POLQA®

Perceptual Objective Listening Quality Analysis

- Technical White Paper -

Updated October 2011

Contents

Introduction to POLQA

POLQA is the next-generation voice quality testing technology for fixed, mobile and IP based networks. POLQA was standardized by the International Telecommunication Union (ITU-T) as new Recommendation P.863 (2011), and can be applied for voice quality analysis of traditional and HD Voice, 2G, 3G and 4G/LTE networks.

POLQA - which stands for "*Perceptual Objective Listening Quality Analysis*", was developed 2006-2011 by leading experts of ITU-T Study Group 12, including well experienced developers, who had already co-authored the preceding standards ITU-T Rec. P.861 (PSQM) and P.862 (PESQ). POLQA – a technology upgrade for PESQ – offers a new level of benchmarking capability to determine the voice quality of mobile and fixed network services.

The POLQA perceptual measurement algorithm is a joint development of OPTICOM, SwissQual and TNO, protected by copyright and patents and available under license as software for various platforms.

New ITU-T Recommendation P.863

Various commercial entities offer voice quality measurement solutions, each claiming that their method offers superior performance. However, only a very few companies were confident enough to allow their algorithms to be tested independently in a public competition, and to have the performance of their particular method disclosed. This is the level of transparency required by the ITU-T if a method is to become a worldwide-recognized ITU-T standard. The developers of POLQA faced this challenge – and clearly won. The advantage for end users of such an internationally recognized standard is that the measurement method is robust, well known and documented. Only measurement results obtained by such exacting methods are 100% comparable between different test labs and test systems, and only ITU-T standards can guarantee this. Furthermore, the ITU-T, being a leading agency of the United Nations, carefully maintains its standards. This guarantees stability over several years, protecting investments and allowing such methods to be referenced in long term SLAs.

POLQA & Evolving Technology

Testing of the new POLQA standard included numerous tests with EVRC-type codecs. This allowed the generation of accurate results with respect to GSM/UMTS coding schemes, where AMR and EFR are widely used. A reliable comparative rating is the key for true benchmarking of GSM/UMTS and CDMA networks and is one of the key success factors of POLQA.

Today, speech quality is no longer determined solely by the speech codec used, or by lost frames. There are also interactions with many other components which automatically control the signal level, applying smart loss concealment and similar strategies, in order to increase intelligibility in case of poor conditions. Those components are integrated within networks and devices, and may modify the speech signal with new and unexpected degradation effects. POLQA is especially trained to handle disruptive effects caused by these multi-component channels.

Along with new voice services, stretching and compression of speech signals in the time domain is becoming common practice. In comparison to time-variant transmissions of years gone by - which usually had a strong, perceptible impact on quality - newer algorithms are typically much smarter and keep quality at a high level. The correct interpretation and scoring of distinct remaining impacts was one of the most complex technical challenges, now solved by the advanced internal scaling and psycho-acoustic modelling of POLQA.

POLQA's radically revised psycho-acoustic and cognitive model allows, for the first time in the history of the ITU-T P.86X series, a true quality prediction for:

- EVRC type codecs
- Noise Reduction and Voice Quality Enhancement
- Time-warping, UCC and VoIP
- Non-optimal presentation levels
- Filtering and spectral shaping
- Recordings made at an ear simulator

POLQA fits well with new transmission technologies in service now or to be launched in the near future, and provides stable and accurate results along with an improvement in performance for existing technologies.

Application and Use Scenarios

PESQ already covers a broad application range, which includes most measurements in fixed and mobile networks, ranging from POTS to VoIP and 3G. But POLQA is a major step forward. The combination of improvements in POLQA

