Perceptual Objective Listening Quality Analysis

POLQA 2015 (V2.4) Investigated

- Technical White Paper -

October 2014

Contents

Summary 3
Objectives for an Advanced POLQA 3
Length Dependency in NB Mode 3
Analysis of the Transparency Behavior 4
Other Changes in POLQA V2.4 5
Analysis of POLQA 2015 (V2.4) 6
POLQA Results for Reference Conditions – Backward Compatibility 6
General Prediction Performance 6
Comparison of EVRC and AMR codec Measurements 7
Analysis of the Sampling Error Detection 7
Shift Performance 7
Processing Requirements 7
POLQA and VoLTE 8
Availability of POLQA V2.4 9
Conclusions 10
Acknowledgments 10
Literature 10

Summary

POLQA, the third generation perceptual voice quality test method standardized as P.863 in 2011, has been widely adopted as the state-of-the-art MOS benchmarking technology for mobile networks. Since its first release (V1.1) in 2011, numerous experiences in the field became available and along with the technical development in mobile communication (VoLTE, 5G) some areas for improvement were identified.

In strong collaboration with SG12, the parties of the POLQA Coalition (OPTICOM, SwissQual and TNO) proposed an evolved version to ITU-T Study Group 12, which was approved as Rec. P.863 Edition 2.4 in September 2014. It is expected that V2.4 will supersede the earlier released V1.1 with product implementing it from 2015 on.

This white paper outlines the objectives for the update, compares those to the achieved improvements and highlights the most important changes for users. It will be demonstrated that this update marks a significant step towards even higher measurement accuracy and a broader range of applications for POLQA, while maximum backward compatibility of the measured scores is maintained. As an addition, an investigation of the applicability of POLQA on live VoLTE networks is presented as well.

Objectives for an Advanced POLQA

POLQA V1.1 was found to have three minor issues, which had very little effect in most real measurement scenarios, but which nevertheless may have caused problems if users were unaware of standard best practices for the use of POLQA. These issues were,

- **Length dependency** of the results in narrowband (NB) mode if the signal duration exceeded approximately 10 s in narrowband mode

- **Transparency problems:** Too often a comparison of the reference signal with itself resulted in MOS scores below the optimum (known as “the transparency problem”)

- **Shift performance:** Small shifts of the starting point of the degraded signal may have had unexpectedly large impacts on the measured MOS

Due to time constraints it was not feasible to tackle all three matters for this update. Where the shift performance is critical there already exists a pragmatic solution, by using the High Accuracy mode (HA-Mode), therefore this matter was deferred for possible future consideration.

The main focus of this update was therefore solving the length dependency of the MOS as well as the transparency problem. Along with this, it goes without saying that, the general behaviour should not become worse and scores predicted by V2.4 should remain as close as possible to those measured with V1.1.

The update also provides several other improvements due to general bug fixes and optimization, as are described herein.

Length Dependency in NB Mode

One problem of V1.1 was that narrowband scores were dependent on the signal length. It was observed that files longer than approximately 10s were consistently scored lower with increasing file length. This problem is completely solved with POLQA v2.4. For illustration the speech samples of P.501 were used in the following example. The four Italian files which are included in that recommendation have durations of more than 15s, while all other files are mostly around 8s long. All samples were processed by a range of standard coding conditions in order to create degraded versions of the reference files. MOS scores were produced by applying either POLQA V1.1 or POLQA V2.4 to these files.

The following two diagrams show the resulting average scores for the Italian (solid line) and the remaining other 28 speech samples (dotted line).

As can be seen in Figure 1, P.863 V1.1 scores the long files significantly lower than the short files, while for P.863 V2.4 there is no such systemic bias of the longer files visible anymore (Figure 2), indicating that the length dependency problem is resolved for POLQA V2.4

© The POLQA Coalition: OPTICOM GmbH, Erlangen, Germany; SwissQual AG, Solothurn, Switzerland; TNO Telecom, Delft, The Netherlands.

