Modellierung der wahrgenommenen Audio- und Sprachqualität

Thomas Sporer[1], Judith Liebetrau[2]

[1] Fraunhofer Institut für Digitale Medientechnologie IDMT, 98693 Ilmenau, Deutschland, Email: spo@idmt.fhg.de

[2] Fraunhofer Institut für Digitale Medientechnologie IDMT, 98693 Ilmenau, Deutschland, Email: ltu@idmt.fhg.de

Einleitung

In technischen Systemen, welche durch Menschen genutzt werden, ist wahrgenommene Qualität ein wesentlicher Faktor. Beispiele hierfür sind Verfahren zur effizienten Übertragung von Sprach- und Audiosignalen (Sprach- bzw. Audiocodierung), zur Verbesserung der Verständlichkeit von Sprache und zur Klangverbesserung von Audio, zur Zerlegung von Sprach- und Audiosignalen (Quellentrennung) sowie Verfahren zur Formatanpassung von räumlichen Audiosignalen (Up- bzw. Down-Mix). Die letzte Instanz ob ein Verfahren für die jeweilige Anwendung geeignet ist, ist immer das menschliche Gehör. Hörtests sind allerdings in der Regel teuer, zeitraubend und daher nicht immer durchführbar. Computersimulationen von Hörtests, im folgenden Messverfahren genannt, sollen daher die Anzahl der nötigen Hörtests reduzieren. Betrachtet man die Einflußfaktoren auf die wahrgenommene Qualität, dann sieht man, dass es eine Reihe von Faktoren gibt, welche bei Anwendungen für Sprache und Audio gleichermaßen auftreten (Störungen sowie Bandbreite bzw. Klangfarbe), und Faktoren, die nur für Sprache (Sprachverständlichkeit, Höranstrengung,...) oder nur für Audio (Lokalisation, Tiefenstaffelung, Einhüllung,...) wichtig sind.

Wahrnehmung von Qualität

Bei der Bewertung der Qualität von Sprach- und Audiosignalen ist ein wesentlicher erster Schritt die prinzipielle Leistungsfähigkeit des Gehörs. Im beschränkten Maße ist diese Leistungsfähigkeit trainierbar. Auf der anderen Seite ist Qualität immer mit einer Erwartungshaltung verknüpft: „Wie soll es klingen?“. Bei Sprache ist hier z.B. die Vorkenntnis „wie klingt diese Sprache“, „wie klingt ein Mann/eine Frau“, „wie klingt eine bestimmte Person“. Bei Audiosignalen ist dies die Kenntnis über den prinzipiellen Klang eines Instruments, eines Musikstücks, eines bestimmten Künstlers usw. Auch bei Kenntnis dieses „Referenzklanges“ist aber ein weiterer wichtiger Schritt, dass sich der Hörer die Unterschiede bewußtmachen muss. Hat er dies getan, folgt als dritter Schritt die Gewichtung der wahrgenommenen Abweichungen: Ist die Abweichung im Anwendungskontext wichtig? Im Beispiel „Freisprechen im Auto“ mag eine Klangfarbenveränderung deutlich hörbar sein, aber durch die dadurch erzeugte Verbesserung der Sprachverständlichkeit mehr als aufgehoben sein. Die Gewichtung wahrgenommener Unterschiede ist daher abhängig von individuellen Vorlieben, der Hörsituation und der Anwendung.

