Quality Model for Monitoring QoE in VoIP Services

Filipe Neves[1,2,3], Simao Cardeal[1,4]
[1]Universidade de Tras-os-Montes e Alto Douro
[2]Polytechnic Institute of Leiria/ESTG
[3]Instituto de Telecomunicacoes
[4]Portugal Telecom Inovacao
simaovertigo@gmail.com, fneves@ipleiria.pt

Salviano Soares[1,5], Pedro Assuncao[2,3],
Filipe Tavares[4]
[5]Instituto de Eng. Electronica e Telematica de Aveiro
Portugal
salblues@utad.pt, paassunc@ieee.org,
filipe-t-tavares@ptinovacao.pt

***Abstract*—This paper presents a no-reference model for monitoring the voice quality experienced by users in VoIP services. The proposed model is based on the E-Model, which is adapted from Recommendation ITU-T G.107 for this purpose. A calibration function was determined for relevant codecs under different packet loss conditions. Extensive field testing was carried out to validate the proposed model at PT Inovação Labs (Portugal). The results show that the mean opinion score (MOS) obtained from the proposed model match class C2 of conformance tests defined in ITU-T Recommendation P.564. The QoE model described in this paper was implemented in a VoIP QoE probe and is currently fully operational at Portugal Telecom.**

I. Introduction

The Internet and its packet based architecture is becoming an increasingly ubiquitous communications resource, providing the necessary underlying support for many services and applications. The classic voice call service over fixed circuit switched networks suffered a steep evolution with mobile networks and more recently another significant move is being witnessed towards packet based communications using the omnipresent Internet Protocol (IP). It is known that, due to real time requirements, Voice over IP (VoIP) needs tighter delivery guarantees from the networking infrastructure than data transmission [1]. While such requirements put strong bounds on maximum end-to-end delay, there is some tolerance to transmission errors and packet losses in VoIP services, providing that users experience a minimum quality. For the purpose of definining a commom set of rules to evaluate voice quality, the Telecommunication Standardization Sector of International Telecommunications Union, ITU-T, has released a set of recommendations that standardize the related procedures. The ITU-T Rec. P.800 introduces the fundamental concepts and terminology to understand the different aspects involved in voice quality assessment, such as the Mean Opinion Scores (MOS) and the associated evaluation methods and requirements to achieve valid results [2]. These are subjective methods, which constitute the *de facto* methods to obtain real MOS, but they have the drawback of requiring several people and restrictive procedures leading to a time-consuming process and an expensive setup [3], [4].

Alternatively, objective methods do not need to involve several people neither to meet restrictive procedures like special rooms satisfying tight reverberation, echo or dimension requirements. This kind of methods are implemented as computational models of the human auditory perceptual system and they provide estimated MOS, approximately as it would be obtained from users. In this context, the "Perceptual Evaluation of Speech Quality" (PESQ) algorithm, described in ITU-T Rec. P.862, is an important benchmark which defines a procedure to estimate the subjective quality of 3.1 kHz (narrow-band) handset telephony and narrow-band speech codecs [5]. As input, two signals are used: the signal under evaluation and the corresponding reference (i. e. the original undistorted signal). The output is an approximate MOS appropriately mapped to MOS_{LQO} [6], [7]. This method takes into account impairments caused by distortion due to high compression codecs, packet loss, channel errors, environment noise and jitter among others. However, impairments due to loudness loss, delay or echoes are not taken into account. Since PESQ needs a reference signal, it is also called an intrusive method.

The E-Model, described in ITU-T Rec. G.107, can also be used to estimate the conversational quality of 3.1 kHz handset telephony [8]. It takes into account the combined effects of variations in several transmission parameters, such as packet loss, loudness rate, equipment parameters or transmission and echo delays, as impairment factors.

In the past several authors have enhanced quality models based on ITU-T Recommendations. In [9] an objective method to assess the speech quality in conversational context, based on both talking and listening qualities, including the impact of delay, is proposed. Two datasets are used: one for training and the other one for validation. Multiple linear regression was used to estimate the model scores with high correlation between subjective and estimated scores. The performance of the proposed model is reported as being high in comparison with the ITU-T G.107 E-Model. In [10] the effectiveness of PESQ for speech with background noise in actual cellular phone systems is evaluated and SNR-dependent background noise compensation methods for reducing estimation errors are proposed. In [11], the E-Model was extended in order to reflect time varying impairments, such as burst, packet loss and recency. This method takes into account that perceptual quality depends on the duration of a given impairment and on the time elapsed between its occurrence and when the actual assessment is done. In [4] the effects of wireless VoIP communications degradations on the performance of PESQ,

P.563 and E-Model are investigated. This comparative study concludes that extended PESQ attains superior overall performance, which is significantly affected by acoustic background noise as well as speech codec and packet loss concealment strategy.

