

The PESQ Algorithm as the Solution for Speech Quality Evaluation on 2.5G and 3G Networks

Technical Paper

1 Objective Metrics for Speech-Quality Evaluation

The deployment of 2G networks quickly identified development of a speech-quality metric as the first priority of the network's performance. Subscribers accustomed to PSTN speech quality obviously demanded the same quality for wireless networks. Networks designers, developers, and operators therefore demand drive-test tools that are capable not only of evaluating the network's RF performance, but also of estimating its speech quality.

Subjective tests providing the Mean Opinion Score of the subscribers on speech quality performed by the network were not capable of dealing with the real-time environment required for the evaluation of the network's performance and the network optimization process. At that time, a set of proprietary algorithms (such as Auryst – TEMS, 1996) and an International Telecommunication Union –Telephony (ITU-T) standard (PSQM, P.861, 1998) had been developed as objective estimators of subjective opinion. The ITU-T algorithm has been mainly designed, tested, and validated for the evaluation of speech quality shown by clean codecs conditions, namely encoding-decoding. Very few channel-error pattern conditions have been tested for ITU standard algorithm. On the other hand, the TEMS solution has been tested and validated for live networks and implemented in drive-test tools.

Years of optimization work performed on 2G networks using drive-test tools such as TEMS products brought the speech quality on these networks to an average quality level of 3.8 MOS. Field measurements showed that, depending on the topology of the covered area, a well-designed 2G network exhibits an average speech-quality ranking from MOS=3.6 to MOS=3.9. These values are fairly close to the ones shown by fixed networks.

For a while, network developers and operators believed that the speech quality provided by their networks was an accomplished task. This proved not to be the case when 2.5G and 3G networks began to become reality.

Evaluating speech quality over complex 2.5G and 3G networks has become one of the most arduous tasks in optimizing the communication environment. The difficulty of maintaining the user's expectation of wireline speech quality has emerged from the interaction of a set of factors such as the convergence of voice and data networks, increased demand of capacity, and low and adaptive bit rate codecs. The interaction among factors such as these upon advanced networks invariably produces new types of distortions that affect speech quality. The most common types of distortions characteristic to 2.5G and 3G networks that affect speech quality are linear distortions, packet loss, and variable delay. In addition, the adaptive and low rate of the codecs used on these networks can generate unexpected artifacts into speech. In order to cope with the new challenges, new devices and algorithms such as noise reduction, automatic gain control, and acoustic and network echo cancellers have been developed. The speech quality is very sensitive to the functionality of these devices and therefore, if not well designed and implemented, they could accentually have the opposite than the desired effect.

Due to this complexity, network developers and operators need to continuously monitor the speech quality provided by their converged voice and data networks.

Extensive work has been performed by both the ITU-T and the telecommunication industry in developing new speech quality (SQ) evaluation algorithms to cope with the degradations and the environment characteristic to 2.5G and 3G networks. Some of these algorithms (such as the ITU-T standard P.862 PESQ approved in 2001) are aimed at end-to-end speech-quality evaluation on all types of networks (wireless, VoIP, fixed). Single-ended algorithms (such as SQI, NiNa, NiQA, 3SQM, VQmon) which can evaluate speech quality at different nodes of the network, regardless of its type, also have been deployed. The main difference between the two types of algorithms is that one is an intrusive metric performing end-to-end speech quality evaluation and the other one is a non-intrusive metric estimating speech quality at the network nodes' level.

Besides the subjective estimation of the speech quality, these algorithms can perform speech cause analysis.

2 Intrusive and Non-intrusive Speech Quality Metrics

Intrusive algorithms provide speech quality scores by comparing reference (sent) and degraded (received) speech samples. Intrusive assessment techniques therefore require access to both transmission and reception ends of communication. Comparing reference and degraded speech samples facilitates an accurate estimation of the subjective perception of speech quality received by the terminal. An accurate estimation, however, is performed at the cost of sending the test samples through the network under test. The connection under test is therefore withdrawn from normal service and rendered unavailable to the customer. During peak hours, and for some technologies and certain areas, this situation may generate artificial lower quality scores.

Intrusive metrics estimate end-to-end speech quality, and thus are useful and meaningful to network operators who need to monitor the performance of their networks end-to-end.

Non-intrusive metrics continuously monitor the quality of speech delivered to the customer or the quality that exists at a particular node in the network. Non-intrusive metrics use the in-service signal to make predictions of speech quality. Using in-service signals instead of test stimuli, however, means that non-intrusive metrics can only *predict* speech quality.

