker kann aber auch ähnliche Ergebnisse zu allen Systemen außer einem System liefern, was vermuten lässt, dass dieses System ein anderes Verhalten zeigt, vielleicht viel weniger oder stärker rauscht, als die übrigen. Hinweis auf einen Fehler bei der Signalgenerierung? Natürlich liegt der Gedanke nahe, diesen Fehler auch ohne Anker erkennen zu können. Aufzuzeigen war, dass der Anker einen Fokus setzt, von einer bekannten Fehlerquelle auf Gemeinsamkeiten oder Exklusionen in den anderen Quellen zu schließen.

1.2.5.3 Anzahl an Stimuli

Um ein Gefühl dafür zu bekommen, wie realistisch die gewünschten Testbedingungen überhaupt durchzuführen sind, ist es sinnvoll, sich zu Beginn des Designprozesses die verschiedenen Kombinationen Systemanzahl, Anzahl unabhängiger Variablen und Anzahl der Abstufungen zu überlegen.

Beispiel vollständiger Paarvergleich (vgl. Abs. 2.2.2.2): man möchte 5 Surround-Systeme miteinander im Referenzraum vergleichen. Diese sollen hinsichtlich 6 verschiedener räumlicher und klangfärbender Attribute betrachtet werden. Zusätzlich sollen 4 verschiedene

Soundbeispiele (3 versch. Musikstile, 1x Musik mit Sprache als Actionfilmsimulation) den möglichen Wertebereich an Musikgeschmack repräsentieren. Der Versuchsleiter überlegt, einen vollständigen Paarvergleich durchzuführen (CCR). Folglich rechnet er sich die Anzahl an Teststimuli wie folgt aus:

$$t = (s \cdot b)$$
$$t = 5 \cdot 4 = 20 \qquad (61)$$

$t \ldots Teststimuli$
$s \ldots Anzahl\ der\ Syteme$
$b \ldots Anzahl\ der\ Testbedingungen$

Der vollständige Paarvergleich besagt, jedes mögliche Paar miteinander zu vergleichen, auch in umgekehrter Reihenfolge, d.h. A-B und im zweiten Durchlauf B-A. Der einseitige Paarvergleich demgegenüber testet lediglich einen Durchlauf A-B.

Für den einseitigen Paarvergleich würde er demnach von jeder Versuchsperson für jedes der 6 Attribute folgende Anzahl an Beurteilungen benötigen:

$$r = \binom{t}{2} = \frac{t!}{2! \cdot (n-2)!} = \frac{t \cdot (t-1)}{2}$$
$$r = \frac{20 \cdot 19}{2} = 190 \qquad (62)$$

$r \ldots Anzahl\ an\ Beurteilungen\ /\ Attribut$

Für den vollständigen Paarvergleich wäre dementsprechend die doppelte Anzahl an Beurteilungen infolge von zwei Durchläufen notwendig. Für jede Versuchsperson würde dies bedeuten, 380 Bewertungen je Attribut, also in Summe 2280 Bewertungen abgeben zu müssen. Er wird sich in diesem Fall gegen den vollständigen Paarvergleich entscheiden. Da es sich in dem Beispiel um Audiomaterial handelt, kann grundsätzlich das MUSHRA-Verfahren oder das Double Blind Triple Stimulus Verfahren angewendet werden. Sind die qualitativen Unterschiede der Surroundsysteme die verschiedenen Attribute betreffend gut erkennbar, ist MUSHRA besser geeignet (vgl. Abs. 2.3.2.3), im Fall geringer qualitativer Unterschiede das Double Blind Triple Stimulus Verfahren (vgl. Abs. 2.3.2.2). Zudem ist noch zu überlegen, ob nicht die Anzahl an abzufragenden Attributen zu reduzieren ist.

1.2.6 „Experiment Design“

Die Literatur [Hun96, Kap2][Næs10, Kap12] unterscheidet im Designprozess eines Versuchs grundsätzlich zwei verschiedene Bereiche. Der eine Bereich beschäftigt sich mit der optimalen Auswahl des Stimulussets (*Factorial design*) und der andere mit der optimalen Zuweisung der Stimuli zu den Vpn im Test selbst (*allocation of stimuli design*). Im Factorial Design liegt der Fokus auf der Anzahl der zu generierenden Stimuli. Dieser Aspekt des Experiment Design wird völlig entkoppelt von den Vpn betrachtet. Ist das Stimulusset

festgesetzt, wird im nächsten Schritt, dem Allocation of Stimuli Design, überlegt, ob die Vpn alle Stimuli oder nur einen Teil dieser beurteilen. Im Experiment Design wird also der Umgang mit den bekannten Variablen, den Faktoren, betrachtet, deren Einfluss auf die Messdaten innerhalb des Versuchs untersucht wird.

Die Intention innerhalb dieses Abschnitts ist es, einen prägnanten Überblick gängiger Varianten im Bereich von Hörtests zu liefern und auf die Wichtigkeit des Experiment Design hinzuweisen.

Anmerkung: Auch im Allocation of Stimuli Design können Ansätze, welche im Factorial Design beschrieben wurden, verwendet werden. Demnach ist beispielsweise der Latin Square Ansatz nicht auf den Einsatz im Factorial Design beschränkt.

1.2.6.1 Factorial design

Um überhaupt eine geeignete Auswahl an Stimuli treffen zu können bzw. diese generieren zu können, ist neben der zugrunde liegenden Hypothese (vgl. Abs. 1.2.1) vor allem wichtig, dass auch das geeignete Antwortattribut bereits definiert wurde. Dieses stellt den Ausgangspunkt dar, um sämtliche potentielle, zu überprüfende Faktoren (*factor*) für den Test vorerst zu identifizieren. Es wird in diesem Schritt folglich festgelegt, welche unabhängigen Variablen variiert werden, und welche Variablen konstant gehalten werden, weil sie für den Informationsgewinn von untergeordneter Bedeutung sind. Somit kann der Einfluss der veränderlichen Faktoren innerhalb des festgelegten Wertebereichs mit entsprechender Auflösung getestet und bewertet werden.

Die Anzahl der Möglichkeiten und damit der Stimuli steigt mit der Anzahl der zueinander in Bezug gesetzten, variierten Faktoren. Zusätzlich spielt die Anzahl der Zustände (und damit der Wirkungsbereich, *range*), die jede Variable annehmen kann eine erhebliche Rolle [Mon97, Kap1].

$$x = z^{f} \tag{63}$$

$$x \ldots Anzahl\ der\ Stimuli$$
$$z \ldots Anzahl\ der\ treatments$$
$$f \ldots Anzahl\ der\ Faktoren / unabh.Var.$$

Obige Formel verdeutlicht, dass sobald mehrere unabhängige Variablen bzw. Faktoren getestet werden sollen und zudem die Anzahl der Abstufungen (*treatment, level*; *z.B. Lautstärke: 1 dB-Schritte von 60 bis 85 dB*) eine bestimmte Größe erreicht, ein vollständiges factorial design (*full factorial*) als Stimuliset in einem Hörtest weder zeit- noch kosteneffizient realisierbar ist. Aus diesem Grund wird versucht, eine optimale Kombination aller Parameter miteinander zu finden (*fractional factorial design*). Dazu stehen drei grundlegende Prinzipa im Design zur Verfügung: Wiederholung, Randomisierung und Blockbildung [Mon97, Kap2]. Einen grundlegenden Ansatz eines fractional factorial designs

stellt der Latin Squares Ansatz dar, bei dem die Randomisierung und die Blockbildung kombiniert werden.

Latin Squares[9]

Eine Blockbildung (*block design*) innerhalb der Randomisierung hat den Zweck, eine Restriktion innerhalb des Zufallsprozesses zu bilden und somit die Variabilität einer nun kontrollierten Störvariable zu unterbinden. Das hat die Verringerung des Restfehlers (*residual error*) zur Folge. Montgomery ist entgegen Bech [Bec06, Kap6] der Meinung, dass die Faktorenfindung zur Blockbildung relativ einfach unter der Voraussetzung ist, dass das Problem (Hypothese bzw. Fragestellung), die Faktoren und die Antwortvariable bekannt sind. Weiters empfiehlt er, diesen pre-experimentellen Prozess iterativ und gruppendynamisch zu gestalten [Mon97, Kap1].

Der Ansatz über Latin Squares erlaubt eine Blockbildung von gleichzeitig zwei Dimensionen (entspricht zwei zu wählenden Variablen/Faktoren für die Blockbildung) und somit die Kontrolle über zwei Störeinflüsse. Er fordert für jede unabhängige Variable dieselbe Anzahl an treatments und legt fest, dass jede Kobinationsmöglichkeit der insgesamt drei vorkommenden Faktoren untereinander nur einmal verwendet wird [Mon97, Kap5]. (*Anmerkung: die Anwendung dieses Ansatzes ist dabei grundsätzlich nicht auf die Stimulusauswahl, also das Factorial Design, beschränkt.*)

Beispiel: getestet werden sollen 5 verschiedene Text-to-Speech Programme. Als Testmaterial werden 5 völlig verschiedene Sätze ausgewählt. Zusätzlich sollen 5 verschiedene Abhörvarianten getestet werden. Dementsprechend würde man für ein full factorial design 125 Stimuli benötigen, die von jedem Hörer bewertet würden. Der Forschungsleiter wählt den Ansatz eines Latin Squares, und beschließt dabei, die beiden Faktoren Abhörvarianten und verschiedene Sätze konstant zu halten, wie folgt:

[9] Latin Square: die Bezeichnung stammt von der quadratischen Anordnung der Elemente und Benennung dieser mit A, B, C, D,…

Sätze	Abhörvarianten				
	1	**2**	**3**	**4**	**5**
1	**A**	**B**	**C**	**D**	**E**
2	**B**	C	D	E	A
3	**C**	D	E	A	B
4	**D**	E	A	B	C
5	**E**	A	B	C	D

Tab. 8: Latin Square für fünf Text-to-speech Programme

Die Buchstaben in Tab. 8 *repräsentieren dabei die fünf verschiedenen zu beurteilenden Programme. Zu erkennen ist, dass Programm A lediglich einmal in der Kombination Satz 1 mit Abhörvariante 1 vorkommt. Durch dieses Design ergibt sich also ein Stimulusset von lediglich 25 Stimuli.*

Ein Standard Latin Square, bei dem die erste Reihe und Spalte die exakte alphabetische Reihenfolge repräsentieren, kann analog zu Tab. 8 auch für eine andere Anzahl an Treatments z konstruiert werden.

Das Latin Square Design basiert auf dem nachfolgenden, rein additiven, statistischen Modell (vgl. Abs. 1.1.3, Regressionsmodell), [Mon97, Kap5]:

$$Y_{t,ij} = \mu + \alpha_i + \tau_t + \beta_j + \varepsilon_{t,ij} \qquad (64)$$

$Y_{t,ij} \ldots Beob.d.i-ten\,\mathrm{Re}ihe,\ j-ten\ Spalte, t-ten\ Stimulus$

$\mu \ldots arithmetischer\ Mittelwert\ aller\ Bewertungen\ (idealisiert)$

$\alpha_i \ldots Effekt\ d.i-ten\,\mathrm{Re}ihe$

$\beta_j \ldots Effekt\ d.\ j-ten\ Spalte$

$\tau_t \ldots j-ter\ treatment\ effect$

$\varepsilon_{t,ij} \ldots zufälliger\ Fehler$

Statistisch betrachtet verliert man natürlich an Informationsgehalt aufgrund der Reduktion der Stimuli. Zudem kann man keine Aussage mehr über die Interaktion der gewählten Faktoren treffen. (*Beispiel: eine Aussage über den Zusammenhang der Auswirkung der verschiedenen Abhörvarianten in Bezug auf die verschiedenen Programme oder die verschiedenen Sätze ist nicht mehr möglich!) Allerdings lässt sich der Einfluss der Abhörvarianten sehr wohl über die Mittelwertbildung der Spalten abschätzen. Ebenso kann*

man eine Abschätzung über den Einfluss der Sätze über die Mittelwertbildung der Reihen treffen.) Zudem geht die Anwendung eines additiven Modells davon aus, dass sich ein Faktor unabhängig von den beiden anderen immer gleich verhält, was nicht korrekt ist. (*Beispiel: Das Programm wird vielleicht Sätze mit vielen Plosivlauten schlechter umsetzen, als solche mit vielen Zischlauten.*)

Ein klarer Vorteil dieses geblockten Designansatzes besteht in der Kosten- und Zeitersparnis gegenüber einem full factorial design und überdies in der Möglichkeit, mehrere unabhängige Variablen innerhalb eines Testdurchlaufs bewerten zu können. Die Schwierigkeit liegt sicherlich in der Beurteilung, welche der unabhängigen Variablen sich überhaupt für die Blockung eignen [Bec06, Kap6].

Graceo Latin Square[10]

Dieser Design Ansatz unterscheidet sich vom Vorherigen durch die Kontrollfunktion (*blocking*) einer dritten Variable aufgrund der Superposition zweier Latin Squares derselben Dimension, die zueinander orthogonal stehen. Er findet bei Fragestellungen Verwendung, in denen vier verschiedene Faktoren von Bedeutung sind.

Reihe	**Spalte**			
	1	**2**	**3**	**4**
1	Aα	Bβ	Cγ	Dδ
2	Bδ	Aγ	Dβ	Cα
3	Cβ	Dα	Aδ	Bγ
4	Dγ	Cδ	Bα	Aβ

Tab. 9: Graceo-Latin Square

Youden Square

Hat man für mehrere Faktoren die Anforderung an eine verschiedene Anzahl an treatments (Abstufungen der unabhängigen Variable), kann der Ansatz über die restriktive Variante des Latin Squares, die Youden Squares gelöst werden.

Der Youden Square zählt zu den balanced incomplete block designs[11] und besteht im Wesentlichen aus dem Latin Square Ansatz, bei dem eine Reihe oder Spalte entfernt wurde.

[10] Graceo Latin Square: die griechischen Buchstaben geben dem Design seinen Namen.

Dementsprechend ist dieser nicht mehr quadratisch, und somit auch für Faktoren mit unterschiedlich vielen Skalierungsstufen anwendbar.

Sätze	Abhörvarianten				
	1	2	3	4	5
1	A	B	C	D	E
2	B	C	D	E	A
3	D	E	A	B	C

Tab. 10: Youden Square für fünf Text-to-speech SW-Tools mit nur 3 Sätzen

Weitere Ansätze zum Treatment Design oder Fractional Design können Montgomery [Mon97] entnommen werden.

1.2.6.2 Allocation of Stimuli Design

In diesem Abschnitt wird bereits von einem bestehenden Stimulusset ausgegangen und nun der Fokus auf die Präsentation dieser gelegt. Dabei wird sowohl die Abspielreihenfolge der Stimuli betrachtet als auch die Frage, ob die Vpn alle oder nur eine bestimmte Auswahl an Stimuli vorgeführt bekommen.

Innerhalb eines Hörversuchs kann die Auswirkung unterschiedlicher, unabhängiger Variablen auf die abhängige Variable getestet werden. Einschränkende Faktoren innerhalb des Designs sind mitunter durch die Versuchspersonen selbst festgelegt. Je nach Komplexität der dargebotenen Stimuli, ist die Zeitdauer innerhalb welcher keine Ermüdungseffekte auftreten begrenzt und dementsprechend im Design zu berücksichtigen (vgl. Abs. 1.2.8.3, Dauer des Versuchs). Bei einem Versuch mit großem Stimulusset ist wiederum zusätzlich der Aspekt, nicht kontrollierbare Variablen (vgl. Abs. 1.1.3, Regressionsmodell) möglichst zu eliminieren, zu berücksichtigen.

Within subject design über den Balanced Latin Square[12]

Das Optimum bezüglich statistischer Aussagekraft wird durch ein full factorial design in Kombination mit dem vollumfänglichen Versuchspersonendesign (*within subject design*)

[11] Balanced incomplete block d.: werden im Allgemeinen dann verwendet, wenn nicht alle treatment Kombinationen innerhalb eines Blocks realisierbar, jedoch gleich bedeutend sind. Jedes Treatmentpaar tritt gleich oft auf → balanced.

[12] Balanced Latin Square: gehört zu den randomized complete block designs.

erzielt. Dabei wird jedem Hörer jeder Stimulus vorgespielt und die Auslastung der Versuchspersonen somit optimiert [Bec06, Kap6]. Hat der Versuchsleiter aufgrund des zu großen Stimulisets lediglich den fractional factorial design – Ansatz gewählt, ist jedenfalls ein within subject design – Ansatz zu empfehlen, um weitere Restriktionen in der statistischen Aussagekraft zu vermeiden.

Der Design-Ansatz, der nun auf die Versuchspersonen referenziert, kann wiederum über einen Latin Square gelöst werden. Als geblockte Variablen kommen hier die Versuchspersonen zum Einsatz, oder bei mehrtägiger Versuchsdurchführung auch der Tag als eigener Block (unabhängige Variable). Diese gezielte Randomisierung in der Zuordnung der Stimuli zu den Versuchspersonen führt somit zur Minimierung systematischer Fehler bei gleichzeitiger Minimierung von nicht kontrollierbaren Variablen [Har05, KapV.5].

Um zusätzlich zu vermeiden, dass immer nach Programm B das Programm C angehört wird, kann auch der Latin Square ausgeglichener (*balanced*) konstruiert werden [Mon97, Kap5]. Mit *n* und *i* als Laufvariablen für die Anzahl an Sätzen *i* und die Anzahl an Abhörvarianten *n* im Versuch, wobei *n=i*, folgt für den Designansatz:

Sätze	**Abhörvarianten**					
	1	2	3	4	n-1	n
1	**A**	**B**	**z**	**C**	**z-1**	**D**
2	**B**	C	A	D	z	z-1
3	**C**	D	B	z-1	A	z
4	**D**	z-1	C	z	B	A
i-1	**z-1**	z	D	A	C	B
i	**z**	A	z-1	B	D	C

Tab. 11: Balanced Latin Square

Die Konstruktion aus Tab. 11 erfolgt nach der ersten Reihe und ersten Spalte. Die zweite Reihe ergibt sich aus der um eine Position nach rechts verschobenen ersten. Zudem folgt auf z wieder 1 (*wrap*). Der Ansatz funktioniert für eine gerade Anzahl von z Treatments.

Mit dem Balanced Latin Square wird also die Zuordnung der Reihenfolge der Stimuli zu den Versuchspersonen gezielt variiert und die Stimulireihenfolge als Fehlerquelle vermieden (vgl. Abs. 1.2.9, Kontexteffekt).

1.2.6.3 Fertigstellung des Experiment Design

Letztendlich stellt sich dem Versuchsleiter die Frage, wie viele Wiederholungen pro Stimulus erforderlich sind, unter der Annahme dass nicht zwangsläufig ein standardisiertes Verfahren angewendet wird. Die Repetition kann dabei beispielsweise über eine Gruppe von Versuchspersonen erfolgen, die je eine Beurteilung zum Stimulus abgeben, oder aber durch mehrere Durchläufe mit einer Versuchsperson abgehandelt werden. Dazu gilt es, folgende Punkte zu beachten:

- ***Auflösung des subjektiven Antwortverhaltens***: ist diese fein genug? Darüber gibt ein Vorversuch (Pilottest) Aufschluss

- ***Varianz der Beurteilung***: diese ist im Allgemeinen, speziell aber bei naiven Hörern, nicht bekannt. Hier liegt gleichzeitig ein großer Vorteil eines Expert Listening Panels, dessen Varianzeigenschaften über laufendes Screening der vorangegangenen Hörversuche gut vorausgesagt werden können. Im Audiobereich wird als Varianz bei einer Skala von 0-10 mit 0,1-Skalierungsstufen $\sigma_{\bar{x}}^{2} = 0{,}4$ für Hörer mit Expertise verwendet. Dieser Wert wurde für ein ELP empirisch ermittelt und kann allgemein für ein ELP angenommen werden [Bec06, Kap6][Fra12]. Durch laufendes Screening ist dieser Wert für das jeweilige ELP zu validieren und gegebenenfalls anzupassen.

- Wahrscheinlichkeit für ***Fehler erster und zweiter Art*** (vgl. Abs. 1.1.2.1, Hypothesenprüfung und Signifikanzniveaus)

- ***Ziel*** der Fragestellung eines Versuchs ist stets, die Variation des subjektiven Eindrucks den kontrollierbaren, unabhängigen Variablen zuschreiben zu können und damit jegliche Einflüsse durch zufällige Fehler auf das Ergebnis weitgehend ausschließen zu können.

1.2.7 Pilottest

Der Pilottest ist ein essentieller Bestandteil eines Versuchsdesigns. Er hat die Aufgabe, das Testsetup zu überprüfen, Indizien zu Verbesserungen und Adaptionen im Design und der Spezifizierung der Variablen zu geben, die Wahl der Rahmenbedingungen zu untersuchen und ein Vorhandensein von Störvariablen zu detektieren. Ein Pilotversuch kann jedoch nur richtungweisend fungieren, zumal das Pilotsetup viel restriktiver als der Aufbau des eigentlichen Versuchs ist.

Üblicherweise wird für den Vorversuch eine kleine Versuchsgruppe von ungefähr fünf Teilnehmern gewählt. Diese sollen derselben Grundgesamtheit entstammen wie die Personen im eigentlichen Test. Sind naive Hörer für den Hörtest vorgesehen, sollten also keine Experten im Pilottest teilnehmen (Bias). Im Allgemeinen gilt es, die Rahmenbedingungen den tatsächlichen Rahmenbedingungen möglichst getreu nachzuahmen, um Effekte durch Störvariablen, die einzig im Pilottest auftauchen, weitgehend ausschließen zu können.

Während des Versuchs soll der Versuchsleiter, wie im tatsächlichen Ablauf auch, nicht anwesend sein, es sei denn, das Versuchsdesign ist darauf ausgelegt worden. Zum Ende des Durchlaufs kann man die Versuchsperson zu Dauer und Anstrengung der Durchführung befragen und so Kenntnis über notwendige Adaptionen im Design gewinnen.

Aus den gewonnenen Daten und deren Darstellung (vgl. Abs. 1.1.1, Deskriptive Statistik) kann man feststellen, ob die bis dato angenommene Teilnehmeranzahl und -auswahl korrekt war, oder, ob die perzeptiven Unterschiede zu fein sind um mit der gewählten Anzahl an Versuchspersonen zu validen Ergebnissen zu führen. Nicht nur Tendenzen im Antwortverhalten für den eigentlichen Versuch sind erkennbar, sondern auch Fehler können erkannt werden. Dabei sei darauf hingewiesen, dass Ausreißer nicht zwangsläufig als Fehler zu interpretieren sind und dass das a posteriori „Schönen“ von Daten keinen echten Wissenszuwachs ermöglicht. (vgl. Abs. 1.1.4, Ausreißer).

Die Überprüfbarkeit der Fragestellung an sich wird zudem getestet, das heißt, ob sich die aufgestellte Hypothese überhaupt zur Überprüfung eignet, oder, ob weitere Restriktionen in der Fragestellung notwendig sind.

Streng genommen ist die Pilotphase als repetitiver Adaptionsprozess zu betrachten, der durchaus mehrere Pilottests beinhalten kann. Eine genaue Deskription sämtlicher Rahmenbedingungen von Versuchen dient gemeinsam mit einem ELP als fundierte Basis für jegliche weitere Designprozesse. Die detaillierte Dokumentation sei als weiterführendes Hilfsmittel auch im Forschungsbereich nicht unterschätzt.

1.2.8 Rahmenbedingungen

Ziel eines Hörversuchs ist das Erlangen von Daten, die objektiv und reproduzierbar sind (vgl. Abs. 1.2). Um diesen Anforderungen zu genügen, ist es relevant, sämtliche Parameter soweit möglich zu kontrollieren, um den Fehler durch unkontrollierte und störende Variablen möglichst zu reduzieren (vgl. Abs. 1.2.9.). Nachfolgend sind die wichtigsten Randbedingungen, diese entsprechen den unabhängigen und kontrollierten Variablen im Regressionsmodell von Abs. 1.1.3, gelistet, die zusätzlich zu den vorangegangenen Punkten zu beachten und bei Bedarf in der Sekundärliteratur detailliert vorzufinden sind [Bec06, Kap5],[Fel12, Kap5.4].