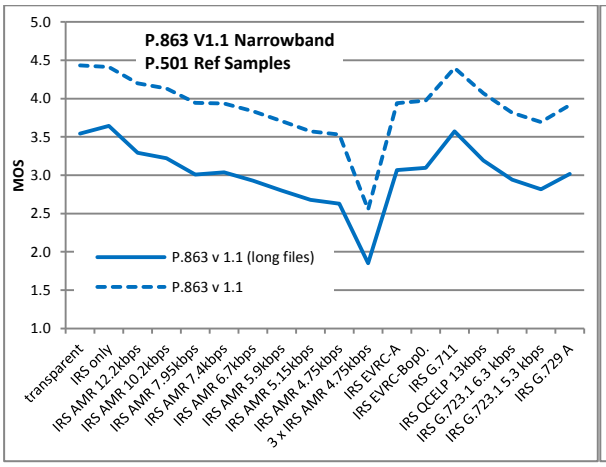

Figure 1: P.863 V1.1 results in NB mode for long speech files compared to 8s regular speech files

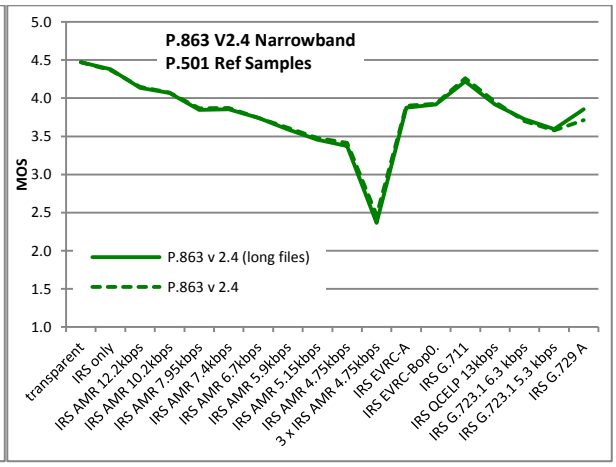

Figure 2: P.863 V2.4 results in NB mode for long speech files compared to 8s regular speech files

Analysis of the Transparency Behavior

A general feature of POLQA is that reference signals which are not optimal are idealized before they are compared to the degraded signal. The reasoning behind this is that in an ACR test, subjects will attribute all audible distortions to the degraded signal, even if they were already part of the reference signal. A consequence of this is that a sub-optimal reference signal compared to itself will be scored <4.75 (or <4.5 in NB mode). Such reference signals are generally described as non-transparent. While this is the desired behaviour of POLQA, V1.1 was apparently too sensitive in this aspect and consequently too many samples were considered as non-transparent. V2.4 addresses this topic and a revision of the internal reference signal handling reduced the effect without any negative impact on the prediction performance. The following table (Table 1) illustrates the increased number of samples considered as transparent in POLQA V2.4. For public repeatability reasons, this analysis was made on the 32 samples of P.501. All samples were filtered to NB or SWB (using the G.191 SWB band-pass filter) and levels adjusted prior to use.

As reported earlier [C0085], not all P.501 speech samples are consistent with the requirements of P.863 and P.863.1. Mainly, the signal length, the duration of leading silence and especially the noise floor are seen to violate these requirements. As a consequence, for an additional analysis all 32 speech samples were manually cleaned and edited correctly according to the following specification:

Nr. of transparent samples (32 P.501 references)		
	NB	SWB
P.863 V1.1	**14**	**10**
P.863 V2.4	**19**	**26**

Table 1: Transparent P.501 samples in V1.1 and V2.4

- Leading silence duration ~0.5s
- Silence between sentences ~1s
- File length 8s
- Speech pauses muted and file interlaced with -85dB white SWB noise

As Table 2 shows, the strict requirements on the reference samples lead to a further, significant improvement of the number of transparent samples in NB mode.

© The POLQA Coalition: OPTICOM GmbH, Erlangen, Germany; SwissQual AG, Solothurn, Switzerland; TNO Telecom, Delft, The Netherlands.

