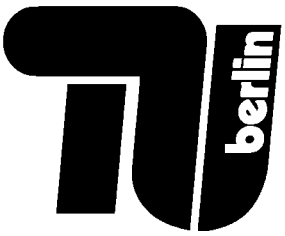
Technische Universität Berlin



Labor Kommunikationstechnik

Prof. Dr. Stefan Weinzierl

Perzeptive Messung und Evaluation

Dozent: Dr. Hans-Joachim Maempel

Kommunikationstechnik-Labor

WS 08/09

Hörversuche

Skript

Inhalt Seite

1 Der Mensch als Messinstrument

Das Ziel von Audiokommunikation ist letztlich immer die Wahrnehmung durch ein Lebewesen, wobei uns vor allem der Mensch interessiert. Die Wahrnehmung ist insoweit das endgültige Prüfkriterium für Audioübertragungssysteme und Audioinhalte. Damit stellt sich die Frage, wie wir Wahrnehmungsinhalte, die ja von außen kaum zugänglich sind, messen können. Der Mensch ist allein schon im auditiven Bereich ein sehr vielseitiges Messinstrument, denn er kann beispielsweise Auskunft geben über

- eigene basale Empfindungen ("Ton gehört"),
- eigene Gestalt- u. Objekterkennungsleistungen ("Ich habe ein Auto vorbeifahren gehört")
- eigene ästhetische Einschätzungen ("gefällt mir")
- eigene Emotionen ("Was ich höre, macht mich fröhlich")
- die Beschaffenheit der akustisch passiven Umgebung ("der Raum ist groß")
- Typ und Beschaffenheit von natürlichen Klangeerzeugern ("Pauke mit weichem Schlegel")
- Typ und Beschaffenheit von künstlichen Schallquellen ("ein kleiner Lautsprecher")
- die Qualität von akustischen Übertragungssystemen ("schlechte Leitung")
- die inhaltliche Bedeutung von Gesprochenem ("es ging um Leistungsscheine")
- die Befindlichkeit eines anderen sprechenden Menschen ("die Person ist aufgeregt")
- den beabsichtigten künstlerischen Ausdruck eines Musikers ("es sollte leicht klingen")

Indem der Mensch also über innere Zustände berichten, äußeren Objekten Eigenschaften beimessen und sogar technische Übertragungsprozesse erkennen und in ihrer Qualität einschätzen kann, kommen folgende Untersuchungsobjekte in Frage:

- Die menschliche Wahrnehmung selbst (z.B. Hörschwelle, Emotionen)
- Audioinhalte (z.B. Musikaufnahmen, Sprecher/innen)
- Übertragungssystemkomponenten (z.B. Codecs, Lautsprecher, Räume)

Die klassische Psychoakustik begann naheliegenderweise mit der Erforschung des menschlichen Hörvermögens in bezug auf seine Empfindungsgrenzen; Kommunikationswissenschaft und Audiogeräteindustrie interessieren sich heute naturgemäß für die Qualität von Übertragungskomponenten; und die psychologische und ästhetische Forschung möchte vor allem ermitteln, wie bestimmte Audioinhalte wirken.

2 Methodik

Hörversuche sind grundsätzlich experimentelle Verfahren: Eine mutmaßliche Einflussgröße (unabhängige Variable) wird gezielt durch Manipulation oder Selektion verändert (z.B. der Signalpegel), und die Ausprägung einer anderen Größe (abhängige Variable) wird gemessen (z.B. der Lautstärkeeindruck). Soll dieses Vorgehen zielgerichtet sein, setzt dies eine begründete Vermutung über den Wirkungszusammenhang voraus. Außerdem muss mit anderen Einflussgrößen (moderierenden Variablen) umgegangen werden: Sie werden entweder nicht beachtet (Störvariablen), ebenfalls gemessen (Kontrollvariablen), ausgeschaltet oder konstant gehalten.

Aus der klassischen Testtheorie ergeben sich die drei Testgütekriterien Objektivität, Validität und Reliabilität. Objektivität bezeichnet die unverzerrte, u.a. vom Forscher unbeeinflusste, 'reale' Beschreibung des Untersuchungsgegenstands. Die Validität (Gültigkeit) fordert, dass eine Messung auch die Merkmale erfasst, die man zu erfassen beansprucht. Reliabilität (Zuverlässigkeit) meint die Wiederherstellbarkeit eines Messergebnisses unter denselben Bedingungen, was u.a. eine bestimmte Messgenauigkeit voraussetzt.

2.1 Objektivität

Eine Besonderheit psychologischer Messmethoden gegenüber physikalisch-technischen ist die vergleichsweise hohe Unterschiedlichkeit individueller Wahrnehmungen. Diese Subjektivität läuft zunächst dem wissenschaftlichen Objektivitätsanspruch zuwider. Allerdings kann gemäß dem Kritischen Rationalismus Objektivität auch in den Natur- und Ingenieurswissenschaften nur ein unerfüllbares Ideal sein, da dieses Kriterium selbst der subjektiven Einschätzung unterliegt. Daher wird Objektivität in den empirischen Wissenschaften als Intersubjektivität aufgefaßt und hergestellt. Im Falle menschlicher Wahrnehmungen kann diese Intersubjektivität durch Gruppierung einer Vielzahl von Individuen zu einem einzigen Messinstrument erreicht werden, das bei richtiger Benutzung eine ähnlich hohe Objektivität aufweist wie technische Messinstrumente. Die verbleibende Versuchsleitersubjektivität spielt bei standardisierten und eindeutig ablesbaren Erhebungsinstrumenten in beiden Bereichen kaum eine Rolle. Die in der Literatur häufig anzutreffende Bezeichnung "subjektive Messung" ist insoweit unzutreffend, denn es handelt sich um eine insgesamt weitgehend objektive Messung auf der Basis mehrerer Subjekte. Sie suggeriert darüber hinaus eine geringere wissenschaftliche Belastbarkeit, eine Ansicht, die ebenfalls nicht haltbar ist.

