Qualitätsbeurteilung von Audiosignalen – Vom Hörtest zum Messverfahren

Thomas Sporer[1]

[1] *Fraunhofer IDMT, 98693 Ilmenau, spo@idmt.fraunhofer.de*

Einleitung

Bedingt durch den Einsatz psychoakustischer Modelle in Sprach- und Audiocodierung ist die Beurteilung der empfundenen Audioqualität mittels klassischer Messwerte wie z.B. SNR und Klirrfaktor nicht mehr sinnvoll. Die letzte Instanz ist daher auch heute noch der Hörtest. Damit die Ergebnisse von Hörtests untereinander vergleichbar sind, braucht es standardisierte Testmethoden.

Diese Veröffentlichung beschreibt zunächst die grundsätzlichen Ansätze heutiger Hörtestmethoden und ihre Abbildung in Standards der ITU. Der Schwerpunkt liegt hierbei bei Methoden aus dem Bereich der Sprachqualität (ITU-T P.8xx) bzw. der hohen Audioqualität (ITU-R BS.1116). Im zweiten Teil wird das auf psychoakustischer Modellierung beruhende und standardisierte Messverfahren PEAQ (perceptual evaluation of audio quality) beschrieben. Im Zuge der Standardisierung in der ITU-R wurde verifiziert, dass PEAQ alle damals bekannten Audiocodier-Verfahren gehörrichtig bewerten kann. Den Abschluss bildet eine Zusammenfassung des aktuellen Standes der Forschung bezüglich der Erweiterungen und Anpassungen von PEAQ für neuere Codierverfahren, neue Hörtestmethoden sowie für räumliche Audiosignale.

Dimensionen der Audioqualität

Bei der Bewertung von Audiosignalen sind verschiedene Betrachtungsweisen sinnvoll: Während es bei der Bewertung der Qualität von Produkten oder auch von Musik um rein ästhetische Gesichtspunkte geht, sind bei der Bewertung von Lärm neben der reinen, durchaus physikalisch motivierten, Lautstärke der zeitliche und spektrale Verlauf zu berücksichtigen. Die vorliegende Veröffentlichung betrachtet aber die Bewertung von Audioübertragungsstrecken für Sprach- oder Audiosignale. Wichtige Kriterien bei Sprachsignalen sind Sprachverständlichkeit, Natürlichkeit der Sprachklanges sowie die Möglichkeit, den Sprecher zu identifizieren und seine emotionale Befindlichkeit bewerten zu können. Bei der Übertragung von beliebigen Audiosignalen steht oft die Forderung im Mittelpunkt, dass Input und Output einer Übertragungsstrecke für den Menschen ununterscheidbar klingen. Sind hörbare Unterschiede unvermeidbar, so ist das Ziel der Codierung die Erhaltung des „Wohlklang“, d.h. ein Klang, welcher ohne genaue Kenntnis des Inputs durch den Hörer als „natürlich“ oder zumindest „plausibel natürlich“ bewertet wird.

Zur kostengünstigen und schnellen Bewertung von Audioqualität sind Computermodelle erwünscht. Die letzte Instanz bleibt aber immer die Bewertung durch Menschen in Hörtests. Je nach Anwendungsgebiet erfolgt in Hörtests ein Vergleich mit einem explizit dargebotenen Referenzsignal („Input“) oder einer internen Referenz. Diese interne Referenz kann eine Vorstellung des gewünschten Signals sein („so sollte der Motor eines Autos der Marke xy klingen“, bzw. „so klingt eine männliche Stimme“) aber auch die Erinnerung an Eigenschaften des Audiosignals („so klingt die Stimme meiner Mutter“).

