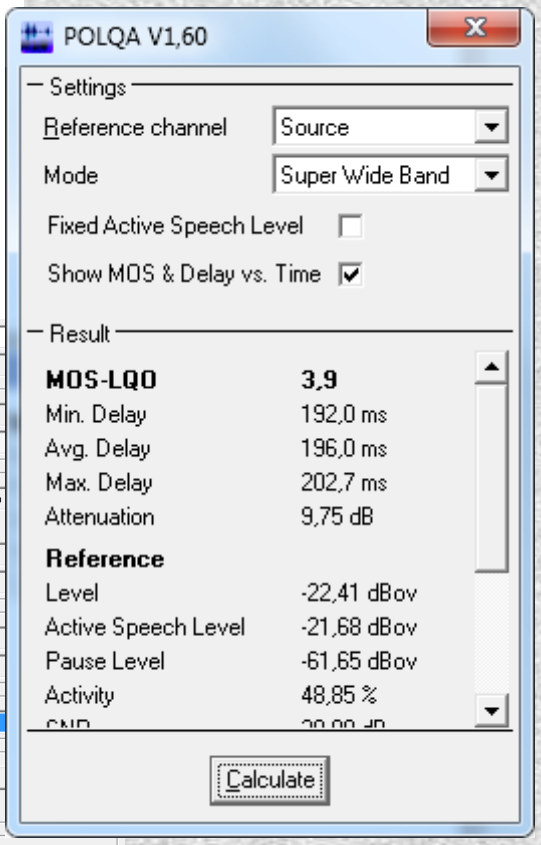
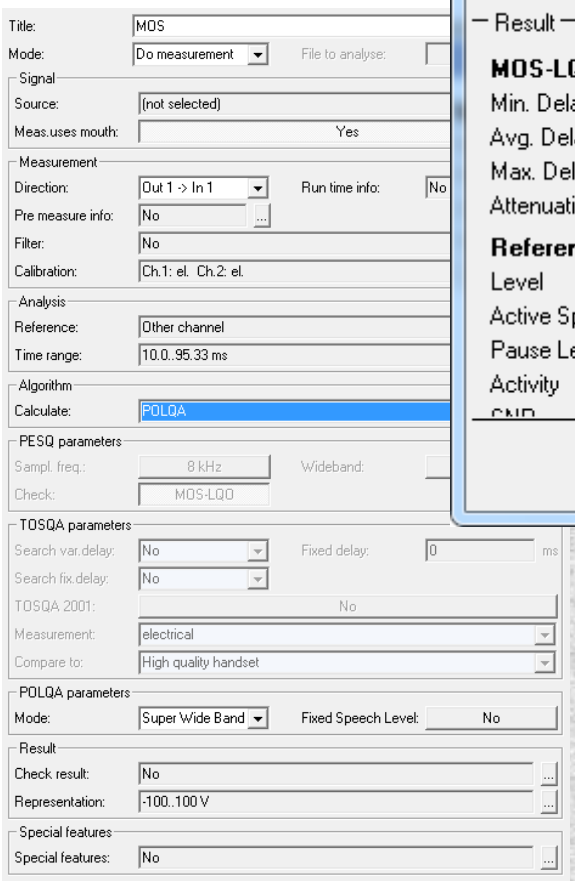
POLQA Application Guide

POLQA®





HEAD acoustics Documentation

This documentation is a copyrighted work by HEAD acoustics GmbH and contains copyrighted material provided by kind permission of OPTICOM GmbH.

The information and artwork in this documentation are the property of HEAD acoustics GmbH and OPTICOM GmbH and shall not be reproduced or copied or used in whole or in part without written permission.

Copyright 2012 by HEAD acoustics GmbH.
All Rights Reserved.

AACHENHEAD® is a registered trademark.
HEAD acoustics® is a registered trademark.

POLQA™ is a registered trademarks of OPTICOM GmbH;
PESQ™ is a registered trademark of OPTICOM GmbH and Psytechnics Ltd..
This information may be subject to change.

All other brand and product names are trademarks and/or registered trademarks of their respective owners.

HEAD acoustics GmbH

Ebertstraße 30a
D-52134 Herzogenrath
Tel: +49 (0)2407-577-0
Fax: +49 (0)2407-577-99
E-mail: telecom@head-acoustics.de
WEB: www.head-acoustics.de

POLQA Application Guide

Contents

1. INTRODUCTION 5

2. LEGAL NOTES 5

3. POLQA™ SMD SETTINGS 6

9. OBJECTIVE LISTENING QUALITY ASSESSMENT WITH POLQA 8

9.1. POLQA MEASUREMENT 8
9.2. DESCRIPTION OF THE POLQA ALGORITHM 8
9.2.1. The Core Model 9
9.3. OPERATING MODES OF POLQA 13
9.4. SOURCE MATERIAL 14
9.4.1. Choice of Source Material 14
9.4.2. Temporal structure and duration of source material 15
9.4.3. Required Filtering, Level Calibration and Sample Rates 16
9.5. IMPORTANT DIFFERENCES BETWEEN PESQ AND POLQA 17
9.5.1. Differences Between the POLQA and PESQ Core Models 18
9.5.2. New Application Areas 21
9.6. PERCEPTUAL RESULTS OBTAINED FROM POLQA 22
9.6.1. MOS-LQO 22
9.6.2. G.107 R-Factor / I_e Value 23
9.6.3. Disturbance Density 23
9.7. NON-PERCEPTUAL RESULTS HELPFUL FOR CAUSE ANALYSIS 23
9.7.1. Delay/Latency 23
9.7.2. Delay Jitter 23
9.7.3. Attenuation 24
9.7.4. Level and Background Noise Measurements 24
9.7.5. Signal to Noise Ratio (SNR) 25
9.7.6. Active Speech Ratio (ASR) 25
9.7.7. Pitch 25
9.7.8. Short Term Spectra 25
9.8. REPORTING AND AVERAGING OF RESULTS 26
9.9. ACCURACY OF POLQA RESULTS (P.863) 26
9.10. LIMITATIONS OF POLQA 26
9.11. HOW TO ASSESS "SIGNAL ENHANCERS" 27
9.12. HOW TO DEAL WITH COMFORT NOISE INSERTION 27
9.13. CONSIDERATIONS REGARDING THE DATA ACQUISITION FOR POLQA 28
9.13.1. Synchronisation of the Sample Clocks 28
9.13.2. Synchronisation of Play and Record 28
9.13.3. Control of Playback and Recording Levels 28

© 2012 HEAD acoustics

© 2012 HEAD acoustics

1. Introduction

This documentation is intended to provide the user of the HEAD acoustics communication analysis system ACQUA with the necessary background information to understand and apply the POLQA algorithm provided by license agreement with OPTICOM GmbH (cf. Legal Notes in chapter 2) and implemented into ACQUA as software option ACOPT 30 (Code 6857).

The fourth chapter is provided by kind permission of OPTICOM GmbH and is identical with chapter 9 of OPTICOM's "User Manual POLQA OEM Library" (Version 1.6). Therefore, the numbering of the fourth chapter starts with 9 instead of 4.

Neither OPTICOM GmbH nor HEAD acoustics GmbH grant any guarantees or assume any liabilities with regard to the contents of this documentation.

2. Legal Notes

Perceptual Objective Listening Quality Analysis (POLQA) according to ITU-T Recommendation P.863 included in this product is protected by copyright and by European, US and other international patents and patent applications and is provided under license from
OPTICOM Dipl.-Ing. M. Keyhl GmbH, Erlangen, Germany, 2011 - www.opticom.de
POLQA™ is a registered trademark of OPTICOM GmbH. Used by permission.
(C) 2011 by the POLQA Coalition of OPTICOM GmbH, Germany - SwissQual AG, Switzerland - KPN, The Netherlands - TNO, The Netherlands
www.polqa.info

- Additional copying of the POLQA software in whole or in part, other than is essential for the proper operation of the POLQA software or for normal security back-up purposes is prohibited.
- Modifying, translating, reverse-engineering or decompiling the POLQA software except to the extent permitted by law is prohibited.
- The acknowledgement of the rights in the POLQA software shall not be removed from the POLQA software or any installation of it.

