THILO THIEDE [TU BERLIN], WILLIAM C. TREURNIET [CRC], ROLAND BITTO, THOMAS SPORER, KARLHEINZ BRANDENBURG [IIS-FHG], CHRISTIAN SCHMIDMER, MICHAEL KEYHL [OPTICOM], JOHN G. BEERENDS [KPN], CATHERINE COLOMES [CCETT], GERHARD STOLL [IRT], BERNHARD FEITEN [BERKOM]

PEAQ - der künftige ITU-Standard zur objektiven Messung der wahrgenommenen Audioqualität

Aufgrund der zunehmenden Verwendung von gehörangepaßten Audiocodierverfahren bei der digitalen Übertragung und Speicherung von Audiosignalen wird ein einheitliches Meßverfahren benötigt, das in der Lage ist, die Qualität solcher Audiocodierer objektiv zu messen. Ende 1994 wurde in der ITU eine Arbeitsgruppe, ITU-R TG 10/4, gegründet, die zum Ziel hatte, ein geeignetes Meßverfahren zur Standardisierung vorzuschlagen. Die Arbeit dieser Gruppe wurde Anfang 1998 abgeschlossen. Das schließlich ausgewählte gehörangepaßte Meßverfahren PEAQ entstand aus einer Zusammenarbeit zwischen mehreren an der Entwicklung früherer Meßverfahren beteiligten Forschungsinstituten. Die Meßmethode basiert auf der Kombination verschiedener objektiver Qualitätsmaße zu einem Schätzwert für die wahrgenommene Audioqualität. Die von diesem Verfahren gelieferten Modellvorhersagen zeigen eine deutlich höhere Korrelation mit den Ergebnissen subjektiver Hörtests als alle bisherigen Meßverfahren.

0 EINLEITUNG

Bei der digitalen Übertragung und Speicherung von Audiosignalen werden in zunehmendem Maße Datenreduktionsverfahren verwendet, die Eigenschaften des menschlichen Gehörs ausnutzen. Dabei wird versucht, die spektrale Verteilung der entstehenden Quantisierungsfehler so zu beeinflussen, daß sie unterhalb der Hörschwelle liegen. Die auf diese Weise unhörbar gemachten Störungen sind jedoch immer noch physikalisch vorhanden. Die wahrgenommene Qualität solcher gehörangepaßten Codierverfahren kann somit mit konventionellen Meßverfahren, die lediglich die insgesamt vorhanden Störungen erfassen, nicht bestimmt werden. Daher wird die Qualität von gehörangepaßten Codierverfahren üblicherweise mittels subjektiver Hörtests bestimmt. Solche Hörtests müssen unter optimalen Abhörbedingungen und mit einer großen Anzahl von Testhörern durchgeführt werden, so daß dieser Weg der Qualitätsbestimmung in vielen Fällen zu aufwendig ist. Ein objektives Meßverfahren, das die zum subjektiven Qualitätseindruck führenden physiologischen und kognitiven Vorgänge

simuliert, kann in vielen Fällen Abhilfe schaffen. Da hierzu verschiedene Vorschläge existierten, wurde in der ITU-R eine Arbeitsgruppe gegründet, die zum Ziel hatte, diese Vorschläge zu untersuchen und eine Empfehlung für ein Meßverfahren zur „objektiven Messung der wahrgenommenen Audioqualität" zu erarbeiten. Das für diese Empfehlung vorgesehene Meßverfahren enthält sowohl Teile von zuvor existierenden Meßverfahren, als auch eine Reihe von neuen Modellierungsansätzen.

Der grundsätzliche Aufbau des Meßverfahrens ist in Abb. 1 gezeigt. Die Qualität des zu bewertenden Testsignal wird anhand seiner Abweichungen von dem als Referenz dienenden Originalsignal bestimmt. Dazu werden beide Signale in eine gehörangepaßte Darstellung umgeformt (peripheres Gehörmodell). Anschliessend werden durch einen Vergleich im Zeit- und Frequenzbereich verschiedene Abstandsmaße bestimmt. Diese Ausgangswerte („model output values" - MOVs) werden zu einer einzelnen Kenngröße („distortion index" - DI) zusammengefaßt, der sich auf einen Schätzwert für die subjektiv empfundene Audioqualität abbilden läßt. Entsprechend der Bezeichnung der aus Hörtests gewonnenen Qualitätsbewertung als „subjective difference grade" (SDG) wird dieser Ausgangswert als „objective difference grade" (ODG) bezeichnet.

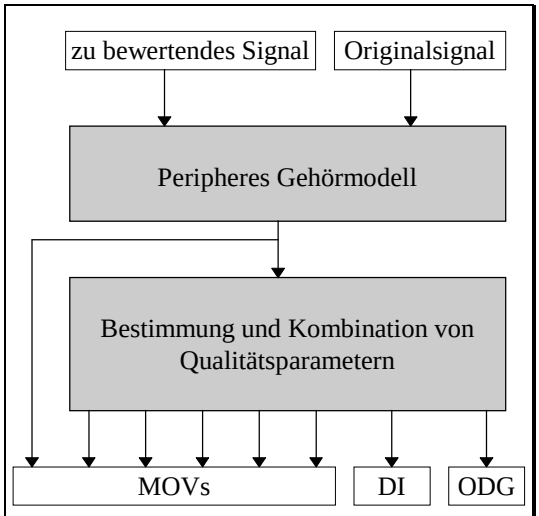

Abb. 1: Grundsätzlicher Aufbau des Meßverfahrens

1 ARBEITSWEISE GEHÖRANGEPAßTER MEßVERFAHREN

Es existieren zwei unterschiedliche Grundprinzipien, nach denen gehörangepaßte Meßverfahren arbeiten können: der Vergleich der Störung mit einer aus dem Originalsignal berechneten Maskierungsschwelle („*Masked Threshold Concept"*) und der Vergleich zwischen gehörangepaßten Signaldarstellungen von Originalsignal und zu bewertendem Signal („*Comparison of Internal Representations"*). Als dritter Ansatz kann der direkte Vergleich der Spektraldarstellungen beider Signale (ohne Verwendung eines Gehörmodells) betrachtet werden.

1.1 Vergleich zwischen Maskierungsschwelle und Störung

Das Prinzip des Vergleichs der Störung mit einer Maskierungsschwelle („*Masked Threshold Concept"*, auch: „*Noise Signal Evaluation*") wurde in den ersten bekannten gehörangepaßten Meßverfahren (z. B. [SCH79], [BRA87]) verwendet. Dabei wird durch Subtraktion des Originalsignals vom zu bewertenden Signal das Fehlersignal berechnet und mit einer aus dem Originalsignal bestimmten Maskierungsschwelle verglichen. Vorteile dieses Konzeptes sind der relativ einfache Abgleich mittels aus psychoakustischen Experimenten gefundener Daten und die Verwendbarkeit des zugehörigen psychoakustischen Modells für Audiocodierverfahren.

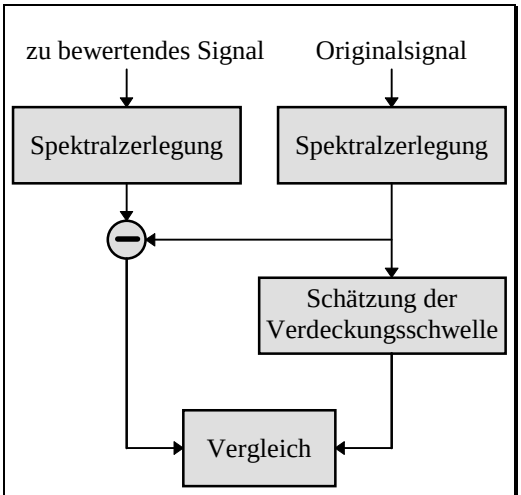

Abb. 2: Vergleich zwischen Maskierungsschwelle und Störung.

1.2 Vergleich zwischen gehörangepaßten Signaldarstellungen

Das Konzept des Vergleichs zwischen gehörangepaßten Signaldarstellungen („*Comparison of Internal Representations"*, auch: „*Comparison in the Cochlear Domain*"), wurde erstmals 1985 von Karjalainen [KAR85] verwendet und bildet die Grundlage für die meisten neueren gehörangepaßten Meßverfahren (z. B. [BEE92], [PAI92], [COL93], [THI96], [SPO97]). Dabei werden sowohl aus dem Originalsignal als auch aus dem zu bewertenden Signal gehörangepaßte Darstellungen (sogenannte Erregungsmuster) bestimmt. Die Bewertung der Qualität erfolgt aus dem Vergleich dieser Erregungsmuster. Diese Vorgehensweise kommt der physiologischen Arbeitsweise des Gehörs sehr viel näher als das zuvor beschriebene Konzept. Es bietet daher eine bessere Ausgangsbasis für komplexere Gehörmodelle.

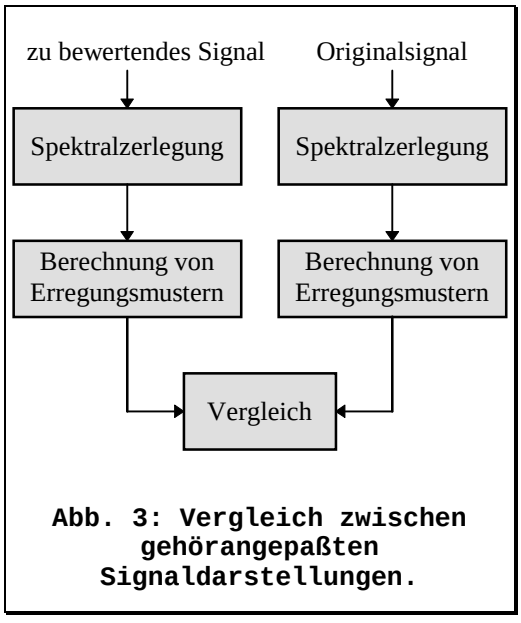

Abb. 3: Vergleich zwischen gehörangepaßten Signaldarstellungen.

1.3 Analyse von Fehlerspektren

Einige Effekte, wie z. B. die Wahrnehmung einer Grundfrequenz in Tonkomplexen, lassen sich anhand linearer Spektraldarstellungen einfacher modellieren als anhand einer gehörrichtigen Darstellung. Ein solcher Ansatz kann zwar wegen des fehlenden Gehörmodells nicht die alleinige Grundlage eines Meßverfahrens sein, er kann ein solches Modell aber ergänzen, da er zusätzliche Informationen über das Testsignal liefert, die aus einem Gehörmodell nur schwer gewonnen werden können.

1.4 Peripheres Gehörmodell

1.4.1 Übertragungseigenschaften des Gehörs

Schallsignale werden durch Außen- und Mittelohr in ihrer Bandbreite beschränkt. In der Gehörschnecke erfolgt die eigentliche Analyse der Schallsignale und ihre Umsetzung in Nervensignale. Ein Grundrauschen, welches zum Teil durch die Strömung des Blutes im Kopf, zum Teil durch spontane Aktivität von Nervenfaser erzeugt wird, verdeckt sehr leise Schallsignale. Die Ruhehörschwelle wird überwiegend durch diese beiden Effekte erklärt.

1.4.2 Frequenzskalen des menschlichen Gehörs

Die Umsetzung von mechanischen Schwingungen in elektrische Signale erfolgt im Innenohr, genauer in der Gehörschnecke. Abhängig von der Frequenz wird die dort befindliche Basilarmembran an unterschiedlichen Orten maximal ausgelenkt. Die Haarzellen, d. h. die Sinneszellen, welche sich auf der Basilarmembran befinden, werden durch diese Auslenkung erregt. Jede Haarzelle reagiert auf einen Bereich benachbarter Frequenzen. Ein konstanter Abstand auf der Basilarmembran ist eng verbunden mit der Wahrnehmung der Tonheit. Die Aufteilung des gesamten Hörbereichs erfolgt dabei nicht linear. Je nach psychoakustischem Experiment findet man unterschiedliche Übertragungsfunktionen von der Frequenzskala auf eine Tonheitsskala. Die am häufigsten genutzte Skala, die Bark-Skala, geht auf Zwicker [ZWI67] zurück: Der Bereich von 0 Hz bis 15000 Hz wird durch sie in 24, einander nicht überlappende Bereiche geteilt. Eine Diskussion unterschiedlicher Messungen von Tonheitsskalen findet sich in [COH92].

1.4.3 Erregung

Jede einzelne Haarzelle reagiert auf einen Bereich benachbarter Frequenzen. Dies kann als eine Bandpaßfilterung betrachtet werden. Die Flankensteilheit der resultierenden Filter ist konstant, wenn man die Filterkurven in Pegelschreibweise über der Tonheitsskala aufträgt. Die Steilheit der unteren Flanke der Filter ist unabhängig vom Pegel des erregenden Signals (ca. 27 dB/Bark). Die Steilheit der oberen Flanke ist für leise Signale betragsmäßig größer als für laute Signale (-30 dB/Bark bis -5 dB/Bark)[TER79]. Diese Pegelabhängigkeit der Steilheit wird durch neuronale Rückkoppelungsmechanismen beeinflußt. Diese Regelvorgänge benötigen eine Einschwingzeit. Die beste Frequenzselektivität wird deshalb erst einige Millisekunden nach Beginn eines Schallreizes erreicht.

Das Erregungsmuster eines Schallreizes, der aus mehreren Komponenten besteht, ergibt sich durch nichtlineare Addition der Erregungsmuster der Einzelkomponenten.

Nach dem Ende eines Schallreizes benötigen die Haarzellen und die nachfolgenden neuronale Verarbeitungsstufen eine Erholungszeit bis sie ihre volle Empfindlichkeit wieder erreichen. Die Dauer dieser Erholungszeit ist abhängig von Pegel und Dauer des ersten Schallreizes und kann bis zu mehreren hundert Millisekunden betragen. Signale mit hohem Pegel werden auf dem Weg zwischen Innenohr und Gehirn schneller verarbeitet als leise Signale. Ein lautes Signal kann dadurch ein vorhergehendes leises Signal unhörbar machen (verdecken).

Ein alternativer Ansatz die Erregung zu modellieren basiert auf der sogenannten ERB-Skala („equivalent rectangular bandwidth") und einer ROEX („rounded exponential filter") genannten Filterfamilie [MOO86]. Einige einfache psychoakustische Experimente lassen sich mit diesem Ansatz besser erklären. Im Bereich der Bewertung der Audioqualität wurden allerdings bessere Ergebnisse mit dem oben beschriebenen Ansatz, der auf Arbeiten von Zwicker beruht, erzielt.

1.4.4 Wahrnehmung von Unterschieden

Die Erregungsmuster von Schallsignalen werden im menschlichen Gehirn weiterverarbeitet und gespeichert. Es werden dabei drei verschiedene Arten der Speicherung unterschieden: Langzeit-Gedächtnis, Kurzzeit-Gedächtnis und Ultra-Kurzzeit-Gedächtnis. Hörtests zur Bewertung von Audiosignalen basieren überwiegend auf dem Ultra-Kurzzeit-Gedächtnis, welches die meisten Feinheiten eines Signals speichern kann. Seine Dauer ist allerdings auf

Signalabschnitte mit einer maximalen Dauer von fünf bis acht Sekunden beschränkt. In der ITU-R Empfehlung BS.1116 ist dies dadurch berücksichtigt, daß die Testpersonen kurze Abschnitte eines Audiosignals auswählen und genauer „betrachten“ können.

