

# **The Performance of the ITU-T P.862.1 Standard (PESQ-LQO) on AMR Live Networks**

Technical Paper

© Ascom 2009. All rights reserved.

TEMS is a trademark of Ascom. All other trademarks are the property of their respective holders.

No part of this document may be reproduced in any form without the written permission of the copyright holder.

The contents of this document are subject to revision without notice due to continued progress in methodology, design and manufacturing. Ascom shall have no liability for any error or damage of any kind resulting from the use of this document.

**Contents**

Abstract ..... 4

1. Myth and reality about the AMR codec’s expected speech quality  
..... 4

2. Evaluation test of the ITU-T speech quality standard in AMR live  
network conditions ..... 7

2.1. Test design ..... 7

2.2. Evaluation procedure ..... 8

2.3. Results of the PESQ algorithm performance on AMR live networks 10

2.4. Comments on the performance and the accuracy of the PESQ algorithm  
as implemented in TEMS Automatic..... 11

2.5. Comments on the AMR PESQ tuned solution ..... 11

3. Conclusions ..... 12

## Abstract

A comprehensive test has been performed in order to evaluate the PESQ performance on AMR live networks. The test was necessitated by the fact that the speech quality standard has not been tested and validated on AMR live networks yet.

### 1. Myth and reality about the AMR codec's expected speech quality

The operability of 3G networks is complex. It has been demonstrated that some estimated performance levels described in 3GPP documents are only met with great difficulty or at lower values than expected.

The functionality of the AMR codec within a live UMTS network might be one of these cases. Designed to ensure high capacity, both the AMR FR and more importantly the AMR-HR codecs are expected to perform higher speech quality than the EFR codec, especially at low C/I values (below 6-8dB). The 3GPP document [1] provides some informative speech quality values for both the AMR-FR and the HR codecs. These values have been obtained based on a single listening test on different simulated RF scenarios. The results presented in the 3GPP document are shown in Charts 1 and 2.

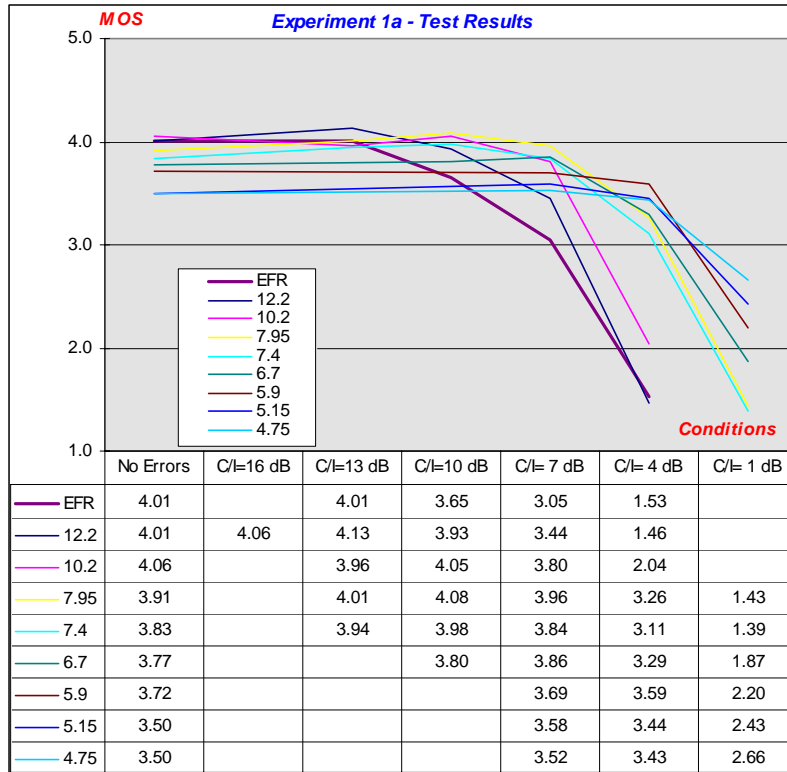


Chart 1. Family of curves for Experiment 1a (Clean speech in Full Rate) (3GPP TR 26.975)

