Journal of Computer Networks and Communications

Improving QoS and QoE for Mobile Communications

Guest Editors: Pedro Merino, Maria G. Martini, Raimund Schatz, Lea Skorin-Kapov, and Martin Varela

Improving QoS and QoE for Mobile Communications

Journal of Computer Networks and Communications

Improving QoS and QoE for Mobile Communications

Guest Editors: Pedro Merino, Maria G. Martini, Raimund Schatz, Lea Skorin-Kapov, and Martin Varela

Editorial Board

Contents

Hindawi Publishing Corporation
Journal of Computer Networks and Communications
Volume 2013, Article ID 645174, 2 pages
http://dx.doi.org/10.1155/2013/645174

Editorial
Improving QoS and QoE for Mobile Communications

Pedro Merino,[1] Maria G. Martini,[2] Raimund Schatz,[3] Lea Skorin-Kapov,[4] and Martin Varela[5]

[1] *Technical High School of Telecommunications Engineering, University of Málaga, Campus de Teatinos, 29071 Málaga, Spain*
[2] *Faculty of Science, Engineering, and Computing, Kingston University London, Surrey KT1 2EE, UK*
[3] *Telecommunications Research Center Vienna (FTW), 1220 Vienna, Austria*
[4] *Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 10000 Zagreb, Croatia*
[5] *VTT Technical Research Centre of Finland, P.O. Box 1100, 90571 Oulu, Finland*

Correspondence should be addressed to Pedro Merino; pedro@lcc.uma.es

Received 1 August 2013; Accepted 1 August 2013

The proliferation of fast, cheap, and ubiquitous network access, in particular on mobile devices, has significantly changed the way users access online services on a day-to-day basis. Along with the increased availability of an always-on Internet, new technologies such as LTE have enabled the use of all types of services, including those that, such as video streaming, come with stringent requirements in terms of network performance and capacity demands. Consequently, the issue of adequately supporting all those users and all those services is a nontrivial one for network operators, as users expect certain levels of quality to be upheld.

In this special issue, we explore the relation between the Quality of Service (QoS) that operators monitor and manage and the Quality of Experience (QoE) that the users actually get. A good understanding of this relation is critical for network and service providers alike in order to properly provision and truly enhance their offerings so that end users receive the quality levels that satisfy or even delight them while at the same time enabling the efficient use of network resources.

The paper *"Dealing with energy QoE trade offs in mobile video,"* authored by F. Liberal et al., addresses the transmission of scalable video over mobile networks. In particular, the focus is on the trade-off energy versus quality: depending on the specific type of service and network scenario, end users and/or operators may decide to choose among different energy versus quality combinations. In order to deal with this trade-off, the paper first proposes a single-objective optimization that is able to minimize the energy consumption with the constraint of a minimum acceptable quality. In order to reflect the fact that the same increment of energy consumption may result in different increments of visual quality, a multiobjective optimization is also proposed, with the associated utility function. Finally, in order to reduce complexity, a heuristic algorithm is proposed.

"Angry apps: the impact of network timer selection on power consumption, signalling load, and web QoE," authored by C. Schwartz et al. discusses the trade-off between energy consumption in smartphones and the generated signaling traffic in the mobile network. They use measurements made in public 3G networks to characterize the impact of inactivity timers in the network in energy consumption for different kinds of applications: bandwidth insensitive, interactive, and background applications. With this information, the authors study how the network operator could optimize the timers to keep the quality of service of a given application. However, such optimizations could also provoke more signaling traffic due to frequent connection reestablishments, and the authors discuss some of the mechanisms that try to resolve such conflicts. Finally, they specifically study the impact of the timer settings on QoE for web browsing.