1.2.8.1 Anzahl der Versuchsteilnehmerinnen

Die benötigte Teilnehmeranzahl wird vor der Versuchsdurchführung definiert. Folgende Aussagen über die benötigte Anzahl an Probandinnen sind in der Literatur zu finden:

- Laut Fellbaum: „Je nach Messverfahren benötigt man 5-50 Personen. Da die Auswertung auf statistischen Parametern basiert, gilt meist die einfache Regel: je mehr Testpersonen, desto besser.“, [Fel12, Kap5.4].

- ITU BS 1284-1: ELP mindestens 10, naive Hörer: mindestens 20. Dabei wird keine weitere Begründung oder Grundlage für die Zahlen geliefert.

- ITU BS 1534-1: Die Daten von 20 Versuchspersonen sind oft ausreichend.

Um diese Aussagen interpretieren zu können, wird ein statistischer Ansatz beschrieben, der auf benötigte Voraussetzungen für eine sinnvolle Abschätzung der Teilnehmeranzahl verweist. Zur weiterführenden Information bezüglich Stichprobenverfahren wird auf Cochran verwiesen [Coc72].

Nach Cochran stellt sich die Frage nach der Erwartungshaltung gegenüber der Stichprobe bezüglich des zu testenden Merkmals. Diese Erwartung wird zumeist durch zu definierende Fehlergrenzen, sowie den angedachten Anwendungsbereich spezifiziert. Dabei sind Konsequenzen verschiedenartiger Fehlergrenzen, sowie die a posteriori Erweiterung des Anwendungsbereichs der Stichprobe nur schwer abschätzbar [Coc72, Kap4].

Das Signifikanzniveau wird bei Hörversuchen zumeist auf α=5% festgesetzt. Dabei wird gleichzeitig versucht, das Niveau für β möglichst gering zu halten. Bortz empfiehlt, β viermal so groß wie α zu wählen (nach [Bor04]). Damit sind die beiden Risiken, die in indirekter Abhängigkeit zueinander stehen, ausgewogen. Einerseits möchte man eine richtige H_0 nicht ablehnen, was der Wahl von α gleichkommt, gleichzeitig möchte man deswegen auch nicht eine falsche H_0 annehmen.

Eine Möglichkeit der Abschätzung der Anzahl an Versuchspersonen wird über den Einsatz einer Effektgröße[13] ε getätigt. Ist ε =0 liegt kein Effekt vor und die Nullhypothese wird angenommen, wird sie hingegen abgelehnt, ist die Effektgröße ein Maß für die Aussagekraft der Annahme von H_1. Es sei angemerkt, dass die Auswahl bzw. Abschätzung von ε einen komplexen Vorgang darstellt, der eine gute Kenntnis der Problemstellung erfordert [Bor04, Kap4.6]. Die Effektgröße für nachfolgendes Beispiel wurde willkürlich gewählt. Demnach ist das Beispiel demonstrativ zu betrachten und keinesfalls als Empfehlung für eine Teilnehmeranzahl miss zu verstehen.

Beispiel: Ausgegangen wird von einem Test, in dem Experten ein System auf einer Skala von 0-10 bewerten sollen. Die zugrundeliegende Nullhypothese lautet wie folgt:

$$H_0 : \mu = 5 \tag{65}$$

Die Varianz von Experten wird mit σ^2=0,4 abgeschätzt (vgl. Abs. 1.2.6.3). Das Signifikanzniveau α wird mit 5% festgesetzt und daraus β=20% festgelegt. Die Effektgröße ε wird zudem mit ε=0,5 als mittlerem Effekt gewählt.

Die Prüfgrößen für die Standardnormalverteilung lauten wie folgt:

[13] Effektgröße: „Mit der Effektgröße wird also festgelegt, wie stark der H_1-Parameter μ_1 (mindestens) von μ_0 abweichen muss, um von einem praktisch bedeutsamen Effekt sprechen zu können." [Bor04, S120] …bezogen auf den Vergleich eines Stichprobenmittelwertes mit dem Erwartungswert der Grundgesamtheit.

$$z_{(1-\alpha)} = \frac{\bar{x} - \mu_0}{\hat{\sigma}_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$$
$$z_{\beta} = \frac{\bar{x} - \mu_1}{\hat{\sigma}_{\bar{x}}} = \frac{\bar{x} - \mu_1}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \qquad (66)$$

Damit kann die Stichprobenkenngröße ohne Kenntnis der Mittelwerte abgeschätzt werden. Formel (65) und (66) werden über $\bar{x}$ gleichgesetzt und auf folgende Form gebracht:

$$\frac{\mu_1 - \mu_0}{\hat{\sigma}} = \frac{z_{(1-\alpha)} - z_{\beta}}{\sqrt{n}} \qquad (67)$$

Beide Seiten werden nun mit $\sqrt{2}$ erweitert, um die linke Seite durch die Effektgröße ersetzen zu können. Letztendlich erfolgt die Umformung auf n zu:

$$n = \frac{2 \cdot \left(z_{(1-\alpha)} - z_{\beta}\right)^2}{\varepsilon^2} \qquad (68)$$

$$mit \quad \varepsilon = \frac{\sqrt{2} \cdot \left(\mu_1 - \mu_0\right)}{\sigma} \quad \left(\mu_1 > \mu_0\right) \qquad (69)$$

$\varepsilon = 0{,}2 \ldots schwacher\ Effekt$
$\varepsilon = 0{,}5 \ldots mittlerer\ Effekt$
$\varepsilon = 0{,}8 \ldots sctar\,\mathrm{ker}\,Effekt$

Für die Veranschaulichung des Beispiels bedeutet das, folgende Werte aus der Tabelle auszulesen:

$$z_{0,95} = \left|-1{,}645\right| = 1{,}645$$
$$z_{0,2} = -0{,}84$$
$$n = \frac{2 \cdot (1{,}645 + 0{,}84)^2}{0{,}5^2} = 49{,}402 \approx 50 \qquad (70)$$

Der Vorteil dieser Abschätzung liegt in der Tatsache, die Varianz und den Erwartungswert nicht kennen zu müssen.

Konkludierend kann man sagen, dass der Anzahl an Teilnehmerinnen und der damit verbundenen Aussagekraft des jeweiligen Versuchs immer der Aufwand in der

Durchführung gegenübersteht. Der Leserin soll bewusst sein, wie viele Personen notwendig wären um die gewünschten Fehlergrenzen α und β zu erreichen, selbst wenn für die Durchführung eine geringere Anzahl an Teilnehmern zur Verfügung steht.

1.2.8.2 Auswahl der Versuchspersonen

Versuchspersonen werden entsprechend dem Anwendungszweck des zu prüfenden Systems und entsprechend der Zielgruppe ausgewählt. Demzufolge spielt in der Versuchsplanung die Grundgesamtheit, auf die Rückschlüsse gemacht werden soll, eine wesentliche Rolle. Die jeweilige Stichprobe, seien es nun naive Hörer, erfahrene oder gar Hörer mit spezieller Expertise, soll diese Grundgesamtheit durch sämtliche Einflussfaktoren auf die Population bestmöglich abbilden.

Zu den verschiedenen Arten von Versuchspersonen und ihren Merkmalen wird auf das Kapitel Versuchspersonen 1.2.2 verwiesen. Welche Gruppe jeweils eingesetzt werden soll wird in den Anwendungen im methodischen Teil dieser Arbeit angeführt (vgl. Abs. 2, Methoden). Werden Experten eingesetzt, ist zusätzlich die Entscheidung zu treffen, ob und wie viele Trainingseinheiten benötigt werden (vgl. Abs. 1.2.2.2, Experten).

1.2.8.3 Dauer des Versuchs

Die Versuchsdauer ist ein viel diskutiertes Thema und eindeutige Empfehlungen existieren praktisch nicht. Verschiedene Faktoren sind für ein variables Design bezüglich der zeitlichen Länge eines Versuchs verantwortlich. Zu diesen gehören unter anderen:

- Dauer des einzelnen Stimulus im Set
- Anzahl der zu verwendenden Stimuli innerhalb eines Versuchs
- Komplexität der Aufgabenstellung und Stimuli
- Ermüdungseffekte aufgrund der Monotonie der Aufgabenstellung
- Konzentrationsdauer, Motivation der Versuchspersonen

Es empfiehlt sich, innerhalb des Pilottests die Gesamtdauer zu dokumentieren und zu Versuchsende die Testpersonen bezüglich Anstrengung und Testdauer zu befragen.

Bezüglich der Gesamtdauer eines Hörtests sind in der Literatur teils widersprüchliche oder lediglich vage Aussagen zu finden. Während beispielsweise in der ITU-T P.800 ein Versuch eine Dauer von 20 bis maximal 45 Minuten nicht überschreiten soll, empfiehlt die ITU-R BS 1284 nach 15 bis 20 Minuten eine Pause einzulegen. Dem entgegen beschreibt Risch bei Zuhilfenahme von EL und fünfminütigen Pausen zwei bis drei Mal pro Stunde eine Dauer von 2 ½ Stunden pro Tag als erschöpfend [Ris91].

Schatz et al. untersuchten die Auswirkung der Testdauer auf Ermüdungserscheinungen und Zuverlässigkeit von subjektiven Qualitätsbeurteilungen. Die Bewertung der Quality-of-Experience (QoE) wurde dabei begleitet von Messungen der Häufigkeit des Lidschlags (eye

blink rate, EBR) und der Herzfrequenz durchgeführt. Ergebnisse zeigen, dass selbst nach einer Dauer von 90 Minuten und erkennbaren Ermüdungserscheinungen die Beurteilungen nach wie vor zu zuverlässigen Ergebnissen innerhalb der statistischen Toleranzen führen. Die Autoren schlagen demnach auch für Audioanwendungen eine Gesamtdauer von maximal 90 Minuten mit einer Pause nach den ersten 40 Minuten vor. Dabei sei angemerkt, dass ein signifikanter Erholungseffekt erst ab einer Dauer der Pause von mindestens 10 Minuten eintritt [Sch12].

1.2.8.4 Räumliche Ausstattung und Equipment

Die im Rahmen dieser Arbeit angeführten Methoden und Anwendungen beziehen sich auf den Einsatzbereich im Multimediaraum von Joanneum Research, der gemäß ITU-R BS 1116-1 geplant und konstruiert wurde. Die Betrachtungen schließen die Beachtung elektroakustischer Charakteristika von Referenz-Monitor Lautsprechern und Kopfhörern sowie von weiterem, zugehörigen Equipment (Soundkarte, Verkabelung, etc.) mit ein.

Potentielle Fehlerquellen betreffend die räumliche Umgebung können zusätzlich im Abschnitt Bias nachgelesen werden (vgl. Abs. 1.2.9.6).

Geräuschpegel und Spektrum

Der Grundgeräuschpegel soll vor und nach jedem Versuch gemessen und protokolliert werden.

1.2.8.5 Versuchsablauf

Während der Durchführung der Hörtests, die sich über mehrere Tage erstrecken kann, dürfen sich die festgelegten Rahmenbedingungen nicht verändern (kontrollierte Variablen). Dies schließt auch die die Versuchspersonen betreuende Person, also den Versuchsleiter, mit ein.

Speziell vor fordernden Aufgaben, in denen höchste Konzentration gewünscht wird, ist es ratsam, den Termin für die Versuchsperson auf eine Viertelstunde vor dem tatsächlichen Versuchsbeginn zu fixieren, sie zu begrüßen und ihr für die Wartezeit ein Getränk bereitzustellen.

Jeder Hörer soll zu Beginn des Versuchs aus seiner Alltagssituation abgeholt werden. Er hat dabei nach der Begrüßung durch die Versuchsleiterin die Möglichkeit, sich mit dem Multimediaraum vertraut zu machen und sich am Abhörplatz einzugewöhnen. In dieser Phase wird eine Erklärung über den Ablauf des Experiments in schriftlicher Form bereitgestellt und auch die formellen Angelegenheiten, wie Geheimhaltungserklärung und Entgeltsvereinbarungen werden geklärt. Um jeder Teilnehmerin dieselben Ausgangsbedingungen zu verschaffen, wird eine eventuell zusätzlich erfolgende mündliche Erklärung vorher festgelegt. Dabei ist auch zu bedenken, in welcher Weise auf Fragen reagiert wird. Die Eingewöhnungsphase soll fünf Minuten in Anspruch nehmen.

Hat die Hörerin vorerst keine Fragen, beginnt der eigentliche Versuchsablauf. Bei manchen Versuchen ist eine Eingewöhnungsphase gefordert bzw. gewünscht. Diese verfolgt zumeist

das Ziel, die Versuchsperson mit den Testsignalen oder Artefakten vertraut zu machen. Ist der akustische Teil des Versuchs vorbei, werden der Teilnehmerin noch Fragen zum Komfort usw gestellt, die üblicherweise auf einer Likert-Skala (nicht zu verwechseln mit den anderen Skalierungsmethoden) bewertet werden. Mit Hilfe dieser Daten ist es möglich, innerhalb der statistischen Analyse auf weiterführende, induktive Hypothesen zu testen.

Nach dem Hörtest erfolgt ein kurzes Dankeschön, die Übergabe der Entlohnung und die Vorbereitung des Systems für den nächsten Versuchsteilnehmer (Überprüfung der Funktionstüchtigkeit, etc.).

1.2.8.6 Protokoll- und Berichterstellung

Die Protokollführung wird aus Gründen der Nachvollziehbarkeit eines Versuchs empfohlen. Wesentlich ist es dabei nicht die konstanten Faktoren wie Raum, Equipment usw. zu erwähnen, sondern die besonderen Vorkommnisse, Abänderungen vom Normablauf und Auffälligkeiten. Das Protokoll dient der Nachvollziehbarkeit eines Hörversuchs und ist nicht offizieller Bestandteil des Berichts, das heißt hier kann alles niedergeschrieben werden, Irrelevantes ebenfalls. Oftmals sind es genau diese Kleinigkeiten, die im Nachhinein bei Ergebnissen, die nicht den Erwartungen entsprechen, gegebenenfalls Unregelmäßigkeiten in der Analyse plausibel machen.

Die Erfahrung zeigt, dass die Protokollführung, die während des Projekts geschieht, nicht wie oftmals vermutet im Nachhinein für die Nachwelt noch einmal schön aufbereitet wird. Das hat leider zumeist zur Folge, dass wichtige Details für andere Projektmitarbeiter verborgen bleiben. Aus diesem Grund wird empfohlen, auf handschriftliche Notizen zu verzichten und diese stattdessen formlos, durchaus auch stichwortartig, jedoch mit entsprechender Dokumentbezeichnung zu verfassen. Noch einmal wird darauf hingewiesen, dass das Protokoll vielfach als Informationsträger unterschätzt wird und für weitere Analysen, a posteriori Vergleiche oftmals die entscheidende Grundlage bildet.

Der Bericht ist als offizielles Dokument nach außen nach den Richtlinien des Qualitätsmanagmentsystems, welches der ISO 9001 unterliegt, zu verfassen. Eine Vorlage findet die Leserin im Anhang (vgl. Abs. 3.2.2).

1.2.9 Bias

Ziel des Versuchsdesigns ist es, reproduzierbare, zuverlässige, vergleichbare und aussagekräftige Ergebnisse zu erzielen. Demnach ist es essentiell, potentielle Fehlerquellen zu eliminieren oder zumindest konstant zu halten. In diesem Abschnitt werden verschiedene Arten von Kontext- und Bias-Effekten angeführt, die die Ursache für einen systematischen Fehler (vgl. Abs. 1.1.3, Regressionsmodell) bilden können. Ziel ist es somit, den Fokus auf Komplexitäten und Fehlerquellen in der Versuchsplanung richten und für den Umgang mit dem Setup zu sensibilisieren. Dabei können manche der genannten Fehlerquellen gut umgangen werden, andere Lösungsansätze mögen zu kosten- oder zeitintensiv erscheinen.

Die Definition der Begrifflichkeiten ist in der Literatur nicht konform, weswegen die vom Autor für sinnvoll erachtete Bezeichnung übernommen wurde.

Es sei zudem noch angemerkt, dass im Allgemeinen Effekte von Bias in der untersuchten Datenmenge nur schwer identifizierbar sind, da diese unterschiedlich stark ausgeprägt sein können, in Mischformen oder auch abhängig von der Vpn auftreten können.

1.2.9.1 Antwortattribut

Dumping Bias: tritt auf, wenn die VP in ihrer Beurteilungshaltung eingeschränkt ist aufgrund des begrenzten Wertebereichs der Skala. Dementsprechend neigt er dazu, sämtliche Antworten in der Umgebung der Mitte anzuordnen. Außerdem tritt er auf, wenn für die VP innerhalb des Versuchs Attribute erkennbar sind, die nicht abgefragt werden, dennoch aus Sicht der VP in die Beurteilung mit eingehen [Bec06, Kap4.2]. Im Pilottest können fehlende Attribute am Ende des Tests abgefragt und somit das Testdesign adaptiert werden.

1.2.9.2 Antwortskala

Perzeptive Nichtlinearität: wenn innerhalb einer Population manche Vpn auf Artefakte beispielsweise im low bit range übersensibel reagieren. Dieser Effekt tritt auch auf, wenn innerhalb eines Versuchs die Abstände der gewählten Skala nicht für alle Stimuli konstant sind (keine Intervallskala). In diesem Fall ist ein indirektes Skalierungsverfahren oder eine bipolare Skala zu empfehlen.

Kultureller Bias: Gerade wenn eine Teilnehmergruppe aus Menschen verschiedener Nationalität besteht, ist die Beschriftung und Richtung der Skala nicht zwangsläufig für alle VP als gleichwertig zu betrachten. Diskrepanzen können sich in der verbalen Beschriftung der Skalen zeigen, weil die Hörerin das Attribut nicht dem Sinn entsprechend versteht oder fehlinterpretiert. Ebenso kann es zu Verwechslungen der Endpunkte der Skala aufgrund einer numerischen Verankerung kommen, die kulturell bedingt sind (man denke an das Schulnotensystem, z.B. Schweiz vs Österreich: 6(sehr gut)-1 : 1(sehr gut)-5).

1.2.9.3 Versuchspersonen

Erwartungsbias: Beispielsweise wenn verschiedene, aus dem Hochpreissegment stammende Lautsprecher miteinander verglichen werden. Die Vpn geht davon aus, dass der Lautsprecher der Marke X mit dabei ist und beurteilt seinen vermeintlichen Lieblingslautsprecher am besten, selbst wenn sich diese Marke nicht im Test befindet. Der Bias beinhaltet demnach Erwartungshaltung und Emotionen der Vpn, die sich in Form von Über/Unterbewertung der Stimuli auswirken kann. Abhilfe schafft in diesem Fall eine große Population mit verschiedenem Hintergrund betreffend Audio [Zie08]. Die Autorin schlägt die Verwendung von zwei Populationen, Experten und naive Hörerinnen, sofern dem Anwendungsfall nach möglich, vor. Damit sind in Bezug auf Qualitätsbeurteilungen ähnliche Tendenzen der beiden Gruppen auf Gemeinsamkeiten der Qualität des jeweiligen Systems rückführbar.

Beurteilungsbias (Judgement bias): Die Akzeptanz einer Vpn für einen Stimulus mit einer bestimmten Verzerrung oder Artefakt ist viel geringer oder größer als jene der Norm (also aller übrigen Vpn). Diese Fehlerart ist auch durch eine Trainingsphase nicht behebbar. Der Judgement Bias kann aber konstant gehalten werden, indem innerhalb eines Testdurchlaufs

nur perzeptiv ähnliche Artefakte präsentiert werden [Qua88, Kap 2.1]. Alternativ dazu könnte man das komplette Stimulusset zweimal innerhalb des Tests präsentieren, somit die Vpn screenen und bei geringen Abweichungen in den Bewertungen zwischen den beiden Ergebnissen mitteln.

Contraction Bias: Extrema einer Skala werden im Allgemeinen von Vpn gerne vermieden. Verschiedene Ursachen sind dafür verantwortlich: ist die Vp vorsichtig, weil sie nicht weiß, was noch an Stimuli nachkommt, kann Abhilfe geschaffen werden, indem zu Versuchsbeginn eine kurze Eingewöhnungsphase präsentiert wird. Der Bias tritt demnach häufiger bei Verfahren mit nur einem Stimulus auf (vgl. Abs. 2.2.1) Sind Stimuli perzeptiv ähnlich, ist vorerst die geeignete Wahl an Vpn zu hinterfragen. Weiters kann ein Anker eingebaut werden, der in der Diskriminierung unterstützt (allerdings nur bei vergleichenden Verfahren, vgl. Abs. 2.2.2, und Abs. 2.3.2). Das hier auftretende Problem der zusätzlichen Informationsreduktion einer Antwortskala beschreibt auch Rohrmann [Roh78].

1.2.9.4 Stimuliauswahl

Neuheitseffekt (Recency Effect): gerade bei Klangbeispielen, die mehrere Sekunden dauern, wird letztendlich der Eindruck beurteilt, der als neuester in Erinnerung ist. Abhilfe schafft einerseits das Loopen des Stimulus, andererseits eine sorgsame, konsistente Auswahl des Klangausschnitts [Zie08].

Spacing bias: Dieser Bias ensteht als Mappingfehler in der Übersetzung der (theoretisch) perzeptiven Beurteilung eines Stimulisets in die objektive Beurteilung auf der Skala. Dabei kommt es vor, dass perzeptiv kleine Unterschiede auf einer Skala mit konstanten Abständen aufgeweitet werden, wohingegen perzeptiv große Unterschiede denselben konstanten Abständen zugeschrieben werden [Lie08]. Gerade bei Design des Mushra-Tests (wie auch bei anderen Tests mit Referenz) ist es daher sinnvoll, darauf zu achten, Systeme mit Eigenschaften zu testen, die zueinander in einem ähnlichen Qualitätsverhältnis in Bezug auf die Eigenschaft stehen und die Referenz unter diesem Aspekt sinnvoll zu wählen. Die statistische Aussagekraft einer Referenz, die zwar von hoher Qualität ist, aber in keinem sinnvollen, weil nicht mehr abschätzbaren, Bezug mehr zu den schlechtesten Systemen steht, ist zu hinterfragen. (*hilfreich ist vielleicht die Überlegung, ob man die Systeme von Vornherein in eine eindeutige Reihenfolge bringen kann, wenn man die Referenz wählt. Diese Frage sollte nicht ad hoc zu beantworten sein.*)

Zudem kommt der Aspekt zu tragen, dass von einem konstanten Abstand in der Verbalisierung der Skala auch für die perzeptive Domäne ausgegangen wird. In manchen Fällen mag es aussagekräftiger sein, die Beurteilung über ein indirektes Antwortverfahren (bei dem man auf die Zuweisung der Beurteilung auf eine Skala verzichtet), wie den Paarvergleich (vgl. Abs. 2.2.2.2, CCR) zu lösen.

1.2.9.5 Experiment Design

Kontextbias: die Reihenfolge (*order of presentation effect*) gleichwie der Umfang des Stimulussets beeinflussen die Art der Beurteilung. Demnach besteht eine Art Zusammenhang zwischen hintereinander gehörten Stimuli, der bei perzeptiv ähnlichen

Stimuli zu einer verfälschten Beurteilung des zweiten Stimulus führt (*beispielsweise: beim ACR, wenn viele „schlechte“ oder „gute“ Stimuli kommen, wird der Stimulus besser oder schlechter geratet*). Zudem kommt es zu einer Verfälschung der absoluten Beurteilung, wenn innerhalb eines Tests lediglich Stimuli besserer Qualität getestet werden. (*Dessen sollte man sich zumindest dann bewusst sein, wenn man die Ergebnisse desselben Systems mit den Ergebnissen aus einem anderen Setup vergleicht, indem vielleicht auch qualitativ schlechtere Stimuli vorgeführt wurden.*) Abhilfe kann einigermaßen geschaffen werden indem die Bandbreite an Verschlechterungen (*impairments*) erhöht wird [Moe10, Kap5]. Eine randomisierte Reihenfolge im Experiment Design führt dazu, dass dieser Fehler ein zufälliger Fehler wird.