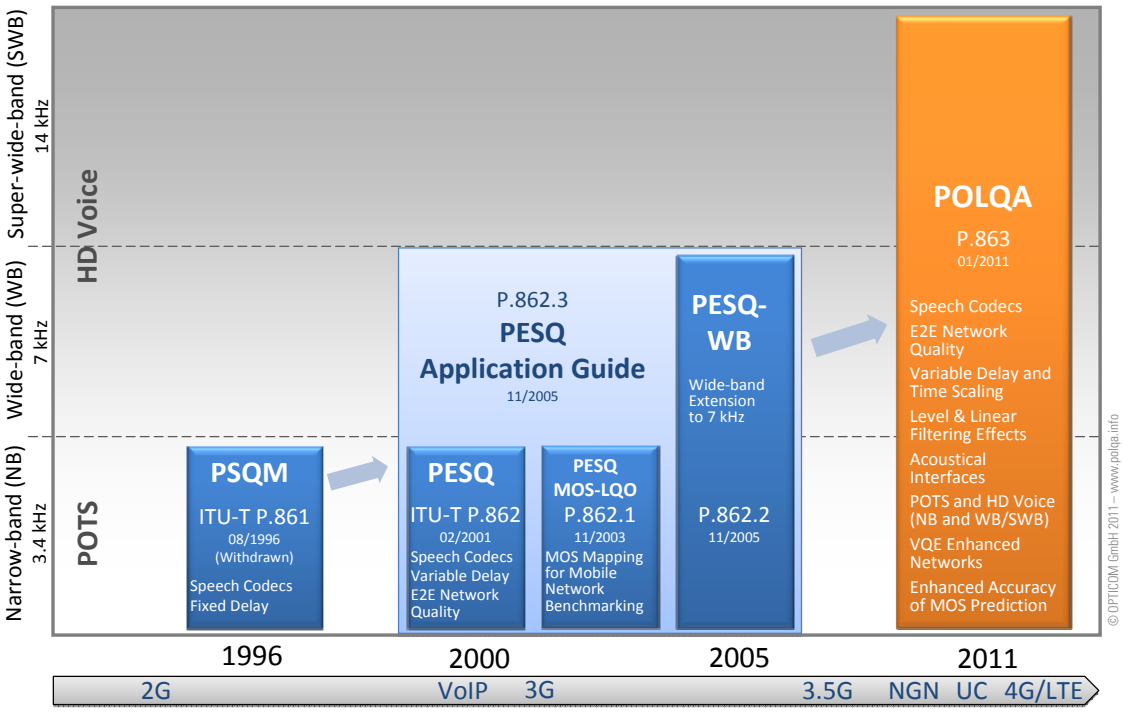
allows for a vast new range of applications that were problematic before, for example, the assessment of Voice Enhancement Devices, Skype calls, benchmarking EVRC with AMR and AAC codecs, or the assessment of terminals using head and torso simulators.

Applying POLQA to today's complex, unified networks will give a significant boost in accuracy and reliability compared to current standards. Due to its ability to handle time scaling effects POLQA can be used in virtually any scenario today; from video telephony to lab testing, from codec or network optimisation to the development and maintenance of Unified Collaboration and Communication services.

As with all methods of ITU-T's P.86X series, POLQA compares a known speech signal to the degraded voice signal by simulating human hearing. Unlike methods estimating quality based on network parameters or packet data analysis (i.e. ITU-T G.107 or P.564), this guarantees utmost accuracy and applicability independent of the underlying network technology and therefore provides measurement of true end-to-end quality.

POLQA is thus not only a full replacement for PESQ, but it is also a significant enhancement to the applicability of end-to-end voice quality testing. Since it is not significantly more complex than PESQ, POLQA is a logical upgrade path for all customers using PESQ today.

Evolution of ITU-T Recommendations for Voice Quality Testing (P.86x - Full Reference MOS-LQO)



Evolution of Network Technologies available at the time of development, i.e. included use cases for each Recommendation

Wideband Audio Transmission

Telecom industries are now initiating the evolution from narrow-band telephony to wideband speech transmission. At the time of writing this 32 public networks support HD voice and 52 devices have been released[1] The codecs for wide band are ready and approved by the standardization bodies, the handsets are not restricted in processing power and the core networks are being upgraded to support new transmission loads.

Within the last few years, efforts were made to standardize wideband versions of common telephony speech codecs like G.722.2 AMR-WB based on AMR, EVRC-WB based on EVRC-B as well as ITU-T G.729.1 which is the wideband extension of ITU-T G.729.

Combined with the evolution of audio codecs, the limitation to 7,000Hz bandwidth became restrictive. Current developments of voice codecs are processing the so-called super-wideband (up

[1] Source: http://www.gsacom.com/news/gsa_339.php4

to 14,000Hz) or even higher ('fullband'), up to approx. 24,000Hz. However, the perceived difference between super-wideband and fullband can be ignored in the case of human speech.