Non-intrusive metrics can be network parameter based (such as SQI, VQmom) or speech based (such as 3SQM, NiNA, NiQA). The parametric methods do not use the speech signal, but only the RF and/or transport parameters for predicting the quality. The speech based methods need to use predictions regarding the sent original speech based on the degraded signal. Strong degradations could easily affect the accuracy of these predictions.

Both of these calculation procedures exhibit therefore lower accuracy than the method used by the intrusive metrics.

Although less accurate than the intrusive metrics, non-intrusive metrics could play an important role in the network's development stage. Once the network is up and running, intrusive metrics are recommended for speech-quality evaluation and for troubleshooting end-to-end performance problems.

3 The TEMS Approach to Speech-Quality Evaluation

For nearly a decade, the TEMS group has understood the importance of speech-quality evaluation in wireless networks. Pioneers in the drive-testing arena – and far ahead of any of the current drive test equipment providers – the TEMS group has implemented both non-intrusive (Speech Quality Index metric) [1] and intrusive (Auryst) [2], [3] metrics.

As a leader in the drive-testing arena, the TEMS group understood the need for a speech-quality metric that can handle conditions associated with 2.5G and 3G networks. As already mentioned at various forums and conferences [4], reduced attention, driven mainly by cost issues, to speech quality will be only an interim phase until the networks are up and running to their full capacity. At that point, network-performance evaluation will require a robust and accurate algorithm such as the PESQ ITU-T standard [5], which has been designed and calibrated [6] for all new types of degradation caused by converged voice and data networks. As an experienced contributor in the arena of speech-quality evaluation on live networks, TEMS staff participated in the standardization process of the calibrated PESQ algorithm [6].

4 The PESQ Algorithm as a 2.5G and 3G Intrusive Method for Speech-Quality Evaluation

The development of the PESQ algorithm was the result of an ITU competition for an objective speech-quality assessment metric to cope with 2.5G and 3G network degradations which affect quality as perceived by subscribers. The ITU competition declared PESQ the winner because it performed best for all the performance requirements and for all the applications imposed by ITU. After a comprehensive process of testing and validation against other SQ metrics on a very large corpus of speech samples, the PESQ algorithm became the ITU-T P.862 standard (February 2001). The test databases contained speech samples provided by different speakers in different languages (British and American English, German, Swedish, Dutch, Italian, French, Japanese). The speech material was represented by samples degraded by different types of network simulators and emulators, such as VoIP, wireless, and fixed. The validation databases contained live degraded VoIP and wireless network speech samples.

4.1 Overview of the PESQ algorithm

The PESQ algorithm was developed to accurately estimate the listening speech quality performed by wireless, VoIP, and fixed networks. PESQ algorithm's accuracy for assessing the degradation of speech quality caused by different types of networks emerged from the complex structure of the algorithm. Figure 1 (Figure 1/ ITU-T P.862 "Overview of the basic philosophy used in PESQ") represents a high-level description of the algorithm and its operability.

The idea of the PESQ algorithm (upper part within Figure 1) is to provide a speech-quality score based on the comparison between the original speech signal and the signal degraded by the network under test. A subject listening to the degraded speech signal (Absolute Category Rating listening test [7]) opinion should provide nearly the same quality score as the PESQ algorithm.

In order to perform this comparison, the PESQ algorithm uses psychoacoustic and cognitive models (lower part within Figure 1). Using a complex time alignment routine, the algorithm synchronizes the original and degraded speech signals. Misalignment of the speech samples could generate a false quality score.