2.2 Validität

Denn die Verwendung einer größeren Anzahl von Individuen und eines sorgfältig erstellten Fragebogens weist wenigstens für die Erfassung auditiver Phänomene naturgemäß eine weitaus größere Validität auf als technische Messungen. So kann z.B. ein technisches Verfahren selbst bei rechnerischer Berücksichtigung von Maskierungseffekten den einfachen Wahrnehmungseindruck *Lautheit* nur unzureichend schätzen; die von Menschen abgefragten Werte haben hingegen maximale Gültigkeit: In sie gehen nicht nur die natürlicherweise wirksamen Maskierungseffekte ein, sondern z.B. auch die Akustik der Abhörsituation und die Bedeutung des Audioinhalts. Erst recht komplexere Klangmerkmale (z.B. Transparenz) können von Modellen nur noch unzureichend erklärt werden. Insoweit können physikalisch-technische Messungen immer nur Indikatoren für Wahrnehmungsinhalte sein. Gründe für deren Einsatz sind vielmehr Praktikabilität, Zeit- und Kostenersparnis, die Anwendbarkeit physikalischer Maßeinheiten und die hohe Reliabilität.

2.3 Reliabilität

Indem die perzeptive Messung eines Merkmals stets die Beteiligung des ganzheitlich arbeitenden Wahrnehmungsapparates erfordert, bestimmen nicht nur die Messbedingung, sondern auch andere Faktoren die Messwerte. Dieser Umstand ist einerseits für die hohe Validität verantwortlich, mindert aber andererseits die Reliabilität: Das Ergebnis einer perzeptiven Messung ist nicht so genau reproduzierbar wie das einer technischen Messung. Daran wird

deutlich, dass auf dem Gebiet der Wahrnehmung, wie überhaupt im Bereich der Lebenswissenschaften, keine deterministischen Zusammenhänge ermittelbar sind (möglicherweise auch gar nicht gelten), sondern probabilistische. Vorhersagen auf der Basis von Hörversuchsdaten werden also zwangsläufig eine gewisse Ungenauigkeit aufweisen.

3 Historischer Hintergrund: Psychophysik

Mitte des 19. Jahrhunderts fragte sich Gustav Theodor Fechner, wie man innere Empfindungen messen könne. Die Lösung bestand in einer in diesem Bereich wohl erstmals bewusst eingesetzten Operationalisierung im Sinne des Wortes: Versuchspersonen müssen beim Zutreffen einer inneren Bedingung eine Operation ausführen, etwas tun. Z.B. "ja" sagen, einen Regler verstellen, einen Knopf drücken. Auf diese Weise lassen sich als Maßeinheiten für Empfindungen einfach die physikalischen Maßeinheiten der entsprechenden Reize verwenden. Diesen Ansatz bezeichnete Fechner als Psychophysik. Das Teilgebiet der Psychophysik, das elementare akustische und auditive Vorgänge auf diese Weise in Beziehung setzt, bezeichnet man als Psychoakustik.
Um eine Skala festzulegen, braucht man einen Nullpunkt. Dieser entspricht der *Reizschwelle* (absolute Schwelle), welche ermittelt werden muss. Eine geeignete Einheit der Skala ist der ebenmerkliche Unterschied (just noticeable difference JND), auch Unterschiedsschwelle genannt. Man stellt empirisch fest, dass das Verhältnis von eben wahrnehmbarer Reizänderung zu Reizschwelle konstant ist (Webersches Gesetz). Allgemein folgt der Zusammenhang zwischen Reiz- und Empfindungsstärke dem Stevensschen Potenzgesetz. Der Exponent für den Zusammenhang von Schalldruck und Lautheit ist n=0,6. Die mit steigender Reizstärke sinkende perzeptive Empfindlichkeit ist Zeichen der Adaptionsfähigkeit der meisten Sinne.

Webersches Gesetz

$$\frac{\Delta S}{S_0} = \text{konst.}$$

ΔS: Eben wahrnehmbarer Stimulusunterschied
S_0: Stimulusstärke

Stevenssches Potenzgesetz

$$E = k \cdot S^n$$

E: Empfindungsstärke
S: Stimulusstärke
k: Sinn-, reizspezifischer Faktor
n: Sinn-, reizspezifischer Exponent

4 Klassische Verfahren zur Bestimmung von Reiz- / Unterschiedsschwellen

4.1 Herstellungsverfahren

Die Versuchspersonen regeln die (meist kontinuierlich variierbare) Reizintensität selbst, bis eine definierte subjektive Empfindung vorhanden ist (z.B. "gerade eben hörbar").

4.2 Grenzverfahren

Der Versuchsleiter regelt die Reizintensität für die Versuchsperson (kontinuierlich oder diskret), bis eine definierte Empfindung vorhanden ist. Man kann sich der definierten Empfindung/Schwelle von unten oder von oben nähern (aufsteigendes/absteigendes Grenzverfah-

ren). Das arithmetische Mittel von auf- und absteigenden Ergebnissen gibt die Reizschwelle an. Ein automatisiertes, auf Rückmeldungen der Versuchsperson reagierendes auf- und absteigendes Grenzverfahren ist die *Békésy-Audiometrie*.