Extrem kleine Unterschiede zwischen Signalen werden nicht wahrgenommen. Die Wahrnehmungsschwelle ist dadurch definiert, daß die Wahrscheinlichkeit der Wahrnehmung an dieser Stelle 50 % beträgt. Die Wahrnehmungsschwelle für Änderungen des Signalpegels (engl. „just-noticeable-level-difference", JNLD) ist abhängig vom Pegel des Eingangssignals. Für leise Signale werden größere Schwellen (z.B. Schalldruckpegel: 20 dB SPL, JNLD: 0,75 dB), für laute Signale niedrigere Schwellen (Schalldruckpegel: 80 dB SPL, JNLD: 0,2 dB) gemessen.

1.4.5 Verdeckung

Ein in vollständiger Ruhe deutlich wahrnehmbares Signal kann in Anwesenheit eines zweiten Signals unhörbar, d. h. verdeckt, werden [ZWI67]. Dieses zweite Signal wird dann Maskierer genannt.

- Simultanverdeckung

  Bei der Simultanverdeckung sind beide Signale gleichzeitig vorhanden. Die Stärke der Maskierung hängt hauptsächlich von Struktur, Pegel und relativer Frequenzlage, teilweise aber auch von der absoluten Frequenz der Signale ab. Die maximale Verdeckung wird jeweils erzielt wenn Maskierer und verdecktes Signal die gleiche Mittenfrequenz haben. Bei abweichender Mittenfrequenz ergibt sich ein der Erregung entsprechender Verlauf. Die maximale Verdeckung liegt bei der Verdeckung von Tönen durch Rauschen unabhängig von der Mittenfrequenz des Rauschens ca. 5 dB unter dem Pegel des Maskierers. Bei der Verdeckung von Rauschen durch Töne ist dieser als *Verdeckungsmaß* bezeichnete Abstand abhängig von der Tonheit z des Maskierers und beträgt ca. *15,5 dB+z/Bark*. Entsprechend der nichtlinearen Addition der Erregungsmuster liegt die Verdeckungsschwelle bei mehreren Maskierern in der Regel höher als die Summe der einzelnen Verdeckungsschwellen.

- Zeitliche Verdeckung

  Zeitliche Verdeckung tritt auf, wenn Maskierer und verdecktes Signal nicht gleichzeitig vorhanden sind. Entsprechend der Erholungszeit bei den Erregungsmustern werden leise

Signale kurz nach lauten Signalen durch diese unhörbar (Nachverdeckung). Laute Signale nach leisen Signalen überholen diese bei der Weiterleitung vom Ohr zum Gehirn und verdecken diese (Vorverdeckung). Die Nachverdeckung ist in der Regel deutlich länger als die Vorverdeckung und kann sich über einige hundert Millisekunden erstrecken [FAS74]. Die minimale Vor- und Nachverdeckung tritt bei der Verdeckung durch Impulse („Klicks“) auf [SPI92]. Im Gegensatz zur Nachverdeckung treten bei der Vorverdeckung sehr große Unterschiede zwischen verschiedenen Hörern auf.

1.4.6 Lautheit und Drosselung

Die wahrgenommen Lautheit eines Audiosignal hängt außer vom Schalldruckpegel auch von seiner Frequenz und Dauer ab. Anteile eines komplexen Schallereignisses können sich gegenseitig verdecken, so daß die resultierende Lautheit geringer sein kann als die Summe der Lautheiten der Einzelschalle („Drosselung ", [ZWI67]). Für die Beurteilung von Audiosignalen im Vergleich zu einem Referenzsignal ist die Lautheit der Störkomponenten wichtig. Die Lautheit dieser Störkomponenten ist auf Grund der Drosselung durch das Nutzsignal reduziert.

1.4.7 Schärfe

Der Parameter Schärfe beurteilt die Klangfarbe eines Audiosignals. Ein Klang wird als „scharf“ wahrgenommen, wenn er viele hochfrequente Komponenten enthält [BIS74]. Am wichtigsten bei der Beurteilung der Schärfe ist dabei der Tonheitsbereich 14 bis 20 Bark [AUR84].

1.5 Modellierung kognitiver Effekte

Bei der subjektiven Bewertung der Qualität eines Audiosignals spielt die Hörerfahrung des Testhörers eine wesentliche Rolle. Diese Hörerfahrung ist zwar zwischen verschiedenen Testhörern nicht völlig identisch, es gibt aber (glücklicherweise) weitgehende Übereinstimmungen. So mag es z. B. verschiedene Meinungen darüber geben, wie ein *gutes* Klavier zu klingen hat, aber es gibt doch eine gemeinsame Vorstellung eines Klavierklanges, und wenn dieser Klang verzerrt wird, werden sich üblicherweise alle Testhörer einig sein, daß die Audioqualität niedrig ist. Computer verfügen dagegen nicht über ein solches Grundwissen, und darum wird ein technisches Meßverfahren immer ein Referenzsignal benötigen.

Das fehlende Grundwissen wird jedoch nur unvollkommen durch das Referenzsignal ersetzt. Dies läßt sich veranschaulichen, wenn man zwei gedachte Geräte miteinander vergleicht: eines, das ein Signal stark verzerrt, und eines, das aus diesem stark verzerrten Eingangssignal wieder das unverzerrte Originalsignal rekonstruiert. Bei einer Betrachtung der Abweichungen zwischen dem jeweiligen Eingangs- und Ausgangssignal würde das letztere Gerät ebenso schlecht abschneiden wie das erste. Ein normaler Testhörer würde dagegen dem Gerät, das die Verzerrungen beseitigt, zweifellos eine bessere Qualität zuordnen als dem Gerät, das die Verzerrungen verursacht.

Auch wenn sich das durch dieses Beispiel aufgezeigte Problem des fehlende Grundwissens nicht vollständig lösen läßt, kann es doch mittels einiger einfacher Annahmen in vielen Fällen umgangen werden. Ein Ansatz ist die Vorstellung, daß alle zu einem Musiksignal gehörenden Anteile ein gemeinsames Hörereignis bilden. Läßt man einzelne Anteile weg, formt der verbleibende Teil weiterhin ein gemeinsames Hörereignis. Wird dagegen eine Verzerrung hinzugefügt, wird diese üblicherweise nicht demselben Hörereignis zugeordnet und wird daher als störend empfunden [MCA84] [BRE81]. Dies kann in vielen Fällen durch eine einfache Unterscheidung zwischen einer Zunahme und einer Abnahme der Energie in einzelnen Frequenzbändern berücksichtigt werden [BEE94]. Auch die in PEAQ vorgenommene Trennung zwischen linearen und nichtlinearen Verzerrungen trägt zur Modellierung dieses Effektes bei (vgl. Abschnitt 3.3).

Auch Lerneffekte spielen eine wichtige Rolle bei der subjektiven Beurteilung von Audiosignalen. Eine unbekannte Störung wird deutlich schwerer erkannt als bekannte Störungsarten. Weiterhin zeigt sich beim Erkennen kleiner Störungen in sehr komplexen Audiosignalen ein deutlicher Trainingseffekt. Es ist bekannt, das sich die Schwelle für die Wahrnehmung einer komplexen Störung in einem komplexen Audiosignal durch Training um bis zu 40 dB verringern läßt [LEE84]. Erste Versuche zur Modellierung dieses Effektes mittels eines Maßes für die zeitliche und spektrale Komplexität [BEE96] ergaben zwar eine Verbesserung der Qualitätsvorhersage für einige Hörtestdatensätze, führten aber bei anderen Datensätzen zu einer Verschlechterung. Da Hörtests nach der ITU-R Empfehlung BS.1116 bereits ein Training der Testhörer vorschreiben, ist es auch denkbar, daß für die Vorhersage dieser Hörtestergebnisse keine Berücksichtigung der zeitlichen und spektralen Komplexität notwendig ist. Ein solches Modell wurde daher nicht in den ITU-Standard aufgenommen.

Weitere kognitive Effekte sind die unterschiedliche Wahrnehmung linearer und nichtlinearer Verzerrungen sowie die veränderliche zeitliche und spektrale Gewichtung von Störungen. Der erstgenannte Effekt wird in PEAQ explizit modelliert (Abschnitt 3.3). Eine veränderliche zeitliche und spektrale Gewichtung hat sich zwar bei der Bewertung codierter Sprachsignale als vorteilhaft erwiesen [BEE98], ergab jedoch bei der Bewertung hochwertiger Musiksignale keine signifikante Verbesserung und wird daher in PEAQ nicht modelliert.

2 VORGESCHICHTE

Das erste gehörangepaßte Verfahren zur objektiven Bestimmung der Qualität von Audiosignalen wurde 1979 von Schroeder, Atal und Hall veröffentlicht [SCH79]. Das „Noise Loudness" (NL) genannte Verfahren simuliert die empfundene Lautheit des durch Sprachcoder erzeugten Quantisierungsgeräusches. Seine Lautheit hängt von der vollständigen oder teilweisen Verdeckung durch das Nutzsignal (Sprache) ab. Die Simulation der Maskierung basiert auf der Simultanverdeckung von Rauschen durch Töne. Die Berechnung der Störlautheit und der „Signaldegradierung" erfolgt alle 20 ms. Eine Zusammenfassung der Meßwerte für längerer Zeitabschnitte erfolgt nicht.

Aufbauend auf dieses Verfahren entwickelte Matti Karjalainen um 1985 das Verfahren „Auditory Spectral Difference" (ASD) [KAR85]. Die zeitliche Auflösung des Verfahrens wurde durch den Einsatz von 40 einander überlappenden FIR-Filtern verbessert. Dies ermöglichte u. a. auch die Modellierung der Nachverdeckung. Im Gegensatz zu NL werden bei ASD sowohl das unverfälschte Eingangssignal (Referenzsignal) als auch das zu bewertende Signal in exakt gleicher Weise verarbeitet. Ergebnis dieser Verarbeitung ist eine „interne Darstellung". Diese repräsentiert im Idealfall die Information, die dem menschlichen Gehör zur Verfügung steht, um zwei Signale zu vergleichen. Eine Zusammenfassung der Meßwerte für längere Zeitabschnitte erfolgt auch bei ASD nicht. Verglichen mit NL bietet ASD eine bessere Modellierung von zeitlichen Eigenschaften des menschlichen Gehörs bei einer höheren Komplexität.

Das von Brandenburg um 1987 entwickelte Verfahren „Noise-to-Mask Ratio" (NMR) baut ebenfalls auf den Ideen von NL auf [BRA87]. NMR wurde als Hilfsmittel für die Entwicklung von Audiocodecs konzipiert. Die Komplexität des Verfahrens ist durch eine Vereinfachung des psychoakustischen Modells reduziert. Ein „worst-case"-Ansatz soll sicherstellen, daß das Verfahren trotzdem alle relevanten Störungen korrekt bewertet und sich robust gegenüber

Veränderungen des Abhörpegels verhält. Im Gegensatz zu NL und ASD basiert die Modellierung der Simultanverdeckung bei NMR auf der Verdeckung von Tönen durch Rauschen. NMR enthält ein einfaches Modell der Nachverdeckung. Verschiedene Methoden der Zusammenfassung der Meßwerte über längere Zeitabschnitte sind in NMR vorhanden.

In den 90er Jahren wurden verschiedene andere Meßverfahren zur Bewertung von Sprach- und Audiosignalen entwickelt. Im folgenden soll auf die in die Entwicklung von PEAQ eingeflossenen Verfahren näher eingegangen werden.

2.1 Vorläufer von PEAQ

2.1.1 Noise-to-Mask Ratio (NMR)

NMR [BRA87] bewertet den Abstand zwischen dem Störsignal und der simulierten Mithörschwelle des Referenzsignals. Eine DFT mit einem Hann-Fenster mit einer Länge von 1024 Abtastwerten wird zur Analyse der Eingangssignale verwendet. Die Spektralkoeffizienten werden zu Bark-breiten Bändern zusammengefaßt. Für jedes der Bänder wird die Verdeckungsschwelle getrennt abgeschätzt. Bei der Verdeckung zwischen Bändern wird berücksichtigt, daß die Maskierungskurve bei niedrigem Schalldruckpegel steiler verläuft als bei hohem Schalldruckpegel, während bei leisen Signalen die Ruhehörschwelle einen größeren Einfluß hat. Die Amplitudenauflösung digitaler Audiosignale (z. B. 16 Bit) wird als grobe Annäherung der Ruhehörschwelle berücksichtigt, eine Annahme über einen Abhörpegel erfolgt nicht. NMR ist dadurch robust gegenüber der Änderung der Abhörlautstärke. Aufgrund der geringen Komplexität des Verfahrens war es bereits 1992 möglich ein Echtzeitmeßsystem zu erstellen [HER92]. Seit 1987 wird NMR bei der Entwicklung von Audiocodierverfahren regelmäßig eingesetzt.

Die wichtigsten Meßgrößen von NMR sind der Anteil der möglicherweise gestörten Zeitabschnitte (engl. „masking flag rate"), sowie der mittlere und der gesamte NMR. Die letzteren Werte werden mittels verschiedener zeitlicher Mittelungsverfahren aus dem NMR, d. h. aus dem Abstand zwischen Störung und Mithörschwelle gewonnen.

2.1.2 Perceptual Audio Quality Measure (PAQM)

PAQM [BEE92] bewertet den Unterschied der internen Darstellungen der beiden Eingangssignale. Dabei werden kognitive Effekte der Wahrnehmung berücksichtigt. Die Eingangssignale werden zunächst mittels einer DFT mit einem Hann-Fenster mit einer Länge

von 2048 Abtastwerten analysiert und auf die Bark-Skala abgebildet. Simultanverdeckung und zeitliche Verdeckung werden durch nichtlineare Faltungsoperationen simuliert. Die resultierende Intensitätsverteilung wird einer nichtlinearen Amplitudenkompression unterzogen. Alle nichtlinearen Abbildungen wurden nicht nur mittels psychoakustischer Experimente (i. A. mit künstlichen Audiosignalen) sondern auch mit den Ergebnissen der ersten MPEG Hörtests abgeglichen. Beim Vergleich der internen Darstellungen werden die kognitiven Effekte „perceptual streaming" und „informational masking" simuliert. „Perceptual streaming" berücksichtigt die Fähigkeit des Gehörs zusammengehörige Schallanteile zu Schallereignissen zusammen zu fassen. Hierdurch bedingt sind Unterschiede zwischen Audiosignalen vor allem dann auffällig, wenn sie zu zusätzlichen Schallereignissen führen. Signale mit hoher zeitlicher und spektraler Komplexität führen zu „informational masking": Je nach vorherigem Training sind Versuchspersonen nicht in der Lage kleine, psychoakustisch durchaus wahrnehmbare, Signalunterschiede richtig wahrzunehmen. „Perceptual Speech Quality Measure" (PSQM), eine teilweise vereinfachte Variante von PAQM, dient zur Bewertung der Qualität von Sprachcodierverfahren. PSQM modelliert zusätzliche kognitive Aspekte die nur bei Sprachsignalen, nicht aber bei Audiosignalen, eine Rolle spielen. PSQM wurde durch ITU-T Empfehlung P.861 standardisiert [ITU96].