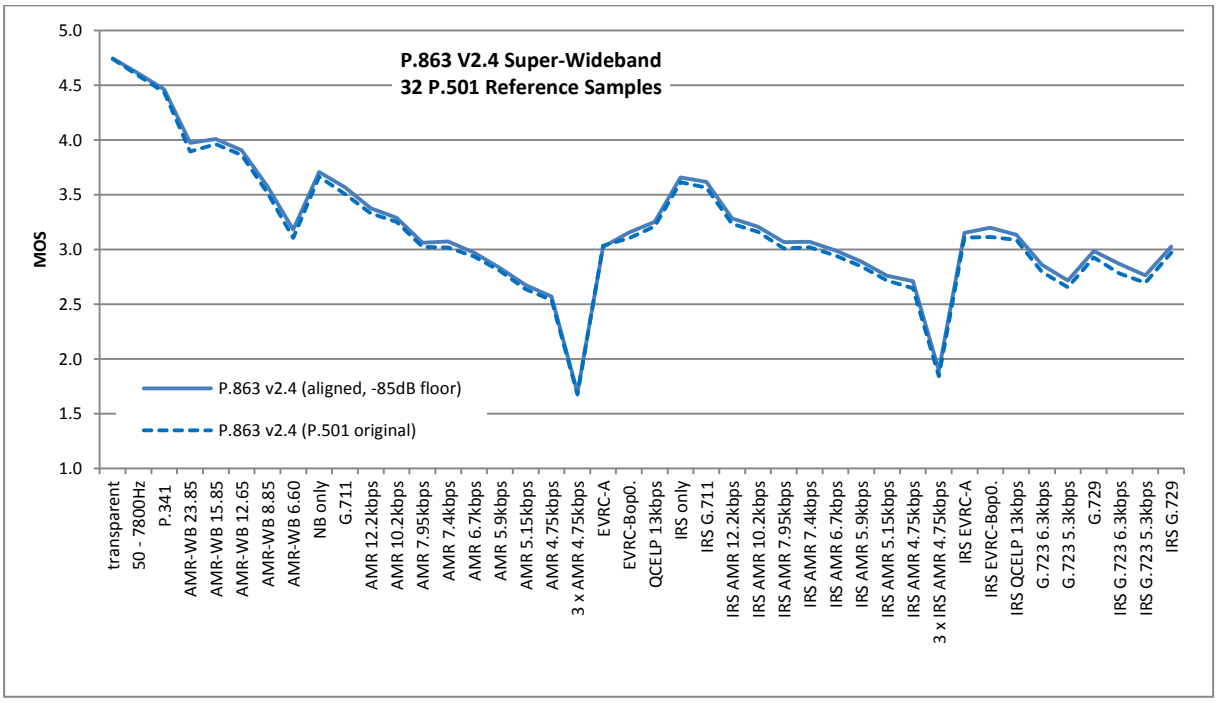

Figure 3: POLQA V2.4 results for P.501 original and cleaned samples.

The remaining non-transparent speech samples are mainly French and Finnish signals which exhibit a relatively strong noise floor (>-75dB OVL) and / or contain significant undesired tonal components during the active speech sections. These should definitely be idealized by the model. Consequently, the behaviour now indicates that the model internal idealisation of the reference signal works as intended within POLQA 2.4.

Note that the use of the edited and cleaned P.501 samples also leads to a very small but visible increase of the predicted MOS scores (Figure 3).

Nr. of transparent samples (32 P.501 references)		
	NB	SWB
P.863 V1.1	**14**	**10**
P.863 V2.4	**19**	**26**
After cleaning (-85dB noise floor)		
P.863 V1.1	**20**	**12**
P.863 V2.4	**27**	**26**

Table 2: Transparent P.501 samples in V1.1 and V2.4 after cleaning P.501 samples

Other Changes in POLQA V2.4

Apart from the mentioned improvements, there are two more important changes.

- The standard now includes an initial sample rate conversion to 8 kHz or 48 kHz, depending on the selected mode (NB or SWB). The sample rate converter and the corresponding filter coefficients are now part of the recommendation.
- The former separately published amendment to P.863 on acoustical measurements using POLQA V1.1 in NB mode is now fully integrated into the revised recommendation.

© The POLQA Coalition: OPTICOM GmbH, Erlangen, Germany; SwissQual AG, Solothurn, Switzerland; TNO Telecom, Delft, The Netherlands.

