The Perceptual Objective Listening Quality Assessment Algorithm in Telecommunication

Introduction of ITU-T new metrics POLQA

Yi Gaoxiong, Zhang Wei
China Telecommunication Technique Labs
China Academy of Telecommunication Research of MIIT
Beijing, China
yigaoxiong@chinattl.com; zhangwei@chinattl.com

Abstract— **In order to accomplish speech quality evaluation tasks in an efficient and economical way, objective computational models which simulate human hearing characteristics of speech perception have been intensively researched in the past decades. Several models have already been developed and standardized in telecommunication industry to assess the listening quality of speech signals transmitted through telecommunication network or generated by voice-communication terminals. Among those standardized models, POLQA, short for Perceptual Objective Listening Quality Assessment, is the newest generation objective metrics designed to address the shortcoming and weakness existed in previous published models such as PESQ and PSQM.**

This paper gives an overview of POLQA characteristics and describes the functional modules utilized in POLQA at the algorithm level. Furthermore the performance of this new model evaluated by standardization group, which reflects the correlation between objective scores outputted by this model to subjective scores, is also shown.

Keywords-POLQA; PESQ; PSQM; Objective speech quality assessment

I. THE HISTORY OF SPEECH QUALITY ASSESSMENT

A. Subjective speech quality assessment

When talking about objective speech quality, subjective speech quality assessment, which is the fundamental basis of objective speech quality assessment, has to be discussed at first. There is no physical and precise definition of what the speech quality is, at most of time what people have when hearing the speech signals is an overall opinion that it sounds good or bad. This judgment is an instantaneous reaction of human brain activities based on its experience, which could be influenced by a large amount of factors including not only the emotional status, the educational background, the expectations of the listener but also the understanding of quality itself. Furthermore the judgment on a same speech signal may vary from people to people and from time to time.

For many years in telecommunication industry the only effective way to get a reasonable assessment of speech quality is asking a panel of telephone users what is their opinion on the speech, quantified with a 0 to 5 point scoring scheme under restricted conditions and afterwards average the output scores to get a Mean Opinion Score (MOS). This so called subjective speech quality assessing process is proved to be useful and have already been standardized in various telecommunication scenarios by ITU-T in [8] [9] [10].

B. Objective speech quality assessment

To ensure delivering good quality speech to telephone users, telecommunication network operators and terminal manufacturers are the natural parties in favor of speech quality assessment. However running subjective assessments requires a large amount of time and effort in planning the test and controlling those factors which might bias the final result, meanwhile the cost to form a listener panel is also a restriction. As a result objective methods which can assess speech quality in a repeatable way with economical cost are certainly highly attractive.

At first the simple sweeping tone-base measurement is used to evaluate the transmission quality before the last 1980's. After that compression technologies and digital codecs were introduced into telecommunication network so that a higher transmission capacity could be achieved. In this new network with digital speech processing techniques it was found out that the tone-base measurement could not be used anymore because the outcome of measurement will usually somehow deviate from user's experience. As a result, a new measurement metric took advantage of the computer-based program, PSQM algorithm [11] was developed in 1996 to produce objective speech quality scores denoted as MOS-LQO (Listening Quality Objective) similar to the subjective MOS scores, with a concept of creating a model that simulates the human hearing to extract the difference between a pair of reference and degraded or transmitted speech signals and afterwards map those differences to subjective scores.

Apparently this model was still far from perfection so based on feedbacks research works continued. In 2001, PESQ algorithm [12] which significantly improved the hearing model was published as a better metrics than PSQM. With the subsequent amendments and algorithm extensions, for several years PESQ was used as the most popular tools in evaluating speech quality in telecommunication networks at that time. However experts found out that it still had limitations in certain scenarios such as variable rate codecs, time warping, frequency response variation and speech enhancement processing. Furthermore, due to the inner hearing model design, this assessment metric was not suitable to evaluate the speech quality at acoustic interface, which referred to the case that the degraded speech was acoustically recorded at the terminal receiver or loudspeaker. This limitation was addressed to some extent by other objective models such as TOSQA [3] and TOSQA 2001 [4]. However these models failed to be accepted as international standards.