Super-wideband and Narrowband MOS Scores

In traditional telephony scenarios, the expectation is set at a perfect narrow-band voice signal. A signal that is close or identical to such a signal is scored subjectively by human listeners with a high quality value (usually a MOS-LQ of around 4.5 on a five-point scale). Additional degradations will decrease the quality value towards to 1.0. A typical value for a perfect mobile connection using AMR 12.2 or EFR is still around 4.0. Interruptions, un-concealed transmission errors and noises easily push the quality into lower regions.

Within a super-wideband scenario the situation is different. The expectation of excellent quality is a perfect super-wideband speech signal. Since the same five-point scale is used, such a perfect super-wideband signal is also subjectively scored close to excellent in the range of 4.5. Obviously, a narrow-band signal in that super-wideband context will not fulfil the expectation of high quality due to its band limitation. Consequently, it will be scored lower in this context.

Since the range of the scores is the same but the meaning is different depending on the context, the two are named as different scales: *narrow-band* or *super-wideband*.

Broadly the main difference is that narrow-band signals will be scored lower in a super-wideband context than in narrow-band experiments, since the band-limitation is scored as degradation. Hence, scores given on the two different scales must not be mixed or directly compared.

Super-wideband and Narrowband in POLQA

To cover both application areas - narrowband telephony and super-wideband communication - POLQA supports two operational modes.

The application of POLQA is exactly the same in both cases; the change between the two modes just requires the use of a control flag. All required adjustments are automatically made by POLQA internally. Consequently, in narrowband mode POLQA scores on a five-point narrow-band scale, in the super-wideband mode on a five-point *super-wideband scale.*

Most important for customers will be typical values and one-to-one comparisons obtained under super-wideband application with common measurements in narrow-band mode.

The following table (Table 1) shows typical values to be expected from POLQA. These were confirmed by subjective auditory experiments.

	MOS-LQ super-wideband 50-14000 Hz	MOS-LQ narrow-band 300-3400 Hz
Transparent transmission 50 – 14000 Hz or wider	**4.75**	**-**
Transparent transmission 50 – 7000 Hz ('old' wideband)	**4.3**	**-**
AMR-WB 12.65 kbps (50 – 7000 Hz)	**3.8**	**-**
Transparent transmission 300 – 3400 Hz ('POTS')	**3.0**	**3.6**
G.711 (A-Law standard PCM)	**3.5**	**4.3**
EFR / AMR-FR 12.2kbps	**3.2**	**4.2**
EVRC 9.5 kbps	**3.0**	**3.9**
EVRC-B 9.5 kbps	**3.0**	**3.9**
AMR-HR 7.95 kbps	**2.9**	**3.9**

Table 1. POLQA MOS Score Comparison

Performance and Accuracy

POLQA is an objective model of subjective Listening Only Tests, where the quality is scored on an Absolute Category Rating Scale. In such experiments, subjects are requested to score a given speech signal on a five-point scale without a direct comparison to a reference. The subjects score in comparison to their expectation and experience in a given context (i.e. narrow-band telephony). Data sets of those subjective experi-

ments are used to develop and evaluate objective models.

More than 150 subjective experiments consisting of more than 10,000 subjectively-scored speech samples were used during the development of POLQA. ITU-T tests used 62 experiments with more than 45,000 speech samples for the evaluation of POLQA. POLQA was expected to show high performance and superior accuracy across all experiments covering an extremely wide range of distortions, as defined in the POLQA scope by ITU-T.

Compared to P.862 a three times larger evaluation set was used, covering a much wider range of distortions. Even though the application range is much wider, residual prediction errors of POLQA are considerable smaller than those from PESQ.

The statistical evaluation made by ITU-T has clearly shown that POLQA significantly outperforms P.862 'PESQ' in both narrow-band and wideband mode.

Technical Structure

POLQA is a model to predict listening quality as it is perceived in an ITU-T P.800 listening experiment using idealized listening devices. POLQA uses an advanced psycho-acoustic model for emulating the human perception and transforming the sound into an internal neuronal representation.

POLQA, as a full reference approach, compares the input or high quality reference signal and the associated degraded signal under test. This process is shown in the overview in Figure 2.

