Revisiting VoIP QoE assessment methods: are they suitable for VoLTE?

Ramon Sanchez-Iborra, Maria-Dolores Cano, Joan Garcia-Haro

Dept. of Information Technologies and Communications, Polytechnic University of Cartagena

Antiguo Cuartel de Antigones. Plaza del Hospital, Nº 1, 30202 Cartagena (Murcia)

Tel: +34 968338872 E-mail: {ramon.sanchez, mdolores.cano, joang.haro}@upct.es

Received: February 29, 2016 Accepted: June 19, 2016 Published: June 30, 2016

DOI: 10.5296/npa.v8i2.9123 URL: http://dx.doi.org/10.5296/npa.v8i2.9123

Abstract

Voice over LTE (VoLTE) emerges as the basis for voice services in future cellular networks. At the same time, Quality of user Experience (QoE) assessment methods have notably evolved in the last years, especially for Voice over IP (VoIP) services. The question that rises is, are these models ready to be used for VoLTE? In this paper, we contribute to answer this question by: (i) providing a high level exploration of current objective non-intrusive models for Voice over IP (VoIP) QoE evaluation, (ii) highlighting the review of works aiming to improve the standardized E-model (ITU-T Rec. G.107 (06/2015)) and the examination of non-standardized solutions proposed by the research community, and (iii) bringing forth the challenges identified for the completion of a successful QoE evaluation approach according to modern VoIP services such as VoLTE.

Keywords: VoLTE, VoIP, QoE, QoS, standards, estimation models, quality

1. Introduction

Despite the outstanding advances on communication technologies, mobile voice quality is still far from the quality provided by landlines. Reasons can be found in cellphones design, (several) voice-data conversions, or lack of radio coverage, among others factors [1], which become important obstacles to achieve the voice quality demanded for a service with more than 5 billion mobile users expected by 2019 [2]. Long Term Evolution (LTE) (and LTE-Advanced, LTE-A) is the latest adopted standard for cellular communications of high speed data for mobile phones and data terminals that defines an all-IP network architecture (current ongoing work is on LTE Release 14 [3]). Compared to its predecessors, LTE aims to provide higher capacity, to reduce network-level energy consumption, to maximize cost efficiency by backing different applications and services, and to offer final users a richer, faster, and more reliable experience. As part of these goals, the voice service in LTE is delivered as data flows in what is called Voice over LTE (VoLTE); i.e., under LTE, voice is just one of the many potential media streams that can be communicated by the use of Voice over IP (VoIP). Although VoLTE has not been fully deployed worldwide yet, it is expected to ensure high levels of quality. The IP Multimedia System (IMS) and the Multi Media Telephony (MMTel) installed on the IMS core are responsible for providing VoLTE with Quality of Service(QoS) by using prioritization, scheduling techniques, and other quality-provision mechanisms [4].

On the other hand, there is consensus on using the Quality of user Experience (QoE) approach in order to estimate the quality achieved by a multimedia service. QoE takes into account classical quality performance metrics such as delay or packet losses, but it goes a step ahead and evaluates the level of quality the customer perceives when consuming a service; thus, shifting quality monitoring from a network-centric to a customer-centric perspective. An extended methodology used to assess QoE is the Absolute Category Rating (ACR) that outputs a Mean Opinion Score (MOS), which is a subjective rating for the service in a scale from 1 (poor quality) to 5 (excellent). QoE assessment methods have notably evolved in the last years, especially for VoIP services. The question that rises is, are these models ready to be used for VoLTE? Can standard (or non-standard) approaches face the challenges of measuring QoE in VoLTE? Would it be necessary to include new key performance metrics in QoE assessment methods? In this paper, we contribute to answer these questions by: (i) providing a high level exploration of current objective non-intrusive models for VoIP QoE evaluation, (ii) highlighting the review of works aiming to improve the standardized E-model (ITU-T Rec. G.107 (06/2015)) and the examination of non-standardized solutions proposed by the research community, and (iii) bringing forth the challenges identified for the completion of a successful QoE evaluation approach according to modern VoIP services.

Other works in the related literature comprise reviews of QoE measurement models, e.g., [5], [6]. Nevertheless, we believe that the capability provided by the single-ended QoE models, which take measurements at real-time (i.e., objective non-intrusive models), is key for the future development of high-quality VoLTE services able to react to QoE variations during the voice call. In this sense and compared with those previous works, we provide a more complete and extensive review of non-standard proposals, we update standard incorporations, and more importantly, we identify the challenges of measuring QoE in VoLTE.

The rest of the work is organized as follows. Section 2 presents a brief overview of VoLTE. Section 3 revises standard models for QoE evaluation proposed by standardization institutions and non-standard QoE assessment methods from the related literature. After reviewing the models, we consider in Section 4 what modifications should be incorporated to convey with VoLTE features. The paper ends summarizing the most important facts in Section 5.

2. VoLTE features

QoS in VoLTE depends heavily on IMS operation. The IMS is in charge of prioritizing traffic according to QoS requirements. To do so, a QoS Class Indicator (QCI) value is assigned to each data connection (see Table 1). VoLTE calls, i.e., conversational voice, have a QCI=1 setting a maximum end-to-end delay of 100 ms, guaranteeing a constant bit rate and a packet delivery rate of 99.99%, and having the second priority among other data flows. Please observe that whereas VoLTE traffic is completely managed by network operators, other VoIP solutions such as Over-The-Top (OTT) services (e.g., Skype or Google Hangout) would run with the default bearer for Internet access (lowest priority). Other performance indicators consider that a VoLTE call is dropped if less than 98% of the VoIP packets were delivered successfully to the user within a one way radio access delay of 50 ms, and a LTE cell is expected to maintain at least 60 VoIP sessions. VoLTE coexistence with previous technologies is assured via the Circuit-Switched fallback (CSFB) scheme.