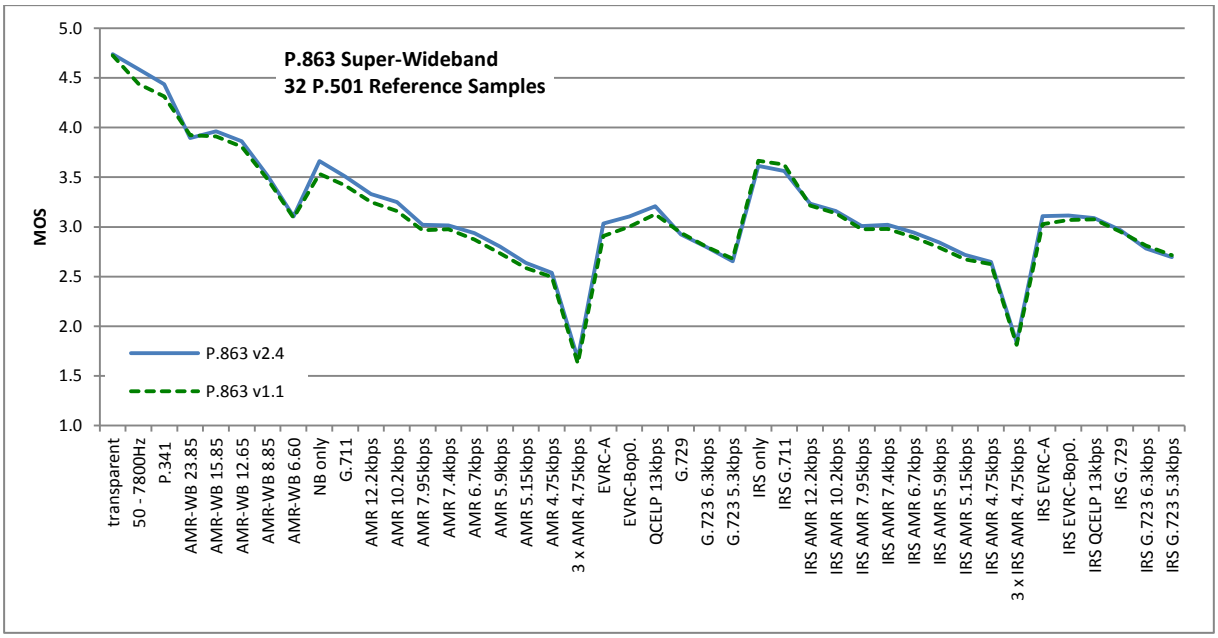

Figure 4: POLQA V2.4 results in SWB mode, average over 32 speech samples (P.501) compared to V1.1 results.

Analysis of POLQA 2015 (V2.4)

POLQA Results for Reference Conditions – Backward Compatibility

Besides the increased prediction performance, it is of interest to know how much the average results of the same coding conditions differ between POLQA V2.4 and V1.1, since this has a direct impact on the comparability of results measured with different versions of POLQA.

A collection of reference samples was therefore taken and processed over a set of standardized codecs. Both versions of POLQA have been applied to this set of "reference conditions". Figure 4 and Figure 5 show the resulting average scores per condition for V1.1 and V2.4. Looking at the results for the super-wideband mode (Figure 4), it can be seen, that the V2.4 scores for conditions which include bandwidth limitations (e.g. NB codecs) are slightly higher than those of V1.1.

The situation for narrow-band measurements is slightly different (Figure 5), since the NB mode in V2.4 is intentionally tuned to score slightly more pessimistic than V1.1, in order to achieve a better discrimination between the different codec rates as well as between transparent narrow-band (50-3800 Hz), IRS only and G.711.

It can be observed that for the most often used codecs of the AMR and EVRC family, the scales are only marginally shifted. A 1:1 comparison of the results is therefore possible if this small offset is kept in mind. However we do not encourage a comparison of scores created with

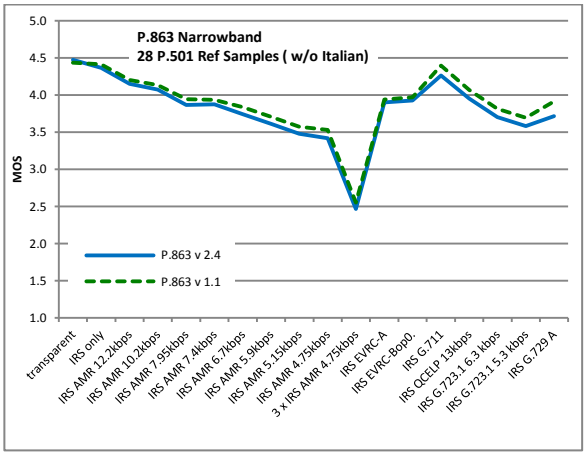

Figure 5: POLQA V1.1 results in NB mode, average over 32 speech samples (P.501) compared to V1.1 results.

different versions of the standard. Please also note that the above is only valid for the average over a large number of measurements. Individual results may differ significantly between V1.1 and V2.4, namely for cases for which V2.4 was enhanced.