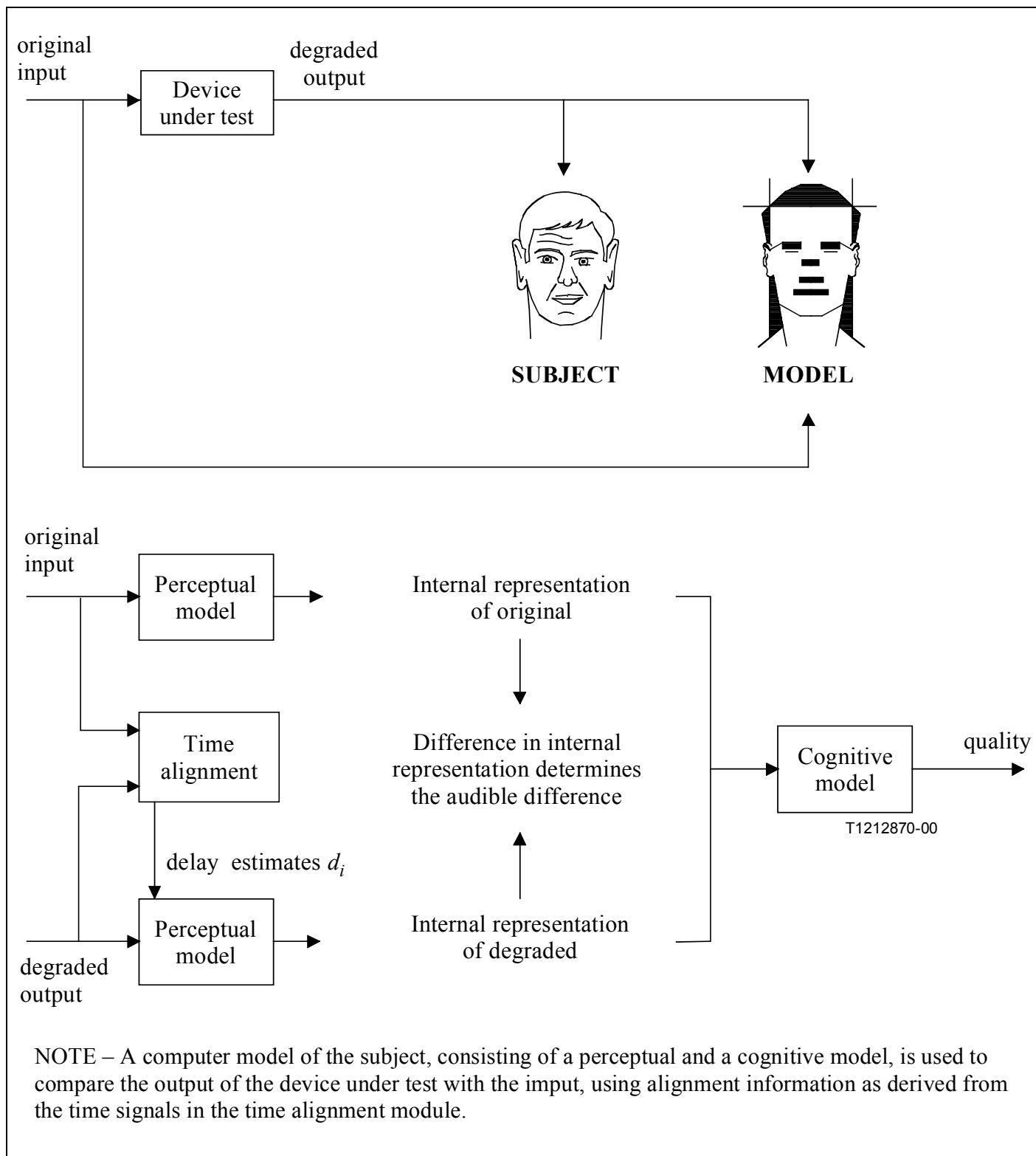


Figure 1 (Figure 1/ITU-T P.862)

The psychoacoustic model applied to both original and degraded signals is used to calculate the internal (or psychophysical) representation of the speech samples. The model performs the transformation from the physical domain represented by the speech signal to the psychophysical domain expressed by the human perception. This transformation is based on comprehensive speech processing, such as time-frequency transformation, Bark transformation, compensation of the transfer function equalization, compensation for the time gain variations between original and degraded signals, and calculation of the loudness transformation with recency effect (degradations heard at the end of the speech sample have a stronger perception impact than the ones heard at the beginning of the sample).

To compute the speech-quality degradation represented by the disturbance metric between the psychophysical representations of the reference and degraded speech samples, the cognitive model performs complex non-linear calculations. Examples of these calculations are time-frequency masking of the inaudible disturbance, asymmetric disturbance metric calculation (weighting of the additive frequencies generally introduced by the codec), weighting of the silence intervals, non-linear average over time and frequency of symmetric and asymmetric disturbance, calculation of a bi-dimensional (considering the symmetric and asymmetric disturbances) figure of merit that gives a single score, which represents the estimator of the speech quality. The raw output of the algorithm lies within the -0.5 to 4.5 range, where -0.5 represents the lowest quality.

4.2 The Calibrated Output

Providing scores in the range -0.5 to 4.5 does not allow a direct comparison to the 1 to 5 Mean Opinion Score (subjective) scale [7], where 1 represents the lowest quality.