Notes on Trademarks:
POLQA™ and the OPTICOM logo are registered trademarks of OPTICOM GmbH;
PESQ™ is a registered trademark of OPTICOM GmbH and Psytechnics Ltd..
This information may be subject to change.
All other brand and product names are trademarks and/or registered trademarks of their respective owners.

© 2012 HEAD acoustics

3. POLQA™ SMD Settings

Signal / Source: Speech material should be selected according to P.863 chapter 8. A signal consisting of two male and two female speakers, two sentences each, are recommended.

Analysis / Reference: Here the clean reference which is used for the POLQA calculation can be selected. Usually the source signal is also used as the reference file. But POLQA always requires a reference without any spectral pre-shaping. When using a pre-filtered source signal (for example for electrical network insertion), an unfiltered reference signal can be selected for the calculation.

POLQA Parameters: A detailed explanation of the basic operational modes of POLQA is given in chapter 1 of P.863. For measuring typical setups in telecommunication applications, the POLQA integration in ACQUA simplifies the use of the bandwidth modes and resampling issues.

Mode:

Narrow Band el.: This mode uses the narrowband mode of POLQA. Internally all signals are resampled to 8 kHz. As stated in P.863, the POLQA narrowband mode is only valid for electrical recordings without any acoustical interfaces (neither sending nor receiving direction). A typical use would be the measurement of IP gateways and/or packet loss concealment in such devices.

Narrow Band ac.: This switch shall be selected when measuring narrowband terminals with at least one acoustical interface (sending or receiving direction), which is not covered by the POLQA narrowband mode. In this mode, the super-wideband mode of POLQA is internally used, but after a resampling to 8 kHz. This method for terminals is described in [1]. In contrast to TOSQA for example, it is not necessary to define how the signal was recorded (electrically or acoustically).

Super Wide Band: This mode shall be used for all other cases, which are not covered by the two modes mentioned before (e.g. WB device in SND and RCV, all acoustic and electric measurements)

Fixed Speech Level: In other algorithms like TOSQA or PESQ, the degraded speech signal is always calibrated to a common active speech level (e.g. 79dB SPL) which is a typical preprocessing for listening tests. In contrast, POLQA is also able to take level differences at listener's side into account. With this switch, the level calibration to the default listening level is activated. Note that this preprocessing step is not provided by P.863 itself. When keeping strictly to the standard, this switch has to be turned off and the user has to check if the active speech level of a measurement reaches a sufficient level.

Depending on the bandwidth mode described before, several aspects have to be regarded:

Narrow Band el.: In this mode, it is strongly recommended to activate the "Fixed Speech Level" switch. The POLQA narrowband mode which is used here is in general able to handle level differences, but it was never validated with these kinds of conditions.

Narrow Band ac.: According to [1], the fixed active speech level should be activated.

Super Wide Band: The fixed active speech level can be activated if wanted.

© 2012 HEAD acoustics

In addition, an automatic sampling rate conversion is always applied to all input signals if necessary. By default, P.863 only supports 8 kHz in narrowband mode and 48 kHz in super-wideband mode. An eventual resampling preprocessing is not part of the P.863 calculation.

[1] "Implementor's Guide to P863 – Prediction of acoustically recorded narrowband speech", Jens Berger, Q9/12 Meeting, Geneva, 31 October - 9 November 2011.

© 2012 HEAD acoustics
Rev0 (03/2012)

9. Objective Listening Quality Assessment with POLQA

9.1. POLQA Measurement

When PSQM was standardized as P.861, the scope of the standard was at that time the assessment of state of the art codecs as they were mainly used for mobile transmission, like GSM. VoIP was not yet a topic at this time. The requirements for measurement equipment have changed dramatically since then. As a consequence, the ITU introduced the P.862 (PESQ) standard in order to cope with the new demands arising from modern networks like VoIP. With these networks the measurement algorithm has to deal with much higher distortions as with GSM codecs, but maybe the most eminent factor is that the delay between the reference and the test signal is not constant anymore, instead it is time varying. However development of network transmission schemes and codecs is an ongoing process. Additional requirements arise for modern measurement schemes. To stay abreast of these changes the new ITU standard P.863/POLQA was developed. POLQA can handle effects caused by new voice services like stretching and compression of speech signals in the time domain. POLQA improves the quality prediction for new and old codecs and allows the direct comparison of AMR and EVRC. POLQA combines an excellent psychoacoustic and cognitive model with a new time alignment algorithm that perfectly handles varying delays.

9.2. Description of the POLQA Algorithm

A general overview on the algorithm is shown in Figure 9-1. The inputs to the algorithm are two waveforms represented by two data vectors containing 16 bit PCM samples. The first vector contains the samples of the (undistorted) reference signal, whereas the second vector contains the samples of the degraded signal. The POLQA algorithm consists of a temporal alignment block, a sample rate estimator of a sample rate converter, which is used to compensate for differences in the sample rate of the input signals, 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 resampling 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 core model, which calculates the perceptibility as well as the annoyance of the distortions and maps them to a MOS scale.

© 2012 HEAD acoustics

A much more detailed and comprehensive description of the algorithm can be found in ITU-T P.863 [ITUT863]. The next few sections are only intended to give an overview on the basics of POLQA's internal structure.