Stimulushäufigkeit (stimulus frequency bias): im Experiment Design soll Sorge getragen, dass sämtliche Stimuli im Test gleich oft abgefragt werden (vgl. Abs. 1.2.6, balanced experiment design). Ansonsten kommt es zu einer Aufweitung in der Beurteilung des häufigeren Stimulus durch die Vpn im Vergleich zu den übrigen Testsignalen [Lie08].

Range Equalizing Bias: Der Bias entsteht wenn für zwei perzeptiv verschiedene Stimulisets (Set 1: Unterschiede klein, Set 2: Unterschiede groß) jeweils der gesamte Wertebereich auf der Skala verwendet wird [Lie08]. In diesem Fall kann Abhilfe durch Anker geschafft werden, die beispielsweise die Endpunkte der Skala repräsentieren.

1.2.9.6 Rahmenbedingungen

Visueller Bias: dieser spiegelt sich in der Repräsentation der Skala wieder. Demnach wird ein Testsignal a priori in diskrete Stufen quantisiert. Gerade im kritischen Umgang mit Datenmaterial als lediglich ordinalskaliert, ist es sinnvoll auf verbale Deskriptoren zu verzichten und lediglich die Endpunkte der Skala zu benennen, um deren Richtung zu definieren.

Cross Modality Bias: Ein zusätzlich zum Hören störender, jedoch nicht zwangsläufig bewusster Empfindungskomplex wird stimuliert und beeinflusst somit das Antwortverhalten. Beispiele hierfür sind ungünstige Lichtverhältnisse, Raumtemperatur, Sitzposition, Gerüche, Haptik des Touchscreens, etc. Diese Fehler werden im Versuchsdesign gerne übersehen oder bleiben unberücksichtigt. In jedem Fall sollten sämtliche Bedingungen nach Möglichkeit für alle Vpn konstant gehalten werden. Es ist beispielsweise sinnvoll, vor jedem Testdurchlauf kurz zu lüften, die Lichtverhältnisse auf angenehmes Licht umzustellen, etc.

1.3 Standardisierung

Wozu Standardisierung?

Standards oder Empfehlungen (*recommendations*) entstehen meist in Kooperation zwischen Wissenschaft und Industrie mit dem Ziel, gemeinsame Ansätze von Experten in eine geeignete Methodik überzuführen und dadurch Objektivierbarkeit der Daten zu erreichen. Die standardisierte Versuchsmethode ermöglicht die Reproduzierbarkeit und Vergleichbarkeit von durchgeführten Versuchen, den daraus gewonnenen Daten und deren Interpretation. Zudem stellt die Standardisierung für ihr Anwendungsgebiet eine wesentliche Erleichterung in Bezug auf Auswahl und Validität der Versuchsmethode, sowie auf den damit verbundenen Aufwand für Planung und Durchführung dar.

Ein weiterer, wesentlicher Vorteil in der Verwendung eines standardisierten Verfahrens ist die Reduktion von Bias, sofern der vorgegebene Anwendungsbereich mit der Aufgabenstellung übereinstimmt. Die beschriebenen Verfahren ebenso wie die Auswahl der Skala für das Antwortformat (sowohl Unterteilung als auch Beschriftung) sind erprobt und sollen nach Möglichkeit genau übernommen werden.

Kann für das gewünschte Vorhaben ein standardisiertes Verfahren gefunden werden, ist die Adaptierung einer Methode mit guter Kenntnis über psychophysikalische Methodik und einer kritischen Auseinandersetzung mit potentiellen Fehlerquellen durchaus möglich. Gerade für Forschungszwecke und Versuche, die effizient geplant werden sollen, wird empfohlen, sich soweit als möglich an die Rahmenbedingungen eines Standardverfahrens zu halten, um Fehler zu vermeiden und zu einer reliablen Aussage zu kommen. Natürlich wird hier auch vorausgesetzt, dass diese Bedingungen dem Geist der Zeit und dem Forschungszweck entsprechen, was mit Sicherheit nicht immer optimal zu kombinieren ist. Die Forscherin möge demnach eine standardisierte Methode als Vorarbeit anderer betrachten, diese kritisch betrachten und niemals leichtfertig verwerfen.

Die akustische Bewertung von Signalen durch das menschliche Individuum wird aus Gründen der Effizienz hinsichtlich Kosten und Zeit oftmals durch eine objektive Methode temporär ersetzt und letztendlich durch subjektive Methoden verifiziert. Diese Gliederung der Methoden anhand der Messeinrichtung in subjektive und objektive Methoden führt jedoch zu Missverständnissen bezüglich der Aussagekraft der jeweiligen Verfahren, weshalb eine Unterteilung nach auditorischen und instrumentellen Methoden sinnvoller erscheint. In auditorischen, Verfahren kommen als Messeinrichtung Versuchspersonen zum Einsatz, wohingegen in instrumentellen Methoden die Messeinrichtung durch ein simulationsbasiertes Modell repräsentiert ist [Cot11, Kap2.1].

Im Rahmen dieser Arbeit wird auf die auditorischen oder subjektiven Methoden näher eingegangen, instrumentelle oder objektive Modelle sind nicht Bestandteil eines Hörversuchs und somit nicht im Umfang enthalten. Eine gute Übersicht über objektive Qualitätsbeurteilung liefert beispielsweise Quackenbush [Qua88].

Ein gut designter und durchgeführter, subjektiver Hörtest ist aus statistischer Sicht jedoch durchaus in der Lage, zuverlässige, reproduzierbare und damit objektive Ergebnisse zu liefern. Eine interessante Darstellung über den teilweise irreführenden Gebrauch der Begriffe subjektiv und objektiv liefert der Artikel von Jules M Rothstein [Rot13].

In einem Standard für subjektive Methoden für Hörversuche sind üblicherweise enthalten:

- Anwendungsgebiet
- Generierung von Signalen
- Rahmenbedingungen wie Raum, Equipment, etc.
- Wahl der geeigneten Population an Versuchspersonen
- Versuchsmethoden und deren Durchführung
- Auswertung der Daten anhand statistischer Analysen

Im Audiobereich sind verschiedene Organisationen wesentlich um Standardisierung bemüht. Dementsprechend wird auf Standards und Papers der ITU (International Telecommunication Union), Audio Engineering Society AES20 und der European Broadcasting Union referenziert. Weitere Organisationen auf diesem Gebiet sind beispielsweise das American National Standards Institute (ANSI S3.1-1999) oder die International Technical Commission (60268-13).

1.3.1 ITU

Die ITU arbeitet in unterschiedlichen Gebieten der Telekommunikation (ITU-T), der Radiokommunikation (ITU-R) und im Entwicklungssektor (ITU-D). Einen Aufgabenbereich stellt die Entwicklung von sogenannten Recommendations dar, die in Forschungsgruppen zu verschiedenen relevanten Themen ausgearbeitet werden. Diese Empfehlungen genießen, sobald sie in Kraft treten, normativen Charakter, werden weltweit anerkannt und können kostenfrei von der Website geladen werden.

In weiterer Folge werden nur Recommendations, die in direktem oder näherem Zusammenhang mit akustischem Versuchsdesign stehen, erwähnt und behandelt.

1.3.1.1 ITU-T

Die ITU-T beschäftigt sich mit Anwendungen im Bereich der Telefonbandbreite (Schmalband 300 – 3400 Hz; Breitband 150 – 7000 Hz) und testet diese ebenso. Ein Fokus liegt dabei auf dem Gebiet der Sprachqualität, Audio und Video werden ebenso behandelt. Die beschriebenen Hörmethoden orientieren sich in diesem Bereich der ITU vorzugsweise an untrainierten Versuchspersonen, um möglichst realistische Ergebnisse erzielen zu können. Es gibt jedoch bereits dokumentierte Versuche [Fra12], die die sinnvolle Anwendbarkeit eines listening panels auch in diesem Einsatzbereich zeigen.

ITU-T P800 (1996)

Die Empfehlung P800, welche Methoden für die subjektive Bewertung von Übertragungsqualität beschreibt, unterscheidet zwischen Konversationstests (*conversation-opinion test*) und Hörtests (*listening-opinion test*) und wurde ursprünglich mit dem Fokus auf Schmalband- und Breitbandcodecs in der Telefonie (analog) erarbeitet.

Während Konversationstests zum Ziel haben, beispielsweise ein Telefongespräch in der Laborsituation möglichst realistisch darzustellen, spielt die exakte Repräsentation der Realität bei Hörtests eine untergeordnete Rolle. Die künstlich generierte Situation bringt aber eine strenge Kontrolle der Rahmenbedingungen und der potentiellen Quellen für Bias mit sich.

ITU-T P830 (1996)

Diese Empfehlung richtet sich an die subjektive Bewertung der Performance von digitalen Schmalband- und Breitbandcodecs (wideband, WB), innerhalb welcher Distortions des Signals mit berücksichtigt werden. Dementsprechend überschneiden sich einige Bereiche mit der P800 und werden erweitert.

1.3.1.2 ITU-R

Dieser Bereich der ITU beschäftigt sich mit dem gesamten hörbaren Frequenzbereich (von 20 Hz bis 20 kHz). Hier sind Methoden, welche sich mit der Bewertung von Audioqualität auseinandersetzen, zu finden. Dementsprechend werden für die Versuchsdurchführung auch trainierte Versuchspersonen bis hin zu EL empfohlen. Auch Audio und Video wird hier behandelt. Im Bereich von subjektiven Hörversuchen ist vorrangig der Bereich BS (*Broadcasting Service - Sound*) und in Bezug auf die Kombination mit Bildern der Bereich BT (*Broadcasting Service - Television*) interessant.

ITU-R BS 1116-1

Nach dieser Empfehlung wurde der Referenzraum am Joanneum Research GmbH geplant und eingerichtet. Abgesehen von den Vorgaben zu den Rahmenbedingungen eines Hörversuchs ist hier auch das Verfahren zur Bewertung von Systemen, die kleine Artefakte produzieren, zu finden. Dieses wird als double blind triple stimulus with hidden reference im methodischen Teil beschrieben.

Zudem wird hier auf Multikanalsysteme (*multi channel systems*) näher eingegangen.

ITU-R BS 1283-1

Diese Recommendation gibt einen schnellen Überblick über die übrigen Empfehlungen im Bereich der Klangqualität (*Sound quality*) der ITU und den zugehörigen Anwendungsbereich. Sie ist sozusagen als Lexikon zu verstehen.

ITU-R BS 1284-1

Die ITU ist als allgemeiner Leitfaden für die Bewertung von Sound quality zu verstehen. Hier sind abzufragende Attribute für Sound quality zu finden.

ITU-R BS 1285

Der (Quasi)Standard ITU-R BS 1285 beschreibt ein Verfahren, welches ein Audiosystem dahingehend überprüft, ob es geringe oder mittlere bis höhere Beeinträchtigungen in Bezug auf die wiedergegebenen Audiosignale aufweist. Dieses Verfahren wird empfohlen, wenn nicht klar ist, in welcher Größenordnung die Artefakte liegen. Grundsätzlich gilt es zu sagen, dass, um ein System mit kleinen Verschlechterungen testen zu können, eine strengere Überwachung der Rahmen- und Testbedingungen notwendig ist, als bei Systemen mit „offensichtlicheren“ Verschlechterungen.

ITU-R BS 1534

Diese Empfehlung kann als Gegenstück zur [1116-1] angesehen werden, da sie Sound Systeme behandelt, die größere Artefakte produzieren. Das MUSHRA-Verfahren ist hier enthalten.

1.3.2 EBU

Die Europäische Rundfunk Union (European Broadcast Union) wurde 1950 unter anderem mit dem Ziel gegründet, ein Nachrichtennetzwerk aufzubauen und darüber hinaus den technischen Fortschritt bei Radio und Film voranzutreiben und zu standardisieren. Heute ist die EBU ein Zusammenschluss von über 70 Sendern (zu denen auch der ORF gehört) aus mehr als 50 Nationen mit Sitz in Genf. Die wohl bekannteste Produktion ist der Eurovision Song Contest.

Technische Empfehlungen sind hier als Technical Review oder Technische Spezifikationen zu finden. Nachfolgend werden die wichtigsten Papers und Dokumente kurz erklärt. Sämtliche beschriebene Dokumente sind im Literaturverzeichnis angeführt und werden innerhalb dieses Kapitels nicht extra zitiert.

1.3.2.1 Technical Specification

Die Spezifikationen, welche hier gelistet sind, beschreiben relevante Aspekte von Hörversuchen mit dem Fokus auf Audio. Dementsprechend sind hier verschiedene Spezifikationen zu finden, die die Kontrollvariablen für Mono-, Stereo, Multikanalwiedergabe listen und somit die Empfehlungen der ITU sinnvoll ergänzen.

Bezüglich der verwendeten Antwortskalen sei hier allerdings vermerkt, dass diese sich von jenen der ITU insofern unterscheiden, als sie 6 anstatt von 5 Kategorien verwenden.

Tech 3276 – 2nd edition, 1998

Das Dokument beschreibt die Rahmenbedingungen für die Bewertung von Klangmaterial bei Mono oder Stereo- Wiedergabe über Lautsprecher. Schallfeldparameter sind hier ebenso enthalten, wie die Konstruktion eines Referenzraums und die Spezifikationen für Monitorlautsprecher und Monitorkopfhörer.

Tech 3276E – Supplement I, 2004

Die Ergänzung zur Spezifikation 3276 beschreibt zusätzliche relevante Aspekte in den Abhörbedingungen für Multikanalwiedergabe.

Tech 3286E, 1997

Inhalt der Empfehlung sind Rahmenbedingungen ebenso wie Beurteilungsmethoden für die Qualität von Klangmaterial mit dem Fokus auf Musikwiedergabe. Attribute sind hier ebenso gelistet.

Tech 3286E – Supplement 1, 2000

Die Ergänzung legt den Fokus zusätzlich auf Multikanalwiedergabe.

2 Methoden für subjektive Hörtests

Bei den nachfolgend beschriebenen Methoden und Anwendungsmöglichkeiten für Hörtests handelt es sich einerseits um standardisierte Verfahren aus den Recommendations (vgl. Abs. 1.3) und andererseits um für spezifische Anforderungen abgeänderte Anweisungen, die nicht standardisiert sind. Es werden die Art der Methode, ihre Anwendung und der sinnvolle Einsatzbereich erklärt, wodurch der Versuchsleiter einen prägnanten Überblick erhält, welcher ihm die Wahl einer geeigneten Methode erleichtern und auf etwaige Problemstellungen hinweisen soll.

Eine Unterteilung der Methoden erfolgt grundsätzlich nach der Bewertung von Sprach- und von Audioqualität und wird weiter gegliedert in Verfahren, die einen Stimulus bewerten und jene, die mehrere Stimuli zueinander in Bezug setzen. In der Benennung der Verfahren und Methoden wurde bewusst auf den Versuch der deutschen Übersetzung der Eigennamen verzichtet, zumal die zugrunde liegende deutschsprachige Literatur nicht zu einheitlichen Begrifflichkeiten gefunden hat und dies nur zu Verwirrungen führt.

2.1 Der Qualitätsbegriff

Das Modell von Blauert und Jekosch (vgl. Abb. 10) zur Definition von Klangqualität basiert auf der epistemologischen Darstellung von Realität und Wahrnehmung nach dem Perzeptionismus oder radikalem Konstruktivismus und leitet analog zum Trialismus (Seele, Körper, Verstand) drei Wahrnehmungsobjekte (Gefühle, Dinge, Konzepte) ab [Bla12]. Konzepte, beispielsweise Ideen oder Gedanken, fungieren dabei als Deskriptoren von Gefühlen und Dingen, indem sie auf deren Eigenschaften hindeuten. Sie spielen bei der Qualitätsbewertung und -beurteilung eine zentrale Rolle. Ausgehend von diesem Ansatz werden vier Qualitätsebenen in eine Rangfolge gebracht:

- Auditive Qualität (Klangqualität an sich): klassische Psychophysik
- Aural - szenische Qualität (Präsentationsqualität)
- Akustische Qualität (Übertragungsqualität)
- Aurale Kommunikationsqualität (Produktklangqualität)

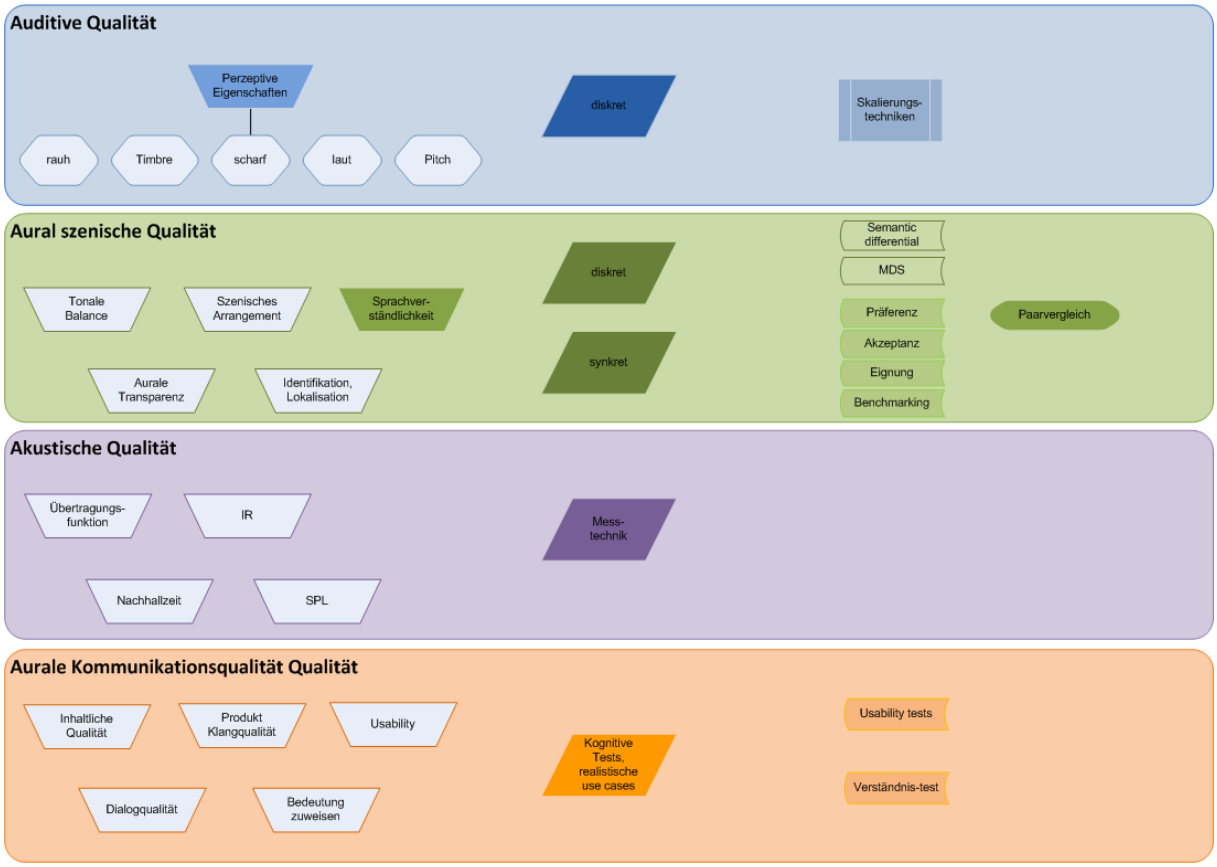

Abb. 10: Layer Modell zum Qualitätsbegriff [Bla12]

Das Layer Modell veranschaulicht die Vielschichtigkeit des Qualitätsbegriffs.

Die Ebene der auditiven Qualität beschäftigt sich mit den auditorischen Ereignissen (*auditory events*, vgl. Abs. 1.2.3.1, Filtermodell), welche analytisch oder diskret über ihre charakteristischen Eigenschaften (z.B. pitch, timbre, Rauigkeit) erfasst werden. Die zugrundeliegende Idee der Psychoakustik hier ist, eine fehlerfreie (*unbiased*), da komplett analytische, Beurteilung der Versuchspersonen zu erhalten, die die auditorische Peripherie repräsentiert. Dazu ist synthetisches Klangmaterial wie Rauschen, Sinus, Tonbursts nötig. Diese Ebene ist vorwiegend für den Auswahlprozess (elicitation methods) neuer Attribute relevant.

Die nachfolgend behandelten Methoden, welche versuchen, Qualität von Sprache und Audio quantitativ erfassbar zu machen, sind überwiegend der zweiten Ebene des Modells, der aural szenischen Qualität, die sich mit der Präsentationsqualität auseinandersetzt, zuzuordnen. Üblicherweise befindet sich ein Individuum, abgesehen von dem Ereignis des „etwas-zum-ersten-Mal-im-Leben-Hören“, in einem integrativen Status des Hörerlebnisses; aus der Holistik von charakteristischen Eigenschaften eines Signals werden aurale Objekte gebildet, die wiederum Szenarios (*aural scenes*) formen. Perzeptive Effekte (Präzedenzeffekt, Cocktail-Party-Effekt, etc.) werden ebenso berücksichtigt, wie eine Wechselwirkung zwischen Modalitäten (z.B. Seh- und Hörsinn).

Die auralen Szenarios werden durch qualitätsbestimmende Faktoren wie tonale Balance, Raumeindruck, Tiefenstaffelung, Präsenz, ease of listening, perzeptive Plausibilität bestimmt und deren Bewertung wird grundsätzlich in zwei methodische Ansätze unterteilt (vgl. Abs. 1.2.3.1)

- Perzeptive Analyse auraler Objekte und Szenarios
    - Semantisches Differential
    - Multidimensional scaling
- Bewertung des Gesamteindrucks
    - Präferenztest
    - Akzeptanztest
    - Benchmarking gegen eine Zielvorgabe

Die dritte Ebene des Modells, welche sich mit akustischer Qualität beschäftigt, repräsentiert die physikalisch erfassbaren Parameter wie Schalldruckpegel, Nachhallzeit usw.

Ebene vier befasst sich mit einem globalen Qualitätsbegriff, der nicht mehr zwangsläufig eine Aussage über die tatsächliche, physikalische Qualität des Signals liefert, sondern Qualität im Kontext oder auch zweckmäßig betrachtet. Als Beispiel liefern die Autoren Sound Design, wo nicht mehr die Abtastrate und Auflösung eines Signals im Vordergrund stehen, sondern beispielsweise die intuitive Aussagekraft als unterstützender auditorischer Faktor eines Displays. Erfüllt das Signal also seinen Zweck und lässt ein Produkt „glänzen“, wird dieses Signal gut bewertet. In diesen Bereich fallen demnach Tests zur Verwendbarkeit eines Gesamtprodukts, Usability, Verständnis (als Ergebnis einer geführten Konversation [Moe10, Kap5.4]), psychologische Tests zu kognitiven Faktoren (unter welchen Bedingungen wird ein Sound für gut befunden), etc.

2.2 Methoden zur Beurteilung der Sprachqualität

Sprachqualität (nach Ute Jekosch):“Ergebnis der Beurteilung der Gesamtheit aller erkannten und benennbaren Merkmale und Merkmalswerte einer betrachteten Sprechprobe bezüglich ihrer Eignung, die Gesamtheit der erkannten und benennbaren Merkmale und Merkmalswerte von individuellen Erwartungen und/oder gesellschaftlichen Forderungen und/oder sachgerechten Erfordernissen zu erfüllen.“[Fel12, Kap5.1]

Sprachqualität ist ein vielfach definierter Begriff. So versteht man darunter beispielsweise:

- eine Zusammenfassung verschiedener Merkmale von Sprache: Verständlichkeit, Lautheit, Klarheit, Verstehbarkeit…
- und ergänzend eine Erfassung indirekter Merkmale: Höranstrengung, Zufriedenheit, Lästigkeit, Annehmlichkeit…

Je nach Fragestellung und Anwendungsbereich der Qualitätsaussage werden spezifische Merkmale des zu testenden Sprachsignals, wie zum Beispiel die Sprachverständlichkeit als ein wesentlicher Teilaspekt der Qualität, zusätzlich in den Fokus gerückt. Genauso ist es aber möglich, lediglich nach der absoluten Qualität oder Gesamtqualität, dem Gesamteindruck als affective measurement zu fragen, der sämtliche Einzelmerkmale für sich unberücksichtigt lässt und eine integrative Betrachtung des Signals wünscht.