Hörtest Methoden zur Evaluation der Sprachqualität

Verfahren zur Bewertung der Sprachqualität werden traditionell durch die ITU-T („International Telecommunication Union – Telecommunication sector“) in den Empfehlungen ITU-T P.8xx standardisiert. Die Kernanwendung ist Telefonie, was eine Signalbandbreite von 300Hz bis 3.4 kHz bedeutet. In diesem Qualitätsbereich werden in der ITU-R Hörtests mit vielen naiven Hörern („max. ein Hörtest im Jahr“) durchgeführt. Das unverarbeitete Inputsignal ist den Hörern dabei nicht bekannt. In letzter Zeit werden auch in der ITU-T höhere Audiobandbreiten verwendet (bis 7 kHz wideband, bis 15 kHz ultra-wideband, bis 20 kHz fullband). Bei diesen Bandbreiten werden statt naive Hörer erfahrene, trainierte Hörer („experts“) eingesetzt.

Im Bereich Hörtests enthält die Empfehlung ITU-T P.800 [1] eine Sammlung der wichtigsten Methoden. An dieser Stelle seien hier nur die auch heute noch am häufigsten verwendeten Verfahren ACR, DCR und CCR kurz erläutert:

Annex B der P.800 beschreibt das Verfahren „Absolute Category Rating“ (**ACR**). Eine Serie von Sprach-Stimuli wird den Hörern in randomisierter Folge dargeboten. Die Folge der Stimuli enthält dabei sowohl die eigentlich zu bewertenden verarbeiteten („kodierten“) Sprachsignale als auch mittels Standardmethoden erzeugte Anker (siehe Empfehlung P.810: „Modulated Noise Reference Unit“ (MNRU)[2]. Nach jedem Stimulus erfolgt die Bewertung mittels der 5-stufigen Qualitätsskala (Tabelle 1, linke Seite).

Tabelle 1: Empfehlung ITU-T P.800**:**
5-stufige Qualitätsskala (quality scale, linke Seite) und
5-stufige Impairment-Skala (rechte Seite)

1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

1	Very annoying
2	Annoying
3	Slightly annoying
4	Audible, but not annoying
5	inaudible

Annex C der P.800 beschreibt das Verfahren „Degradation Category Rating“ (**DCR**). Um Unterschied zu ACR wird hier dem Hörer vor jedem der zu bewertenden Stimuli das unverarbeitete Referenzsignal dargeboten. Die Bewertung

erfolgt mittels der 5-stufigen Impairment-Skala (Tabelle 1, rechte Seite).

Annex D der P.800 beschreibt das Verfahren „Comparison Category Rating (**CCR**). Ähnlich dem Verfahren DCR werden dem Hörer hier zwei Stimuli angeboten. Im Unterschied zu DCR ist hier aber die Reihenfolge der zwei Stimuli randomisiert.

Da dem Hörer nicht bekannt ist welches der beiden Stimuli die Referenz ist, wird hier als Bewertungsskala die 7-stufige Vergleichsskala (siehe Tabelle 2) verwendet.

Tabelle 2: Empfehlung ITU-T P.800:
7-stufige Vergleichsskala (comparison scale)

-3	-2	-1	0	1	2	3
Much worse	Worse	Slightly worse	Same	Slightly better	Better	Much better