II. NO-REFERENCE VOIP QUALITY MODEL

In this paper, a VoIP quality model based on the E-Model is proposed and validated. Like other parametric methods, this model is based on a set of parameters that characterize the subsystems involved in the transmission path. The output is a rating factor, *R*, which can be transformed in order to give estimates of user opinion scores that represent the global quality of a bidirectional communication. The fundamental principle of the E-Model is based on the concept that "psychological factors on the psychological scale are additive". Thus, the *R* factor actually comprises the relevant transmission parameters and is given by Equation (1)

$$R = Ro - I_s - I_d - I_{e\text{-}eff} + A \tag{1}$$

where *Ro* is the basic signal-to-noise ratio including room and circuit noise, I_s are all impairments which occur more or less simultaneously with the noise signal, I_d is the impairment caused by delays and echoes, and $I_{e\text{-}eff}$ is the effect of distortion caused by low bitrate codecs and packet losses. The advantage factor *A* allows for compensation of impairment factors [8].

Since the E-Model was primarily developed as a transmission planning tool, its adaptation to voice quality assessment must be validated against a reference method, as defined in the ITU-T Rec. P.564 requirements [12]. In this work we have used PESQ as the reference method for validation. Although the original E-Model meets both correlation and false positive/false negative error requirements defined in ITU-T Rec. P.564, in regard to error bounds it does not match any class of accuracy defined in that Recommendation. The model proposed in this paper improves the original E-Model in order to accomplish the ITU-T Rec. P.564 conformance accuracy requirements defined for class C2.

Since the intended application of the proposed model is VoIP, the primary concern is to take into account the impairments caused by packet loss and by distortion of low bitrate coding. Thus, a modified version of the E-Model was derived from Equation (1) by considering $I_{e\text{-}eff}$ the most relevant term. Since the reference model (PESQ) does not take the delay into account, this factor cannot be assessed. Hence I_d was set zero. *Ro* and *Is* comprise the default E-Model values. Therefore the proposed model can be simplified and described by the following expression

$$R = 93.355 - I_{e\text{-}eff} \tag{2}$$

As stated before, $I_{e\text{-}eff}$ represents the impairments caused by both low bitrate codecs and packet loss factors. Concerning the first factor, an Equipment Impairment Factor, I_e, is defined [8]. Concerning the second factor (packet loss), this includes three other factors: the Packet-loss Robustness Factor, *Bpl*, the Packet-loss Probability, *Ppl*, and the Burst Ratio, *BurstR*.

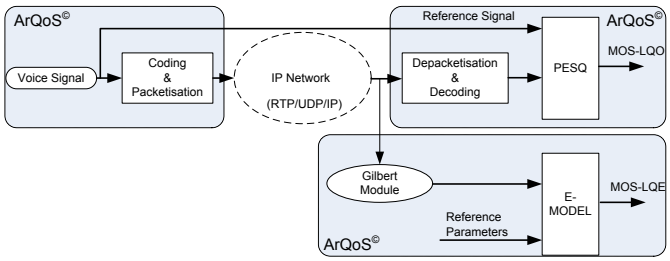

Fig. 1. Experimental setup for calibration and validation

Therefore, according to the relationship between these parameters, defined in ITU-T Rec. G.107, the model expressed by Equation (2) is more explicitly defined as follows,

$$R = 93.355 - \left(I_{\mathrm{e}} + (95 - I_{\mathrm{e}}) \frac{Ppl}{\frac{Ppl}{BurstR} + Bpl} \right) \tag{3}$$

In this work, both values of *Ppl* and *BurstR* were obtained from our experiments. In regard to I_e and *Bpl*, the recommended values referred to in ITU-T Rec. G.113/Appendix I, were used.