Unter Sprachverstehbarkeit (speech comprehensitivity) versteht man die Eigenschaft, Inhalte korrekt wiederzugeben. Der Begriff definiert die Qualität des (Übertragungs-) Systems ohne Kontextbezug. Er ist nicht zu verwechseln mit Sprachverständlichkeit (speech intelligibility), die den Kontext oder die Bedeutung miterfasst und sich auf den Zuhörer bezieht. Bereits 1929 unterschieden Fletcher und Steinberg zwischen Artikulation (korrekte Identifikation von sinnfreien, einsilbigen Worten; Verstehbarkeit) und Sprachverständlichkeit von übertragener Sprache [Fle29].

Zur Ermittlung der Verstehbarkeit wird jegliche Sinnhaftigkeit aus den Teststimuli entfernt (beispielsweise bei Tests mit Logatomen: die Silbenverständlichkeit ist eigentlich eine Silbenverstehbarkeit), wohingegen zur Bewertung der Sprachverständlichkeit Worte und Zahlen verwendet werden. In der Literatur wird häufig nicht zwischen den Begriffen unterschieden [Fel12, Kap5.1]. Dabei ist die Diskriminierbarkeit aus Sicht der Versuchsperson wesentlich schwieriger für Verstehbarkeits- als für Verständlichkeitstests und liefert demnach auch Aussagen verschiedener Gewichtung. Der Hörer referenziert in der Beantwortung bei sinnlosen Logatomen auf Phoneme (Verstehbarkeit), bei „echten“ einsilbigen Worten hingegen auch auf „echte“ sinnvolle Worte (Verständlichkeit). Dementsprechend ist die Betrachtung der Verstehbarkeit als Bestandteil der Verständlichkeit logisch nachvollziehbar. Für weitere Informationen zur Bewertung von Sprachverständlichkeit und Sprachverstehbarkeit sei auf Jekosch [Jek10], Quackenbush [Qua88] etc. verwiesen.

Der gegenüber früheren Definitionen von Qualität von Sprache und auch von Klang wesentlichste, neue Aspekt ist das Messorgan Mensch als unersetzbares Instrument in der Wahrnehmung, Reflexion und Beurteilung von Signalereignissen, wodurch Qualität als solche erst existiert [Moe10, Kap1.2]. Qualität von Sprache und Klang entsteht somit erst durch die Bewertung einer Momentaufnahme als Reaktion auf die Prägung und Erwartungshaltung des Individuums.

Nachfolgend bildet die ITU-T P800 [800], welche die grundlegenden Verfahren beinhaltet, den Ausgangspunkt für die anderen relevanten, standardisierten und adaptierten Verfahren. Es ist dennoch anzumerken, dass die Empfehlung unter einem bestimmten Fokus erarbeitet wurde (vgl. Abs. 1.3.1.1), den es im Rahmen dieses Kapitels zu erweitern gilt. Dieser Aspekt betrifft beispielsweise die Erweiterung des Anwendungsgebietes auf Audiosignale, oder auf spezifische Attribute eines Signals. Umgekehrt werden Verfahren, die dem Audiobereich zugeschrieben sind, auch für Superbreitbandcodecs Einsatz finden.

2.2.1 Methoden zur Bewertung von einem Stimulus (ohne vergleichende Referenz)

Die Methoden in diesem Kapitel beschreiben standardisierte Verfahren zur Beurteilung der Gesamtqualität eines Signals, wenn eine Referenz nicht verfügbar ist oder ein Vergleich mit einer Referenz nicht gewünscht ist. Genau in der Tatsache, nämlich eine Referenz nicht erzeugen oder zur Verfügung stellen zu müssen, liegt bei vielen Anwendungen der große Vorteil, wenn man eine der nachfolgenden Methoden wählt.

2.2.1.1 Absolute Category Rating, (ACR) [800, Annex B]

Der ACR - Test nach ITU-T P.800 führt als standardisierter Hörtest zu einer Bewertung der Gesamtqualität (Gesamteindruck) des Testsignals. Spezifische Attribute, also Teilaspekte des Signals, werden dabei, abgesehen von der Möglichkeit der Lautheitsbewertung, nicht beurteilt.

Ursprünglicher Anwendungsbereich dieses Verfahrens sind analoge Schmalband Codecs für die Telefonie.

Es handelt sich um einen Versuch, bei dem die Testperson Stimuli, ohne Vergleichsreferenz, anhand einer vordefinierten Skala absolut bewerten soll.

Fragestellung und Anwendungsbereich

Der Qualitätsbegriff beschreibt in diesem Zusammenhang die Qualität der gehörten Sprache während des Telefonierens, oder während der Anwendung verschiedener Codecs, die miteinander verglichen werden. Aus Sicht der Versuchsperson ist die Bewertung eine des Gesamteindrucks. Die Fragestellung richtet sich nach dem zu übertragenden System.

Anwendungsbeispiele:

- Gerichtete Zusammenhangshypothese: *Es soll geprüft werden, ab wann ein System zunehmend schlechter in seiner Gesamtqualität beurteilt wird. Der Zusammenhang wird zwischen dem System und Babble Noise hergestellt, die Richtung des Zusammenhangs, wonach die Qualität mit zunehmendem Rauschen sinkt, ist bekannt. Die Hypothese kann lauten: Mit zunehmendem Rauschanteil im Signal wird das System in seiner Gesamtqualität besser, also mit einem höheren Wert auf der Skala, beurteilt. Der Begriff „besser" ist in der Hypothese spezifiziert. Die Frage an die Versuchsperson lautet lediglich: Bewerten Sie die Stimuli hinsichtlich ihrer Gesamtqualität. Dabei soll in den anfänglichen Instruktionen der Begriff Gesamtqualität spezifiziert werden.*

- Ungerichtete Unterschiedshypothese: *Mehrere Übertragungssysteme werden miteinander hinsichtlich ihrer Höranstrengung verglichen. Der Unterschied besteht zwischen den Systemen, allerdings ist nicht bekannt, welches System die besten Bewertungen im Hörkomfort erhalten wird. Die Hypothese H_0 könnte lauten: Es gibt keinen Unterschied zwischen den Übertragungssystemen hinsichtlich einer Bewertung der Höranstrengung. Die Teilnehmerinnen werden über die Art der Übertragungssysteme in Unkenntnis gelassen und bewerten die dargebotenen Stimuli bezüglich ihrer subjektiven Höranstrengung.*

Formate für Antwortskalen

Im Standard sind drei verschiedene Skalen angeführt, die im selben Wortlaut verwendet werden sollen. Es handelt sich dabei um eine kategoriale Skalierungstechnik (Ordinalskala) , die laut ITU jedoch als Quasi-Intervallskala interpretiert werden darf; Jekosch beispielsweise betrachtet dies kritisch. (vgl. Abs. 1.2.4.3, Skalenniveau der Skalierungstechniken).

Listening-quality scale

Skala	Verbale Deskriptoren	
5	Excellent	**Sehr gut**
4	Good	**Gut**
3	Fair	**Annehmbar**
2	Poor	**Mäßig**
1	Bad	**Schlecht**

Tab. 12: Listening Quality Scale, deutsche Begriffe nach [Lie08]

Listening-effort scale

Skala	Verbale Deskriptoren	
5	Complete relaxation possible, no effort required	**Völlig entspannt, keine Anstrengung notwendig**
4	Attention necessary; no appreciable effort required	**Aufmerksamkeit notwendig, aber keine Anstrengung**
3	Moderate effort required	**Mäßige Anstrengung notwendig**
2	Considerable effort required	**Beträchtliche Anstrengung notwendig**
1	No meaning understood with any feasible effort	**Unverständlich, trotz größter Anstrengung**

Tab. 13: Listening-effort scale

Loudness-preference scale

Skala	Verbale Deskriptoren	
5	Much louder than preferred	**Viel lauter als angenehm**
4	Louder than preferred	**Lauter als angenehm**
3	Preferred	**Angenehm**
2	Quieter than preferred	**Leiser als angenehm**
1	Much quieter than preferred	**Viel leiser als angenehm**

Tab. 14: Loudness-preference scale

Man beachte hier, dass die Skalenmitte keine neutrale Stellung darstellt.

Versuchspersonen

Bei der Bewertung von Sprachsignalen wird der naive Hörer empfohlen [800, B4.4]. Dieser soll nicht im Bereich der Sprachkodierung oder der Bewertung von Übertragungssystemen tätig sein. Außerdem darf er in den vergangenen sechs Monaten an keinem subjektiven Test teilgenommen haben und im vergangenen Jahr nicht an einem Hörtest. Entscheidend ist jedenfalls, dass den Versuchspersonen der Sprachkorpus nicht bekannt sein soll, zumindest aber, dass sie noch nie dieselben Sätze aus dem Korpus gehört haben, um zu gewährleisten, dass es zu keiner besseren Bewertung aufgrund von Wiedererkennungseffekten kommt.

Stimuli

Das verwendete Sprachmaterial besteht aus einer Liste mit kurzen, sinnvollen Sätzen mit einer Gesamtdauer von je zwei bis drei Sekunden. Es ist darauf zu achten, dass sowohl männliche, als auch weibliche Stimmen gleichermaßen verwendet werden. Die Generierung dieses Sprachsets ist zu dokumentieren. (siehe auch Teststimuli)

Die Reihenfolge dieser Satzliste wird für jede Versuchsperson randomisiert.

Je nach Anwendung werden zwei bis fünf Sätze zu einer Gruppe als repräsentatives Sprachbeispiel (*speech sample*) zusammengefasst. (Achtung: Nachdem die Satzliste randomisiert ist, sind folglich auch die Speech samples zufällig angeordnet. Innerhalb eines Samples bleiben aber die Sätze dieselben.) Fünf bis zehn dieser Gruppen ergeben dann einen Durchlauf (*run*).

Versuchsablauf

Allgemeines zum Versuchsablauf kann in Abs. 1.2.8.5 nachgelesen werden.

In der Eingewöhnungsphase wird die Vorbereitungsliste vorgespielt, welche für alle Teilnehmerinnen dieselbe ist. Die Teilnehmerin bewertet diese Stimuli nach den Vorgaben. Ist der Durchlauf beendet, werden eventuelle Fragen bezüglich des Ablaufs oder der Instruktionen geklärt, nicht aber technische Details oder Hintergründe des Experiments.

Es folgt der eigentliche Versuch, der aus mehreren Teilen besteht, die sich wiederum in Trials unterteilen. Innerhalb jedes Trials wird jeder Testsatz einmal vorgespielt und von der Vpn bewertet.

Zu beachten

- In den Instruktionen an die Versuchsperson (Eingewöhnungsphase) darf nicht erwähnt werden, ob die Endpunkte der Skala durch Stimuli im Hörtest repräsentiert sind oder nicht [800, B4.6]. Es werden entgegen der Meinung anderer Publikationen [Qua88, Kap2.1], [Rot69] auch keine Stimuli mit zugehöriger Bewertung präsentiert. Die Eingewöhnungsphase dient lediglich der Versuchsperson als Orientierungshilfe, in welcher Größenordnung sich die zu beurteilenden Stimuli befinden werden.

  Die Eingewöhnungsphase ist nicht mit der Trainingsphase von Experten zu verwechseln.

- Damit die Bewertung tatsächlich absolute Ergebnisse liefert, ist darauf zu achten, die qualitative Bandbreite der Stimuli sinnvoll abzudecken. (vgl. Abs. 1.2.9, Kontextbias)

- Die Nummerierung der Skala darf den europäischen Verhältnissen (1=sehr gut, 5=nicht genügend) durchaus angeglichen werden, sofern dies für alle Versuchspersonen im Test erfolgt und in der anschließenden Analyse die Richtung der Skala auch mit berücksichtigt wird.

Ergebnis

Ergebnis des Experiments ist der MOS (mean opinion score) als arithmetischer Mittelwert. Dieser hat für jede der drei Skalen ein Suffix:

- Listening quality: MOS
- Listening effort: MOS_{LE}
- Loudness preference: MOS_{LP}

Der ACR-Test tendiert im Vergleich zu anderen Testmethoden zwar zu einer geringeren Empfindlichkeit aufgrund der fehlenden Vergleichsprobe, ist dafür aber relativ einfach und zeiteffizient durchzuführen und auszuwerten. Aufgrund seines Bekanntheitsgrades erfreut er sich auch großer Beliebtheit, nicht zuletzt weil die Wahrscheinlichkeit, vergleichbare Ergebnisse zu finden, höher ist als bei anderen Verfahren.

Ist das zu testende System von sehr hoher Qualität, tendiert der ACR dazu, wenig sensitiv zu reagieren. In diesem Fall wird zur Verwendung des DCR geraten (vgl. Abs. 2.2.2.1).

Tendenzen sind mit dieser Methode schnell erkennbar. Getestet werden kann dabei nicht nur die Qualität von Sprache, sondern auch von Musik und Klängen, ebenso wie die Höranstrengung bei Verständnis von Sprache oder eine Lautheitspräferenz in einer vorgegeben Hörsituation.

2.2.1.2 Quantal Response Detectability Test (QRDT), [800, Annex C]

Der QRD-Test ist als standardisierter Test der P800 eher unbekannt. Getestet wird eine analoge Eigenschaft x eines Signals die in Abhängigkeit einer anderen Eigenschaft y abgefragt wird.

Dabei geht es um die Wahrnehmbarkeit eines Störanteils im Signal, vorzugsweise im analogen Bereich.

Fragestellung und Anwendungsbereich, Funktionsweise

Dieser Test kann auf zwei verschiedene Arten ausgelegt werden, weshalb auch zwei grundsätzlich verschiedene Skalen zur Verfügung stehen.

Detectability-opinion scales

Einerseits kann die Aufgabenstellung an die Versuchsperson sein, auf einer 3-Punkte Skala zu beurteilen, ob eine Eigenschaft störend, detektierbar oder nicht detektierbar ist. Dabei richtet sich die Fragestellung mehr an die Art der Eigenschaft. Das bedeutet, es ist weniger von Interesse, in welchem Ausmaß (Quantität) sich die Vpn gestört fühlt, als ob sie sich überhaupt gestört fühlt. *Beispiel: Die qualitative Aussage bei einem Echo ist somit eine ganz andere: angenommen zwei Vpn nehmen das Echo beide wahr: während sich die eine gestört fühlt, stört es den anderen nicht und er detektiert das Echo lediglich.*

7-stufige Skala

Speziell wenn die Anwendung Rauschen im Signal erfassen soll, wird eine Skala mit sieben Kategorien verwendet. In diesem Fall liegt der Fokus wiederum ein Stück weit näher an der quantitativen Untersuchung des Störeinflusses.

Geeignet ist diese Methode um Grenzwerte bzw. Grenzbereiche einer bestimmten Menge (auch prozentueller Natur) des zugehörigen Attributs zu ermitteln. Als Beispiel sei der Level genannt, über dem eine einzelne Frequenzinterferenz eine bestimmte Wahrscheinlichkeit besitzt, detektierbar, also hörbar zu sein [800, Kap6]. Ein anderes Beispiel ist die Anwendung der Methode, um die Wahrscheinlichkeit des Auftretens von verständlichem Übersprechen innerhalb eines bestimmten Bereichs zu bestimmen.

Anwendung findet die Methode bei Wahrnehmbarkeitstests wie zum Beispiel Echo, Hall, Nebengeräusche, Crosstalk (Übersprechen).

Formate für Skalen

Detectability-opinion scales

Skala	Verbale Deskriptoren		
A	Objectionable	**Störend**	**Verständlich**
B	Detectable	**Detektierbar**	**Detektierbar**
C	Not detectable	**Nicht detektierbar**	**Nicht detektierbar**

Tab. 15: Detectability-opinion scales

Der mittlere Ankerpunkt B, detektierbar ist dabei folgendermaßen zu verstehen: der Störeinfluss wird bemerkt, aber er stört nicht bzw. nicht verständlich (beispielsweise bei Crosstalk: hier wird nach der Wahrscheinlichkeit getestet, mit der Crosstalk in einem bestimmten Wertebereich klar erfassbar ist [800, Kap6]).

Beide Skalen enthalten zwei dichotome Attributpaare. Anhand der ersten drei Punkte Skala kann dies gut veranschaulicht werden:

- Störend - detektierbar: detektierbar bedeutet hier, dass ein Störsignal erkannt wird, aber gleichzeitig nicht stört. Daraus folgt:

  Störend - nicht störend

- Detektierbar - nicht detektierbar

Anhand der Auswahl der Ankerpunkte ist erkennbar, dass es sich um Attribute mit Präferenzcharakter handelt. Streng genommen, ist die Methode der Literatur entsprechend

als Beurteilung des Gesamteindrucks zu werten [Bec06][Jek10]. Allerdings wird hier ein spezifischer Aspekt des Signals betrachtet und lediglich nach einem präferenzierenden Charakter bewertet. Demnach ist die Beurteilung jene eines Attributs, wenn auch nicht eines perzeptiven Attributs im ursprünglichen Sinn, des Signals innerhalb einer affektiven Methodik.

Skala	Verbale Deskriptoren	
A	Inaudible	**Nicht hörbar**
B	Just audible	**Gerade hörbar**
C	Slight	**Leise**
D	Moderate	**Mäßig**
E	Rather loud	**Eher laut**
F	Loud	**Laut**
G	Intolerable	**Zu laut**

Tab. 16: 7-stufige Skala zur Bewertung von Störgeräuschen

Die Skala wird ähnlich der Loudness Preference Scale (vgl. Abs. 2.2.1.1, Formate für Antwortskalen) als quasikontinuierlich betrachtet.

Versuchspersonen

Grundsätzlich wird bei sämtlichen Bewertungen des Gesamteindrucks von Sprache der naive Hörer empfohlen. Der Grund hierfür liegt in einer sinnvollen Repräsentation der Population Endverbraucher. Handelt es sich jedoch um perzeptiv komplexere Aufgabenstellungen bezüglich der zu untersuchenden Störung, ist es ratsam, Experten heranzuziehen.

Gerade bei Anwendung dieses Verfahrens möchte man zum Beispiel den Grad an Störung eruieren, der noch wahrnehmbar ist. Diese Grenze ist bei Auswahl des ELP strikter.

Stimuli

Als Testsignal kommen je nach Anwendungsbereich beispielsweise Sprache im Rauschen, Sprache mit Crosstalk, oder Musik zum Einsatz.

Versuchsablauf

Vor jedem Durchlauf wird ein Stimulus präsentiert, der die Störung x im Signal offenhörbar erkennen lässt. Anschließend wird der Hörpegel in konstanten Schritten je Sample verringert und bewertet.

Zu beachten

Im Versuchsablauf ist es empfehlenswert, zu Beginn jedes Durchlaufs ein Signal zu präsentieren, welches zweifellos die zu testende Eigenschaft x repräsentiert.

Um einen Bezugspunkt herstellen zu können, wird zusätzlich zu diesem Versuch bei den Testpersonen ein audiometrischer Test durchgeführt. Die Ergebnisse des QRDT werden dann in Bezug auf die Hörschwelle betrachtet.

Ergebnis

Die Ergebnisse werden für die beiden Skalentypen unterschiedlich betrachtet.

Ergebnisse der drei-Punkte Skalen sollen grundsätzlich nicht, wie die 5-Kategorien-Skala, als kontinuierlich betrachtet werden, da nicht davon ausgegangen wird, dass der Abstand zwischen den Punkten konstant ist (Intervallskala) oder sich die Punkte zumindest in einer Rangfolge zueinander befinden (Ordinalskala). Viel hilfreicher ist die getrennte Betrachtung und Auswertung der Wahrscheinlichkeiten der beiden dichotomen Paare als Funktion eines globalen Parameters (Hörpegel). In Ausnahmefällen werden den drei Stufen dennoch Zahlen (2,1,0) zugeordnet.

Die quasikontinuierliche, 7-stufige Skala, wird über den mean opinion score (vgl. Abs. 2.2.1.1, Formate für Antwortskalen) und anschließende ANOVA ausgewertet.

In beiden Fällen ist es möglich, aufgrund der Abstufungen des Störanteils in den Teststimuli, eine Mithörschwelle zu ermitteln oder genauer gesagt, eine Schwelle, ab der ein Störeinfluss eben lediglich noch detektiert, aber von dem Panel nicht mehr als störend wahrgenommen wird.

2.2.1.3 Kombination mehrerer Skalen in einem Test [835]

Ein beliebiger Hörversuch kann aus mehreren Durchläufen bestehen, innerhalb derer verschiedene Aspekte des Stimulus betrachtet werden. Nachfolgende Methode nach ITU-T P.835 [835], sowie P.835 Amendment1 [835-1] verwenden drei verschiedene Skalen zur Bewertung von Sprachqualität in Kombination mit Noise Suppression Algorithmen.

Fragestellung und Anwendungsbereich

Um den Qualitätsbezug vielschichtiger darzustellen, wird die Versuchsperson neben der Frage zur Gesamtqualität zusätzlich zur Qualität des Sprachsignals allein und zur Qualität des Hintergrundrauschens befragt. Die Kombination dieser drei Beurteilungen soll ein repräsentatives Ergebnis zu Aussagen über die subjektive Verbesserung des Sprachsignals im Rauschen geben.

Anwendung finden die Methoden beispielsweise bei Freisprechanlagen, Handys, Headsets beispielsweise in Callcentern, aber auch beim Einsatz von Überwachungsanlagen.

Beispiel Monitoring: Mit diesem Setup ist es auch möglich, das Rauschen über mehrere Einzelsprecher, die wiederum ihrerseits unterschiedlich stark verrauscht sind, zu repräsentieren und so ein Monitoring zu simulieren, bei dem der Hörer verschiedene Sprachsignale simultan möglichst gut verstehen soll. Dabei können durchaus Versuchspersonen zum Einsatz kommen, die in diesem Bereich arbeiten und somit diese Art des selektiven Hörens gewohnt sind. Eine Fragestellung könnte sich nach der Qualitätsverbesserung durch Noise Suppression richten. Ebenso könnten spezifische Klangveränderungen, Lautstärkenveränderungen oder sogar Sprachverständlichkeit getestet werden, wenn man die Methode sinnvoll mit anderen Verfahren koppeln würde. Der Versuch wäre dann allerdings nicht mehr standardisiert, was ja im Forschungsbereich nicht vorrangig ist.

Formate für Antwortskalen

Alle drei Skalen bestehen aus 5 Kategorien.

Skala	Verbale Deskriptoren	
5	Not distorted	**Nicht verzerrt**
4	Slightliy distorted	**Geringfügig verzerrt**
3	Somewhat distorted	**Etwas verzerrt**
2	Fairly distorted	**Ziemlich verzerrt**
1	Very distorted	**Äußerst verzerrt**

Tab. 17: Skala für die Bewertung des Sprachsignals

Skala	Verbale Deskriptoren	
5	Not noticeable	**Nicht bemerkbar**
4	Slightly noticeable	**Geringfügig bemerkbar**
3	Noticeable but not intrusive	**Bemerkbar jedoch nicht aufdringlich**
2	Somewhat intrusive	**Etwas aufdringlich**
1	Very intrusive	**Sehr aufdringlich**

Tab. 18: Skala für die Bewertung des Hintergrundrauschens

Die Skala zur Beurteilung der Gesamtqualität ist die Listening Quality Scale des ACR, vgl Tab. 12.

Versuchspersonen

Zumindest 32 naive Hörer, die in den letzten drei Monaten an keinem Hörtest teilgenommen haben, werden benötigt.

