# QoE Assessment of VoIP in Next Generation Networks

David Rodrigues[1], Eduardo Cerqueira[1,2], and Edmundo Monteiro[1]

[1] University of Coimbra, Department of Informatics Engineering, 3030-290 Coimbra, Portugal,
drod@student.dei.uc.pt, {ecoelho, edmundo}@dei.uc.pt
[2] Federal University of Para, Rua Augusto Corra, 01, 66075-110, Belém, Brazil

**Abstract.** The Voice over Internet Protocol (VoIP) services are currently present in our personal and professional activities and will be key services in Next Generation Networks (NGN). Hence, in order to keep and attract new customers, the quality of delivery for VoIP services needs to be measured and optimized to ensure Quality of Service (QoS) and Quality of Experience (QoE) support to users in future multimedia networking systems. This paper presents the requirements to assess the quality level of VoIP services in NGN and analyzes the limitations of the well-known E-Model and Perceptual Evaluation of Speech Quality (PESQ) metrics for quality evaluation of VoIP services. Additionally, a new QoE metric, named Advanced Model for Perceptual Evaluation of Speech Quality (AdmPESQ), is proposed to overcome the limitations of current proposals concerning packet loss and packet delay awareness and to improve the VoIP assessment process. Performance evaluation was carried out based on simulation experiments to show the benefits of AdmPESQ in assessing the impact of VoIP services on the user's expectation.

**Keywords:** Quality of Service (QoS), Quality of Experience (QoE), Voice over Internet Protocol (VoIP), E-Model, Perceptual Evaluation of Speech Quality (PESQ)

## 1 Introduction

Voice over Internet Protocol (VoIP) services are now offered by different service providers and subscribed by a large number of fixed and mobile users in multimedia-aware networking systems. VoIP brings benefits for both service providers and customers. For service providers, operational costs are diminished due to the distribution of different services, such as voice and data, in a networking infrastructure with shared resources. For customers, VoIP presents the features of a traditional Public Switched Telephone Network (PSTN) phone, such as voice mail and conference, in a ubiquitous way as well as with quality level support and fair rates.

The quality level assessment for VoIP services in Next Generation Networks (NGN) with different Quality of Service (QoS) models (e.g., Differentiated Services (DiffServ) and IEEE 802.11e), wireless access technologies (e.g., from cellular to Wide Local Area Networks (WLAN)), mobility controllers (e.g., Mobile

IP (MIP) and Hierarchical MIP (HMIP)) is still a challenging research goal [1]. Due to the real-time behavior of voice services, implementation of different network technologies and human perception characteristics, the quality level control of VoIP services must be performed taking into account both network/packet requirements, such as packet loss, delay tolerances and user-level requirements related with human perception [2], such as noise.

Due to the limitations of traditional QoS solutions regarding voice-awareness and human perception, Quality of Experience (QoE) assessment approaches have been introduced to estimate the quality level of VoIP services from the user point of view, as presented in the Telecommunication Standardization Sector (ITU-T) Recommendation P.800 [3]. In order to assure QoE support for voice services in NGN, VoIP quality assessment methods and metrics are required to estimate/monitor the quality level of services coded with different CODEC and transmitted over wired and wireless heterogeneous networks with links with different packet loss probability and delay. The output of an assessment solution can be used for both costumers, to know the real quality of the subscribed services, and providers, as input for future network management and optimization procedures such as resource reservation, mobility prediction and QoE routing, as well as for pricing schemes.

Most of the current assessment schemes for controlling the quality level of voice content assume random values for transmission parameters, such as packet delay and loss, as well as do not consider the CODEC types during the evaluation process[4], thus reducing their applicability for a controlled and limited number of scenarios. The basic idea behind assessment models in NGN are twofold: which metric must be used to estimate the quality level of VoIP services taking into consideration different CODEC types, loss and delay values, and how to communicate with other network entities and standards to collect and send voice information.

This paper studies QoE assessment schemes for quality evaluation of voice calls. Moreover, it analyzes the limitations of current QoE metrics with focus on two main standards presented in Section 2, E-Model and Perceptual Evaluation of Speech Quality (PESQ), by using both conceptual and simulation evaluation. Following, a new QoE-aware VoIP assessment solution to be used in next generation multimedia systems is proposed and validated in a heterogeneous scenario. The results show correct results for scenarios with different delays and loss rates, which is not provided by current metrics. The proposed scheme helps providers to develop or adapt management mechanisms to improve the quality level of VoIP services, as well as to enhance pricing schemes. Additionally, it allows costumers to know about the quality of the subscribed services.