III. MODEL CALIBRATION AND VALIDATION

The proposed model described in the previous section was calibrated using PESQ and then validated according to ITU-T Rec. P.564 requirements. Note that the E-Model and PESQ are both sensitive to distortions caused by codecs and packet loss. The model was calibrated by using impairments caused by both low bit-rate codecs and voice packet losses randomly distributed. As pointed out before in regard to Equation (2), the relevant term of Equation (1) is $I_{e\text{-}eff}$, which represents these type of impairments. In our calibration procedure, the monitoring system platform ArQoS©, from PTIn, was used [13]. This system permits to set up, maintain, monitoring and analyze telephony calls over technologies such as PSTN, GSM or IP. It provides QoS and QoE metrics such as MOS computed by the PESQ algorithm.

The calibration test scenario is illustrated in Fig. 1, where the main signal path includes coding and packetisation, transmission over an IP network with preset average random loss and decoding to obtain the degraded. Thereafter, on the one hand, both reference and degraded signals are given as inputs to the PESQ algorithm, whose output is the reference $\mathrm{MOS_{LQO}}$ used to calibrate the proposed model. On the other hand, the loss pattern of the degraded voice stream is given as input to a Gilbert module to find the corresponding transition probabilities in order to compute the *Ppl* and *BurstR* values for $I_{e\text{-}eff}$ [8].

The test samples defined in ITU-T Rec. P.501 were used in the calibration tests [14]. Two male and two female speaker sentences were used, comprising English and Spanish languages downsampled to 8 kHz (16 bits) as required by PESQ. The codecs used in these tests were G.711, G.729 8kbps and G.723.1 6.3kbps. Six different packet loss ratios were used in the experiments: 0%, 2.5%, 5%, 10%, 15% and 20%. The $\mathrm{MOS_{LQO}}$ values obtained from PESQ, as well as $\mathrm{MOS_{LQE}}$ [7] obtained from our model were collected for each packet loss

rate, codec and sentence. This yields a total of 24 different tests for each codec and 24 different MOS scores for each quality evaluation method, i.e, the proposed E-Model and PESQ. Then for each codec, regression analysis was used to calibrate the proposed voice quality model. Based on these two sets of scores (PESQ and E-Model), the coefficients of a polynomial *p(x)* of degree *n* that fits *p*(E-Model MOS) to MOS_{LQO} were derived.

Then the results obtained from the calibrated model were validated by using a new set of sentences and new experiments. The overall test scenario and test conditions were the same as in the calibration tests.

IV. EXPERIMENTAL RESULTS

A. Model calibration

Fig. 2 shows the results obtained from regression analysis, that models the relationship between MOS_{LQO} and MOS_{LQE} scores for G.711 codec. The horizontal axis contains the scores obtained from E-Model while the vertical axis represents the scores obtained from PESQ. For each point in the graph, the difference between the scores is the error between the E-Model and the reference PESQ. For instance, the second point from the left corresponds to MOS_{LQE}=1.5 and MOS_{LQO}=1.8, which means a MOS error of 0.3. At this particular point, E-Model underestimates the MOS score in comparison with PESQ, but in general, this figure shows that the E-Model overestimates MOS relatively to PESQ. Therefore, a calibrating function was derived in order to approximate the E-Model output to that of PESQ. In the graph of Fig. 2, the points over the straight line correspond to a perfect match of both models producing the same result. The figure also shows the trend line that minimizes the RMSE between both MOS scores, which is the polynomial line that best approximates the E-Model to PESQ, for G.711 codec. Such line corresponds to the coefficients of a polynomial which gives the best approximation to PESQ. The resulting polynomial is given by

$$\begin{aligned} MOS_{LQO} &= -0.0058MOS_{LQE}^4 + 0.1252MOS_{LQE}^3 \\ &\quad -0.6467MOS_{LQE}^2 + 1.9197MOS_{LQE} \\ &\quad -0.291 \end{aligned} \tag{4}$$

This polynomial is used as the calibrating function of the proposed model in order to obtain the corresponding MOS_{LQO} scores.

Fig. 3 shows the MOS scores obtained for G.729 codec under the same test conditions as in the previous case. The figure shows that in this case, the E-Model overestimates the MOS, when compared with MOS_{LQO} obtained from PESQ. Fig. 3 also shows the trend line that best approximates the E-Model scores to MOS_{LQO} from PESQ algorithm, for G.729 codec. For this codec, the polynomial function to approximate the E-Model results to those of PESQ was found to be the following one,

$$\begin{aligned} MOS_{LQO} &= -0.0554MOS_{LQE}^5 - 0.7496MOS_{LQE}^4 \\ &\quad +3.9507MOS_{LQE}^3 - 9.874MOS_{LQE}^2 \\ &\quad +11.939MOS_{LQE} - 3.8293 \end{aligned} \tag{5}$$