Stimuli

Stationary

Basis für den Test bilden kurze, deutsche Sätze mit Bedeutung. Dabei kommen wiederum weibliche und männliche Sprecher gleichermaßen vor.

Die weiterverarbeiteten Stimuli, die nun aus dem Sprachsignal und einem Rauschanteil bestehen, werden in zwei Arten innerhalb des Versuchs verändert:

- Zur Beurteilung des Sprachanteils wird die MNRU [810] variiert und der SNR konstant gehalten.
- Zur Beurteilung des Rauschanteils wird der SNR variiert und die MNRU konstant gehalten.

Nonstationary [835-1]

Je nach Anwendungsfall wird der Noise Suppressor gegenüber dem Kopf-Torso-Mund-Simulator entsprechend ausgerichtet. Abtastrate und Bandbreite für die Aufnahmen sind entsprechend zu wählen:

$$\begin{aligned} f_s &= 8kHz \\ B &= 300 - 3400Hz \end{aligned} \quad (71)$$

Für Schmalband-, und für Breitband-Suppressoren:

$$\begin{aligned} f_s &= 16kHz \\ B &= 100 - 7000Hz \end{aligned} \quad (72)$$

Weiterführende Informationen über Aufnahmebedingungen (z.B. 4 Lautsprecher plus Subwoofer für die Aufnahmen) der Sprach- und Rauschquellen sind dem Standard [835-1, Kap4-5] zu entnehmen.

Versuchsablauf

Stationary

Ein Trial besteht aus drei Samples. Jedes Sample wird dabei durch einen anderen Satz repräsentiert, dauert ungefähr 4 s und setzt sich zusammen aus:

- 1 s Hintergrundrauschen
- 2 s Sprache im Rauschen
- 1 s Rauschen

Jedem Sample folgt eine Pause für die Beurteilung.

Die Versuchsperson beurteilt die Hälfte aller Trials im Versuch nach dem Fokus Sprachsignal – Hintergrundrauschen - Gesamtqualität (Sprachsignal und Rauschen) und die andere Hälfte der Trials nach Hintergrundrauschen – Sprachsignal - Gesamtqualität.

Nonstationary

Ein Sample dauert zumindest 8 s und besteht aus folgenden Teilen:

- 1 s Hintergrundrauschen
- 2 s Sprache und Rauschen
- 2 s Rauschen
- 2 s Sprache und Rauschen
- 1 s Hintergrundrauschen

Die Samples werden mit jeder anzuwendenden Rauschquelle bei zumindest drei SNR-Stufen abgefragt, bei 12 dB, 6 dB und 0 dB.

Zu beachten

Stationary

Die Dauer eines Samples darf der sinnvollen Durchführung nach angepasst werden. Es ist jedenfalls Sorge zu tragen, dass alle Samples im Design ungefähr gleich lange dauern und dass sich die gewählte Dauer im Rahmen von 4 – 8 s bewegt.

Nonstationary

Hat man einen nicht stationären noise suppression algorithm, ist das Verfahren etwas anders durchzuführen. Dabei ist es wichtig, das nicht verrauschte Sprachsignal ohne Suppressor-Einfluss ebenfalls abzufragen, um eine konsistente Beurteilung der Versuchspersonen nachweisen zu können. Auch die SNR-Schritte dürfen weiter adaptiert werden, wenn das für den speziellen Fall für sinnvoll erachtet wird.

Beschrieben werden Rauschquellen wie babble, pink, music, voice, car und street noise (das mag auch für andere Anwendung nützlich sein).

Ergebnis

Je nach Fragestellung kann der t-Test, die ANOVA, der Test nach Tukey oder die MANOVA angewendet werden.

2.2.2 Vergleichende Methoden

Die nachfolgend beschriebenen Verfahren basieren auf einem vergleichenden Konzept. Dabei steht beispielsweise eine Referenz zur Verfügung, die die Versuchsperson in ihrer anschließenden Qualitätsbeurteilung auf einer Skala unterstützen soll. Oder der Versuchsteilnehmer wird vor die Aufgabe gestellt, sich für eines von zwei oder auch mehreren Signalen zu entscheiden (*forced choice*). Der Entscheidungsprozess kann zudem noch mit einer Beurteilung auf einer Skala verknüpft sein, um zu erfahren, in welchem Ausmaß der Stimulus nun den zweiten hinsichtlich des interessierenden Attributs übertroffen hat.

2.2.2.1 Degradation Category Rating (DCR) [800, Annex D]

Die Methode vergleicht die Stimuli des Testsystems mit einer hochqualitativen, der Versuchsperson bekannten, Referenz. Die Entartung des Signals wird auf einer Skala mit fünf Punkten von unhörbar bis sehr anstrengend bewertet.

Geeignet ist der DCR wenn die Entartungen oder Artefakte klein sind, also das zu beurteilende System von sehr hoher Qualität ist.

Fragestellung und Anwendungsbereich

DCR eignet sich gut, wenn das Quellmaterial von niedriger absoluter Qualität ist (beispielsweise Sprache mit Hintergrundrauschen) oder auch, wenn die digitalen Entartungen (*impairments*) sehr klein sind [P830].

Anwendung findet die Methode beispielsweise zur Evaluierung verschiedener, jedoch ähnlicher, digitaler Sprachverarbeitungsalgorithmen. Oder auch als Hilfsmittel zur Systemoptimierung, wenn bei ACR herausgefunden wurde, dass sogar die schlechteste Variante im Test innerhalb der subjektiven Toleranzen liegt. Beispiel sind verrauschte, portable Aufnahmegeräte, Mikrofone, etc.

Formate für Antwortskalen

Die Skala hat fünf Punkte und ist diskret. Sie unterscheidet sich lediglich in ihren verbalen Qualifikatoren bzw. Ankerpunkten von jenen des ACR. Das Skalenniveau darf auch in diesem Fall als quasi-intervallskaliert angesehen werden.

Skala	Verbale Deskriptoren	
	Degradation is…	Entartung ist…
5	Inaudible	**Unhörbar**
4	Audible but not annoying	**Hörbar aber nicht störend**
3	Slightly annoying	**Leicht störend**
2	Annoying	**Störend**
1	Very annoying	**Sehr störend**

Tab. 19: Skala zur Bewertung von Entartungen, Artefakten [800, D2.4]

Versuchspersonen

Als Versuchspersonen werden für den spezifischen Anwendungsbereich des Standards naive Hörer empfohlen. Im Bereich von Musik und Soundqualität oder zu Forschungszwecken können auch hier EL verwendet werden.

Stimuli

Mindestens vier verschiedene Sprecher (zwei Frauen und zwei Männer) kommen zur Generierung von Sprachmaterial zum Einsatz. Dabei sollen dieselben vier Sätze von allen Sprechern gesprochen werden. Jeweils zwei Sätze werden mit einer halben Sekunde Pause zu einem Sample (S_1, S_2) zusammengesetzt, siehe nachfolgende Abbildung.

Abb. 11: Sample bei DCR

Als Sprachmaterial sollen phonetisch balancierte Sätze verwendet werden.

Versuchsablauf

Ein Trial besteht aus dem Abspielen des Samples (z.B. S_1) zuerst für das Referenzsystem (A) und dann für das zu beurteilende System (B). Diese Reihenfolge bleibt während des Hörtests bestehen.

Abb. 12: Trial aus Referenz- (S_1, blau) und verarbeitetem Sample (S_1, grün)

Das Trial kann bei schwierigen Aufgabenstellungen auch zweimal abgespielt werden (A-B, A-B), in diesem Fall soll zwischen den Wiederholungen eine Pause von 1 s bis 1 ½ s eingehalten werden.

Zum Screening der Versuchspersonen kann das Referenzsample innerhalb eines Trials wiederholt werden (A-A). Diese Screeningtrials dürfen in die statistische Analyse des zu bewertenden Systems nicht einfließen, werden aber in der ANOVA zur Prüfung der Reliability herangezogen.

Abb. 13: Screeningtrial aus 2x Referenzsample (S_1 oder S_2)

Zu beachten

Die Wiederholung der vier Sätze im Test wird in der Empfehlung als nicht kritischer Faktor angesehen [800, D2.1]. Das mag zwar für Anwendungen mit guter oder sehr guter Sprachverständlichkeit durchaus richtig sein, allerdings ist ein Gewöhnungseffekt und somit die Tendenz zu einem verbesserten Beurteilungsverhalten mit zunehmender Dauer des Versuchs nicht auszuschließen (vgl. Abs. 1.2.9, Bias). Dieser Einfluss auf Audioanwendungen ist kritisch zu hinterfragen. Die ITU lässt jedoch eine Variation der Samples zu.

Der DCR gehört aufgrund der Bewertung des zweiten Stimulus auf einer Skala auch nicht zu den Paarvergleichen im Sinne indirekter Antwortskalierungen.

Ergebnis

Die Beurteilungen liefern eine Entscheidung zum zweiten des dichotomen Entscheidungspaars (Testsystem), also auf Nominalskala befindlich, die zusätzlich auf einer 5-Kategorienskala hinsichtlich der Störung bewertet wird. Das Ergebnis ist der arithmetische Mittelwert (*degradation mean opinion score*, DMOS).

Möchte man zwei oder mehrere Systeme a posteriori miteinander vergleichen, besteht die Möglichkeit, dies über multiple Vergleiche von Erwartungswerten, beispielsweise nach

Tukey zu tun, Signifikanz vorausgesetzt [800, D3] [Sac02, 742], (vgl. Abs. 1.1.2.2, multiple Vergleiche).

2.2.2.2 Comparison Category Rating (*CCR*) [800, Annex E]

Der CCR-Test erscheint dem DCR-Test vorerst ähnlich, da ebenfalls pro Trial zwei Stimuli präsentiert und miteinander verglichen werden. Allerdings ist der Versuchsperson in diesem Fall nicht bekannt, welches von beiden Samples das Originalsample (Referenz-) darstellt und welches das bearbeitete Sample ist. Das bedeutet, dass in dem Versuch die Hälfte der Trials zuerst das Originalsignal, die andere Hälfte zuerst das veränderte Signal abspielen.

Somit repräsentiert diese Methode den ***vollständigen Paarvergleich*** und bildet als solche eine optimale Ausgangssituation für Designs von Paarvergleichen, unabhängig davon, ob diese vollständig oder lediglich einseitig, demnach ohne spiegelbildliche Wiederholung, durchgeführt werden sollen.

Fragestellung und Anwendungsbereich

Anwendung findet der CCR bei Sprachverarbeitungssystemen, die die Qualität von Sprache als Inputsignal verbessern, wie beispielsweise Noise Cancellation oder Noise Reduction.

Außerdem ist es möglich, den Einfluss verschiedener Umgebungsgeräusche auf ein System zu testen, oder verschiedene Systeme miteinander zu vergleichen.

Die standardisierte Methode eignet sich gut als Basis für jegliche Paarvergleiche, selbst wenn kein Referenzsystem zur Verfügung steht. (*Beispiel: Vergleich von Mikrofonen, Lautsprechern, etc.*)

Formate für Antwortskalen

Die Versuchsperson beurteilt die Qualität des zweiten Samples relativ zu jener des ersten anhand nachfolgender Skala.

Skala	Verbale Deskriptoren	
3	Much better	**Viel besser**
2	Better	**Besser**
1	Slightly better	**Ein wenig besser**
0	About the same	**Ungefähr gleich**
-1	Slightly worse	**Ein wenig schlechter**
-2	Worse	**Schlechter**
-3	Much worse	**Viel schlechter**

Tab. 20: 7-stufige, diskrete Skala mit neutralem Mittelpunkt

Die Skala ist 7-stufig, diskret und bipolar. Zudem weist sie im Gegensatz zu den übrigen Skalen einen neutralen Ankerpunkt auf, das bedeutet, die Vpn kann sich für die Gleichheit zweier Stimuli aussprechen und muss sich demnach nicht für einen der beiden entscheiden.

(*Beispiel: Dieser Aspekt ist wichtig für die Vergleichbarkeit von Beurteilungen desselben Systems auf verschiedenartigen Skalen. Demnach kann das System in einem anderen Versuch ein wenig besser abgeschnitten haben als das Vergleichssystem. Dieser Umstand kann aber auch daher rühren, dass in dem anderen Versuch eine Forced Choice Methode verwendet wurde und demnach also eine Skala, die eine Wahl zugunsten eines Systems erzwingt.*)

Versuchspersonen

Die ITU 800 schlägt grundsätzlich naive Hörer als Versuchsteilnehmer vor. Es sei noch einmal ausdrücklich erwähnt, dass die Auswahl einer Teilnehmergruppe zweckgebunden zu erfolgen hat und demnach für andere als die im Standard definierten Bereiche, das Verfahren adaptiert werden kann und darf, natürlich unter dem Aspekt der Einbuße der Standardisierung.

Stimuli

Verwendet wird für beide Kategorien von Samples derselbe Korpus als Ausgangsmaterial. Ein Sample stellt die Referenz dar, das andere das Testsystem. Dabei besteht ein Sample wiederum aus zwei Sätzen mit einer halben Sekunde Pause inzwischen.

Abb. 14: Sample bei CCR, bestehend aus 2 Sätzen mit Pausen von ½ s

Referenzrauschen kann hier auch zur Kalibrierung der Daten verwendet werden, siehe Kapitel Referenzstimulus 1.2.5.1

Versuchsablauf

Ähnlich dem DCR werden innerhalb eines Trials zwei Samples präsentiert. Die Reihenfolge der Stimuli Referenzsignal und verändertes Testsignal ist der Versuchsperson dabei allerdings unbekannt, da diese sich innerhalb des Versuchsablaufs zu gleichen Anteilen aufteilt. Dennoch beurteilt der Hörer immer das zweite Sample in Bezug und somit relativ zum ersten Sample. Die Methode entspricht somit einem vollumfänglichen Paarvergleich.

Dabei trifft die Versuchsperson gleichermaßen eine Entscheidung darüber, welches der beiden Samples das Bessere ist und in welcher Relation die Qualität jene des anderen Samples übersteigt.

Der Standard schlägt vor, die Reihenfolge von Referenz und Teststimulus für jeden Trial zu randomisieren, sodass letztendlich für jedes Paar einmal die Referenz an erster und einmal an zweiter Stelle präsentiert wurde. Zusätzlich sollen Nullpaare, Referenz – Referenz, für jede zu beurteilende Qualität mit dargeboten werden.

Zu beachten

Sorge ist zu tragen, dass aufgrund der zweiseitigen Antwortskala die statistische Analyse korrekt durchgeführt wird, was eine Korrektur der Daten vor Berechnung des MOS mit einschließt.

Zudem ist über den Pilottest im Vorfeld eruierbar, ob die Sensitivität der Methode für den jeweiligen Anwendungsbereich passend ist.

Ergebnisse

Als Ergebnis des CCR wird grundsätzlich der CMOS angegeben. Dieser Wert bezieht sich auf die korrigierte Datenmenge entsprechend der Reihenfolge Referenz – verändertes Signal. Die Hälfte der Daten, deren Trials die vertauschte Reihenfolge hatten, muss dazu in ihren Antwortwerten der VP korrigiert, also invertiert werden.

Beispiel:

- *Reihenfolge der Samples innerhalb eines Trials: Referenz – Testsystem. Die Bewertung bleibt unverändert.*

- *Reihenfolge der Samples innerhalb eines Trials: Teilsystem – Referenz. Die Bewertung der Vpn wird invertiert.*

    - *Bewertet die Vpn das zweite Sample mit +3 (entspricht: das zweite Sample, Referenz, ist viel besser als das erste, Testsystem), wird daraus der Wert -3, was so viel bedeutet wie: in der neuen, vertauschten Reihenfolge ist das Sample 2, Testsystem, viel schlechter als Sample 1, Referenz.*

Der Vorgang der Invertierung ist nur zulässig, wenn die Ankerpunkte der positiven Seite der Skala äquivalente Relative der negativen Seite darstellen und umgekehrt. Überdies müssen die Abstände der Kategorien intervallskaliert sein, andernfalls sind die beiden Hälften getrennt voneinander als ordinalskaliert zu betrachten. In diesem Fall sind lediglich Median und Perzentile berechenbar.

Der vollständige Paarvergleich bietet den Vorteil, den Kontexteffekt zu vermeiden, da sowohl B auf A als auch A auf B getestet wird (vgl. Abs. 1.2.9.5). Dieser Vorteil geht oft auf Kosten des Prüfumfangs (vgl. Abs. 1.2.5.3).

Ein zusätzlicher Vorteil liegt in der Prüfbarkeit der Variabilität der Vpn durch das wechselweise Abspielen der Stimuli (unter Umständen sogar double blind, wenn selbst VL nicht genau weiß, welches Sample zuerst gespielt wird). Die Methode ermöglicht außerdem ein zweiseitiges Testen von Sprachverarbeitungssystemen (besser oder schlechter).

2.2.2.3 Schwellwertmethode (threshold method) [800, Annex F]

Die Versuchsperson vergleicht in diesem Verfahren wiederum eine Referenz mit einem Teststimulus. Dabei entscheidet sie, welche der beiden Stimuli sie präferiert, ohne die Präferenz zusätzlich auf einer Skala zu spezifizieren. Dieses Verfahren spiegelt daher einen reinen Präferenztest wieder (vgl. Abs. 1.2.3.1, Präferenztest)

Fragestellung und Anwendungsbereich

Die Methode ist ganz allgemein zur Systemoptimierung einsetzbar. Das Testsystem wird mit einem Referenzsystem verglichen und der Wert für die Bedingung bestimmt, bei dem die Performances von Referenz- und Testsystem übereinstimmen.

Formate für Antwortskalen

In diesem Verfahren kommt keine Skala zur Anwendung, da die Teilnehmerin lediglich A besser B oder B besser A bewertet.

Versuchspersonen

Der Standard schlägt naive Hörer vor, von denen mindestens 6, vorzugsweise aber mehr als 12. Dem Anwendungsfall angepasst kann auch ein trainiertes Panel zur Beurteilung herangezogen werden.

Stimuli

Ein Trial besteht grundsätzlich aus dem Referenzstimulus und dem Teststimulus mit einer Pause von1-1½ s zwischen den beiden und einem Hinweissignal (*Cue*) zu Beginn des Trials. Als vollständiger Paarvergleich ausgelegt, wird die Reihenfolge A-B für die zweite Hälfte an Trials genau umgekehrt, folglich B-A.

Abb. 15: Abfolge von Referenz- und Testsample mit Pausen

Die Stimuli selbst, also sowohl Referenz als auch Testsignal, dauern für Sprachtests 2,5-5 s und für Musikanwendungen 10-15 s. Dabei werden sukzessive Abstufungen, bsp. in konstanten dB-Schritten, der zu testenden Eigenschaft durchlaufen.

Für das Quellmaterial Sprache sollen zumindest je 2 bis 6 Männer und Frauen voneinander verschiedene kurze Sätze sprechen.

Versuchsablauf

Beurteilt wird innerhalb jedes Trials, welcher Stimulus besser war als der andere. Die Abfolge der Stimulipaare ist dabei randomisiert. Ein Durchlauf soll eine Zeitdauer von 6 Minuten nicht übersteigen, um Ermüdungseffekte zu vermeiden. Dabei können mehrere Durchläufe innerhalb eines Versuchs realisiert werden, eine Wiederholung des Durchlaufs wird je nach Komplexität der Aufgabenstellung empfohlen.

Zu beachten

Bei der Auswahl der Teststimuli soll ein Wertebereich von 20% bis 80% der Eigenschaft abgedeckt werden. Der Anspruch ist allerdings kritisch in Bezug auf die Gesamtdauer des Versuchs und in Bezug auf Ermüdungserscheinungen der Vpn aufgrund anspruchsvoller Stimuli zu betrachten.

Dieser Aspekt mag nicht so gravierend sein, wenn es darum geht, Rauschen als störende Eigenschaft zu detektieren. Man stelle sich nun zusätzlich zu einem störenden Artefakt ein perzeptiv komplexes Signal vor, beispielsweise der neue Rasenmäher der Firma XY rasselt. Zu detektierende Eigenschaft ist das Rasseln, Test- und Referenzsignal sind Rasenmähergeräusche. Für einen derartigen Versuch ist es ratsam, von Vornherein nahe der Schwelle, also dem 75% Threshold, an der man Schwierigkeiten hat, das Rasseln noch zu identifizieren, Teststimuli zu generieren und auf deutlich voneinander zu unterscheidende Stimuli zu verzichten. Der Versuch ist mit Sicherheit ohnehin anstrengend genug.

Ergebnisse

Die Schwelle für die Eigenschaft, bei der kein Unterschied zwischen Test und Referenzsignal erkennbar ist, stellt sich aufgrund der sukzessiven Approximation durch die variierten Abstufungen der Eigenschaft bei 50% ein. Dieser Punkt subjektiver Gleichheit wird graphisch auf einer psychometrischen Funktion, siehe Abb. 16, dargestellt. Die Ordinate dieser Funktion wird durch den Prozentsatz der Präferenz oder auch korrekten Antworten gebildet und die Abszisse je nach Fragestellung beispielsweise durch den SNR des Referenzsignals oder die Level der Signalintensität. Nach Gelfand wird der 75%-Punkt als Schwelle zur tatsächlichen Wahrnehmung eines Unterschieds bezeichnet. Dementsprechend leitet sich daraus dann die Unterschiedsschwelle als Differenz zwischen dem Punkt der Gleichheit zweier Signale und dem Punkt der Unterschiedswahrnehmung ab [Gel04, Kap7].

Die Streuungsbreite r zum Signifikanzniveau α für einen Wert u berechnet sich aufgrund der Annahme einer t-Verteilung zu:

$$r = \pm t(df, \alpha) \cdot \sqrt{\frac{u(1-u)}{df}} \quad \text{mit} \begin{cases} 0 \le u \le 1 \\ df = n-1 \end{cases} \tag{73}$$

nach [800, F.7].

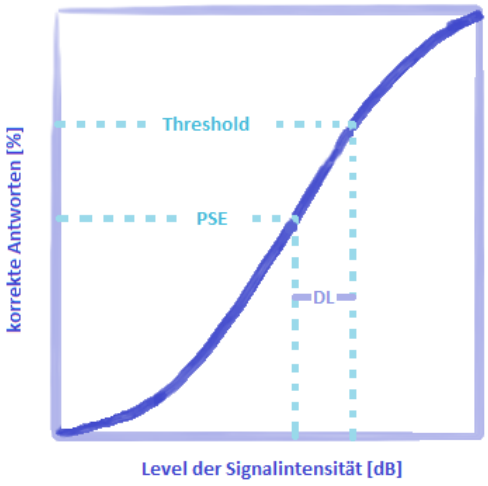

Abb. 16: psychometrische Funktion [Gel04, Kap7]

Bezug zur Psychophysik

Die Methode ähnelt der aus der Psychophysik bekannten Methode der konstanten Stimuli (*method of constant stimuli*) mit differentieller Schwelle zur Ermittlung der Hörschwelle von Menschen. Als differentielle Schwelle gilt im besprochenen Verfahren die festgelegte Referenz, um die herum die Teststimuli generiert und getestet werden. Damit kann die

Referenz aber auch Informationsträger für eine beliebige Eigenschaft eines auditiven Signals sein, solange diese Eigenschaft eindeutig erkennbar ist. Vorrangig bei dem Verfahren ist nicht die Benennung des Attributs, sofern der Versuchsperson unmissverständlich vorgeführt werden kann, worauf, also auf welches Attribut, er hören muss.

(*Beispiel: wenn sich die Frage an die Hörerin lediglich danach richtet, ob sie die Eigenschaft noch hört, bzw. ob sie den Unterschied zur Referenz noch hört, kann somit ausgetestet werden, um welchen Anteil ein System tatsächlich bezüglich der Eigenschaft verändert werden muss, ohne diese verbal weiter definieren zu müssen, wie es sonst bei Attributauswahlprozessen der Fall ist. Diese Eigenschaft des Tests macht ihn flexibel einsetzbar, auch für Forschungszwecke.*)

Die Methot of Constant Stimuli stellt zudem eine Alternative zur Anpassungsmethode (*method of adjustment*). Die Method of Adjustment ist dabei jene Methode, bei der der Teilnehmer mit einem taktilen Werkzeug, beispielsweise einem Einstellrad, die zu testende Eigenschaft der Referenz subjektiv gleichsetzt.