4.3 Konstanzverfahren

Es werden diskret und äquidistant variierte Reize, deren Intensitäten um die erwartete Schwelle verteilt sein müssen, in zufälliger Reihenfolge dargeboten. Die Versuchsperson soll angeben, ob eine definierte Empfindung vorhanden ist. Es erfolgen 10 bis 15 Durchgänge. Trägt man die Trefferquote über die Bedingungsvariation auf, ergibt sich eine psychometrische Funktion (Abb. 1). Die gesuchte Schwelle ist diejenige Bedingung, die 50% Trefferquote aufweist. Bei der Ermittlung der Unterschiedsschwelle soll die Versuchperson bestim-

Psychometrische Funktion
Bestimmung einer Reizschwelle

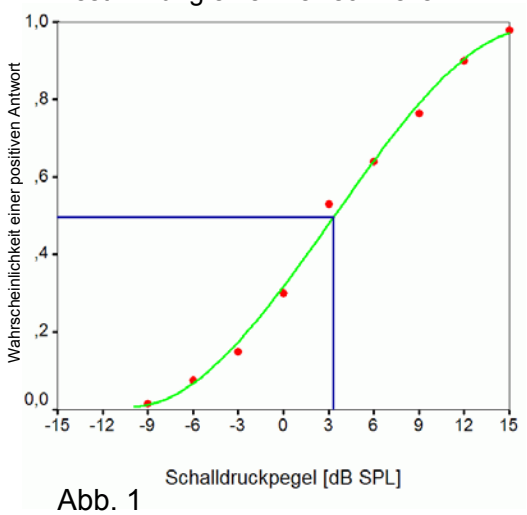


Abb. 1

Psychometrische Funktion
Bestimmung einer Unterschiedsschwelle

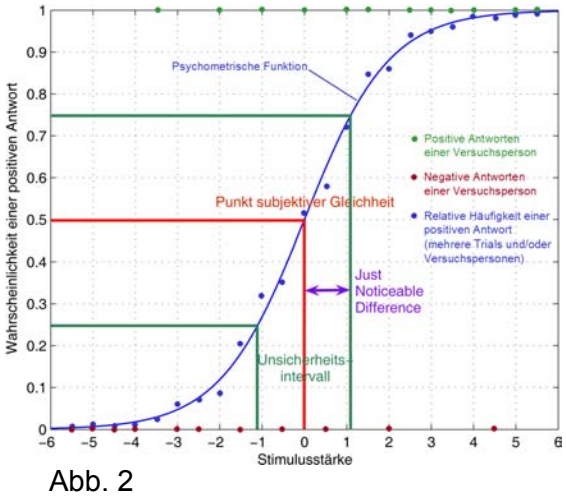


Abb. 2

men, ob ein Testreiz z.B. lauter oder leiser als ein konstanter Vergleichsreiz (Standardreiz) ist. Man kann aus 10-15 solcher *trials* nun den Punkt Subjektiver Gleichheit (PSG) bestimmen (Abb. 2). Die Abweichung dieses Wertes vom Vergleichsreiz ist der konstante Fehler. Er beträgt im Beispiel 0,5dB bei einem Vergleichsreiz von 3dB (Abb. 2). Als Unsicherheitsintervall UI ist der Bereich zwischen 25% und 75% der positiven Antworten definiert. Das halbe Unsicherheitsintervall ist die Just Noticeable Difference JND. Sie beträgt im Beispiel 4,5dB (Abb. 2).

4.4 Probleme der klassischen Verfahren

Die drei Verfahren weisen unterschiedliche Vor- und Nachteile auf. Beim Herstellungs- und Grenzverfahren kennen die Versuchspersonen die Spielregeln (die Richtung der Reizänderung) und neigen zu Antizipationsfehlern (zu frühe Entscheidung) oder Habituationsfehlern (zu späte Entscheidung). Beim Herstellungsverfahren beeinflusst außerdem die Gewissenhaftigkeit der Versuchspersonen das Ergebnis. Das Konstanzverfahren ist zufällig und damit unwissentlich. Es vermeidet dadurch obige Fehler. Es ist allerdings sehr zeitintensiv. Ein Problem ist die Vorgabe geeigneter Reizstärken im interessierenden, aber vorab unbekann-

ten Bereich. Dies kann durch Verfahren minimiert werden, die auf die Erkennungsleistung der Versuchsperson reagieren, sogenannte adaptive Verfahren.
Ein grundsätzliches Problem der drei klassischen Schwellenmessverfahren ist das Kriterienproblem. Die Versuchsperson kann ein strenges (konservatives) oder laxes (progressives) Kriterium setzen. Von diesem Kriterium hängen Fehlerhäufigkeit und ermittelte Schwelle ab. Es resultiert eine Antworttendenz (response bias). Die Messung weist also einen sensorischen Anteil und einen Entscheidungsanteil auf, die jedoch beide konfundiert sind. Für die Lösung dieses Problems gibt es zwei Ansätze:

- Criterion-free procedures (Forced-Choice-Verfahren) minimieren die Störeinflüsse.
- Getrennte Erhebung von sensorischer und Entscheidungs-Komponente nach der Signalentdeckungstheorie. Hierauf wird in diesem Skript nicht weiter eingegangen (vgl. hierzu Bortz 2005, Gelfand 2004 und Hellbrück & Ellermeier 2004).

5 Typische moderne Hörversuchsverfahren (Auswahl)

Mittlerweile existieren eine Vielzahl von Bezeichnungen für Hörtests (z.B. Forced Choice, ABX, ABCHR, 4AFC, MUSHRA, PEST). Die verschiedenen Verfahrenstypen lassen keine handhabbare Systematisierung zu, denn durch sie werden ganz verschiedene experimentelle Aspekte festgelegt, impliziert oder offen gelassen: Bedingungsvariation, Versuchsdesign, Versuchsablauf, erhobene Merkmale, Skalenniveau, Auswertungsmethode. Dies gilt auch für die unter Abschnitt 6 genannten Verfahren.

5.1 Einfache Forced-Choice-Verfahren

Forced-Choice-Verfahren erzwingen durch die Darbietung mindestens zweier Reize, von denen stets einer der experimentell veränderte ist, eine Entscheidung der Versuchperson. Es gibt also objektiv richtige und falsche Antworten. Die Richtigkeit kann der Versuchsperson zurückgemeldet werden (feedback). Sie muss kein Kriterium setzen, ab welchem Reizunterschied sie "ja" sagt. So kann sie auch kleinste Reizunterschiede angeben, da unter der Reizschwelle Fehler sowieso unvermeidlich sind. Die Versuchsperson ist also von dem Druck befreit, keine Fehler machen zu dürfen. Das Problem der Wahl geeigneter Reizstärken bleibt jedoch erhalten. Daher werden Forced-Choice-Verfahren heute in der Regel mit automatischer Reizstärkenanpassung (adaptiv) durchgeführt (vgl. 5.2).