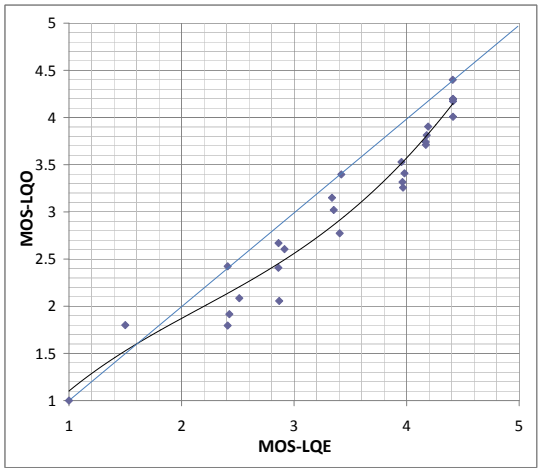

Fig. 2. Regression for calibration of the proposed model (G.711)

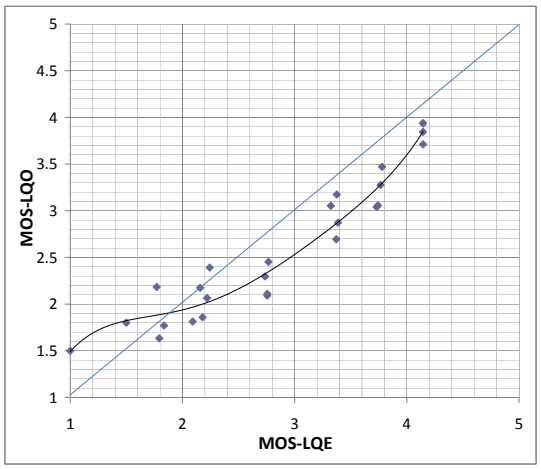

Fig. 3. Regression for calibration of the proposed model (G.729)

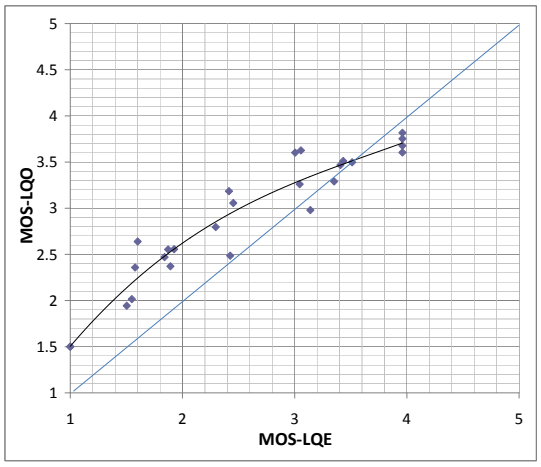

Fig. 4. Regression for calibration of the proposed model (G.723.1)

Finally, Fig. 4 shows the results for G.723.1 codec. In this case, the E-Model underestimates MOS, in comparison with MOS_{LQO} from PESQ. The figure also shows the polynomial trend line that best approximates the E-Model scores to

those from PESQ, for G.723.1 codec. From these results, the polynomial function that best approximates PESQ is given by the following expression

$$\begin{aligned} MOS_{LQO} &= 0.0018MOS_{LQE}^4 + 0.0248MOS_{LQE}^3 \\ &\quad -0.4262MOS_{LQE}^2 + 2.1953MOS_{LQE} \\ &\quad -0.2914 \end{aligned} \tag{6}$$

B. Model validation

In the validation stage, a different set of sentences were used in the ArQoS© test system to obtain PESQ MOS_{LQO} and the MOS_{LQE} scores of our calibrated model by using Equations 4, 5, and 6. Then the correlation factor, error and false positive/negative analysis between MOS_{LQO} scores and those of the proposed model were determined as defined in Recommendation ITU-T Rec. P.564. Table I, Table II and Table III show the results obtained from the tests and the conformance accuracy requirements defined in ITU-T Rec. P.564. The tables show the correlation factor, percentage of errors and false negative/false positive measures, respectively. The error boundaries in Table II are defined in Rec. ITU-T P.564 [12].