Regarding the protocol architecture, VoLTE encapsulates voice payload in RTP (Real-time Transport Protocol) / UDP (User Datagram Protocol) / IP. Adaptive Multi-Rate (AMR), AMR Wideband (AMR-WB), and Enhanced Voice Services (EVS) are the selected codecs for VoLTE, the latter able to provide a full high definition voice service. To reduce overhead, header compression is mandatory by using the RObust Header Compression Protocol (ROHC). As indicated in 3GPP specifications, header compression is carried out in the Packet Data Convergence Protocol (PDCP) layer.

3. QoE evaluation models

In this section, we address both standard and non-standard objective non-intrusive models. Objective models, i.e., not based on customer surveys, can be broadly classified into intrusive (*aka* comparison-based, full-reference, or double-ended models), which use the original speech-signal to compare it with the degraded one when it arrives to its destination, and non-intrusive models. Given that intrusive methods are not useful at real-time, we concentrate on non-intrusive models (*aka* single-ended or output-based methods). Service quality is hence assessed according to objective parameters at any point of the VoIP communication path, without needing a reference signal. These methods allow detecting bottlenecks and generating real-time data about QoE variations.

In turn, non-intrusive models can be classified as signal-based models, parametric models, and packet-layer models. Signal-based models process directly the human speech, analyzing the distortion introduced in the voice signal (e.g., ITU-T Rec P.563 (05/2004)). Parametric models (e.g., ITU-T Rec. G.107 (06/2015)) compute QoE from different impairments introduced by the network and the encoding schemes. Parametric models need a full knowledge of the system, from end-terminals to network equipment and data-links employed in the VoIP transmission to tabulate the different sources of impairments affecting communication. Finally, packet-layer models only use the information that can be extracted from the different headers of the multimedia packet (e.g., ITU-T Rec. P.564 (11/2007)).

3.1 Standard Models

3.1.1 E-model

The E-model (ITU-T Rec. G.107 (06/2015)), originally designed as a telecommunication transmission planning tool, has become one of the most popular methods to evaluate the quality of a voice transmission system. This parametric model takes into account several tabulated transmission impairments, such as delay, echo, codec distortion, etc., measured by In-Service, Non-Intrusive Measurement Devices (INMDs) (ITU-T Rec. P.561 (07/2002)). Impairments assessed by INMDs form part of an additive rating scale, called R, which estimates the conversational quality (MOS-CQE) of a voice call. R can be directly mapped to a MOS scale, so it is also employed to predict customer's conversational QoE (MOS-CQS) (please see Table 2). As described in ITU-T Rec. G.107 (06/2015), R is obtained through expression (1),

$$\mathrm{R} = \mathrm{R}_0 - \mathrm{I}_\mathrm{s} - \mathrm{I}_\mathrm{d} - \mathrm{I}_{\mathrm{e-eff}} + \mathrm{A} \tag{1}$$

where R_o is the basic signal-to-noise ratio, including background noise and transmission noise; I_s represents impairments that occur simultaneously with the voice signal, such as a non-optimal side-tone level; I_d includes impairments caused by delay; $\mathrm{I}_{\mathrm{e\text{-}eff}}$ denotes impairments due to low bit-rate audio codecs and randomly distributed packet losses, and A is the advantage factor, which allows for compensation of impairment factors when the user benefits from other types of access advantages (e.g., a mobile environment). As shown in Table 2, R ranges from 0 (lowest possible quality) to 100 (optimum quality).

However, calculations to obtain the aforementioned impairments involve many input parameters and complex mathematical formulas. For those reasons, the ITU-T proposed a reduced E-model (ITU-T Rec. G.109 (09/1999)), taking into consideration just the impairments related to the transmission over the network, and setting the rest of parameters to their default values (please see Table 2 in ITU-T Rec. G.107 (06/2015)). Using this simplified model, the expression for R is reduced as shown in (2),

$$\mathrm{R} = 93.4 - \mathrm{I}_{\mathrm{dd}} - \mathrm{I}_{\mathrm{e-eff}} \tag{2}$$

where I_{dd} represents the impairments due to transmission delay in echo-free connections, being calculated as a function of the absolute transmission delay. $I_{e\text{-}eff}$ includes the same effects as in the original formula taking into account that, for each codec, this impairment is calculated as a function of packet loss probability.

Table 2. R to MOS mapping

R	User Satisfied	MOS
90 - 100	Very Satisfied	4.34 – 4.50
80 - 90	Satisfied	4.03 – 4.34
70 - 80	Some Users Dissatisfied	3.60 – 4.03
60 - 70	Many Users Dissatisfied	3.10 – 3.60
50 - 60	Nearly All Users Dissatisfied	2.58 – 3.10
0 - 50	Not Recommended	1 – 2.58

The description given in ITU-T Rec. G.107 (06/2015) presents the E-model as a planning tool for narrow band transmissions (300–3400 Hz), but an extended version for wideband (50-7000 Hz) has been recently released (ITU-T Rec. G.107.1 (06/2015)). Although some wideband E-model predictions are currently under study, this extended model captures the effects of several factors previously ignored, such as low-bitrate wideband coding or VoIP packet loss degradations.

Based on this transformation, the updated maximum value for R (wideband) is 129 [7]; thus, the quality improvement introduced by wideband transmission is quantified on the R-scale. Expression (1) applied to wideband transmission undergoes some changes as following; the impairment parameter I_s and the advantage factor A are not adequately analyzed yet, so both take a value equal to 0. I_d takes into account talker/listener echo in addition to the absolute transmission delay and, finally, $I_{e\text{-}eff}$ includes the effect of random or bursty packet loss and speech coding. An extended expression for R in wideband is shown in (3),

$$R_{WB} = R_{o,WB} - \{I_{dte,WB} - I_{dle,WB} - I_{dd}\} - \left\{(I_{e,WB} + (95 - I_{e,WB}) \cdot \frac{P_{pl}}{\frac{P_{pl}}{BurstR}+B_{pl}}\right\} \quad (3)$$

where $R_{o,WB}$ is the basic wideband signal-to-noise ratio, including noise sources such as circuit noise and room noise; $I_{dte,WB}$ and $I_{dle,WB}$ give an estimate for the impairments due to talker/listener echo; I_{dd} includes the impairments caused by delay; and $I_{e,WB}$ and B_{pl} are codec-specific values and represent the impairment due to low bit-rate coding and the packet-loss robustness factor, respectively. Values for these parameters can be found in Appendix IV of the ITU-T Rec. G.113 (11/2007). P_{pl} is the packet-loss probability and *BurstR* is the burst ratio, which represents the bursty pattern of the transmission.

