Next-Generation (3G/4G) Voice Quality Testing with POLQA®

White Paper

POLQA® (Perceptual Objective Listening Quality Analysis) is the next-generation mobile voice quality testing standard according to ITU-T recommendation P.863 and has been especially developed for the super wideband requirements of HD Voice, 3G, VoLTE (4G), VoHSPA and VoIP. This white paper describes the POLQA® algorithm implemented in the R&S® UPV Audio Analyzer and shows an example hardware setup for standard independent audio measurements.

1MA202

Ottmar Gerlach
03.2012 - 1e

ROHDE&SCHWARZ

Table of Contents

1 Introduction

POLQA is a next-generation mobile voice quality testing standard according to ITU-T recommendation P.863[2]. It has been especially developed for super wideband (SWB) requirements of HD Voice, 3G, VoLTE (4G), VoHSPA and VoIP (Voice over Internet Protocol). This white paper describes the POLQA algorithm implemented in the R&S® UPV Audio Analyzer, points out the enhancements compared to the PESQ[2] (Perceptional Evaluation of Speech Quality) measurement and shows an example hardware setup for speech quality testing.

2 Overview

A migration to POLQA became necessary since certain conditions in current and emerging networks had not been considered in PESQ ITU-T P.862 recommendation. The performance of POLQA has been enhanced to allow for:

- New types of speech codecs as used in 3G/4G/LTE and audio codecs, e.g. AAC and MP3.
- Voice Enhancement (VQE/VED) systems using non-linear processing.
- Codecs that modify the audio bandwidth, e.g. SBR (Spectral Band Replication).
- Measurements on signals with very high background noise levels
- Correct modeling of effects caused by variable sound presentation levels.
- Support of NB (narrowband, 300 to 3400 Hz) and SWB (super-wideband, 50 to 14000 Hz) mode.
- Handling of time-scaling and –warping as seen in VoIP and 3G packet audio.
- Evaluation of signals recorded with acoustic interfaces.
- Correct weighting of reverberation, linear and non-linear filtering.
- Direct comparison between AMR (GSM/UMTS) and EVRC (CDMA2000) coded transmissions.

Possible applications for POLQA are:

- Codec evaluation.
- Terminal testing with or without influence of the acoustical path and electro-mechanical transducers in sending and receiving directions.
- Bandwidth extensions.
- Live network testing using digital or analog connection to the network.
- Testing of emulated and prototype networks.
- UMTS, CDMA2000, GSM, TETRA, WB-DECT, VoIP, POTS, Video, Telephony, Bluetooth.
- Voice Activity Detection (VAD), Automatic Gain Control (AGC).
- Voice Enhancement Devices (VED), Noise Reduction (NR).
- Discontinuous Transmission (DTX), Comfort Noise Insertion.

2.1 POLQA Algorithm

The POLQA algorithm compares a reference signal X(t) with a signal Y(t) which is degraded from passing a communication system with coding, decoding, LAN and RF components. The algorithm output is a prediction of the perceived quality as would be given to Y(t) by persons in a subjective listening test.

In a first step the reference and degraded signal are split into very small time slices referred as frames. Then the delay of each reference signal frame relative to the associated degraded signal frame is calculated. The sample rate of the degraded signal is then estimated. If the estimated sample rate significantly differs from the reference signal sample rate, the signal with the higher sample rate will be down sampled and the delays re-determined.

Based on the found delay set, POLQA compares the reference (input) with the aligned degraded (output) signal of the SUT (system under test) using a perceptual model as shown in Figure 1.

The key to this process is the transformation of both signals to an internal representation analogous to the psychophysical representation in the human auditory system, taking into account the perceptual pitch (Bark) and the loudness (Sone). This is achieved in several consecutive stages:

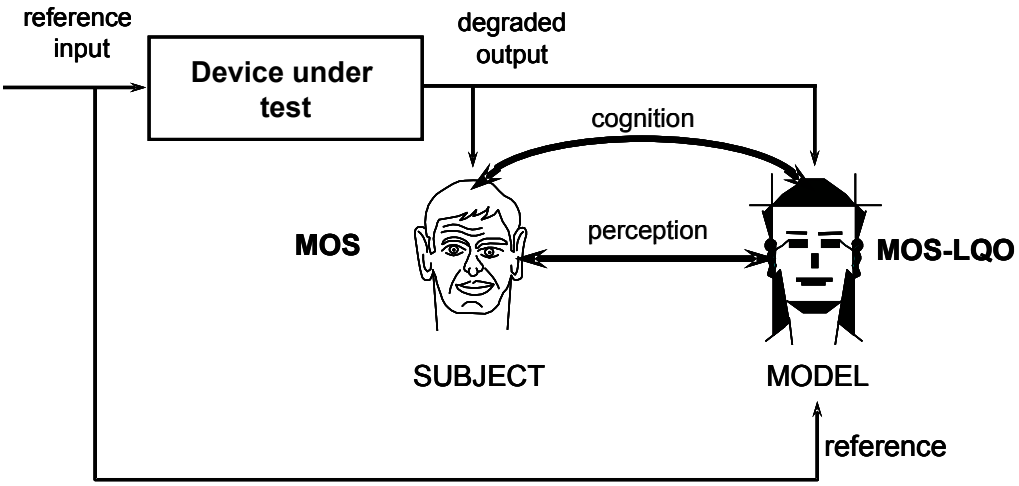
- Time alignment
- Level alignment to a calibrated listening level
- Time-frequency mapping
- Frequency warping
- Compressive loudness scaling

POLQA takes the playback level for the perceived quality prediction into account in SWB mode. In NB mode the speech quality is determined with a constant listening level. By processing the internal representation level, local (rapid) gain variations and linear filtering effects can be taken into account.

POLQA also eliminates low noise levels in the reference signal and partially suppresses noise in the degraded output signal. Operations that change the characteristics of the reference and degraded signal are used for the idealization process. This subjective testing is carried out without direct comparison with to the reference signal (Absolute Category Rating). It supplies six quality indicators that are computed in the cognitive model

- Frequency response indicator (FREQ)
- Noise indicator (NOISE)
- Room reverberation indicator (REVERB)
- Three indicators describing the internal difference in the time-pitch-loudness domain

These indicators are combined to give an objective listening quality MOS. POLQA always expects a clean (noise-free) reference signal.