In their paper *"Survey and challenges of QoE management issues in wireless networks,"* S. Baraković and L. Skorin-Kapov provide a state-of-the-art survey on research in the field of QoE management for wireless networks. By identifying and discussing the key aspects and challenges of QoE modeling, monitoring, adaptation, and optimization from a wide interdisciplinary perspective, the paper enables a

better understanding of QoE (and its relationship to QoS) as well as the process of its management in converged wireless environments. In particular, their work focuses on questions of what QoE influence factors to control, where to implement control mechanisms (e.g., network, client device), when to invoke them, and finally how to provide such control mechanisms, with the ultimate goal of realizing QoE optimization strategies that balance end-user as well as network provider interests.

The paper *"Towards a QoE-driven resource control in LTE and LTE-A networks"* by G. Gómez et al. focuses on meeting QoE requirements from a mobile network operator point of view. The authors propose an architecture for achieving QoE-driven resource control in LTE and LTE-advanced networks, based on the introduction of a centralized QoE server responsible for collecting relevant performance indicators, making QoE estimations, and triggering potential actions towards standardized 3GPP Evolved Packet System entities. The authors identify key performance indicators at both the application and network levels to be used for QoE estimation across different types of applications and highlight potential use cases for their proposed approach.

Finally, the paper *"Obtaining more realistic cross-layer QoS measurements: A VoIP over LTE Use Case,"* by F. J. Rivas et al., presents a real-time testbed enabling realistic cross-layer measurements over LTE. This testbed enables the correlation of information at different layers, from service and IP levels to radio and protocol parameters. The analysis of the interlayer dependencies will support the identification of optimal settings for the radio access network and service parameters. New cross-layer optimization strategies could be suggested based on the observations provided. As a use case, VoIP over LTE is addressed.

Acknowledgments

This special issue cannot be produced without the effort of all the authors that responded to the call for papers and the work of the external reviewers. We are grateful to all of them.

Pedro Merino
Maria G. Martini
Raimund Schatz
Lea Skorin-Kapov
Martin Varela

Hindawi Publishing Corporation
Journal of Computer Networks and Communications
Volume 2013, Article ID 405858, 10 pages
http://dx.doi.org/10.1155/2013/405858

Research Article

Obtaining More Realistic Cross-Layer QoS Measurements: A VoIP over LTE Use Case

F. Javier Rivas, Almudena Díaz, and Pedro Merino

Department of Languages and Computer Science, University of Malaga, UMA, Campus Teatinos, 29071 Malaga, Spain

Correspondence should be addressed to Almudena Díaz; almudiaz@lcc.uma.es

Received 12 December 2012; Revised 4 June 2013; Accepted 9 July 2013

Academic Editor: Maria G. Martini

We introduce a real-time experimentation testbed in this paper which enables more realistic analysis of quality of service (QoS) in LTE networks. This testbed is envisioned for the improvement of QoS and quality of experience (QoE) through the experimentation with real devices, services, and radio configurations. Radio configurations suggested in the literature typically arise from simulations; the testbed provides a real and controlled testing environment where such configurations can be validated. The added value of this testbed goes a long way not only in the provision of more realistic results but also in the provision of QoS and QoE cross-layer measurements through the correlation of information collected at different layers: from service and IP levels to radio and protocol parameters. Analyzing the interlayer dependencies will allow us to identify optimal settings for the radio access network and service parameters. This information can be used to suggest new cross-layer optimizations to further improve quality of experience of mobile subscribers. As a use case, we examine VoIP service over LTE, which is currently an open issue.

1. Introduction

The enhancement of QoS in a sustainable manner is a critical goal for network operators as management tasks are becoming increasingly complex. Although some initial efforts have been carried out by the standardization bodies, there is still a significant gap to be covered in QoS and also in QoE optimization. Actually, current efforts towards improving QoS and QoE are typically based on estimations derived from costly drive test campaigns. Furthermore, involvement of human expertise is required to manually tune network configurations. On the other hand, most of the service and network configurations available in the literature are derived from simulations [1–6]. As is widely known, in the process of modeling communication systems to simulate them, some details may be missed, and thus, misleading results may be derived. For example, it is very common to find that the consumption of control resources is ignored when evaluating different scheduling methods. In this context, providing optimized network configurations based on measurements obtained directly from the subscribers' terminals and correlated with the information collected at the network will pave the way for a reduction of costs and more accurate tuning of network operation from the point of view of the QoS perceived by final users. Moreover, as stated by standards organizations (SDOs) or alliances with the participation of network operators such as NGMN [7], *the optimization of QoS still requires "real" developments to further study the direction in which to move forward*.