TABLE I
RESULTS FOR THE CORRELATION FACTOR

Measures	Results			Requirements (P.564)	
	G.711	G.729	G.723.1	Class C1	Class C2
Correlation	0.956	0.964	0.887	> 0.900	> 0.850

TABLE II
RESULTS FOR THE PERCENTAGE OF ERRORS

Errors within standard bounds	Results			Requirements (P.564)	
	G.711	G.729	G.723.1	Class C1	Class C2
Quality band *B*=1 ($MOS_{LQO} \geq 2.8$)					
Boundary 1 (%)	81	90	67	≥ 95.0	≥ 75.0
Boundary 2 (%)	100	100	100	≥ 97.9	
Boundary 3 (%)	100	100	100		≥ 95.0
Boundary 4 (%)	100	100	100	≥ 99.0	
Boundary 5 (%)	100	100	100		≥ 97.9
Boundary 6 (%)	100	100	100		≥ 99.9
Quality band *B*=2 ($MOS_{LQO} < 2.8$)					
Boundary 7 (%)	75	86	78	≥ 90.0	
Boundary 8 (%)	88	100	89		≥ 90.0
Boundary 9 (%)	100	100	100	≥ 95.0	
Boundary 10 (%)	100	100	100		≥ 95.0
Boundary 11 (%)	100	100	100	≥ 99.0	
Boundary 12 (%)	100	100	100		≥ 99.0

Most of the results in Table I and Table III, match both the correlation and false negative/false positive requirements for the class C1. However, according to the results shown in Table II, the percentage of errors falls within boundaries 7 and 8, which makes the proposed model to be included into class C2.

TABLE III
RESULTS FOR THE CORRELATION FACTOR

Measures	Results			Requirements (P.564)	
	G.711	G.729	G.723.1	Class C1	Class C2
False negatives (%)	0	0	0	< 5	< 5
False positives (%)	0	0	0	< 3	< 3

Based on these results, the voice quality evaluation model based on the modified E-Model along with the respective calibration functions was considered a useful voice quality evaluation tool and is currently in production at Portugal Telecom, SA. Compliance with ITU-T requirements validates the proposed voice quality evaluation model, which was already integrated in the passive probes of ArQoS© system currently in use at Portugal Telecom SA.

V. CONCLUSION

A no-reference model for voice quality monitoring was derived, based the ITU-T E-Model and taking into account the most relevant factors in the context of VoIP communications. The proposed model was validated according to ITU-T Rec. P.564 requirements and the results fit in standard accuracy class C2. Overall the proposed model exhibits good accuracy and low complexity for practical implementation.

ACKNOWLEDGMENT

This work was partially supported by Portugal Telecom Inovação - Project eVoIP.

REFERENCES

[1] T. Zourzouvillys and E. Rescorla, "An Introduction to Standards-Based VoIP: SIP, RTP, and Friends," March/April 2010.

[2] ITU-T, "P.800: Methods for subjective determination of transmission quality," 1996.

[3] A. Rix and M. Hollier, "The perceptual analysis measurement system for robust end-to-end speech quality assessment," vol. 3, pp. 1515–1518 vol.3, 2000.

[4] T. H. Falk and W.-Y. Chan, "Performance Study of Objective Speech Quality Measurement for Modern Wireless-VoIP Communications," 2009.

[5] ITU-T, "P.862: Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb 2001.

[6] ITU-T, "P.862.1: Mapping function for transforming P.862 raw result scores to MOS-LQO," Nov 2003.

[7] ITU-T, "P.800.1: Mean Opinion Score (MOS) terminology," Jul 2006.

[8] ITU-T, "G.107: The E-Model, a computational model for use in transmission planning," Mar 2005.

[9] M. Guguin, R. L. Bouquin-Jeanns, V. Gautier-Turbin, G. Faucon, and V. Barriac, "On the evaluation of the conversational speech quality in telecommunications," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, 2008.

[10] K. Fujita, T. Kato, H. Yamada, and H. Kawai, "SNR-dependent Background Noise Compensation of PESQ Values for Cellular Phone Speech," *Interspeech 2005*, pp. 3165–3168, Sep. 2005.

[11] A. Clark, "Modeling the effects of burst packet loss and recency on subjective voice quality," 2001.

[12] ITU-T, "Rec. P.564 - Conformance testing for voice over IP transmission quality assessment models," Nov. 2007.

[13] Portugal Telecom Inovacao SA, *ArQoS, Network and Services Perfomance Monitoring System*. Portugal Telecom Inovacao SA, Rua Eng. Jos Pinto Basto, 3810-106 Aveiro, Portugal.

[14] ITU-T, "ITU-T P.501: Test signals for use in telephonometry," Jun 2007.