General Prediction Performance

Due to the modifications and applied bug-fixes in V2.4, the average and the worst case prediction performance has been improved. Table 3 shows the rmse* values according to P.1401 for V1.1 compared to V2.4. The underlying subjective experiments were all 64 experiments from the POLQA pool, which resulted in roughly 47000 file pairs.

The rmse* is similar to the root mean square error (rmse) and indicates the average error of the model predictions compared to subjective test results, while taking the confidence of the

subjective test into account. The unit is MOS (P.800 1 to 5 scale). Note that very small values like 0.01 are already significant.

The reported minimum value is the lowest (best) rmse* measured value for any of the experiments. The reported average value is the average rmse* across all 64 rmse* values (the average across all experiments). The most important value is probably the maximum rmse* since it indicates the worst case and thus the reliability of the model.

As can be seen in Table 3, while the average performance in NB mode remained the same as in V1.1, the worst case behaviour could be improved significantly. In SWB mode the improvement is even stronger, since not only the worst case was improved, but also the average performance is now clearly better.

	Rmse* V1.1	Rmse* V2.4
NB Avg	**0.1380**	**0.1308**
NB Min	**0.0553**	**0.0466**
NB Max	**0.2844**	**0.2618**
SWB Avg	**0.1940**	**0.1674**
SWB Min	**0.0713**	**0.0265**
SWB Max	**0.2809**	**0.2775**

Table 3: Prediction performance of POLQA V2.4 compared to V1.1, lower values mean better performance.

Comparison of EVRC and AMR codec Measurements

V1.1 of P.863 contained a remark that comparisons of POLQA measurements for EVRC and AMR codecs “are for further study” [P863V1]. Since its release, further experience has become available and the latest results with V2.4 confirm that such restriction is no longer required. The according sentence has therefore been deleted from P.863 and no limitations regarding its use in combination with AMR or EVRC codecs exist.

Analysis of the Sampling Error Detection

One of the major advantages of POLQA is that it can handle small differences between the sampling rate of the reference and the degraded signal (aka sampling error or time scaling). In practice this occurs when different sampling clocks are used at the sending side and on the receiving side, typical ranges which must be handled without significant drop of the MOS-LQO are in the range of +/-3% sampling error. Time scaling above +/-5% of the nominal sample rate without pitch preservation is rare. Nevertheless, POLQA should be able to handle this gracefully. Between POLQA V1.1 and V2.4 the concept of the sampling error detection has not changed. It is still based on a linear regression in order to estimate the slope of the delay vs. time curve. What has changed though is the way in which this regression is performed.

V1.1 made use of an extensive histogram analysis, while in V2.4 a least squares approximation is used, which proved to be far more robust for larger sampling errors. Typically end users will not notice this difference for their applications. However, this is an important basis for future improvements of the prediction accuracy.

Shift Performance

It is known that the POLQA MOS-LQO may vary more than expected if the starting point of the degraded signal varies slightly. This was tested with all samples from the POLQA pool and by shifting the degraded signals sample wise over an entire FFT window (128, 256 or 1024 samples, depending on the sample rate). A histogram of the resulting MOS differences compared to the median of all values is shown in Figure 6 below. The ideal shape of the curve would be a sharp peak in the centre. As can be seen, there is a small improvement in V2.4 compared to v1.1. It was not expected to solve this with the current update to V2.4, but this small improvement is of course most welcome. As can also be seen from the chart, larger deviations due to the shifting of the start point are very rare for both versions of POLQA.

Processing Requirements

Some of the changes to POLQA have saved a lot of processing time, while others have increased the same. To what degree the two factors balance each other depends on the input signals and cannot be generalized. On average, however, it is expected that V2.4 will require

© The POLQA Coalition: OPTICOM GmbH, Erlangen, Germany; SwissQual AG, Solothurn, Switzerland; TNO Telecom, Delft, The Netherlands.

slightly more processing power than V1.1. Benchmarks will be published as soon as they become available.