Hörtestmethoden zur Evaluation der Audioqualität

Seit Beginn der Forschung zur Speicherung und Übertragung von beliebigen Audiosignalen war es das Ziel eine von der Natur nicht unterscheidbare Audioqualität zu erreichen. Die Compact Disk, mit einer Auflösung von 16 bit und 20 kHz Audiobandbreite, wurde Anfang der 80er als das bestmögliche Format betrachtet. Die dabei verwendete Datenrate beträgt 1,4 Mbit/s. Mitte der 80er begannen Arbeiten, die gleiche Audioqualität auch über Kanäle mit niedrigerer Kapazität zu übertragen. Ein wichtiges Stichwort war dabei der Begriff „transparent quality“: Kein Hörer soll bei keinem Audiosignal in keiner Umgebung einen Unterschied zwischen Original und dekodierten Signal wahrnehmen können. Bei genauerer Betrachtung stellt sich aber heraus, dass Transparenz nicht wirklich beweisbar ist, denn dafür müsste jedes beliebige Audiostück mit jedem Menschen getestet werden. Im Umkehrschluss ist es allerdings sehr einfach, die „Nicht-Transparenz“ zu beweisen: Es genügt eine Testperson, die bei einem Testsignal zuverlässig Unterschiede wahrnimmt. Alle (Hörtest-) Designs haben daher gemein: Die Originale sind mit mindestens 20 kHz Audiobandbreite aufgenommen, als Hörer werden Personen mit überdurchschnittlichem Hörvermögen („goldene Ohren“) verwendet und es wird ausführlich nach möglichst schwierig zu codierendes Audiomaterial gesucht. Die Anforderungen an die Hörumgebung (bzgl. Hintergrundgeräuschen, Raumakustik sowie an die Wiedergabeeinrichtungen aka Verstärker, Lautsprecher, Kopfhörer) sind in der Regel standardisiert. Bei allen Testmethoden im Audiobereich ist außerdem festgeschrieben, dass Hörer trainiert werden damit sie ihre maximal mögliche Fähigkeit zur Erkennung kleinster Unterschiede erreichen. In der Regel sind in die Tests auch Methoden eingebaut, welche die Überprüfung der Expertise jeder Testperson ermöglichen.

Eine besonders präzise Hörtestmethode ist der **Paartest** [3]: Jeder Testperson werden für jeden zu bewertenden Stimulus zehn zufällige Paare bestehend aus Referenz R und Coder C dargeboten. Es gibt also die vier Möglichkeiten RR, RC, CR oder CC. Aufgabe der Testperson ist es jeweils zu entscheiden, ob sie einen Unterschied zwischen den Stimuli eines Paares wahrnimmt („gleich oder ungleich“). Die 95%-Konfidenz wird bei mindestenst acht korrekten Antworten erreicht. Die Methode ermöglicht eine Auswertung für jeden einzelnen Hörer. Sie ist allerdings sehr zeitraubend und damit teuer. Bei Verwendung einer großen Anzahl von Hörern steigt die Wahrscheinlichkeit der fälschlichen Ablehnung der Hypothese „Übertragung ist transparent“ „false positive“. Im Bereich der „kleinen wahrnehmbaren Unterschiede“ ist mit dieser Methode keine Aussage bezüglich der relativen Qualität verschiedener Audiocoder möglich.

Der **AB-X** Test versucht den DCR Test aus ITU-T für kleine Unterschiede empfindlicher zu machen [4]: Dem Hörer werden drei Stimuli dargeboten. Der erste Stimulus (A) ist immer die Referenz, der zweite (B) immer das zu bewertende Signal und der dritte (X) eine zufällige Auswahl aus A und B. Aufgabe der Testperson ist die Bewertung von X (Impairment-Skala) unter Kenntnis von A und B. Vorteil des Verfahrens ist, dass durch mehrmaliges Hören von A und B die Unterschiede gelernt (trainiert) werden können. Ergebnis des Tests ist ein Zahlenwert für jeden Stimulus (z.B. Mittelwert der Bewertung aller Hörer). Eine Bewertung der Expertise eines Hörer ist möglich durch die Auswertung wie oft der Hörer die versteckte Referenz (X=A) abwertet (Falschurteil). Nachteil von AB-X ist, dass jede Testperson AB-A und AB-B bewerten muss, aber nur eine dieser beiden Bewertungen Informationen über B trägt.