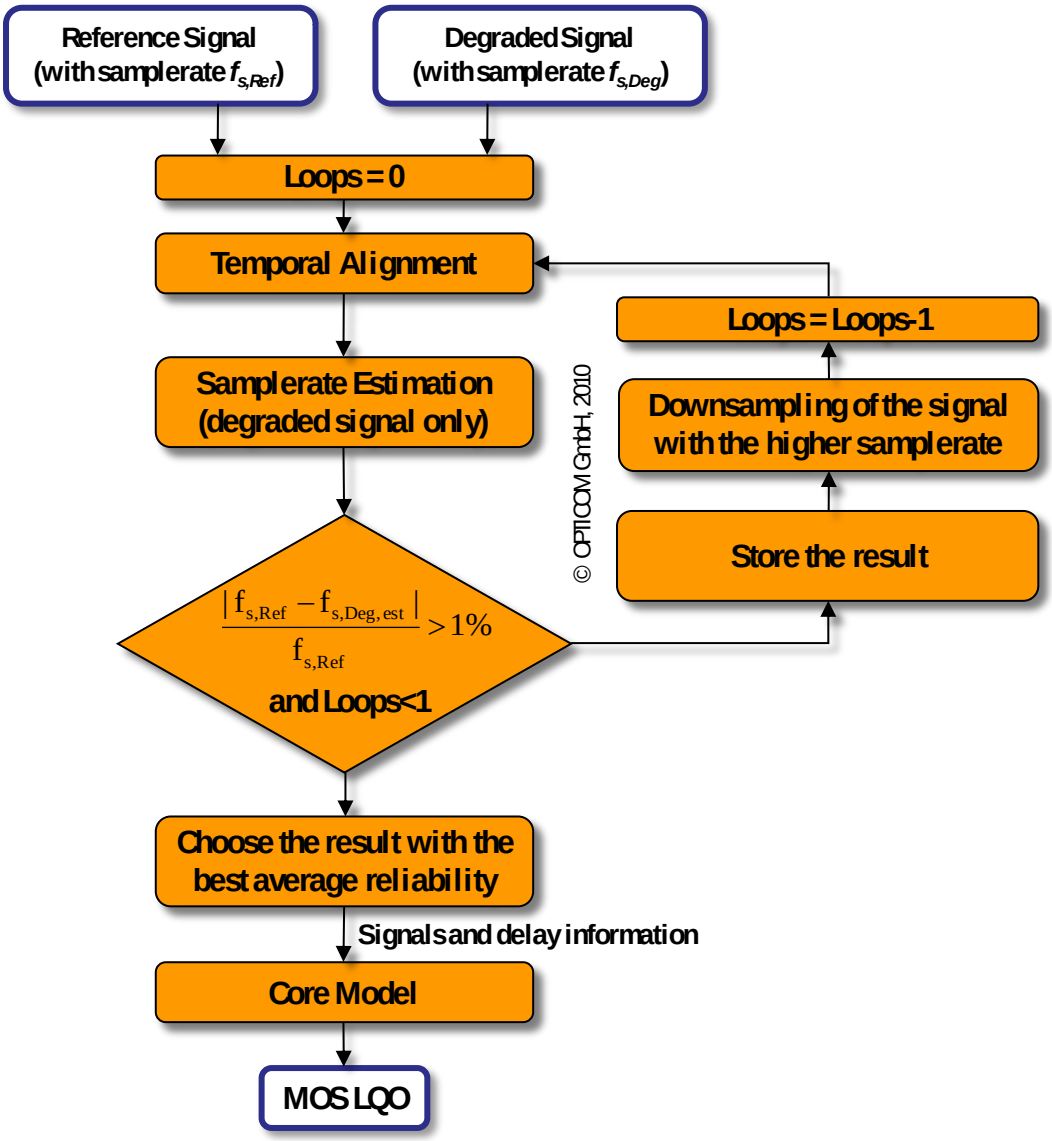


Figure 9-1, general overview of POLQA

9.2.1. The Core Model

An overview of the core model is given in Figure 9-3. The main element here is the perceptual model which is calculated four times using different parameters in order to cope with different major distortion types. Those distortion types can be split into additive distortions and subtracted distortions. For both types a further distinction is made between very strong and weaker effects. The inputs to the perceptual models are waveforms and the delay information. The output is the "Disturbance Density", which is a measure for the perceptibility of

© 2012 HEAD acoustics
Rev0 (03/2012)

distortions in the signals. The perceptual model for the main branch also produces indicators for Frequency distortions, Noise and Reverberation distortions. A subsequent switch which is triggered by a detector for very strong distortions reduces the four Disturbance Density values down to two, one for added and one for subtracted distortions. So far the Disturbance Density is an indicator for the perceptibility of distortions only and cognitive effects are not yet taken into account. Cognitive aspects are however important when human beings are asked to score the quality of what they can perceive. Essentially they convert the perceptibility measure Disturbance Density into an annoyance measure. This conversion is performed by correcting the Disturbance Density values for situations with:

- Significant level variations
- Many frame repetitions
- Strong timbre
- Spectral flatness
- Noise switching during speech pauses
- Many delay variations
- Strong variations of the Disturbance Density over time
- Strong variations of the loudness of the signals

Two further indicators, one for spectral flatness ad one for level variations are also calculated in this step.

So far all operations were performed on frames with a duration of approximately 32 and 43ms duration (depending on the sample rate and using an overlap of 50%) and for each Bark band separately. In a final step all indicators are integrated over time and frequency in order to compute the final MOS LQO value.

© 2012 HEAD acoustics
Rev0 (03/2012)

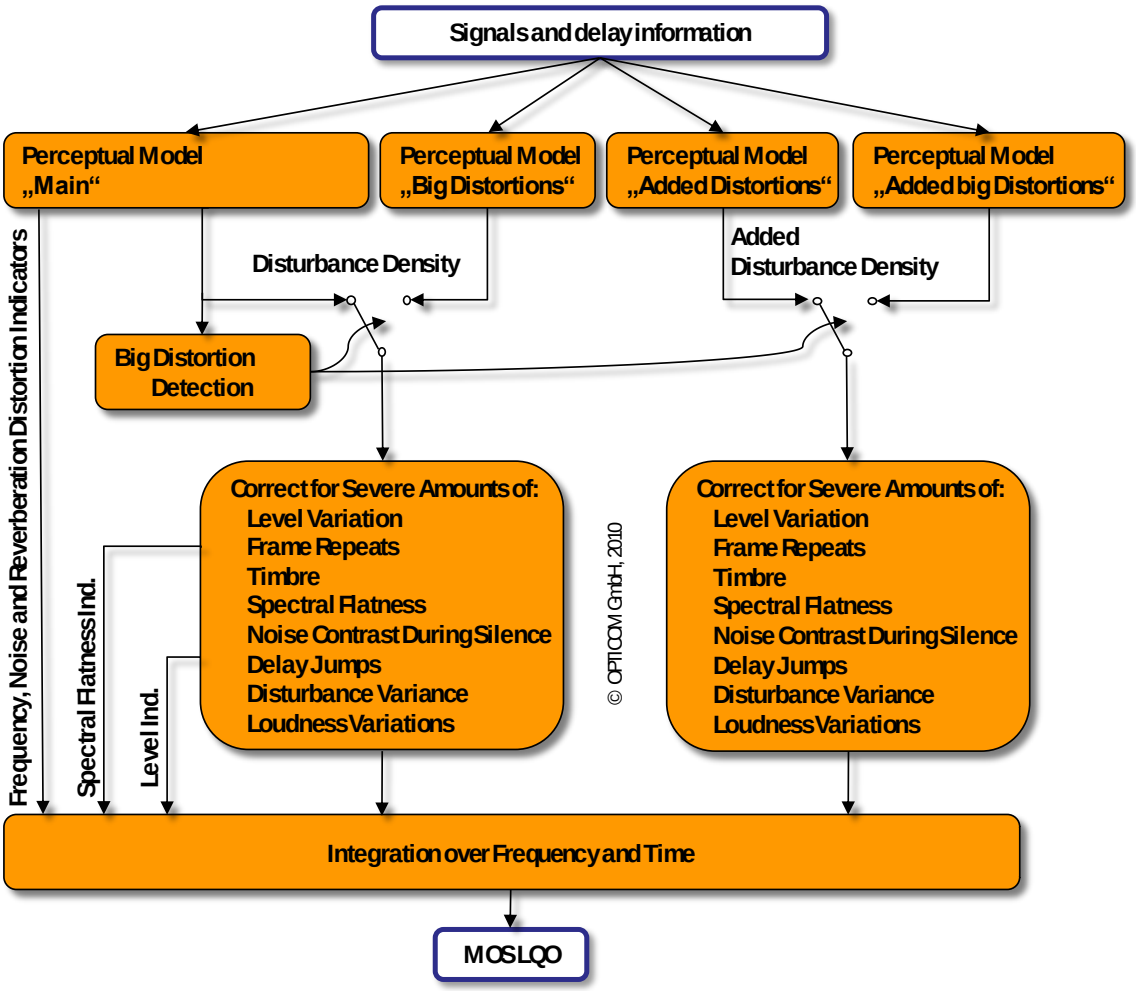


Figure 9-2, overview of the POLQA core model. Calculation of final disturbance densities from the four different variants of the internal representations.

The Perceptual Model

The key concept behind the perceptual model (Figure 9-3) is “Idealisation”. The idea behind this is, that POLQA is supposed to simulate ACR tests. In an ACR test however, subjects have no comparison to the actual reference signal when they score a speech signal. Instead, it is assumed that subjects have an understanding of what an ideal signal sounds like and they use this as their own reference. Consequently, if they are asked to score a reference signal which is not absolutely perfect (e.g. it has the wrong volume or contains too much timbre, noise or reverberation), it will be scored worse than perfect. In its idealization step POLQA therefore corrects small imperfections of the reference signals in order to derive the same ideal reference for the comparison to the degraded signal as human subjects would use in their minds. Similar to the idealization of the reference signal, some distortions present in the degraded signal which are hardly perceptible in an ACR test will be partially compensated (e.g. small pitch shifts, linear frequency distortions).

© 2012 HEAD acoustics
Rev0 (03/2012)

The perceptual model starts with scaling the reference signal to an ideal average active speech level of approximately -26dBov. No such scaling is performed on the degraded signal. It is assumed that any deviation of the level of the degraded signal from the ideal -26dBov is to be scored as a degradation of the signal.