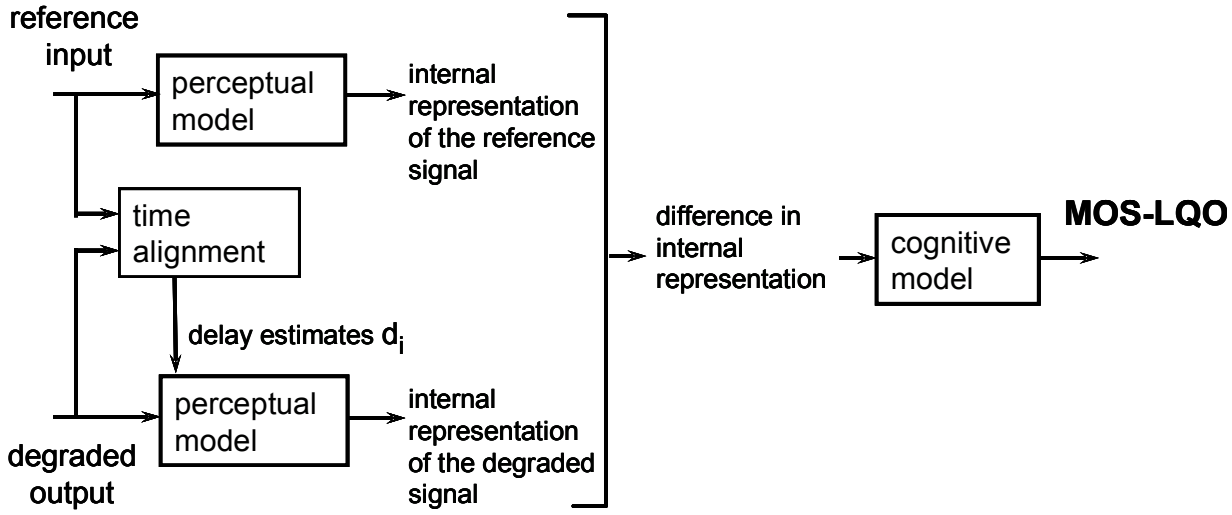


Figure 1: Basic POLQA Philosophy

The difference between subjective and objective listening quality scores is that the subjective score depends on the listener group and design of the test. The objective measure is independent from the test context and individual behavior of the listening panel. It reflects an *'average test scored by an average group of listeners'*. An objective model cannot exactly reproduce the absolute scores of an individual experiment, but it reproduces the relative quality ranking. A good objective quality measure should have a high correlation with many different subjective experiments.

In daily practice, no special mappings to the objective scores must be applied, since the POLQA scores are already mapped to a MOS scale reflecting the average over a huge amount of individual data sets.

In SWB mode POLQA always requires a mono signal with i.e. 48 kHz sampling rate which has to be pre-filtered with a 50 to 14000 Hz band-pass filter. This signal can also be used for NB mode. It can alternatively be down sampled to 16 or 8kHz.

The speech files used in the POLQA evaluation phase had following attributes:

• Each reference speech file should consist of two or more sentences separated by a gap of at least 1 s but not more than 2 s.
• The minimum amount of active speech in each file should be 3 s.
• Reference speech files should have a sufficient leading and trailing silence intervals to avoid clippings of the speech signal, e.g. 200 ms of silence each.
• For SWB reference speech samples the noise floor of the reference files should not exceed -84 dBov(A)[1)] in the leading and trailing parts as well as in the gaps between the sentences.
• The room used for recording reference material must have a reverberation time below 300 ms above 200 Hz (e.g. an anechoic chamber).

The degraded signal that has passed through the SUT was captured either at the electrical interface or at the acoustical interface.

The MOS scale ranges from 1 to 5 and the predicted scores reach a maximum value MOS-LQO = 4.75 for SWB and MOS-LQO = 4.5 for NB due to saturation (see 2.1.2.10).

[1)] The unit dBov (= overload) is the amplitude of a (usually audio) signal compared with the maximum which a device can handle before clipping occurs. Similar to dBFS, but also applicable to analog systems. The decibel A filter is widely used. The unit dB(A) roughly corresponds to the inverse of the 40 dB (at 1 kHz) equal-loudness curve for the human ear. A sound level meter is less sensitive to very high and low frequencies with a dBA filter.

2.1.1 Technical Overview

An overview on the POLQA algorithm is shown in Figure 2.

The inputs are two 16 bit waveforms. The first one contains the (undistorted) reference and the second one the degraded signal.

The POLQA algorithm consists of a **sample rate converter** used to compensate differences in the input signal sample rates, a **temporal alignment block**, a **sample rate estimator**, and the actual **core model**, which performs the MOS calculation.

In a first step, the delay between the two input signals is determined and the sample rate of the two signals relative to each other is estimated. The sample rate estimation is based on the delay information calculated by the temporal alignment.

If the sample rate differs by more than approximately 1%, the signal with the higher sample rate is down sampled. After each step, the results are stored together with an average delay reliability indicator, which is a measure for the quality of the delay estimation. The result from the re-sampling step, which yielded the highest overall reliability, is finally chosen.

Once the correct delay is determined and the sample rate differences have been compensated, the signals and the delay information are passed on to the perceptual model, which calculates the perceptibility as well as the annoyance of the distortions and maps them to a MOS scale.

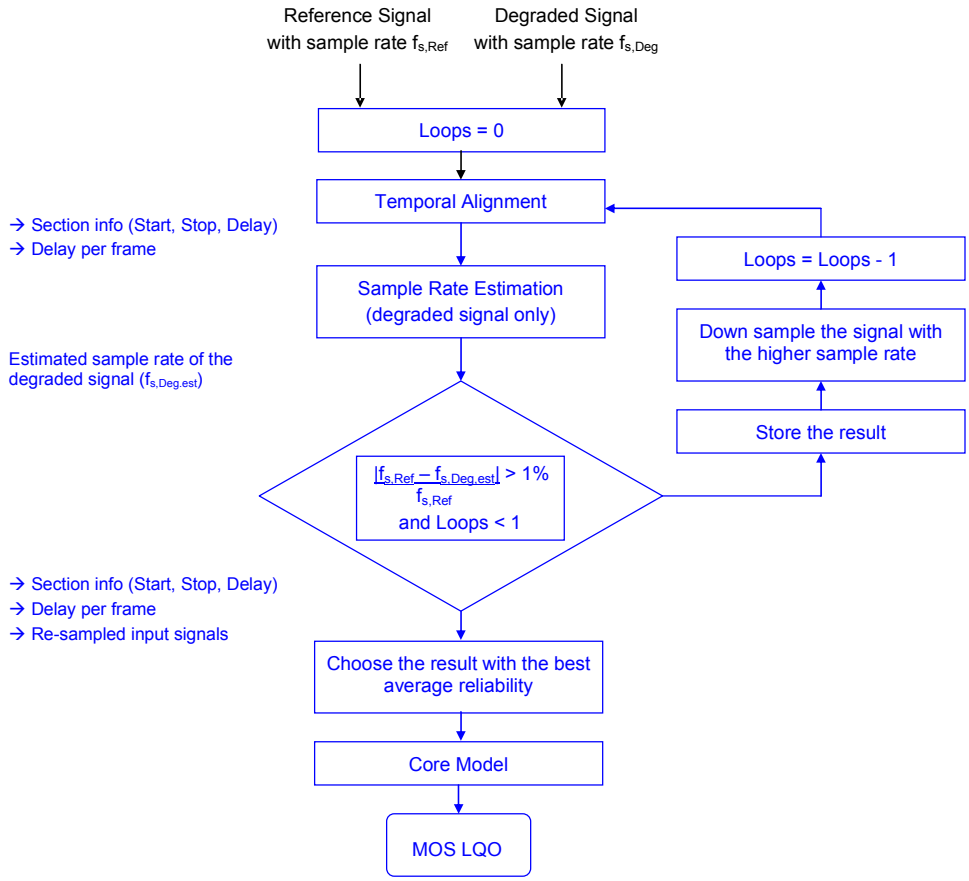


Figure 2: POLQA Overview

2.1.1.1 Temporal Alignment

The temporal alignment:

- Splits the signals into equidistant pairs of frames and calculate a delay for each frame pair.
- Searches the matching counter parts of the degraded signal sections in the reference signal and not vice versa, when possible.
- Stepwise refines the delay per frame to avoid long search ranges which require high computing power and are critical in combination with time scaled input signals.