Hörtests

In den Anwendungsbereichen Sprach- und Audiocodierung ist oft ein Vergleich verschiedener Verfahren und Geräte nötig. Zur Sicherstellung der Vergleichbarkeit sind standardisierte Hörtests nötig. Derzeit sind Hörtests nach Empfehlungen der International Telecommunication Union (ITU) verbreitet. Hörtests für Sprache finden sich in den Empfehlungen ITU-T P.8xx [1]. Hörtests für Audio finden sich vor allem in den Empfehlungen ITU-R BS.1116 und BS.1534 MUSHRA [2]. Hörtestverfahren wie z.B. ITU-T P.800 Annex B „Absolute Category Rating (ACR)“ haben nur eine interne Referenz: Der Hörer hat keinerlei Kenntnis über das „Original“ vor der Verarbeitung. Andere Verfahren, wie z.B. ITU-R P.800 Annex D „Degradation Category Rating (DCR)“ und ITU-R BS.1116 bieten dem Hörer eine offene Referenz mit der das zu bewertende Signal verglichen werden soll. Die Hörtests der ITU-T decken mit unterschiedlichen Verfahren einen breiten Bereich der Sprachbandbreite von Sprachqualität (3,5kHz), Wideband (7,0kHz), Ultrawideband (15kHz) bis Fullband (20kHz) ab. In der ITU-R ist die Audiobandbreite in der Regel mindestens 20kHz. Die dort heute am häufigsten verwendeten Verfahren ITU-R BS.1116 „tripple stiumulus with hidden reference“ und ITU-R BS.1534 „multiple stimulus with hidden reference and anchors (MUSHRA)“ unterscheiden sich imbezüglich des zu evaluierenden Qualitätsbereichs: BS.1116 wurde entwickelt, um kleinste Unterschiede zur Referenz zu evaluieren. BS.1534 hat als Zielbereich die moderaten Qualitäten. Während beim Ersteren Unterschiede oft nicht oder nur mit großer Mühe von wenigen Hörern wahrnehmbar sind, ist bei MUSHRA die Schwierigkeit für die Hörer auch deutlich hörbare Unterschiede zu gewichten.

Messverfahren

Ähnlich der Standards für Hörtests wurden die Standards für Messverfahren zunächst für die Bereiche Sprach- und Audiocodierung entwickelt. In der Regel gibt es eine enge Beziehung zwischen den Messverfahren und dem von diesem Verfahren simulierten Hörtest. In der Regel wird eine offene Referenz vorausgesetzt, mit der das Messverfahren das zu bewertende Signal vergleicht. Zur Erleichterung des Vergleichs dieser Eingangssignale wird oft eine zeitliche Synchronität vorausgesetzt. Eventuelle Pegelunterschiede der Eingangssignale werden vor der Messung ausgeglichen.

Prinzipieller Aufbau

Der grundsätzliche Aufbau von Messverfahren folgt dem Hörprozess der wahrgenommenen Qualität. Referenz

und zu bewertendes Signal werden jeweils mittels eines psychoakustischen Modells in eine interne Darstellung transformiert. Diese Darstellung versucht die Information nachzubilden, welche dem menschlichen Gehirn zum Vergleich zur Verfügung steht. Aus dem Vergleich der internen Darstellungen werden im Folgenden Qualitätskenngrößen berechnet, welche unterschiedliche Aspekte des Unterschiedes messen. Beispiele hierfür sind z.B. lineare Verzerrungen (Klangfarbenunterschiede), Wahrscheinlichkeit der Hörbarkeit des Unterschiedes, Lautheit des Unterschiedes. In einem letzten Schritt werden diese Parameter mit einer mehrdimensionalen Gewichtungsfunktion zusammengefasst um einen Einzahlqualitätswert entsprechend dem Hörtestergebnis zu berechnen. Diese Gewichtungsfunktion ist im einfachsten Fall eine lineare Abbildung, kann aber durchaus nichtlinear bis hin zum neuronalen Netz sein.

Der erste Schritt, die psychoakustische Modellierung, ist dabei der einfachste, da hier auf 150 Jahre Forschung zurückgegriffen werden kann. Die hierbei verwendeten psychoakustische Modelle gehen von trainierten Hörern aus, d.h. gehen an die Grenze des durch Menschen eben noch Wahrnehmbaren. Der zweite Schritt, der Vergleich der internen Darstellungen und die Berechnung von Qualitätskenngrößen ist etwas schwieriger: Es gibt eine große Anzahl möglicher Kenngrößen, und nur wenige davon lassen sich mittels Probandentests verifizieren. Die besondere Schwierigkeit hierbei ist insbesondere, dass einige der Eigenschaften von Sprach- und Audiodateien sich nicht in reiner Form erzeugen lassen, und Probanden bereits hier ein bestimmte Eigenschaft mit unterschiedlichen Worten beschreiben. Der dritte Schritt ist der schwierigste: In Abhängigkeit von der Applikation ist die Gewichtung unterschiedlicher Qualitätskenngrößen unterschiedlich. Ein Messverfahren muss daher an die jeweilige Applikation angepasst werden. Dieses Training adaptiert das Gesamtsystem, d.h. viele Skalierungsfehler der Schritte 1 und 2 werden automatisch ausgeglichen. Wie im folgenden noch gezeigt werden wird, ist es aber nicht möglich Defizite in der zeitlichen Auflösung des ersten Schrittes bzw. fehlende Qualitätskenngrößen des zweiten Schrittes vollständig zu korrigieren.

