POLQA und VQuad-HD
Neue ITU-T Empfehlungen für instrumentelle Qualitätsvorhersage

Jens Berger

SwissQual AG, CH-4528 Zuchwil, Schweiz, E-Mail:jens.berger@swissqual.com

Einleitung

Zur Qualitätsbeurteilung von technischen Systemem zur Sprach-, Audio oder Videoübertragung werden Versuchspersonen in einer simulierten experimentellen Umgebung zu ihrer Qualitätswahrnehmung befragt. Zur Durchführung dieser sogenannten subjektiven Experimente stehen eine Vielzahl normierter Testverfahren zur Verfügung, die oft auf spezielle Anwendungsfälle abgestimmt sind (z.B. Bewertung von Codierverfahren oder Geräuschreduktionssystemen). Sie definieren Randbedingungen des Experiments, wie z.B. die Fragestellung, die Skala, aber auch die Nutzerschnittstelle, z.B. einen Telefonhandapparat. Diese Limitierungen erlauben eine Vergleichbarkeit von Ergebnissen gleichartiger Experimente auch im internationalen Rahmen.

Experimente mit Personen sind auf speziell eingerichtete Labore beschränkt, für Qualitätsaussagen in Netzen im Wirkbetrieb oder Optimierungen von Übertragungssystemen im Labor sind sie meist nicht praktikabel. Bereits seit einigen Jahrzehnten werden daher Qualitätsschätzer entwickelt, die basierend auf einer Signalanalyse Abschätzungen der wahrgenommenen Qualität vornehmen. Dabei wird das zu bewertende Signal mit einer ‚Erwartung' verglichen. Diese Erwartung kann eine modellhafte Beschreibung sein, wird aber meist aus einem Referenzsignal abgeleitet. Solche Verfahren werden als ‚full-reference' Verfahren bezeichnet.

Die Evolution solcher ‚objektiver' oder instrumenteller Verfahren ist nicht nur durch die fortschreitende Analysetechnik und detailliertere Modelle der menschlichen Wahrnehmung gekennzeichnet, sondern auch durch eine stärkere Kopplung an individuelle Experimente. Ein instrumentelles Verfahren ist das Modell eines ‚subjektiven' Testverfahrens. Es modelliert dessen experimentelle Umgebung und nutzt dessen Skala.

Die ITU-T hat zu Beginn 2011 zwei neue instrumentelle Verfahren standardisiert, zum einen P.863, auch bekannt als POLQA, zur Beurteilung der Sprachqualität in einer reinen Hörsituation und J.341, VQuad-HD, zur qualitativen Beurteilung von High Definition Video Sequenzen [1], [2].

Im folgenden wird zunächst POLQA vorgestellt. Am Ende des Beitrages wird das Sprachqualitätsverfahren mit dem Verfahren für Videobeurteilungen verglichen und es werden Gemeinsamkeiten und Unterschiede dargestellt.

Motivation für ITU-T Rec. P.863 ‚POLQA'

Bereits 2005 initiierte ITU-T den Wettbewerb für ein neues instrumentelles Sprachqualitätsschätzverfahren. Die Motivation war in erster Linie eine gegenüber dem existierenden Verfahren P.862 ‚PESQ' [3] erweiterte Audiobandbreite, die Möglichkeit der Analyse von Aufzeichnungen mit einem Ohrsimulator (Endgerätemessungen), der Berücksichtigung des akustischen Präsentationspegels und des Frequenzganges sowie die Korrektur von Unzulänglichkeiten von PESQ bei der Bewertung neuerer Codier- und Sprachverarbeitungsverfahren. Mitte 2010 fand das Selektionsverfahren statt, im Februar 2011 wurde ein kombiniertes Modell der Firmen OPTICOM (DE), SwissQual (CH) und TNO (NL) als neuer Standard ITU-T P.863 verabschiedet [1].

P.863 ‚POLQA' als Modell von Hörtests

POLQA ist das Modell eines Hörtests zur Bestimmung der Sprachqualität gemäß ITU-T P.800/P.830 [4] mit ‚Absolute Category Rating'. Für die Entwicklung und Spezifikation wurden ausschließlich derartige auditive Experimente herangezogen. Insgesamt standen 62 Experimente mit über 45'000 Sprachbeispielen zur Verfügung. POLQA unterstützt zwei verschiedene ‚operational modes', der sogenannte superwideband Modus bewertet Signale gegenüber einer Referenz mit einer Audiobandbreite 50 bis 14'000Hz. Der ‚narrowband' Modus nimmt eine Qualitätsschätzung gegenüber einer klassischen telefonband-begrenzten Referenz vor und dient auch der Kompatibilität mit existierenden Verfahren wie z.B. PESQ.