ABX-Test
Eine populäre Variante der Forced-Choice-Verfahren ist der ABX-Test. Durch diesen einfachen Blindtest kann ermittelt werden, ob eine Versuchsperson zwei verschiedene Reize zuverlässig unterscheiden kann. Es werden drei Stimuli angeboten: A und B unterscheiden sich, bilden also zwei gekennzeichnete und klingende Bezugssignale, während X zufällig entweder mit Reiz A oder Reiz B identisch ist. Die Versuchsperson soll nun die Identität aufgrund ihres Höreindrucks zuordnen, indem sie X=A oder X=B wählt (vgl. Abb. 3). Dies kann korrekt oder falsch geschehen. Der Versuch wird wieder-

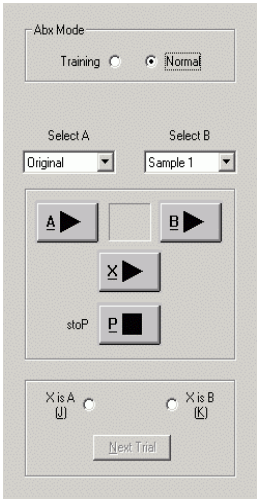


Abb. 3

holt und sollte insgesamt etwa 15-20 mal durchgeführt werden. Die mehrmalige Durchführung ermöglicht die Berechnung einer statistischen Signifikanz aus den Trefferhäufigkeiten und der Anzahl der *trials*. Es läßt sich also ermitteln, ob die Versuchsperson geraten oder aber den Unterschied der Stimuli überzufällig erkannt hat. Der ABX-Test ist relativ schnell durchführbar und sehr empfindlich. Er liefert aber keine skalierten Daten, sondern nur die Aussage "Unterschied erkannt" versus "Unterschied nicht erkannt" für ein Stimuluspaar. Soll eine mehr als zweistufige Bedingungsvariation getestet werden, wird der ABX-Test zu einem mehrfachen Paarvergleich und mit steigender Anzahl zu vergleichender Stimuli schnell sehr zeitintensiv.

ABC/HR-Test

Der auch sogenannte Triple-Stimulus-Hidden-Reference-Test wird häufig für die Evaluation von Codecs eingesetzt. Optisch ist nur der Vergleichsreiz kenntlich gemacht (Referenz), nicht der (schlechtere) Testreiz. Außerdem stehen die nochmalige Referenz und der Testreiz in zufälliger Reihenfolge zur Verfügung. Die Versuchsperson soll entscheiden, welcher der beiden ungekennzeichneten Reize die veränderte Version ist. Üblicherweise wird außerdem der Grad der subjektiven Störung miterhoben (vgl. ITU 1997), der mittels eines Schiebereglers quantitativ festgelegt werden kann (vgl. Abb. 4). Ist ein solches Rating vorgesehen, werden die zu beurteilenden Bedingungen meist gleichzeitig (und zufällig gemischt) angeboten (vgl. die acht Gruppen in Abb. 4). Die Versuchsperson kann beliebige Stimuli so direkt miteinander vergleichen, was für ein stabiles Gefüge von Beurteilungswerten förderlich ist. Das Forced-Choice-Paradigma wird in diesem Verfahren also mit der Erhebung eines absoluten Urteils verknüpft. Da eine 'falsche' Wahl möglich ist, die auch das Rating betrifft (die Referenz wird als schlechter eingestuft), wird für die Auswertung zunächst die Differenz zwischen Testreiz- und Referenzbeurteilung gebildet (*diffgrade*). Positive Differenzgrade zeigen an, dass der Vergleichsreiz besser als die Referenz beurteilt wurde. Die Differenzgrade werden in der Regel für die Gruppen getrennt deskriptiv-statistisch ausgewertet (arithmetisches Mittel, Standardabweichung, Konfidenzintervall), aber auch der Einsatz inferenzstatistischer Methoden (z.B. Varainzanalyse) ist möglich. Kritisch zu thematisieren sind dabei u.a. die Vertretbarkeit der Intervallskalenannahme, wenn alle fünf Ratingstufen bezeichnet sind, und die Dimensionalität des Merkmals.

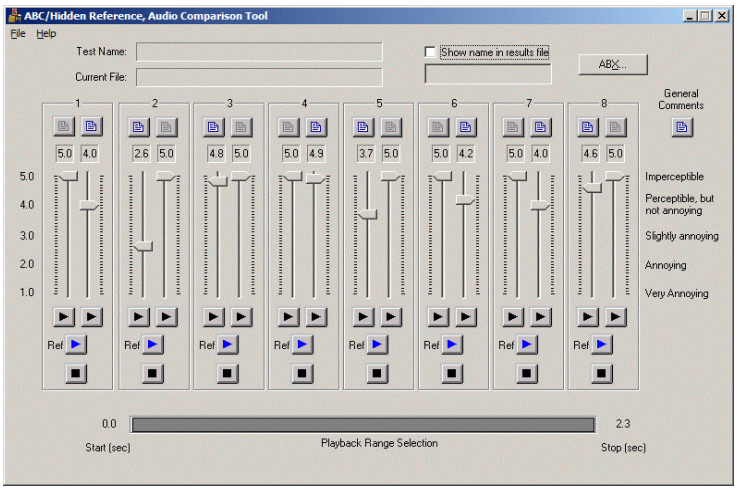


Abb. 4

5.2 Adaptive Verfahren

Adaptive Verfahren sind in der Regel als wiederholte Forced-Choice-Versuche ausgelegt, die dann einzeln als trials bezeichnet werden. In einem trial werden zwei oder mehrere Reize (sog. Intervalle oder Alternativen) vollständig blind angeboten, Vergleichsreiz(e) und Testreiz

sind also zufällig versteckt (vgl. Abb. 5 als Beispiel für eine Versuchsoberfläche eines trials mit 3 Intervallen). Die Anzahl der Intervalle und die Aufgabe der Versuchsperson (Erkennung) werden als *Versuchsparadigma* bezeichnet. Das einem adaptiven Forced-Choice-Verfahren zugrunde liegende Paradigma wird abgekürzt nach der Anzahl der Stimuli/Intervalle/Alternativen benannt, z.B. als 2AFC (zwei Intervalle, Forced Choice). Über die Bedeutung des "A" gibt es unterschiedliche Auffassungen: Nach Gelfand (2004) steht "A" für "alternative" (S.259), nach Hellbrück und Ellermeier (2004) für "adaptive" (S.230). Es wird daher empfohlen, das verwendete Versuchsparadigma auszuschreiben.