In this paper, we propose the use of an experimental testbed [8] implemented by our research group to carry out specific long term evolution (LTE) experiences in a real context and to extract the correlation between LTE radio configurations and QoS parameters perceived at the application level. The execution of exhaustive measurements campaigns using this testbed will enable the identification of specific performance counters, correlations between them, and use cases for QoS and QoE optimization in LTE networks.

The focus of the paper is on VoIP calls over LTE, which pose new challenges over previous technologies. In LTE, voice calls are now delivered through an all-IP network (VoIP) instead of a circuit switched one, which means that voice has to compete for bandwidth with other services provided in the network. It is vital to at least guarantee the same QoE for VoIP calls that was available in pre-LTE technologies such as global system mobile communications (GSM) and

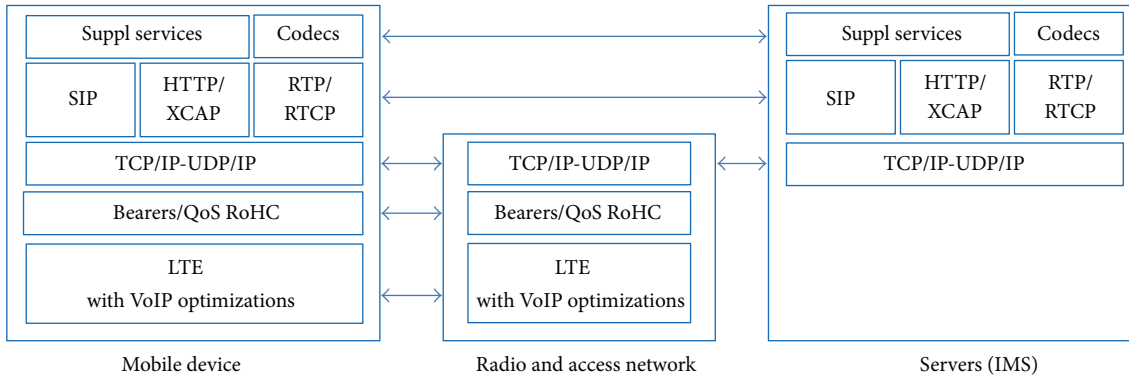


Figure 1: Scope of VoIP deployment over LTE by GSMA PRD IR.92 v6.0.

universal mobile telecommunications (UMTS). This will be required to avoid significant impact on customers, who will demand a good service in all-IP mobile networks. Due to situations like this, the testbed has been conceived to validate the performance of the network configurations and problems presented in the research literature. Specifically for VoIP service we have correlated layer 1 and layer 2 LTE radio parameters with IP performance parameters and mean opinion score (MOS) measurements base on perceptual evaluation of speech quality (PESQ) algorithm.

The organization of the paper is as follows. Section 2 introduces the necessity of obtaining performance measurements which capture the QoS and QoE as perceived by final users and the new challenges presented in the provision of voice call services in LTE. This section provides also a brief state of the art on VoIP over LTE and some noteworthy LTE configurations proposed in the literature to optimize its performance. Section 3 presents the testbed and the configuration under which the measurements were collected. Section 4 provides the results obtained during the analysis of IP performance parameters and Section 5 provides their correlation with LTE parameters. Finally, in Section 6, we present conclusions and future work.