Training

Ein wesentliches Element der Anpassung eines Messverfahrens an die Anwendung ist die Auswahl des zu modellierenden Hörtestverfahrens sowie geeigneter Hörtestdaten als „Ground-Truth“. Um „Over-Fitting“ zu vermeiden, ist es wichtig, ausreichend viele Hörtestdaten zu nutzen. In der Regel werden diese Tests mit unterschiedlichen Rohdaten (Sprach- bzw. Audiodaten), unterschiedlichen Bedingungen (z.B. verschiedene Codecs, Bitraten, Bandbreiten) und an verschiedenen Orten durchgeführt. Ein wesentlicher erster Schritt ist die Sichtung der Güte der Trainingsdaten: wurden die Tests standardkonform durchgeführt? Fragen dabei sind z.B.: Wurde die Anforderungen an den Raum und Geräte erfüllt? Wurden Rohdaten korrekt produziert bzw. ausgewählt? Wurde eine ausreichende Anzahl Hörer verwendet? Wurde die Zuverlässigkeit der Hörer untersucht und ggf. die Daten unzuverlässiger Hörer aus der statistischen Analyse entfernt? Gibt es signifikante Unterschiede zwischen den Bewertungen von Stimuli, die in mehreren Hörtests verwendet wurden? Gerade dieser letzte Punkt kann ein Indiz sein, dass einer der möglichen Hörtests sich nicht für das Trainieren eignet.

Ein zweiter Schritt ist die Definition der Gütefunktion: Was ist ein gutes Messverfahren? Einfache Beispiele für gebräuchliche Gütefunktionen sind die Korrelation zwischen Hörtestdaten und Messergebnissen, der mittlere quadratische Fehler und der maximale Fehler. Besser sind allerdings Gütefunktionen, die den Anwendungsbereich berücksichtigen bzw. welche, die auch die Zuverlässigkeit der Hörtestergebnisse mit einbeziehen: Wenn sich die Hörer nicht einig sind, dann muss ein Messverfahren auch nicht genau den Mittelwert der Hörer voraussagen.

Im eigentlichen Trainingsprozess ist es ratsam, nicht alle Trainingsdaten zum Abgleich der freien Parameter der mehrdimensionalen Gewichtungsfunktion zu verwenden: Man sollte einen Teil der Hörtestdaten zum Training verwenden und dann überprüfen, ob das Messverfahren mit der so optimierten Gewichtungsfunktion den anderen Teil der Hörtestdateien vorhersagen kann (Generalisierung). Durch Vertauschen von Trainings- und Überprüfungsdaten und Vergleich der so berechneten freien Parameter der Gewichtungsfunktion kann auch die Robustheit überprüft werden. Bei der Aufteilung der Hörtestdateien sind verschiedene Methoden üblich: Aufteilung entlang der Grenzen unterschiedlicher Hörtests, oder Aufteilung der Daten jedes Hörtests in Trainings- und Überprüfungsdaten. In jedem Fall ist es wichtig, dass in beiden Teilmengen alle Arten von Störungen, alle Arten von Rohdaten und alle Qualitätsstufen vorhanden sind.

Ergebnis des Trainings ist ein Messverfahren, welches in der Lage ist, die Ergebnisse aller Hörtests, die zum Training verwendet wurden, zu erklären. Ob das Messverfahren auch in der Lage ist, neue, unbekannte Hörtests vorherzusagen, ist nach dem Training noch nicht sicher.

Verifikation

In der Standardisierung wird daher eine komplett neue Datenbasis zur Verifikation verwendet. In der Regel wird durch ein von den Entwicklern der Messverfahren unabhängiges Team ein neuer Hörtest zusammengestellt und durchgeführt. Vor der Bekanntgabe der Hörtestergebnisse werden Messergebnisse erstellt und hinterlegt. Diese Messergebnisse werden dann mit den Hörtestergebnissen verglichen. Insbesondere wenn mehrere Messverfahren verglichen werden sollen, ist die Definition der zu verwendenden Gütefunktion vor der Durchführung dieses Vergleichs wichtig.