The temporal alignment consists of the major blocks:

- filtering
- pre-alignment
- coarse alignment
- fine alignment
- section combination

The input signals are split into equidistant macro frames which length is dependent on the input sample rate. The delay is determined for each "macro frame". The calculated delay is always the delay of the degraded signal relative to the reference signal.

The **pre-alignment** determines the active speech sections of the signals, calculates an initial delay estimate per macro frame and an estimated search range required for the delay of each macro frame.

The **coarse alignment** performs an iterative refinement of the delay per frame, using a multidimensional search and a Viterbi-like backtracking algorithm to filter the detected delays. The resolution of the coarse alignment is increased from step to step in order to keep the required correlation lengths and search ranges small.

The **fine alignment** finally determines the sample exact delay of each frame directly on the input signals with the maximum possible resolution. The search range in this step is determined by the accuracy of the last iteration of the coarse alignment. In a final step all sections with almost identical delays are combined to form the so-called "Section Information".

This temporal alignment procedure has the following characteristics:

- No hard limit for the static delay.
- Designed to handle a variable delay of less than 300 ms around the static delay, but no hard limit exists.
- Delay may vary from frame to frame.
- Small sample rate differences (less than approx. 2%) can be handled well, larger differences will be detected and compensated outside of the temporal alignment.
- Time stretched or temporally compressed signals with or without pitch correction are handled well.
- Alignment works well even under very noisy conditions with an SNR below 0 dB.
- No problems observed with signal level variations.

General Delay Search Method
Most modules related to temporal alignment use the same method to find the delay between two signals. This method is based on histogram analysis created by:

- Calculating the cross correlation between two signals.
- Centering the found peak value into the histogram.
- Shifting both signals by a small amount.
- Repeating this step again.

Once the histogram contains enough values, it is filtered and the peak determined. The position of this peak in the histogram is equivalent to the delay offset between the two signals.

General Delay Reliability Measure
In most temporal alignment steps the simple Pearson correlation is used as a reliability measure for a found delay between two signals.

Bandpass Filter
Both input signals are bandpass filtered before any further step. The filter shape depends on the model operating mode (SWB or NB).

In SWB mode, the signals are bandpass filtered from 320 Hz up to 3400 Hz. In NB mode, the signals are bandpass filtered from 290 Hz up to 3300 Hz.

Please note that those filtered signals are only used for the temporal alignment. The perceptual model uses differently filtered signals.

Pre-Alignment
The pre-alignment first identifies reparse points in the degraded signal. Reparse points are positions where the signal makes a transition from speech pause to active speech.

The reparse points mark the beginning of active speech sections, while reparse sections describe the entire active speech segment beginning at a reparse point.

The reparse section information is calculated for each reparse point. The section information stores the section's beginning and end position as well as an initial delay value, an reliability indication of the found delay and its accuracy, i.e. upper and lower limit in which the accurate delay is expected to be.

Coarse Alignment
The coarse alignment performs a stepwise refinement of the delay per frame. This is implemented by splitting each signal into small subsections (feature frames) and by calculating a characteristic value (feature) for each subsection.

The resulting vectors are called **feature vectors**. Feature frames are equidistant and their length is reduced from iteration to iteration. Their length is independent from the macro frame length. The iterative length reduction increases the accuracy of the estimated delay with each iteration, but at the same time the search range is reduced.

Multiple feature vectors are calculated and the feature which is best suitable for each macro frame is used to determine the current frame's final delay value.

The coarse alignment result is a vector with the delay per macro frame expressed in samples and an accuracy which depends on the feature frame length in the final iteration.

Fine Alignment
The fine alignment operates on the reference and degraded signal at the maximum possible resolution and determines each frame's precise delay expressed in samples. The required search range is drastically limited due to the previous alignment steps. Therefore, it is possible to predict the accurate delay values using very short correlations without losing accuracy. The fine alignment result is the sample accurate delay value of each macro frame.

Joining Sections with Constant Delay
In this step all sections with identical delay are combined, meaning one set of information (delay, reliability, start, stop, speech activity) is stored for the entire section.

In a second step each section *n+1* is combined with section *n*

if section *n+1* contains active speech **and** the delay for both sections differs by less than 0.3 ms or
if section *n+1* consists of a speech pause **and** the delay for both sections differs by less than 15 ms.

The resulting section information is passed to the psychoacoustic model.

2.1.1.2 Sample Rate Estimation

The important part is to separate delay variations caused by sample rate differences from those caused by distortions like packet loss or jitter buffer adjustments. POLQA performs this by calculating a histogram of all delay variations that might be caused by a sample rate difference.

The **sample rate ratio detection** is required to compensate perceptually irrelevant differences in the play speed of both reference and degraded signal. Such differences may have various reasons and may be intentional or not intentional.

The resulting effect in both cases is the same and can be described as a difference in the sample rate of two signals in the range of very few percent. This is not about the nominal but effective sample rate relative to another signal.

The detection of this effect implemented in POLQA is based on the delay per frame vector and detected active sections of the speech signals determined by temporal alignment. The algorithm is based on the theory that sample rate differences will lead to delay changes, which are proportional to the ratio of the effective sample rates. Only relative small changes are accepted since sample rate differences cause more short rather than few large delay variations.

The calculated histogram describes the distribution of delay variations per frame, meaning that each detected delay variation is divided by the duration of the preceding section without delay variation.

After filtering unreliable peaks from the histogram the position of the peak value indicates the ratio of sample rates. In order to calculate the exact value, the number of samples *NumAvg* stored in the histogram is counted, the weighted average of all values calculated (*AvgBin*) and the sample rate ratio *SRRatio* derived from this value.

If the detected sample rate ratio is larger than 0.01, the signal with the higher sample rate will be down sampled and the entire processing started from the beginning. This happens maximally one time to avoid excessive looping with signals for which the sample rate ratio cannot be reliably determined.

Even if the sample rate cannot be determined perfectly, e.g. in case of signals with additional variable delay, the detected sample rate ratio is still accurate enough to return the signals to the safe operating range of the temporal alignment.

2.1.2 Perceptual Model

Figure 3 shows a simplified block diagram of the perceptual model used to calculate the internal representation.