With the development of even powerful speech processing algorithm used in telecommunication and the fast growing technique trending that telephone speech communication is moving from narrowband to wideband and even super wideband, PESQ again showed its weakness and incapability in handling these new speech transmission scenarios. At 2006 ITU-T initiated a research program to address all the remaining problem of PESQ and subsequently outputted a new ITU-T speech quality assessing model in 2012 and published as P.863 (Perceptual Objective Listening Quality Assessment, POLQA), which is introduced in this paper. With this new model, more accurate results could be expected when assessing the speech quality delivered by telecommunication network and terminals.

II. Overview of POLQA Algorithm

Traditionally the POLQA algorithm still makes use of the comparison between the reference and degraded speech signals. To overcome the limitations of former generations of objective assessment models, POLQA works in two different operational modes: Super wideband mode, and Narrowband mode.

In super wideband mode, the model will compare the degraded speech signal with a super wideband reference. The model will output an objective score as the subjective listening quality perceived by human listeners when the degraded speech signal is presented at both ear using a diffuse-field equalized headphone. Any speech degradation including band-limiting would be treated as adverse impacts and taken into consideration by the model. Theoretically degraded speech from acoustic interface will be evaluated accurately. With this mode speech quality delivered by terminals could be assessed.

In narrowband mode, the degraded speech signal would be compared with a narrowband reference speech. Consequently, the band-limitations from 300 to 3400 Hz would not be treated as degradations anymore. This narrowband mode offers the backward compatibility to PESQ, which modeled the listening quality as perceived by human listeners when the degraded speech signal is presented at one ear using an IRS type handset. With this mode speech quality delivered by traditional networks and narrowband digital codecs could be assessed.

Overview of this algorithm is shown in Figure 1.

The first step of POLQA processing is temporal alignment of the reference and degraded signal to ensure the following processing in core model is based on an accurate comparison of the same speech segment in two signals. To deal with the piece-wise fixed delay and time-variant delay, the temporal alignment is carried out in several steps:

- Firstly the initial delay of the reference signal to degraded signal is found out with a histogram based cross-correlation function at frame level, and general speech section information is obtained with a VAD algorithm;
- Secondly, based on the initial delay and section information, a coarse alignment is performed using a Viterbi-like backtracking algorithm around the beginning and ending part of speech sections. It still operates at a frame level however at a step by step increasing frame resolution;
- At last, a fine alignment is performed with the maximum possible resolution with delay express by samples based on the result given by coarse alignment. After this operation, a section grouping is carried out by joining consecutive small sections with same delay.

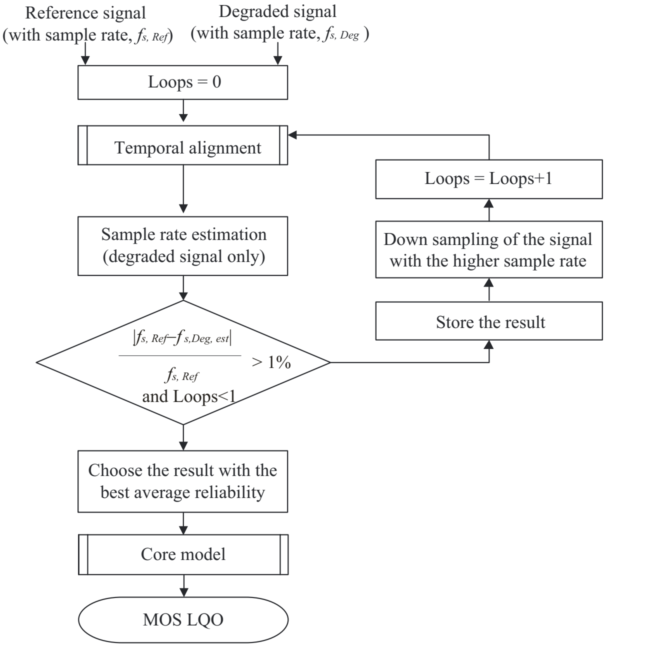
Reference signal
(with sample rate, $f_{s,Ref}$)
Degraded signal
(with sample rate, $f_{s,Deg}$)
Loops = 0
Temporal alignment
Loops = Loops+1
Sample rate estimation
(degraded signal only)
Down sampling of the signal with the higher sample rate
Store the result
$\frac{|f_{s,Ref}-f_{s,Deg,est}|}{f_{s,Ref}} > 1\%$
and Loops<1
Choose the result with the best average reliability
Core model
MOS LQO