2.1.3 Perceptual Evaluation (PERCEVAL)

PERCEVAL [PAI92] bewertet den Unterschied der internen Darstellungen der beiden Eingangssignale. Auch PERCEVAL benutzt eine DFT mit Hann-Fenster mit einer Länge von 2048 Abtastwerten. Die Spektralkoeffizienten werden auf die mel-Skala abgebildet. Mittels einer Faltungsoperation wird die Energieverteilung auf der Basilarmembran simuliert. Ein frequenzabhängiger Beitrag des internen Rauschens wird zu jedem Band addiert. Aus dem Unterschied der internen Darstellung des Referenzsignals und des zu bewertenden Signals wird u.a. die Wahrscheinlichkeit der Wahrnehmbarkeit berechnet. Diese und andere aus dem Vergleich der internen Darstellungen gewonnenen Kenngrößen werden mittels eines neuronalen Netzes zusammengefaßt. Dieses neuronale Netz wurde mit einer großen Anzahl von Ergebnissen aus Hörtests trainiert.

2.1.4 Perceptual Objective Measure (POM)

POM [COL93] bewertet den Unterschied der internen Darstellungen der beiden Eingangssignale. Auch POM benutzt eine DFT mit Hann-Fenster mit einer Länge von 2048

Abtastwerten. POM verwendet ein Frequenzskala mit 620 Bändern, welche an das Frequenzunterscheidungsvermögen des menschlichen Gehörs angepaßt ist. Die Faltungsoperation zur Simulation der Simultanverdeckung berücksichtigt die unterschiedliche Steilheit der Verdeckungskurven bei unterschiedlichen Schalldruckpegel. POM simuliert die Wahrscheinlichkeit der Wahrnehmung und berechnet den Abstand der internen Darstellung. Ein neuronales Netz faßt diese Rohwerte zusammen.

2.1.5 Disturbance Index (DIX)

DIX (Disturbance Index) [THI96] bewertet den Unterschied der internen Darstellungen der beiden Eingangssignale. Das Verfahren verwendet eine gehörangepaßte Filterbank mit hoher zeitlicher Auflösung, die (verglichen mit FFT-basierten Verfahren) eine genauere Modellierung zeitlicher Effekte, wie z. B. Vor- und Nachverdeckung ermöglicht. Darüber hinaus werden aus der zeitlichen Feinstruktur der Hüllkurven in den einzelnen Filterbändern zusätzliche Informationen über die Art des Testsignals und die vorhanden Störungen gewonnen. Die Mittenfrequenzen und Bandbreiten sind gleichmäßig über eine Bark-Skala verteilt. Der hörbare Frequenzbereich wird mit 40 Filterbändern abgedeckt, was einer Auflösung von etwa 0,6 Bark entspricht. Die verwendete Filterbank ist linearphasig und erfordert verglichen mit anderen gehörangepaßten Filterbänken einen relativ geringen Rechenaufwand.

DIX nimmt eine getrennte Bewertung von linearen und nichtlinearen Verzerrungen vor und modelliert Verdeckungsschwellen in Abhängigkeit von der zeitlichen Struktur der basilaren Erregungsmuster. Zu den berechneten Ausgangsparametern gehören u. a. die partielle Lautheit von nichtlinearen Verzerrungen, sowie Maße für lineare Verzerrungen und für Änderungen der zeitlichen Struktur einzelner Signalanteile.

2.1.6 Objective Audio Signal Evaluation (OASE)

OASE [SPO97] bewertet den Unterschied der internen Darstellungen der beiden Eingangssignale. OASE verwendet eine Filterbank mit 241 einander überlappenden Filtern. Die Mittenfrequenzen der Filter haben gleichen Abstand auf der Bark-Skala. Die Pegelabhängigkeit der Maskierung ist beim Entwurf der einzelnen Filters berücksichtigt. Hierzu wurde ein worst-case Ansatz ähnlich dem von NMR verwendet. Auch die Außen- und Mittelohrübertragungsfunktion ist bereits in der Filterbank berücksichtigt. Nach der Modellierung zeitlicher Vorgänge kann die Abtastrate in allen Bändern reduziert werden. Der Vergleich der so gewonnenen internen Darstellung berücksichtigt den pegelabhängigen „gerade

wahrnehmbaren Pegelunterschied" (engl. just noticable level difference, JNLD). Eine Wahrscheinlichkeitsfunktion gibt die Wahrscheinlichkeit der Wahrnehmbarkeit des Unterschieds für jedes Band an. Im Falle der Wahrnehmbarkeit wird die Differenz der internen Darstellungen mit dem aktuellen JNLD normiert und ergibt die „schwellennormierte Störlautheit". OASE gewinnt aus der Wahrnehmungswahrscheinlichkeit und der Störlautheit eine große Anzahl verschiedener Qualitätsparameter.

Die Filterbank von OASE kann mittels einer Multirate-Struktur realisiert werden, was zu einer im Vergleich zu anderen Filterbänken moderaten Komplexität führt.

2.1.7 Toolbox

Toolbox basiert auf der Kombination von psychoakustischen Parametern, welche sich aus jedem einzelnen der Eingangssignale bestimmen lassen, mit Parametern die aus beiden Signalen gewonnen werden. Das hierzu verwendete psychoakustische Modell verwendet eine DFT mit Hann-Fenster mit einer Länge von 2048 Abtastwerten. Aus jedem der Eingangssignale werden Maße für die Schärfe, die Rauhigkeit und die spezifische Lautheit gewonnen. Aus dem Vergleich der Signale wird die Störlautheit bestimmt. Für alle diese Parameter werden Mittelwert, Minimum, Maximum und Standardabweichung berechnet.

2.2 ITU-R Arbeitsgruppe zum Thema gehörangepaßte Meßverfahren

1994 wurde in der ITU-R die Arbeitsgruppe TG10/4 unter dem Vorsitz von Thomas Rydén gegründet. Ein „Call for Proposals" führte zu sechs Vorschlägen. Eine Validierungsprozedur sollte diese Vorschläge bezüglich ihrer Leistungsfähigkeit untersuchen. Hierzu wurden zunächst die Ergebnisse vergangener Hörtests in einer Datenbank gesammelt und den Proponenten zur Verfügung gestellt. Das Datenmaterial bestand aus den Audiosignalen zusammen mit den dazugehörigen Hörtestergebnissen. Diese Hörtests wurden überwiegend im Rahmen der Standardisierung von ISO/IEC MPEG bzw. für die ITU-R im Rahmen der Erstellung einer Empfehlung für den digitalen Rundfunk durchgeführt. Alle diese Hörtests waren nach ITU-R BS.1116 [ITU97] durchgeführt. Die untersuchten Audiocodierverfahren waren MPEG1 Layer 2 und 3, Dolby AC2, ATRAC (MiniDisk), MUSICAM, ASPEC und NICAM. Überwiegend befanden sich Stücke mit mittlerer oder guter Qualität in dieser Datenbasis. Diese Datenbank wird im folgenden als DB1 bezeichnet. Aufgabe der DB1 war es, allen Entwicklern von Meßverfahren die gleichen Möglichkeiten des Abgleichs zu geben und gleichzeitig die universelle Nutzbarkeit eines Verfahrens zu überprüfen. Bereits während der

Arbeit an der DB1 wurde auffällig, daß die Ergebnisse von Hörtests nach BS.1116 kontextabhängig sind: Die Testpersonen tendieren dazu ihre individuellen Bewertungsskalen an das Testmaterial zu adaptieren. Bei der späteren Validierung von Meßverfahren mit unbekannten Datenbasen wurde diese Erkenntnis durch eine Aufteilung in zwei Phasen berücksichtigt: In einer ersten Phase erfolgte jeweils eine „blinde" Voraussage der Hörtestergebnisse. Nach einer „Eichung" des Meßverfahrens mit einem Teil der Hörtestergebnisse erfolgte dann anschliessend die Vorhersage des weiterhin unbekannten Teils der Datenbasis.

Eine angemessene Validierung eines objektiven Meßverfahrens erfordert eine Datenbasis mit Material, das dem Meßverfahren „unbekannt" ist. Daher wurden Hörtests notwendig, die speziell für den Validierungsprozeß ausgelegt wurden. Da das Verfahren im Idealfall sämtliche bei Rundfunkanwendungen mögliche Artefakte anzeigen können soll, wurden zunächst nicht nur Codierartefakte sondern auch Störungen wie Klirren und Rauschen berücksichtigt. Unter diesen Gesichtspunkten wurde im Jahr 1996 von der TG 10/4 die Datenbasis 2 (DB2) zusammengestellt und in Hörtests entsprechend BS.1116 bewertet. Im weiteren Verlauf der Aktivitäten wurde für die abschließende Validierung im Jahr 1997 eine dritte Datenbasis (DB3) notwendig. Im Gegensatz zu DB2 enthielt DB3 wieder zum überwiegenden Teil Codierartefakte, wobei neben Codecs, die schon in DB1 und DB2 enthalten waren, auch neuere Codecs, wie AC-3 und AAC, enthalten waren.

Im Rahmen der ITU-R TG 10/4 gab es zwei Validierungstests: zunächst wurde 1996 ein Test zum Vergleich bisher bestehender Meßverfahren durchgeführt. Anschließend kam es zur Entwicklung des neuen Meßverfahrens PEAQ, und es wurde ein weiterer Validierungstest zur Auswahl der besten Modellversion von PEAQ und zum Vergleich von PEAQ mit bisherigen Verfahren durchgeführt.

2.2.1 Vergleich verschiedener Meßverfahren

Die sechs vorgeschlagenen objektiven Meßverfahren (DIX, NMR, PAQM, PERCEVAL, POM, Toolbox) wurden 1996 mit Hilfe der Datenbasis 2 und einem Teil der Datenbasis 1 auf ihre Leistungsfähigkeit untersucht. Die Testbeispiele und die Codecs in Datenbasis 2 wurden gemeinschaftlich von SR (Schweden) und BBC (England) zusammengestellt. Die Hörtests wurden von NRK (Norwegen), DR (Dänemark) und NHK (Japan) durchgeführt. Vertiefende statistische Betrachtungen wurden von Teracom (Schweden) und Deutsche Telekom (Deutschland) erarbeitet. Die konkurrierenden Proponenten erhielten die erste Hälfte der

Datenbasis 2 für eine letzte Adaption der Meßverfahren. Von Swisscom (Schweiz) wurden vor und nach der Adaption mit den Meßverfahren Messungen mit den Beispielen aus der gesamten Datenbank durchgeführt und für die Auswertung zur Verfügung gestellt.

Die Analyse der Leistungsfähigkeit der Meßverfahren wurde von Teracom und den Proponenten selber durchgeführt. Auch wenn einige der vorgeschlagenen Verfahren bereits eine hohe Korrelation zu den SDG-Werten produzierten, kam die TG 10/4 zu dem Schluß, daß keine der Methoden die Anforderungen der Anwender erfüllt. Die statistische Auswertung ergab weiterhin, daß keines der vorgeschlagenen Verfahren signifikant besser ist als die anderen. Es wurde daher entschieden, gemeinsam eine verbesserte Methode zu entwickeln. Zur Überprüfung der Leistungsfähigkeit der neuen Methode wurde das Verfahren, das in diesem ersten Vergleichstest die höchste Korrelation gezeigt hatte, als Referenzmodell für die weiteren Tests ausgewählt.

2.2.2 Entwicklung eines kombinierten Meßverfahrens

In Folge der unbefriedigenden Ergebnisse des ersten ITU Vergleichstests wurde beschlossen, die besten Elemente der verschiedenen Meßverfahren zu einer neuen Methode zu kombinieren. Das Validierung für diese neue Methode wurde ähnlich wie die Validierung in der vorangegangenen Wettbewerbsphase durchgeführt. Eine neue Datenbasis (DB3) mußte geschaffen werden. Die Beispiele und Codierungsbedingungen wurden im Frühjahr 1997 definiert und von SR, Swisscom und BBC zusammengestellt. Die BS.1116 konformen Hörtests wurden von SR, NHK und Deutsche Telekom durchgeführt. Eine ausführliche statistische Untersuchung der Hörtestergebnisse erfolgte u. A. durch Teracom. Die Ergebnisse wurden zu der 84 Teststücke enthaltenden „Datenbasis 3" (DB3) zusammengefaßt [ITU98].

Nach einem ersten Vergleich zwischen Modellvorhersagen und Hörtestergebnissen wurden den Entwicklern im Herbst 1997 52 Beispiele aus Datenbasis 3 überlassen, um eine letzte Anpassung des neuen Verfahrens zu ermöglichen. Zur Validierung wurden die verbleibenden 32 Teststücke aus DB3, sowie die Ergebnisse eines neuen Hörtests, der am CRC (Kanada) [SOU98] durchgeführt wurde, verwendet. Abschnitt 5 enthält eine detaillierte Beschreibung der Analyse der Leistungsfähigkeit des neuen Verfahrens anhand dieser Testergebnisse.

2.2.3 Anwendungen und Testsignale

Eine der ersten Arbeiten der ITU-R Task Group 10-4 war es, die Anwendungen für das Meßverfahren zu identifizieren und zu beschreiben. Eine Klassifizierung der Anwendungen ist in Tabelle 1 gegeben:

Tab. 1: Anwendungen

	Anwendungen	**Beschreibung**
1	Bewertung von Implementierungen	Charakterisierung verschiedener Implementierungen von Audiotechnik und Übertragungsstrecken, meist Codiersysteme
2	Kurzprüfung der Sendestrecke (Line-Up)	Schnelle Überprüfung der Strecke bevor der Service aufgenommen wird
3	On-Line Monitor	Kontinuierliche Überwachung der Audioübertragung während des Betriebes
4	Systemstatus	Detaillierte Analyse eines Teils der Ausrüstung oder einer Übertragungsstrecke
5	Codec-Identifizierung	Prozedur zur Identifizierung des Types und der Implementierung eines bestimmten Codecs
6	Codec-Entwicklung	Detaillierte Charakterisierung der Leistungsfähigkeit von Codecs
7	Netzwerk-Planung	Optimierung der Kosten und der Leistungsfähigkeit von Übertragungsnetzwerken unter gegebenen Einschränkungen
8	Unterstützung von subjektiven Bewertungsverfahren	Analyse großer Materialbestände auf kritische Audiosignale, die im Hörtest verwendet werden sollten.

Einige der Anwendungen erfordern Echtzeit-Implementierungen des Meßverfahrens, für andere Anwendungen ist die Bestimmung der Audioqualität in Nicht-Echtzeit ausreichend. Weiterhin muß zwischen Off-Line- und On-Line-Messung unterschieden werden. Bei Off-Line-Messungen kann die Strecke mit beliebigen Signalen getestet werden, während bei On-Line-Messungen das laufende Programm nicht unterbrochen werden darf.