Ergebnis der Verifikation ist ein Messverfahren, welches in der Lage ist, die Ergebnisse von Hörtests im verifizierten Bereich vorherzusagen. Ob das Messverfahren auch in der Lage ist, in anderen Anwendungsgebieten oder mit andern Störungsklassen sinnvolle Ergebnisse zu liefern,

ist auch nach der Verifikation nicht sicher.

Die folgenden Ausführungen am Beispiel des in der Empfehlung ITU-R BS.1387 beschriebenen Messverfahrens gelten in sehr ähnlicher Weise auch für die Messverfahren für Sprache, welche in den Empfehlungen ITU-T P.862 „Perceptual evaluation of speech quality“ (PESQ) und P.863 „Perceptual objective listening quality assessment“ (POLQA) beschrieben sind.

Beispiel: ITU-R BS.1387 PEAQ

Die Empfehlung ITU-R BS.1387 PEAQ [3] wurde entwickelt, um Hörtests gemäß der Empfehlung nach ITU-R BS.1116 vorherzusagen. Der Qualitätsbereich für BS.1116 ist die Evaluation von Audiocodierungsverfahren bei höchster Qualität, d.h. in einem Bereich, in dem nur mit Mühe hörbare Störungen wahrgenommen werden können. PEAQ wurde in den Jahren 1994 bis 1998 entwickelt. Zu dieser Zeit war die Rechenleistung ein wichtiger Faktor. Die Empfehlung enthält daher zwei verschiedene Messverfahren: die weniger rechenintensive „Basic Version“ und die genauere „Advanced Version“. Abbildung 1 zeigt das allgemeine Blockschaltbild, welches für beide Versionen gültig ist: Das Original („Referenz“) und das Test-Signal („zu bewertendes Signal“) werden mittels einer Filterbank in den Frequenbereich transformiert und mittels eines Gehörmodells in eine interne Darstellung gebracht. Aus der internen Darstellung sowie den Frequenzbereichsdaten werden die sogenannten „model output values“ (MOV) berechnet und diese mittels eines neuronalen Netzes zum „objective differential grade“ (ODG) abgebildet.

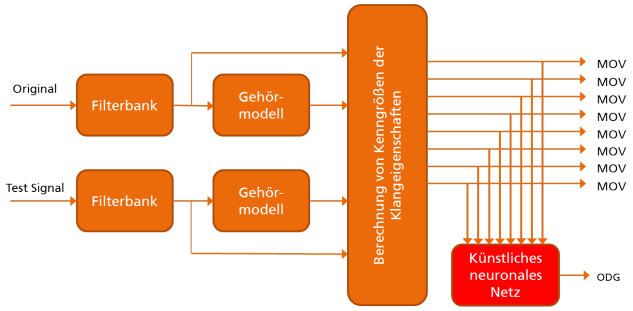

Abbildung 1: Blockschaltbild von ITU-R BS.1384 PEAQ.

Ein wichtiger Unterschiede zwischen der Basic Version und der Advanced Version ist die verwendete Filterbank: Die Basic Version nutzt eine FFT, die Advanced Version zusätzlich eine adaptive Filterbank [3]. Die Kombination der spektral höher auflösenden FFT und der zeitlich besser auflösenden adaptiven Filterbank ermöglicht eine genauere Modellierung. Dadurch benötigt die Advanced Version weniger MOVs, um alle Eigenschaften der Trainingsdaten evaluieren zu können. Bei der Entwicklung von PEAQ zeigte sich, dass Unterschiede im psychoakustischen Modell (unterschiedliche Bandaufteilung, Annahmen bzgl. Ruhehörschwellen, u.ä.) nur einen sehr geringen Einfluß auf die Genauigkeit des Modells haben. Die wesentlichen Einflußfaktoren waren die Auswahl der MOV und die Gewichtungsfunktion.