Next, the spectra of both signals are computed using an FFT with 50% overlapping frames with a duration of between 32ms and 43ms duration (depending on the sample rate). Subsequently small pitch shifts of the degraded signal will be eliminated ("Frequency Dewarping"). Now, the spectra will be transformed to a psychoacoustically motivated pitch scale, by combining individual spectral lines (FFT bins) to so-called critical bands. The pitch scale used is similar to the Bark scale with an average resolution of 0.3 Bark per band. The result is the "Pitch Power Density". At this stage the first three distortion indicators for frequency response distortions, additive noise and room reverberations are calculated.

After this, the excitation of each band is derived. This includes the modeling of masking effects in the frequency as well as in the temporal domain. The result is for each frame of each signal a head-internal representation which indicates roughly how loud each frequency component would be perceived.

Now, a further idealization step of the reference signal takes place by filtering out excessive timbre and low level stationary noise. At the same time, linear frequency distortions and stationary noise are partially removed from the degraded signal.

A subtraction of the idealized excitations finally leads to the "Distortion Density", which is measure for the audibility of distortions.

© 2012 HEAD acoustics

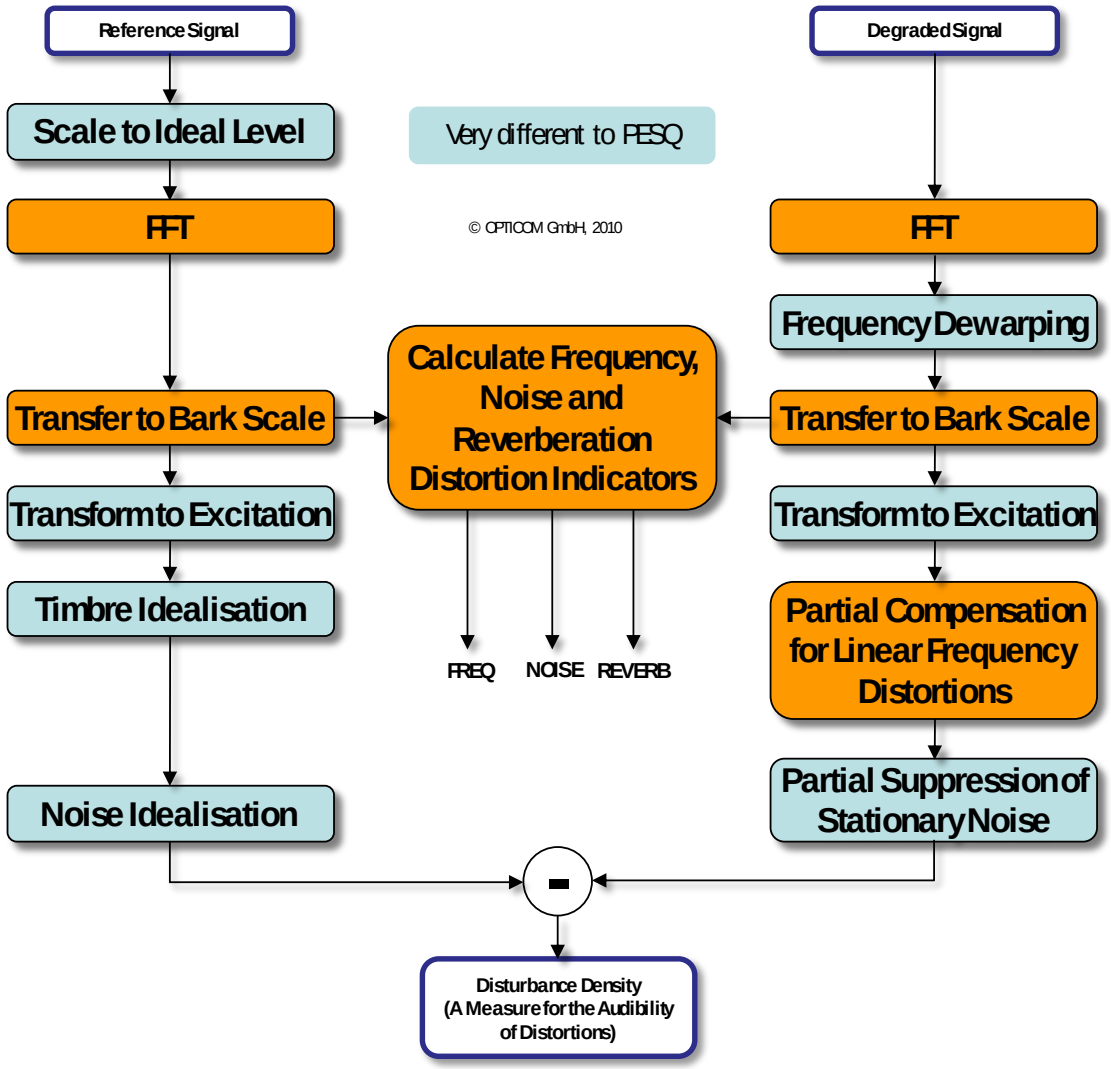


Figure 9-3, overview of the POLQA perceptual model, which performs the calculation of the Disturbance Density. Four different variants this model are calculated, each focussed on a specific set of distortions ("Main", "Big Distortions", "Added Distortions" and "Added big Distortions").

9.3. Operating modes of POLQA

OPTICOM's POLQA implementation offers two operating modes, which allow the assessment of super-wideband and narrowband speech signals.

By choosing the super-wideband mode a flat input filter and a different mapping function to the MOS scale than narrow band speech is used. Consequently in super-wideband mode, bandwidth limitations are considered as degradations and scored accordingly. The listening quality is modeled as perceived by a human listener using a diffuse field equalized headphone with diotic (a mono signal is presented to both ears) presentation. If POLQA is used in super-wideband mode, then the usage of super-wideband reference signals is mandatory.

© 2012 HEAD acoustics
Rev0 (03/2012)

In narrowband mode the received speech signal is compared to a narrowband reference signal. Consequently, normal telephone bandwidth limitations are not considered as severe degradations and are scored much less than in super-wideband mode. The listening quality is modeled as perceived by a human listener using a loosely coupled IRS type handset at one ear.

When interpreting POLQA results it is important to know which version of the algorithm was used. Although the narrowband and the wideband version of POLQA will generate MOS scores on the same five point scale, it is strictly forbidden to mix super-wideband and narrowband results. If super-wideband networks have to be compared to narrowband networks, then the super-wideband version of POLQA has to be used in both cases. It is also not allowed to use bandwidth limited reference signals in super-wideband mode.

The ITU Rec. P.863 strictly requires a sample rate of 48kHz for super-wideband measurements. For narrowband applications, 8, 16 and 48kHz sample rate are supported. As an extension to this, the OPTICOM POLQA Library provides the capability to resample the input signals to those sample rates if required.

9.4. Source Material

It is important that test signals for use with POLQA are representative of the real signals carried by communications networks. Networks may treat voice and silence differently and coding algorithms are often highly optimized for voice – and so may give meaningless results if they are tested with signals that do not contain the key temporal and spectral properties of voice. Further pre-processing is often necessary to take account of filtering in the send path of a handset, and to ensure that power levels are set to an appropriate range.

9.4.1. Choice of Source Material

Reference speech or reference signal is the original speech signal without any degradation. This should be recorded and stored in conformance with ITU-T Recommendation P.830 [11]. In the case of an acoustical sending path, this signal is used for feeding the artificial mouth. This speech signal is used by the POLQA algorithm as a reference against which the effects of the system under test are revealed. Examples of such speech files are to be found in ITU-T Recommendation P.501 [6].

POLQA was tested with human speech material. For a consistent speech quality prediction, active speech parts and speech pauses are required in the speech sample. It is recommended to use typical spoken sentences with typical syllable and word structures. It is not recommended using single word samples only (i.e. counting).