The remainder of this paper is organized as follows. Section 2 discusses the actual QoE metrics used for VoIP evaluation. Section 3 proposes a new metric that combines the positive aspects of both E-Model and PESQ in order to give a better overall evaluation. Section 4 presents the simulation results and analyzes the Advanced Model for Perceptual Evaluation of Speech Quality (AdmPESQ) improvements. Conclusions and future work are summarized in Section 5.

## 2 Related Work

This section presents current studies of VoIP assessment schemes with focus on PESQ and E-Model, because they are widely used and well-known standards. A detailed description on VoIP subjective and objective metrics can be found in [2].

Subjective metrics assess how audio calls are perceived by users, where the Mean Opinion Score (MOS) is widely used. These assessment schemes are carried out by participants, who evaluate prerecorded audio signals with different impairments. ITU-T recommends the evaluation of voice taking into account three opinion scales: quality, effort and loudness[3]. Each opinion scale is expressed from 1 (worst result) to 5 (best result).

Objective metrics allow a quality evaluation from the user point of view by using mathematical models. Among several existing metrics, ITU-T introduces two main methods for objective evaluation of VoIP calls, named E-Model and PESQ. The E-Model is based on the concept that impairments which affects voice calls are independent[5]. Five factors are considered: the basic signal-to-noise ratio ($Ro$), which includes sources of noise such as the environment, the impairments which occur more or less simultaneously with the voice signal ($Is$), the impairments caused by delay ($Id$), the impairment introduced by the equipment ($Ie_{eff}$), such as losses, and the advantage factor ($A$)[6][7]. The $A$ factor allows the compensation of impairment when there are other advantages of access to the user. Therefore, a conventional wired access has no compensation, while a wireless access in remote areas includes a high $A$ factor. Each parameter is calculated separately and the final result is obtained by Equation 1, where $R$ is the evaluation of the transmission on a scale from 0 (poor quality) to 100 (excellent quality).

$$R = Ro - Is - Id - Ie_{eff} + A \tag{1}$$

The E-Model result can be mapped into MOS evaluation using Equation 2.

$$\text{MOS(R)} = \begin{cases} 1 \quad \text{if } R < 0 \\ R(R-60)(100-R)7 \times 10^{-6} + 1 + 0.035R \quad \text{if } 0 < R < 100 \\ 4.5 \quad \text{if } R > 100 \end{cases} \tag{2}$$

Although the E-Model equation contain several variables, the amount of information required for its evaluation is minimal. Because they are essential to the calculation of $Id$ and $Ie_{eff}$, the delay and packet loss are always necessary. However, the $Is$ and $Ro$ factors have a low variation and default values can be used. Regarding $A$, which is define by the advantages of access to the user, presents a constant or non-dynamic values. Therefore, the calculation of the $R$ factor can be simplified according to Expression 3.

$$R = 93.2 - Id - Ie_{eff} + A \tag{3}$$

Based on [6] and [8] recommendations, an extension to E-Model was proposed in [5] in order to include a new concept, called recency effect. The recency effect takes into account the moment (time) when the losses occur, where losses at the end of the call lead to further degradation compared to losses at the beginning of the conversation. However, the proposed solution does not take into consideration human psychological aspects related with the impact of losses on the user and is not suitable for QoE-aware multimedia systems.

Another important metric is called PESQ. PESQ evaluates the QoE of VoIP service by comparing the original and processed signals. This metric can be used in different environments, from analog to digital multimedia networks. PESQ evaluation includes factors of distortion due to channel/encoder, losses and jitter. The effects of delay, echo, loudness loss, sidetone and impairments related to two-way interactions are not reflected in the PESQ scores[9][10]. PESQ presents values from -0.5 (lower value) to 4.5 (higher value), although for most cases the output range will be a listening quality MOS-like score between 1.0 and 4.5. The PESQ score cannot be directly mapped to MOS, but can be approximated to it[11].