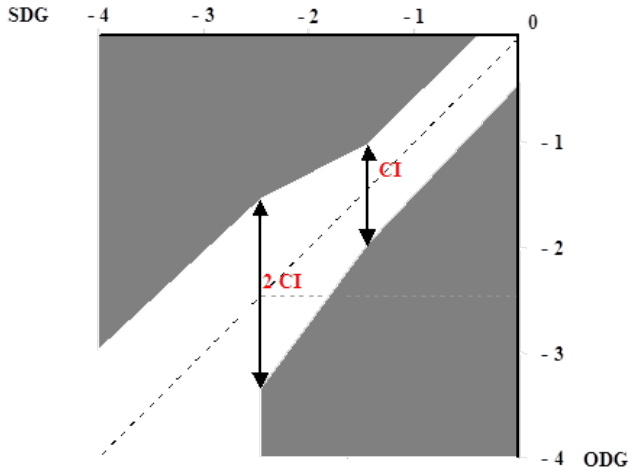

Abbildung 2: Toleranzschema für Training und Verifikation von ITU-R BS.1384 PEAQ. Subjective Differential Grade (SDG) ist das Hörtestergebnis. ODG ist das Messergebnis. Das Convidence Interval (CI) ist ein Maß für die Güte des Mittelwertes der Hörer.

Abbildung 2 zeigt das in Training und Verifikation von PEAQ verwendete Toleranzschema: Nachdem der Hörtest BS.1116 nur im höheren Qualitätsbereich verwendet werden soll, ist hier eine genauere Modellierung erforderlich. Bei größeren Störungen (SDG $< -2,5$) wird BS.1116 unzuverlässig und daher sind dort auch für das Messverfahren größere Abweichungen erlaubt. Zur Berücksichtigung der Zuverlässigkeit der Hörtestergebnisse wurde das Confidence Interval (CI) verwendet. Als numerische Umsetzung des Toleranzschemas wurde der „Absolute Error Score“ (AES) definiert. Dieser wird berechnet nach der Formel

$$\mathrm{AES} = 2 * \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{(\mathrm{ODG}_i - \mathrm{SDG}_i)}{\max[\mathrm{CI}_i; 0.25]}\right)^2} \qquad (1)$$

Nach heutigem Wissen würde man eher die Standardabweichung verwenden, da sich über die Anzahl der Hörer das CI beliebig verkleinern lässt, die Standardabweichung dagegen ein robustes Maß darstellt.

PEAQ wurde trainiert mit über 600 Hörtestergebnissen und verifiziert mit 84 Hörtestergebnissen. Alle Daten bezogen sich auf Mono und Stereo, die Abtastrate war 48 KHz (Audiobandbreite 20 kHz). In den Trainigsdaten waren sowohl Kopfhörer- als auch Lautsprechertests enthalten, die Verifikationsdatenbasis bestand ausschliesslich aus Lautsprechertests. Zur Verifikation wurden die folgenden Codierverfahren bzw. Störklassen verwendet: NICAM, MiniDisc, Dolby AC2, MPEG1 Layer II/II, MPEG2 AAC, Dolby AC3, Quantization Distortion, THD, Noise. Bei den Codecs wurden unterschiedliche Bitraten, Stereomodi und auch Kaskadierungen verwendet.

Limitierungen von PEAQ

Neue Audiocodierverfahren

Bereits wärend der Standardisierung von PEAQ, aber nach Erstellung der Verifikationsdatenbank wurde die neue Codiermethode „temporal noise shaping“ (TNS) erfunden [4]. TNS verschiebt die Störung innerhalb eines Codierblockes zeitlich. Dadurch ist die zeitliche Struktur des Fehlers der zeitlichen Struktur des (Nutz-)Signals besser angepasst. Ein Seiteneffekt von TNS ist, dass die Fehlerenergie leicht ansteigt. Insbesondere bei Sprache führte dieses neue Werkzeug aber zu einer Verbesserung der wahrgenommenen Audioqualität. Auf Grund der geringeren zeitlichen Auflösung ist die Basic Version von PEAQ aber nicht in der Lage, diese Verbesserung festzustellen und bewertet das Codierergebnis fälschlicherweise als schlechter. Die Advanced Version erkennt diese Verbesserung.

1998 wurde die neue Codiermethode „perceptual noise substitution“ (PNS) beschrieben [5]: Teile des Kurzzeitspektrum eines Codierblockes, welche nur Rauschkomponenten enthalten, werden parametrisch kodiert: Statt der exakten Struktur wird nur die spektrale Einhüllende übertragen. Ein Rauschgenerator im Dekoder rekonstruiert dann die grobe Struktur dieses Bereichs. Durch diese effiziente Codierung von Rauschkomponenten wird insgesamt eine verbesserte wahrgenommene Audioqualität erreicht. PNS zerstört aber die Kurvenform des Audiosignals. Daher bewerten beide Versionen von PEAQ dies als schlechter.