POLQA and VoLTE

Currently, VoLTE networks are not yet widely deployed and very little field data exist which can be used for POLQA measurements. Consequently, even the updated V2.4 of P.863 correctly includes a remark that the use of POLQA with VoLTE networks must be further studied.

As a first step [C229] presented some POLQA measurement results for a limited set of field collected data (around 2000 files) which indicate the safe use of POLQA for VoLTE:

In a first analysis (Figure 7) the variability of the delay vs. time is investigated and compared for VoLTE (upper chart) and 3G (middle chart), using a typical signal as it is applied for drive testing (lower chart). It is obvious, that the VoLTE case exhibits not only far more delay variations than the 3G case, but also larger delay steps. This is further analysed in Figure 8, which is a histogram of the distribution of the delay steps. As can be seen, in the 3G case delay steps typically do not exceed 20 ms, which corresponds to the frame size used in 3G networks. Usually these delay changes occur during handovers between cells only. Please note that in this chart the 0 to 20 ms range includes many cases where no delay variation is present at all. For the VoLTE case instead, the majority of the delay steps (50%) are between 20 and 40 ms and range up to 80 ms. It is assumed that these variations coincide with jitter buffer adaptations and variations of the playout speed in the VoIP like architecture of VoLTE.

It is now interesting to see how POLQA scores the VoLTE conditions compared to the 3G cases. A histogram of the MOS values is presented in Figure 9. For the 3G case, the result is clear; in roughly 30% of the cases a clean channel is found and the maximum score possible for the used codec is achieved, which is in the range of 4.0 to 4.2 MOS. With increasing amounts of transmission errors, the quality degrades rapidly. For VoLTE the situation is slightly different. Here the peak in the histogram is at a slightly lower MOS of 3.8 to 4.0 and only 20% of the cases reach the maximum quality.

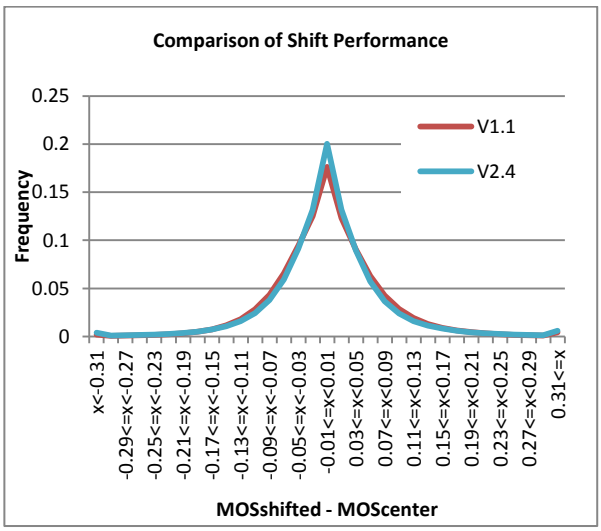

Figure 6: Comparison of the shift performance

This can be well explained by the fact that 50% of the cases show potentially audible delay variations of 20 to 40 ms.

The idea behind [C229] was based on the assumption, that the main difference between cases for which POLQA is known to behave well (e.g. 3G networks) and the still little known behaviour on VoLTE is related to the delay variability. If it can be shown that POLQA works nicely for these delay variations, then this can be seen as a clear indication for the applicability of POLQA in VoLTE networks. The subsequent analysis therefore focuses on only those samples where delay variations actually occur. From subjective experience it is expected that the delay variation in VoLTE has some, but limited effect on the resulting POLQA score since in contrast to 3G networks, the system tries to conceal the audible effect.

The outcome of this analysis can be seen in Figure 10, where the MOS for different amounts of delay variation is presented. The blue bars indicate the VoLTE case where it can be seen that the effect of the delay variability is small, but clearly increasing as the amount of delay variation increases. For the 3G case (orange bars), the effect is much stronger since the changes typically happen quite uncontrolled. For delay variations of more than 40 ms the amount of data for 3G networks is very small and does not allow drawing conclusions. It is assumed that these few cases are typically more related to

© The POLQA Coalition: OPTICOM GmbH, Erlangen, Germany; SwissQual AG, Solothurn, Switzerland; TNO Telecom, Delft, The Netherlands.