Figure 1. Overview of the algorithm of POLQA.

If sampling rate difference is detected between both signals base on the result of temporal alignment, the signal with higher sampling rate will be down sampled towards another signal until an acceptable sampling rate difference is reached. This re-sampling operation will compensate the perceptually irrelevant differences in speech play out speed, which may be resulted by various intentional or unintentional reasons. (e.g., time scaling by jitter buffer adaptation or unsynchronized A/D or D/A converters).

III. The Detailed Processing of POLQA Perceptual Model

After the signal alignment step, the reference and degraded speech are sent to the core model of POLQA. The two signals will be transformed into internal expressions which approximate the human hearing characteristics to speech. Based on this expression a disturbance analyzing will be made to reflect the difference between two signals. The detailed block diagram of POLQA core model is shown in Figure 2.

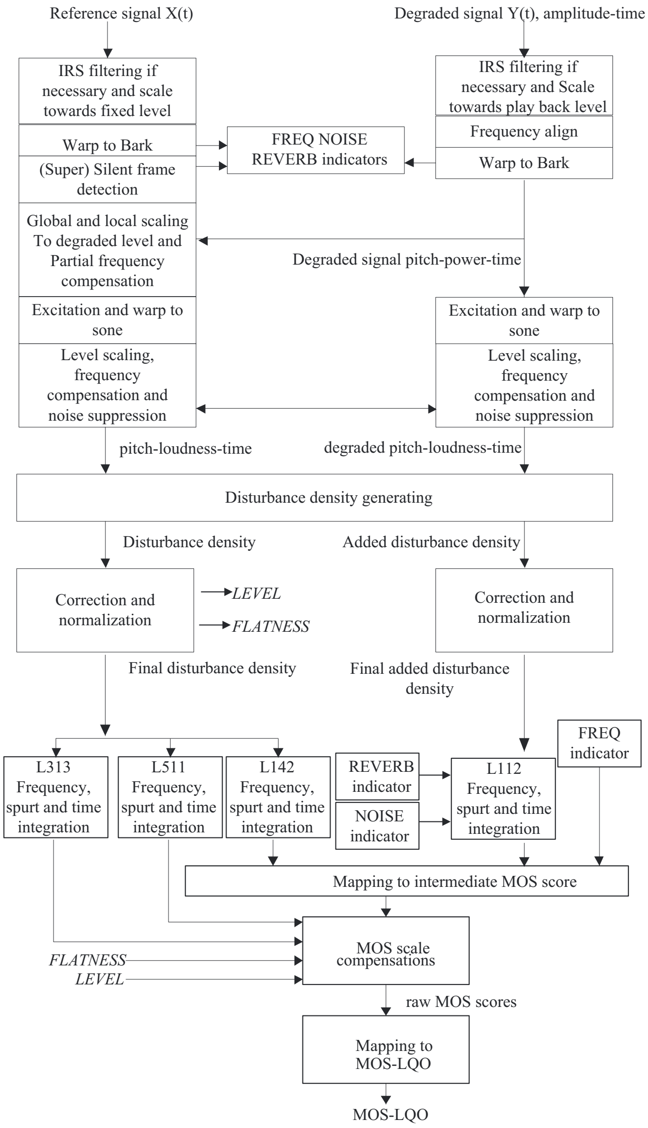
Reference signal X(t)
Degraded signal Y(t), amplitude-time
IRS filtering if necessary and scale towards fixed level
IRS filtering if necessary and Scale towards play back level
Warp to Bark
FREQ NOISE REVERB indicators
Frequency align
(Super) Silent frame detection
Warp to Bark
Global and local scaling To degraded level and Partial frequency compensation
Degraded signal pitch-power-time
Excitation and warp to sone
Excitation and warp to sone
Level scaling, frequency compensation and noise suppression
Level scaling, frequency compensation and noise suppression
pitch-loudness-time
degraded pitch-loudness-time
Disturbance density generating
Disturbance density
Added disturbance density
Correction and normalization
LEVEL
FLATNESS
Correction and normalization
Final disturbance density
Final added disturbance density
FREQ indicator
L313 Frequency, spurt and time integration
L511 Frequency, spurt and time integration
L142 Frequency, spurt and time integration
REVERB indicator
NOISE indicator
L112 Frequency, spurt and time integration
Mapping to intermediate MOS score
MOS scale compensations
FLATNESS
LEVEL
raw MOS scores
Mapping to MOS-LQO
MOS-LQO