The E-Model and PESQ metrics have several drawbacks. On the one hand, the E-Model metric does not take into account packet loss proprieties during the content distribution, as shown by our simulation results in Section 4. On the other hand, the PESQ shows a good accuracy for different losses, but does not take into consideration the packet delay factor along the end-to-end communication path. Both metrics are implemented and evaluated by several proposals to control the quality level of VoIP services in networks[12][13][14]. However, the use of current PESQ and E-Model versions are not suitable for networking environments with variable packet loss and delay values as expected for NGN.

From the related work analysis it is concluded that both E-Model and PESQ metrics have key drawbacks and needs to be improved to operate in real networking systems. Therefore, this paper proposes a new metric for VoIP assessment that combines both E-Model and PESQ aspects as presented in Section 3.

## 3 Advanced Model for Perceptual Evaluation of Speech Quality (AdmPESQ)

In the face of network resource restrictions, voice delivery through NGN leads to unavoidable quality degradation and a solution to assess how good audio services meet the user's expectation is a key requirement for multimedia systems. The AdmPESQ proposal assesses the quality level of VoIP services along heterogeneous wired and wireless networks by using a new QoE metric for voice calls that extends PESQ and E-Model models.

AdmPESQ is a full reference metrics implemented at end-hosts to produce a final score about the VoIP quality level. Since the AdmPESQ voice assessment results can be used for optimization and mobility schemes, open interfaces are defined to allow a tight communication with existing standards and solutions. The use of interfaces increases the system flexibility and the inclusion (or change) of network control mechanisms or policies. Moreover, an interface with Real-time Control Protocol (RTCP) is used to collect information about the end-to-end delay along heterogeneous communication paths. Other interface with Session Description Protocol (SDP) (transported in Real-time Streaming Protocol (RTP) or Session Initiation Protocol (SIP)) is implemented to acquire information about the VoIP CODEC negotiated during the session establishment process.

As presented before, E-Model and PESQ metrics have several limitations to be used in heterogeneous networks with dynamic packet loss and delay parameters. To overcome the above limitations and to provide an efficient assessment model, AdmPESQ was designed. The proposed solution combines important characteristics of both E-Model and PESQ, namely the impact of delay during the evaluation process used by E-Model, with the impact of packet loss, packet loss concealment, transmission channel errors and jitter used by PESQ.

When all default values are used, the E-Model result is $R = 93.2$. However, the ITU-T Recommendation G.113 [15] suggests different values of $Ie$ and $Bpl$ depending on the CODEC used. Thus, for different codifiers, the E-Model is calculated by:

$$R = 93.2 - Ie_{eff} \tag{4}$$

Because at this point only the impact of delay will be computed, losses will not be considered ($Ie_{eff} = Ie$) and therefore

$$R = 93.2 - Ie \tag{5}$$

In order to take into account the delay, the $Id$ must be added:

$$R = 93.2 - Ie - Id \tag{6}$$

where

$$Id = Idte + Idle + Idd \tag{7}$$

Since PESQ returns inaccurate values on the existence of echo, and therefore its use is not recommended in these conditions, only the $Idd$ will be taken into account. Thus,

$$R = 93.2 - Ie - Idd \tag{8}$$

where, for a given one-way delay $ta$, $Idd$ is calculated by

$$\mathrm{Idd} = \begin{cases} 0 \quad \text{if } ta \leq 100 \\ 25((1 + X^6)^{\frac{1}{6}} - 3(1 + (\frac{X}{3}^{6})^{\frac{1}{6}} + 2) \quad \text{if } ta > 100 \end{cases} \tag{9}$$

with

$$X = log_2 \frac{ta}{100} \tag{10}$$

Then the ratio between $MOS(93.2 - Ie - Idd)$ and $MOS(93, 2 - Ie)$ is calculated in order to take into account the impact of delay on the user perspective. Finally, $AdmPESQ$ measures the overall quality by combining the delay impact and $PESQ$ results:

$$AdmPESQ = PESQ\frac{MOS(93.2 - Ie - Idd)}{MOS(93, 2 - Ie)} \tag{11}$$

where MOS is defined in Equation 2.