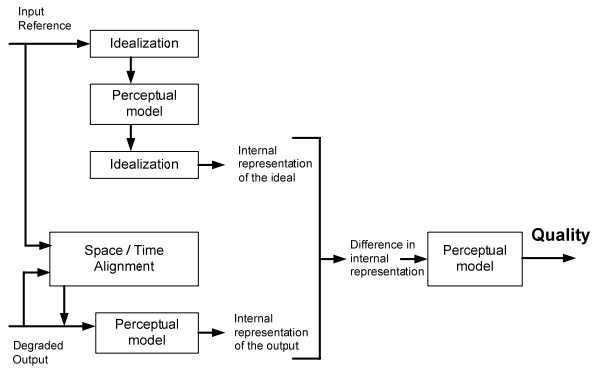


Figure 2. Overview of POLQA Algorithm

POLQA uses the concept of idealization of both input signals in multiple steps. This ensures that only the relevant perfect speech information is used for comparison and any unwanted signal components are discarded.

Psycho-acoustics and Cognitive Modelling

One important component of POLQA is the perceptual model, which takes into account masking effects of the human hearing. The signal is subdivided into short segments and transformed to the spectral domain (Figure 3).

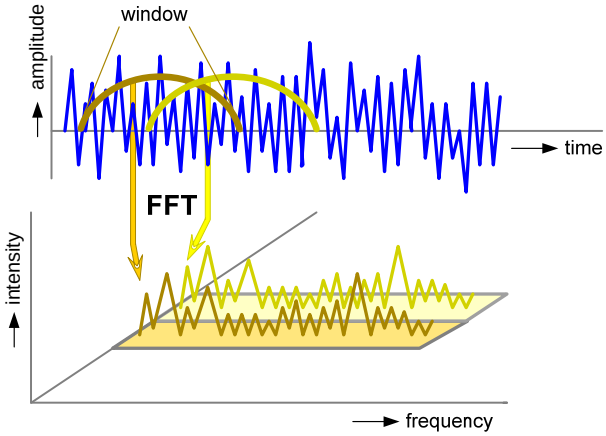


Figure 3. Waveform to Intensity Warping

Further psychoacoustic processing and considered masking follow common psychoacoustic models. A transformation to the Bark scale (critical bands) and an intensity warping to the perceptual-based 'sone' scale are applied. In the Bark domain, masking thresholds are calculated. POLQA then applies both spectral and temporal masking (Figure 4).

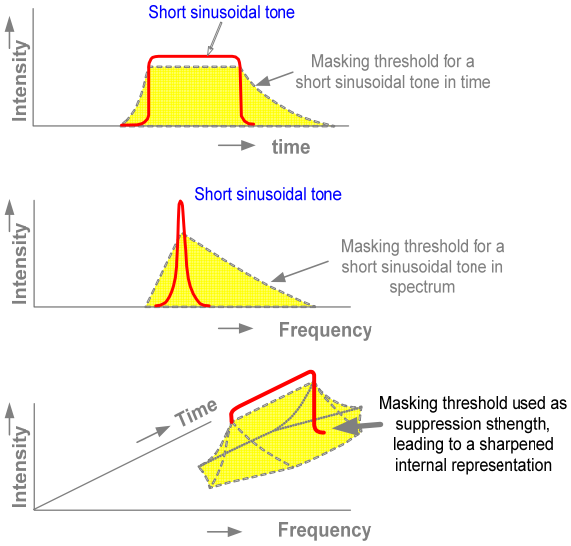


Figure 4. Calculation of Masking Thresholds

The following high level diagram shows all the important processing steps (Figure 5). POLQA follows the approach of detecting individual, orthogonal distortions in the perceptual model and combining those later in the cognitive model.