Figure 2. Detailed processing diagram of POLQA.

In the processing chain, the core perceptual model is break down into three major steps: expression transformation, disturbance analyzing and arithmetic integration.

A. Step 1: Expression transformation

In this step, the reference and degraded speech signal is transformed from temporal domain to internal pitch-loudness-time domain expression.

At first the reference and degraded speech are filtered with IRS receiving characteristics if POLQA operate in narrowband mode. The reference signal is scaled to a fixed level of -26dBov as measured by [7] and meantime the degraded signal is scaled to a play back level. This play back level is the same with the level used by the listening panel when creating subjective training database, which is 79dBSPL for narrowband speech with monotic presentation and 73dBSPL for super-wideband speech with diotic presentation.

After level adjustment the two temporal signals are transformed by FFT, and frequency alignment is performed in spectra-time domain to avoid very narrow frequency response distortion by adding a sliding smoothing window of 100Hz length at the power spectra of both signals. To reflect the human hearing system which has a finer frequency resolution at low frequencies than at high frequencies, the frequency axis is warped towards pitch scale in Bark as described in [1]. From the resulted two pitch-power-time expressions three indicators are derived. The FREQ indicator which quantifies the impact of frequency response distortions, the NOISE indicator which quantifies the impact of additive noise calculated from the non-speech sections and the REVERB indicator which reflects the impact of room reverberations.

Base on the previously obtained pitch-power-time expressions, partial compensation of the original pitch power density for linear frequency response is performed to deal with the non-audible linear frequency response distortions which are introduced by filtering in the system under evaluation. Both time and frequency domain hearing masking scheme are taken into account using a convolution approach as given in [2]. The pitch-power-time expressions are transformed to pitch-loudness-time domain by warping the power intensity into loudness scale using a modified version of Zwicker's power law [17]. After that small compensations are performed on pitch-loudness-time domain to eliminate differences resulted by inaudible effects such as slow gain variation and linear frequency distortion.

B. Step 2: Disturbance analyzing

Base on the pitch-loudness-time expression, two functions indication the difference between the reference and degraded speech signal is calculated. The first one is called disturbance density obtained from the difference between the reference pitch-loudness-time and degraded pitch-loudness-time expression. The second one called added disturbance density and is derived from the reference pitch-loudness-time and a degraded pitch-loudness-time functions optimized only within speech sections where the degraded power density is larger than the reference power density.

An indicator LEVEL is derived from the signal level of the degraded signal to quantify severe level deviations from the optimal listening level as used in creating subjective training database. Another indicator FLATNESS is derived from the ratio of the upper frequency band loudness and the lower frequency band loudness to quantify the severe deviations from the optimal timbre. After that, the final disturbance and added disturbance densities are compensated for these severe amounts of specific distortions besides level and timbre such as the noise level variation and jumps in loudness.