Despite of its standardized status, Grah and Radcliffe [8] questioned the applicability of the E-model to VoIP transmissions, as it was designed for systems operating in different scenarios to those currently used for VoIP. A new expression for R, called VoIP E-model (VoIP-EM), was proposed. VoIP-EM includes relevant impairments for current real-time systems such as coding scheme, packet loss, and delay, and ignores those impairments whose impact could be dismissed in VoIP quality such as echo, quantization, and loudness. Thus, the expression proposed for VoIP-EM is shown in (4),

$$VoIP - EM = N_o - SE \quad (4)$$

where N_o is the default initial quality value according to the network setup and codec type and *SE* represents the sum of all errors, which are calculated as a linear (or non-linear) function of the aforementioned delay, packet loss, and codification impairments. However, this work did not provide any analytical result or performance evaluation to assess the accuracy of the proposed model. Specific configuration parameter values were not included either, so proposal efficiency was not verified.

A more complete study was presented by Meddahi and Afifi [9], where the E-model was also redefined to work in packet-switched networks. Authors assumed that the main factor affecting the speech quality in VoIP environments are: Analog/Digital (A/D) and Digital/Analog (D/A) conversions, coding algorithm, bandwidth, jitter, delay, and packet loss. By analyzing and adapting these elements to the classical parameters in the R additive scale, a derived model for datagram transport, called Packet-E-model (P-E-model), was obtained. The proposed expression to estimate R is shown in (5).

$$R_p = R_{op} - I_{dp} - I_{ep} + A \quad (5)$$

Observe the similarity of this formula with the reduced E-model expression (2). In this case, R_{op} includes a new factor to characterize the effect of packet switched noise, and I_{dp} and I_{ep} are obtained as indicated in the standard ITU-T Rec. G.107 (06/2015). R_p is calculated for every VoIP packet and it may be averaged over a longer period, e.g., during a phone call. Authors tested the P-E-model through simulation and in a real scenario. Results showed the evolution of the estimated MOS under several packet loss and delay conditions, evidencing good model response against impairment variations.

Falk and Chan [10] investigated the impact of wireless-VoIP degradation on the performance of the E-model, among other standards such as ITU-T Rec. P.563 (05/2004) and PESQ (ITU-T Rec. P.862 (02/2001)). Through factorial analysis of variance tests, authors found that the performance of the aforementioned algorithms is sensitive to several degradation sources such as noise level and codec-PLC type. Focusing on the E-model, they also suggested several significant two-way interaction effects, such as codec and noise type or codec and noise level, concluding that the E-model accuracy is limited for wireless-VoIP scenarios. Work by Picovici and Nelson [11] dealt with the main impairment introduced by the wireless networks, i.e., the variability of packet loss. Authors added a new parameter, I_p, into the R scale with the aim of accounting for the perceptual relevance of packet loss variations. I_p is calculated by computing the Euclidean-based median distance between a vector representing the perceptual features of the clean signal and another including the speech features of the received signal. An important drawback is found in this proposal, regarding the need of the clean signal, which transforms this model into a reference-based methodology.

Another work considering the impact of wireless-system issues, such as temporal disconnection, was proposed in [12]. This model, so-called PEVOM (Perceptual Evaluation of Voice over MANETs), detects the loss of nodes connectivity and estimates separately the quality experienced in connected and disconnected periods. When users are connected, the instantaneous transmission quality is calculated by using the reduced version of the E-model (ITU-T Rec. G.109 (09/1999)). On the other hand, when users loose connectivity, PEVOM predicts the instantaneous transmission quality employing a formula taking into account users' dissatisfaction due to issues during the call. The performance of PEVOM was compared against that attained by VQMON in ad-hoc scenarios, achieving a higher level of accuracy on the predicted quality. According to the authors, VQMON underestimates the final quality of the communication when disconnections happen during the call; however, no other comparative results (e.g., subjective test, or PESQ) were showed to confirm this hypothesis.

Finally, an improvement of the E-model also for wireless networks, particularly for satellite-based radio calls in air traffic control, was described in [13]. Authors proposed to adapt the E-model for radio call quality analysis by taking into account radio call features such as half-duplex communication, no-echo possibilities, absence of sidetone during reception, voice signal level at the pilot side larger than in the telephone networks, or the prevalence of the delay as impairment over other conditions that affects the end-to-end voice quality in this scenario. Their proposal was successfully validated by field-testing.

To end this section, we would like to note that one of the last works on the E-model [14] suggests the need of modifying it to better fit the effect of burstiness, packet-loss robustness, or jitter-buffer behavior. As indicated by the authors, MOS scores output by the E-model will be below the real quality level since the E-model was not designed as a monitor tool but as a planning one.

3.1.2 P.563

The ITU-T Rec. P.563 (05/2004) describes a single-ended, signal-based method for objective speech quality assessment in narrowband telephony applications. P.563 is able to predict the listening quality (MOS-LQO) in a perception-based scale considering the full range of distortions occurring in public switched telephone networks; therefore, this model allows real-time measurements to estimate the MOS of a voice call at any point of the path between users. The P.563 model is formed by the series of blocks shown in Figure 1. As explained in [15], the first step consists of pre-processing the received signal aiming at preparing it for the next steps. Once the signal is pre-processed, it passes by different blocks that generate different quality estimations. The first block uses a vocal track model in order to extract parameters from the distorted speech, which allows measuring the unnaturalness of the speech. The second block uses a full reference perceptual model (PESQ) for estimating the quality of the degraded speech. The last block focuses on the detection of particular degradations such as noise estimation and robotization detection.