The reference signals should be leveled to -26 dBov rel. to OVL. Other signal levels will be accepted by POLQA to and internally equalized; however different levels than -26 dBov could cause amplitude clipping or lower SNR.

© 2012 HEAD acoustics

The use of artificial speech signals and concatenated real speech test signals is recommended only if they represent the temporal structure (including silent intervals). Artificial speech test signals can be prepared in several ways. A concatenated real speech test signal may be constructed by concatenating short fragments (e.g. one second) of real speech while retaining a representative structure of speech and silence. Alternatively, a phonetic approach may be used to produce a minimally redundant artificial speech signal which is representative of both the temporal and phonetic structure of a large corpus of natural speech. Test signals should be representative of both male and female talkers. It should however be noted that for artificial speech a direct subjective assessment is not possible as subjects are not able to determine the quality correctly when artificial speech samples as described above are used. Therefore the use of real speech samples is preferred to also allow subjective assessment of the test data. No tests were performed yet for POLQA using artificial speech.

If natural speech recordings are used, the guidelines given in clause 7 of P.830 [ITUT830] should be followed, and it is recommended that a minimum of two male talkers and two female talkers be used for each testing condition. If talker dependency is to be tested as a factor in its own right, it is recommended that more talkers be used: eight male, eight female and eight children.

9.4.2. Temporal structure and duration of source material

Test signals should include speech bursts separated by silent periods of at least 1s but not more than 2s to be representative of natural pauses in speech. As a guide, 1s to 3s is a typical duration of a speech burst, although this varies considerably between languages. A minimum of 3s of active speech should be included in each file. A sufficiently long silent period should be included between speech bursts as certain types of voice activity detectors are sensitive only to silent periods that are longer than 200 ms.

Most of the experiments used in calibrating and validating POLQA contained pairs of sentences separated by silence, totaling 8s in duration; in some cases three or four sentences were used, with slightly longer recordings (up to 12s). Recordings made for use with POLQA should be of similar length and structure.

Thus, if a condition is to be tested over a long period, it is most appropriate to make a number of separate recordings of around 8s to 20s of speech and process each file separately with POLQA. This has additional benefits: if the same original recording is used in every case, time variations in the quality of the condition will be very apparent; alternatively, several different talkers and/or source recordings can be used, allowing more accurate measurement of talker or material dependence in the condition.

Note that the non-linear averaging process in POLQA means that the average score over a set of files will usually not equal the score of a single concatenated version of the same set of files.

© 2012 HEAD acoustics

9.4.3. Required Filtering, Level Calibration and Sample Rates

Signal Levels and Filtering in Super-wideband Mode

In its super-wideband operational mode POLQA always requires a super-wideband reference signal. It has to be provided in mono and in a sampling frequency that allows the full spectral range of super-wideband. The signal has to be pre-filtered by a 50 to 14'000 Hz band-pass filter (according to ITU-T Rec. G.191, 14 kHz).

The only sample rate supported by P.863/POLQA in super-wideband mode is 48kHz. OPTICOM's POLQA library however provides an option to automatically up sample signals with lower sample rates to the required 48kHz. Please note, the minimum sample rate for the reference signal is 32kHz. Otherwise the requirements regarding the spectral content of up to 14000Hz cannot be met.

For all signals delivered for super-wideband operational mode, a digital level of -26 dBov (obtained in accordance with ITU-T Rec. P.56) corresponds to the nominal presentation level (73 dB (A) SPL in case of diotic presentation). The actual presentation level can be directly derived from the P.56 level of the degraded signal (e.g., a level of -34 dBov corresponds to a presentation level 8 dB below the nominal level).

The nominal level for POLQA reference signals in super-wideband operational mode is therefore -26 dBov.

As far as degraded signals are concerned, potential differences between the recording level and the presentation level are allowed and are part of the test condition. These level differences have to be restricted to a range of +5 dB to -20 dB relative to the nominal level which is again -26dBov (this results in a range from -21 dBov to -46 dBov).

Signal Levels and Filtering in Narrowband Mode

In narrowband operational mode, the chosen sampling frequency has to allow the presentation of the full spectral range of narrow-band telephony (50 to 3800 Hz). The reference signal must not be limited in bandwidth below 3.8 kHz. The reference signal used by POLQA must not be IRS pre-filtered. This clean reference signal should be used for POLQA. Before this signal is sent through the system under test an IRSsend filter should be applied. This filter forms part of the channel under test.

A sampling rate of 8kHz is preferred for narrow-band operational mode, but 48kHz is also supported. OPTICOM's POLQA library however provides an option to automatically resample signals to 8kHz if they are not presented at that sample-rate. The OPTICOM POLQA Library does also support 16kHz, but the this not part of the standard. Please note, that signals provided at 16kHz sample rate will also not be resampled if the optional automatic sample rate conversion is switched on.

The narrowband operational mode of the POLQA algorithm pre-assumes that a digital level of -26 dBov corresponds to a presentation level of 79 dB(A) SPL at the ERP (ear reference

© 2012 HEAD acoustics

point) in monotic presentation. Deviations from this nominal level are taken into account as non-optimal presentation level.

Optional Automatic Level Alignment

Since the level of a speech signal is a very important factor in human perception as well as in modern telecommunication networks, POLQA takes them into account when calculating the MOS score. For this to work however, some very strict requirements on measurement systems regarding the control of record and playback levels have to be met. In most cases it will be required to either amplify or attenuate the signals in the analog domain, or to use AD/DA converters with a resolution >= 24Bits if the level control shall be performed digitally.

For systems and applications where the signal levels cannot be controlled, the OPICOM POLQA library provides an optional automatic level alignment. This level alignment will scale both input signals to an average active speech level of approximately -26dBov. Please note that this is only a workaround and cannot replace correct tuning of the hardware. This feature is enabled by setting a flag during the initialization of the Library.

Sample Rates

The ITU Rec. P.863 specifies very few sample rates at which POLQA can operate (8 and 48kHz for NB and 48kHz only for SWB measurements). To ease the application and implementation of POLQA, the OPTICOM POLQA Library can automatically resample the input signals the correct required sample rate if the input signals are provided with any non-standard sample rate. This feature is enabled by setting a flag during the initialization of the Library. The resulting sample rate used for the processing depends on the selected mode (NB/SWB) and the sample rate in which the signals are presented. In NB mode the behavior is such that if both signals are of the same sample rate and this sample rate is 8, 16 or 48kHz, no resampling is performed. In all other cases, both signals will be resampled to 8kHz. In SWB mode all signals will be resampled to 48kHz.

9.5. Important Differences Between PESQ and POLQA

This chapter focuses on the main differences between PESQ and POLQA which have a direct impact on the application and the interpretation of the measurement results. This section is directed to experienced users which have dealt with PESQ in the past and are now investigating the usage of POLQA.

© 2012 HEAD acoustics

9.5.1. Differences Between the POLQA and PESQ Core Models

Transparency - or why Comparing a Signal to Itself gives Unexpected Results

In PESQ, a comparison of a file with itself will always result in a perfect MOS-LQO. For POLQA this is not necessarily the case.

The background to this lies in the behavior of listeners in a subjective experiment. Listeners do also not score all reference signals as perfect. Especially, timbre in voices or background noise during the silent intervals is critical for listeners. If any of those exceeds a certain amount, then listeners will score the signal below perfect. In order to cope with this effect, POLQA performs an idealization of the reference signal before it is compared to the not idealized degraded signal.

Due to this asymmetry in the processing of the reference and the degraded signal, the MOS measured when comparing the reference signal to itself may be slightly smaller than expected, but this is in line with subjective experiments, where subjects as well make the comparison to an ideal which exists in their minds only (there is no direct comparison of two files in the common ACR method). This principle of idealization is also one of the reasons for the improved accuracy of POLQA compared to PESQ.