2. Voice Calls over LTE: A Regular Data Service, the Same Quality as before

Third generation partnership project (3GPP) has standardized two solutions for the deployment of voice call service over LTE. The first is circuit-switched fallback (CSFB) [9], which implies a shift of the user equipment (UE) access from LTE to 2G/3G during a voice call. The second is VoLTE [10] which on the contrary is based on IP multimedia subsystem (IMS) and does not require the use of legacy technologies. Another alternative available in the market is voice over LTE via generic access (VoLGA) [11]. That specification has been developed by the VoLGA forum, based on the existing 3GPP Generic Access Network (GAN) standard [12]. CSFB and VoLGA provide interim solutions for early LTE deployments, while VoLTE offers a long term opportunity for mobile operators. VoLTE allows integrating voice and Internet services and delivering new multimedia services in a permanent environment. This will enable the exploitation of the potential offered by mature LTE networks. In this context, VoLTE has emerged as the preferred solution by carriers and the GSM association (GSMA) is developing a specification for delivering integrated telephony services over LTE.

Specifically, the GSMA defines in [10], the minimum mandatory set of features that a wireless device and a network should implement to support a high quality IMS-based telephony service over LTE radio access. The scope provided by GSMA is shown in Figure 1.

Standardization bodies are confident about the necessity of introducing specific LTE configurations for the deployment of VoIP service. In [10] GSMA proposes a list of LTE configurations which we aim to extend with new ones obtained and validated in the testbed proposed.

As stated in the 3GPP initiative, the multiservice forum has already demonstrated successful VoLTE calls, and also multi media telephony (MMTel) services [13]. During these tests, equipment from 19 manufacturers was used. The tests performed focused on validating the interoperability between the interfaces defined in the 3GPP technical specifications.

Network operators have also performed testings experiments for quality of service (QoS) measuring. For example, different performance metrics such as latency or throughput are evaluated for the TeliaSonera network in [14]. However, our work aims to go a step beyond and not only measure the performance of interfaces and terminals using individual metrics but also correlate all these measurements with specific LTE parameters in order to identify optimum configurations.

2.1. Some Considerations about the Transport of VoIP over LTE. As proposed by GSMA in [10] session initiation protocol (SIP) is the protocol used to register UE in the IMS server. real-time transport protocol (RTP) and user datagram protocol (UDP) are the protocols recommended to voice transport, and RTP control protocol (RTCP) to provide link aliveness information, while the media are on hold.

The most restrictive performance indicators for interactive real-time services such as VoIP are the end-to-end delay and jitter. The maximum allowed one-way delay for voice service is 300 ms as stated in [15], with a recommended value lower than 150 ms. The low latency of LTE access

(20–30 ms) reduces the end-to-end delay obtained in previous cellular technologies. However, LTE radio bearers do not employ fixed delay. Instead, fast retransmissions are used to repair erroneous transmissions, and uplink and downlink transmissions are controlled by schedulers. Consequently, LTE transmissions introduce jitter, which implies that UE must implement efficient dejitter buffers. The minimum performance requirements for jitter buffer management of voice media are described in [16]. In Section 4, IP parameters will be analyzed in more detail, while in this section, we will continue with the introduction of some concepts which will identify exactly what is standardized in VoIP over LTE, what is not, and how to generate specific configurations to improve the performance of VoIP over LTE.

The objective of radio resource management (RRM) procedures in LTE is to ensure an efficient use of the resources [17]. RRM algorithms at the eNodeB involve functionalities from layer 1 to layer 3. Admission control mechanisms, QoS management, and semipersistent scheduling are deployed at layer 3, while hybrid adaptive repeat and reQuest (HARQ) management, dynamic scheduling, and link adaptation are in layer 2 and channel quality indicator (CQI) manager and power control in layer 1.