C. Step 3: Arithmetic integration

From the final disturbance density and added disturbance density, four disturbance indicators are calculated from frequency, spurt and time integration. This integration converts the three dimensional disturbance functions into one explicit numeric indicator. Denote the integration method as “Lxyz” and the disturbance density as $D(f)_n$, the disturbance indicators D are derived as equation (1) (2) and (3):

$$D_n = \sum_{f=1,..Numberof\ Barkbands} | D(f)_n |^x W_f \qquad (1)$$

With W_f denoting a series of constants proportional to the width of the Bark bins, and

$$DS_n = \sqrt[y]{\frac{1}{6} \sum_{m=n,..n+6} D_m{}^y} \qquad (2)$$

$$D = \sqrt[Z]{\frac{1}{numberOfFrames} \sum_{n=1,..numberOfFrames} DS_n{}^Z} \qquad (3)$$

The disturbance indicator L112 obtained from added disturbance is compensated using the REVERB and NOISE indicators, and then combined with FREQ and another indicator L142 to get an internal indicator derived by a third order regression polynomial to get a MOS-like intermediate indicator.

After getting this intermediate indicator, the raw POLQA score is obtained by take into consideration of other compensation indicators as follows:

- Two disturbance indicators, one calculated with an L511 aggregation over frequency, spurts and time, and one calculated with an L313 aggregation;
- The LEVEL indicator;
- The FLATNESS indicator.

During the developing phase of POLQA, the training of the model parameters is carried out on a large speech database including various kinds of subjective degradations. The compensated raw MOS scores are then mapped to MOS-LQO using a third order polynomial which is optimized on the ITU-T evaluating dataset. In narrowband mode, the maximum POLQA MOS-LQO score could reach is 4.5 if comparing two identical ideal reference speech signals, however in super-wideband mode, this possible maximum score is 4.75.

IV. Performance of POLQA

A. Evaluating critera

During the developing phase of POLQA model, evaluation was carried out based on a so-called epsilon-insensitive r.m.s.e criteria denoted as rmse*, which took into consideration of the confidence interval of the subjective MOS score obtained by listening panel. The rmse* is calculated as equation (4) and (5) through the entire evaluating database and gives the impression how the MOS-LQO prediction error exceeds the ci95 of MOS.

$$P_{error}(i) = \max(0, \left| MOS(i) - MOS_{LQO}(i) \right| - ci_{95}(i)) \qquad (4)$$

$$rmse* = \sqrt{\left(\frac{1}{N-d} \sum_{N} Perror(i)^2 \right)} \qquad (5)$$

Due to the fact that subjective tests will always exhibit a broad range of scores for a given speech condition in different experiments despite that reference conditions and careful test design balancing have already been taken to ensure a stable subjective score range, the very nature of objective assessing model, that gives only one absolute score on a specific condition, decided that a compensation must be taken to eliminate the methodology bias between subjective and objective scores on specific speech database. Before calculating the prediction error, the subjective scores must be mapped to the scale of objective scores using a linear-regression function.

$$y = a + bx \qquad (6)$$

Where a, b are the one-order linear-regression coefficients between subjective scores and objective scores on the specific database. The effect of this mapping is shown in figure below:

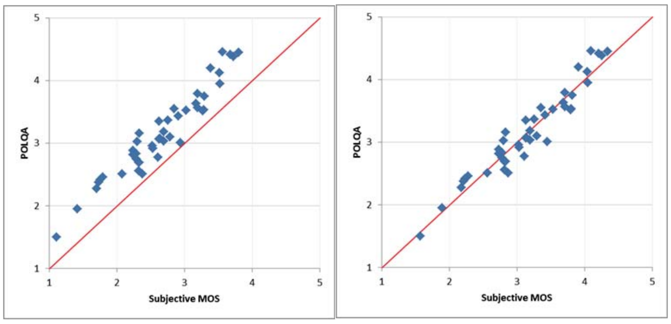

Figure 3. Example of subjective scores before and after 1-order mapping.