Die Testsignale können in zwei Gruppen eingeteilt werden: natürliche und synthetische Testsignale. Natürliche Testsignale sind kritische Audiosequenzen, wie sie auch in Hörtests zur Bewertung der Audioqualität von Codecs verwendet werden. Die Dauer der natürlichen Testbeispiele beträgt meist 5 bis 20 Sekunden, wobei oft schon deutlich kürzere Teile des Signals die wahrnehmbaren Artefakte aufdecken. Im Testsystem sollte die Länge des Signals ähnlich abgestimmt sein, wie im Hörtest. Die Testbeispiele müssen sowohl auf der Sender-

seite als auch auf der Empfängerseite zur Verfügung stehen. Dementsprechend benötigt das Meßsystem für einige Anwendungen einen Speicher für das Referenzsignal.

Synthetische Signale sind mathematisch definiert und können sowohl auf der Sende- als auch auf der Empfängerseite bei Bedarf generiert werden. Zusätzlicher Speicher ist nicht erforderlich. Allerdings ist die Bestimmung der subjektiven Qualität von synthetischen Signalen sehr schwierig. Bislang liegen noch keine Ergebnisse aus BS.1116-konformen Hörtests mit synthetischen Signalen vor. Aus diesem Grund wurde auch das neue objektive Meßverfahren nicht mit synthetischen Meßsignalen verifiziert.

3 PEAQ - PERCEPTUAL EVALUATION OF AUDIO QUALITY

PEAQ (Perceptual Evaluation of Audio Quality) entstand in einer Zusammenarbeit aller an der Entwicklung der oben beschriebenen Meßverfahren beteiligten Organisationen. Im peripheren Gehörmodell sind Teile aller genannten Verfahren enthalten; die Qualitätsparameter stammen größtenteils aus DIX und NMR, aber auch aus PERCEVAL und OASE. Das neue Meßverfahren enthält sowohl ein FFT-basiertes, als auch ein filterbankbasiertes Gehörmodell. Die Qualitätsparameter werden zum Teil aus berechneten Verdeckungsschwellen und zum Teil aus dem Vergleich zwischen gehörangepaßten Signaldarstellungen gewonnen. Daneben sind auch Ausgangsparameter enthalten, die kein Gehörmodell verwenden, sondern auf einem direkten Vergleich von FFT-Spektren beruhen. Die einzelnen Qualitätsparameter werden mit Hilfe eines einfachen künstlichen neuronalen Netzes zu einer die globale Audioqualität beschreibenden Kenngröße zusammengefaßt.

3.1 Skalierungen

Um eine exakte Modellierung der Hörschwellen zu ermöglichen, wird aus dem Abhörpegel des Testsignals ein Skalierungsfaktor für die Eingangssignale berechnet. Sofern der Abhörpegel nicht bekannt ist, wird ein Pegel von 92 dB SPL für einen vollausgesteuerten Sinuston angenommen. Weiterhin muß sichergestellt werden, daß Testsignal und Referenzsignal keinen Zeitversatz aufweisen. Der Algorithmus zur Bestimmung und Ausgleich eines evtl. vorhanden Zeitversatzes ist kein Bestandteil des in der ITU-R Empfehlung beschriebenen Modells.

3.2 Peripheres Gehörmodell

3.2.1 FFT-basiertes Gehörmodell

In dem FFT-basierten Gehörmodell des Meßsystems, wird die spektrale Darstellung des Audio Signals anhand einer Diskreten Fouriertransformation, DFT berechnet. Mit Hilfe eines Hann-Fensters werden aus dem Zeitsignal jeweils um 50 % überlappende Blöcke von 2048 Abtastwerten ausgeschnitten. Bei einer Abtastrate von 48 kHz entspricht dies einer Zeitauflösung von ungefähr 21 ms.

Anschließend wird das Spektrum mit einer frequenzabhängigen Funktion (Gl. 1) gewichtet:

$$A(f)/dB = -0,6 \cdot 3,64 \cdot (f/kHz)^{-0,8} + 6,5 \cdot e^{-0,6 \cdot (f/kHz - 3,3)^2} - 10^{-3} \cdot (f/kHz)^4 . \quad (1)$$

Zusammen mit dem später hinzugefügten Innenohrrauschen (Gl. 3) modelliert diese Funktion die Ruhehörschwelle.

Im nächsten Schritt erfolgt eine Abbildung der gewichteten spektralen Energien von der Frequenzskala auf eine dem physiologischen Gegebenheiten des Ohrs angepaßte Bark-Skala.

Die Abbildung erfolgt nach einer von Schroeder [SCH79] ermittelten Näherung:

$$f/kHz \approx \sinh\left(\frac{z/Bark}{7}\right) \quad (2)$$

Um die Rechenleistung niedrig zu halten, wurde eine Auflösung von 0,25 Bark in der Basic Version und 0,50 Bark in der Advanced Version gewählt, was zu 109 bzw. 55 Analysebändern führt. Zur Modellierung eines Grundrauschens wird ein von der Mittenfrequenz f_c abhängiger Wert

$$Grundrauschen/dB = 0,4 \cdot 3,65 \cdot (f/kHz)^{-0,8} . \quad (3)$$

addiert.

Dieses Ergebnis ist Ausgangspunkt für die Berechnung der Verteilung der Energie entlang der Basilarmembran, die in zwei Schritten erfolgt. Im ersten Schritt erfolgt eine Verschmierung der in einem Frequenzband vorhandenen Energie über die gesamte Skala. Das hierfür eingesetzte Filter hat eine konstant ansteigende Flanke von 24 dB/Bark. Die abfallende Flanke wird nach einer Näherung von Terhardt [TER79] energie- und frequenzabhängig angesetzt:

$$\frac{Steigung}{dB\,/\,Bark} = -24 - \frac{230Hz}{fc} + 0.2 \cdot L\,/\,dB, \tag{4}$$

wobei L die lokale Energie und *fc* die Mittenfrequenz im aktuellen Analyseband ist.

Im zweiten Schritt werden die Einzelwerte nichtlinear anhand eines Potenzgesetzes („power-law model") nach dem Modell von Lufti [LUF83] verknüpft.

$$E_k = norm_k \cdot \left[\sum_{\forall i} {E_{i,k}}^{\alpha} \right]^{\frac{1}{\alpha}}, \tag{5}$$

Für den Faktor α gibt Lufti den Bereich von 0,3 bis 0,4 an. Aufgrund von experimenteller Optimierungen wurde für α der Wert 0.4 gewählt.

Der so errechnete Erregungspegel ist nur für stationäre Signale gültig. Vor- und Nachverdeckungseffekte sind noch nicht berücksichtigt. Da die zeitliche Auflösung des FFT-basierten Modells ungefähr 20 ms beträgt, wird die Vorverdeckung im Modell nicht berücksichtigt.

Zur Modellierung der Nachverdeckung werden die Erregungspegel durch IIR-Filter erster Ordnung zeitlich verschmiert. Die Zeitkonstanten des Tiefpaßfilters sind abhängig von der Mittenfrequenz der Analysebänder und werden anhand der folgenden Formel bestimmt:

$$\tau\left(f_{mitten}\right) = \tau_{\min} + \frac{100Hz}{f_{mitten}} \cdot \left(\tau_{100} - \tau_{\min}\right) \tag{6}$$

Die Konstanten $\tau_{\min}$ und τ_{100} haben den Wert von 30 ms und 8 ms.

Um sicherzustellen das bei einem plötzlichen Energieanstieg (Anschläge) die zeitliche Auflösung möglichst groß ist, wird als aktueller Wert für die Erregungspegel das Maximum zwischen den gefilterten und nicht gefilterten Werten genommen.

3.2.1.1 Eigenschaften des FFT-basierten Gehörmodells

In diesem Abschnitt sollen die Eigenschaften des Gehörmodells anhand von zwei Signalen demonstriert werden.

■ **1 kHz Sinus Ton**

Abbildung 4 zeigt die Erregungspegel für 1 kHz Sinustöne für unterschiedliche Pegel. Das Maximum der Kurven entspricht der Frequenz der Sinustöne. Der Anstieg der Erregung in den unteren Frequenzen entspricht dem Eigenrauschen des Gehörs. Die unterschiedliche Flankensteilheit der oberen Flanken ergibt sich durch die energieabhängige Verschmierung.

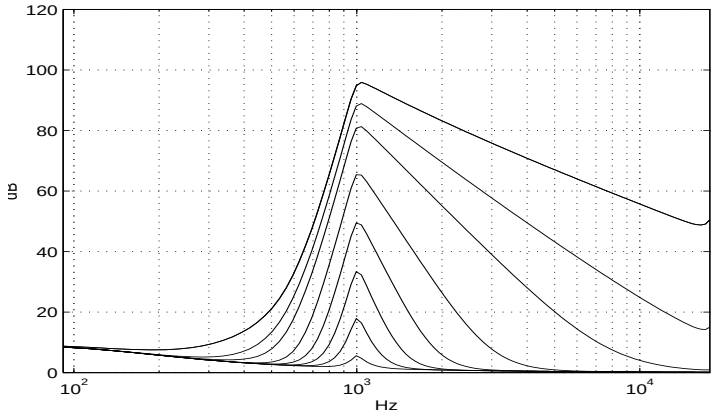

Abb. 4: Erregungspegel von 1 kHz Sinustönen unterschiedlicher Pegel

■ 100 ms Rauschimpuls

Abbildung 5 zeigt die zeitliche Verschmierung anhand eines 100 ms Rauschimpulses vor und nach der zeitlichen Filterung dar.

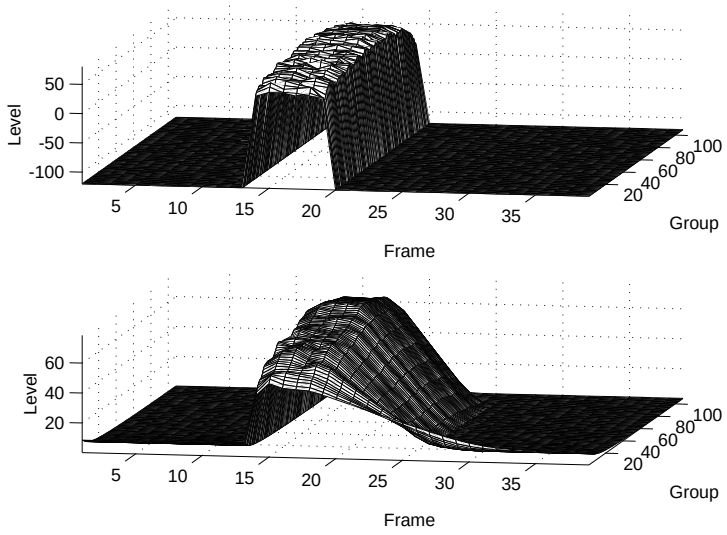

Abb. 5: Erregungspegel eines 100 ms Rauschimpulses als Funktion der Frequenz und Zeit. Oberes Bild: ohne Berücksichtigung der Nachverdeckung. Unteres Bild: mit Berücksichtigung der Nachverdeckung.

An der linken Seite, beim Anstieg des Signals, haben die Erregungspegel den gleichen Wert. An der rechten Seite, nach dem Abschalten des Signals, ist die modellierte Nachverdeckung als Verschmierung der Erregungspegel erkennbar.

3.2.2 Filterbankbasiertes Gehörmodell

Der filterbankbasierte Teil des Gehörmodells verwendet rekursiv berechenbare Filter mit zeitlich begrenzter Impulsantwort. Die Filter sind quasi-signalabhängig und wurden bereits in dem Meßverfahren DIX [THI96] verwendet. Eine detaillierte Beschreibung ist aber bisher nicht veröffentlicht worden und wird daher im Folgenden gegeben.

3.2.2.1 Verwendete Filterstruktur

Eine Sinusschwingung kann durch Berechnung der Ausdrücke $a_{n+1}=a_n \cdot cos(\varphi) - b_n \cdot sin(\varphi)$ und $b_{n+1}=a_n \cdot sin(\varphi) + b_n \cdot cos(\varphi)$ rekursiv fortgesetzt werden. Diese Berechnungsvorschrift kann als IIR-Filter erster Ordnung mit einem komplexen Koeffizient $e^{j\varphi}$ gedeutet werden, das eine unendlich lange Impulsantwort in Form einer Sinusschwingung hat (d. h. das Filter ist nicht stabil). Indem am Filtereingang das um einen geeigneten Phasenwinkel gedrehte Eingangssignal mit einer Verzögerung von N Abtastwerten zum ursprünglichen Eingangssignal addiert wird, kann die Impulsantwort des Filters nach N Samples künstlich abgebrochen werden. Bei Vernachlässigung der begrenzten Rechengenauigkeit kann das Filter dann als theoretisch stabil betrachtet werden. Das Filter hat keine Polstellen aber N-1 Nullstellen. Es verhält sich damit wie ein FIR-Bandpaß N-ter Ordnung obwohl es als rekursives Filter aufgebaut ist. Aufgrund dieser Eigenschaft wird dieses (linearphasige) Filter im Folgenden als RFIR (rekursives FIR-Filter) bezeichnet. Wenn die komplexe Multiplikation in Real- und Imaginärteil aufgeteilt wird, ergibt sich die in Abb. 6 gezeigte Filterstruktur.

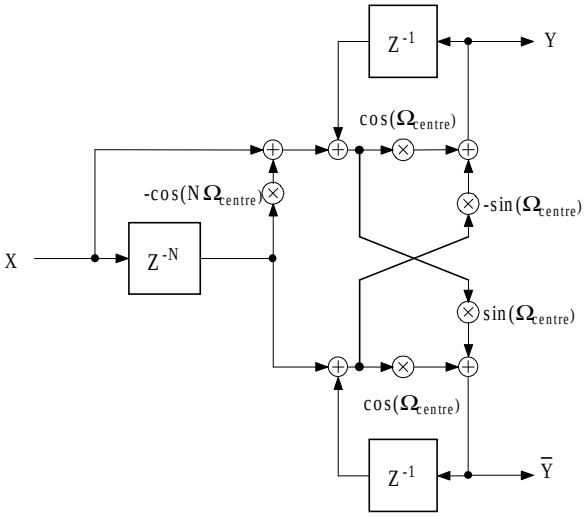

Abb. 6: Rekursive Filterstruktur zur Realisierung eines linearphasigen Bandpasses mit Real- und Imaginärteilausgang.