Abb. 5

Adaptive Verfahren reagieren nun auf die Detektionsleistung der Versuchsperson, indem sie die Reizstärke für den nächsten trial entsprechend anpassen. Die Anpassung kann schrittweise durch eine adaptive Regel erfolgen (sog. Staircase-Verfahren) oder aber ggf. auch sprungweise durch Darbietung der statistisch wahrscheinlichsten Schwelle (Maximum-Likelihood- und Bayes-Verfahren, z.B. The Best PEST, QUEST, ZEST), für deren Schätzung allerdings eine psychometrische Modellfunktion und deren Parameter angenommen werden müssen (daher auch als parametrische Verfahren bezeichnet). Die Messstrategie (Startbedingung, Adaptionssregel und Abbruchkriterium) und die bei den Maximum-Likelihood-Verfahren nötige Berechnungsvorschrift für die Schwellenschätzung werden zusammenfassend als *Versuchsmethode* bezeichnet. Im folgenden werden nur die (nichtparametrischen) Staircase-Verfahren beschrieben.

Bei adaptiven Forced-Choice-Verfahren ist eine zufällige Erkennung des Testreizes durch Raten möglich. Daher verschiebt sich gegenüber einem adaptiven Ja-Nein-Paradigma (kein Forced-Choice-Paradigma wegen nur einem Intervall) die untere Asymptote um die nun von der Intervallzahl abhängige Ratewahrscheinlichkeit γ nach oben, und die psychometrische Funktion wird gestaucht. Die gesuchte sensorische Schwelle θ, der Wendepunkt, liegt daher bei einer höheren Häufigkeit richtiger Antworten. Abb. 6 zeigt drei psychometrische Funktionen, jeweils mit Wahrscheinlichkeit positiver bzw. korrekter Antwort am Wendepunkt und Ratewahrscheinlichkeit. Dabei ist zusätzlich die Wahrscheinlichkeit λ von Lapsus-Fehlern (falsche Wahl trotz richtiger Erkennung) berücksichtigt. Will man von der beobachteten Antwortwahrscheinlichkeit $\hat{\psi}(x)$ auf die eigentlich interessierende Wahrscheinlichkeit $\hat{\psi}^*(x)$ der sensorischen Detektion schließen, muss erstere 'ratekorrigiert' werden nach der Formel $\hat{\psi}*(x) = (\hat{\psi}(x) - \gamma)/(1 - \gamma - \lambda)$.

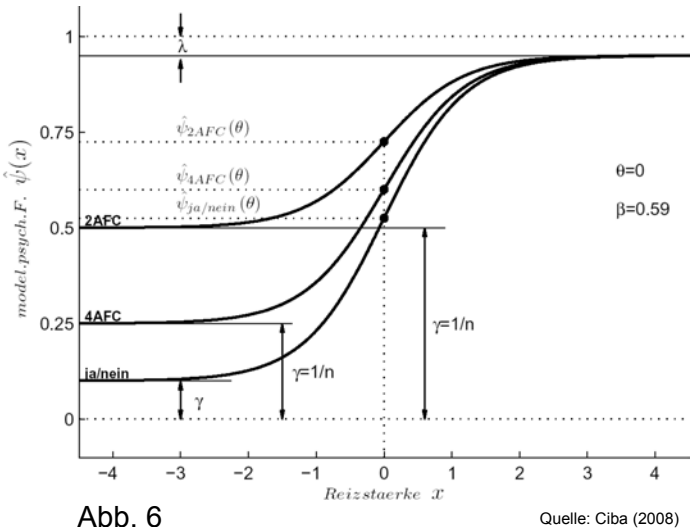


Abb. 6 Quelle: Ciba (2008)

Einfache Up-Down-Verfahren

Einfache Up-Down-Verfahren erhöhen die Reizstärke bei einer falschen Entscheidung der Versuchsperson und vermindern sie bei einer richtigen Entscheidung. Diese Regel ist allerdings für die Versuchspersonen leicht durchschaubar.

Transformed Up-Down-Verfahren

In transformierten Up-Down-Verfahren kommen unsymmetrische Adaptionsregeln zur Anwendung: Beispielsweise wird die Reizstärke schon bei einer falschen Entscheidung der Versuchsperson erhöht, jedoch erst bei zwei aufeinanderfolgenden richtigen Entscheidungen vermindert (2-Down/1-Up-Regel). Einen entsprechenden Messverlauf zeigt Abb. 7. Die Schwelle wird durch den Mittelwert der Umkehrpunkte geschätzt, wobei man sich auf die letzten Umkehrpunkte beschränken kann.