Das in der Empfehlung ITU-R BS.1116 [5] spezifizierte Verfahren **„tripple stimulus with hidden reference“** vermeidet diesen Nachteil. Hierbei werden drei Stimuli R-A-B dargeboten. R ist immer die offene Referenz. A-B sind eine zufällige Sequenz aus der Referenz und dem zu bewertenden Signal. Aufgabe der Testperson ist zu entscheiden, ob A oder B sich von R unterscheidet und diesem Stimulus eine Bewertung nach der Impairment-Skala (mit einer Nachkommastelle, d.h. insgesamt 41 Stufen) zu geben. Der nicht ausgewählte Stimulus erhält automatisch die Bewertung 5,0. Üblicherweise wird zur Vermeidung von Sequenzeffekten jeder Testperson RRC und RCR angeboten, C wird somit von jeder Testperson zweimal bewertet. Zur Auswertung kann einerseits der Prozentsatz der richtigen Entscheidungen verwendet werden. Häufiger wird aber die Differenz der Bewertungen (differential grade), welche für R und C eines Stimulus gegeben werden, verwendet. Zur Prüfung der Expertise jeder Testperson werden mittels t-Test die Rate und Stärke der Urteile bewertet. Daten von unzuverlässigen Hörern werden aus der weiteren statistischen Auswertung ausgeschlossen. Der t-Test zur Überprüfung der Expertise von Hörern wird nur mit Stimuli mit kleinen Störungen (Mittelwert des „differential grade“ C-R >-2,5) durchgeführt. Gegebenenfalls ändert sich hierbei die Auswahl der Stimuli durch den Ausschluss der Ergebnisse von Hörern. Ist dies der Fall, wird die Überprüfung der Expertise der verbleibenden Hörer mit der neuen Auswahl wiederholt. Bei hohen Audioqualitäten ist der BS.1116 Test sehr zuverlässig. Er wurde daher über lange Zeit als alleiniger Test im Bereich der

Standardisierung von Audiocodern (MPEG) bzw. der Spezifikation von Rundfunksystemen (DAB, DVB) verwendet. Bei hörbaren Störungen sinkt die Einigkeit der Hörer bezüglich der absoluten Werte und auch bezüglich der Rangliste verschiedener Codierverfahren.

Eine Hörtestprozedur, welche in diesem Qualitätsbereich zuverlässigere Ergebnisse erreicht, ist in der Empfehlung ITU-R BS.1534 **„multi-stimulus with hidden reference and anchors“** (MUSHRA) beschrieben [6]. Bei dieser Methode werden alle zu bewertenden Coder parallel verglichen und bewertet.

Messverfahren zur Evaluation der Audioqualität

Hörtests zur Evaluation der Audioqualität sind teuer und zeitraubend. In den 90ern wurde daher in der ITU-R an einem Verfahren zur Bewertung der wahrgenommenen Audioqualität mittels eines Computermodels gearbeitet. Wissenschaftler aus acht Nationen und mehr als 12 Forschungslabore arbeiteten zusammen an einen gemeinsamen Standard. Ziel war eine Modellierung von Hörtests nach BS.1116 für Mono- und Stereo-Signale. Ergebnis ist die Empfehlung ITU-R BS.1387 „Perceptual Evaluation of Audioquality“ (**PEAQ**), welche 1998 verabschiedet wurde [7]. Zur Entwicklung des Verfahrens wurden die Hörtestergebnisse einer Vielzahl von Hörtests, welche durch ITU, MPEG und andere Organisationen durchgeführt wurden, verwendet. Um 1998 war die nötige Rechenleistung ein wesentlicher Gesichtspunkt. PEAQ ist daher in zwei Varianten standardisiert: Die „Basic Version“ basiert auf einer FFT zur Frequenzanalyse. Die „Advanced Version“ nutzt zusätzlich eine adaptive Filterbank mit höherer zeitlicher Auflösung. Abbildung 1 zeigt ein Blockschaltbild beider Verfahren.