The $A$ factor of E-Model is also added to take into account different users' tolerance and improve the AdmPESQ results:

$$AdmPESQ = PESQ\frac{MOS(93.2 - Ie - Idd + A)}{MOS(93, 2 - Ie + A)} \tag{12}$$

With AdmPESQ better VoIP quality evaluation results are achieved than with pure E-Model or PESQ schemes (results are presented in Section 4). Only a single metric is needed to perform QoE-based assessment for voice services in NGN. Providers can also use AdmPESQ as a manner to optimize network management operations, detect network impairments and define QoE-based pricing schemes.

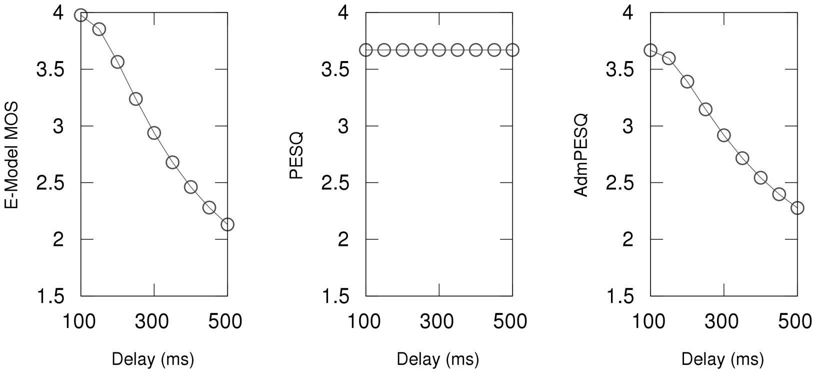
## 4 Performance Evaluation

To analyze the limitations of existing QoE metrics and the benefits of the AdmPESQ metric in dynamic heterogeneous environments, several simulation experiments were performed using the Network Simulator 2 (NS2)[16] and the VoIP module developed by the Technical University of Berlin[17]. Additionally, bugs were fixed, an intelligent drop mechanism was added and the calculation of E-Model and AdmPESQ was introduced. The DiffServ and IEEE 802.11e QoS models were used to provide QoS assurance in wired and wireless links respectively as expected for NGN.

The topology was generated by Boston University Representative Internet Topology Generator (BRITE)[18]. The simulated scenario is composed by two networks with sixteen routers each. One network hosts the source and another hosts the wireless receiver. The propagation delay is assigned by BRITE according to the distance between each device. The bandwidth capacity of wired and wireless links is of 100 Mb/s and 11 Mb/s, respectively. The source transmits a VoIP flow coded with ITU-T G.729 [19] to a wireless receiver. The network load, loss and delay are randomly changed to simulated the characteristics of different networking scenarios. The default $Ie$ and $Bpl$ values for ITU-T G.729 codifier are used in accordance with the ITU-T Recommendation G.113.

### 4.1 End-to-End Delay Variation

The delay is an important parameter in real-time VoIP services and directly affects the quality of service from the user point of view. High delays induce high



(a) E-Model (b) PESQ (c) AdmPESQ

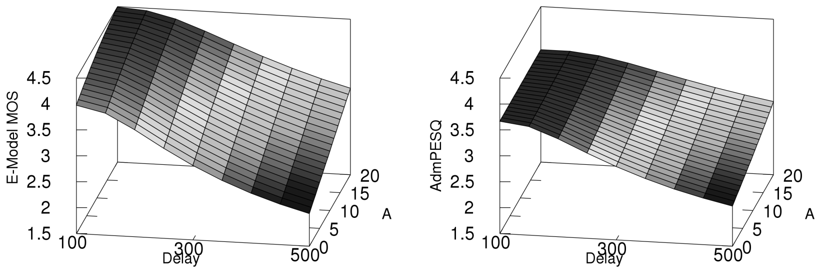
**Fig. 1.** QoE metrics variation with the increase of the delay

gaps between interlocutors responses and lead to discontentment. Several sources influence the delay, such as propagation, transmission, queuing and processing delay.

In order to analyze the impact of delay on user side and the performance of E-Model, PESQ and AdmPESQ metrics, several experiments were carried out in a scenario with assured bandwidth in the VoIP class (no loss) and the one-way delay varying. Since the E-Model can provide inaccurate predictions for values exceeding 500 ms, the delay ranges from 100 to 500 ms. Because there is no loss, for smaller values than 100 ms, the assessments are identical to the quality evaluation of 100 ms.