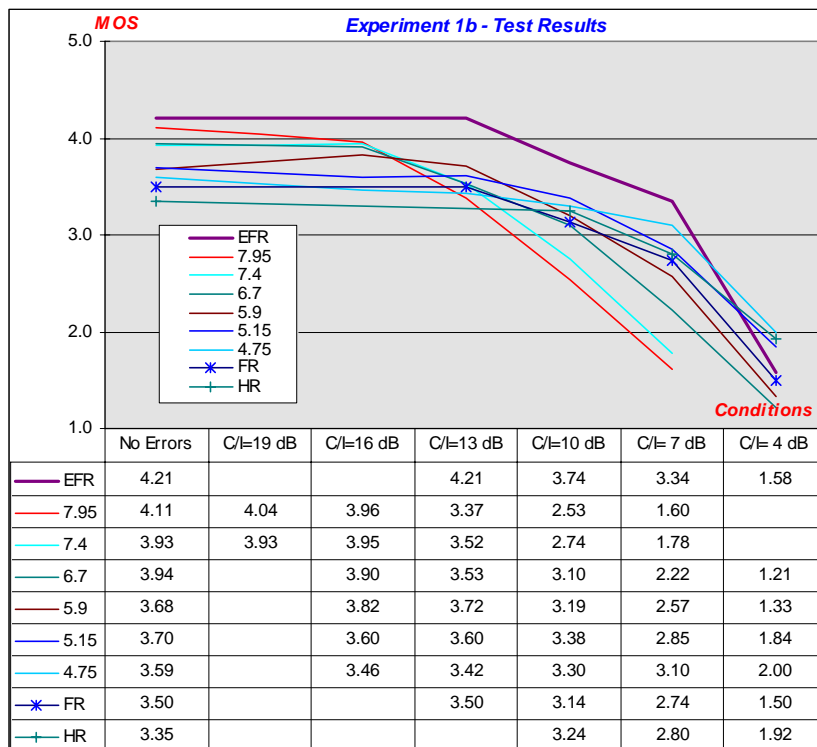


Chart 2. Family of curves for Experiment 1b (Clean Speech in Half Rate) (3GPP TR 26.975)

It should be noted that the 3GPP document states that the subjective test is valid only on the used database and that another database could exhibit different results. This statement becomes important especially if the databases represent live network conditions instead of simulated conditions.

**“Important Note:** MOS values are provided in these figures for information only. Mean Opinion Scores can only be representative of the test conditions in which they were recorded (speech material, speech processing, listening conditions, language, and cultural background of the listening subjects, etc.). Listening tests performed with other conditions than those used in the AMR Characterization phase of testing could lead to a different set of MOS results. On the other hand, the relative performances of different codec under test conditions is considered more reliable and less impacted by cultural difference between listening subjects. Finally, it should be noted that a difference of 0.2 MOS between two test results was usually found not statistically significant” (3GPP TR26.975).

The ITU-T standard P.862/P.862.1 (PESQ-LQO) has been developed, tested, validated, and calibrated for all types of applications (wireless, VoIP, fixed networks) using different speech codecs.

Since the standards P.862 and P.862.1 have been empowered by the ITU-T (in February 2001 and September 2003, respectively), various 3G networks (UMTS) using the AMR codec have been deployed and are currently running in different markets.

PESQ-LQO has been thoroughly analyzed within AMR simulated conditions, such as all codec bit rates with different error patterns generated by a large scale of C/I values. However, none of the databases used for the standard's development, training, testing, validation, or calibration contained live AMR network conditions.

Our tests, however, have been performed on different live UMTS networks, in various markets. Using PESQ-LQO as the speech quality evaluation metric, the results showed slightly lower performance speech quality than the results presented in the 3GPP document [1]. This could have been caused by either the fact that the 3GPP document regards only simulated conditions, or by the fact that the PESQ algorithm exhibits limitations when the degraded conditions are characteristic to the AMR live networks.

Since the tests on the UMTS network's speech quality have been performed with an objective speech quality metric, characterized by a defined accuracy ([2], [3]), which has not been evaluated on AMR live networks before, it was decided that a custom subjective test must be performed in parallel in order to evaluate the PESQ algorithm's accuracy on these networks.

## **2. Evaluation test of the ITU-T speech quality standard in AMR live network conditions**

### **2.1. Test design**

#### **Speech material**

The test databases consist of speech samples collected in AMR live networks. Four Harvard Sentence pairs, spoken by four talkers (two males and two females using American English) were used as source material and run through the networks being tested.