The raw output of the PESQ algorithm shows a high correlation with the subjective opinion. However, to be implemented in drive-test tools for field measurements, the algorithm has to provide an estimator of the MOS score, which is on a scale 1 to 5 [8]. As a result, in September 2003, ITU approved the new ITU-T P.862.1 recommendation [6], which provides a universal calibration function of the PESQ algorithm to the subjective domain: the MOS scale 1 to 5.

The calibration process has these important advantages:

- The output of P.862.1 directly represents an MOS estimate.
- The PESQ algorithm's training has been extended to conditions for which it is used extensively and will continue to be used. This advantage emerged from the fact that the corpus of samples used for the calibration process contained real networks degradation conditions in a percentage higher than 50%. In addition, a large set of speech samples represented AMR coding and AMR error pattern conditions.

- The calibration process has increased the accuracy of the PESQ algorithm. Details regarding the performance are presented below.
- Different tools running PESQ produce comparable measurement results.

4.3 Algorithm's Performance

4.3.1 Evaluation Metrics

The evaluation of the algorithm's performance relies on two sets of data: the subjective scores obtained in a subjective test and the objective scores obtained by running the PESQ algorithm on a large corpus of speech samples.

Three statistical metrics [9] have been calculated in order to perform the evaluation.

- The Pearson correlation coefficient R with the subjective opinion MOS

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 $y=ax+b$.

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

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

- The residual error distribution that 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 0.5MOS wide. An accurate algorithm should exhibit errors below 0.75MOS in at least 90% of the situations.

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 prediction error E_P that gives the average standard error of the objective estimator of the subjective opinion

The prediction error is given by (3), below:

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

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

The prediction error evaluative statistic emerged from wireless market demand [8]. Network providers, designers, and operators are users of drive-test tools who need to have not only an estimator for the perceived speech quality, but the average evaluation error as well. The E_p statistic is normally calculated for the specific service under test, that is, over a range of impairments.

Market performance requirements for the prediction error are very tough, especially when it comes to drive-test tools used for network comparisons. Besides knowing the network performance within a 95% confidence interval, users definitely want to know how their network is ranked in comparison with the others or how is one market performing versus another. Benchmarking is also used to assess which of the network's link directions performs better. An acceptably accurate ranking requires an objective estimator with an average prediction error, at 0.5MOS or lower. The release of new models of wireless phones also requires a low E_p and a fine rank-discrimination capability in order to accurately evaluate its perceived impact on the network.

4.3.2. PESQ Performance Results

Table 1 and Table 2 present PESQ performance results for a large corpus of samples, which comprise all types of applications (VoIP, wireless-all technologies, fixed networks) along with simulated and field-collected samples (in each table, see the column labeled "Overall"). Details regarding the content of the databases are presented in [10].

The results for calibrated PESQ (ITU-T P.862.1) are presented to show the improvement of the accuracy determined by the calibration process. The calibrated values are directly comparable with MOS. The performance of the calibrated PESQ is the only means of showing the accuracy of the implemented PESQ algorithm in the drive-test tools.

In Table 1, it should be noted that the correlation coefficient drops below 90% and that the 95% Confidence Interval (CI) lower limit is 85.5%. Due to the very broad range and multiple types of conditions, this result is expected. The prediction error E_p, however, shows values below 0.45MOS. The highest 95%CI limit also does not exceed 0.5MOS.

Table 1 also presents the results for databases containing only field data collected in VoIP and wireless networks. The field-collected samples cover live network conditions; they have real significance from the perspective of a practical implementation. In addition, the calibrated PESQ performs even better for field-collected data than for data in the "overall" category. The correlation coefficient higher than 90% exceeds the ITU-T recommended value of 85% for field collected data. A prediction error of 0.42MOS is recorded for P.862.1 on the field collected speech corpus.

Table 2 shows the residual error distribution for the two databases presented in Table 1. The residual error distribution shows the high accuracy of the calibrated PESQ algorithm for both cases, overall and field. Absolute errors below 0.75MOS are exhibited in more than 90% of the situations. In addition, a fairly high percentage (above 70%) is recorded for residual errors below 0.5MOS.

These performances make calibrated PESQ an accurate metric, not only for evaluating real networks, but for benchmarking networks and markets from the perspective of speech quality. Due to its high accuracy, PESQ is a perfect fit for drive-test, autonomous, and benchmarking tools.