The behavior of E-Model and PESQ for different values of delay can be observed in Figure 1(a) and 1(b), respectively. The E-Model shows a continuous decrease on the VoIP quality level. Regarding PESQ, the values stay unchanged for the different values of delay, because the PESQ metric does not take into account this important parameter during the assessment process. The improvement of AdmPESQ is shown in Figure 1(c) when the delay parameter is considered to define the service quality level. When compared to PESQ, AdmPESQ has similar results for low delays values, but has a difference up to 60% when the delay is 500 ms. Because the impact of the delay on E-Model and AdmPESQ is calculated the same way, their variation are similar.

The variation of the E-Model and AdmPESQ metrics with the delay and the $A$ factor is depicted in Figure 2(a) and 2(b), respectively. The $A$ factor allows assessment models to take into account the tolerance of different users, which varies according to the access scheme used by them. In this context, PESQ performs poorly in dynamic and heterogeneous multimedia systems because it does not assume the existence of packet delay neither the $A$ factor in a network.



(a) E-Model (b) AdmPESQ

**Fig. 2.** QoE metrics variation with the the delay and $A$

### 4.2 Selective Loss Variation

The packet loss rate is another important metric to assessment models. Some losses are more harmful than others, depending on the time they occur and on the signal proprieties. This is clear in a situation when the interlocutor make pauses in its speech. For example, if losses occur in periods of silence, they have low impact on the VoIP quality.

To verify the impact of various types of losses on the user experience, three experiments were performed and evaluated with E-Model, PESQ and AdmPESQ. The links have sufficient bandwidth in the VoIP classes in both wired and wireless interfaces and a constant delay of 100 ms was defined. In order to verify the impact of losses in different VoIP packets (packets with content related to silence periods and not) a random packet drop scheme was accomplished.

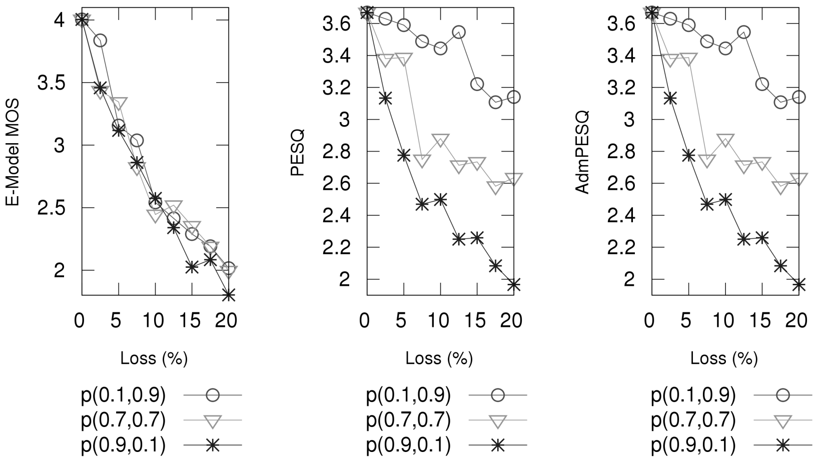
In the first scenario, a packet has a probability of 70% to be discarded, regardless of its content. In the second scenario, the drop is preferably done in packets that include voice content (no silence), according to Expression 13. Finally, the last scenario performs the drop of packets with silence content, and the probability defined in Expression 14.

$$\text{p(drop packet)} = \begin{cases} 0.9 & \text{if voice packet} \\ 0.1 & \text{if silence packet} \end{cases} \tag{13}$$

$$\text{p(drop packet)} = \begin{cases} 0.1 & \text{if voice packet} \\ 0.9 & \text{if silence packet} \end{cases} \tag{14}$$

Figure 3(a) illustrates the E-Model scores according to the number of losses occurred in the three different scenarios. The MOS calculated by the E-Model is directly influenced by losses, but the distinction between the various losses is not realized. Therefore, for each loss value, the E-Model only varies due to the different mean delays and fails in predicting the VoIP quality level in real networking environments.