The pitch power densities (power as function of time and frequency) of reference and degraded are derived from the time and frequency aligned time signals. These densities are then used to derive the first three POLQA quality indicators for frequency response distortions (FREQ), additive noise (NOISE) and room reverberations (REVERB).

The internal representations of reference and degraded signal are derived from the pitch power densities in several steps. Four different variants of these densities are calculated, one representing the main branch, one the main branch for big distortions, one focused on added distortions and one focused on added big distortions.

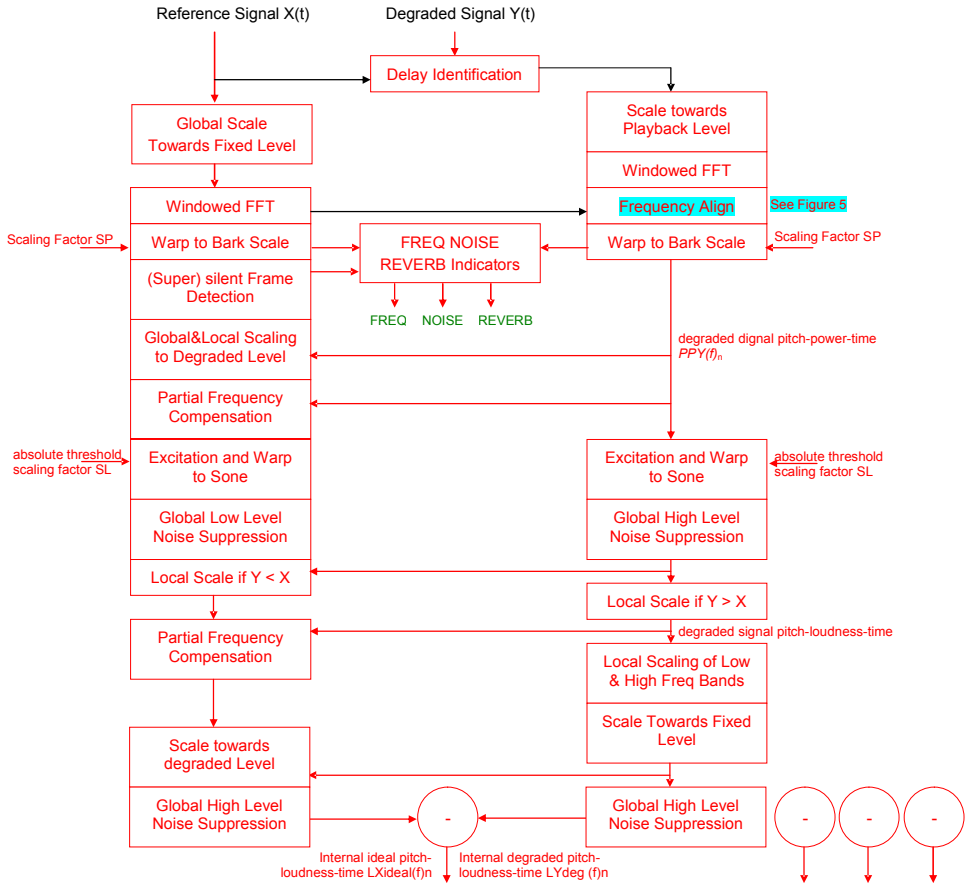


Figure 3: Overview of the first part of the POLQA Perceptual Model

Bark Scale

The Bark Scale is a psycho acoustical scale which ranges from 1 to 24 corresponding to the first 24 critical bands of hearing. The band edges in Hz are 20, 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, 9500, 12000 and 15500.

$$Z / Bark = 13 * \arctan(0.00076 * f / Hz) + 3.5 * \arctan((\frac{f}{7.5kHz})^2)$$

Excitation

The loudness density is calculated from the excitation level which is the difference between the level in a frequency group and the absolute threshold of hearing in this frequency group.

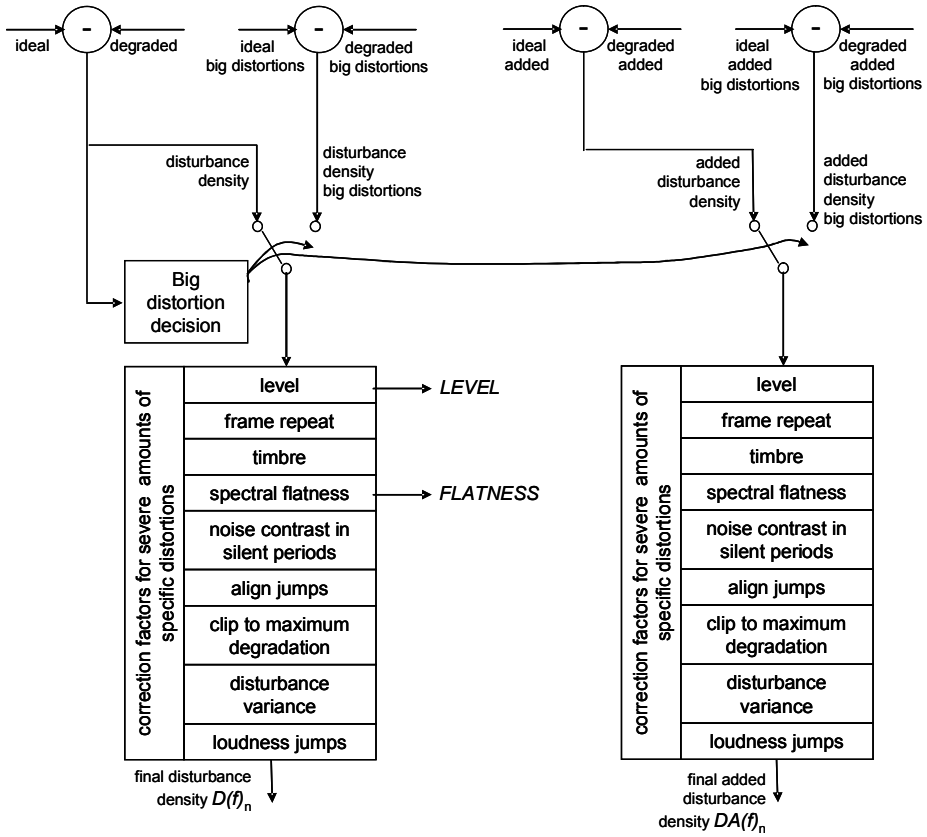


Figure 4: Overview of the second part of the POLQA Perceptual Model

In Figure 4 the final disturbance densities are calculated from the four different variants of the internal representations.

Level Variation

Level variation in digital transmission usually means that the signal to noise ratio (SNR) varies due to disturbances. Automatic Gain Control (AGC) circuits in the DUT increase or decrease the volume depending on the total power measured.

Noise Contrast in Silent Periods

Noise Contrast is the sudden change of the noise timbre, e.g. when simulated background noise (comfort noise) is turned on in DTX (Discontinuous Transmission).

Align Jumps

The delay between the original and transmitted blocks may vary due to missing blocks that need to be retransmitted and blocks that appear numerous times in IP-based Transmission. These effects may appear when the packet delay changes.