Das in Abb. 6 gezeigte Filterelement weist noch keine sehr selektive Bandpaßcharakteristik auf. Die Amplitude der Nebenmaxima fällt nur umgekehrt proportional zum Abstand von der Mittenfrequenz ab. Durch Parallelschalten von *K+1* Filtern mit leicht versetzten Mittenfrequenzen kann die Selektivität des Filters jedoch erhöht werden. Bei einer geeigneten Gewichtung der parallelgeschalteten Teilfilter läßt sich ein Abfall der Nebenmaxima des so

definierten RFIR-Filters K-ter Ordnung mit der (K+1)-ten Potenz des Abstands zur Filtermittenfrequenz erreichen. Die Impulsantwort des Filters hat dann die Form

$$a_K(n) = \sin^K\left(\frac{\pi}{N}n\right) \cdot \cos\left(\frac{\Omega_{centre}}{f_{samp}} \cdot n\right) \quad \Big| \quad 0 \le n < N \tag{7}$$

für den Realteil und

$$a_K(n) = \sin^K\left(\frac{\pi}{N}n\right) \cdot \sin\left(\frac{\Omega_{centre}}{f_{samp}} \cdot n\right) \quad \Big| \quad 0 \le n < N \tag{8}$$

für den Imaginärteil. Mit *K=2* entspricht die Filtercharakteristik damit der Frequenzselektivität einer Spektrallinie einer DFT mit Kosinusquadratfenster. Bei höherer Filterordnung nähert sich die Hüllkurve der Impulsantwort einer Gaußkurve. Allerdings wird der beschriebene Algorithmus ineffizient wenn *K* sehr groß wird. Für das beschriebene Gehörmodell hat sich eine Filterordnung von *K=2* als ausreichend erwiesen.

Die gewichtete Summation von *K+1* parallelen Bandpaßfiltern kann als Faltung des Amplitudenfrequenzganges des Filters mit der Fouriertransformierten eines Zeitfensters aufgefaßt werden. Daher wird ein solches Filter im Folgenden als *FDC (frequency domain convolution)-Filter* bezeichnet.

3.2.2.2 Aufbau der Filterbank

Die Filterbank besteht aus 40 FDC-Filtern zweiter Ordnung zwischen deren Ausgängen eine nach Real- und Imaginärteil getrennte gewichtete Summation vorgenommen wird, um die für das Gehörmodell erforderlichen exponentiellen Filterflanken zu erzeugen. Dieses Vorgehen entspricht formal der im FFT-basierten Gehörmodell vorgenommenen spektralen Verschmierung, jedoch mit dem entscheidenden Unterschied, daß diese Verschmierung *vor* der Gleichrichtung vorgenommen wird. Es handelt sich daher um eine quasi-lineare Operation, und die aus der Filtertheorie bekannte Relation zwischen Frequenzgang und Impulsantwort bleibt erhalten. Die durch die Verschmierung verringerte spektrale Auflösung führt daher zu einer Verkürzung der Impulsantwort und somit zu einer erhöhten zeitlichen Auflösung. Dies gilt näherungsweise auch noch bei der Modellierung pegelabhängiger Filterflanken, wobei die flacheren Filterflanken bei hohen Signalpegeln gleichzeitig zu einer erhöhten zeitlichen Auflösung führen.

Die Aufteilung der Mittenfrequenzen und Bandbreiten der Filter erfolgt entsprechend der von Zwicker [ZWI67] vorgeschlagenen Bark-Skala, wobei eine von Schroeder et. al. [SCH79] vorgeschlagene Näherung verwendet wird und eine Filterbreite von 0,6 Bark verwendet wird (Anm.: dieser experimentell gefundene Wert entspricht relativ genau der von Glasberg und Moore [GLA90] gefunden Filterbreite des Gehörs). Die exponentiellen Filterflanken werden entsprechend der von Terhardt [TER79] vorgeschlagenen Näherung pegelabhängig modelliert, wobei wegen der hohen zeitlichen Auflösung der Filterbank die zeitliche Änderung der Flankensteilheit durch einen Tiefpaß erster Ordnung begrenzt wird.

3.2.2.3 Gleichrichtung

Für ein physiologisch korrektes Modell der Signalverarbeitung im menschlichen Gehör wäre eigentlich eine Halbwellengleichrichtung die adäquate Gleichrichtungsmethode. Dies würde jedoch zu einer Reihe von Problemen bei der Weiterverarbeitung führen. Es wurde daher statt dessen die auch bei der Berechnung von DFT Spektren verwendete Methode der Betragsbildung über die Hilbert-Transformierte gewählt. Dabei werden die Betragsquadrate des Signals und seiner Hilbert-Transformierten (die auch als Imaginärteil eines komplexen Signals oder als 90° phasenverschobenes Signal interpretiert werden kann) addiert, woraus sich, wie sich mit Hilfe der Additionstheoreme für trigonometrische Funktionen leicht zeigen läßt, für den Fall einer Sinusschwingung das Amplitudenquadrat des Signals ergibt. Als besonders günstig erweist es sich hier, daß die in Abb. 6 gezeigte Filterstruktur bereits die Hilbert-Transformierte des Signals liefert. Der Hauptvorteil dieser Gleichrichtungsmethode liegt in der möglichen Unterabtastung der Filterausgänge und darin, daß für stationäre Eingangssignale auch ohne nachfolgende Glättung konstante Ausgangswerte vorliegen.

3.2.2.4 Zeitliche Verschmierung

Zeitliche Verdeckungseffekte werden durch eine Tiefpaßfilterung der gleichgerichteten Bandpaßsignale modelliert. Das Filter besteht aus zwei Teilfiltern: ein IIR-Tiefpaß erster Ordnung, der hauptsächlich Nachverdeckung modelliert, und ein FIR-Tiefpaß mit kosinusquadratförmiger Impulsantwort, der hauptsächlich Vorverdeckung modelliert. Die Zeitkonstanten für den IIR-Tiefpaß hängen von der Mittenfrequenz des zugehörigen Bandpasses ab, während die Länge des FIR-Tiefpasses für alle Frequenzbänder gleich ist. Die Koeffizientenzahl des FIR-Tiefpasses entspricht einer Impulsantwortlänge von 8 ms, was wiederum einer Vorverdeckungsdauer von ca. 2-4 ms entspricht. Für den IIR-Tiefpaß

erwiesen sich Zeitkonstanten von 50 ms bei 100 Hz (τ_{100}) und 4 ms in den obersten Frequenzbändern (τ_{min}) als optimal.

3.2.2.5 Modellierung der Ruhehörschwelle

Die Ruhehörschwelle wird in zwei Anteile aufgeteilt: eine Außen- und Mittelohrübertragungsfunktion, die durch eine spektrale Gewichtung der Frequenzbänder modelliert wird, und Eigenrauschen des Gehörs, das durch Addition eines frequenzabhängigen Offsets modelliert wird. Hierzu wird die in [TER79] gegebenen Näherung in zwei Teile aufgespalten. Das Eigenrauschen des Gehörs entspricht etwa 40% des tieffrequenten Anteils der Ruhehörschwelle (vgl. Gl. 3). Die Außen- und Mittelohrübertragungsfunktion entspricht dem verbleibenden Teil der Ruhehörschwelle (vgl. Gl. 1).

3.2.2.6 Eigenschaften des filterbankbasierten Gehörmodells

Dieser Abschnitt veranschaulicht das Verhalten der Filterbank anhand ihrer Antwort auf Impulse und reine Töne.

- **Antwort auf Sinustöne**

Abbildung 7 zeigt das beim Einschalten eines Sinustones bei 1 kHz entstehende Erregungsmuster. Die Position des Maximums der Erregung entspricht der Frequenz des Tones. Der Anstieg der Erregung in den unteren Filterbändern entspricht dem modellierten Eigenrauschen des Gehörs. Der vordere Teil gibt die Antwort der Filterbank auf stationäre Signale wieder, während der hintere Teil die Antwort auf die Sprungfunktion beim Einschalten des Tones zeigt. Der Übergang zwischen beiden Bereichen gibt die frequenzabhängige zeitliche Verschmierung wieder.

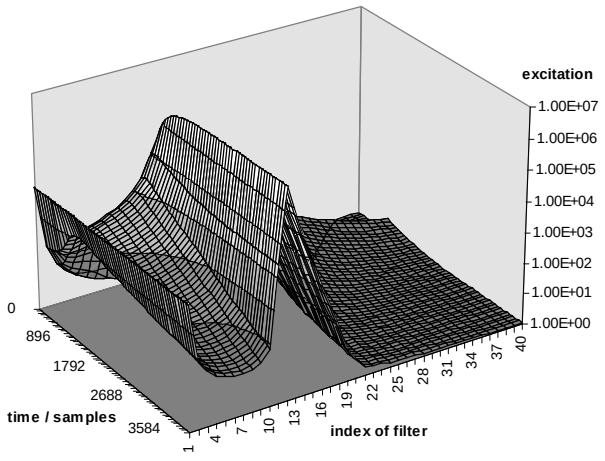

Abb. 7: Erregungsmuster nach dem Einschalten eines Sinustones bei 1 kHz.

- **Impulsantwort**

Abbildung 8 zeigt die Impulsantwort der Filterbank ohne zeitliche Verschmierung. Man erkennt, daß die zeitliche Auflösung im oberen Frequenzbereich um ein Vielfaches höher ist, als bei niedrigen Frequenzen (was einer der wichtigsten Vorzüge von Filterbänken gegenüber einer DFT ist).

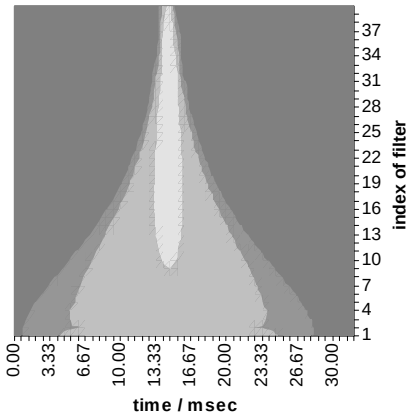

Abb. 8: Hüllkurve der Impulsantworten der FDC-Filterbank ohne zeitliche Verschmierung.

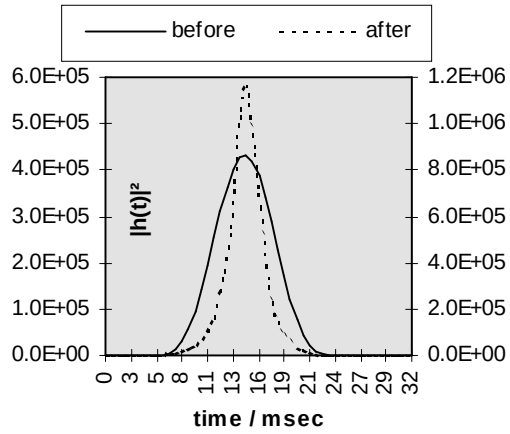

Abb. 9: Hüllkurve der Impulsantwort des zehnten Filterbandes der FDC-Filterbank vor und nach der Modellierung exponentieller Filterflanken (ohne zeitliche Verschmierung).

Abbildung 9 zeigt die Impulsantwort für ein einzelnes Filterband vor und nach der Modellierung der exponentiellen Filterflanken. Man erkennt, daß sich die Impulsantwort des Filters tatsächlich entsprechend der durch die exponentiellen Filterflanken verringerten spektralen Auflösung deutlich verkürzen.

3.3 Unterscheidung zwischen linearen und nichtlinearen Verzerrungen

Nicht jeder hörbare Unterschied zwischen dem Testsignal und dem Original wird tatsächlich als Fehler wahrgenommen. Dies gilt beispielsweise für einen konstanten Zeitversatz oder einen gleichbleibenden Amplitudenfehler. Auch langsame Amplitudenänderungen werden nicht unbedingt als Fehler wahrgenommen, oder sind zumindest weniger störend als additive Störungen. Es ist daher erforderlich, einen eventuell vorhandenen Zeit- und Pegelversatz auszugleichen, ehe nach hörbaren Störungen gesucht wird.

Darüber hinaus werden einige andere Arten von Störungen zwar als Fehler wahrgenommen, sind aber weniger lästig als additive Störungen. Dies gilt insbesondere für eine spektrale Verfärbung des Testsignals (lineare Verzerrungen). Dies wird in dem vorliegenden Meßverfahren berücksichtigt, indem die gehörangepaßte Signaldarstellung in zwei Teile aufgespalten wird: eine Signaldarstellung, in der lineare Verzerrungen kompensiert worden sind, und eine Signaldarstellung, in der die linearen Verzerrungen noch vorhanden sind.

Die Hauptschwierigkeit bei der Kompensation linearen Verzerrungen liegt in der Unterscheidung zwischen verstärkten bzw. abgeschwächten Komponenten des Originalsignals (lineare Verzerrungen) und neuen Signalanteilen (nichtlineare Verzerrungen), wobei letztere nicht mit kompensiert werden dürfen. Diese Unterscheidung erfolgt im Modell über die Annahme, daß durch lineare Verzerrungen entstandene Fehler innerhalb kleiner Ausschnitte der Zeit-Frequenz Ebene weitgehend dasselbe Muster aufweisen wie das Originalsignal, während sich für die durch nichtlineare Verzerrungen entstandenen Anteile eine deutliche geringere Ähnlichkeit ergibt. Die Anteile der linearen Verzerrungen werden daher unter Ausnutzung der Orthogonalitätsbeziehung

$$m = \frac{\int_{-\infty}^{\infty} A(x) \cdot B(x)\,\mathrm{dx}}{\int_{-\infty}^{\infty} [A(x)]^2\,\mathrm{dx}} . \qquad (9)$$

ermittelt und anschließend kompensiert. Die in Gl. (9) angegebene Orthogonalitätsbeziehung liefert den Anteil eines Signals A(x) am Signal B(x) unter der Voraussetzung, daß B(x) aus einem gewichteten Teil von A(x) und einer zu A(x) orthogonalen Fehlerkomponente besteht. Sie kann daher unter den oben genannten Voraussetzungen zur Bestimmung der durch lineare Verzerrungen entstandenen Fehleranteile verwendet werden. Der Angleich zwischen Original und Testsignal erfolgt dabei immer durch eine Abschwächung des jeweils stärkeren Signalanteils, da andernfalls Signalanteile, die eigentlich unterhalb der Ruhehörschwelle liegen, in den hörbaren Bereich angehoben werden könnten.

3.4 Bestimmung von Qualitätsparametern

Die globale Audioqualität des Testsignals wird bestimmt, indem zunächst eine Reihe verschiedener Qualitätsparameter berechnet werden, die abschließend mit Hilfe eines

künstlichen neuronalen Netzes [RUM86] auf eine Kenngröße zur Abschätzung der wahrgenommenen Qualität abgebildet werden. Die Berechnung der einzelnen Qualitätsparameter wird im Folgenden näher beschrieben.

3.4.1 Analyse von Hüllkurvenmodulationen

Die Bestimmung der Hüllkurvenmodulation innerhalb der einzelnen Filterbänder kann einen wichtigen Beitrag zur Modellierung verschiedener Hörphänomene liefern. Hierzu wird ein Modulationsmaß bestimmt, dessen spektrale Verteilung im Folgenden als Modulationsmuster bezeichnet wird. Das Modulationsmaß wird bestimmt, indem zunächst die Erregungsmuster ohne Modellierung der zeitliche Verschmierung über eine Potenzfunktion in vereinfachte Lautheitsmuster umgerechnet werden. Diese Werte, *E'(f, t)*, sowie die Beträge ihrer zeitlichen Ableitungen werden durch einen einfachen Tiefpaß zeitlich geglättet. Aus den sich ergebenden Größen, E_Δ und $\overline{E}$, wird das Modulationsmaß durch Normierung der mittleren zeitlichen Änderung der Hüllkurve auf ihren Mittelwert gewonnen:

$$Mod(f,t) = \frac{E_\Delta(f,t)}{1 + \frac{1}{c_1} \cdot \overline{E}(f,t)} \quad . \tag{10}$$

Dieses Modulationsmaß, *Mod(f, t)*, wird hauptsächlich zur Bestimmung des Schwellenfaktors verwendet (siehe Abschnitt 3.4.3). Es wird aber auch zur Berechnung eines eigenständigen Qualitätsmaßes verwendet (siehe Abschnitt 3.4.2).