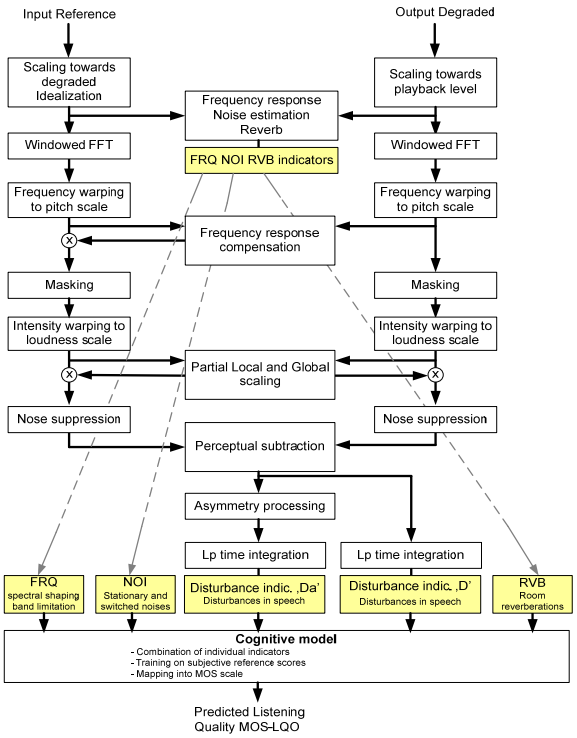


Figure 5. Perceptual Model

Indicators for specific degradations such as noise, spectral shaping and reverberations are derived in the early stages of the algorithm and are weighted by specialized perceptual approaches. The overall disturbance indicator, however, is derived by applying an advanced hearing and masking model to the speech content of the signal.

After a first idealization process which aligns the reference and degraded signal in level and spectral shape, the psycho-acoustic modelling is applied.

The second step of idealization consists of a local scaling and more importantly a noise suppression step. This ensures that only the speech content of the two signals will be compared in the internal representation domain. The degradation of interlaced noises is already considered by the separate noise degradation indicator and therefore not taken into account in the disturbance indicator.

All indicators are used in the final cognitive model. While the perceptual model determines which distortions can be perceived by the listener, the cognitive model decides how annoying those distortions are. The output of the cognitive model forms the final overall quality score.

Implementation of POLQA

POLQA is commercially available in the form of libraries for integration into OEM measurement equipment, as well as in end user products. The OEM libraries are very easy to use and require only the PCM samples as input (in a file or in memory), plus some very simple control parameters. The measured KPIs (e.g. MOS, delay, level) will be returned as a data structure in memory. The provided POLQA OEM implementation can operate with speech signals independent from their sampling frequency. The POLQA OEM libraries are guaranteed to be 100% in conformance with the ITU reference implementation.

POLQA is available for Windows, Linux and Android operating systems. POLQA libraries supporting other operating systems are planned for future release.

The consequential run-time optimization – especially for Intel CPUs – decreases the processing time to less than 20% of the speech signal duration on a year 2010 off-the-shelf PC.

Conclusions

The upcoming new ITU-T Recommendation P.863 ‘POLQA’ is a technology update for the current P.862 ‘PESQ’ and also a significant enhancement to the applicability of end-to-end voice quality testing.

Since POLQA applies the same scale and scores common techniques in the same quality range as PESQ today, users will be able to obtain backward compatible results in applications where PESQ has been delivering accurate results today. Concerning implementation efforts, POLQA is not considerably more complex than PESQ.

The combined advancements in POLQA allow for a vast new range of applications that were somehow problematic to assess with PESQ, e.g. the assessment of Voice Enhancement Devices, Skype calls, benchmarking EVRC with AMR and AAC codecs, or the assessment of terminals using head and torso simulators.

It is worth noting that the use of the advanced psycho-acoustics model in POLQA is not restricted to voice quality only; it may also become the heart of other applications in the future that include testing of intelligibility, audio quality and perceptual-based echo/noise ratings.

Published by:

OPTICOM GmbH
Naegelsbachstrasse 38
D - 91052 Erlangen
GERMANY

Phone: +49 (0) 91 31 - 5 30 20 - 0
Fax: +49 (0) 91 31 - 5 30 20 - 20
info@opticom.de
http://www.opticom.de

VAT ID No. DE 194 631 268
Managing Directors:
Dipl.-Ing. Michael Keyhl, CEO
Dipl.-Ing. Christian Schmidmer, CTO
Register:
Amtsgericht Fürth (Bay.) HRB 7169

SwissQual AG
Allmendweg 8
4528 Zuchwil/Solothurn
Switzerland

Phone: +41 32 686 65 65
Fax: +41 32 686 65 66
info@swissqual.com
http://www.swissqual.com

VAT ID No. CH 554 715
Directors:
John May, CEO
Martin Coates, Chairman
Registration no.
CH-241.3.002.491-9

For inquiries on POLQA Licensing please contact OPTICOM GmbH or visit www.polqa.info for further details.
For an updated reference list of available POLQA products and solutions please refer to our website:

www.polqa.info