B. The performance on evaluation database

The performance results based on rmse* shown here is carried out by ITU-T experts on 64 speech databases. Table I and II reporting the narrowband and super wideband test results are extracted from meeting documents [5] and [6] of ITU-T POLQA developing group.

TABLE I. Narrowband Test Results from ITU-T Evaluation

Database(NB)	rmse* 1st	Database(NB)	rmse* 1st
BT_P862_BGN_ENG	0.1284	ITU_SUPPL23_EXP1o	0.1221
BT_P862_PROP	0.1691	ITU_SUPPL23_EXP3a	0.2036
16kHz_HUAWEI_1	0.1304	ITU_SUPPL23_EXP3c	0.0944
16kHz_HUAWEI_2	0.2226	ITU_SUPPL23_EXP3d	0.0654
8kHz_TEMS_ASCOM	0.1491	ITU_SUPPL23_EXP3o	0.0623
8kHz104_ERICSSON	0.2798	LUC_P563_PROP	0.1202
8kHz404_PSYTECHNICS	0.1657	NTT_PTEST_1	0.0908
8kHz504_SWISSQUAL	0.2386	OPT_P563_PROP	0.1188
ATT_iLBC	0.2305	PSY_P563_PROP	0.1848
DT_P862_1st	0.1827	QUALCOMM_EXP1b	0.1248
DT_P862_BGN_GER	0.1124	QUALCOMM_EXP2b	0.1493
DT_P862_Share	0.0886	QUALCOMM_EXP3w	0.0982
ERIC_AMR_4B	0.1558	QUALCOMM_EXP4	0.1275
ERIC_Field_GSM_EU	0.1546	QUALCOMM_EXP6a	0.2136
ERIC_Field_GSM_US	0.1533	QUALCOMM_EXP6b	0.1372

Database(NB)	rmse* 1st	Database(NB)	rmse* 1st
ERIC_P862_NW_MEAS	0.1767	SQ_P563_PROP	0.1800
FT_P563_PROP	0.0644	TNO_P862_KPN_KIT97	0.2074
GIPS_EXP1	0.1267	TNO_P862_NW_EMU	0.1563
ITU_SUPPL23_EXP1a	0.1213	TNO_P862_NW_MEAS	0.1722
ITU_SUPPL23_EXP1d	0.0665		
Average POLQA rmse* 1st : 0.1473			

TABLE II. Super Wideband Test Results from ITU-T Evaluation

Database(SWB/WB)	rmse* 1st	Database(SWB/WB)	rmse* 1st
48kHz101_ERICSSON	0.2870	48kHz601_TNO	0.2175
48kHz103_ERICSSON	0.2270	48kHz602_TNO	0.1881
48kHz201_FT_DT	0.2950	48kHz603_TNO	0.1580
48kHz202_FT_DT	0.2461	GIPS_EXP4	0.0794
48kHz203_FT_DT	0.2875	WB_16kHz204_FT_DT	0.2338
48kHz301_OPTICOM	0.2844	WB_16kHz402_PSYTECHNICS	0.1831
48kHz302_OPTICOM	0.1971	WB_48kHz102_ERICSSON	0.1926
48kHz303_OPTICOM	0.1774	WB_GIPS_EXP3	0.1522
48kHz401_PSYTECHNICS	0.1486	WB_NTT_PTEST_2	0.2719
48kHz403_PSYTECHNICS	0.1671	WB_QUALCOMM_EXP1w	0.1386
48kHz501_SWISSQUAL	0.1914	WB_QUALCOMM_EXP3w	0.1130
48kHz502_SWISSQUAL	0.2597	WB_QUALCOMM_EXP5	0.1422
48kHz503_SWISSQUAL	0.1949		
Average POLQA rmse* 1st : 0.203			

From Table I and II, i the average POLQA rmse* is nearly no more than 0.2 point of MOS on these evaluation database.