One way to avoid confusing results which might result from this behavior is to use selected reference signals which *sound* perfect. Such signals can often be constructed from existing reference files by filtering them slightly. For samples consisting of concatenated voices this is however difficult due to the timbre variations between the individual voices. Comparing an ideal reference signal with itself will then lead to the expected maximum MOS.

Effects of Level

PESQ is almost completely insensitive to level differences between the reference and the degraded signal. Also, the perceptual effect of level variations is very often under predicted by PESQ. Both factors are corrected in POLQA, which scores these effects in the same way as they are perceived in subjective listening only tests. Consequently, some rules have to be followed in order to obtain correct results:

- POLQA strictly assumes that **the reference signals** used have an active speech level of -26dBov (measured acc. to P.56) and are presented at a nominal level of 73dB SPL at the reference point of **both** ears in super- wideband mode (diotic presentation!) and 79dB SPL at the reference point of **one** ear in narrowband mode (monotic presentation!).

- In super-wideband mode, the **degraded signals** used for the validation of POLQA had an active speech level in the range between -21dBov to -46dBov for which POLQA worked reliably. A level of -26dBov would represent the ideal normal listening level. Further investigations in the future may extend the operating range.

© 2012 HEAD acoustics
Rev0 (03/2012)

- In narrowband mode, the **degraded signals** should be scaled towards 26dBov. Further investigations on the behavior of POLQA for severe deviations from this level in narrowband mode have not yet been performed.

Please note that the frequently used tool "Cool Edit Pro" does not measure according to P.56 and that the resulting levels may vary by up to approximately 3dB.

Under all circumstances the correct scaling between digital levels and sound pressure levels as described for the reference signals above should be maintained.

Narrowband – Wideband – Super-Wideband

For PESQ, there are two operating modes, narrowband (NB, 300-3400Hz) and wideband (100-7000Hz). For POLQA, also two modes exist, but they cover narrowband and super-wideband (SWB, 50-14000Hz). This means, there is an overlap between the NB modes, but no direct correspondence exists for the WB respectively SWB modes.

For NB, the results obtained by either algorithm are usually similar, but they will hardly ever be exactly the same. Comparing individual results from PESQ measurements to results obtained by POLQA should clearly be avoided and certainly those comparisons make no sense at all for applications for which PESQ fails.

PESQ's WB mode is completely replaced by the SWB mode of POLQA. The major consequence of this is that in order to correctly apply POLQA for WB scenarios, SWB reference files must be used. The achieved scores will then be on a slightly different range of the MOS scale as they used to be for PESQ in WB mode. This is caused by the different usage of the five point scale which is applied in all modes and which must now accommodate for scores given to NB, WB and SWB conditions in the same experiment. Nevertheless, care was taken to place the scores obtained by POLQA for typical WB conditions in a *similar* range as those obtained by PESQ WB for the same conditions. If a more exact match of the scales is required, then a transformation of one of the scales must be used. The maximum score which can be achieved by using POLQA to compare a SWB reference signal to its WB equivalent is around 4.5, the same as with PESQ WB.

Using WB reference signals with POLQA is technically possible, but a scale transformation will definitely be required. **Please note, that using WB or NB reference signals in SWB mode are not part of P.863 and should be avoided.**

In order to perform a direct comparison between PESQ and POLQA results, a scale transformation of the POLQA scores is required in order to compensate for the different context (WB vs. SWB). Such a transformation is usually performed by applying a monotonic third order polynomial function to one of the dataset.

Compared to PESQ, POLQA always shows a much higher accuracy for wideband scenarios, no matter, which reference signals are used.

© 2012 HEAD acoustics
Rev0 (03/2012)

Range of MOS values

Table 1 shows the limits of the MOS scales used by the PESQ (P.862.1/2) and POLQA, depending on the operating mode of the algorithm. Note that there is no overlap for WB and SWB mode. In SWB mode, POLQA will typically score an excellent WB signal compared to a SWB reference signal at around 4.5 as well. In the same way, a perfect NB signal compared to the corresponding SWB reference signal will be scored at around 4.0.

Mode	P.862.1/2 MOSmin	P.862.1/2 MOSmax	POLQA MOSmin	POLQA MOSmax
NB	1	4.5	1	4.5
WB	1	4.5	-	-
SWB	-	-	1	4.75

Table 1, limits of the used MOS scales per mode.

Full Scale Context

One consideration that led to the design of PESQ was to predict the results of either plain NB or plain WB experiments. In NB mode, POLQA does exactly the same. In SWB however, POLQA is trained to predict full scale experiments. These are experiments which contain a balanced amount of NB, WB and SWB conditions. The effect of this is that for the first time a measure can be used to directly compare all those data. On the other side, a direct comparison with plain NB or WB experiments will show a skew in the scales which are used and will require a first or better third order regression to be applied to the MOS scores before they can be compared (in most cases the Excel "Trendline" will do a sufficient job).

Effects of Linear Frequency Distortions

PESQ was very insensitive to linear frequency distortions. They only had an impact if they were extreme. This was caused by the equalization built into PESQ which compensated such distortions almost completely. Especially for handset testing this meant a significant drawback. In POLQA those issues have been corrected and the weighting of such distortions is now in line with human perception.

Requirements for Recordings

While PESQ was very tolerant towards incorrect recording levels (and thus often wrong in such cases…), POLQA requires fairly exact control of the levels as mentioned above. If the ratio between the digital level and the sound pressure level is unknown, then it is better to scale both signals to -26dBov, rather than performing a measurement with a "by chance"

© 2012 HEAD acoustics

recording of only 40dBov. Note, that care has to taken when performing such scaling operations on digital signals with an insufficient bit resolution.

On the other side, PESQ was very sensitive to clock differences between the D/A and A/D converters involved in the transmission. This often led to too pessimistic results. POLQA is however designed to handle such cases and it will easily compensate for such clock differences and other time scaling effects in most cases. Of course, this does not mean that test equipment should now use cheaper converters/sound cards….

Another issue could rise from POLQA being a super-wideband speech quality measure. To fully exploit this, the recording equipment used must be capable of recording signals at least 32kHz sample rate with very high quality.

9.5.2. New Application Areas

Compared to PESQ, POLQA does not only provide significantly enhanced accuracy. It also extends the application range for which perceptual measurements can be applied now. Many applications for which PESQ was unsuitable are now within the scope of POLQA.

Acoustical Measurements

While PESQ could only be used to assess acoustical recordings if you knew exactly what you were doing, this is different for POLQA. POLQA comes with full support for acoustical measurements. Those acoustical measurements include the correct weighting of reverberations as well linear and non-linear filtering characteristics caused by handsets and transducers. Also, super-wideband noise superimposed on a narrowband speech signal was part of the POLQA validation. Significant background noise at the listener side however has not yet been validated.

Time Scaling

Time scaling is in principle a variation of the playout speed of the signal. The causes for time scaling can be intentional, e.g. to compensate for packet losses or to avoid jitter buffer over- or underruns. This intentional time scaling comes in two flavors, either with or without pitch correction. In addition, time scaling effects can also be introduced unintentionally, due to poor playback or due to recording devices with inaccurate sample rates. At sample rates above 16kHz it also becomes important to synchronize the sample clocks of the recording and the playing device. Using PESQ, the only chance was to avoid time scaling at all. POLQA instead handles all described effects without difficulties and can be safely applied in such situations. However, it is still recommended to synchronize sample clocks higher than 16kHz in order to improve the stability of measurement results. Below 16kHz, it is generally enough to use sufficiently accurate sample clocks, but the higher the sample rate gets, the more important becomes the synchronization of those clocks. The actual maximum amount of time scaling that can be handled by POLQA is signal dependent, but tests showed that up to 5% are typically handled well, even in the presence of a significant amount of other distortions, like e.g. background noise.

© 2012 HEAD acoustics

High Background Noise

For signals with very high background noises PESQ simply failed since its temporal alignment could not handle such signals. POLQA instead has been tested with very high background noise levels, and in many cases it even worked at an SNR below 0dB. A fixed lower limit for which SNR POLQA still works can't be given since this is again highly signal dependent.