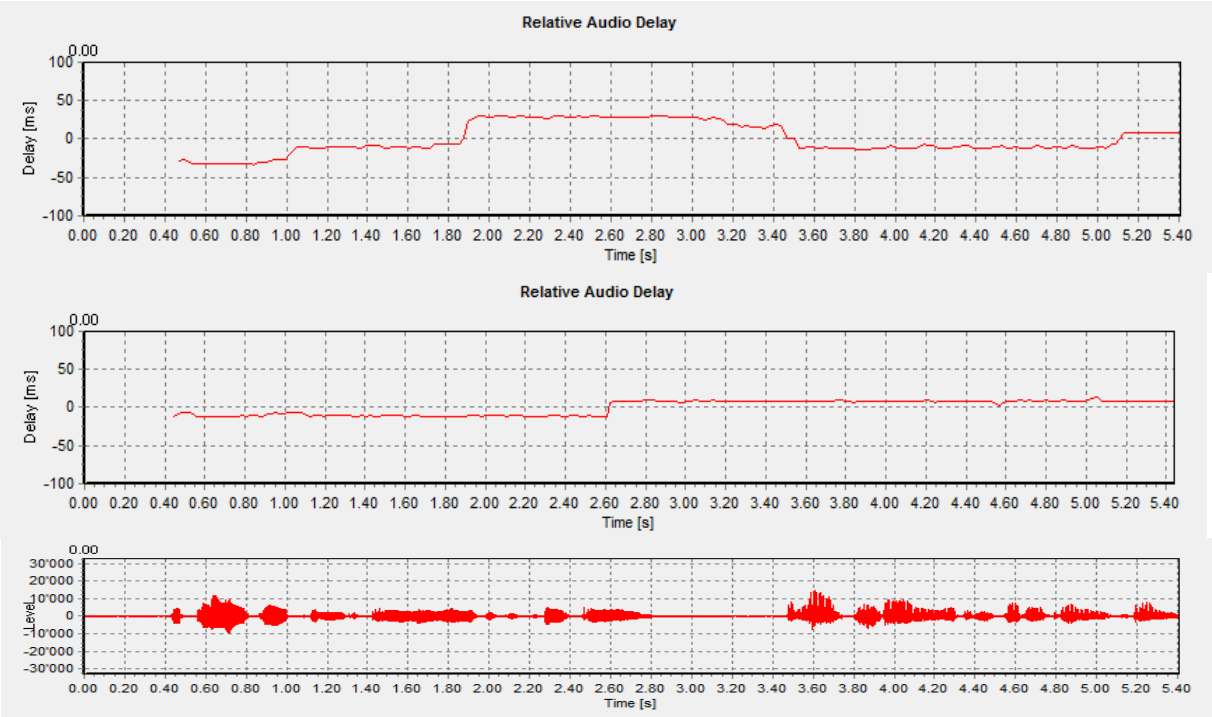

Figure 7: Comparison of the delay variability during one speech sample in VoLTE (top) and 3G (middle) networks (live recording).

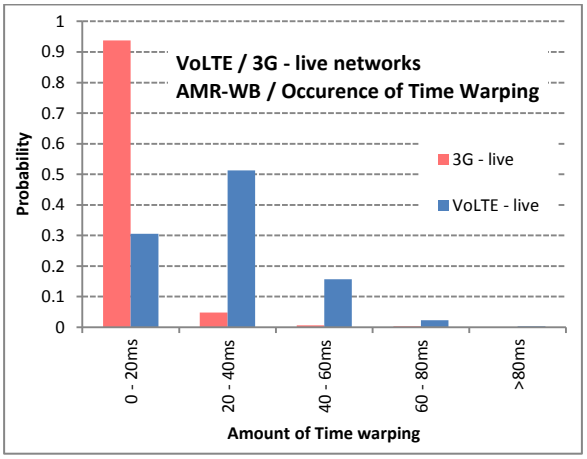

Figure 8: Histogram of delay variations in live VoLTE and 3G networks, for field collected data.