3.4.2 Modulationsabweichung

Aus dem oben definierten Modulationsmuster kann eine einfaches Maß zur Charakterisierung der durch das Testobjekt verursachten Veränderung der zeitlichen Hüllkurven einzelner Signalanteile gewonnen werden. Hierzu wird zunächst für jedes Frequenzband eine lokale relative Modulationsdifferenz bestimmt (Gl. 11), die dann zeitlich und spektral gemittelt wird.

$$ModDiff(f,t) = w \cdot \frac{\left| Mod_{proc}(f,t) - Mod_{orig}(f,t) \right|}{offset + Mod_{orig}(f,t)} \quad . \tag{11}$$

Der im Nenner addierte Offset dient dabei in erster Linie zur Begrenzung der Werte bei sehr kleinem Modulationsmaß des Originalsignals, und der Gewichtsfaktor *w* modelliert eine unterschiedliche Gewichtung zwischen einer Zunahme und einer Abnahme des Modulationsmaßes.

3.4.3 Partielle Störlautheit

Als eine der wichtigsten Eigenschaften der im Signal vorliegenden Verzerrungen wird die partielle Lautheit der vorliegenden Störungen bestimmt. Die Formel zur Berechnung dieser Störlautheit wurde dabei so entworfen, daß die folgenden Forderungen erfüllt werden:

- Bei nicht vorhandenem oder nicht wirksamem Maskierer sollte die berechnete Störlautheit die in [ZWI67] angegebene Formel zur Lautheitsberechnung approximieren.
- In der direkten Umgebung der Verdeckungsschwelle sollte die berechnete Störlautheit näherungsweise proportional zum Energieverhältnis zwischen Störung und Maskierer sein.
- Der Grad der Lautheitsdrosselung bzw. die Höhe der Schwelle sollte sich in Abhängigkeit von den spektralen Modulationsmaßen von Test- und Originalsignal ergeben.

Eine Formel, die diese Forderungen erfüllt ist durch Gl. (12) gegeben:

$$N' = k \cdot \left(\frac{1}{s_{proc}} \cdot \frac{E_{thres}}{E_0} \right)^{\gamma} \cdot \left[\left(1 + \frac{\max\left(s_{proc} \cdot E_{proc} - s_{orig} \cdot E_{orig}, 0 \right)}{E_{thres} + s_{orig} \cdot E_{orig} \cdot e^{-\alpha \cdot \frac{E_{proc} - E_{orig}}{E_{orig}}}} \right)^{\gamma} - 1 \right]. \quad (12)$$

Die Gleichung enthält drei freie Parameter: den Faktor α, der die Lautheitsdrosselung bestimmt, sowie zwei Koeffizienten, die die Abbildung zwischen Modulationsmaß und Verdeckungsmaß definieren. Diese Abbildung wird durch Gl. (13) modelliert:

$$\begin{aligned} s_{proc} &= m_s \cdot Mod_{proc}(f,t) + c_s \\ s_{orig} &= m_s \cdot Mod_{orig}(f,t) + c_s \end{aligned}, \quad (13)$$

wobei m_s etwa 0.2 Sekunden beträgt und c_s Eins ist. Die übrigen Konstanten sind entweder reine Skalierungskonstanten, die keinen Einfluß auf die Funktion des Modells haben (wie E_0 und k), oder sind bereits fest vorgegeben (wie der Exponent γ, der nach [ZWI67] den Wert *0.23* hat).

Durch die Bildung des Maximums im Zähler von Gl. (12) werden Störkomponenten, bei denen dieser Ausdruck negativ wird nicht bewertet. Dieses Problem wird dadurch berücksichtigt, daß Gl. (12) zusätzlich auch mit vertauschtem Test- und Originalsignal berechnet wird und das Ergebnis mit einer Gewichtung von 0.5 zu dem ursprünglichen Ergebnis addiert wird.

Durch Anwendung des bis hierher beschriebenen Algorithmus auf die nach Kompensation der linearen Verzerrungen vorliegenden Erregungsmuster ergibt sich die *partielle Lautheit additiver Störungen*.

3.4.4 Hörbare Lineare Verzerrungen

Wenn man statt der Erregungsmuster von Test- und Originalsignal die vor und nach der Kompensation linearer Verzerrungen vorliegenden Erregungsmuster des Originalsignals miteinander vergleicht, liefert der in Abschnitt 3.4.3 beschriebene Algorithmus ein Maß für die *partielle Lautheit linearer Verzerrungen*.

3.4.5 Noise-to-Mask Ratio

Basierend auf dem „Masked Threshold Concept“ kann der NMR-Wert („Noise-to-Mask Ratio“) verwendet werden, um den Sicherheitsabstand zwischen den maximalen nicht hörbaren Fehler und den tatsächlichen Fehler zu schätzen.

NMR ist definiert als das Verhältnis von Fehlersignal zur Verdeckungsschwelle. Das lineare Mittel der NMR Werte über alle Analysebänder ist das lokale NMR eines Blocks. Zur Bestimmung der NMR Werte wird im Meßsystem aus dem Erregungspegel durch Gewichtung mit einer frequenzabhängigen Funktion die Verdeckungsschwelle geschätzt. Zur Abschätzung des Fehlersignals wird die Differenz der Beträge der Spektralkoeffizienten des Referenz- und Testsignals bestimmt. Diese Werte werden dann auf eine Bark-Skala abgebildet. Drei verschiedene Qualitätsparameter werden aus der NMR Berechnung abgeleitet:

- Der gesamte NMR - das arithmetische Mittel des lokalen NMR
- Der segmentierte NMR - das geometrische Mittel des lokalen NMR
- Die relative Anzahl der gestörten Blöcke, d. h. der Blöcke, in welchem in mindestens einem Band der lokale NMR größer als 1,5 dB ist.

Negative Werte des gesamten NMR und des segmentierten NMR geben eine Schätzung des vorhandenen Sicherheitsabstandes ab. Positive Werte geben eine Schätzung der hörbaren Fehlerenergie ab. Die relative Anzahl Blöcke mit NMR Wert größer als 1,5 dB ist ein Maß für den Anteil der gestörten Zeitabschnitt im Signal.

3.4.6 Relative Bandbreite

Audioübertragungssysteme reduzieren häufig die Bandbreite des Signals, was oft zu einer hörbaren Signalverfälschung führt. Um diesen Effekt zu erfassen, wird eine Schätzung der Signalbandbreite wie folgt durchgeführt:

Im oberen Frequenzbereich von 21,5 kHz bis 24 kHz wird im DFT-Spektrum die Frequenzlinie mit dem maximalen Wert bestimmt. Die Energie dieser Linie wird als Schätzung für die Sperrdämpfung des Tiefpaßfilters verwendet. Die oberste Frequenzlinie, die im Bereich von 0 bis 21,5 kHz diese Sperrdämpfung um 10 dB übersteigt, definiert die Bandbreite im aktuellen Block. Der Mittelwert über alle Blöcke ergibt die *relative Bandbreite* des Signals.

3.4.7 Detektionswahrscheinlichkeit

Die Wahrnehmungsschwelle und damit auch die Wahrscheinlichkeit der Wahrnehmung eines Unterschiedes zwischen den Eingangssignalen hängt vom Erregungspegel der Eingangssignale ab. Das Eingangssignal mit dem höheren Pegel bestimmt hierbei grundsätzlich den Arbeitspunkt. Bei natürlichen Signalen ist eine Verringerung des Pegels allerdings üblicherweise weniger deutlich wahrnehmbar als eine Vergrößerung des Pegels [SPO97]. In PEAQ wird deshalb ein gewichtetes Mittel gebildet:

$$L[k,n] = 0{,}3 \cdot \max(\tilde{E}_{ref}[k,n], \tilde{E}_{test}[k,n]) + 0{,}7 \cdot \tilde{E}_{test}[k,n] \qquad (14)$$

Für den so bestimmten Wert wird der kleinste wahrnehmbare Pegelunterschied s(k,n) berechnet [ZWI90]. Abhängig vom Vorzeichen der Fehlers werden unterschiedliche Kurvenformen für die Berechnung der Detektionswahrscheinlichkeiten verwendet: bei einer Verringerung des Pegels ist der Übergangsbereich von „keine Störung wahrnehmbar“ zu „Störung deutlich wahrnehmbar“ größer als bei einer Vergrößerung des Pegels [SPO97]. Beide Verteilungen erreichen bei einem Pegelunterschied von s(k,n) die Detektionswahrscheinlichkeit 50%. Bei binauralen Signalen wird für jedes Band der Kanal mit der höheren Wahrscheinlichkeit ausgewählt. Die Gegenwahrscheinlichkeiten der Wahrnehmung in den einzelnen Bändern eines Blockes werden zur gesamten Gegenwahrscheinlichkeit zusammengefaßt.

Die Wahrscheinlichkeit der Detektion aller Blöcke wird mittels einer Glättung und Maximumbildung zusammengefaßt. Erstere vermeidet eine Übergewichtung extrem kurzer seltener Fehler, letztere trägt dem Effekt Rechnung, daß in Hörtests der Zeitabschnitt mit den

am deutlichsten hörbaren Fehlern das Gesamtergebnis dominiert. Dies ergibt die *maximale geglättete Detektionswahrscheinlichkeit*.

Ein Maß für die Größe einer Störung ergibt sich durch die Normierung des absoluten Fehlers auf die Wahrnehmungsschwelle s(k,n). Bei binauralen Signalen wird hier für jedes Band der Kanal mit dem größeren normierten Fehler ausgewählt. Die normierten Fehler werden auf die nächst kleinere ganze Zahl abgerundet. Dadurch wird verhindert, daß eine große Zahl unhörbarer Fehler einen gleich großen Beitrag zum Gesamtergebnis liefert wie ein einzelner sehr großer Fehler. Die normierten Fehler aller Bänder eines Blocks werden addiert und ergeben so die schwellennormierte Lautheit des Fehlers.

Die schwellennormierten Lautheit aller Blöcke wird addiert und durch die Anzahl der Blöcke, in denen die Detektionwahrscheinlichkeit größer als 50% ist, dividiert. Der Logarithmus dieses Wertes ergibt die *schwellennormierte Störlautheit des mittleren gestörten Blockes*.

3.4.8 Harmonische Struktur des Fehlersignals

Unter bestimmten Umständen können bei der Audiocodierung Verzerrungen entstehen, die eine ausgeprägte harmonische Struktur, d. h. ein Linienspektrum mit regelmäßigen Abständen zwischen den einzelnen Spektrallinien, aufweisen. Die Wahrnehmbarkeit und insbesondere die Lästigkeit solcher Störungen kann in vielen Fällen gegenüber anders geformten Störungen gleicher Intensität stark erhöht sein. Daher wird mit Hilfe einer Vorgehensweise, die der in der Sprachverarbeitung oft verwendeten Cepstralanalyse ähnlich ist, ein Maß für die Ausprägung der harmonischen Struktur des Fehlersignals bestimmt. Hierzu wird die Autokorrelations-funktion der spektralen Pegeldifferenz zwischen Test- und Originalsignal berechnet. Die Höhe des Maximums der Fouriertransformierten dieser Funktion bildet das *Maß der harmonischen Fehlerstruktur*.

3.5 Kombination verschiedener Modelle

Beide in der gehörrichtigen Qualitätsbewertung von Audiosignalen üblichen Konzepte - Abstandsmaße zur Verdeckungsschwelle (siehe 1.1) und Abstandsmaße zwischen gehörange-paßten Signaldarstellungen (siehe 1.2) - haben ihre eigenen Vor- und Nachteile. PEAQ kombiniert daher Qualitätsparameter, die aus beiden Ansätzen gewonnen werden (vgl. 3.4). Je nach Anwendung werden dabei zwei verschiedene Modellvarianten vorgesehen: die

sogenannte „Basic Version" ist für Anwendungen gedacht, die eine geringe Rechenzeit erfordern, und die „Advanced Version" liefert die bestmögliche Vorhersage der empfundenen Audioqualität auf Kosten eines deutlich erhöhten Rechenaufwands.

3.5.1 Basic Version

Die Basic Version von PEAQ benutzt ausschließlich das FFT-basierte Gehörmodell für die Bestimmung der Qualitätsparameter. Die wegen der gröberen zeitlichen Auflösung der FFT fehlende Detailinformation wird zum Teil durch Verwendung einer größeren Anzahl von Qualitätsparametern ausgeglichen. Es werden alle oben beschriebenen Qualitätsparameter verwendet.

3.5.2 Advanced Version

Die Advanced Version von PEAQ benutzt das filterbankbasierte Gehörmodell für die Bestimmung aller Qualitätsparameter, die durch einen Vergleich gehörangepaßter Signaldarstellungen gewonnen werden, und das FFT-basierte Gehörmodell für die übrigen Qualitätsparameter. Zu der erstgenannten Gruppe gehören die partielle Lautheit additiver Störungen, die partielle Lautheit linearer Verzerrungen und das Maß für die Veränderung der zeitlichen Hüllkurven. Zur zweiten Gruppe gehören die „Noise-to-Mask Ratio" und das Maß für die Harmonische Fehlerstruktur. Die in der Basic Version verwendeten Maße für die Wahrscheinlichkeit und Häufigkeit von hörbaren Verzerrungen sind in der Advanced Version nicht notwendig.

4 BESTIMMUNG DER GLOBALEN AUDIOQUALITÄT

Eine objektive Meßmethode, die die Gehörwahrnehmung widerspiegelt, muß sich bei der Entwicklung, Anpassung und Validierung zwangsläufig sehr eng an subjektive Hörtests anlehnen. Zum einen ist eine klare und eindeutige Definition der subjektiven Qualitätsparameter notwendig, zum anderen muß die Anpassung des objektiven Verfahrens an die subjektiven Daten eine Generalisierung bewerkstelligen.

4.1 Basic Audio Quality

Die Festlegung der Meßparameter und Prozeduren für Hörtests ist selbst eine diffizile Aufgabe. Die Frage nach der Angemessenheit der Kriterien sowie nach der Zuverlässigkeit

und Gültigkeit der Ergebnisse stellt sich immer wieder erneut. Die Entwicklung von PEAQ fokussierte sich auf die globale Audioqualität („Basic Audio Quality" - BAQ). Dieser Parameter sowie eine Testmethode zu seiner Bestimmung durch Hörtests ist in der ITU-R Empfehlung BS.1116 definiert [ITU97]. Das grundlegende Prinzip dieser Testmethode wird im Folgenden kurz beschrieben. Der BS.1116 Test ist ein Triple-Stimulus-Test mit verdeckter Referenz. Der Hörer kann beliebig zwischen den drei Quellen „A", „B" und „C" wählen. Quelle „A" ist immer das Referenzsignal und somit bekannt. Quelle „B" und „C" ergeben sich durch eine Zuordnung des getesteten Signals und der jetzt versteckt dargebotenen Referenz. Die Zuordnung erfolgt zufällig und ändert sich von Versuch zu Versuch.