Struktur P.863 ‚POLQA'

P.863 ‚POLQA' ist ein ‚full-reference' Verfahren. Es vergleicht das zu bewertende Signal mit einem Referenzsprachsignal (dem Eingangssignal in das zu bewertende System). Nach einer Transformation auf eine interne Perzeptionsebene werden Unterschiede zwischen beiden Signalen berechnet, gewichtet und zu einem integralen Qualitätsmaß zusammengefaßt.

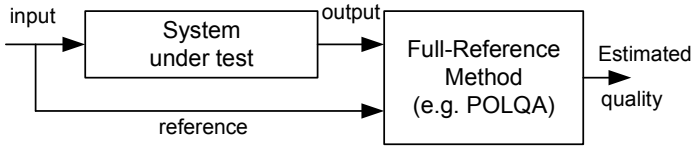

Abbildung 1: Prinzip ‚full-reference' Verfahren

Vorverarbeitung der Sprachsignale

Zuerst erfolgt eine globale Pegelangleichung, die später im Kernmodel um lokale Pegel- und Lautheitsangleichungen ergänzt wird.

POLQA vergleicht beide Signale in kurzen Abschnitten (32ms). Deshalb ist zunächst eine Synchronsation beider Signale bzw. aller enthaltenen Signalabschitte erforderlich (‚time alignment'). Neben der Kompensation einer statischen Verzögerung ist eine komplexe Verarbeitung bei variabler Laufzeit (z.B. VoIP) erforderlich.

Nach einer Unterabtastung erfolgt zunächst ein ‚prealignment', bei dem anhand sicher erkannter ‚landmarks' (grüne Abschnitte in Abb. 2) im Signal grobe Zuordnungen

getroffen werden können. Basierend auf dieser groben Zuordnung werden dann die noch undefinierten Bereiche zugeordnet (blaue Abschnitte in Abb. 2). Dazu werden verschiedenen möglichen Laufzeiten Wahrscheinlichkeiten ermittelt und in einer Sparse-Matrix gespeichert. Mittels eines Viterbi-ähnlichem Algorithmus wird das wahrscheinlichste Laufzeitprofil in einer Korrepondenztabelle erstellt.

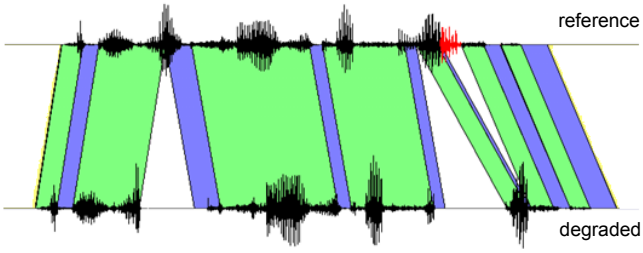

Abbildung 2: POLQA Laufzeitausgleich

Am Ende dieses Prozesses steht zu jedem Abschnitt des zu bewertenden Signals ein Abschnitt der Referenz zur Verfügung. Zusätzliche und fehlende Abschnitte werden gesondert behandelt (z.B. roter Abschnitt in Abb. 2).

Idealisierung – Ein neuer Ansatz

Bekannte ‚full-reference' Verfahren betrachten das Eingangssignal in das zu bewertende System, das Referenzsignal, als ideal und wichten wahrnehmbare Unterschiede zu diesem als Störung. Ist das zu bewertende Signal diesem Referenzsignal ähnlich, dann resultiert dies in einem hohen Qualitätswert, sind beide Signale identisch, wird der höchstmögliche Wert erreicht. Gerade im praktischen Einsatz sind diese Referenzsignale nicht immer als ideal anzusehen. Sie können ein geringes Maß an Geräuschen enthalten oder eine individuelle, durch Aufnahme oder Vorverarbeitung verfälschte spektrale Charakteristik aufweisen. Solche Sprachsignale werden auch in einem Hörtest als nicht hochqualitativ erkannt und entsprechend niedriger bewertet.