Direct Comparisons between EVRC and AMR type Codecs

PESQ is known to score EVRC and AMR type codecs with a very small, but systematic offset. It is unknown if PESQ over predicts AMR or under predicts EVRC and the effect is generally so small that it is negligible in field measurements, but it is annoying if a direct comparison between those two codec types shall be made. To avoid such a bias was one of the main targets for POLQA and this was successfully achieved.

9.6. Perceptual Results Obtained from POLQA

This chapter is intended to give a broad overview of the measured parameters.

9.6.1. MOS-LQO

The most eminent result of POLQA is the MOS-LQO. It directly expresses the voice quality on the MOS scale. This score is defined by ITU-T Recommendation P.863 which uses a MOS-like scale ranging from 0 (worst) up to 4.5 (best) for the narrowband mode and 0 (worst) up to 4.75 (best) in super-wideband mode.

Please note that the MOS determined by P.862 has an offset of app. 0.5MOS compared to the results obtained by PESQ and most subjective tests. This is due to the fact that the validation of POLQA required a comparison over an extremely wide range of subjective experiments which all had to be squeezed with the [1:4.5] (or [1:4.75] for SWB) scale which was a requirement by ITU-T Q9/12. For practical purposes this is of course not very useful. The POLQA coalition therefore agreed to develop a mapping to a more realistic MOS scale and to only supply versions of POLQA to the market which include this mapping. This additional mapping is directly applied to the MOS produced by the strict P.863 version and uses the following polynomials:

Narrowband: $MOS_{LQO} = 0.79 + 0.0036*MOS_{P863} + 0.2117* MOS_{P863}^2 - 0.0065* MOS_{P863}^3$

Super-wideband: $MOS_{LQO} = 0.276 + 0.7203*MOS_{P863} - 0.00756* MOS_{P863}^2 + 0.01141* MOS_{P863}^3$

The OPTICOM OEM Library provides the MOS after this remapping only.

© 2012 HEAD acoustics

9.6.2. G.107 R-Factor / I_e Value

The POLQA library also provides a mapping of the MOS-LQO score to the scale used by G.107 (e-model). The resulting parameter is equivalent to an I_e – Value. Many people also refer to it as an R-factor. The scale ranges from 0 (bad) up to 100 (best). All values below 60 indicate unacceptable quality. The mapping between R Scale and MOS is implemented as defined in G.107.

Please note that this mapping is valid for narrowband measurements only. Currently there is no such conversion standardized for super-wideband measurements.

9.6.3. Disturbance Density

The disturbance density is an indicator of how loud a signal degradation is perceived in a specific frequency range and at a specific time. The amplitude is in Sone and the frequency scale uses 1/3 Bark bands. For each frame a disturbance density vector is calculated. The disturbance density is a perceptual measure and takes frequency as well as temporal masking into account.

9.7. Non-Perceptual Results Helpful for Cause Analysis

Apart from the perceptual KPIs mentioned above, many other values are reported by the listening quality test module. Most of them are not directly related to listening quality, but they give further insight in the functionality and characteristics of a system.

9.7.1. Delay/Latency

As soon as a signal is processed by any piece of equipment, it will be slightly delayed. The resulting delay is also frequently called latency. While delay does not affect the listening quality, it is an important factor when assessing telecommunication systems because its value has a major impact on the interactivity of a conversation. The longer the delay is, the more discipline is required from both parties involved in a conversation. Delays larger than approx. 300ms are generally unacceptable. While the delay for the old POTS is usually in the range of a few milliseconds, it is typically around 150ms for VoIP systems, sometimes even much longer.

9.7.2. Delay Jitter

Delay jitter in the voice signal may have various reasons. The most frequent cause for this is the dynamic adaptation of the jitter buffers built into modern VoIP equipment. The purpose of these buffers is to assemble a continuous voice stream out of the RTP (=voice) packets which arrive in bursts with non-deterministic timing. The longer these buffers are, the more packet jitter they may compensate for, but the overall latency of the speech signal is then also increasing. On the other side, if the jitter buffer is shorter, the latency is shorter too, but the danger of packet loss is significantly higher. The optimum length of the jitter buffer is de-

© 2012 HEAD acoustics
Rev0 (03/2012)

pending on the network itself and the load on the network. In order to optimize the latency of VoIP equipment, adaptive algorithms are used to automatically adjust the size of the jitter buffer to whatever is required by the network.

Please note that in this context delay and jitter are both referring to the speech signals. They must not be confused with the packet delay and jitter in IP networks even so they are of course resulting from these effects. There might be a significant network jitter in a system under test which will be translated into a large overall delay of the voice without any jitter because a sufficiently large jitter buffer is available in the receiving terminal.

OPTICOM's POLQA implementation offers you information not only about the overall delay but also about the jitter in terms of the minimum and maximum delay which occurs in the test sample. The positive delay jitter is the difference between the maximum delay and the average delay, while the jitter in the negative direction is the difference between the average delay and the minimum delay.

The delay measurement in POLQA is based on the comparison of the two input files. The accuracy of the delay measurement is therefore only limited by the accuracy with which the recording of the degraded file was synchronized to the playback of the reference file. Ideally, recording and playback start simultaneously. While this can be easily accomplished using one soundcard in one test system, it becomes quite difficult across different interfaces or even machine boundaries.

9.7.3. Attenuation

Especially all analog equipment modifies the level of the speech signal. A high attenuation generally leads to a worse perception of voice. In contrast to PESQ, POLQA does weight this as degradation of the signal. Knowing the value of the attenuation is also important for optimizing the overall system design. Attention should be paid to signals which show either a negative attenuation, or attenuations larger than approximately 10dB. In the first case, the signal was amplified instead of attenuated. This may eventually lead to level clipping during the transmission. In the second case, the quantization noise may become an important source of degradation, if low level analog signals are converted to the digital domain and are subsequently amplified in the digital domain. Depending on the test setup, both cases may be ok and intended, but this has to be decided on a case by case basis.

In order to calculate the attenuation, POLQA computes P.56 like active speech levels of the reference as well as the degraded signal in dB. The level of the degraded signal minus the level of the reference signal is then used as the attenuation. Please note that this excludes the pause intervals and that this procedure is different from the one used in OPTICOM's PESQ OEM Library.

9.7.4. Level and Background Noise Measurements

In transmission systems it is frequently important to know the exact levels of the signals. Especially for VoIP systems and voice activity detection (VAD) it becomes also important to know

© 2012 HEAD acoustics

the signal level during the silent intervals as well as during active speech. It is important, that the received background noise does not exceed a certain limit. Levels can be measured in dB if you want to relate the level directly to a sound pressure or electrical level, or as loudness levels. Since the latter takes human perception into account it should be used in preference to the levels in dB. With the OPTICOM POLQA implementation both options are available. Please note, that the loudness as presented by POLQA is directly taken from the intermediate processing steps of POLQA and differs from other loudness definitions as e.g. the Zwicker Loudness. This may change in future versions of the OPTICOM POLQA implementation. All levels are derived using a P.56-like filtering. In order to discriminate between active speech and pause, information from POLQA's temporal alignment is used.

9.7.5. Signal to Noise Ratio (SNR)

OPTICOM's POLQA OEM DLL also calculates the SNR for the reference and the degraded signal independently. The noise as well as the signal level is calculated by the VAD which POLQA uses for the temporal alignment. Please note that this is completely different from the way in which the SNR is calculated in OPTICOM's PESQ OEM DLL. The result obtained by PESQ indicates the SNR between the reference and the degraded signal, while in POLQA the SNR is calculated for each of the two signals separately based on that signal only (without relation to the other signal).

9.7.6. Active Speech Ratio (ASR)

ASR is calculated by the POLQA OEM DLL based on the information calculated by the Voice Activity Detection (VAD) which is part of the temporal alignment. The ASR defines the ratio between active speech and the overall signal length.