Table 1

Application	Metric	P.862 (raw PESQ)	P.862.1 (calibrated PESQ or PESQ-LQO)
Overall of 2493 samples (wireless, VoIP, fixed, simulated and field)	R	0.876	0.879
	CI95%-lower limit	0.855	0.86
	Ep	0.492	0.441
	CI95%-upper limit	0.501	0.449
Field collected data of 1135 samples (Wireless networks: GSM US and EU, CDMA- US, TDMA-US, iDEN-US, AMPS-US; VoIP networks)	R	0.925	0.926
	CI95%-lower limit	0.897	0.901
	Ep	0.479	0.42
	CI95%-upper limit	0.492	0.431

Table 2

Application	MOS bins	<0.25	<0.5	<0.75	<1	<1.25	<1.5	<1.75	<2
Overall of 2493 samples	P.862 (raw PESQ) (%)	36.1	66.63	87.44	96.95	99.56	99.96	99.96	100

Application	MOS bins	<0.25	<0.5	<0.75	<1	<1.25	<1.5	<1.75	<2
(wireless, VoIP, fixed, simulated and field)	P.862.1 (calibrated PESQ) (%)	41.92	72.64	91.22	98.4	99.64	99.88	99.96	100
Field collected data of 1135 samples	P.862 (raw PESQ) (%)	32.51	66.52	90.84	97.97	99.38	99.91	99.91	100
(Wireless networks: GSM US and EU, CDMA- US, TDMA-US, iDEN-US, AMPS-US; VoIP networks)	P.862.1 (calibrated PESQ) (%)	40.44	70.48	90.33	97.71	99.3	99.7	99.91	100

4.3.3. PESQ on UMTS Networks

The PESQ standard (P.862 and P.862.1) 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. Since the standard has 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.

Therefore, the evaluation of the PESQ algorithm’s performance on live UMTS is required. TEMS designed and performed an evaluation test [12]. Speech samples in US UMTS networks running both AMR HR and AMR FR codecs and both 850Mhz and 1900Mhz bandwidths have been collected. PESQ and subjective speech-quality scores obtained for the live UMTS databases have been compared.

The calculated correlation coefficient met the ITU-T recommended value of 85%. In addition the prediction error showed that the PESQ algorithm meets the ITU-T expected value of 0.45MOS characteristic to the wireless live networks (see Table 2). The evaluation of the residual error distribution (Table 3) shows that the error is lower than the expected ITU-T value for all MOS bins (Table 2).

Table 3

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 database of 420 speech samples	PESQ – LQO (CDF%)	>40	>80	>95	>98	100	100	100	100

Database	Metric	MOS bins							
		<0.25	<0.5	<0.75	<1	<1.25	<1.5	<1.75	<2
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 Table 2, [6]).

4.4 The PESQ Algorithm Implemented in TEMS Products

It has been shown that the PESQ algorithm represents the most accurate solution for the implementation of drive-test and autonomous tools. As a result, TEMS selected PESQ as the solution for speech-quality evaluation on 2.5G and 3G networks. Details regarding the implementation and its extensive testing are presented in [13]. Implementing the PESQ algorithm in TEMS products has some valuable advantages for network-performance evaluation:

- The ability to digitize and record speech files by configuring the measurement setups. This feature allows creating valuable speech databases that could be used for troubleshooting, optimization and further improvement of the speech quality metric.
- The ability to configure and control the recordings so that whenever the scores fall below a defined quality threshold, the files start to be recorded. Once the scores reach above the threshold, the recording stops. Accurate and straightforward troubleshooting is possible by correlating the speech quality results with RF and/or transport parameters information. Replaying the files that were recorded in a problem-generating area helps evaluate and identify the magnitude and the type of the impact on the subscribers' perception of the network performance.
- The ability to generate advanced statistical reports of the speech quality performed by the network under test within a market, a submarket and/or a specific area of interest, during different time windows of the day, of the week and/or of the month. Average values along with their 95% confidence intervals; the time, space and statistical distribution of the speech quality scores within an area during a defined time window allow accurate and consistent monitoring of the network and its benchmarking across markets as well. Implemented statistical significance tests of the speech quality measurements allow a controlled monitoring, ensuring work efficiency and low costs.

5 Additional Features of the PESQ Algorithm

Network designers develop new devices and algorithms, such as noise reduction, automatic gain control, and acoustic and network echo cancellers, aimed at coping with the new challenges generated by the 2.5G and 3G networks. Therefore, estimating the perceived speech quality for these networks requires more than a single number that represents subjective opinion, at which point a cause analysis of the speech-quality degradation starts to become more useful.