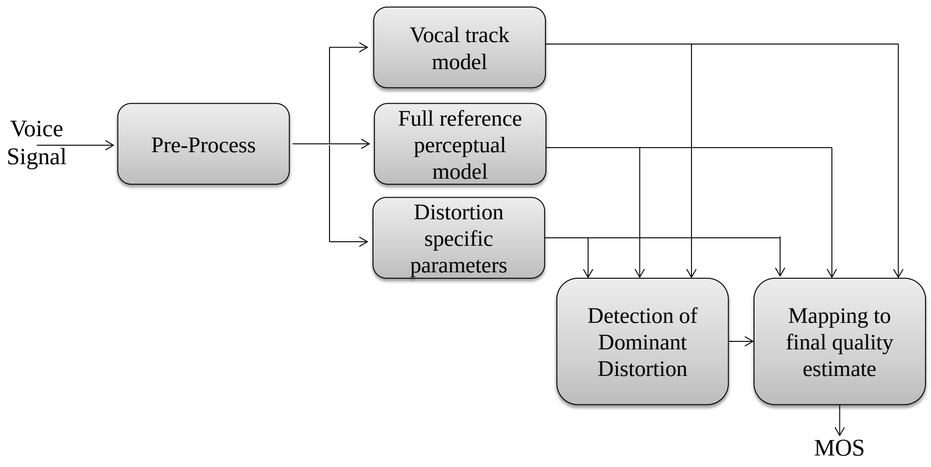
Voice Signal
Pre-Process
Vocal track model
Full reference perceptual model
Distortion specific parameters
Detection of Dominant Distortion
Mapping to final quality estimate
MOS

Figure 1. P.563 blocks. Extracted from [15].

Once this complex analysis is done, a dominant distortion class is determined and a class-specific subset of the extracted parameters is used to estimate signal quality. Abareghi *et al.* [16] introduced an enhancement to the P.563 model to adapt its features to VoIP conditions. Authors declared mute-length, sharp-decline, and speech-interruptions as the most sensitive parameters to network variations. Consequently, a new distortion class and a priority for this new class are defined using those parameters. This new class is added to the P.563 quality estimator. However, this work did not provide any comparison between the accuracy obtained by the standard and the proposed model, so additional work would be needed to demonstrate the validity of this enhancement.

3.1.3 ANIQUE+

The Auditory Non-Intrusive Quality Estimation Plus (ANIQUE+) model is an ANSI standard for non-intrusive, signal-based, estimation of narrowband speech quality. ANIQUE+ estimates the listening quality (MOS-LQO) of a voice call based on the functional roles of human auditory systems and the characteristics of human articulation systems [17]. The ANIQUE+ algorithm measures the overall distortion affecting the voice signal and maps this distortion to a MOS value (see ANIQUE+ block structure in Figure 2). In the first block, the input is pre-processed in order to reflect the frequency features of the handset used in ITU-T listening tests. This pre-processed speech signal is passed to the next step where other types of distortion are analyzed. The ANIQUE+ algorithm uses three distortion measurement modules. As described in [17], the articulation analysis block separates the original speech signal into time frames and the perceptual distortion to estimate the overall distortion of the input speech signal. The mute detection module detects unnatural mutes in speech signals, obtaining a speech activity profile. These measurements are passed to the mute impact module that estimates the impact on the speech quality degradation of the unnatural mutes in the speech signals. Finally, the non-speech modules detect the effects of annoying non-speech activity and quantify its impact on speech quality. Kim and Tarraf [17] presented results in which ANIQUE+ overcame P.563's accuracy, reaching very close performance to intrusive PESQ model. Particularly, making use of 10 non-training databases, ANIQUE+ reached an accuracy of the 93.9% while P.563 only reached the 84.8% of performance. However, the best results were still obtained by PESQ, attaining a 95.3% of accuracy.

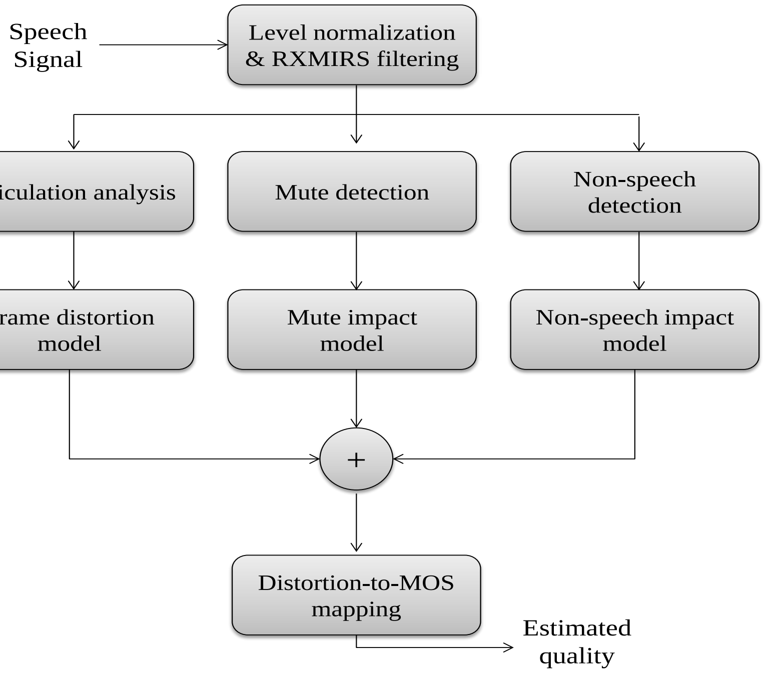

Figure 2. ANIQUE+ blocks diagram. Extracted from [17].

3.1.4 P.564

All the measurement methods mentioned so far, but the E-model, evaluate the quality of the VoIP service by analyzing speech parameters, e.g., SNR, echo, silent periods, coding distortion, etc. Thereby, aforementioned algorithms need to de-packetize vocal signal contained in IP flows to evaluate the speech payload. This task requires high computational processing compared with parametric models. For that reason, ITU-T SG12 launched the P.VTQ (Voice Transmission Quality) competition during the period 2002-2004. Two different methodologies were presented: Telchemy's VQMon [18] (not included here due to space limitations) and Psytechnics' PsyVoIP [19].