9.7.7. Pitch

POLQA bases some of its analysis on the pitch of the reference and the degraded signal. The pitch frequency of the signals is calculated by analysis of the subharmonics [BEER89] in the signals and subsequently averaging the pitch frequency of individual voiced signal sections. The pitch is given in Hz and separately for the reference and the degraded signal.

9.7.8. Short Term Spectra

The POLQA OEM DLL also calculates the short term spectrum of each frame. The spectra are available either as line spectra, using a linear Hz scale, or in so-called critical bands on a psycho acoustically motivated pitch scale, using 1/3 bark bands as resolution. The amplitude of the line spectra is the power of the signal at a given frequency in dB. In case of the Bark spectra, the amplitude is the power density of the signal at a given Bark band in dB.

© 2012 HEAD acoustics

9.8. Reporting and Averaging of Results

Whenever possible, POLQA results should be reported **after** averaging some data points. As a rule of thumb averaging should at least be performed over sentences spoken by two male and two female talkers. The averaging should preferably happen in the MOS domain, although a concatenated speech sample using different voices also provides accurate results. It must however be noted that averaging the MOS LQO of four files leads to slightly different results than concatenating these four files and performing one measurement. The type of averaging should thus be reported (averaging in the MOS domain or in the signal domain).

Nevertheless, there are some applications that do not allow for such averaging. These include scenarios like drive testing and most tests run on live networks with non-deterministic behaviour. In these cases no averaging can be performed and per file results must be reported.

If averaging was possible, at least the average, maximum and minimum MOS values should be reported along with the number of samples used for calculating the average. In addition, the 95% confidence intervals should also be mentioned.

If, due to the nature of the experiment, averaging across files was not allowed, the report should contain more elaborated statistics on the MOS values and ideally the distribution of the scores in a graphical form.

9.9. Accuracy of POLQA Results (P.863)

One very common mistake when interpreting POLQA results is to overestimate the accuracy of the results. When presenting MOS-LQO values with a resolution of three decimals, one should always be aware, that the subjects in a listening test have zero decimals available for their votes. Without any further uncertainties and other subjective factors this would result in a theoretical maximum resolution of 0.03 points on the MOS scale if 30 subjects were participating scoring the same speech sample. In reality, the accuracy of a subjective test will be much worse since very often the subjects disagree on the correct vote. POLQA is however trained on such data and the resulting errors will propagate. The absolute accuracy of the POLQA results depends on the application as well as the number of measurements which are averaged. For a single measurement the Prediction Error will usually be much better than 0.3 MOS.

9.10. Limitations of POLQA

Static delay differences can't be perceived in a listening only experiment. Delay is therefore although measured by POLQA, not taken into account as a degradation of the signal. Delay variations that happen during active speech will however be weighted properly.

© 2012 HEAD acoustics

9.11. How to Assess "Signal Enhancers"

Signal enhancers or VEDs (Voice Enhancement Devices) are pieces of equipment that try to make the processed signal sound better than the original signal. Examples are e.g. noise reduction systems. If you take the input signal of the enhancer as the reference and the output signal as the test signal of any perceptual measure, the result will usually be the opposite of what you would expect. In general the enhanced signal will be graded down more, the better your enhancer works. This is due to the fact, that perception based measurement algorithms assume that any audible difference between the two input signals is a distortion, and by definition the "enhanced" signal will sound different than the unprocessed signal.

In order to come around this, we recommend a setup as shown in Figure 9-4. Here a clean signal is chosen as the reference file and afterwards the signal is distorted artificially, which results in signal D. Subsequently the distorted signal D is sent to the enhancer. The output of the enhancer will be E, the enhanced signal.

If voice quality is assessed now, one should choose the clean reference R and the enhanced signal E as the input signals of POLQA, which will result in MOS_E. The value of MOS_E now indicates how similar the enhanced signal sounds to the clean reference.

Going one step further, one can calculate the gain achieved by the enhancer by relating the final MOS derived this way, to the MOS achieved by comparing the clean reference (R) to the distorted reference (D), resulting in MOS_D. A comparison of MOS_E **and** MOS_D **is then a good measure for the effectiveness f the VED.**

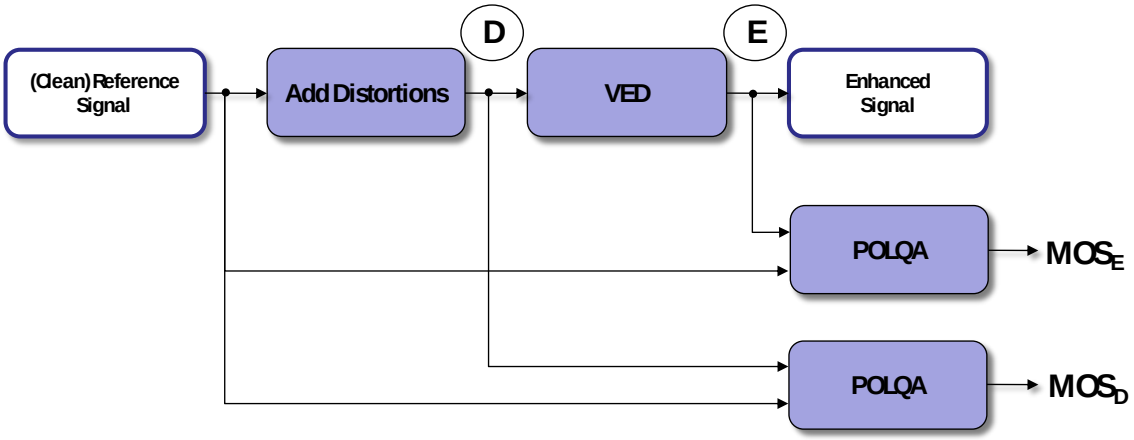


Figure 9-4 Measurement setup for the assessment of Voice Enhancement Devices (VEDs)

9.12. How to deal with Comfort Noise Insertion

Many modern telephony systems use noise substitution or comfort noise insertion. Especially the second case may lead to worse measurement results than expected when reference signals are used that contain digital zero during the silent intervals. POLQA will compare this digital silence to the comfort noise and detect the comfort noise as a distortion.

© 2012 HEAD acoustics
Rev0 (03/2012)

One can partially compensate for this issue by avoiding digital zero in the silent periods of the reference signals and using a low level noise floor instead. The level of the noise floor should be in the same range as the expected comfort noise level is. The exact value is best determined in practical experiments.

9.13. Considerations Regarding the Data Acquisition for POLQA

There are two major issues which have to be taken into account when designing a test system that shall perform data acquisition for POLQA. The first issue is related to the accuracy of the clocks used for A/D or D/A converters and the second issue is related to the synchronization of the moments when playing the speech files starts on one side of the connection and recording starts on the other side.

9.13.1. Synchronisation of the Sample Clocks

If the playing and recording side uses different clocks to generate the sample rate, then this appears to POLQA like a variable delay of the system under test. Although POLQA is excellent in handling larger delay variations that occur from time to time as well as drifting delays, we strongly advise using professional soundcard equipment as it is used in studio environment as well. For sample rates higher than 8kHz it is also strongly recommended to synchronize the clocks of the soundcards.

9.13.2. Synchronisation of Play and Record

Besides the quality scores, the OPTICOM POLQA OEM Library also provides delay measurements between the sender and the receiver. This measurement is based on the comparison of the two signals provided to POLQA. It is obvious, that the accuracy of the delay measurement therefore depends on the accuracy of the synchronization of playing and recording the speech samples. If the transmission started too late, then the delay will be measured larger than it is and if playing the sample started too early, the beginning of the signal will be cut off in the recorded file and the measured delay will therefore be too short.

9.13.3. Control of Playback and Recording Levels

POLQA imposes some very strict constraints regarding the signal levels. Please refer to chapter 9.4.3 for details on this.

© 2012 HEAD acoustics

© 2012 HEAD acoustics