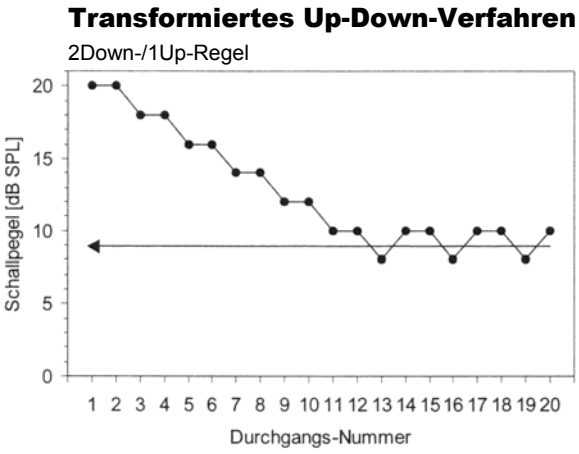


Abb. 7 Quelle: Hellbrück und Ellermeier (2004)

Die Adaptionsregel beeinflusst das Konvergenzniveau des Testverfahrens, d.h. die Wahr-Wahrscheinlichkeit positiver bzw. richtiger Antworten (auch *Korrekt-Schwelle* genannt), bei der sich der Versuchsverlauf einpendelt, und die die Schätzung des Schwellwertes mit sich bringt. Beispielsweise ist es nach der 2-Down/1-Up-Adaptionsregel leichter, einen Fehler zu machen, als zweimal nacheinander richtig zu urteilen. Das Verfahren konvergiert daher nicht bei 0,5. Konvergenz liegt vor, wenn die Wahrscheinlichkeit, sich an die Schwelle anzunähern (p^2 entsprechend 2 richtigen Antworten in Folge), gleich groß ist wie die Wahrscheinlichkeit, sich zu entfernen ($1 - p$ entsprechend einer falschen Antwort + $p \cdot (1 - p)$ entsprechend einer richtigen und einer falschen Antwort in Folge). Nach Gleichsetzung beider Terme und Umformung ergibt sich für das Beispiel $p = \sqrt{0{,}5} = 0{,}707$ als Konvergenzniveau (vgl. hierzu Hellbrück und Ellermeier 2004, S.228-229).

Adaptionsregel	Konvergenzniveau	'ratekorrigiertes' Konvergenzniveau		
	ja/nein	2AFC	3AFC	4AFC
1-Down/1-Up	0,5	0	0,25	0,333
2-Down/1-Up	0,707	0,414	0,561	0,609
3-Down/1-Up	0,794	0,588	0,691	0,725

Tab. 1 Quelle: Ciba (2008)

Da die gesuchte Schwelle bei $\hat{\psi}^*(\theta) = 0{,}5$ liegt, verschätzen sich Verfahren mit einem abweichenden Konvergenzniveau. Bei Forced-Choice- bzw. nAFC-Paradigmen ist außerdem die Ratekorrektur zu berücksichtigen. Die sich so für verschiedene Paradigmen ergebenden 'ratekorrigierten' Konvergenzniveaus zeigt neben den einfachen Konvergenzniveaus Tab. 1, gültig für $\gamma = 1/n$ und $\lambda = 0$. Ein einfaches 2AFC-Verfahren (1-Up/1-Down) pendelt sich demnach bei $p = 0{,}5$ ein, bestimmt damit aber den nicht interessierenden 0%-Punkt der sensorischen psychometrischen Funktion $\hat{\psi}^*(x)$ (vgl. Abb. 6). Das 3AFC-Paradigma bestimmt unter Verwendung einer 2-Down/1-Up-Regel hingegen mit $\hat{\psi}^*(\theta') = 0{,}561$ einen Punkt dieser Funktion, der nah am idealen 50%-Punkt liegt.

6 Allgemeinere psychologische Verfahren (Auswahl)

Je nach Aufgabenstellung werden neben den klassischen und modernen psychoakustischen Testverfahren auch allgemeinere psychologische Verfahren für Hörversuche eingesetzt. Hierbei geht es um die Erfassung von Merkmalen, die Ergebnis höherer psychischer Verarbeitungsprozesse sind, z.B. komplexe Klangeigenschaften, ästhetische Eindrücke, Präferenzen, Ähnlichkeiten, Assoziationen oder Emotionen – ggf. auch unter Berücksichtigung visueller Inhalte bzw. der Bearbeitung multimodaler Fragestellungen, z.B. zur Synchronizität (vgl. Rudloff 1997) oder zum *matching* (vgl. Iwamiya 1994). Hörbeispiele sind hier zumeist nicht nur Reize, die Zustände im Untersuchungsobjekt Mensch auslösen, sondern selbst Untersuchungsobjekte, denen vom Menschen Eigenschaften zugeschrieben werden. Dies bedingt stets eine dreidimensionale Datenstruktur *Merkmale* × *Versuchspersonen* × *Konzepte (Bedingungen/Objekte)*. In der Regel wird in diesem Bereich nicht mit wenigen Versuchspersonen gearbeitet, die eine Vielzahl von *trials* mit denselben oder ähnlichen Stimuli durchlaufen, sondern mit vergleichsweise vielen Versuchspersonen und wenigen *trials* oder gar nur einem. Dementsprechend werden Siginifikanztests nicht für die Beobachtungen einer Versuchsperson, sondern für die gesamte Anzahl von Versuchspersonen unter den interessierenden Versuchsbedingungen gerechnet. Man unterscheidet dabei zwischen anderen Versuchspersonen unter jeder Versuchsbedingung (unabhängige Stichproben) und denselben Versuchspersonen unter jeder Bedingung (abhängige Stichproben, Messwiederholung). Je nach Fragestellung können die Hörbeispiele auch so variiert sein, dass sie sich in einem komplexen Merkmal unterscheiden. Im Sinne der Testgütekriterien (vgl. S. 2) kommt der Auswahl geeigneter Merkmale, Versuchspersonen und Audioinhalte eine besondere Bedeutung zu.