In addition, the source speech samples have been encoded-decoded with the AMR codec, running all its eight bit rates (4.75kbps, 5.15kbps, 5.9kbps, 6.7kbps, 7.4kbps, 7.95kbps, 10.2kbps, 12.2kbps). The clean codec conditions were used in order to introduce the high end of speech quality into the test.

Five MNRU conditions (6dBQ, 12dBQ, 18dBQ, 24dBQ, 30dBQ) of the source speech have been used as references for the subjective test.

**Test conditions**

Besides the clean AMR coding conditions and the MNRU conditions, uplink and downlink field recordings of the source speech running through AMR networks have been used. In order to get the complete behavior of the ITU-T speech quality standard with AMR live network conditions, speech samples in AMR HR, AMR FR, both 850Mhz and 1900Mhz bandwidths, have been collected. The Cingular/AT&T networks in and around Reston, VA have been used for the tests.

Care has been taken in order to ensure that the whole speech quality range has been covered. With this in mind, data collection routes were used in areas ranging from very good RF coverage to very poor coverage.

Within each network and link, 4 speech samples per talker have been collected per each 0.25MOS bin.

Thus, approximately 80 speech samples have been collected for each network.

**Subjective data**

The subjective test has been performed by Dynastat Lab, an ITU-T certified lab with more than 20 years of experience in subjective speech quality evaluation.

The lab performed an ACR test designed to feature four MOS panels that consisted of eight voters each. Therefore thirty-two voters scored each speech sample tested. This resulted in an average standard per individual MOS score of about 0.7MOS.

Details regarding the test procedure are presented in the DYNASTAT report [4].

## 2.2. Evaluation procedure

The evaluation considered three issues for analysis: the performance of the PESQ algorithm as implemented in TEMS Automatic; the accuracy of the PESQ implementation in TEMS Automatic; and PESQ tendency (behavior) within AMR live networks.

The evaluation procedure relies on two sets of data: the subjective scores obtained in the Dynastat test, and the objective scores obtained by the implementation of the ITU-T standard in the TEMS Automatic tool, on all the tested speech samples.

Three statistical metrics have been calculated in order to perform the evaluation. They are the correlation coefficient between the two sets of data, the prediction error, and the residual error distribution. [5]. A fourth metric, the residual error sign, has been introduced in order to obtain information regarding the algorithm's tendency for the AMR live network conditions, respectively overprediction or underprediction.

The Pearson correlation coefficient R (1) gives a basic measure for goodness of fit between objective and subjective scores, describing the extent to which data points are scattered with respect to the linear mapping  $y=ax+b$ .

$$r = \frac{\sum (SQM_i - \overline{SQM})(MOS_i - \overline{MOS})}{\sqrt{\sum (SQM_i - \overline{SQM})^2 \sum (MOS_i - \overline{MOS})^2}}, \quad i=1\dots N. \quad (1)$$

$MOS_i$  denotes the subjective score, and  $SQM_i$  denotes the objective score for the sample  $i$ .

The residual error (absolute error) (2) is calculated as the result of the application of the correlation line  $y=x$  to a measurement data set

$$Er = MOS_i - SQM_i. \quad (2)$$

The residual error distribution expresses the percentage of the situations for which the algorithm exhibits errors with values within a certain MOS bin. The MOS bins are uniformly distributed on the 1 to 5 subjective scale and are 0.25MOS wide. An accurate algorithm should exhibit errors below 0.75MOS in at least 90% of the situations.

The **prediction error** is given by (3) and it gives the average standard error of the objective estimator of the subjective opinion

$$E_p = \sqrt{\frac{\sum (MOS_i - SQM_i)^2}{N-1}} \quad i=1\dots N \quad (3)$$

$N$  denotes the number of samples considered in the analysis.  $MOS_i$  and  $SQM_i$  represent the subjective and objective scores, respectively.

#### **Accuracy of the PESQ implementation in TEMS Automatic**

In order to also verify the PESQ algorithm's implementation in TEMS Automatic, the PC version has been run on the same speech database and the obtained scores have been compared to the TEMS Automatic measurements.

#### **PESQ tuning on AMR live networks**

A tuning of the raw PESQ scores on the AMR live data has been performed and the improvements of the AMR tuned scores have been analyzed.

## 2.3. Results of the PESQ algorithm performance on AMR live networks

Three cases have been analyzed: the PESQ algorithm's output from TEMS Automatic (denoted pesqTA), the PESQ PC version's output (denoted pesqLQO) and the AMR tuned PESQ version's output (denoted pesqamr).