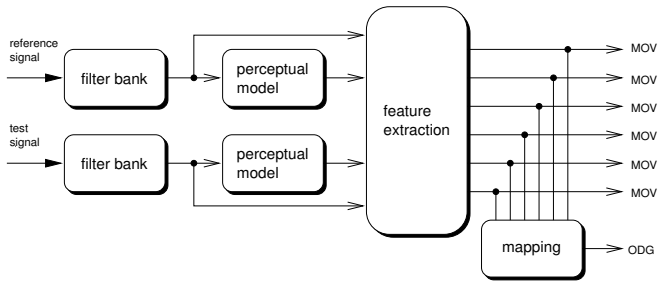

Abbildung 1: Blockschaltbild von ITU-R BS.1387 (PEAQ)[8]

Referenz und zu bewertendes Signal werden mittels einer Filterbank und eines Gehörmodells in eine interne Repräsentation überführt. Aus dem Vergleich der internen Repräsentationen werden Features, die sogenannten „model output variables (MOV), erzeugt. Ein künstliches neuronales Netz erzeugt aus den MOVs den Einzahlqualitätswert ODG („objective differential grade“), welcher eine Simulation des Hörtestergebnisses SDG („subjective differential grade“) darstellt. Grob lassen sich die MOVs in die Kategorien „Veränderung der Modulation (Rauheit)“, „Störungslautheit“, „Häufigkeit von Störungen“, „mittlerer Abstand von Störung zur Maskierungsschwelle“, „Wahrscheinlichkeit der Wahrnehmung von Unterschieden“ und „harmonische Struktur von Störungen“ einteilen. In der Basic Version werden elf verschiedene MOVs verwendet, in der Advanced Version nur fünf. Diese geringere Anzahl ergibt sich aus der besseren Güte der einzelnen MOVs, bedingt durch die bessere Zeitauflösung.

Im Zuge des Standardisierungsprozesses war eine Reihe von Aspekten zu beachten:

- Training und Verifikation von Algorithmen müssen auf unterschiedlichen Daten beruhen. Zur Selektion der Teilalgorithmen und zur Verifikation des Standards mussten Hörtestergebnisse unbekannter neuer Hörtests vorhergesagt werden.
- Zum Training standen mehr als 600 Items aus Hörtests zur Verfügung (DB1). Da diese Items aus unterschiedlichen Tests stammten, waren dabei die verwendeten Abhörpegel unterschiedlich bzw. unbekannt. Die Daten enthielten Ergebnisse von Hörtests mit Lautsprecher und Kopfhörern. Bei Kopfhörertests war es innerhalb einer Testsession teilweise möglich, die Abhörlautstärke zu ändern. Zu Beginn der Arbeiten war unbekannt, in wie weit die Ergebnisse dieser verschiedenen Hörtests von ihren absoluten Werten zusammenpassen.
- Zur Selektion und Verifikation wurden neue Stücke von neutralen Laboren ausgewählt und im Hörtest bewertet (DB2 und DB3). Zur Produktion von DB2 und DB3 wurde der Abhörpegel in allen Laboren gleich eingestellt. In DB2 und DB3 wurden größere Standardabweichungen als in DB1 beobachtet. Es wurde vermutet – aber nie verifiziert –, dass Hörer, welche ihren individuell optimalen Abhörpegel für jedes Audiostück wählen, besser hören und damit einheitlichere Bewertungen abgeben.
- Ein wichtiger Gesichtspunkt sind die Qualitätskriterien – Was ist eine gute Vorhersage eines Hörtests?

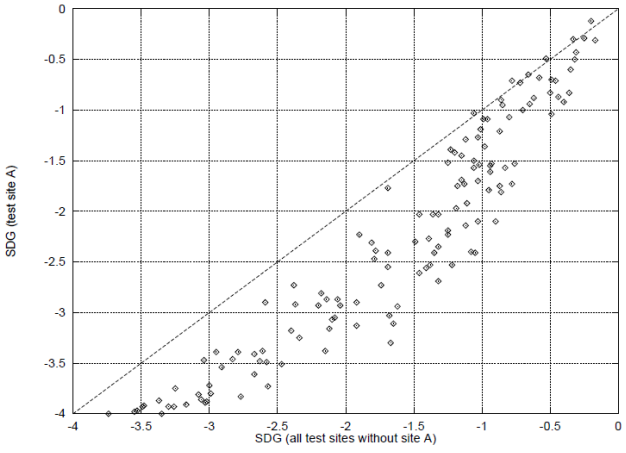

Abbildung 2: Vergleich der Mittelwerte eines Testlabors (15 Hörer) mit den Ergebnissen mehrerer Testlabors (64 Hörer) [9].