3GPP specifies RRM signaling, but the actual RRM algorithms are not provided [17]. The combination of radio bearers that a UE must support for voice over IMS profile is defined in [18] Annex B. Concretely, the voice traffic requires a guaranteed bit rate (GBR) bearer, as described in [19]. The network resources associated with the evolved packet system (EPS) bearer supporting GBR must be permanently allocated by admission control function in the eNodeB at bearer establishment. Reports from UE, including buffer status and measurements of UE's radio environment, must be required to enable the scheduling of the GBR as described in [20]. In uplink, it is the UE's responsibility to comply with GBR requirements.

In the following subsections, we will analyze separately those parameters and configurations which have been identified in the current state of the art as optimized LTE solutions for VoIP.

2.1.1. Quality Class Indicator. The characteristics of the bearers are signalized with a QoS class identifier (QCI). The QCI is a pointer to a more detailed set of QoS attributes, including layer 2 packet delay budget, packet loss rate, and scheduling priority. As defined in [21], QCI 1 is intended for conversational voice.

2.1.2. RLC Mode Configuration. As specified in [10], the unacknowledged mode (UM) should be configured at radio link control (RLC) layer for EPS bearers with QCI 1 to reduce traffic and latency.

2.1.3. DRX Mode. Support of LTE discontinuous reception (DRX) methods for both UE and network is mandatory to reduce power consumption on mobile devices. The idea behind DRX methods is that the terminal pauses the monitorization of control channels during some periods of time, allowing it to turn the radio off. DRX parameters can be tuned depending on radio resource control (RRC) status or service. Decisions about when the radio should be activated again can be based on QoS indicators. The simulations conducted in [1] give, for VoIP applications, a potential saving of about 60 percent.

2.1.4. Compression. In order to optimize, radio resources the UE and the network must support robust header compression (RoHC) to minimize the size of IP packets during VoIP calls [10]. As we have already said, the use of UM at RLC and reduced sequence number sizes also decrease overhead. The reduction of packet size will enable the improvement of coding efficiency which is especially important for uplink scenarios and to improve the quality of data connection in areas with poor coverage.

2.1.5. Semipersistent Scheduling. The mechanisms involved in packet scheduling at the eNode-B are three: dynamic packet scheduling, link adaptation, and hybrid adaptive repeat and request management (H-ARQ). Semipersistent scheduling significantly reduces control channel overhead for applications that require persistent radio resource allocations such as VoIP. There are many different ways in which semi-persistent scheduling can configure persistent allocations.

Simulated results obtained in [2] show that in uplink direction, semi-persistent scheduling can support higher capacity than dynamic scheduling while at the same time guaranteeing VoIP QoS requirement, but with the cost of sacrificing some statistical multiplexing gains from HARQ. As dynamic scheduling is already needed for other services, they suggest using dynamic scheduling by default for VoIP. As an exception, semi-persistent scheduling should be applied for some VoIP users only in situations where the signaling load becomes too high.

2.1.6. Admission Control. Admission control mechanisms determine whether a new evolve packet system (EPS) bearer request should be admitted by checking that QoS requirements of at least all the bearers with high priority are fulfilled. Specific rules and algorithms for admission control are not specified by 3GPP. An interesting technique is proposed by AT&T Labs Research in [3] based on an intelligent blocking algorithm (IBA) for the admission process of VoIP calls subject to individual customers' blocking objectives and which is only invoked when the total bandwidth in use is close to its engineering limit. The algorithm has also been validated with simulations.

2.1.7. Packet Bundling in Downlink. Traffic generated by voice codecs can be very bursty due to the nature of the service, that is, silences in the conversations, and the nature of the codecs, for example, traffic from the adaptive multi-rate (AMR) codec. The combination between VoIP traffic patterns with dynamic packet scheduling can lead to inefficient resource utilization that might be improved by the use of packet bundling in the downlink. Several simulation studies have been done for both universal terrestrial radio access network (UTRAN) [4] and evolved universal terrestrial radio