Ähnliche Effekte treten z.B. auch bei den in den späteren Jahren entwickelten Audiocodiermethoden „Spatial audio object coding“ (SAOC) [6] und „universal speech audio codec“ (USAC) [7] auf.

Neue Anwendungsgebiete

Im wesentlichen wurde PEAQ nur mit Audiocodecs verifiziert. Eine wesentliche Eigenschaft von Audiocodern ist, dass die Quantisierungsfehler sich zeitlich und spektral dort konzentrieren wo Energie im Nutzsignal vorhanden ist. Korrelation von Nutz- und Störschall ist daher immer sehr hoch. Die Struktur unerwünschter Komponenten in anderen Anwendungsgebieten ist deutlich anders. Dies sei an einem Beispiel erklärt: Bei der Evaluation von Algorithmen zur Trennung von Audioobjekten ist der übliche Versuchsaufbau, dass man aus einzelnen Audioobjekten eine Mischung erstellt, der Quellentrennungsalgorithmus diese Mischung wieder zerlegt, und die ursprünglichen Objekte mit dem Ergebnis der Trennung verglichen wird. PEAQ ist weder trainiert noch verifiziert zur Bewertung von unkorrelierten Signalen und sollte dort ohne Verifikation durch zusätzliche Hörtests nicht verwendet werden.

Bei der Bewertung von Up- und Downmixing Algorithmen durch PEAQ stellen sich ähnliche Probleme: Zum einen ist teilweise unklar, was überhaupt die Referenz ist, zum anderen können neue Fehlerklassen wie Kammfiltereffekte, Fehlortungen und Änderungen der wahrgenommenen Quellengröße auftreten. Auch hier sollte PEAQ nur zusammen mit Hörtests zur Verifikation verwendet werden.

Hörtestmethoden

In vielen heutigen Studien wird als Hörtest ITU-R BS.1534 MUSHRA eingesetzt. MUSHRA ermöglicht auch die Bewertung von deutlich hörbaren, aber angenehm klingenden Unterschieden. PEAQ wurde nur im Bereich der hohen Qualitäten trainiert und verifiziert. Der Versuch der Erweiterung von PEAQ auf MUSHRA Tests scheiterte allerdings daran, dass keine ausreichende Trainingsdatenbank zur Verfügung stand, und dass es nicht gelang, eine dazu passende Verifikationsdatenbank zu schaffen.

Zusammenfassung

Standardisierte Messverfahren zur Messung der wahrgenommen Sprach- und Audioqualität bieten bei ihren verifizierten Störungsklassen und Anwendungsgebieten einem kostengünstigen Ersatz für Hörtest. Treten andere Störungsklassen auf bzw. werden andere Anwendungsgebiete untersucht, ist eine Verwendung ohne zusätzlich stattfindenden Hörtest äußerst zweifelhaft.

Literatur

[1] ITU-T Recommendations P-Series, URL: `https://www.itu.int/itu-t/recommendations/index.aspx?ser=P`

[2] ITU-R Recommendations BS-Series, URL: `https://www.itu.int/rec/R-REC-BS/en`

[3] Thiede, T.;Treurniet, W.C.; Bitto, R.; Sporer, T.; Brandenburg, Kh.; Schmidmer, C.; Keyhl, M.; Beerends, J.G.; Colomes, C.; Stoll, G.; Feiten, B.: PEAQ-der künftige ITU-Standard zur objektiven Messung der wahrgenommenen Audioqualität. Bericht der 20.Tonmeistertagung (1998), 724-766

[4] Herre, J.; Johnston, J.D.: Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS). Audio Engineering Society Convention 101 (1996).

[5] Herre, J.; Schultz D.: Extending the MPEG-4 AAC codec by perceptual noise substitution. Audio Engineering Society Convention 104. (1998).

[6] Engdegård, J. et.al.: Spatial audio object coding (SAOC)-The upcoming MPEG standard on parametric object based audio coding. Audio Engineering Society Convention 124. (2008)

[7] Neuendorf, M, et al.: The ISO/MPEG unified speech and audio coding standard—consistent high quality for all content types and at all bit rates. Journal of the Audio Engineering Society (2013), 956-977.