PsyVoIP block structure is shown in Figure 3. When a call stream is detected, PsyVoIP assigns it a specific ID, and pre-processes the signal, discarding the useless fields of the headers. Then, the packets are realigned and a Voice Activity Detection (VAD) algorithm is used to mark them with a “voice” or “silence” flag. This process yields to more accurate results, since distortion in silence packets has lower impact on quality than impairments on voice packets. Finally, the quality estimation is calculated as a function of network parameters extracted from the processed stream. It is also remarkable that this model takes into account the particular features of the different manufactured edge-devices, such as VoIP phones or gateways, by using calibrated formulas and weighting coefficients to each specific device. No winner for the VTQ quest was selected, but this competition served to develop the recommendation P.564, which specifies the minimum criteria for objective speech quality assessment models that predict the impact of observed IP network impairments on the one-way listening quality experienced by the end-user in IP/UDP/RTP-based telephony applications (ITU-T Rec. P.564 (11/2007)). Originally specific to narrowband (3.1 KHz), the P.564 Recommendation also includes an extension for wideband (7 KHz) telephony.

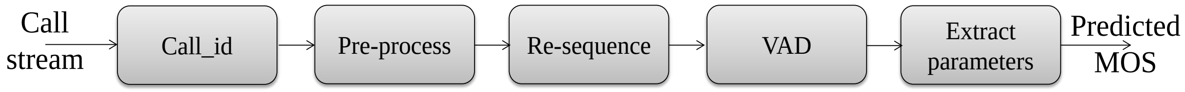

Figure 3. PsyVoIP blocks diagram. Extracted from [19].

To the authors’ knowledge, there is still no model elected as standard for this recommendation. Compliant candidate models should produce a MOS estimation as defined in the ITU-T Rec. P.800 (08/1996). This evaluation should be done ignoring the voice payload; thus, conversational impairments such as speech level, background noise, side-tone level, or echo have to be disregarded. On the other hand, the voice codec used should be always taken into account. The accuracy criteria set in this recommendation for proposed models is based on a comparison of the proposal’s performance with PESQ.

Given that PESQ has been recently upgraded by POLQA (ITU-T Rec. P.863 (09/2014), a newer full-reference method for QoE evaluation of high definition (HD) voice in a broad range of networks, an update is needed in P.564 to accommodate these improvements. Finally, a model compliant with recommendation P.564 should be able to be deployed in endpoint locations as embedded monitoring agents, at mid-network monitoring locations, or a combination of both.

3.2 Non-standard Models

In addition to the standard models, there is an increasing number of algorithms proposed by several authors aiming to improve the accuracy of standards by using different techniques. In the following, we will discuss the most extended techniques employed. For simplicity, we have classified the non-standardized proposals into four categories: based on Gaussian Mixture Models (GMM), based on neural networks, based on exponential functions, and finally, hybrid models.

3.2.1 Gaussian Mixture Models

Falk *et al.* [20] made an extensive use of GMMs. In their first works [20], GMMs were used to generate artificial reference models of speech behavior. This technique compares distortion features introduced in these reference models with those affecting the real signal stream. Therefore, a double-ended quality estimation algorithm is emulated. In addition to these works, authors introduced an enhancement to improve the quality estimation accuracy when noise suppression algorithms are incorporated. Afterwards, they proposed a modification that includes additional information related to the transmission and coding schemes employed in the communication, showing a better performance in terms of accuracy. Results in their works revealed better performance than P.563, being remarkable the decrease obtained in processing-time.

In the same vein, Wang *et al.* proposed two different quality estimators. First, a quality estimation algorithm based on GMM and on Support Vector Regression (SVR) was proposed [21]; then, an enhanced non-intrusive objective speech quality evaluation method based on Fuzzy Gaussian Mixture Model (FGMM) and Fuzzy Neural Network (FNN) was also presented [22]. In the former, authors used GMM to form an artificial reference model of the behavior of Perceptual Linear Predictive (PLP) features of clean speech. Consistency measures between the degraded speech and the reference model were utilized as indicators of speech quality. The effective least square SVR arithmetic was used to map the consistency values to the predicted MOS. In the latter work [22], an improved version of the previous method was proposed based on FGMM and FNN. FGMM was employed, instead of GMM, to form the artificial reference model. FNN regression algorithm was used to map the consistency values to the predicted MOS. Results in both works outperformed the standard P.563 for several coding schemes, such as G.711 and G.729; but additional tests under different conditions and multilingual databases would be necessary, in order to lead to more robust and comprehensive algorithms.

3.2.2 Neural Networks

It is well known that Neural Networks (NN) have been extensively applied to emulate human behavior. Following this direction, several authors have developed quality-estimation models for voice communications based on Artificial Neural Networks (ANN) [23] and Random Neural Networks (RNN) [24]–[27]. Sun and Ifeachor [23] analyzed the effect on the call quality of four different parameters, namely, codec type, gender, loss pattern, and loss burstiness. In order to model the relationships between these elements and perceived speech quality, a neural network model was developed to learn the non-linear mapping from these parameters to a MOS score. Results showed good accuracy and correlation with PESQ, demonstrating that packet loss has a severe impact on perceived quality, and that female voices tend to be worse evaluated than male voices.