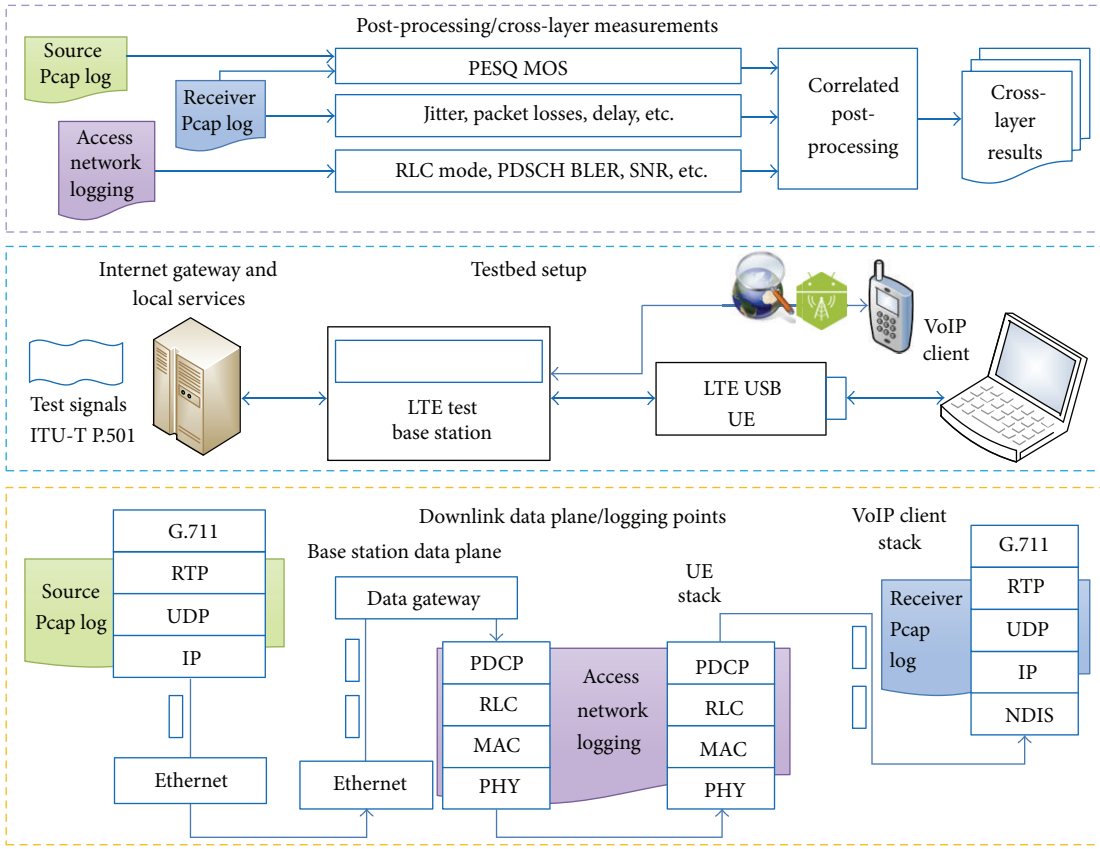


Figure 2: Testbed configuration.

access (E-UTRAN) [5] probing that dynamic packet bundling approaches improve VoIP services.

2.1.8. RLC Segmentation and TTI Bundling in Uplink. UE has limited transmission power, and at the edge of LTE cells, there is a high probability of obtaining an error in the transmission of VoIP packets because the device is not able to gather enough energy during one transmission time interval (TTI) (1 ms) to send the packet. In the case of unsuccessful transmission, HARQ retransmissions are required, which will imply the introduction of 8 ms delay per retransmissions. A large number of retransmissions will involve an intolerable increase in the delay for a conversational voice service and reduction of the transmission efficiency.