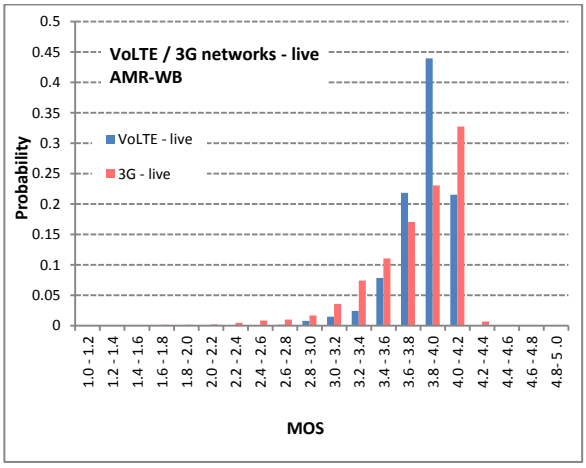

Figure 9: Histogram of MOS-LQO for live VoLTE and 3G networks.

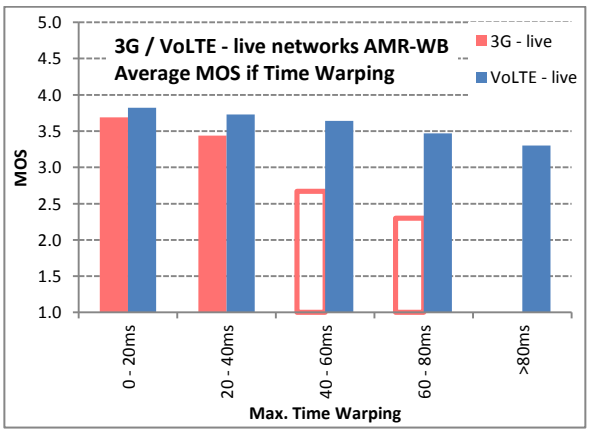

Figure 10: MOS scores in dependence of delay variability for live VoLTE and 3G networks. Non-filled bars indicate insufficient data.

other problems like e.g. bad coverage and transmission errors It can thus be concluded that, at least for the investigated cases, POLQA scores VoLTE as expected and no problems due to the increased delay variability, which is the main difference to 3G networks with regard to voice quality, are observed.

Availability of POLQA V2.4

POLQA V2.4 is available from OPTICOM for integration in OEM products starting October 2014. This should lead to updated POLQA products in early 2015. The available library for OEM integration will be able to produce results for V1.1 as well as for V2.4.

Conclusions

As explained in this white paper, POLQA V2.4 is a major milestone in the development of perceptual measurement methods. Not only have the targeted objectives for this update been met, but also the general performance has been significantly improved and the application range was extended. It is also shown that despite the limited amount of field data, early indications are that POLQA can work very well for VoLTE conditions.

Acknowledgments

The POLQA Coalition would like to thank the following partners who have supported the development of POLQA V2.4 by extensive Beta testing and useful comments: ASCOM, Dolby, Head Acoustics, Malden and Orange

Literature

[P863V1] Recommendation ITU-T P.863 (2011), *Perceptual objective listening quality assessment*

[C0085] ITU-T SG12 C229 (2014), *Reference Speech Samples for POLQA, Selection Method and Available Samples*, OPTICOM GmbH, Rohde & Schwarz, TNO

[C229] ITU-T SG12 C229 (2014), *P.863 under live VoLTE conditions*, Rohde & Schwarz

© The POLQA Coalition: OPTICOM GmbH, Erlangen, Germany; SwissQual AG, Solothurn, Switzerland; TNO Telecom, Delft, The Netherlands.

Published by:

OPTICOM GmbH
Naegelsbachstrasse 38
D - 91052 Erlangen
GERMANY

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

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

SwissQual AG

Allmendweg 8
4528 Zuchwil
Switzerland

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

VAT ID No. CHE-113.795.335
Directors:
Hanspeter Bobst, CEO

Dietmar Vahldiek, Chairman
Registration no.
CHE-113.795.335

For inquiries on POLQA Licensing please contact OPTICOM GmbH or visit www.polqa.info for further details.

For an updated reference list of available POLQA products and solutions please refer to our website:

www.polqa.info

Copyright and Trademark Information

© 2011 The POLQA Coalition: OPTICOM GmbH, Erlangen, Germany; SwissQual AG, Solothurn, Switzerland; TNO Telecom, Delft, The Netherlands.

POLQA®, PESQ® and the OPTICOM logo are registered trademarks of OPTICOM GmbH; All other brand and product names are trademarks and/or registered trademarks of their respective owners.

This information may be subject to change. All rights reserved.