Loudness

Loudness is the quality of a sound that is primarily the psychological perception of the amplitude. Loudness is a subjective measure and is often confused with objective measures of sound strength such as sound pressure, sound pressure level (in dB), sound intensity or sound power.

Filters such as A-weighting attempt to adjust sound measurements to correspond to loudness as perceived by the typical human. However, loudness perception is a much more complex process than A-weighting. Furthermore, as the perception of loudness varies from person to person it cannot be universally measured using any single metric. Loudness is also affected by parameters other than sound pressure, including frequency, bandwidth and duration.

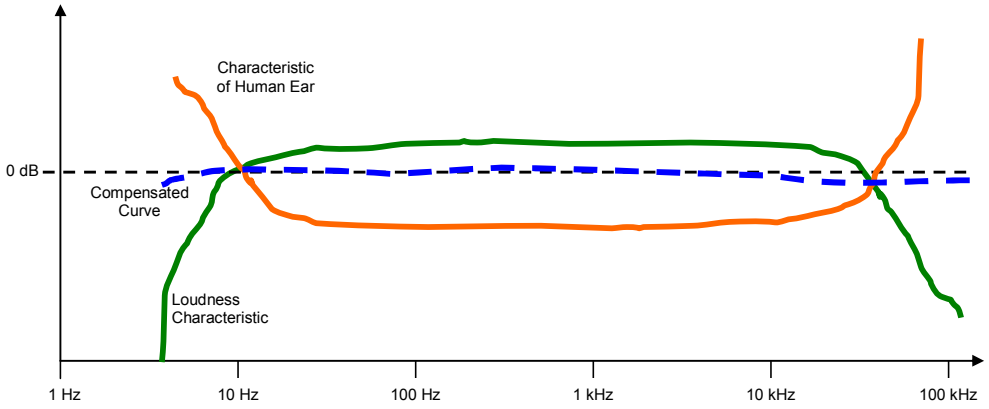


Figure 5: Loudness Compensation

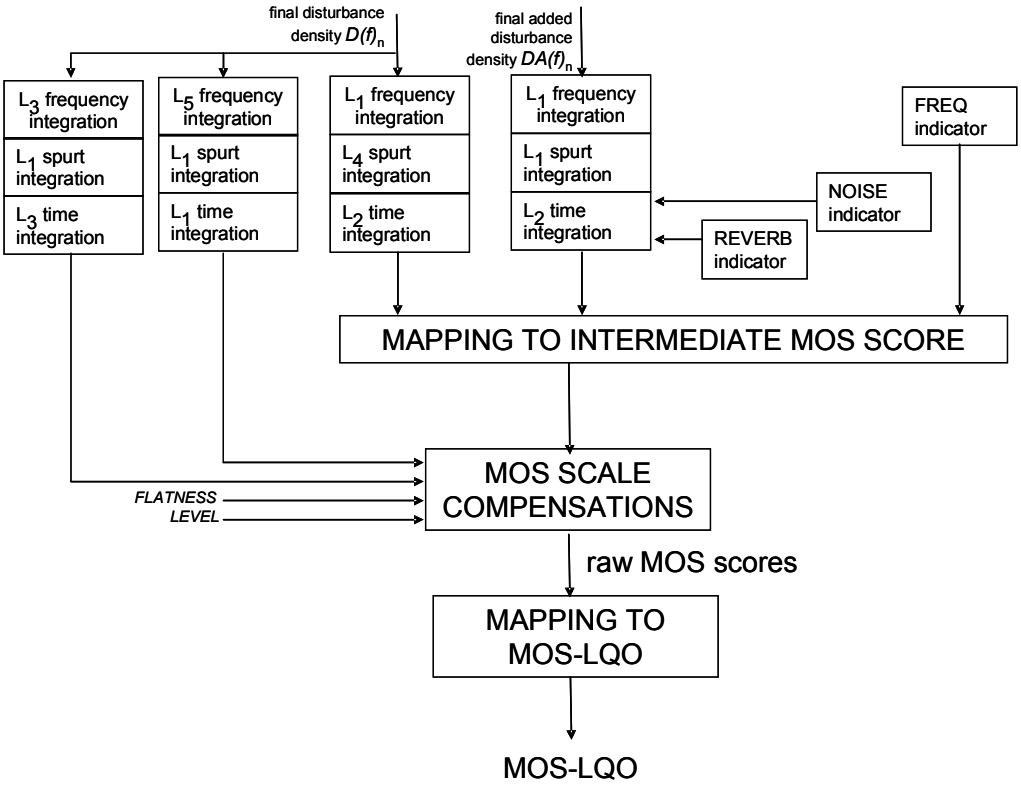


Figure 6: Overview of the third part of the POLQA Perceptual Model

In Figure 6 the MOS-LQO is calculated from the final disturbance densities.

2.1.2.1 Pre-Computation of Constant Settings

FFT Window Size Depending on Sample Frequency
The window size *W* depends on the sampling frequency f_s :

$$0 < f_s \le 9kHz \rightarrow W = 256$$
$$9 < f_s \le 18kHz \rightarrow W = 512$$
$$18 < f_s \le 36kHz \rightarrow W = 1024$$
$$36 < f_s \le 72kHz \rightarrow W = 2048$$

POLQA was tested with 8, 16 and 48 kHz sampling rate. Re-sampling will not reproduce exactly the same MOS score as it would in a subjective test, especially if the re-sampling deviates significantly from a factor of 2.

2.1.2.2 Pitch Power Densities

The degraded signal Y(t) is multiplied by the calibration factor *C* ($C = 2.8 * 10^{(-26-dBov)/20} * 10^{(73-dB(A))/20}$) and transformed to the frequency domain with 50% overlapping FFT frames. The reference signal is scaled towards a fixed optimum level.

A de-warping in the frequency domain is carried out on the FFT frames for files where the frequency axis is warped when compared to the reference.

First the reference and degraded FFT power spectra are preprocessed to reduce the influence of very narrow frequency response distortions and overall spectral shape differences on successive calculations. The preprocessing consists of performing a sliding window average of length equivalent to 100 Hz over both power spectra, taking the logarithm, and performing a sliding window normalization, using a window length equivalent to 218.75 Hz.

The current frame's reference to degraded pitch ratio is computed and used to determine the warping factor's search range, which lies between 1 and the mentioned pitch ratio. If possible, the search range is extended by the minimum and maximum pitch ratio found for one preceding and subsequent frame pair.

The function iterates through the search range and warps the degraded power spectrum with the current iteration's warping factor and processes the warped power spectrum as described above.

The correlation of the processed reference and warped degraded spectrum is then computed for bins between the common lower frequency limit and 1500 Hz. After complete iteration the "best" (highest correlation) warping factor is retrieved. The correlation of the processed reference and best warped degraded spectra is then compared against the correlation of the original reference and degraded spectra.

The "best" warping factor is then kept if the correlation increases by a set threshold. The warping factor is limited by a maximum relative change to the one determined for the previous frame pair, if necessary.