Speech-quality diagnosis helps to detect problems associated with the operation of the network's devices, such as Voice Quality Enhancement devices (noise reduction, automatic gain control), adaptive and low rate codecs, and mobile terminals (generally the frequency shaping implemented in the phone). Details are presented in [11].

A cause analysis could be performed by post-processing interim measurements of the PESQ algorithm. Examples of these types of calculations, potential troubleshooting uses for them, and parties that might be interested in this feature are presented in Table 4.

Table 4

Analysis Type	Troubleshooting Uses	Interested Parties
Signal level analysis	<ul style="list-style-type: none"> • Detecting and analyzing the VQE devices operation • Analyzing codecs • Detecting circuit noise 	<ul style="list-style-type: none"> • Network designers • Developers of speech processing algorithms
Temporal analysis	<ul style="list-style-type: none"> • Detecting malfunctioning Voice Activity Devices and AMR vocoders • Analyzing HOs in wireless networks • Analyzing muting caused by VAD or packet loss • Evaluating the impact of the jitter and its effect on the end-to end audio delay (VoIP networks) 	<ul style="list-style-type: none"> • Network designers, developers and operators • Developers of speech processing algorithms
Spectral analysis	<ul style="list-style-type: none"> • Analyzing codecs • Evaluating low bit codecs • Detecting packet loss • Evaluating handset characteristics 	<ul style="list-style-type: none"> • Network designers, developers and operators • Developers of speech processing algorithms • Phone developers and vendors
Speech quality distribution within speech samples	<ul style="list-style-type: none"> • A very accurate procedure for troubleshooting network parameters that affect the speech quality 	<ul style="list-style-type: none"> • Network designers, developers and operators
Processing of PESQ output and some of its interim calculations	<ul style="list-style-type: none"> • Correlating speech quality measurements with the values predicted by planning tools (E-model planning tool, ITU-T G.107) 	<ul style="list-style-type: none"> • Network designers

As can be seen in Table 4, speech diagnosis could represent a powerful and valuable feature for network designers, developers, and operators, as well as for developers of speech-processing algorithms. Reference [11] presents results of speech diagnoses using the PESQ algorithm and discusses solutions for implementing these features of the PESQ algorithm in TEMS products.

References

- [1]. A.Karlsson, G.Heikkila, T.B. Minde, M.Nordlund, B.Timus, "Radio Link Parameter Based Speech Quality Index-SQI", IEEE Workshop on Speech Coding, 1999, Finland
- [2]. S. Quakenbush, T.Barnwell, "Objective Measures of Speech Quality", Georgia Institute of Technology, Prentice Hall, 1993
- [3]. Irina Cotanis, "Speech Quality Evaluation for Mobile Networks" IEEE conference – ICT 2001, June 2001
- [4]. Qualcomm, Nortel, Agilent, "End to End Voice Performance on CDMA EV-DO networks", Plenary Session, CDMA2000 Forum, 8-10 Dec. 2003, Miami, FL.
- [5]. J.Beerends, A.Rix "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs", ITU-T P.862 rec., February 2001.
- [6]. I.Cotanis, T.Goldstein, V. Matilla, "Mapping function for transforming P.862 raw result scores to MOS-LQO", ITU-T P.862.1 rec., September 2003
- [7]. ITU-T P.800 Rec. "Methods for subjective determination of transmission quality"
- [8]. Irina Cotanis, John Morfit, "Mapping the PESQ Algorithm to the MOS domain", ITU-T white paper, January 2003.
- [9]. Irina Cotanis, "The ITU-T P.862 Standard Metric and Other Speech Quality Metrics and Tools", TEMS white paper, November 2003.
- [10]. I.Cotanis, T.Goldstein, V. Matilla, "Results of the PESQ algorithm mapping", ITU-T white paper, September 2003
- [11]. Irina Cotanis, "Speech in the VQE Environment" TEMS white paper, July 2002 and presentation to CDG2002 Forum, IEEE-WCNC2003 Conference-CTIA
- [12] Irina Cotanis, "The Performance of the ITU-T P.862.1 Standard (PESQ-LQO) on AMR Live Networks", TEMS technical paper, Nov. 2005
- [13]. Per Johansson, "PESQ Algorithm in TEMS Automatic", TEMS technical paper, 2003.