6.1 Dominanzpaarvergleich

Versuchspersonen bekommen Stimuluspaare vorgespielt und sollen angeben, bei welchem Reiz ein bestimmtes Merkmal stärker ausgeprägt ist. Das Merkmal kann z.B. die Lautstärke, das Gefallen oder die Präferenz sein. Nach der Häufigkeit der Dominanzurteile wird eine Rangfolge erstellt. Dominieren bestimmte Elemente gleich häufig über andere, so müssen Verbundränge gebildet werden. Inkonsistenz von Urteilen kann zu zirkulären Triaden führen, etwa wenn B>A und C>B, aber zugleich A>C. Dies ist meist ein Hinweis auf eine Mehrdimensionalität des untersuchten Merkmals. In der Psychoakustik werden Dominanzurteile oft einer probabilistischen Skalierung nach Thurstone unterzogen und so in intervallskalierte Daten überführt. Dabei werden jedoch ungeprüfte Annahmen über Normalverteilung und Intervallskalenniveau gemacht. Dies ist messtheoretisch nicht haltbar.

6.2 Ähnlichkeitspaarvergleich

Beim Ähnlichkeitspaarvergleich wird in der Regel kein spezifisches Beurteilungsmerkmal vorgegeben. Die Versuchspersonen sollen vielmehr die globale Ähnlichkeit von Objektpaaren beurteilen, anhand eigener, frei wählbarer Kriterien. Die Beurteilung erfolgt meist auf mehrstufigen Ratingskalen. Ergebnis ist eine Ähnlichkeitsmatrix, die einfach in eine Unähnlichkeits- oder Distanzmatrix überführt werden kann, indem man die Ähnlichkeitswerte von 1 abzieht. Die Distanzen der Objekte können in euklidischer Metrik (Exponent 2) oder aber in anderen (allgemein: Minkowski-)Metriken dargestellt werden, wobei in der Hörpsychologie

v.a. die City-Block-Metrik (Exponent 1) eine Rolle spielt. Die Auswertung erfolgt aufwändig durch Metrische oder Nonmetrische Multidimensionale Skalierung (MDS oder NMDS). Ziel dieser Verfahren ist die Ermittlung der minimalen Anzahl von Dimensionen, die eine Darstellbarkeit der empirischen Objektdistanzen ermöglichen, sowie einer entsprechenden additiven Konstante. Die NMDS berücksichtigt dabei nur ordinale Information. Der sog. Stress ist ein Maß für die Güte der Dimensionslösung. Die gefundenen Dimensionen müssen anhand der Objekteigenschaften inhaltlich interpretiert werden.

6.3 Semantisches Differenzial

Das semantische Differenzial geht auf Charles E. Osgood zurück und ist in Deutschland auch als Polaritätsprofil oder Eindrucksdifferenzial bekannt. Es wird v.a. für die Erfassung von Eigenschaften externer Objekte eingesetzt. Die Beurteilungsobjekte können Personen, Begriffe, Gegenstände, Kunstwerke oder sonstige Konzepte sein. Die Eigenschaften sind in der Regel konnotative Bedeutungen; sie hängen mit emotionalen Qualitäten zusammen (vgl. Ertel 1964). Aber auch die Ausprägung denotativer Bedeutungen kann abgefragt werden. 10 bis 30 Merkmale werden nach Validitätskriterien für die jeweilige Untersuchungsfrage und die getesteten Beurteilungsobjekte zusammengestellt, als 5- oder 7stufige Ratingskalen mit gegensätzlich bezeichneten Skalenpolen operationalisiert und zu einem Differenzial gruppiert. Die Versuchspersonen sollen nun auf diesen Skalen die jeweilige Merkmalsauspräung des Hörbeispiels quantitativ einstufen, es werden also absolute Urteile abgefragt. Die Vielzahl der Variablen kann in der Auswertung durch datenreduzierende Verfahren verringert werden. Hierfür werden überwiegend die Hauptkomponenten- oder Faktorenanalyse angewandt. Ergebnis sind meist 2-4 künstliche Variablen (Komponenten oder Faktoren), die orthogonal unabhängig sind (also nicht miteinander korrelieren) und daher als Dimensionen bezeichnet werden können. Inhaltlich besitzen sie eine neue integrative Bedeutung, die im Einzelfall zu ermitteln ist. Häufig ergibt sich eine dreidimensionale Struktur mit den Bedeutungen Evaluation, Potency, Activity, die auch als EPA-Struktur oder semantischer Raum bezeichnet wird. Das semantische Differenzial ist eine flexible Methode, die stets der Fragestellung bzw. Bedingungsvariation, den Beurteilungsobjekten und den Versuchspersonen angepasst werden sollte (Konzept- und Zielgruppenspezifität). So lässt sich typischen Problemen wie Merkmalsverständlichkeit und *rater-concept-scale-interaction* effektiv entgegenwirken.

6.4 Repertory Grid Technique

Zentrales Problem bei der Erfassung psychischer Prozesse ist die Validität der abgefragten Merkmale: Sind z.B. die für ein semantisches Differenzial ausgewählten Merkmale wirklich geeignet, um einen vermuteten Effekt zu messen und wird ihre Bezeichnung von den Versuchspersonen 'richtig' verstanden? Die auf die Konstruktpsychologie von George Alexander Kelly zurückgehende Repertory Grid Technik (RGT) ermöglicht es, dem quantitativen Untersuchungsschritt einen qualitativen voranzustellen. Nach Kelly ist jeder Mensch ein Forscher, der Hypothesen bildet und prüft. Ergebnis sind sogenannte Konstrukte, Modelle der Wirklichkeit. Denn die Realität bleibt dem interpretationsfreien, objektiven Zugang ja verschlossen. Die Grid-Technik gibt nun einer Versuchsperson die Möglichkeit, seine ganz persönlichen Konstrukte mit eigenen Worten zu formulieren. Das Verfahren ist relativ offen und kann an