Die Korrelation zwischen SDG und ODG gibt Hinweise, ob das Ranking von Hörtest und Computermodell gleich ist. Wie oben erwähnt, ergibt sich bei BS.1116 nur für die hohen Audioqualitäten eine große Einigkeit der Testpersonen. Abbildung 2 zeigt die Gegenüberstellung der Mittelwerte eines Testlabors (site A, 15 Hörer) mit den Mittelwerten

mehrerer Testlabors (64 Hörer). Bei Werten >-1 ist eine gute Übereinstimmung zwischen den Hörergruppen feststellbar. Bei schlechterer Qualität ergibt sich ein Offset zwischen den Gruppen und eine insgesamt größere Streuung. Zur Bewertung von Messverfahren ist dieses Verhalten bei Training und Verifikation zu berücksichtigen. Abbildung 3 zeigt das Toleranzschema zur Bewertung der Abweichungen. Zur Vermeidung unrealistisch kleiner Toleranzen bei besonders guten Items wurde vor Berechnung des Toleranzschemas das Konfidenzintervall (CI) auf den Minimalwert von 0.25 limitiert.

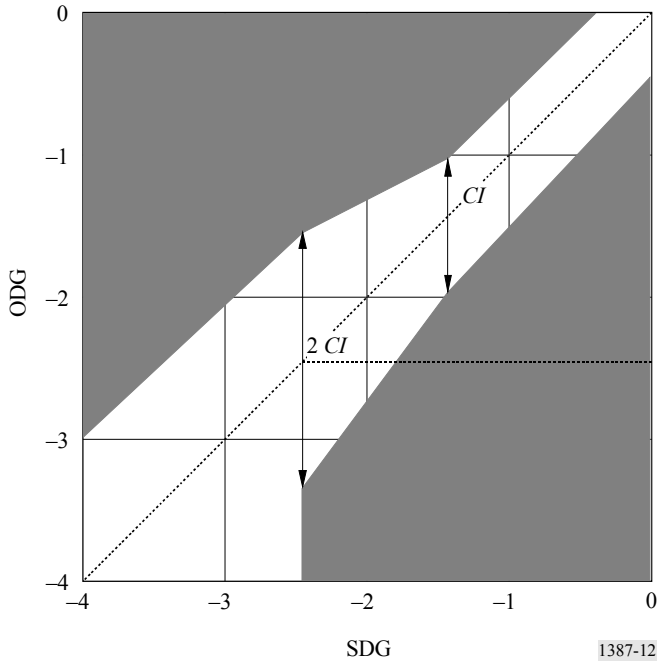

Abbildung 3: Kriterium zur Bewertung der Güte der Vorhersagen durch PEAQ. CI ist das 95% Konfidenzinterval der Hörtests, SDG (subjective differential grade) sind die Hörtestergebnisse, ODG (objective differential grade) die Ergebnisse der Modellierung [7].