As presented in Figure 3(b), PESQ performs well in scenarios with different type of packets dropped, because it uses a comparison between the sent and received voice signals. Regarding the scenario where the drop of packets containing



(a) E-Model (b) PESQ (c) AdmPESQ

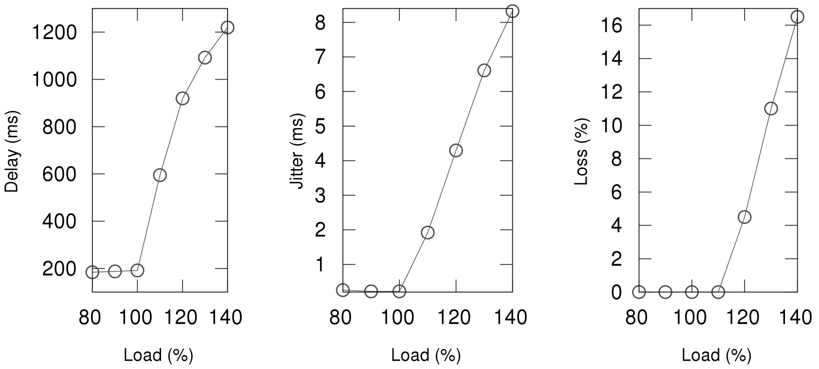
**Fig. 3.** QoE metrics behavior for different values of losses and scenarios with different probability of discarding packets p(x, y), where x and y represent the discard probability of a packet containing voice and silence, respectively

silence is analyzed, the quality evaluation assessed by PESQ shows approximatively the same results through the different loss rates. The scenario where the loss probability is high in period of speeches shows the worst results. Finally, the scenario where the discard is realized totally randomly presents the intermediate values. AdmPESQ shows similar results compared to PESQ, as depicted in Figure 3(c). The AdmPESQ scores presents a significant difference relatively to E-Model when the discard focuses packets containing silence, presenting a more accurate assessment.

### 4.3 Load Variation

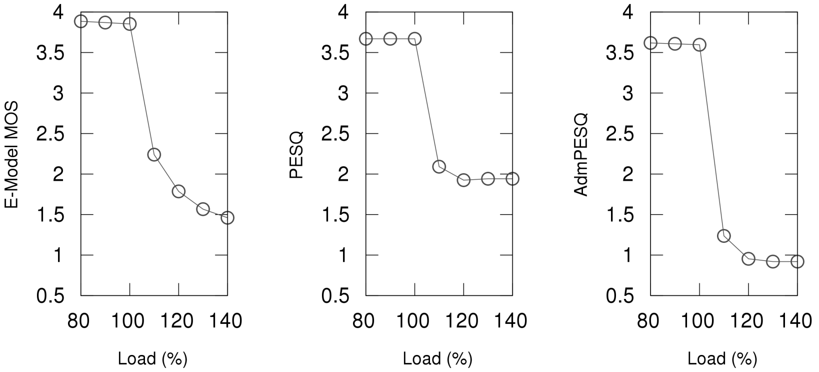
To verify the impact of the load variation on each QoE metric, several experiments were carried out with a fixed end-to-end delay of 150 ms and different loads causing a congestion up to 140% in the VoIP class in both wired and wireless links. With increasing of background traffic in a class, the number of packets present in queues also raises. This factor causes higher delays and jitters in VoIP services as shown in Figures 4(a) and 4(b), respectively. When the load is too high, the maximum number of packet in queues is exceeded, and some packets are discarded, as illustrated in Figure 4(c). Therefore a metric which combines both E-Model for delay and PESQ for loss performs better as presented by AdmPESQ.

Figure 5 depicts the behavior of E-Model, PESQ and AdmPESQ in the load variation scenario. For the increasing values of load, the E-Model shows a constant decrease in quality estimation process. This behavior is due to its mathe-



(a) Delay (b) Jitter (c) Loss

**Fig. 4.** QoS metrics variation with the load increase



(a) E-Model (b) PESQ (c) AdmPESQ

**Fig. 5.** QoE metrics variation with the load increase

matical model, which is focused on two metrics: the delay and loss. Since both metrics decrease, the E-Model also decrease steadily. However, the results can be incorrect because different types of packets are discarded. Regarding PESQ, the decline of its values is not linear. This because PESQ takes into account several factors of the audio signal. However, because it does not take into account the delay, it fails in predicting the overall VoIP quality level. Finally, AdmPESQ presents more accurate results than E-Model and PESQ, due to the combination of both metrics. When compared to PESQ, AdmPESQ has an accuracy up to 45% when the load is higher than 120%.