Der Hörer wird aufgefordert die Beeinträchtigung der Quelle „B" im Vergleich zu „A" und die Beeinträchtigung der Quelle „C" im Vergleich zu „A" auf einer fünfstufigen Beeinträchtigungsskala (Abb. 10) zu bewerten. Eine der Quellen „B" oder „C" ist die versteckte Referenz und sollte demnach als nicht beeinträchtigt bewertet werden. Jeder wahrgenommene Unterschied zwischen Referenz und einer anderen Quelle muß als eine Beeinträchtigung interpretiert werden. Üblicherweise wird nur das Attribut „Basic Audio Quality" benutzt. Es ist als globales Attribut definiert, das all wahrnehmbaren Unterschiede zu einem Kriterium zusammenfaßt.

Die Bewertungsskala wird als eine kontinuierliche Skala mit Ankerpunkten behandelt. Die Ankerpunkte bei 5.0 und 1.0 bilden die Grenzen der Bewertungsspannbreite (Abb. 10). Die Analyse der Ergebnisse aus Hörtests nach BS.1116 erfolgt üblicherweise auf der Basis der „Subjective Difference Grades" (SDGs):

$$SDG = Grade_{Testsignal} - Grade_{Referenz}$$

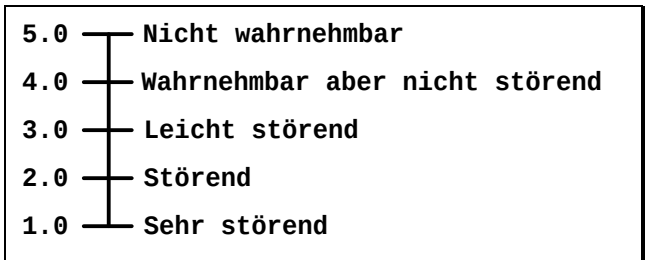

Abb. 10: Die ITU-R fünfstufige Bewertungsskala

Der SDG erhält, sofern der Hörer das versteckte Referenzsignal richtig zuordnet, einen negativen Wert zwischen 0 und -4. Ein SDG von 0 entspricht einer nicht wahrnehmbaren Beeinträchtigung, ein Wert von -4 entspricht einer Beeinträchtigung die als „sehr störend" wahrgenommen wird. Die Mittelwerte, die aus den Ergebnissen der beteiligten „erfahrenen Testhörer" (engl.: expert listeners) gebildet werden, charakterisieren zusammen mit den 95 % Vertrauensintervallen die empfundene Audioqualität für die einzelnen Testbeispiele.

Der „Objective Difference Grade" (ODG) ist der wesentlichste Ausgangsmeßwert des neuen Verfahrens und entspricht direkt dem SDG. Idealerweise wäre der ODG-Wert und der SDG-Wert für ein bestimmtes Signal im Test identisch. In der Praxis ist das schon aufgrund der statistischen Streuung der Ergebnisse subjektiver Tests nicht möglich [SPO96]. Zur Bewertung der Qualität des objektiven Verfahrens müssen die Abweichungen durch eine geeignete Metrik bewertet werden. Die inhärenten statistischen Eigenschaften, wie sie zum Beispiel in den Konfidenzintervallen erkennbar werden, müssen bei der Validierung eines objektiven Verfahrens berücksichtigt werden. In Abschnitt 5 werden verschiedene Bewertungsmethoden für die Validierung beschrieben.

4.2 Kalibrierung des Modells

Zur Kalibrierung des Modells wurde eine mehr als 600 codierte Musikstücke umfassende Datenbasis verwendet, die unter anderem die Daten fast aller zwischen 1990 und 1997 von MPEG und der ITU durchgeführten Codec-Vergleichstests enthielt. Zwei weitere Datensätze, ein von der ITU-R TG 10/4 speziell für die Validierung der Meßverfahren durchgeführter Test (DB3) und ein 1997 am CRC durchgeführter Test (CRC'97), wurden unter Verschluß gehalten und dienten der nachträglichen Überprüfung des Modells.

Das für die Abbildung der berechneten Qualitätsparameter auf einen Schätzwert für die Basic Audio Quality verwendete künstliche neuronale Netz [RUM86] enthält eine versteckte Schicht mit drei (Basic Version) bzw. fünf (Advanced Version) Knoten. Es verwendet eine unsymmetrische Sigmoidfunktion als Aktivierungsfunktion. Alle Eingangs- und Ausgangswerte wurden zum Trainieren des Netzes auf einen Wertebereich von Null bis Eins normiert, und es wurde ein Lernalgorithmus nach [JAC88] verwendet. Um eine Überanpassung des neuronalen Netzes zu verhindern, wurde ein Teil der verfügbaren Daten vom Training ausgenommen und diente zum Testen der Generalisierungsfähigkeit des Netzes. Das Trainieren des Netzes wird abgebrochen, wenn der Vorhersagefehler für diese Auswahl ein Minimum erreicht. Da sich durch Verwenden dieser Auswahl als Abbruchkriterium zumindest ein indirektes Training auf diese Testdaten ergibt, wurden die sich ergebenden Netze nach Abschluß des Trainings anhand der unter Verschluß gehaltenen Datensätze (DB3 und CRC'97) einem „echten" Generalisierungstest unterzogen.

5 VALIDIERUNG UND LEISTUNGSFÄHIGKEIT DES VERFAHRENS

Zur Auswahl der besten Modellversion(en) von PEAQ, sowie zur Überprüfung, ob PEAQ tatsächlich den bisher existierenden Meßverfahren überlegen ist, wurden die bis zum Ende der Trainingsphase unter Verschluß gehaltenen Datenbasen DB3 und CRC'97 verwendet. Dabei diente das in einem 1996 von der ITU-R durchgeführten Vergleichstest am besten bewertete Meßverfahren als Referenzmodell. Die Validierung wurde in zwei Stufen durchgeführt: zunächst war der gesamte Validierungsdatensatz unter Verschluß und die Leistungsfähigkeit der Verfahren wurde anhand des Datensatzes DB3 verglichen. Hierbei ergab sich eine deutliche Überlegenheit von PEAQ gegenüber dem Referenzmodell. In der zweiten Stufe wurde ein Teil des Datensatzes DB3 zur Kalibrierung der Modelle freigegeben und der verbleibende Teil diente zusammen mit dem Datensatz CRC'97 zur Überprüfung der Modelle (vgl. 2.2.2). Der Vergleich wurde anhand einer Reihe verschiedener Kriterien vorgenommen und führte zur Auswahl der letztendlich in die ITU-R Empfehlung aufgenommenen Modellversionen.

5.1 Vergleichskriterien

Die Entscheidung, welcher Meßalgorithmus der „beste" ist, läßt sich oft nicht anhand einer einzelnen Eigenschaft treffen. Wenn z. B. in einem Modell viele geringfügige Abweichungen zwischen Modellvorhersage und tatsächlich empfundener Qualität auftreten und in einem anderem Modell wenige, aber dafür deutlichere Abweichungen, welches Modell ist dann „besser"? Um den vielfältigen möglichen Betrachtungsweisen gerecht zu werden, wurden bei der Entscheidung über das in die ITU-Empfehlung aufzunehmende Modell eine Reihe verschiedener Vergleichskriterien berücksichtigt.

5.1.1 Toleranzschema

Ein möglicher Weg zur Auswahl eines Meßverfahrens ist die Festlegung eines Toleranzbereiches um die aus Hörtests ermittelten Qualitätsbewertungen. Ein Modell wird als ausreichend genau betrachtet, wenn seine Modellvorhersagen innerhalb dieses Toleranzbereiches liegen.

Das in der ITU-R TG 10/4 verwendete Toleranzschema ist in Abb. 11 dargestellt. Die Breite des Toleranzbereichs hängt vom Vertrauensintervall der Hörtests sowie vom Qualitätsbereich der Teststücke ab. Da die Abweichungen zwischen Modellvorhersagen und Hörtestergebnissen - bei einer sehr großen Anzahl von Testdaten - näherungsweise normalverteilt sind (vgl. Abb. 18 - 19), kann allerdings nicht erwartet werden, daß *alle* Modellvorhersagen innerhalb des Toleranzbereiches liegen. Als weiteres Vergleichskriterium dient daher der mittlere Abstand der außerhalb des Toleranzbereiches liegenden Modellvorhersagen von den Grenzen des Toleranzbereichs.

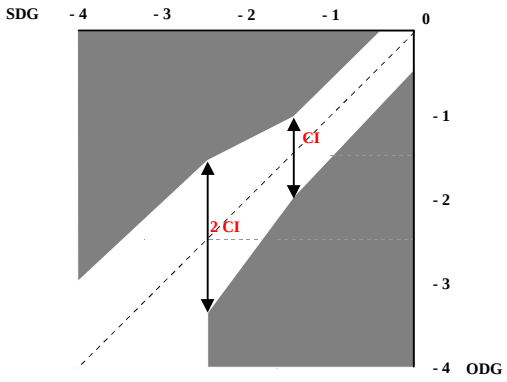

Abb. 11: Toleranzschema bei einem konstanten Vertrauensintervall (CI) von 0,25.

5.1.2 Korrelation und mittlerer Fehler

Die einfachsten und am weitesten verbreiteten Kenngrößen zur Messung der Übereinstimmung zwischen Modellvorhersagen und tatsächlichen Ergebnissen sind Korrelation und mittlerer quadratischer Fehler. Die Verwendung dieser Werte setzt jedoch eigentlich eine äquidistante Skala voraus, was für die in der Codecbewertung übliche fünfstufige Bewertungsskala nach [ITU90] offensichtlich nicht der Fall ist. Da aber auch die anderen Beurteilungsmethoden nicht unproblematisch sind, wurde die Korrelation dennoch berücksichtigt.

5.1.3 Gewichteter Fehlerbetrag

Der gewichtete Fehlerbetrag (AES - „Absolute Error Score") wurde eingeführt, um die vom Modell erwartete Genauigkeit ins Verhältnis zu der durch die Vertrauensintervalle gegebenen Zuverlässigkeit der aus Hörtests gewonnenen Bewertungen zu setzen. Er wird durch die folgende Gleichung definiert:

$$AES = 2 \cdot \sqrt{\frac{1}{N} \cdot \sum_{n=0}^{N-1} \left(\frac{ODG(n) - SDG(n)}{\max[CI(n), 0.25]} \right)^2} \,, \qquad (15)$$

wobei *CI* das Vertrauensintervall ist.

Für ein Modell, in dem der Vorhersagefehler immer innerhalb des Vertrauensintervalls liegt, bleibt der AES deutlich unter 2,0. Das Problem bei Verwendung des AES besteht in der in

manchen Bereichen zu starken Abhängigkeit vom Vertrauensintervall - insbesondere, da Teststücke mit kleinem Vertrauensintervall nicht immer die Teststücke sind, für die die größtmögliche Vorhersagegenauigkeit erforderlich ist (z. B. wird das Vertrauensintervall am unteren Ende der Bewertungsskala durch Sättigungseffekte sehr klein). Konkret zeigte sich, daß der AES vieler Meßverfahren bei Verwendung der Daten aus der ITU-R Validierungs-datenbasis überwiegend durch ein einzelnes Teststück bestimmt wurde, was offenbar keine sinnvolle Bewertung darstellt.

5.1.4 Ausreißerhäufigkeit

Ein weiteres verwendetes Vergleichskriterium ist die Anzahl von Teststücken, in denen der Vorhersagefehler eine bestimmte Grenze überschreitet. Dabei wurde diese Grenze einmal in Abhängigkeit vom Vertrauensintervall der subjektiven Bewertungen festgelegt und einmal durch einen absoluten Wert auf der SDG-Skala. Es wurden sowohl einfache Ausreißer, bei denen der Vorhersagefehler größer als das einfache Vertrauensintervall war, gezählt, als auch verschiedene Gruppen von weiten Ausreißer, bei denen die Abweichungen größer als das zweifache Vertrauensintervall bzw. größer als eine, 1,5 oder 2 Stufen auf der Bewertungsskala waren.

5.2 Ergebnisse

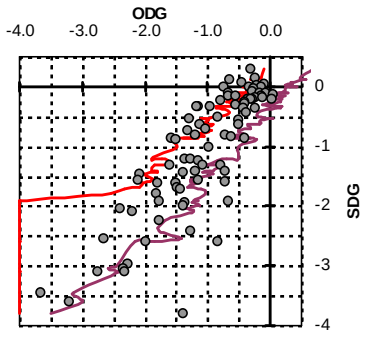

Abb. 12: Ergebnisse der Advanced Version von PEAQ für Datenbasis 3.

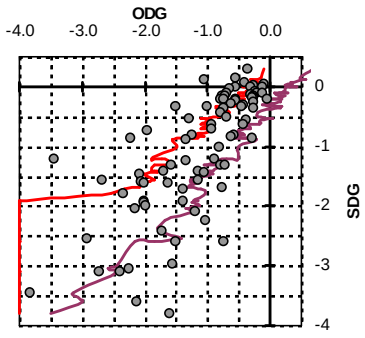

Abb. 13: Ergebnisse der Basic Version von PEAQ.

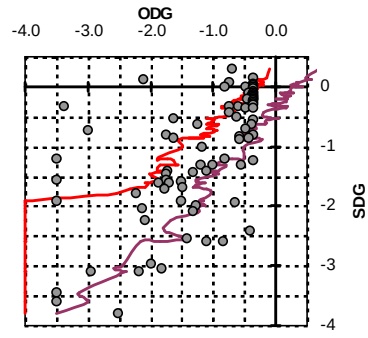

Abb. 14: Ergebnisse des Referenzmodells.

Wie die Abbildungen 12 - 14 zeigen, ergibt sich anhand der Ergebnisse für die ITU-R Validierungsdatenbasis (DB3) eine eindeutige Rangfolge zwischen beiden Versionen von PEAQ und dem Referenzmodell (die durchgezogenen Kurven markieren den im Idealfall

einzuhaltenden Toleranzbereich). Die Advanced Version von PEAQ ist dem Referenzmodell offenbar deutlich überlegen, was sich auch anhand der meisten der oben beschriebenen Vergleichskriterien bestätigte. Auch die Basic Version von PEAQ zeigte noch eine deutlich bessere Übereinstimmung mit den Hörtestdaten als das Referenzmodell, wobei der Unterschied aber etwas weniger deutlich ist.

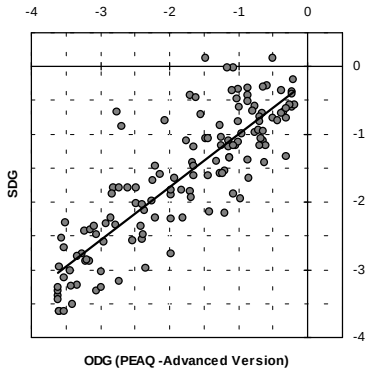

Abb. 15: Ergebnisse der Advanced Version von PEAQ für die CRC'97 Datenbasis.