The statistical evaluation metrics calculated for all the three test cases showed that the PESQ algorithm meets the ITU-T expected performance for wireless live networks. Its performance lies within the ITU-T 95% confidence intervals of the statistical metrics (Table 1 and 2).

**Table 1**

Database	Metric	pesqTA	pesqLQO (PC version)	pesqLQO_AMR
AMR HR&FR (850Mhz&1900Mhz) live network and AMR clean coding conditions	Correlation	>0.85		
	Prediction error	< 0.45		
ITU-T expected performance for wireless live networks*	Correlation	0.85 (lower limit of the 95% confidence interval)		
	Prediction error	0.45 (upper limit of the 95% confidence interval)		

\* Values based on the performance of the P.862.1 on wireless live networks (see [2]).

**Table 2**

Database	Metric	MOS bins							
		<0.25	<0.5	<0.75	<1	<1.25	<1.5	<1.75	<2
AMR HR&FR (850Mhz&1900Mhz) live network and AMR clean coding conditions	pesqTA ; pesqLQO (PC version) ;pesqamr CDF(%)	>40	>80	>95	>98	100	100	100	100
ITU-T expected performance for wireless live networks*	CDF(%)*	40.44	70.48	90.33	97.71	99.3	99.7	99.91	100

\* Values based on the performance of the P.862.1 on wireless live networks (see [2]).

In order to evaluate the algorithm's tendency within AMR live networks, the over- and under- prediction error percentages have been determined based on the absolute residual error vales. The results are presented in the Table 3.

**Table 3**

Prediction percentage	pesqTA	pesqLQO	pesqamr
Under prediction	50.54%	51.08%	51.34%
Over prediction	48.66%	47.58%	48.12%

## 2.4. Comments on the performance and the accuracy of the PESQ algorithm as implemented in TEMS Automatic

The PESQ implementation in TEMS Automatic (pesqTA) shows a prediction error on AMR live networks within the ITU-T expected 95% confidence limits. The residual error distribution exhibits higher CDF(%) values than the ITU-T expected distribution, indicating that the algorithm performance on AMR live networks stays close to its performance on other wireless live networks tested within the ITU-T.

The analysis showed the same performance results for the PC version, which means that the TEMS Automatic implementation works properly, in accordance with the ITU-T implementation requirements.

The analysis of the algorithm's tendency, respectively the over- and under-prediction percentages (Table 3), showed that the algorithm is well balanced. Therefore, it is expected that on average, PESQ implemented in TEMS Automatic will not exhibit speech quality values that are either too pessimistic or too optimistic.

## 2.5. Comments on the AMR PESQ tuned solution

The results of the AMR tuned solution of the PESQ algorithm (Table 1, 2 and 3) show that the re-tuning (re-calibration) of the PESQ algorithm on the AMR live networks does not bring significant improvement of the results. In addition, since the test showed that PESQ in TEMS Automatic exhibits the ITU-T expected performance values, there is no reason to alter the standard.

### 3. Conclusions

A comprehensive test has been performed in order to evaluate the PESQ performance on AMR live networks. The evaluation showed that PESQ-TEMS Automatic performs within the ITU-T's expected 95% confidence interval limits. In addition, it has been verified that the PESQ-TEMS Automatic matches the PESQ-PC version and it is meeting therefore the ITU-T implementation requirements.

The PESQ algorithm exhibited very well balanced behavior, with equal over- and under-predicting percentages across the entire AMR live network database.

A PESQ AMR tuned version has been created and tested. The results showed that the improvements are not significant. Any re-tuning of the standard is therefore not meaningful, especially since the test showed that the performance values lie within the ITU-T expected limits.

#### References

- [1] 3GPP TR 26.975, "Technical Specification Group Services and System Aspects; Performance characterization of the Adaptive Multi-Rate AMR speech codec", Release 1999.
- [2] ITU-T SG12, P.662.1, February 2003
- [3] ITU-T SG12, P.862.2, October 2005
- [4] Dynastat Lab, "MOS Test Results for Ericsson", August 2005
- [5]. I.Cotanis, " ", ITU-T SG12 white contribution, January 2003.
- [6]. Cingular, "MOS Lab Test-RF Planning and Standard", report presented to TEMS, April 2005.