To illustrate the highest possible deviations of objective scores from subjective results, the best case and worst case performance of POLQA in narrowband mode and super wideband modes on the 64 databases are also shown respectively in Figure 4 and 5.

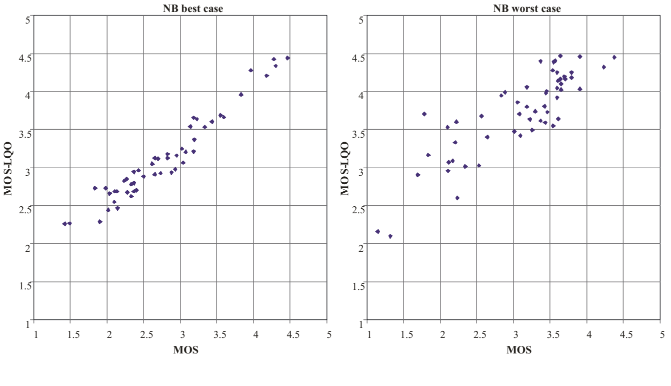

Figure 4. Best and worst case POLQA performance in Narrow band.

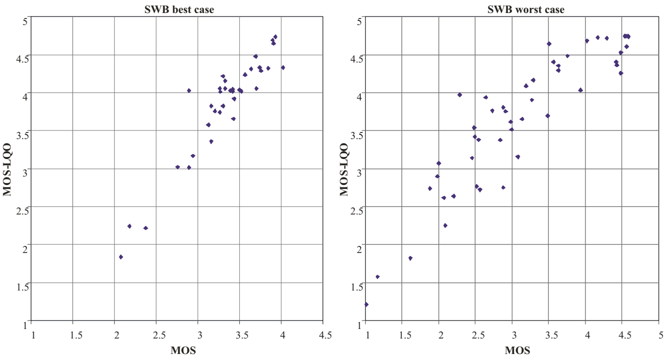

Figure 5. Best and worst case POLQA performance in Super wideband.

V. Conclusion

In this paper the basic concept and process diagram of POLQA is briefly introduced. Base on this model, performance and application scenario validation works are still ongoing in ITU-T standardization group.

However, from the results already reported by developing group, this model shows promising performance and hopefully in the near future is the best tool for objective speech quality assessment in contemporary telecommunication scenarios.

References

[1] Beerends, J.G. (1989), *Pitches of simultaneous complex tones, Chapter 5: A stochastic subharmonic pitch model*, Ph.D. dissertation, Technical University of Eindhoven, April 1989.

[2] Beerends, J.G. and Stemerdink, J.A. (1994), *A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation*, Journal of the Audio Engineering Society, vol. 42, No. 3, March, pp.115-123.

[3] Berger, J. *TOSQA – Telecommunication objective speech quality assessment*, ITU-T COM 12-34, December 1997.

[4] Berger, J. *Results of objective speech quality assessment of wideband speech using the advanced TOSQA2001*, ITU-T COM 12-19, December 2000.

[5] Berger, J, Folkesson, M. and Grancharov, V. *P.OLQA validation cross check*, ITU-T TD 12-376, September 2010

[6] Berger, J and Varga, I. *P.OLQA validation cross check results*, ITU-T TD 12-375, September 2010

[7] Recommendation ITU-T P.56 (2011), *Objective measurement of active speech level*.

[8] Recommendation ITU-T P.800 (1996), *Methods for subjective determination of transmission quality*.

[9] Recommendation ITU-T P.830 (1996), *Subjective performance assessment of telephone-band and wideband digital codecs*.

[10] Recommendation ITU-T P.835 (2003), *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*.

[11] Recommendation ITU-T P.861 (1998), *Objective quality measurement of telephone-band (300-3400 Hz) speech codecs*.

[12] Recommendation ITU-T P.862 (2001), *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*.

[13] Zwicker, E. and Feldtkeller, R. (1967), *Das Ohr als Nachrichtenempfänger*, S. Hirzel Verlag, Stuttgart.