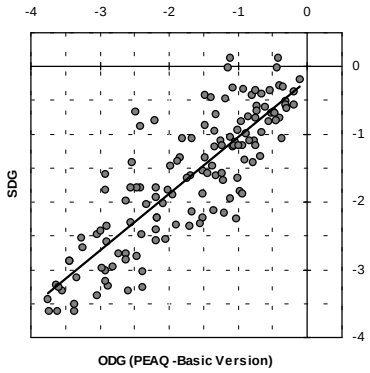

Abb. 16: Ergebnisse der Basic Version von PEAQ.

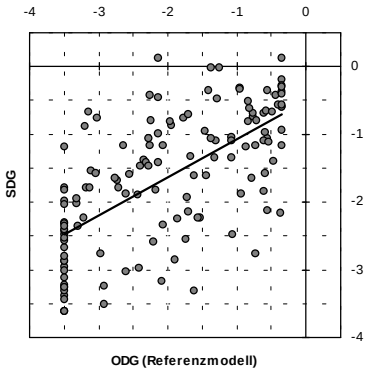

Abb. 17: Ergebnisse des Referenzmodells.

Der Vergleich der Meßergebnisse mit der zweiten unter Verschluß gehaltenen Datenbasis ergibt einen noch deutlicheren Vorsprung von PEAQ zu dem Referenzmodell, während die Unterschiede zwischen Advanced Version und Basic Version hier nur sehr gering sind (Abb. 15 - 17). Die Ergebnisse für alle verfügbaren Datensätze (einschließlich der zur Modellkalibrierung verwendeten Daten) werden in den Abbildungen 18 - 19 gezeigt. Hier ist allerdings zu beachten, daß diese Darstellungen wegen der darin enthaltenen zur Modell-kalibrierung verwendeten Daten keinen Aufschluß über die „tatsächliche" Leistungsfähigkeit des Verfahrens geben.

Anhand der im den Abbildungen zu beobachtenden Streuung wird deutlich, daß PEAQ trotz der insgesamt hervorragenden Ergebnisse und seiner Überlegenheit über andere Meßverfahren dennoch keine 100%-ige Übereinstimmung mit subjektiven Hörtestergebnissen liefert. In der Anwendung zum Vergleich verschiedener Codecs sollte daher immer eine ausreichende Anzahl verschiedener Teststücke verwendet werden, wodurch sich die verbleibende Unsicherheit weitgehend eliminieren läßt.

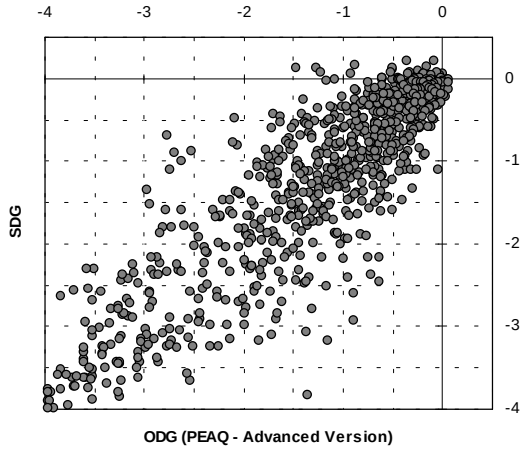

Abb. 18: Modellvorhersagen versus Hörtestergebnissen für alle vorhandenen Testdaten (Advanced Version).

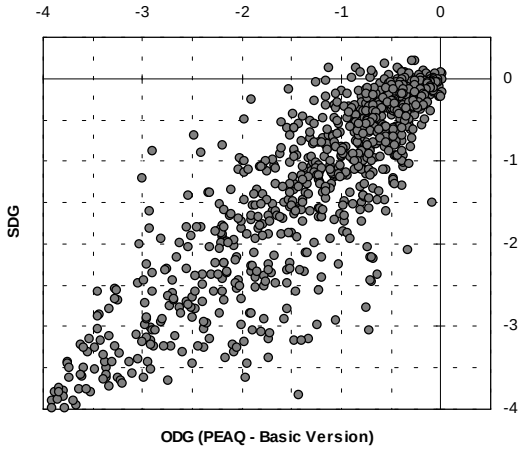

Abb. 19: Modellvorhersagen versus Hörtestergebnissen für alle vorhandenen Testdaten (Basic Version).

6 ZUSAMMENFASSUNG

Das Meßverfahren PEAQ bildet den künftigen ITU-Standard zur objektiven Messung der wahrgenommen Audioqualität. PEAQ ist eine internationale Gemeinschaftsentwicklung mehrerer Firmen und Forschungsinstitute und kombiniert Konzepte aus verschiedenen bisher bekannten Meßmethoden. Es mißt lineare und nichtlineare Verzerrungen, Abstände zu einer berechneten Verdeckungsschwelle und Änderungen der zeitlichen Struktur des Testsignals. Dabei wird u. a. auch die harmonische Struktur der Verzerrungen berücksichtigt. Mittels eines künstlichen neuronalen Netzes wird aus diesen verschiedenen Qualitätsparametern ein globales Maß zur Abschätzung der empfundenen Audioqualität gewonnen.

Im ITU-Standard sind zwei Versionen der Meßmethode enthalten: die sogenannte Basic Version ist für Anwendungen gedacht, die eine geringe Rechenzeit erfordern, und die Advanced Version liefert die bestmögliche Vorhersage der empfundenen Audioqualität auf Kosten eines deutlich erhöhten Rechenaufwands. Die von der ITU durchgeführten Vergleichstests haben gezeigt, daß beide Versionen von PEAQ eine genauere Abschätzung der empfundenen Audioqualität ermöglichen als die bisher existierenden Meßverfahren, wobei die Advanced Version eine gegenüber der Basic Version nochmals leicht verbesserte

Genauigkeit aufweist. PEAQ stellt zwar das zur Zeit bestmögliche Verfahren zur objektiven Bestimmung der empfundenen Qualität codierter Audiosignale dar, eine zuverlässige Einschätzung der Qualität einzelner Codecs ist aber nur bei Verwendung einer ausreichenden Zahl verschiedener Teststücke möglich.

7 DANKSAGUNG

Wir danken allen Personen und Institutionen, die die Entwicklung von PEAQ unterstützt und begleitet haben, insbesondere dem Chairman der TG10/4 Thomas Rydén und den an der Validierung beteiligten Institutionen British Broadcasting Corporation (BBC), Danmarks Radio (DR), Deutsche Telekom AG Berkom, Nippon Hoso Kyoka (NHK), Norsk Rikskringkasting (NRK), Sveriges Radio (SR), Swisscom und Teracom. Besonderer Dank gilt auch der ITU, die den Rahmen für eine fruchtbare Zusammenarbeit zur Verfügung gestellt hat.

8 LITERATUR

[AUR84] W. Aures: „BERECHNUNGSVERFAHREN FÜR DEN WOHLKLANG BELIEBIGER SCHALLSIGNALE, EIN BEITRAG ZUR GEHÖRBEZOGENEN SCHALLANALYSE". Dissertation an der Fakultät für Elektrotechnik der Technischen Universität München, September 1984.

[BEE92] J.G. Beerends and J.A. Stemerdink: „A PERCEPTUAL AUDIO QUALITY MEASURE BASED ON A PSYCHOACOUSTIC SOUND REPRESENTATION". J. Audio Eng. Soc., Vol. 40, S. 963-978, December 1992.

[BEE94] J.G. Beerends and J.A. Stemerdink: „A PERCEPTUAL SPEECH QUALITY MEASURE BASED ON A PSYCHOACOUSTIC SOUND REPRESENTATION". J. Audio Eng. Soc., Vol. 42, S. 115-123, March 1994.

[BEE96] Beerends, J. G.; van den Brink, W. A. C.: THE ROLE OF INFORMATIONAL MASKING AND PERCEPTUAL STREAMING IN THE MEASUREMENT OF MUSIC CODEC QUALITY. Contribution to the 100th AES Convention, Copenhagen, May 1996, Preprint 4176.

[BEE98] J. G. Beerends: „AUDIO QUALITY DETERMINATION BASED ON PERCEPTUAL MEASUREMENT TECHNIQUES". in M. Kahrs and K. Brandenburg (editors): *Applications of Digital Signal Processing to Audio and Acoustics*. The Kluwer International Series in Engineering and Computer Science, Volume 437, Kluwer Academic Publishers, Boston, March 1998.

[BIS74] G. von Bismarck: „SHARPNESS AS AN ATTRIBUTE OF THE TIMBRE OF STEADY SOUNDS". *Acustica* 30, 1974, S. 159 - 172.

[BRA87] K. Brandenburg: „EVALUATION OF QUALITY FOR AUDIO ENCODING AT LOW BIT RATES". Proceedings of the 82nd AES Convention, London 1987, Preprint 2433.

[BRE81] A. S. Bregman: „ASKING THE "WHAT FOR" QUESTION IN AUDITORY PERCEPTION". in M. Kubovy and J. R. Pomerantz (editors): *Perceptual organization*. Hillsdale, New York, 1981, S. 99-118.

[COH92] E. A. Cohen and L.D. Fielder: „DETERMINING NOISE CRITERIA FOR RECORDING ENVIRONMENTS". *J. Audio Eng. Soc*. Vol. 40, S. 384-402, May 1992.

[COL93] Colomes, C.; Lever, M.; Rault, J. B.; Dehery, Y. F.: A PERCEPTUAL MODEL APPLIED TO AUDIO BIT-RATE REDUCTION. Contribution to the 95th AES Convention, New York, October 1993, Preprint 3742.

[FAS74] Fastl, H.: „ MITHÖRSCHWELLEN ALS MAß FÜR DAS ZEITLICHE UND SPEKTRALE AUFLÖSUNGSVERMÖGEN DES GEHÖRS". Dissertation an der Fakultät für Maschinenwesen und Elektrotechnik der Technischen Universität München, München, 1974.

[GLA90] Glasberg, B. R.; Moore, B. J.: „ DERIVATION OF AUDITORY FILTER SHAPES FROM NOTCHED NOISE DATA". Hearing Research, *Vol. 47*, 1990, S. 103-138.

[HER92] J. Herre, E. Eberlein, H. Schott and Ch. Schmidmer: ANALYSIS TOOL FOR REAL TIME MEASUREMENTS USING PERCEPTUAL CRITERIA. *AES 11th International Conference*, Portland, Oregon, USA, 1992, S. 169-179.

[ITU90] Recommendation ITU-R BS.562-3, *Subjective assessment of sound quality*, 1990.

[ITU96] Recommendation ITU-T P.861, „OBJECTIVE QUALITY MEASUREMENT OF TELEPHONE-BAND (300-3400 HZ) SPEECH CODECS". 1996.

[ITU97] Recommendation ITU-R BS.1116(Rev.1), „METHODS FOR THE SUBJECTIVE ASSESSMENT OF SMALL IMPAIRMENTS IN AUDIO SYSTEMS INCLUDING MULTICHANNEL SOUND SYSTEMS". 1997.

[ITU98] Chairman, ITU-R Task Group 10/4: „REPORT ON THE SIXTH MEETING OF ITU-R TASK GROUP 10/4". Doc. 10-4/21, Geneva, 1998.

[JAC88] Jacobs, R.A.: „INCREASED RATES OF CONVERGENCE THROUGH LEARNING RATE ADAPTATION". *Neural Networks*, 1:295-307, 1088.

[KAR85] M. Karjalainen: „A NEW AUDITORY MODEL FOR THE EVALUATION OF SOUND QUALITY OF AUDIO SYSTEM", Proceedings of the ICASSP, Tampa, Florida, S. 608-611, March 1985.

[LEE84] M. R. Leek and C. S. Watson: „LEARNING TO DETECT AUDITORY PATTERN COMPONENTS". JASA, 1984, vol 76, S. 1037-1044.

[LUF83] R. A. Lufti: „ADDITIVITY OF SIMULTANEOUS MASKING". JASA, 1983, vol 73, S. 262-267.

[MCA84] S. McAdams: „THE AUDITORY IMAGE: A METAPHOR FOR MUSICAL AND PSYCHOLOGICAL RESEARCH ON AUDITORY ORGANIZATION". in: W. R. Crozier and A. J. Chapman (editors): *Cognitive Processes in the perception of art.* Elsevier Science Publishers, North-Holland, 1984, S. 289-323.

[MOO86] B.C. Moore: „FREQUENCY SELECTIVITY IN HEARING". Academic Press, London, 1986.

[MOO89] B.C. Moore: „AN INTRODUCTION TO THE PSYCHOLOGY OF HEARING". Academic Press, London, 1989.

[PAI92] B. Paillard, P. Mabilleau, S. Morisette, and J. Soumagne: „PERCEVAL: PERCEPTUAL EVALUATION OF THE QUALITY OF AUDIO SIGNALS". *J. Audio Eng. Soc.*, Vol. 40, 21-31, 1992.

[SCH79] M.R. Schroeder, B.S. Atal, and J.L. Hall: „OPTIMIZING DIGITAL SPEECH CODERS BY EXPLOITING MASKING PROPERTIES OF THE HUMAN EAR". *J. Acoust. Soc. Am.*, Vol. 66, S. 1647-1652, 1979.

[SOU98] Soulodre, G.A., Grusec, T., Lavoie, M., and Thibault, L.: „SUBJECTIVE EVALUATION OF STATE-OF-THE-ART 2-CHANNEL AUDIO CODECS". *J. Audio Eng. Soc.*, March, 1998.

[SPI92] J. Spille: „MESSUNG DER VOR- UND NACHVERDECKUNG BEI IMPULSEN UNTER KRITISCHEN BEDINGUNGEN". Internal report, Tomson Consumer Electronics, Hannover 1992.

[SPO96] Sporer, Th.: „EVALUATING SMALL IMPAIRMENTS WITH THE MEAN OPINION SCALE - RELIABLE OR JUST A GUESS?". Contribution to the 101st AES Convention, Los Angeles, October 1996, Preprint 4396.

[SPO97] Sporer, Th.: „OBJECTIVE AUDIO SIGNAL EVALUATION-APPLIED PSYCHOACOUSTICS FOR MODELING THE PERCEIVED QUALITY OF DIGITAL AUDIO". Contribution to the 103rd AES Convention, New York, September 1997, Preprint 4512.

[TER79] Terhardt, E.: CALCULATING VIRTUAL PITCH. Hearing Research, *Vol. 1,* 1979, S. 155-182.

[THI96] T. Thiede and E. Kabot: „A NEW PERCEPTUAL QUALITY MEASURE FOR BIT RATE REDUCED AUDIO". Proceedings of the 100th AES Convention, Copenhagen 1996, Preprint 4280.

[ZWI67] Zwicker, E.; Feldtkeller, R.: „DAS OHR ALS NACHRICHTENEMPFÄNGER". Stuttgart: Hirzel Verlag, 1967.

[ZWI90] E. Zwicker and H. Fastl: „PSYCHOACOUSTICS, FACTS AND MODELS". Berlin; Heidelberg: Springer Verlag, 1990.