Vorteil dieser Methode sind hohe Präzision und die Möglichkeit der Integration von Nullpaaren (*Catch-Trial*), die eine nachfolgende Analyse der Ratewahrscheinlichkeit einer Versuchsperson ermöglichen. Nachteilig wirkt sich die Inefffizienz der Methode dann aus, wenn sich der Großteil der Stimuli nicht in der Nähe der zu ermittelnden Unterschiedsschwelle befindet. Zur weiterführenden Information sei auf Green verwiesen [Gre88].

2.2.3 Anwendungsbeispiele

Nachfolgende Beispiele geben Anwendungsmöglichkeiten im Bereich der Evaluierung von Sprachqualität an.

- Systeme hoher Qualität: DCR (eigentlich Codecs, P800, S22)

Codecs (Schmalband 300-3400 Hz, Breitband 100-7000 Hz)

- Bewertung des Effekts von Multiple talkers bei Codecs
    - z.B. Konferenz: DCR, QRD (P830, S.7)
- Bewertung der Robustheit gegen error conditions: (P830, S.7)
    - Festnetz: randomly distributed bit errors (Bit error rate BER)
    - Radio environment (mobile radio): burst error type
    - Beide mit ACR bewertet, dort mit listening quality scale, außer die Qualität ist schlecht, dann eher listening effort scale verwenden.
- Bewertung der Abhängigkeit von Sprecherstimmen:

    - In diesem Fall ganz wichtig: 8 Frauen, 8 Männer, 8 Kinder als Korpus

- Quellmaterial verrauscht:
    - DCR (P830, S.1)
- Systemoptimierung:
    - DCR, threshold method (P830, S.1)
- Use of information signals (ring tone…) bzw recognition
    - DCR (P830, S.11)

2.3 Methoden zur Beurteilung von Audioqualität

Die vorangegangenen Methoden (ACR, CCR, DCR, etc. vgl. Abs. 2.2) der ITU-T P800 und P830 beschäftigten sich mit der Bewertung von Sprachmaterial im Umfeld der Telefonie. Um die Verwendbarkeit jener Verfahren für Audiosignale im Bereich des Rundfunks zu überprüfen, führte die EBU zahlreiche Versuche durch mit dem Ergebnis, dass keine dieser Methoden alle gewünschten Anforderungen gleichzeitig erfüllt [1534, S.2]. Zu diesen gehören:

- Absolute Skala
- Vergleichbarkeit mit einem Referenzsignal
- Kleine Konfidenzintervalle
- Vernünftige Anzahl an Versuchspersonen

Aus diesem Grund werden nachfolgend spezielle Verfahren, die sich für die Vergleichbarkeit und Beurteilung von Audiosignalen eignen, näher erläutert. Dabei kann nicht nur ein System (Mikrofon, oder ein Lautsprecher, oder Codec), sondern auch das Klangmaterial selbst beurteilt werden (*beispielsweise ein Mix kann bezüglich Tiefenstaffelung, Stereobreite etc. beurteilt werden*).

Grundsätzlich können alle Methoden aus der ITU P800 für Audioanwendungen verwendet werden, deren Durchführung ist für Sound quality aber nicht näher in den Empfehlungen spezifiziert. Lediglich in der ITU BS 1284-1 sind wichtige Aspekte für die Bewertung von Sound quality im Allgemeinen angeführt. Deren Details werden in die nachfolgenden Verfahren mit eingearbeitet, beziehungsweise sind in den allgemeinen Rahmenbedingungen aus Abs. 1.2.8 angeführt. In diesem Fall wird auf relevante Aspekte verwiesen.

2.3.1 Methoden zur Bewertung von einem Stimulus (ohne vergleichende Referenz)

Verfahren, die lediglich einen Stimulus ohne Bezugspunkt absolut bewerten, sind in den Methoden zur Beurteilung von Sprachqualität zu finden, vgl. Abs. 2.2.1. Grundsätzlich ist anzumerken, dass die Beurteilung von Audioqualität und Audiomaterial von den Gremien kritischer betrachtet wird und wurde. Das schließt die Komplexität des Ausgangsmaterials, die damit einhergehende erhöhte Sensitivität in der Beurteilung, den Anwendungsbereich und auch die Methode selbst mit ein (vgl. Abs. 2.3). Jedenfalls sollen ergänzend zu den Inhalten aus Abs. 2.2.1 die Betrachtungen aus Abs. 2.3.2.1 gelesen werden, die in Bezug auf Audioqualität allgemeine Hinweise zur Durchführung anhand der ITU-R BS.1284-1 geben.

2.3.2 Vergleichende Verfahren

2.3.2.1 Paarvergleich (Paired Comparison) [1284-1, 4]

Die Durchführung des Paarvergleichs für Audiomaterial erfolgt im Wesentlichen ähnlich dem Comparison Category Rating (vgl. Abs. 2.2.2.2, CCR).

Fragestellung und Anwendungsbereich

Der Paarvergleich ist als direkte Vergleichsmöglichkeit ähnlicher Systeme sehr beliebt. Er eignet sich dabei gut um zwei Systeme bezüglich verschiedener Eigenschaften genauer zu befragen, oder aber die Vielseitigkeit von Systemen zu testen.

Das Verfahren findet aber auch Anwendung bei Benchmarking Tests, deren Geräusche als anstrengend empfunden werden und dementsprechend hörerfahrene Versuchsteilnehmerinnen erfordern.

Formate für Antwortskalen

Abgesehen von der 7-stufigen diskreten Vergleichsskala können auch die Listening Quality Scale vgl. Tab. 12) oder nachfolgende Impairment Scale verwendet werden.

Skala	Verbale Deskriptoren	
5	Imperceptible	**Nicht wahrnehmbar**
4	Perceptible, but not annoying	**Wahrnehmbar, aber nicht störend**
3	Slightly annoying	**Leicht störend**
2	Annoying	**Störend**
1	Very annoying	**Sehr störend**

Tab. 21: Impairment Skala [1284, 4]

Die Skalen liefern im Allgemeinen, abgesehen von den verbalen Deskriptoren, die verschiedene Fragestellungen voraussetzen, unterschiedliche Ergebnisse. Die siebenstufige Skala stellt einen direkten Vergleich zweier Systeme an, die beiden anderen Skalen setzen eine Verschiedenheit der beiden Systeme a priori voraus (vgl. Abs. 2.2.2.2, Formate für Antwortskalen: Beispiel).

Versuchsablauf

Wird eine der beiden 5-stufigen Skalen in Kombination mit einer Referenz verwendet, werden Referenz- und Teststimulus mit ½ s Pause zu einem Sample zusammengefasst und wiederholt präsentiert. Zwischen den beiden Samples soll die Pause 1½ s betragen.

Abb. 17: zwei Samples beim Paarvergleich

Abb. 17 zeigt den Ablauf innerhalb eines Trials bestehend aus: Referenz-Test-Referenz-Test.

Wird hingegen die 7stufige Skala verwendet, werden zwei Systeme miteinander verglichen und der Ablauf sieht an sich gleich aus. Zusätzlich kann jedoch zu Beginn jedes Trials eine Referenz einmalig dargeboten werden.

Stimuli

Wird musikalisches Material verwendet, sollen die Abschnitte nicht länger als 15-20 s dauern und eine Phrase bilden. Für die Musikstelle bedeutet dies, einen Spannungsbogen nicht abzuschneiden, sondern stattdessen eine passendere Stelle im Stück zu finden. Das heißt nicht, dass es nicht auch kurze Abschnitte mit einer Dauer von wenigen Sekunden geben darf.

2.3.2.2 Double blind Triple-Stimulus with hidden reference [1116-1]

Diese Versuchsmethode kommt als standardisiertes Verfahren bei jenen Hörtests zum Einsatz, in denen Audiosysteme mit geringen Qualitätsbeeinträchtigungen (*small impairments*) getestet werden sollen. Das bedeutet, das Verfahren ist sensitiv für kleine Klangunterschiede im zu testenden Signal.

Fragestellung und Anwendungsbereich

Getestet werden die Basic Audio Quality und zusätzlich für Mehrkanalsysteme die Qualität des Front Image und der Eindruck der Surround Qualität.

Formate für Antwortskalen

Die Skala für die Bewertung ist die kontinuierliche Impairment Skala. Die Deskriptoren sind dieselben wie für den Paarvergleich im vorherigen Abschnitt. Allerdings wird hier eine zusätzliche Gliederung der Kategorien in Form einer Auflösung von einer Dezimalstelle gefordert. Die Dezimalstelle in nachfolgender Tabelle soll dies verdeutlichen.

Skala	Verbale Deskriptoren	
5.0	Imperceptible	**Nicht wahrnehmbar**
4.0	Perceptible, but not annoying	**Wahrnehmbar, aber nicht störend**
3.0	Slightly annoying	**Leicht störend**
2.0	Annoying	**Störend**
1.0	Very annoying	**Sehr störend**

Tab. 22: kontinuierliche Impairment Skala [1116-1,4]

Auch bei dieser Skala kann auf die Deskriptoren verzichtet werden, sofern die Richtung der Skala ersichtlich bleibt. Die nachfolgende Normalisierung erfolgt nach Formel (21), (vgl. Abs. 1.2.4.3).

Versuchspersonen

Als Vpn werden in der ITU-R BS 1284-1 Experten immer den naiven Hörern vorgezogen.

Stimuli

Als Quellmaterial soll dem Standard nach kritisches Material verwendet werden, das das jeweilige System strapaziert. Der Prozess der Materialauswahl ist zeitintensiv und komplex, nicht zuletzt auch deswegen, weil es keinerlei Restriktion bezüglich des Ursprungs gibt. Diese Gegebenheit eröffnet aber auch vielerlei Anwendungsmöglichkeiten (vgl. Abs. 1.2.5, EBU Audiofiles). Es sollen mindestens fünf Audioabschnitte verwendet werden, idealerweise jedoch 1½ mal die Anzahl an Systemen. Die Abschnitten dauern zwischen 10-25 s. Beispielstimuli für die Methode sind im Anhang zu finden.

Versuchsablauf

Der Versuchsperson werden innerhalb eines Durchlaufs (Trial) drei Stimuli präsentiert. Stimulus A ist dabei stets der Referenzstimulus und der Vpn als solcher bekannt, wohingegen Stimulus B und C randomisiert das Testsignal und die versteckte Referenz darstellen. Der Hörer hat nun die Aufgabe, B versus A und C versus A zu bewerten, nicht aber B und C zueinander. Dabei ist der Teilnehmerin aus den Instruktionen bekannt, dass es sich bei Stimulus B oder C um eine versteckte Referenz handelt, eines von den beiden Testsignalen demnach also mit *nicht wahrnehmbar* bzw. mit der Note *5* zu beurteilen ist. Das Bewusstsein um das Vorhandensein einer versteckten Referenz hat zur Folge, dass die Testperson eine Annahme über das veränderte Material zu treffen hat, die sich von jener des versteckten Originals unterscheidet. (*bspw. Wenn das veränderte Material für subjektiv besser befunden wird als das Original, so kann es mit einer Note von 4.0 bis 4.9 - als*

merkbarer aber nicht störender Unterschied - bewertet werden und ist für die Versuchsleiterin als Impairment detektiert.).

Zu beachten

In den einleitenden Erläuterungen an die Versuchsperson ist zu erwähnen, dass ausdrücklich ein Vergleich der verschiedenen Systeme zueinander gewünscht ist. Die Referenz soll bei Bedarf angehört werden, um den Bezug zur absoluten Qualität aufrecht zu erhalten, soll aber nicht zu einem direkten Vergleich zwischen Referenz und Teststimulus verleiten. Der direkte Vergleich zwischen den Teststimuli der Audiosysteme soll erhalten bleiben.

Ergebnisse

Die ANOVA kann wiederum angewendet werden um signifikante Unterschiede zwischen den Systemen festzustellen.

Ein Vorteil dieses Verfahrens ist die Möglichkeit, das Antwortverhalten von Versuchspersonen relativ einfach zu überprüfen (*screening*). Dies geschieht mit einem einseitigen t-Test. Für jedes Trial wird dazu die Differenz der beiden Beurteilungen, immer in die richtige Richtung also Beurteilung des Referenzsamples minus Beurteilung des Testsamples gebildet. Die Nullhypothese postuliert, dass der Durchschnitt der Differenzen aller Trialbewertungen dieser Teilnehmerin annähernd Null ist, das bedeutet, dass die Teilnehmerin raten würde.

2.3.2.3 Multi stimulus with hidden reference and anchor (MUSHRA) [1534]

Fragestellung und Anwendungsbereich

Die Versuchsmethode eignet sich zur Prüfung von Audiosystemen mit mittleren bis hohen Qualitätsverlusten, beispielsweise Low Bit-rate Audio Codecs. Das bedeutet, dass Artefakte als Höreindruck nicht im gerade noch wahrnehmbaren Bereich der Unterschiedsschwelle liegen, sondern eindeutig diskriminierbar sind.

Folgendes ist hier von Interesse:

- wenn verschiedene Systeme verschiedene Stimuli mit verschiedenartigen Artefakten oder ähnlichen Artefakten verschiedener Intensität repräsentieren, und
- wenn die Versuchsperson diese Störungen eindeutig wahrnehmen kann,
- welche dieser Fehler werden dann für am Angenehmsten oder am wenigsten störend empfunden?

Formate für Antwortskalen

Die Antwortskala verwendet dieselben verbalen Deskriptoren, wie die Listening Quality Scale des ACR, jedoch mit äquidistanten, kontinuierlichen Intervallen.

Skala	Verbale Deskriptoren	
5.0	Excellent	Sehr gut
4.0	Good	Gut
3.0	Fair	Annehmbar
2.0	Poor	Mäßig
1.0	Bad	Schlecht

Tab. 23: kontinuierliche Listening Quality Scale bei MUSHRA

Vorgeschlagen wird eine Gesamtlänge der Skala von 10 cm. Allerdings ist diese Vorgabe bei Touchscreens der entsprechenden Bildschirmdiagonale und Ergonomie anzupassen.

Versuchspersonen

Als Vpn kommen Experten zum Einsatz.

Für dieses Verfahren wird eine Trainingsphase empfohlen um die Zuverlässigkeit im Antwortverhalten der Vpn zu erhöhen. Der Grund liegt darin, dass die perzeptiven Unterschiede der Teststimuli zueinander gering sein können, bzw. sollen, wenngleich diese gleichzeitig im Vergleich zur Referenz groß sind.

Das Stimuliset soll hierzu den gesamten, innerhalb des Tests zu beurteilenden, Bereich an Störungen abdecken.

Stimuli

Um die Wahrnehmbarkeit der Fehlerhaftigkeit im Audiosignal zu festigen, kommt in dieser Methode eine qualitativ hochwertige Referenz zum Einsatz. Diese wird durch das unbearbeitete Originalsignal repräsentiert. Damit ist ein deutlich hörbarer Unterschied zu den Testsignalen, die mit Artefakten versehen sind, festgelegt.

Zusätzlich zum Referenzsignal wird als Anker das tiefpassgefilterte und auf 3,5 kHz bandbegrenzte Originalsignal verwendet. Der spezifische Anwendungsfall kann den Einsatz eines passenderen Ankers – Beispiele hierzu sind in Kapitel Anker 1.2.5.2 zu finden – notwendig machen.

Ein Stimulus soll nicht länger als 20 s dauern, um den Ermüdungseffekten und zu langer Testdauer vorzubeugen, wobei gleichzeitig die Charakteristik des Audioabschnitts repräsentiert werden soll. Als Abschätzung für die Anzahl an auszuwählenden Musikabschnitten gilt 1½-mal die Anzahl an Testsystemen.

Wird ein Multikanalsystem über einen 2-Kanal Down-Mix getestet, ist die Eignung des Audiomaterials vorab über einen Referenz-Down-Mix nach ITU-R BS.775 zu prüfen.

Versuchsablauf

In den Instruktionen an die Versuchsperson wird darauf hingewiesen, dass sich im Set eine versteckte Referenz befindet (vgl. Abs. 1.2.5.1, Referenzstimulus).

Innerhalb des Versuchs sollen maximal 12 verschiedene Systeme miteinander verglichen werden. Insgesamt werden somit, abgesehen von der bekannten Referenz, welche lediglich angehört wird, 14 Stimuli pro Trial bewertet, da ja eine versteckte Referenz und ein versteckter Anker zusätzlich als Teststimuli beurteilt werden.

Die Versuchsperson hat über eine Schaltfläche zusätzlich die Möglichkeit, sich das Referenzsignal während des Versuchs beliebig oft anzuhören (diese wird nicht bewertet). Es erfolgt eine Bewertung nach der relativen Annoyance aller vorkommenden Artefakte, und auch der beiden versteckten Stimuli, in Form einer Gewichtung im Set.

Die Bewertung des zu testenden Attributs erfolgt über den Vergleich der Stimuli von verschiedenen Audiosystemen. Idealerweise sind die qualitativen Unterschiede der Testsignale relativ zur Referenz groß, wohingegen sie untereinander im direkten Vergleich klein sind [1534, 5.4].

Zu beachten

Führen die verlustbehafteten Teststimuli des getesteten Audiosystems entgegen der Erwartung des Versuchsleiters zu einer Verbesserung der subjektiven Qualitätsbeurteilung, soll eine andere Methode zur Beurteilung herangezogen werden [1534, 3]. Eine Ursache für die subjektive Verbesserung der Qualität kann in einer zu geringen Auflösung der Diskriminierbarkeit der Qualitätseigenschaften im Vergleich zur Referenz liegen, weshalb das Verfahren zu wenig sensitiv reagiert. In diesem Fall ist die Methode zur Untersuchung von geringen Qualitätseinbußen anzuwenden, vgl. Abs. 2.3.2.2.

Während des Versuchs soll die Vpn in der Lage sein, den Abhörpegel in einem Bereich von ±4 dB nachzuregeln. Dieses Bedienelement darf während des Anhörens eines Samples nicht verfügbar sein [1534, 8].

Ergebnisse

Die Versuchsperson hat bei diesem Verfahren die Möglichkeit, zwei oder mehrere Audiosysteme miteinander hinsichtlich ihrer Qualitätseinbußen im Audiosignal zu vergleichen, ohne dabei die absolute Qualität als Bezugspunkt zu verlieren.

Innerhalb dieses Versuchs findet neben der Bewertung der einzelnen Artefakte zusätzlich eine Reihung der verschiedenen Testsysteme statt.

Dieses Verfahren repräsentiert einen vollständigen Paarvergleich, da jederzeit sämtliche Systeme miteinander in Bezug gesetzt werden. Das hat den entscheidenden Vorteil, dass die Auflösung im Einstufungsprozess viel höher ist als bei Verfahren, bei denen beispielsweise

lediglich zwei Stimuli miteinander in Bezug gesetzt werden oder bei denen gar kein Vergleichswert zur Beurteilung herangezogen werden kann. Daraus resultieren direkt kleinere Konfidenzintervalle und konsistentere Antworten im Vergleich zur ITU BS1116-1 [1116-1].

Das Verfahren ist durch eine relativ kurze Testdauer ausgezeichnet.

2.3.3 Anwendungsbeispiele

Nachfolgende Beispiele geben einen Einblick in mögliche Anwendungsbereiche der behandelten Methoden.

- Küchengeräte: Mixer, Pürierstäbe, Smoothiemaker, etc.
- Geräte im Bad: Rasierer, Epiliergerät, elektrische Zahnbürste, Haartrockner
- Technische Geräte: Diktiergerät, Lüftergeräusche im Arbeitsmodus, Radio, etc.
- Audio: Lautsprecher, Miniboxen, Pc-Lautsprecher, Portable Lautsprecher, Handylautsprecher, Surround Anlagen, Kopfhörer, etc.
- Beispiele aus der Literatur:
    - Kleine Lautsprecher: [Bah12]
    - Basic Audio Quality: [Sch12]
    - Internet Audio Codecs: [Sto00]
    - Paarvergleich und Reihenfolge der Stimuli, Kontexteffekt: [Mar06]

3 Ergänzungen zum Handbuch

Bei der Ausarbeitung der Teilbereiche, aus denen sich das Handbuch zusammensetzt, wurde der Fokus auf die den jeweiligen Bereich betreffenden Aspekte gelegt. Die Herangehensweise, Gedankengänge und Hintergrundinformation zu den Vorlagen für das Testdesign, die Berichterstellung, sowie die Kalkulation eines Hörversuchs werden in diesem dritten Teil behandelt.

3.1 Checkliste

Das Versuchsdesign ist als iterativer Prozess anzusehen. Das bedeutet doch zu Beginn, dass man sich einer Fragestellung klar wird, die man im Rahmen eines Versuchs möglichst beantwortet haben möchte. Mit dieser Fragestellung im Hinterkopf stellt man die Parameter auf, die im Zuge des Designprozesses umgesetzt werden können. Nachfolgende Fragen sollen eine Hilfestellung zur Durchführung eines Versuchsdesigns darstellen. Dabei ist es durchaus möglich, dass man beispielsweise erst bei Punkt 3, wenn sich die Frage nach dem Antwortformat stellt, herausfindet, dass das gedachte Format nicht die gewünschte Antwort bereitstellt, demnach ein anderes präferiert wird und diesem angepasst die Variablen umstrukturiert werden müssen. Diese Iterationen können bis zur Umformulierung der ursprünglichen Fragestellung zurückgehen.

In der Literatur wird manchmal zu einem gruppendynamischen Prozess geraten. Die Autorin sieht die Diskussion in der Planungsphase, auch mit Laien, als durchaus produktiven und zeiteffizienten Prozess an, wenn es darum geht, Störfaktoren aufzuzeigen und aus einem anderen Blickwinkel das Design kritisch hinterfragen zu können.

Sämtliche, nachfolgende Fragen sind in ICH-Form verfasst, was bedeutet, dass das ICH sowohl auf den Forscher als auch auf eine Kundin auslegbar ist. Die Fragen sollen das Augenmerk auf verschiedene Aspekte die innerhalb dieser Arbeit abgehandelt wurden, lenken, was keinesfalls bedeutet, dass diese vollumfassend sind. Man möge sich von der ICH-Form geleitet und gleichermaßen inspiriert, hoffentlich nicht irritiert fühlen.

3.1.1 Fragestellung, Hypothese

Was will ich herausfinden? Was will ich wissen? Kann ich, was ich wissen will tatsächlich in einem Hörtest erfahren, oder vielleicht einfacher berechnen, simulieren, etc?

Wie kann ich, was ich wissen will, in eine Frage umformulieren?

Weiß ich, in welche Richtung mein Ergebnis gehen wird? Rechne ich damit dass ein System am Besten abschneidet? Oder kann ich den Ausgang des Experiments gar nicht abschätzen?

Habe ich die Frage danach gerichtet/ungerichtet formuliert? Kann ich genau diese Frage verwerfen?

3.1.2 Antwortattribut

Wie konkret ist mein Anwendungsfall?

Welche Variablen spielen mit?

Wie allgemein soll die Anwendung gültig sein?

Wie viele verschiedene Eigenschaften will ich prüfen?

Wie viele Kombinationsmöglichkeiten ergeben sich durch die verschiedenen, involvierten Variablen? Ist das Szenario realisierbar?

Soll ich mich vielleicht auf einen Ausschnitt davon beschränken? Was will ich wirklich/unbedingt wissen?

Welche von den Variablen kann ich konstant halten? (Rahmenbedingungen, z.B. Referenzraum)

Was berücksichtige ich gerade nicht, was will ich nicht berücksichtigen (konstant halten, Kontrollvariable)?

3.1.3 Antwortformat

Mit welcher Art von Antwort bin ich zufrieden?

Will ich einen schnellen und einfachen Versuch?

Reicht mir eine Zahl als Ergebnis des Versuchs aus?

Will ich nur eine grobe Abschätzung? Eine Reihenfolge?