In turn, Mohammed *et al.* made a deep study of their Pseudo-Subjective Quality Assessment (PSQA) technique [24], [25]. This model uses an RNN–based quality assessment mechanism in order to evaluate the influence of certain quality-affecting parameters, such as coding scheme, redundancy, packet loss rate, Mean loss Burst Size (MBS), and packetization interval, on real-time listening quality. PSQA allows evaluating the effects of these parameter interactions as a whole. An enhancement of this model was also presented [26]. In this work the effect of delay, jitter, and Forward Error Correction (FEC) mechanisms were also taken into account. Consequently, PSQA was extended to assess not only the listening quality, but the conversational quality. In their results, authors showed that the key parameters affecting the conversational quality are packet loss rate, coding bit-rate, and the FEC mechanism. The impact of other impairments, such as delay or jitter, is low and subordinated to that of the packet loss process. In contrast to previously-discussed works, authors claimed that the mean loss burst size does not play a significant effect on the QoE. This fact was explained because of the good performance of the employed FEC mechanism and the PLC algorithm implemented in the codec used in their experiments. PSQA quality estimation showed good correlation with subjective tests. Cherif *et al.* also made use of RNN to capture the non-linear relation between network parameters that cause voice distortion and the perceived quality [27]. The proposed model, so-called A_PSQA, receives as input just two parameters, namely, the packet loss rate and the Mean Loss Burst Size (MLBS) of the VoIP communication. The latter is employed, as in other reviewed models, to have an idea about the burstiness of losses. In this work, the effect of jitter is ignored as, with the absence of a de-jittering buffer, it is considered like additional packet losses. In order to train the RNN, authors developed a database of MOS scores for different speech samples transmitted under different loss conditions, characterized by the Gilbert model. These scores were attained by using the PESQ model. After training the RNN, results showed very good correlation with PESQ, beating the standard E-model and the IQX model, which will be analyzed in Section 3.2.3. As stated in their conclusions, an extension of this algorithm including additional coding schemes would be desirable, as well as the inclusion of the delay impact.

3.2.3 Exponential functions

The IQX hypothesis, developed by Hoßfeld *et al.*, assumes an exponential functional relationship between QoE and QoS. Authors claimed that the more sensitive the subjective sensibility of QoE is, the higher the experienced quality. Under this assumption, they assumed that the change of QoE depends on the current level of QoE given the same amount of change of the QoS value. Thus, the IQX hypothesis is formulated as shown in (6),

$$QoE = \propto \cdot\, e^{-\beta \cdot QoS} + \gamma \tag{6}$$

where α, β, and γ take different values depending on the codec employed, and the QoS parameter is measured in terms of packet loss rate, delay, and jitter.

In a first study [28], authors concentrated on the packet loss probability to measure the quality of service using a well known commercial VoIP application; an extension of this work was then presented [29], in which delay and jitter were also taken into account to assess the QoS. The results included in these works showed good accuracy, verifying the exponential relationship between QoE and QoS, but authors did not evaluate the combined effect of the different impairments assessed, e.g., packet loss rate and jitter, which could introduce important variations in the measured QoE. In addition, authors did not estimate the effect of bursty losses that, as demonstrated by other authors [18], [30], has a notable effect on the VoIP QoE. Aiming to simplify the parametric approach as much as possible and at the same time further explore the burstiness-relating metrics as recommended in [14], Jung and Mazano [31] presented three parametric models obtained through regression analysis, namely Model A, Model B, and Model C, whose unique input parameter is packet loss and the output a MOS value in the range 1-5. The simplest model, Model B, is expressed as shown in (7), where P_{loss} corresponds to packet loss probability.

$$Model\_B(P_{loss}) = 3.108 \cdot e^{-0.06561 \cdot P_{loss}} \tag{7}$$

In turn, Models A and C are enhanced versions of Model B and include the effect of two additional burst-related metrics, specifically burst density and the fraction of burst loss within loss. Model C was designed as a quality model to be incorporated as VoIP QoE monitoring method in mobile energy-constrained end-points. Due to space limitations, equations for Models A and C are not shown but can be found in [31]. The accuracy of the models was validated using PESQ, but no comparison with IQX or any other parametric models was carried out.

3.2.4 Hybrid algorithms

The approaches discussed above can be categorized as signal-based models, parametric models, or packet-layer models. The former, estimate the QoE by processing directly the human speech, analyzing the distortion introduced in the voice signal. These models have shown to be sensitive to bursty packet loss and PLC algorithms. On the other hand, parametric models base its QoE estimation on the assessment of different impairments introduced by network and encoding schemes, which make them sensitive to background noises or noise suppression strategies. Finally, the packet-layer models have not been fully developed, because of the difficulty of representing the complex interactions among the different impairments affecting the VoIP QoE just from parameters extracted from the packet headers. Thus, there is a current trend to join the best characteristics of previous approaches. These methods, so-called hybrid models, are gaining momentum and several proposals have been presented.

Jelassi *et al.* [32] extended the conventional parametric speech quality estimation models by considering the voicing feature of lost packets. This model builds a voicing-aware speech quality model that allows to accurately quantifying the effect of lost packets according to their voicing property. By using multiple regression analysis, the speech quality estimation is obtained as a function of the different values of quality obtained for voiced and unvoiced frames, weighted by fitting coefficients. Further information is gathered from the packet loss pattern, distinguishing drops of voice and unvoiced frames. To do that, a sender-based notification scheme is adopted, in which additional data is introduced into the VoIP packets by the transmitter. This information allows the receiver to differentiate between voiced or unvoiced lost packets, but with an extra-consumption of bandwidth.

A listening-only model based on both network impairments (packet loss), and voice distortions (temporal clipping and noise) was presented in [33]. First, the aforementioned impairments are detected. In order to evaluate (i) the occurrences of packet loss, (ii) the loss pattern and (iii) the employed coding scheme, an analysis of the IP headers is conducted. Additionally, a differentiation of silenced, voiced, or unvoiced packets is done, through voice payload analysis. The detections of temporal clipping and noise are also based on processing the voice payload. In the second step, the impact of individual impairments is quantified. Finally, the overall quality evaluation model is built by integrating individual impairments and employing the E-model. Results exhibited high correlation with PESQ (mapped onto listening-only MOS), but several important impairments were ignored, such as delay, jitter, echo, etc. Further work is needed to extend this model to a conversational quality estimator, improving also its predictions by comparing them to those obtained by subjective tests.

In another work [34], authors extended the use of their GMM proposals, by integrating packet header analysis. GMM are used, again, to generate an artificial reference model that is compared with the transmitted speech signal. On the other hand, a parametrical analysis similar to the previous one [32] is performed, evaluating the VoIP header. Consequently, this model inherits the limitations discussed above, related to the relevance of ignored impairments such as delay or jitter. A key characteristic for this methodology is its low computational complexity, which is 88% lower than the ITU-T standard P.563. Additionally, the proposed algorithm improves both, pure parametric approaches by measuring distortions that are not captured by connection parameters and pure signal based models by reaching lower per-call estimation errors.