After de-warping the frequency scale in Hertz is warped towards the pitch scale in Bark reflecting that the human hearing system has a finer frequency resolution at low than at high frequencies. This is implemented by binning FFT bands and summing the corresponding powers of the FFT bands with a normalization of the summed parts. The resulting reference and degraded signals are known as the pitch power densities $PPX(f)_n$ and $PPY(f)_n$ with *f* the frequency in Bark and index *n* representing the frame index.

2.1.2.3 Computation of Speech Active, Silent and Super Silent Frames

POLQA operates with three frame classes:

• Speech active frames	where	reference signal frame level > average level – 20 dB
• Silent frames	where	reference signal frame level < average level – 20 dB
• Super silent frames	where	reference signal frame level < average level – 35 dB

2.1.2.4 Computation of Frequency, Noise and Reverb Indicators

An indicator from the average spectra of reference and degraded signals for the impact of overall global frequency response distortions is calculated. To estimate the impact for frequency response distortions independent of additive noise, the average noise spectrum density of the degraded signal over the silent frames of the reference signal is subtracted from the pitch loudness density of the degraded signal.

The resulting pitch loudness density of the degraded and reference is then averaged in each Bark band over all speech active frames for the reference and degraded file. The difference in pitch loudness density between these two densities is then integrated over the pitch to derive an average frequency response difference indicator. This indicator is combined with the rate of change over consecutive Bark pitch bands to obtain the indicator for quantifying the impact of frequency response distortions (FREQ).

An indicator is calculated from the average spectrum of the degraded signal over the silent frames of the reference signal for the impact of additive noise. The difference between the average pitch loudness density of the degraded signal over the silent frames and a zero reference pitch loudness density determines a noise loudness density function quantifying the impact of additive noise. This noise loudness density function is integrated over the pitch to derive an average noise impact indicator (NOISE).

For the impact of room reverberations, the energy over time function (ETC) is calculated from the reference and degraded time series. The ETC represents the impulse response envelope. First the loudest reflection is calculated by simply determining the maximum value of the ETC curve after the direct sound (sounds that arrive within 60 ms). Next a second loudest reflection is determined over the interval without the direct sound and reflections arriving within 100 ms from the loudest reflection. Then the third loudest reflection is determined over the interval without the direct sound and reflections arriving within 100 ms from the first and second loudest reflection. The energy of the three loudest reflections is then combined into a single reverb indicator (REVERB).

2.1.2.5 Scaling the Reference

The reference is now at the ideal level while the degraded signal is represented at a level coinciding with the play back level.

Before a comparison is made between the reference and degraded signal the overall level and small changes in local level are compensated to the extent that is necessary for the quality calculation. Global level equalization is carried out on the basis of average power of reference and degraded in the speech band between 400 and 3500 Hz.

The reference is scaled towards the degraded signal and the level difference impact compensated thereby. For correct modeling of slowly varying gain distortions a local scaling is carried out for level changes up to approximately 3 dB.

2.1.2.6 Partial Compensation of Original Pitch Power Density for Linear Frequency Response Distortions

To deal with SUT filtering which introduces non-audible linear frequency response distortions the reference signal is partially filtered with the SUT transfer characteristics.This is carried out by calculating the average power spectrum of the original and degraded pitch power densities over all speech active frames.

A partial compensation factor per Bark bin is calculated from the degraded to original spectrum ratio.

2.1.2.7 Modeling Masking Effects, Calculating Pitch Loudness Densities

Masking is modeled by calculating a smeared representation of the pitch power densities. Both time and frequency domain smearing is taken into account.

Time-frequency domain smearing uses a convolution approach. From this smeared representation the reference and degraded pitch power density representations are re-calculated suppressing low amplitude time-frequency components, which are partially masked by loud components in the time-frequency plane neighborhood. This suppression is implemented in two different manners, a subtraction of the smeared from the non-smeared representation and a division of the non-smeared by the smeared representation.

The resulting pitch power density representations are then transformed to pitch loudness density representations using a modified version of Zwicker's power law.

2.1.2.8 Noise Compensation in Reference and Degraded Signals

Low reference signal noise levels not affected by the SUT (e.g. a transparent system) will be attributed to the SUT by subjects and must be suppressed in the calculation.

This is carried out by calculating the average steady state noise loudness density of the reference signal $LX(f)_n$ over the super silent frames as a function of pitch.

This average noise loudness density is then partially subtracted from all pitch loudness density frames of the reference signal The result is an idealized internal representation of the reference signal.

Audible steady state noise in the degraded signal has lower impact than non-steady state noise. This applies for all noise levels. This effect's impact can be modeled by partially removing steady state noise from the degraded signal. It is performed by calculating the average steady state noise loudness density of the degraded signal $LY(f)_n$ frames for which the corresponding frame of the reference signal is classified as super silent as a pitch function.

The average noise loudness density is then partially subtracted from all pitch loudness density frames of the degraded signal. The partial compensation uses different strategies for low and high noise levels. For low noise levels the compensation is only marginal while the suppression becomes more aggressive for loud additive noise.

The result is an internal representation of the degraded signal with additive noise adapted to the subjective impact as observed in a listening test.

2.1.2.9 Calculation of Final Disturbance Densities

Two final disturbance densities are calculated. The first is derived from the difference between the ideal pitch-loudness-time and degraded pitch-loudness-time function. The second is derived from the ideal pitch-loudness-time and a degraded pitch-loudness-time function. The resulting disturbance density is referred to as the added density.

Two density flavors are calculated to deal with a large range of distortions. One derived from the difference between $LX_{ideal}(f)_n$ and $LY_{deg}(f)_n$ calculated with a perceptual model focused on small to medium distortions and one derived from the difference between $LX_{ideal}(f)_n$ and $LY_{deg}(f)_n$ calculated with a perceptual model focused on medium to big distortions. The switching between the two is performed with a first estimation from the disturbance focused on small to medium level of distortions.

In the next steps the final disturbance and added disturbance densities are compensated for severe amounts of specific distortions. Severe deviations of the optimal listening level are quantified by an indicator derived from the degraded signal level.

This global LEVEL indicator is also used to calculate MOS-LQO. Severe distortions introduced by frame repeats are quantified by an indicator derived from a correlation comparison of consecutive reference signal with consecutive degraded signal frames.

Severe deviations from the optimal timbre are quantified by an indicator derived from the upper frequency band to lower frequency band loudness ratio. Compensations are performed per frame on a global level.

The global level of timbre deviation is quantified in the FLATNESS indicator also used in MOS-LQO calculation. Severe noise level variations focusing the attention of subjects towards noise are quantified by a noise contrast indicator derived from the reference signal's silent parts.

Finally the disturbance and added disturbance densities are clipped to a maximum level and the disturbance and jumps variance in the loudness are used to compensate for specific disturbance time structures.