Oder will ich mehr wissen? Was brauche ich dann dazu? Mindestens?

3.1.4 Experiment Design

Kann ich die subjektive Differenz auflösen oder bräuchte ich dazu zu viele Stimuli?

Habe ich so einen ähnlichen Test oder genau so einen Test schon einmal durchgeführt? Wenn ja, was war dann gut/ nicht gut? Was will ich diesmal besser machen? Kann ich Lessons learned heranziehen?

Was ist mir wichtig zu wissen? Auf welche Aussage kann ich verzichten?

Ist es mit den vielen Variablen realistisch, ein zuverlässiges Ergebnis zu bekommen?

Muss ich überhaupt verzichten, oder geht sich die Umsetzung gut aus? (weil ich nicht so viele verschiedene Szenarien teste; weil ich genau weiss, welches Szenario ich möchte)

Kann ich die Stimuli schnell generieren? Habe ich schon eine Datenbank die ich verwenden kann?

Wie viele Stimuli benötige ich?

Sind diese Stimuli den Versuchspersonen schon bekannt?

Kann ich zeitlich betrachtet jeder Versuchsperson alle Stimuli vorspielen?

3.1.5 Versuchspersonen

Welche will ich? Für welchen Anwendungsbereich ist meine Frage gedacht?

Wie viele nehme ich? Wie viele kann ich mir zeitlich und finanziell maximal leisten?

Hat mit den finanziellen Mitteln das Design Sinn? Sind vielleicht Restriktionen in der Fragestellung zugunsten eines umfassenderen Testdesigns sinnvoller, damit ich zuverlässigere Antworten bekomme?

Habe ich Screening Daten von den Experten? Wenn ja, welche waren denn bei dem von mir gewählten Verfahren am zuverlässigsten?

3.2 Projektabwicklung

Dieser Abschnitt beschreibt wesentliche Aspekte der Akquisitionsphase und der Berichterstellung.

3.2.1 Akquisition

Nicht immer ist dem Kunden völlig klar, welche Art von Information oder Datenerhebung er genau benötigt bzw. welche Ergebnisse sie sich von der Durchführung eines Versuchs versprechen kann. Dazu sei auf die beiden folgenden Unterkapitel mit möglichen und auch unmöglichen Ergebnissen eines subjektiven Hörtests nach Bech verwiesen [Bec06, Kap1].

Zudem kann es sinnvoll sein, Mockups bereits durchgeführter Tests aus einem ähnlichen Anwendungsgebiet vorzustellen.

3.2.1.1 Ergebnisse eines Hörversuches

- ✓ Reihung von mehreren Audiosystemen nach Priorität anhand eines festgelegten Attributs
- ✓ Ermitteln, ob ein akustisches Ereignis gleich gute, bessere oder schlechtere subjektive Ergebnisse liefert und in welchem Ausmaß
- ✓ Feststellen, ob gewünschte Performance des Audiosystems in einem bestimmten Anwendungsfall zufriedenstellende Ergebnisse liefert, dabei muss „zufriedenstellend“ definiert werden ebenso wie ein zugehöriges Attribut
- ✓ Erforschen: Akustisches Ereignis wird erkannt/nicht erkannt (2 Stimuli führen zur selben Wahrnehmung oder nicht)
- ✓ Performance eines Audiosystems detaillierter ermitteln anhand mehrerer, perzeptiver Attribute

3.2.1.2 Ergebnisse eines Hörversuchs sind NICHT

- ✓ Identifikation von problematischen Systemdesignparametern
- ✓ Identifikation, welche Parameter das Konkurrenzprodukt besser machen
- ✓ Vorhersage, welches Produkt im Wettbewerb als Testsieger hervorgeht
- ✓ Maßnahmen, die das Produkt signifikant verbessern

Subjektive Wahrnehmung kann durchaus Aussagen über Tendenzen machen und gemeinsam mit den technischen Daten eines Produkts zu Verbesserungsmaßnahmen und dergleichen seinen Beitrag leisten, allerdings geht dies über das Aufgabengebiet des Hörversuchs selbst hinaus. Dieser zeigt lediglich like/dislike von perzeptiven beziehungsweise affektiven Attributen auf. Daher sollen Ergebnisse auch mit Vorsicht interpretiert und zur weiteren Analyse herangezogen werden.

Vorrangiges Ziel der Akquise ist - einmal abgesehen vom wichtigsten Punkt, dem Projektzuschlag – eine konkrete Fragestellung gemeinsam mit dem Kunden zu erarbeiten. Diese dann in eine zu testende Hypothese umzuformulieren, ist vorerst von untergeordneter Bedeutung. Dem Versuchsleiter mag hilfreich erscheinen, sich die Attributlisten für den etwaigen Anwendungsbereich a priori anzusehen um sein Portfolio zu erweitern. Abgesehen davon ermöglicht die Benennung zu testender Eigenschaften eines Signals, abgesehen von Gesamtqualität, eine klarere und anschaulichere Vorstellung vom potentiellen Informationsgewinn für beide Parteien.

Gegen Ende der Akquisitionsphase ist es entscheidend zu fixieren, welche Information sich der Kunde erhofft und auch erwarten kann, um den Projektablauf nicht im Design- oder Adaptionsprozess zu stören. Dementsprechend sei der Fokus auch auf Informationsbeschaffung gelegt. Umgekehrt ist es für die Durchführung essentiell, dem Kunden klar zu vermitteln, welche Informationen man benötigt und bis zu welchem Zeitpunkt. Erst wenn das zu testende System analysiert werden kann und die Komponenten bekannt sind, können alle Rahmenbedingungen aufgestellt werden.

3.2.2 Berichterstellung

Die Erstellung eines Berichts stellt die wichtigste Aufgabe in Hinblick auf die Kundenzufriedenheit dar. Der Abschlussbericht jedes durchgeführten Versuchs repräsentiert das messbare Ergebnis der Projektarbeit. Zudem hinterlassen ein ansprechendes Design und eine vollständige Ergebnisdarstellung ein professionelles Bild und sollen den Gegenwert der Kosten widerspiegeln. Der Bericht ist das Stück Arbeit, das der Kunde sieht, bewertet, lobt oder kritisiert. Dementsprechend wichtig ist das Schriftstück als Präsentationsvorlage nach außen, auch in Bezug auf zukünftige Projektanfragen. Ein zufriedener Kunde gibt womöglich Empfehlungen ab, oder bleibt ein Kooperationspartner.

Dementsprechend soll einerseits das Augenmerk auf firmenrepräsentative Eigenschaften, andererseits auf für den Kunden relevante Inhalte gelegt werden. Jene Eigenschaften, die das Unternehmen widerspiegeln werden vielfach in einem Qualitätsmanagementsystem überwacht und vereinheitlicht. Dazu gehören beispielsweise Arbeitsanweisungen, welche die einheitliche Formatierung aller Berichte anhand einer Vorlage beschreiben oder die fortschreitende Namensgebung von Dokumenten (damit ein Kunde auf eine bestimmte Version des Berichts referenzieren kann). Aus Sicht des Kunden betrachtet ist vorrangig das Ergebnis eines Versuchs von Interesse. Dementsprechend gibt ein Überblick zu Beginn des Berichts - zumeist in tabellarischer Form - die wichtigsten Punkte der neuen Erkenntnisse mit entsprechender Referenz auf das jeweilige Kapitel bekannt. Auch das Versuchsdesign, die Beschreibung des Aufbaus und Ablaufs, sowie der Pilottest sind wichtiger Bestandteil des Schriftstücks, zumal der Kunde während des Projekts zumeist nur überblicksmäßig

informiert ist. Die Formulierung steht hierbei im Mittelpunkt, der Kunde muss in der Lage sein, die sachlichen Inhalte als externer Beobachter klar erfassen zu können.

Zumeist ist die Berichterstellung ein zeitintensiver Prozess, der im Fall von Hörversuchen jedoch großteils automatisiert werden kann. Der Hörraum und die projektleitende Firma sind konstante Faktoren, ebenso wie das Expert Listening Panel, das Testdesign und die Methode, sofern diese wiederholt zum Einsatz kommen. Die Dokumentation des Ablaufs, das Protokoll, welches für die interne Nachvollziehbarkeit geführt wird, liefert die entscheidenden Details für den jeweiligen Bericht, die in einer vorgefertigten Maske lediglich ergänzt werden müssen. Eine Berichtvorlage mit relevanten Kapiteln, wie Besprechungsergebnisse mit dem Kunden, Versuchsaufbau und –ablauf, Ergebnisdarstellung und analyse, etc. ist dieser Arbeit beigelegt.

3.3 Kostenkalkulation

3.3.1 Allgemeines zum Kostenaufwand

Die Kalkulation basiert auf dem Stundenaufwand der im Projekt beteiligten Mitarbeiter, was dem Projektleiter die Möglichkeit gibt, die geführte Stundenaufzeichnung (tatsächlicher Aufwand) innerhalb des Projekts mit den jeweils anberaumten Stunden zu vergleichen und somit zukünftige Projektbudgetierungen sinnvoll zu adaptieren. Für die Stundensätze in der Vorlage (gehört zu den im Anhang befindlichen Dokumenten) wird nach beruflichen Positionen innerhalb des Projekts unterschieden. Die Unterscheidung beispielsweise nach der Projektleiterin, die auch die Versuchsleiterin darstellt und dem Projektmitarbeiter erscheint mir aus Gründen der Nachvollziehbarkeit in der Kalkulation sinnvoll. Gleichzeitig wird dadurch ein Rahmen für die Zuständigkeitsbereiche innerhalb des Projekts geschaffen. Es ist aber auch möglich, einen mittleren Satz zur Berechnung heranzuziehen und durch Namensgebung für eine gute Übersicht zu sorgen.

Zusätzlich zu den Personalkosten werden auch Hardwarekosten, Durchlaufposten und Reisekosten berücksichtigt.

3.3.1.1 Personalkosten

Der Gruppenleiter setzt im Allgemeinen den anzuführenden Stundensatz für das Team fest. Ein Gewinnaufschlag von 20 % bei Personalkosten wird empfohlen und in der Vorlage ermittelt. Der Gesamtpreis plus Gewinnaufschlag gibt die Summe des anberaumten Stundenaufwands mit dem zugehörigen Stundensatz inklusive Gewinnaufschlag wieder. Es liegt ein Vergleich mit der Summe ohne Gewinnaufschlag vor.

3.3.1.2 Hardwarekosten

Mit Hardwarekosten sind sämtliche verwendete Materialien gemeint, auch Software und Raummiete. Materialien können entweder stückweise mit einer Pauschale verrechnet werden oder aber, bei größeren Anschaffungen, prozentuell. In diesem Fall ist mit dem Gruppenleiter Rücksprache zu halten. Die Raummiete und sonstige Abschreibungen von Hard- und Software werden anteilig und an den jeweiligen Projektumfang angepasst verrechnet.

3.3.1.3 Durchlaufposten

Hier werden sämtliche Positionen gelistet, die für das Testdesign von außen benötigt werden, an die aber im Rahmen des Projekts nur ein geringer Gewinnaufschlag geknüpft ist. Ein Durchlaufposten im Bereich des Versuchsdesigns wäre beispielsweise die Beauftragung der Statistikabteilung mit der Analyse der Testdaten, vgl. Abs. 3.3.8.1 *Statistik*.

3.3.1.4 Reisekosten

Dieser Bereich steht für Reisezeit, Taggeld, Übernachtungen, Kilometergeld etc. Dabei ist beim Taggeld auf den Tarif für das jeweilige Land, in dem man sich aufhält, und auf das aktuelle Jahr Rücksicht zu nehmen, ebenso beim Kilometergeld.

3.3.2 Allgemeines zur Vorlage

Die Kalkulation eines Versuchsdesigns erfolgt in der Regel nach dem ersten Kundenkontakt. Nichtsdestotrotz ist die Akquisitionsphase ein Bestandteil der im Anschluss durchgeführten Kalkulation. Diese besteht aus einem Richtpreisangebot, welches noch verhandelbar ist und weiteren Kalkulationsblättern, innerhalb derer Zeitaufwand, Gewinnaufschlag, Summe etc. berechnet werden.

Aus Gesprächen und Schriftverkehr sind Kundenwünsche bekannt und erste Informationen zum zu untersuchenden Produkt wurden eingeholt. Der Projektleiter soll zum Zeitpunkt der Erstangebotserstellung bereits einen Überblick über die Versuchsdurchführung, besser gesagt, die anwendbaren Methoden haben und diese dann im Zeitaufwand entsprechend einberechnen können. Dabei müssen unterschiedliche Überlegungen angestellt werden, die nachfolgend anhand der wichtigsten Eckpfeiler näher erklärt werden.

Die Überschriften decken sich mit den Positionen im Excel-Sheet, welches als Grundlage für Kostenkalkulationen auch in anderen Anwendungsbereichen dienen kann und beliebig erweiterbar ist.

3.3.3 Vorbereitung – Projektmanagement

Die gesamte Planungsphase des Projekts fällt in diesen Bereich. Der Projektleiter berücksichtigt hier einen Teil des Zeitaufwands der Projektakquise und stellt Überlegungen

zur Versuchsdurchführung an. Dabei können durchaus verschiedene Vorschläge, die für den Kunden von näherem Interesse sind, berücksichtigt werden. Diese kann man in getrennten Richtpreisangeboten ausarbeiten oder aber man stellt dem Kunden ein Basisangebot dar, welches durch mehrere Optionen erweitert werden kann (das würde bedeuten, dass die Optionen etwas günstiger kalkuliert werden, dann allerdings nur in Kombination mit dem Basisangebot erwerbbar sind).

Beispiel: Der Kunde bekommt einen Test mit Absolute Category Rating angeboten, schließlich möchte er nicht viel Geld ausgeben und lediglich wissen, welcher der beiden ANC-Ansätze besseren Anklang findet. Der ACR Test ist zwar ein kostengünstiger Test, da relativ unkompliziert, aber eben auch nicht sehr präzise in der Aussagekraft, da er statistisch betrachtet nur ordinal skaliert ist. Daher wird dem Kunden als Option ein zweiter Test angeboten, ein Degradation Category Test, der eine Aussage darüber liefert, in welchem Ausmaß der eine Algorithmus den anderen überwiegt. Dieser Test erfordert zwar ebenfalls eine Planungs-, Durchführungs- und Auswertephase, allerdings fällt der Part der Signalgenerierung weg und die Versuchsdauer verdoppelt sich nicht zwangsläufig. Der DCR kann also als Option etwas günstiger angeboten werden, als er als Standalone Variante angeboten würde. Bei Protokoll und Berichterstellung kann weiter an Zeit gespart werden.

3.3.3.1 Positionen

Nachfolgende Positionen sind in der beiliegenden Kostenkalkulation zum Projektmanagement angeführt.

Vorbereitung

Der Zeitaufwand der Projektplanungsphase wird erfasst. Versuchsplanung, Terminabstimmung, Besprechungen mit dem Kunden, etc. fallen in diese Kategorie.

Protokoll/ Testdesign erstellen

Die Zeit, die benötigt wird, um einen Versuchsablauf zu planen und wichtige Eckpfeiler bei der Durchführung schriftlich festzulegen, wird hier berechnet.

Diese Position eignet sich sehr gut, um im Preis zu „jonglieren“, sofern bereits eine vernünftige Vorlage (beispielsweise durch vorangegangene, gleiche oder ähnliche Testabläufe) für ein Protokoll existiert (vgl. Abs. 3.3.7, Dokumentation).

3.3.4 Expert Listening Panel

Den sinnvollen Einsatzbereich von Experten gegenüber den naiven Hörern behandelt das Kapitel 1.2.2 Versuchspersonen. Es ist durchaus möglich, dass sich der Kunde dezidiert naive Hörer aus Gründen der Empfehlungen von Standards wünscht. In diesem Fall fällt die Trainingszeit der Versuchspersonen jedenfalls weg. Andernfalls ist die zusätzliche Überlegung anzustellen, ob denn Training sinnvoll ist und auch in welchem Ausmaß. Diese Entscheidung wird zumeist durch die Art der Anwendung bestimmt und im Kapitel der

jeweiligen Methode unter Versuchspersonen besprochen. Jedenfalls ist es ratsam, bei komplexen oder ausgefallenen Problemstellungen an den Hörer, diesen im Vorfeld, wenngleich auch nur kurz, zu trainieren.

3.3.4.1 Positionen

Folgende Positionen werden für das ELP näher betrachtet.

Trainingszeit

Die Anzahl der Stunden, die trainiert wird und die Anzahl der Versuchspersonen (als Stück) sind zu berücksichtigen.

Versuchsdauer

Diese kann entweder pauschal abgegolten werden (1 Stunde, Stundenpreis = Pauschale), oder nach einem Stundenlohn und der jeweiligen Dauer des Versuchs. Die Einheit Stück steht hier wieder für die Anzahl der Versuchspersonen.

Pauschale für Panelrekrutierung

Die laufenden organisatorischen Tätigkeiten das Panel betreffend wie beispielsweise Terminplanung, Aussendung von Trainingseinheiten, werden in einer Pauschale abgegolten.

3.3.5 Signalgenerierung, Instrumentierung

In diesem Bereich werden sämtliche Vorbereitungen für den eigentlichen Versuch getroffen, mit Ausnahme von den Dokumentationen. Diese Vorbereitungen betreffen die zu generierenden Teststimuli genauso, wie die Hardwarekomponenten, die im Multimediaraum aufgebaut und miteinander verkabelt werden müssen. Der fertige Aufbau muss dann auf Funktionalität hin überprüft und kalibriert werden. Der Pilottest und die Datendarstellung fallen genauso in diesen Bereich wie anteilige Kosten für Hard- und Softwarekomponenten.

3.3.5.1 Positionen

Um die Signalgenerierung und Instrumentierung zu kalkulieren, werden die folgenden Positionen betrachtet.

Generierung der Teststimuli

Sobald diese Problemstellung softwaretechnisch zu realisieren ist, sollte die Dauer dieses Arbeitsprozesses mit den durchführenden Projektmitarbeitern abgesprochen werden.

Instrumentierung und Aufbau des Versuchs

Diese Position ist preislich sehr variabel gestaltbar. Bei einem Standardversuchsaufbau und unter der Voraussetzung, dass der Multimediaraum in seiner Standardaufstellung vorgefunden wird, sinkt der Aufwand auf ein Minimum.

Plausibilitätsprüfung des Aufbaus, Kalibrierung, Testmessungen

Je nach Komplexität des Versuchsaufbaus wird diese Position mehr oder weniger Zeit in Anspruch nehmen.

Pilottest

Bestandteil vom Pilottest sind nicht nur dessen Dauer und die Kosten der Versuchspersonen, sondern auch die graphische Darstellung der Ergebnisse der gewonnenen Daten im Vergleich zueinander nach dem Pilotversuch. Jene Graphiken sollen Aufschluss über mögliche Fehlerquellen geben und dann in Verbesserungsmaßnahmen ausgemerzt werden können. Dieser Bereich der Kalkulation ist als äußerst komplex anzusehen, da man bedenken muss, dass sich an diesem Punkt entscheidet, wie gut und ob der Versuch funktioniert. Es ist daher sinnvoll, hier etwas mehr Zeitaufwand einzukalkulieren.

Anteilige Kosten für Mikrofone, Verstärker, Software, Verbrauchsmaterial, etc.

Diese Kostenstelle eignet sich gut, um Abschreibungen neu gekaufter Hard- und Software zu tätigen. Soll der Versuch kostengünstiger ausfallen, kann auch eine geringe Pauschale verrechnet werden.

3.3.6 Versuchsdurchführung

Dieser Abschnitt erfasst den Zeitaufwand des Versuchsleiters, der jenem der Gesamtdauer aller Versuchspersonen gleichkommt, plus Vor- und Nacharbeiten. Die Experten oder naiven Hörer werden hier stundenmäßig nicht erfasst.

3.3.6.1 Positionen

In der Versuchsdurchführung wird lediglich die Arbeitszeit kalkuliert.

Arbeitszeit Versuchsleiter/Mitarbeiter

Je nach Komplexität und Intention des Versuchs kann es sogar erwünscht sein, einen Mitarbeiter mit versuchsleitender Funktion zu bestimmen, der nicht mit der Planung des Tests beschäftigt war. Die Mitarbeiterin sollte gut instruiert werden und es soll Sorge getragen werden, dass sich keine Fehler durch fehlende Instruktionen ergeben.

3.3.7 Dokumentation

Eine gut geführte Dokumentation ist Bestandteil jedes Protokolls und weiterführend auch jedes Abschlussberichts. Dabei ist es ratsam, bereits bei Protokollerstellung Randbedingungen und Besonderheiten, potentielle Fehlerquellen und Überlegungen, die zu diversen Entscheidungen führen, schriftlich zu erfassen, um zu einem späteren Zeitpunkt reproduzierbare Handlungsschritte aufweisen zu können. Das erleichtert nicht nur die Nachvollziehbarkeit durch Kolleginnen, sondern vereinfacht möglicherweise notwendige Verbesserungsmaßnahmen und verkürzt überdies den Zeitaufwand der Berichterstellung.

3.3.7.1 Positionen

Die Dokumentation wird in nachfolgende zwei Positionen aufgeteilt.

Dokumentation Versuchsaufbau/-durchführung

Hier wird die Zeit erfasst, die benötigt wird, um den Versuch mit seinen Randbedingungen (Raum, Lautsprecheranordnung, Signale, SW, etc.) und den Ablauf des Tests selbst zu beschreiben.

Die Position stellt einen Buffer für die Messberichterstellung (3.3.3.2 verweisen) dar und kann Fehleinschätzungen in der Zeitplanung teilweise kompensieren.

Dokumentation Datenanalyse und Ergebnisse

In dieser Position werden zwei Aspekte berücksichtigt. Der Versuchsleiter dokumentiert die Daten, beispielsweise mit Hilfe von graphischen Darstellungen, die Ergebnisse werden von der Statistik und/oder dem Versuchsleiter gemeinsam dokumentiert.

3.3.8 Datenanalyse, Ergebnis, Bericht

Die Berichterstellung nimmt im Allgemeinen recht viel Zeit in Anspruch und soll den Anforderungen eines Dokuments, das nach außen gerichtet ist, genügen. Der Bericht soll im selben Maß repräsentativ wie auch gut verständlich sein und den Kunden als „greifbares Endprodukt“ zufriedenstellen.

3.3.8.1 Positionen

Die Berichterstellung und Analyse kann je nach Anwendungsfall in folgenden Positionen genauer kalkuliert werden.

Graphiken

Der Zeitaufwand für die Darstellung der Datenmengen in Form von vergleichenden Graphiken kann hier erfasst werden.

Statistik

Der Zeitaufwand zur Datenaufbereitung und Berechnung der gesuchten Ergebnisse wird ermittelt. Erfolgt die Analyse extern oder wird intern an die Abteilung für Statistik weitergeleitet, fällt die Position nicht unter Personalkosten, sondern wird als Durchlaufposten eingeordnet.

Interpretation der Daten

Hier wird der Zeitaufwand für die eventuelle gemeinsame Interpretation von Ergebnissen und Diskussion der Ergebnisse ermittelt.

Berichterstellung, Ergebnispräsentation

Die Position dient der Zeiterfassung der formalen Niederschrift des kompletten Versuchsdesigns inklusive Ablauf, besondere Vorkommnisse und Ergebnisse.

3.3.9 Nachwort zur Kostenkalkulation

Verschiedene Positionen überschneiden einander in der Arbeitsausführung und dementsprechend auch im prognostizierten Stundenaufwand. Da gerade in der anfänglichen Planungsphase die Tendenz dahingeht, den Arbeitsaufwand geringer einzuschätzen als den tatsächlichen aufzuwendenden, ist diese aus der Kalkulation selbst entstehende Überbewertung als Puffer zu sehen. Dieser liegt in der Regel in einem engen Toleranzbereich, der nicht für Vergünstigungen und Rabatte verwendet werden sollte. Die Preisreduktion soll vielmehr folgendermaßen entstehen können: die Berichterstellung wird mit einem Arbeitsaufwand von einer Woche, also 38,5 Stunden, bewertet und in die Kalkulation auf 40 Stunden aufgerundet hinzugefügt. Korrekturarbeiten, die üblicherweise 2 Stunden benötigen, werden mit 4 Stunden anberaumt. Die Kalkulation fällt dementsprechend etwas großzügiger aus und kann bei Preisreduktionen, die direkt zu einer Stundenreduktion führen, einen realistischen Gesamtaufwand darstellen.