P.863 nutzt erstmals eine interne Idealisierung des Referenzsignals. In erster Linie werden additive Geräusche minimiert und spektrale Charakteristika an Idealverläufe angeglichen. Das idealisierte Referenzsignal stellt mit Einschränkungen das Signal dar, welches ein Hörer als interne Referenz für diesen Sprecher und dieses Textbeispiel aufbauen würde.

Als Konsequenz bewertet P.863 ‚POLQA' nicht grundsätzlich mit dem höchsten Wert, wenn das Referenzsignal selbst zur Bewertung übergeben wird, da es ja mit seiner idealisierten Form verglichen wird. Diese wurde ‚bereinigt' und kann Unterschiede zum Referenzsignal aufweisen.

Psychoakustisches Modell

Das POLQA zugrunde liegende psycho-akustische Kernmodell kann teilweise als Weiterentwicklung von P.862 ‚PESQ' betrachtet werden. Es sind aber zwei entscheidende Unterschiede festzustellen.

Einzelne Störungsarten werden separat. Das betrifft additive Geräusche, spektrale Verformungen und Hall- und Echoanteile. Diese Signalstörungen bilden Teilqualitätsdimensionen, die im Qualitätsmodell berücksichtigt werden. Geräusche und spektrale Störungen werden dann sogar kompensiert und das eigentliche psycho-akustische Modell wird an Sprachsignalen angewandt, die diese Störungen kaum noch enthalten.

Das psycho-akustische Modell weist zudem gegenüber dem von PESQ eine höhere spektrale Schärfe auf und verwendet eine abweichende spektrale und temporale Maskierung.

Ergebnisse und Genauigkeit der Vorhersage

Im Rahmen der Evaluierung von POLQA wurde eine Bewertung auf Basis des r.m.s.e. vorgenommen. Generell läßt sich sagen, dass der Prädiktionsfehler im Mittel aller 37 ‚narrowband' Experimente gegenüber PESQ um >25% reduziert wurde. Für einzelne Experimente, speziell für neuere Übertragungstechniken und Übertragungen in komplexen Netzwerken sogar um bis zu 50% (s. Appendix I [1]).

Vergleich ITU-T P.863 und J.341

Der Vergleich von P.863 und J.341 ist interessant, da beide Verfahren dem ‚full-reference' Ansatz folgen. Auch J.341 vergleicht das zu bewertende Signal mit dem Referenzsignal. Zunächst erfolgt eine globale Luminanzanpassung und eine Größenreduzierung der Bildauflösung (vglb. zur globalen Pegelanpassung und Unterabtastung bei POLQA). Danach wird eine Zuordnungstabelle, hier für die einzelnen Bilder des Videos, erstellt. Der dazu verwendete Ansatz entspricht dem des ‚pre-alignments' unter Berücksichtigung der ‚landmarks'. Die Auswertung erfolgt ebenfalls in kurzen Abschnitten, hier in Videoframes. Auch hier erfolgen lokale Luminanzanpassungen und Differenzberechnungen zur Referenz. Temporale Effekte werden über Inter-Frame-Differenzen berücksichtigt.

Der wesentliche Unterschied liegt in der deutlich rudimentäreren Modellierung psycho-visueller Effekte verglichen mit dem psycho-akustischen Modell von POLQA. Das Videoqualitätsschätzverfahren basiert auf eher technisch motivierten Differenzberechnungen der beiden Pixelmatrizen sowie in der gezielten Suche nach bekannten Artefakten der Videocodierung und -übertragung, wie z.B. Blockstörungen und Slicingeffekten. Zusätzlich kann im Videomodell keinerlei Annahme über eine Abgrenzung des Inhalts getroffen werden, dieser kann beliebig sein. Für POLQA gilt dies nicht, die Signale sind Sprachsignale und Begrenzungen des menschlichen Sprachproduktionstraktes unterworfen.

Literatur

[1] ITU-T Rec. P.863: Perceptual Objective Listening Quality Asessment. Genf, 2011

[2] ITU-T Rec. J.341: Perceptual Objective Listening Quality Asessment. Genf, 2011

[3] ITU-T Rec. P.862: Perceptual Evaluation of Listening Quality Asessment. Genf, 2001-2002

[4] ITU-T Rec. P.800: Methods for subjective determination of transmission quality. Genf, 1996-2008

http://www.itu.int/en/ITU-T/publications/Pages/recs.aspx