verschiedene Fragestellungen angepasst werden. Von Interesse sind hier Konstrukte zu akustischen Reizen.
Zunächst werden 10-20 Stimuli (systematisch oder zufällig) ausgewählt und daraus zufällig drei ausgelost. Diese Stimulus-Triaden spielt man der Versuchsperson vor. Sie soll nun entscheiden, welche zwei Stimuli einander am ähnlichsten sind. Dann soll sie überlegen, welches Merkmal die beiden Stimuli so ähnlich macht und zugleich von dem dritten unterscheidet, und die beiden dazugehörigen Merkmalsausprägungen nennen. Die der Versuchsperson auf diese Weise 'entlockten' Begriffe sind nicht nur für sich absolut valide, sondern bilden auch ein Gegensatzpaar, das die betreffende Person versteht. Der Vorgang wird nun so lange mit neuen Triaden wiederholt, bis etwa so viele Konstrukte (Begriffspaare) wie Stimuli vorliegen. Triviale Konstrukte dürfen vom Versuchsleiter abgelehnt werden. Schließlich werden aus den Gegensatzpaaren mehrstufige Ratingskalen konstruiert, auf denen die Versuchsperson in einem zweiten Untersuchungsteil die Stimuli quantitativ beurteilen soll. Die Datenerhebung ist damit abgeschlossen. Ergebnis ist ein individuelles Gitter mit den Konstrukten als Zeilen und den Stimuli als Spalten, in dessen Zellen die numerischen Werte der jeweiligen Merkmalsausprägung stehen.
Als Auswertungsmethoden kommen Faktoren- und clusteranalytische Verfahren in Betracht. Die Faktoren oder Cluster müssen inhaltlich interpretiert werden. Die Ergebnisse gelten nur für das Individuum. Ein Problem der RGT ist die Vergleichbarkeit der Grids verschiedener Versuchspersonen: Für eine Gruppenauswertung können strenggenommen nur Strukturmerkmale herangezogen werden, z.B. der Varianzanteil des ersten Faktors einer Faktorenanalyse, oder Korrelationen zwischen ausgewählten und vorgegebenen Konstrukten. In einer Zusammenschau lassen sich allerdings strukturelle Ähnlichkeiten (z.B. EPA-Struktur), interindividuell relevante Bedeutungsfelder und Begriffe sowie deren Rangfolgen erschließen – wichtige Informationen, die z.B. eine Konstruktion eines validen Semantischen Differenzials ermöglichen. Probeweise können auch Korrelationen zwischen den Konstrukten verschiedener Personen gerechnet werden, wenn man im Auge behält, dass dabei Konzept- und Urteilervarianz nicht getrennt sind.
Alternative Möglichkeiten der Gewinnung qualitativer Daten sind Verfahren der Textproduktion wie Interviews, offene Befragungen und Niederschriften, die anschließend inhaltsanalytisch ausgewertet werden können.

7 Literatur

Bortz, Jürgen (2005). *Statistik für Human- und Sozialwissenschaftler.* 6., vollst. bearb. u. akt. Aufl. Heidelberg: Springer Medizin Verl.

Bortz, Jürgen und Nicola Döring (2005). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler.* 3., überarb. Aufl., Nachdr. Berlin et al.: Springer.

Brunner, Stefan, Hans-Joachim Maempel und Stefan Weinzierl (2006). "On the Audibility of Comb Filter Distortions". In: *Bericht der 24. Tonmeistertagung 16.-19.11.2004 Congress Center Leipzig (CD-ROM).* o.O., o.V. In print.

Ciba, Simon (2008). *Erstellung einer Software-Bibliothek für Hörversuche. Programmkonzept und zu implementierende Testverfahren.* Mag.-Arb. Berlin, Techn. Univ. FG Audiokommunikation..

Ertel, Suitbert (1964). "Die emotionale Struktur des ‚semantischen' Raumes". In: *Psychologische Forschung* 28 (1). S.1-32.

Gelfand, Stanley A. (2004): *Hearing. An Introduction to psychological and physiological acoustics.* 4. überarb. u. erw. Aufl., New York: Dekker.

Hellbrück, Jürgen & Wolfgang Ellermeier (2004). *Hören. Physiologie, Psychologie und Pathologie.* 2., akt. u. erw. Aufl. Göttingen: Hogrefe.

International Telecommunication Union (Hg.) (1997). *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems.* Recommendation ITU-R BS.1116-1.

International Telecommunication Union (Hg.) (2003a). *Method for the subjective assessment of intermediate quality levels of coding systems.* ITU-R, BS.1534.

International Telecommunication Union (Hg.) (2003b). *General methods for the subjective assessment of sound quality.* ITU-R, BS.1284.

Iwamiya, Shin-ichiro (1994). "Interaction Between Auditory and Visual Processing When Listening to Music in an Audio Visual Context: 1. Matching 2. Audio Quality". In: *Psychomusicology* 13 (1/2). S.133-153.

Kelly, George A. (1986): *Die Psychologie der persönlichen Konstrukte.* Paderborn: Junfermann. (Enthält die ersten 3 Kapitel der Originalschrift von Kelly 1955).

Leek, Marjorie R. (2001). "Adaptive procedures in psychophysical research". In: *Perception & Psychophysics* 63 (8), S.1279-1292.

Marvit, Peter, Mary Florentine & Søren Buus (2003). "A comparison of psychophysical procedures for level-discrimination thresholds." In: *Journal of the Acoustical Society of America* 113. S.3348-3361.

Osgood, Charles E. et al. (1978). *The measurement of meaning.* 4. Aufl. d. Paperback-Ausg. Urbana/IL et al.: University of Illinois Press. 1. Aufl. 1957.

Otto, Stefanie (2008). *Vergleichende Simulation adaptiver, psychometrischer Verfahren zur Schätzung von Wahrnehmungsschwellen.* Mag.-Arb. Berlin, Techn. Univ. FG Audiokommunikation.

Rode, Martin (2005). *Evaluation von Kammfiltereffekten bei Lauzeitstereophonie mit mehr als zwei Mikrophonen.* Mag.-Arb. Berlin, Techn. Univ. FG Kommunikationswissenschaft.

Rudloff, Ingo (1997). *Untersuchungen zur wahrgenommenen Synchronität von Bild und Ton bei Film und Fernsehen.* Dipl.-Arbeit. Bochum, Univ.

Treutwein, Berhard (1995). "Adaptive Psychophysical Procedures". In: *Vision Research* 35. S.2503-2522.