Neuere Arbeiten an Messverfahren

Zum Zeitpunkt der Standardisierung von PEAQ war das Ziel „Transparenz“ und das einzige verbreitete Hörtestverfahren ITU-R BS.1116. Mit der Verbreitung von Audiocodecs in mobilen Geräten erwuchs der Wunsch nach „plausibler“ Qualität, was zum neuen Hörteststandard MUSHRA und zu neuen Werkzeugen in den Audiocodern führte. Beispiele für solche neuen, im Allgemeinen nicht mehr wellenformerhaltende Werkzeuge, sind „Temporal Noise Shaping (TNS)“, „Perceptual Noise Substitution PNS“ und verschiedene Methoden der Stereocodierung. Erste Tests im Jahr 1998 zeigten schon kurz nach Finalisierung von PEAQ, dass z.B. TNS durch die „Basic Version“ komplett falsch bewertet wird, während die „Advanced Version“ hier noch korrekte Ergebnisse erreichte. Generell ist allerdings bei allen nach 1998 entstandenen Codecs die Messung durch PEAQ noch nicht verifiziert.

Im Bereich der höheren Qualitäten verbreiteten sich in den Anwendungen auch Mehrkanalverfahren für welche es noch kein standardisiertes Messverfahren gibt.

Die ITU-R arbeitete an einer Erweiterung von PEAQ für MUSHRA, neue Codecs und Multikanal. Hierzu wurde die Umrechnung von beliebigen Lautsprecherpositionen mittels HRTFs oder BRIRs auf eine binaurale Darstellung sowie die Ergänzung durch weitere MOVs zur Bewertung von räumlichen Eigenschaften untersucht. Im Zuge der Arbeiten stellte sich ein unerwartetes Problem: nur wenige Datenbanken mit Hörtestergebnissen neuerer Codecs waren verfügbar und diese waren so unterschiedlich, dass für jede dieser Datenbanken ein eigenes Modell nötig gewesen wäre. Letzteres deutete auf Schwächen im Design der Hörtestverfahren hin. Revisionen der Empfehlungen BS.1116 (2015) und BS.1534 (2014) sollen diese Probleme lösen. Anders als in den 90ern gab es nur wenige Beteiligte, die an den Modellen für Messverfahren arbeiteten und sehr wenig Unterstützung durch neutrale Labore zur Erzeugung von Datenbanken für Training, Selektion und Verifikation. Die Arbeiten in der ITU-R sind daher momentan wieder eingestellt.

Zusammenfassung und Ausblick

Zur Bewertung Audiosignalen gibt es eine Reihe von etablierten Hörteststandards. Für Mono- und Stereo-Signale gibt es in der Empfehlung ITU-R BS.1387 ein Messverfahren, welches für Audiocodecs (vor 1998) zuverlässige Ergebnisse erzielt. Im Bereich Hörtests für Multikanal ist noch zu prüfen ob die Revisionen der Hörteststandards die erwartete Verbesserung in der Vergleichbarkeit von Tests bringen. Für neuere Audiocodecs sowie für 3D- Audio sind noch keine verifizierten Messverfahren bekannt.

Literatur

[1] ITU-T: Rec. P.800 Methods for subjective determination of transmission quality, 1996

[2] ITU-T: Rec. P.810 : Modulated noise reference unit (MNRU), 1996

[3] Brandenburg, Kh.: Ein Beitrag zu den Verfahren und der Qualitätsbeurteilung für hochwertige Musikcodierung, Erlangen, Nürnberg, Univ., Diss., 1989

[4] ITU-R: Rec. BS.1284: General methods for the subjective assessment of sound quality, (1978)/1997/2003

[5] ITU-R: Rec. BS.1116 : Methods for the subjective assessment of small impairments in audio systems, 1994/2015

[6] ITU-R: Rec. BS.1534: Method for the subjective assessment of intermediate quality levels of coding systems. 2001/2014

[7] ITU-R: Rec. BS.1387: Method for objective measurements of perceived audio quality, 1998/2001

[8] Thiede, T et al: PEAQ - The ITU standard for objective measurement of perceived audio quality. JAES 48 (2000), Nr.1/2, S.3-29

[9] Sporer, T: Evaluating Small Impairments with the Mean Opinion Scale – Reliable or just a Guess? 101st AES Convention 1996, Preprint #4396