## 5 Conclusion

This paper analyzed the main limitations of well-known VoIP assessment metrics, namely E-Model and PESQ. In addition, a new metric, called AdmPESQ, was proposed to be used in NGN, where communication paths can have different packet loss and delay parameters. AdmPESQ combines the main attributes of E-Model and PESQ in a single metric, reducing the system complexity and optimizing the assessment process in hetereogeneous scenarios. The proposed solution can also be used for providers to adapt or develop new network resource and mobility controllers. Furthermore, simulation results demonstrate that AdmPESQ presents accurate results compared with E-Model and PESQ.

Future works will evaluate AdmPESQ in an experimental network and subjective QoE tests based on ITU-T recommendation will be performed to verify the efficiency of the proposed solution with real users.

## Acknowledgment

The work presented in this paper is supported by the European Commission, under the Grant No. FP6-0384239 (Network of Excellence CONTENT).

## References

1. S. Uemura, N. Fukumoto, H. Yamada, and H. Nakamura. QoS/QoE measurement system implemented on cellular phone for NGN. In *Consumer Communications and Networking Conference, 2008. CCNC 2008. 5th IEEE*, pages 117–121, Las Vegas, NV, 2008.
2. Marie Guguin, Rgine Le Bouquin-Jeanns, Valrie Gautier-Turbin, Grard Faucon, and Vincent Barriac. On the evaluation of the conversational speech quality in telecommunications. *EURASIP J. Adv. Signal Process*, 8(2):1–15, 2008.
3. ITU-T p.800.1 : Mean opinion score (MOS) terminology, July 2006.
4. S. Sengupta, M. Chatterjee, and S. Ganguly. Improving quality of VoIP streams over WiMax. *Computers, IEEE Transactions on*, 57(2):145–156, February 2008.
5. L. Carvalho, E. Mota, R. Aguiar, A.F. Lima, and J.N. de Souza. An e-model implementation for speech quality evaluation in VoIP systems. In *Computers and Communications, 2005. ISCC 2005. Proceedings. 10th IEEE Symposium on*, pages 933–938, June 2005.
6. ITU-T g.107 : The e-model: a computational model for use in transmission planning, March 2005.
7. Yoanes Bandung, Carmadi Machbub, Armein Z.R. Langi, and Suhono H. Supangkat. Optimizing voice over internet protocol (VoIP) networks based-on extended e-model. In *Cybernetics and Intelligent Systems, 2008 IEEE Conference on*, pages 801–805, Chengdu, China, September 2008.
8. A. D. Clark. Modeling the effects of burst packet loss and recency on subjective voice quality. In *IP telephony Workshop*, pages 123–127, New York, NY, USA, April 2001.
9. 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, February 2001.

10. A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, volume 2, pages 749–752, Salt Lake City, UT, May 2001.
11. H.M. Liang, C.H. Ke, C.K. Shieh, W.S. Hwang, and N.K. Chilamkurti. Performance evaluation of 802.11e EDCF in the ad-hoc mode with real audio/video traffic. In *Wireless and Optical Communications Networks, 2006 IFIP International Conference on*, April 2006.
12. M. Baratvand, M. Tabandeh, A. Behboodi, and A.F. Ahmadi. Jitter-Buffer management for VoIP over wireless LAN in a limited resource device. In *Networking and Services, 2008. ICNS 2008. Fourth International Conference on*, pages 90–95, Gosier, March 2008.
13. E.W.C. Peh, W.K.G. Seah, Y.H. Chew, and Y. Ge. Experimental study of voice over IP services over broadband wireless networks. In *Advanced Information Networking and Applications, 2008. AINA 2008. 22nd International Conference on*, pages 834–839, Okinawa, March 2008.
14. Zizhi Qiao, Lingfen Sun, and E. Ifeachor. Case study of PESQ performance in live wireless mobile VoIP environment. In *Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008. IEEE 19th International Symposium on*, pages 1–6, Cannes, September 2008.
15. ITU-T g.113 : Transmission impairments due to speech processing, November 2007.
16. The network simulator - ns-2. http://www.isi.edu/nsnam/ns/, May 2009.
17. Predicting the perceptual service quality using a trace of VoIP packets. http://www.tkn.tu-berlin.de/research/qofis/, May 2009.
18. BRITE: boston university representative internet topology generator. http://www.cs.bu.edu/brite/, May 2009.
19. ITU-T g.729 : Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP), 2007.