2.1.2.10 Final MOS-LQO POLQA calculation

The raw POLQA score is derived from the MOS like intermediate indicator using four different compensations:

• two compensations for specific time frequency characteristics of the disturbance. One calculated with an L_{511} (L_5-L_1-L_1) aggregation over frequency, spurts and time and one calculated with an L_{313} aggregation over frequency, spurts and time (see Figure 6)
• one compensation for very low presentation levels using the LEVEL indicator
• one compensation for big timbre distortions using the FLATNESS indicator

This mapping is trained on a large set of degradations, including degradations not belonging to the POLQA benchmark. These raw MOS scores are mostly already linearized by third order polynomial mapping used in calculating the MOS like intermediate indicator.

The raw POLQA MOS scores are finally mapped towards the MOS-LQO scores using a third order polynomial optimized for the POLQA database set.

In NB mode the maximum POLQA MOS-LQO score is 4.5 and in SWB mode 4.75. An important consequence of the idealization process is that under some circumstances, when the reference signal contains noise or the voice timbre is severely distorted, a transparent chain won't provide the maximum MOS score 4.5 in NB mode or 4.75 in SWB mode.

3 From PESQ to POLQA

3.1 Enhanced Features of POLQA

- Maintains correct scoring also at high background noise levels.
- Comparison of AMR (Adaptive Modulation Rate) codec used in GSM/3G and EVRC (Enhanced Variable Rate Codec) used in CDMA2000 possible.
- Representative scoring of reference signals.
- Effects of speech level in samples.
- SWB with 50 Hz – 14 kHz frequency range.
- Linear Frequency distortion sensitivity.

In NB the relative measurement uncertainty of POLQA measurements decreases by 27% compared to PESQ.

3.2 POLQA as Substitute for PESQ?

- Backward compatible MOS scale in NB for major speech codecs (AMR, GSM). PESQ can easily be migrated to POLQA, 1…4.5 for PESQ NB and POLQA NB.
- Extended MOS-scale for SWB takes HD-Voice into account: 1…4.75 for POLQA-SWB
- There are two MOS scales for all sample frequencies: Fs = 8 kHz → MOS NB Fs = 16 kHz → MOS SWB

4 Test Solution

The following schematic shows a possible POLQA test solution for LTE downlink and uplink with fading.

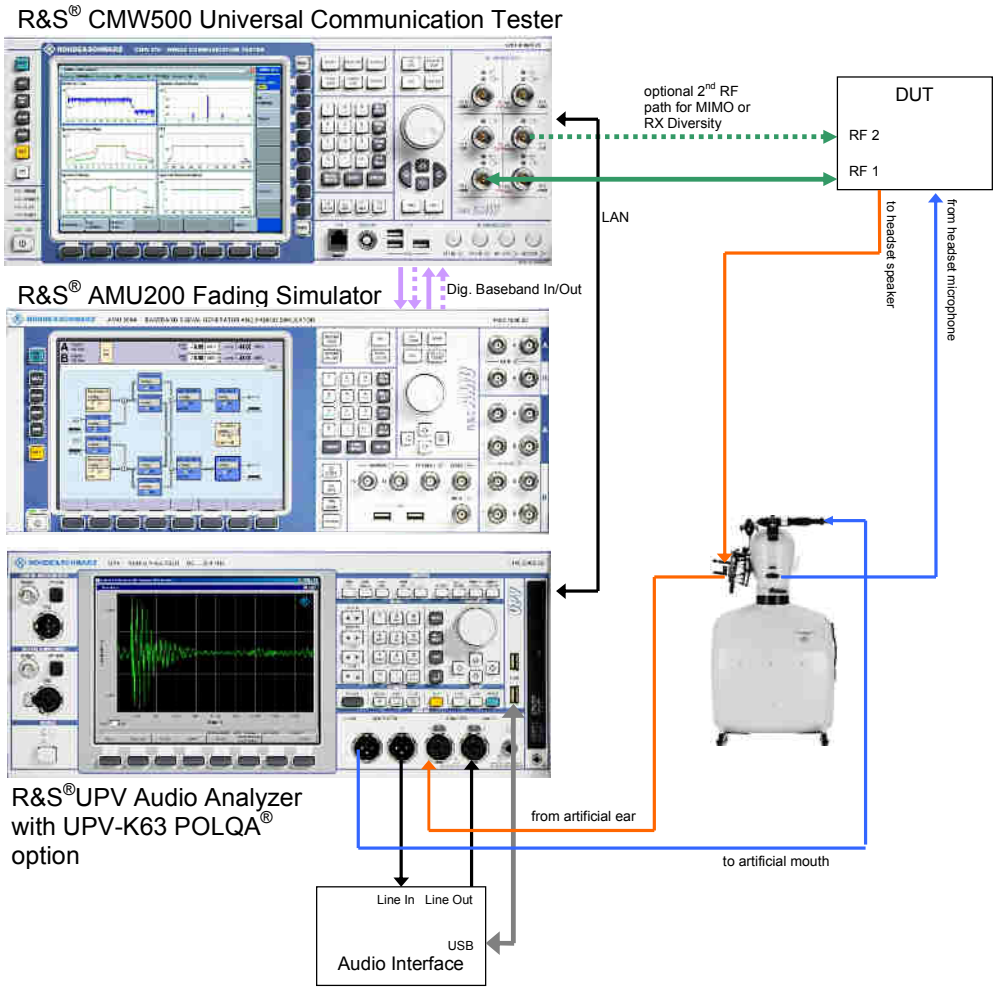


Figure 7: POLQA Test Configuration

The test configuration consists of a R&S®CMW500 Universal Communication Tester simulating a base station, an optional R&S®AMU200 Fading (Channel) Simulator and an R&S®UPV Audio Analyzer for performing the POLQA measurement.

The external audio interface is necessary for transferring the digitized audio data to a VoIP or IMS (Internet Media Service) server running either on the audio analyzer or external PC.

For acoustic measurements you may use a dummy head with artificial ear and mouth. For electrical measurements connect the DUT[1)] speaker output directly to the audio analyzer input and the microphone output directly to the audio analyzer output.

[1)] The DUT is the mobile device while SUT means the complete transmission chain of the audio signal (audio analyzer output to input).

4.1 Downlink POLQA Measurement

The simplified analog and digital routing for downlink POLQA measurements can be seen in Figure 8.

The audio analyzer generates an analog test signal (speech) which is fed to the line input of the audio interface. The analog signal is converted to a digital one and sent to the VoIP or IMS server running on either on the audio analyzer itself or an external PC. From there it is transferred to the communication tester via LAN.

The VoIP or IMS coded data packages are then transmitted via RF to the mobile (DUT) where they are decoded and converted into an analog signal at the earphone plug.

This (degraded) signal is fed to the audio analyzer's input and is needed together with the original (reference) signal calculating the MOS-LQO score with the POLQA algorithm.