Zusammenfassung

Zu Beginn von Teil 1 der Arbeit, welcher Grundlagen und Voraussetzungen für das Versuchsdesign in der Psychoakustik behandelt, wurde auf für Hörversuche relevante Grundzüge aus der Statistik eingegangen. Neben Kennwerten zur Interpretation von Datensets wurden Möglichkeiten zur Datendarstellung aufgezeigt. Tests aus der parametrischen und nicht parametrischen Statistik wurden beispielhaft zur Analyse von Beurteilungen erläutert. Der nachfolgende Abschnitt behandelte Parameter des Versuchsdesigns, die innerhalb eines Projektablaufs benötigt werden. Die richtige Formulierung sowie der Unterschied zwischen Fragestellung und Hypothese wurden aufgezeigt und auf die Bedeutung von verschiedenen Arten von Versuchspersonen als Messinstrument wurde eingegangen. Das Expert Listening Panel liefert konsistente Urteile und deren Eignung für standardisierte Verfahren in denen naive Hörer gefordert werden, ist in Zukunft zu validieren. Weitere Parameter wie die korrekte Attributauswahl und die Deskription von Skalen wurden ebenso behandelt wie etwa die Rolle von spezifischen Teststimuli im Versuch. Auf die Wichtigkeit der korrekten Abschätzung der Teilnehmerinnenanzahl wurde in diesem Abschnitt ebenso hingewiesen wie auf die Rolle des Experiment Design als Ansatz zur Vermeidung von Fehlerquellen. Der letzte Abschnitt des ersten Teils beschäftigte sich mit der Rolle der Standardisierung und den im Rahmen des Handbuchs verwendeten Standards.

In Teil 2 der Arbeit liegt der Fokus auf der praxisnahen Deskription von standardisierten Methoden zur Beurteilung der Qualität von Sprach- und Audiosignalen. Dazu wurde neben der Unterteilung von Sprache und Audio eine Unterscheidung zwischen vergleichenden Verfahren und jenen Verfahren getroffen, die eine absolute Bewertung ohne Vergleich erzielen. Innerhalb der jeweiligen Methode wurde auf Besonderheiten in der Durchführung hingewiesen und der sinnvolle Anwendungsbereich, sowie der Versuchsablauf wurden beschrieben. Anwendungsbeispiele aus der Literatur, eine Bibliothek von Teststimuli und Attributlisten auf Deutsch und Englisch ergänzen diesen Teil der Arbeit.

Teil 3 der Arbeit liegen Vorlagen zur Berichterstellung und Kostenkalkulation zugrunde. Diese vervollständigen den Projektzyklus gemeinsam mit Erläuterungen zur Akquisitionsphase, Kalkulation und Berichterstellung. Zudem ist hier eine Checkliste enthalten, welche wegweisende Fragen zu den einzelnen Parametern des Versuchsdesigns aufweist.

Das Handbuch für Versuchsdesign in der Psychoakustik unterstützt damit den Versuchsleiter während des gesamten Projekts. Das Handbuch dient als Nachschlagewerk gleichermaßen wie als einführende Anleitung für Forscher mit wenig oder keiner Erfahrung auf diesem Gebiet und ermöglicht damit die Aneignung der Grundlagen zur Methodendurchführung und –auswertung. Es kann zur Methodenfindung ebenso herangezogen werden wie als Leitfaden während der Versuchsdurchführung und als Kommunikationshilfe in interdisziplinären Anwendungsbereichen.

Der technische Fortschritt, die Wissenschaft und auch die Unterhaltungsindustrie liefern ständig neue Aufgabenstellungen für die Bewertung psychoakustischer Faktoren durch den Menschen. Demnach sind auch die Methoden und Verfahren, Attributlisten und Skalierungsmethoden neuen Gegebenheiten anzupassen. Die Verwendung der Borg-Skala aus der Psychologie stellt beispielsweise eine vielversprechende Alternative zu den in der Arbeit beschriebenen Verfahren für die Zukunft dar, da sie eine feinere Auflösung der Sensitivität verspricht. Auch eine Erweiterung rein akustischer Beurteilungsverfahren auf den visuellen und haptischen Bereich ist realisierbar und damit die Beurteilung von neuen Medien und Technologien.

Literaturverzeichnis

[Asu12] ASUTAY E., VÄSTFJÄLL D., et al. Emoacoustics: *A Study of the Psychoacoustical and Psychological Dimensions of Emotional Sound Design.* Journal of the Audio Engineering Society, San Francisco, Vol. 60, No. 1/2, 2012, pp. 21-28

[Bec06] BECH S., ZACHAROV N.: *Perceptual Audio Evaluation – Theory, Method and Application.* John Wiley & Sons Ltd, England, 2006

[Bla12] BLAUERT Jens, JEKOSCH Ute: *A Layer Model of Sound Quality*. Journal of the Audio Engineering Society, Vol. 60, No. 1/2, 2012, pp. 4-12

[Bor04] BORTZ Jürgen: *Statistik für Human- und Sozialwissenschaftler*. Springer Medizin Verlag, Heidelberg, 6.Auflage, 2004

[Bra96] BRADLOW Ann R., TORRETTA Gina M., GISONI David B.: Intelligibility of normal speech I: Global and fine - grained acoustic - phonetic talker characteristics. Speech Communication, Vol 20, Issue 3-4, 1996, pp 255-272

[Coc72] COCHRAN William G.: *Stichprobenverfahren.* Walter de Gruyter Verlag, Berlin New York 1972

[Cot11] CÔTÉ Nicolaos: *Integral and diagnostic intrusive prediction of speech quality*. Springer Verlag, France, 2011

[Fel12] FELLBAUM Klaus: *Sprachverarbeitung und Sprachübertragung*. Springer Verlag Heidelberg Berlin, 2.Auflage, 2012

[Fle29] FLETCHER H., STEINBERG J.C.: *Articulation Testing Methods*. Bell System Technical Journal, Vol. 8, 1929, pp. 806-854

[Fra12] FRANK Matthias, SONTACCHI Alois: *Performance Review of an Expert Listening Panel.* DAGA 2012

[Gel04] GELFAND Stanley A.: *Hearing: An introduction to Psychological and Physiological Acoustics, 4th edition.* Marcel Dekker, New York, 2004

[Har05] HARTUNG Joachim: *Statistik. Lehr- und Handbuch der angewandten Statistik.* R. Oldenbourg Verlag München Wien, 14. Auflage, 2005

[Hay70] HAYS William Lee, WINKLER Robert L.: *Statistics, vol I & II.* Holt, Rinehart and Winston, New York, 1970

[Hoe97] HOEG W. et al.: *Subjective assessment of audio quality – the means and methods within the EBU.* European Broadcast Union Technical Review, Switzerland1997

[Hof04] HOFFMANN Jörg: *Taschenbuch der Messtechnik*. Fachbuchverlag Leipzig im Hanser Verlag, 4. Auflage, 2004

[Hun96] HUNTER Edward A.: *Experimental design*. Multivariate analysis of data in sensory science, Vol 16, Data handling in Science and Technology, Elsevier Verlag, 1996, pp 37-69

[Itu11] ITU-T HANDBOOK: *Practical procedures for subjective testing.* ITU 2011

[Jac13] [JACOBS] JACOBS Jörg: *Einführung in die Methoden der empirischen Sozialforschung. Quantitative Verfahren.* Skriptum zur Vorlesung, Universität Frankfurt 2013

[Jek10] JEKOSCH Ute: *Voice and speech quality perception. Assessment and Evaluation.* Springer Heidelberg Verlag, 2010

[Moe10] MÖLLER Sebastion: *Quality Engineering. Qualität kommunikationstechnischer Systeme.* Springer Verlag, Berlin Heidelberg, 2010

[Mon97] MONTGOMERY Douglas C.: Design and analysis of experiments. John Wiley & Sons, Inc. Verlag, USA, 4th edition, 1997

[Næs10] NÆS T., BROCKHOFF Per B., TOMIC: *Statistics for sensory and consumer science.* John Wiley & Sons Ltd. Verlag, Chichester, 2010

[Pfe11] PFEFFER Peter, HARRER Manfred: *Lenkungshandbuch*. Vieweg und Teubner Verlag, Springer Fachmedien Wiesbaden GmbH, Stuttgart, 2011

[Qua88] QUACKENBUSH S. R., BARNWELL T.P., CLEMENTS M.A.: Objective measures of speech quality, Prentice Hall Advanced Reference Series, New Jersey, 1988

[Raa12]] RAAKE A., WÄLTERMANN M., WÜSTENHAGEN U., FEITEN B.: *How to talk about speech and audio quality with speech and audio people.* Journal of the Audio Engineering Society, Vol. 60, No. 3, 2012, pp. 147-155

[Ris91] RISCH Jon M.: *A user-friendly methodology for Subjective Listening Tests.* AES Convention, New York, 4.-8.10. 1991; paper no: 3178

[Rot13] ROTHSTEIN Jules M: *On defining subjective and objective measurements*. Physical Therapy, Journal of the American Physical Therapie Association, Vol. 69, No. 7 , 1989, pp. 577-579, http://ptjournal.apta.org/content/69/7/577, verifiziert am 18. Juni 2013

[Rot69] ROTHAUSER E. H., CHAPMAN W. D.,IEEE Subcommittee: *IEEE Recommended practice for speech quality measurements.* IEEE Trans. Audio Electroacoust., Standards Publication No. 297, Vol. AU-17(3), 1969, pp. 225-246

[Sac02] SACHS Lothar: *Angwandte Statistik. Anwendung statistischer Methoden.* Springer Verlag, Berlin Heidelberg New York, 11. Auflage, 2002

[Sch12] SCHATZ Raimund, EGGER Sebastian, MASUCH Kathrin: The impact of test duration on user fatigue and reliability of subjective quality ratings. Journal of the Audio Engineering Society, Vol 60, No. 1/2, 2012, pp. 63-73

[Ste46] STEVENS S.S.: *On the theory of Scales of Measurement*, Science, New Series, Vol. 103, No. 2684 (1946), pp. 677-680, http://www.jstor.org/stable/1671815, verifiziert am 22. Mai 2013

[Zie08] ZIELINSKI S., RUMSEY F., BECH S.: On some biases encountered in modern audio quality listening tests – a review. Journal of the Audio Engineering Society, Vol. 56, No. 6, 2008, pp. 427-451

[800] RECOMMENDATION ITU-T P800: *Methods for subjective determination of transmission quality*. Geneva, 1996

[830] RECOMMENDATION ITU-T P830: *Subjective Performance Assessment of Telephone-Band and Wideband Digital Codecs.* Geneva 1996

[835] RECOMMENDATION ITU-T P835: *Subjective test methodology for evaluation speech communication systems that include noise suppression algorithm.* Geneva, 2003

[835-1] RECOMMENDATION ITU-T P835 *Amendment 1: New Appendix III: Additional provisions for nonstationary noise suppressors.* Geneva, 2007

[1116-1] RECOMMENDATION ITU-R BS.1116-1: *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems.* Geneva 1997

[1284-1] RECOMMENDATION ITU-R BS.1284-1: *General methods for the subjective assessment of sound quality.* Geneva, 2003

[3276] EBU TECH 3276 – 2nd edition: *Listening conditions for the assessment of sound programme material: monophonic and two-channel stereophonic.* European Broadcast Union, Technical Specification, Switzerland 1998

[3276s] EBU TECH 3276E – Supplement I: *Listening conditions for the assessment of sound programme material: Multichannel sound.* European Broadcast Union, Technical Specification, Switzerland 2004

[3286] EBU TECH 3286E: *Assessment methods for the subjective evaluation of the quality of sound programme material – Music.* European Broadcast Union, Technical Specification, Switzerland 1997

[3286s] EBU TECH 3286E Supplement I*: Assessment methods for the subjective evaluation of the quality of sound programme material – Multichannel.* European Broadcast Union, Technical Specification, Switzerland 2000

[5492] ÖNORM EN ISO 5492:2008 *Sensorische Analyse - Vokabular*., mehrsprachige Fassung Ausgabe 2009-12-01

[8586] ÖNORM EN ISO 8586-2:2008 *Sensorische Analyse - Allgemeiner Leitfaden für die Auswahl, Schulung und Überprüfung von Prüfpersonen. Teil 2 - Sensoriker.*; Anmerkung: zum Zeitpunkt der Verfassung der Diplomarbeit ist die aktuelle Version dieser Norm die ISO 8586:2012 Sensory analysis - General guidelines for the selection, training and monitoring of selected assessors and expert sensory assessors

Weiterführende Literatur

Anwendungsbeispiele

[Bah12] BAHNE Adrian: Perceived Sound Quality of Small Original and Optimized Loudspeaker Systems. Journal of the Audio Engineering Society, Vol. 60, No. 1/2, 2012, pp. 29-37

[Sch12] SCHINKEL-BIELEFELD Nadja, LOTZE Netaya, NAGEL Frederik: Does understanding of test items help or hinder subjective assessment of basic audio quality? Audio Engineering Society 133rd Convention, San Francisco, USA, 2012

[Sto00] STOLL G., KOZAMERNIK F.: EBU listening tests on Internet Audio codecs. EBU Technical Review, 2000

[Mar06] MARTENS William L., et al.: *Investigating Contextual Dependency in a Pairwise Preference Choice Task.* Audio Engineering Society 28th International Conference, Sweden, 2006

Attributlisten

[Fra09] FRANK Matthias: Perzeptiver Vergleich von Schallfeldreproduktionsverfahren unterschiedlicher räumlicher Bandbreite. Diplomarbeit, Graz 2009

[Sem08] SEMMLER Barabara: Subjektive Evaluierung von Mikrofonen. Diplomarbeit, Institut für elektronische Musik, Graz, 2008

Audio

[Too82] TOOLE Floyd E.: Listening Tests: Turning Opinion into Fact. . Journal of the Audio Engineering Society, Vol. 30, No. 6, 1982, pp. 431-445

Bradely-Terry-Modell und Bradley-Terry-Luce Modell

[Col80] COLONIUS Hans: Representation and uniqueness of the Bradley-Terry-Luce model for pair comparisons. British Journal of Mathematical and Statistical Psychology, Vol. 33, Issue 1, 1980, pp. 99-103

[Bra52] BRADLEY Ralph A., TERRY Milton E.: Rank analysis of incomplete block designs.: I. The method of paired comparisons. Biometrika, Vol. 39, 3/4, 1952, pp. 324-345

[Bra55] BRADLEY Ralph A.: Rank analysis of incomplete block designs.: III. Some large-sample results on estimation and power for a method of paired comparisons. Biometrika, Vol. 42, 3/4, 1955, pp. 450-470

[Wic04] WICKELMAYER Florian, SCHMID Christian: A Matlab function to estimate choice model parameters from paired-comparison data. Behavior Research Methods, Instruments, & Computers, Vol 36 (1), 2004, pp. 29-40

Flash Profile

[Lor05] LORHO Gaëtan*: Individual Vocabulary Profiling of Spatial Enhancement Systems for Stereo Headphone Reproduction.* Audio Engineering Society Convention Paper, No.:6629, USA, 2005

Kategorien, Skalierungstechniken

[Pre00] PRESTON C., COLMAN A.: Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta psychologica, No. 104, 2000, pp.1-15

[Roh78] ROHRMANN Bernd: Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. Zeitschrift für Sozialpsychologie, No. 9, 1978, pp. 222-245

[Roh07] ROHRMANN Bernd: Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data. Project Report, University of Melbourne, 2007

Multidimensional Scaling

[Wic03] WICKELMAIER Florian: An introduction to multidimensional scaling. Aalborg University, Denmark, 2003

Perceptual Structure Analysis

[Cho05] CHOISEL Sylvain, WICKELMAIER Florian: *Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound.* Audio Engineering Society Convention Paper, No.: 6369, Spain, 2005

Psychoakustik und Signal Detection Theory

[Gre88] GREEN David M., SWETS John A.: *Signal detection Theory and Psychophysics.* Peninsula Publishing, Los Altos Hills, USA, 1988

Reptory Grid Technique

[Ber99] BERG Ian, RUMSEY Francis: *Identification of Perceived Spatial Attributes of Recordings by Repetory Grid Technique and other Methods.* Audio Engineering Society Convention Paper, Munich, 1999

Sprache

[Laz07] LAZARUS H. SUST C., STECKEL R, KULKA M., KURTZ P.: *Akustische Grundlagen sprachlicher Kommunikation.* Springer Verlag Berlin Heidelberg 2007

Standards

[810] ITU-T Recommendation P 810: *Modulated Noise Reference Unit (MNRU).* 1996

[56] ITU-T Recommendation P 56: *Objective measurement of active speech level.* 2011

Statistik

[Mut06] MUTZ Michael: *Deskriptiv- und inferenzstatistische Modelle der sozialwissenschaftlichen Datenanalyse.* Seminarunterlagen aus dem Bereich Methoden der empirischen Sozialforschung, Universität Potsdam WS 2006/07, http://www.uni-potsdam.de/u/soziologie/methoden/mitarbeiter/shk/Michael/, verifiziert am 27.08.2013

[Ott10] OTT Lyman, LONGNECKER Michael: An Introduction to Statistical Methods and Data Analysis. 6th International Edition, Brooks/Cole, Canada, 2010

Thurstone V Modell

[Bra07] BRACHMANSKI Stefan: *Subjective Assessment ofQuality of Multimedia Signals by Means of A-B-Test.* Audio Engineering Society Convention Paper, No.: 7118, Vienna 2007

Glossar

Abszisse: x-Achse im kartesischen Koordinatensystem.

dBov: dB overload. Amplitude eines Signals im Vergleich zum Max. des Systems vor Clipping. (ähnlich dBFS)

Durchlauf (Run): Der Durchlauf testet ein spezifisches Attribut, oder eine Eigenschaft eines Systems und besteht aus mehreren Trials, innerhalb derer Grenzwerte der Wahrnehmung durch sukzessive Approximation (beispielsweise des Pegels) bestimmt werden.

Epistemologie: auch Erkenntnistheorie, ist ein Hauptgebiet der Philosophie und beschäftigt sich unter anderem mit der Frage: Wie kommt es zu Wissen an sich und neuen Erkenntnissen.

Grundgesamtheit: die zugrundeliegende, interessierende Menge, aus der die Stichprobe gezogen wird. Hier: ein Zielpublikum an Hörern, kann auch „die Menschheit“ sein.

Homomorph: gleichwertig, ähnlich. Die Überführung vom empirischen in das numerische Relativ erfolgt die Struktur erhaltend.

Konstruktivismus, radikaler: Strömung in der Philosophie des 20. Jahrhunderts, die postuliert, dass die Realität des Menschen nicht als objektiv erachtet werden kann, sondern lediglich eine Hirngeburt, also biologische Hirnfunktionen in Form von bijektivem Mapping darstellt [Bla12].

Korpus: Eine Bibliothek von Stimuli. Bei Sprache eine umfassende Darstellung der charakteristischen Merkmale von Sprache, die auch deren Variabilität durch Sprecherinnen einschließt.

Logatom: in Zusammenhang mit Hörtests versteht man darunter ein einsilbiges Kunstwort, das aus Sprachlauten zusammengesetzt wird und zur Silbenverstehbarkeitsmessung verwendet wird. [Vgl. Wikipedia, Logatom]

Metrische Daten: andere Bezeichnung für numerische, intervallskalierte Daten.

Ordinate: y-Achse im kartesischen Koordinatensystem.

Panel, Expert Listening Panel: Innerhalb dieses Dokuments eine Gruppe an Versuchsteilnehmern mit Expertise, die an einem Hörversuch teilnehmen.

Perzeptionismus: in diesem Zusammenhang: alles Existente ist eine Empfindung die mit den Hirnfunktionen in bijektivem Zusammenhang steht, siehe auch Konstruktivismus [Bla12]

Phonem: kleinste, bedeutungsunterscheidende Einheit der Sprache. Beispiel: rot - tot, für die inhaltliche Unterscheidung der beiden Worte sind die Phoneme „r“ und „t“ am Wortbeginn verantwortlich. [Vgl. Wikipedia, Phonem]

Prüfperson (sensory assessor): jede Person, die an einer sensorischen Prüfung teilnimmt. Anmerkung 1: Ein Laie ist eine Person, die kein bestimmtes Kriterium erfüllt. Anmerkung 2: Eine eingeführte Prüfperson hat bereits an einer sensorischen Prüfung teilgenommen. [5492]

Quality-of-Experience (QoE): die Gesamtakzeptanz eines Systems, vom Endabnehmer in Form einer Testperson subjektiv bewertet. [vgl. ITU-T P.10/G.100, Ammendment 2]; diese Beurteilungen entsprechen dem untersten Layer (aurale Kommunikationsqualität) des Layer Modells zum Qualitätsbegriff, siehe Kapitel 2.1.

Sample: Beispielstimulus, der innerhalb eines Hörtests abgefragt wird. Die Stimuli werden hierzu in ihren Merkmalen verändert und entstammen demselben Korpus. Das Sample verknüpft einen Stimulus mit einem System (Referenzsystem oder zu beurteilendes System).

Sensoriker (expert sensory assessor): Subst. ausgewählter Prüfer mit nachgewiesener sensorischer Empfindlichkeit und mit umfassender Schulung und Erfahrung hinsichtlich der sensorischen Prüfung, der in der Lage ist, verschiedene Prüfmaterialien widerspruchsfrei und wiederholbar sensorisch zu beurteilen. [5492]

Stimulus: Überbegriff zur Beschreibung von Signalen im hörbaren Frequenzbereich und deren Eigenschaften.

Trial: Ein Trial besteht aus dem Abspielen der Samples plus Beantwortung der Versuchsperson. Mehrere Trials formen einen Durchlauf (run).

Zufälliger Fehler: bei wiederholter Messung streut das Ergebnis um einen Mittelwert (hebt sich bei ∞ Wiederholung auf). Der z.F. kann in mindestens einem der Merkmale Amplitude, Vorzeichen oder Zeitpunkt seines Auftretens nicht vorhergesagt werden.[Hof04, Kap7]

Akronymliste

ACR	absolute category rating
ANC	acitve noise cancellation
CCR	comparison category rating
DCR	degradation category rating
EBU	European Broadcast Union
EL	exper listener
ELP	expert listening panel
H_0	Nullhypothese
H_1	Alternativhypothese
ITU	International Telecommunication Union
Vpn	Versuchspersonen
QRDT	quantal response detectability test
MUSHRA	multi stimulus with hidden reference and anchor

Index

Anhang Deutschsprachige Attribute

Nachfolgend sind deutschsprachige Attribute gelistet, sowie deren zugehörige Antwortattribute. Diese sollten jedenfalls mit den Versuchspersonen besprochen und adaptiert werden. Zudem kann in den Referenzen nachgelesen werden, für welchen Anwendungszweck sie ursprünglich verwendet wurden. Außerdem wird ausdrücklich darauf hingewiesen, dass die Qualität der ausgewählten Attribute kritisch zu hinterfragen ist.

Klangfarbe [Fra09]

Antwortattribute	
dünner	Voller
Weicher	härter
natürlicher	Künstlicher
dunkler	Voller

Räumliche Unterschiede [Fra09]

Antwortattribute	
Weniger räumlich	Stärker räumlich
Näher	Ferner
Schmaler	Breiter
Weiter links	Weiter rechts
Besser lokalisierbar	Schlechter lokalisierbar

Klangbild Mikrofon [Sem08]

Tiefenbetonung		
schwach	neutral	stark
Hohe Frequenzen		
dumpf	neutral	höhenbetont
Klangbild		
dumpf	neutral	scharf
dünn		Satt
verwaschen		transparent