4. Challenges in VoLTE QoE assessment

Not all QoE models take into account delay. However, due to VoLTE characteristics, e.g., the use of CSFB, delay effects could become more difficult to manage. To optimize the trade-off between voice quality and voice capacity, delay is a key parameter to manage, and consequently, a proper jitter buffer operation management is needed, too. In this sense, QoE assessment/monitoring tools should not ignore either jitter.

Although the use of a flat all-IP network eliminates the need of several voice-data conversions, it should be investigated the effect of compression on the voice quality, e.g., decompression failures due to burst errors; thus, considering its inclusion in QoE metrics. Similarly, further work is needed to evaluate the novel EVS codec and its derived effect on QoE models/monitoring tools (e.g., updated values for $I_{e\text{-}eff}$ impairment factor or extending the maximum value for *R* on E-model).

Handovers, in all their forms (e.g., from/to indoor cells, from/to non-LTE coverage, from/to Voice over WiFi, etc.), might be responsible of dropped calls, service interruptions, re-buffering, disconnection times, or higher failure rates. The low-explored *A* factor from the E-model could be a suitable candidate to accommodate handovers effect on voice quality assessment, and we believe that their impact on QoE models should be parameterized.

Likewise, the effect of techniques such as Discontinuous Transmission (DTX) and Discontinuous Reception (DRX) introduced in VoLTE to reduce energy consumption, or the use of Packet Loss Concealment methods in VoLTE should be investigate to gather their implications in QoE models.

In order to achieve the expected QoE, carriers will have to respect priority codes, which could become more complex with the coexistence of network operators and *virtual* network operators. A proper selection of monitoring points for QoE evaluation/monitoring could facilitate this compromise. Finally, the accuracy of most QoE measurement methods should be now contrasted with POLQA instead of PESQ; but POLQA has not been validated in live VoLTE networks yet.

5. Conclusion

In this work we have presented a detailed review of current standards and non-standard proposals related to objective non-intrusive QoE estimation for VoIP services, aiming at deriving the challenges that these models face to be used as QoE assessment methods in VoLTE. We focused on these methodologies, instead of the reference-based models, because obtaining QoE estimations at real-time assists network operators and service providers in performing efficient QoE management. First, we drew a general view of the standardized models by both ITU-T and ANSI standardization bodies. Next, as an essential part of this survey, we compiled and presented a great number of non-standard methodologies, aiming to improve the accuracy of the standards. Our conclusion is that there is not available yet an objective non-intrusive QoE assessment method that encompasses all VoLTE characteristics due to the new features that introduces (e.g., CSFB, compression/decompression, EVS codec, novel handovers, etc.). New approaches for QoE assessment could come from improvements of current proposals or breakthrough methods based on data mining. In either case, it will be a highly-valuable tool to achieve the expected voice quality in next generation networks.

Acknowledgement

This work was supported by the MINECO/FEDER project grant TEC2013-47016-C2-2-R (COINS).

References

[1] J. Hecht, “Why mobile voice quality still stinks—and how to fix it,” *IEEE Spectrum*, pp. 30–35, 2014.

[2] Cisco, “Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014–2019. White Paper,” *White Pap.*, 2015.

[3] “3GPP,” 2015. [Online]. Available: http://www.3gpp.org.

[4] S. Wang, L. Sun, Q. Sun, X. Li, and F. Yang, “Efficient service selection in mobile information systems,” *Mob. Inf. Syst.*, vol. 2015, pp. 1 – 10, 2015,

http://doi.org/10.1155/2015/949436.

[5] S. Jelassi, G. Rubino, H. Melvin, H. Youssef, and G. Pujolle, "Quality of Experience of VoIP service: a survey of assessment approaches and open issues," *IEEE Commun. Surv. Tutorials*, vol. 14, no. 2, pp. 491–513, 2012, http://doi.org/10.1109/SURV.2011.120811.00063.

[6] S. Moller, W.-Y. Chan, N. Cote, T. H. Falk, A. Raake, and M. Waltermann, "Speech quality estimation: models and trends," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 18–28, 2011, http://doi.org/10.1109/MSP.2011.942469.

[7] S. Möller, A. Raake, N. Kitawaki, and A. Takahashi, "Impairment factor framework for wide-band speech codecs," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 14, no. 6, pp. 1969–1976, 2006, http://doi.org/10.1109/TASL.2006.883262.

[8] M. Grah and P. Radcliffe, "Dynamic QoS and network control for commercial VoIP systems in future heterogeneous networks," in *Tenth IEEE International Symposium on Multimedia*, 2008, pp. 356–363, http://doi.org/10.1109/ISM.2008.87.

[9] A. Meddahi and H. Afifi, "'Packet-E-model': E-model for VoIP quality evaluation," *Comput. Networks*, vol. 50, no. 15, pp. 2659–2675, Oct. 2006, http://doi.org/10.1016/j.comnet.2005.10.008.

[10] T. Falk and W.-Y. Chan, "Performance study of objective speech quality measurement for modern wireless-VoIP communications," *EURASIP J. Audio, Speech, Music Process.*, vol. 2009, pp. 1–11, 2009, http://doi.org/10.1155/2009/104382.

[11] D. Picovici and J. Nelson, "Time-varying quality estimation for VoIP over Wireless Networks," in *9th IFIP International Conference onMobile Wireless Communications Networks*, 2007, pp. 91–95, http://doi.org/10.1109/ICMWCN.2007.4668187.

[12] S. Jelassi and H. Youssef, "Connectivity Aware Instrumental Approach for Measuring Vocal Transmission Quality Over a Wireless Ad-Hoc Network," in *New Technologies, Mobility and Security*, 2008, pp. 1–5, http://doi.org/10.1109/NTMS.2008.ECP.32.