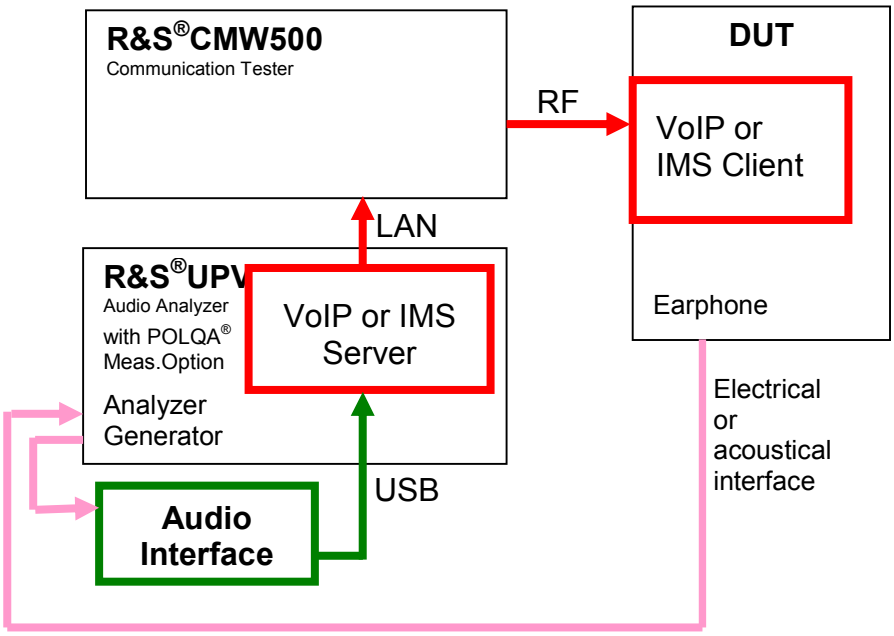


Figure 8: Downlink POLQA Measurement

4.2 Uplink POLQA Measurement

The simplified analog and digital routing for uplink POLQA measurements can be seen in Figure 9.

A reference speech signal is fed from the audio analyzer's generator to the mobile (DUT) microphone input. The signal is coded into VoIP or IMS packets and modulated to the RF carrier.

The IMS data packets are demodulated in the communication tester and fed to the audio analyzer or external PC via LAN where it is decoded by the VoIP or IMS server.

The audio interface converts the digital speech data into an audio signal which is fed to the audio analyzer input for the POLQA measurement.

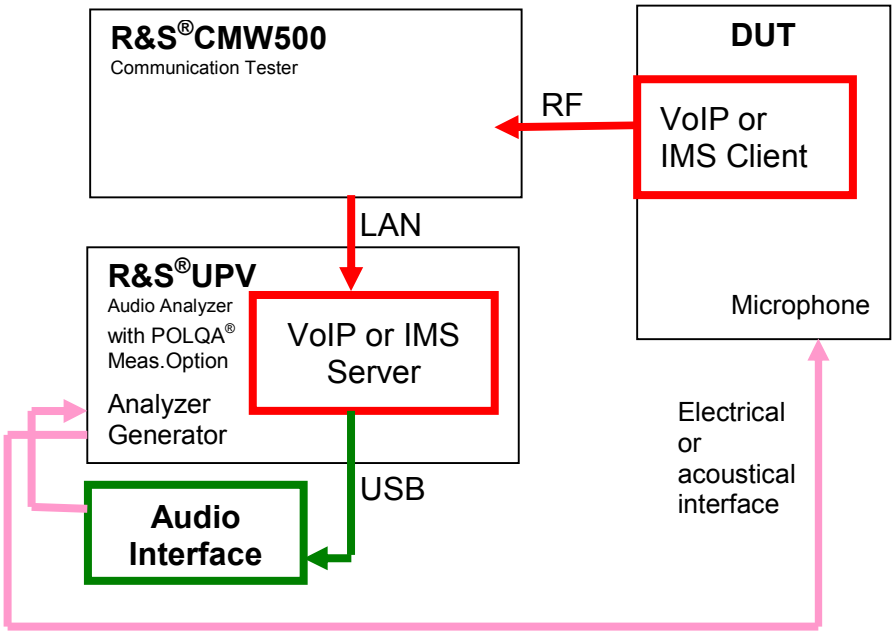


Figure 9: Uplink POLQA Measurement

5 Literature

[1] POLQA® Introduction – Jochim Pomy, OPTICOM GmbH
[2] Draft New Recommendation ITU-T P.863
[3] Recommendation ITU-T P.862
[4] United States Patent US 8,032,364 B1
[5] Application Note 1MA149 – "VoIP Measurements for WiMAX" - Ottmar Gerlach, Rohde&Schwarz GmbH & Co KG
[6] Application Note 1MA164 – "VoIP PESQ® Measurements for WiMAX with R&S®CMWrun" – Ottmar Gerlach, Rohde&Schwarz GmbH & Co KG
[7] Psychoacoustics – Facts and Models, E.Zwicker and H.Fastl, Springer Verlag 1990

6 Additional Information

Please contact TM-Applications@Rohde-Schwarz.com for comments and further suggestions.

7 Abbreviations

3G	- 3rd Mobile Generation
4G	- 4th Mobile Generation
AMR	- Adaptive Multirate Codec
AMR-NB	- Adaptive Multirate Codec – Narrow Band
AMR-WB	- Adaptive Multirate Codec – Wide Band
FFT	- Fast Fourier Transformation
IMS	- Internet Media Service protocol used in LTE
LQO	- Listening Quality, Objective
LTE	- Long Term Evolution
MOS	- Mean Option Score
NB	- Narrowband
PESQ	- Perceptual Evaluation of Speech Quality
POLQA	- Perceptual Objective Listening Quality Analysis
RMSE	- Root Mean Square Error
SNR	- Signal to Noise Ratio
SWB	- Super Wideband
UMTS	- Universal Mobile Telecommunications System
VoIP	- Voice over Internet Protocol
VoHSPA	- Voice over High Speed Packet Access
VoLTE	- Voice over Long Term Evolution

About Rohde & Schwarz
Rohde & Schwarz is an independent group of companies specializing in electronics. It is a leading supplier of solutions in the fields of test and measurement, broadcasting, radiomonitoring and radiolocation, as well as secure communications. Established more than 75 years ago, Rohde & Schwarz has a global presence and a dedicated service network in over 70 countries. Company headquarters are in Munich, Germany.

Environmental commitment
- Energy-efficient products
- Continuous improvement in environmental sustainability
- ISO 14001-certified environmental management system

Certified Quality System
ISO 9001

Regional contact
Europe, Africa, Middle East
+49 89 4129 12345
customersupport@rohde-schwarz.com

North America
1-888-TEST-RSA (1-888-837-8772)
customer.support@rsa.rohde-schwarz.com

Latin America
+1-410-910-7988
customersupport.la@rohde-schwarz.com

Asia/Pacific
+65 65 13 04 88
customersupport.asia@rohde-schwarz.com

Rohde & Schwarz GmbH & Co. KG
Mühldorfstraße 15 | D - 81671 München
Phone + 49 89 4129 - 0 | Fax + 49 89 4129 – 13777

www.rohde-schwarz.com