[13] S. Apostolacos, A. Meliones, S. Badessi, and G. Stassinopoulos, "Adaptation of the E-model for satellite internet protocol radio calls in air traffic control," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 51, no. 1, pp. 81–96, 2015, http://doi.org/10.1109/TAES.2014.130064.

[14] M. Soloducha and A. Raake, "Speech quality of VoIP: bursty packet loss revisited," in *ITG-Fachbericht 252: Speech Communication*, 2014, pp. 1–4.

[15] L. Malfait, J. Berger, and M. Kastner, "P.563 - The ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 6, pp. 1924–1934, 2006, http://doi.org/10.1109/TASL.2006.883177.

[16] M. Abareghi, M. M. Homayounpour, M. Dehghan, and A. Davoodi, "Improved ITU-P.563 non-intrusive speech quality assessment method for covering VOIP conditions," in *10th International Conference on Advanced Communication Technology*, 2008, pp. 354–357, http://doi.org/10.1109/ICACT.2008.4493777.

[17] D.-S. Kim and A. Tarraf, “ANIQUE+: a new American national standard for non-intrusive estimation of narrowband speech quality,” *Bell Labs Tech. J.*, vol. 12, no. 1, pp. 221–236, May 2007, http://doi.org/10.1002/bltj.v12:1.

[18] A. D. Clark, “Modeling the effects of burst packet loss and recency on subjective voice quality,” in *IP Telephony Workshop*, 2001.

[19] S. Broom and M. Hollier, “Speech quality measurement tools for dynamic network management,” *MESAQIN*, 2003.

[20] T. H. Falk and W.-Y. Chan, “Nonintrusive speech quality estimation using Gaussian mixture models,” *IEEE Signal Process. Lett.*, vol. 13, no. 2, pp. 108–111, Feb. 2006, http://doi.org/10.1109/LSP.2005.861598.

[21] J. Wang, J. Luo, S. Zhao, and J. Kuang, “Non-intrusive objective speech quality measurement based on GMM and SVR for narrowband and wideband speech,” in *11th IEEE Singapore International Conference on Communication Systems*, 2008, pp. 193–198, http://doi.org/10.1109/ICCS.2008.4737170.

[22] J. Wang, Y. Zhang, Y. Song, S. Zhao, and J. Kuang, “An improved non-intrusive objective speech quality evaluation based on FGMM and FNN,” in *3rd International Congress on Image and Signal Processing*, 2010, pp. 3495–3499, http://doi.org/10.1109/CISP.2010.5646757.

[23] L. Sun and E. C. Ifeachor, “Perceived speech quality prediction for voice over IP-based networks,” in *IEEE International Conference on Communications. ICC’02*, 2002, vol. 4, pp. 2573–2577, http://doi.org/10.1109/ICC.2002.997307.

[24] S. Mohamed, G. Rubino, and M. Varela, “Performance evaluation of real-time speech through a packet network: a random neural networks-based approach,” *Perform. Eval.*, vol. 57, no. 2, pp. 141–161, Jun. 2004, http://doi.org/10.1016/j.peva.2003.10.007.

[25] S. Mohamed, G. Rubino, and M. Varela, “A method for quantitative evaluation of audio quality over packet networks and its comparison with existing techniques,” in *Measurement of Speech and Audio Quality in Networks. MESAQIN’04*, 2004.

[26] A. P. C. da Silva, M. Varela, E. de Souza e Silva, R. M. M. Leão, and G. Rubino, “Quality assessment of interactive voice applications,” *Comput. Networks*, vol. 52, no. 6, pp. 1179–1192, Apr. 2008, http://doi.org/10.1016/j.comnet.2008.01.002.

[27] W. Cherif, A. Ksentini, D. Negru, and M. Sidibe, “A_PSQA: PESQ-like non-intrusive tool for QoE prediction in VoIP services,” in *IEEE International Conference on Communications (ICC)*, 2012, pp. 2124–2128, http://doi.org/10.1109/ICC.2012.6364004.

[28] T. Hoßfeld, P. Tran-Gia, and M. Fiedler, “Quantification of quality of experience for edge-nased applications,” in *20th International Teletraffic Congress*, 2007, pp. 361 – 373, http://doi.org/10.1007/978-3-540-72990-7_34.

[29] T. Hoßfeld, D. Hock, P. Tran-Gia, K. Tutschku, and M. Fiedler, “Testing the IQX hypothesis for exponential interdependency between QoS and QoE of voice codecs iLBC and G.711,” in *18th ITC Specialist Seminar on Quality of Experience*, 2008.

[30] H. Z. H. Zhang, L. X. L. Xie, J. B. J. Byun, P. Flynn, and C. S. C. Shim, “Packet loss burstiness and enhancement to the E-model,” in *6th Int. Conference on Software Engineering Artificial Intelligence Networking and ParallelDistributed Computing and First ACIS Int. Workshop on SelfAssembling Wireless Network*, 2005, pp. 214 – 219, http://doi.org/10.1109/SNPD-SAWN.2005.57.

[31] Y. Jung and C. Manzano, “Burst packet loss and enhanced packet loss-based quality model for mobile voice-over Internet protocol applications,” *IET Commun.*, vol. 8, no. 1, pp. 41–49, Jan. 2014, http://doi.org/10.1049/iet-com.2011.0701.

[32] S. Jelassi, H. Youssef, C. Hoene, and G. Pujolle, “Voicing-aware parametric speech quality models over VoIP networks,” in *Global Information Infrastructure Symposium*, 2009, pp. 1–8, http://doi.org/10.1109/GIIS.2009.5307097.

[33] L. Ding, Z. Lin, A. Radwan, M. S. El-Hennawey, and R. A. Goubran, “Non-intrusive single-ended speech quality assessment in VoIP,” *Speech Commun.*, vol. 49, no. 6, pp. 477–489, Jun. 2007, http://doi.org/10.1016/j.specom.2007.04.003.

[34] T. H. Falk and W.-Y. Chan, “Hybrid signal-and-link-parametric speech quality measurement for VoIP communications,” *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 16, no. 8, pp. 1579–1589, Nov. 2008, http://doi.org/10.1109/TASL.2008.2004524.

Copyright Disclaimer