The conventional approach used to reduce delays and improve the coverage at the cell edge is RLC segmentation. It consists of the segmentation of service data unit (RLC SDU) and their transmission in consecutive TTIs. However, this is not an optimum solution from the point of view of the overhead introduced in the control signaling and the increase of the vulnerability of packet loss due to HARQ feedback errors. The idea behind TTI bundling is that for a given transport block a fixed number of transmissions is done in consecutive TTIs without waiting for the HARQ feedback. The eNodeB sends the corresponding HARQ feedback only when it has received the whole bundle of transmissions. This approach reduces the amount of HARQ feedback significantly [6]. The L2 header overhead is also reduced because there is a lower need for segmentation, as well as the signaling overhead required for uplink grants (i.e., resource allocations) because a single grant is required for each TTI bundle.

3. Testbed Configuration for VoIP Testing

We have composed an experimental testbed [8] with the aim of providing a realistic test scenario where previous and new radio configurations could be deployed. Additionally, it is possible to analyze their interactions and to verify cross-layer performance of Internet applications and services over LTE. Moreover, the testbed can be used to reproduce, in a controlled environment, behaviors captured in field test campaigns [8].

The testbed includes an LTE test base station from AT4 wireless which provides high performance protocol and radio capabilities behaving as an actual LTE radio access network (RAN), as shown in Figure 2. It also includes features such as emulation of channel propagation that allows modeling fading and additive white gaussian noise impairments, in addition to a high degree of configurability of the LTE stack and logging functions. The LTE test base station supports the connection of real LTE terminals and the transport of IP traffic generated by commercial applications installed on

them. The non access stratum (NAS) signalling exchange is provided by a core network emulation. Although the effect of core network transportation is also important for QoS, it has not been analyzed in the present work because of the focus on RAN and will be addressed in the future. The mobile terminals also incorporate advanced monitoring software [22, 23]. Finally, the testbed includes postprocessing tools which enable the testing and identification of IP connectivity issues and LTE mismatches through the correlation of logs collected at different points as shown in Figure 2.

During the experiment, VoIP calls are initiated by a commercial VoIP client running in a laptop. The laptop uses the Samsung GT-B3730 USB LTE modem connected, via a radio frequency (RF) wire, to the E2010 eNodeB emulator from AT4 wireless. The emulator is connected to the Internet and to a local Asterisk server via a proprietary data gateway. The core network is not presented in the current version of the testbed. In this work, we focus on the study of radio access interface performance from the point of view of QoS and QoE perceived at the user equipment (UE). In order to automate the establishment of the calls, the Asterisk server is configured to provide a callback service, so that this service reproduces a 30-second recording each time a call is received in a preconfigured VoIP extension. Records have been extracted from audio samples provided in the (international telegraph union telecommunication standardization sector (ITU-T) recommendation P.501 to speech quality evaluation on telephone networks. The codec used during the transmission is the G.711 because it is well known and its constant bit rate eases the initial analysis of the impact in throughput and similar metrics. Other codecs such as AMR (codec recommended for VoLTE profile in [10]) or SILK (Skype) will be addressed in future experiments.

Wireshark is used to capture the IP traffic on both sides. The emulator also provides low level EUTRAN traces that are valuable for detailed examination of behaviors of interest. The postprocessing of the results collected is carried out using a tool developed in our research group which, among other things, obtains delay, packet losses, jitter, and MOS values of the VoIP calls. perceptual evaluation of speech quality (PESQ) algorithm defined in [24] is used to calculate MOS. PESQ analyzes relative degradation between the original and the received voice signals. In order to apply the algorithm, both waveforms have to be provided as input. The source signal is directly fed to the server, but for comparison purposes, we obtain it from the generated IP traffic to isolate it from server encoding effects, whereas the received voice is reconstructed from the IP traffic recorded by Wireshark at the destination end.

Table 1 contains an example configuration deployed in the LTE test base station.

Different fading and noise propagation conditions have been applied. Multipath fading conditions are typically experienced in mobile environments as a result of the user mobility. A typical environment with low delay spread is represented with the EPA5 profile as defined in [25]. EPA stands for Extended Pedestrian A channel model, which contains 7 channel taps with an average delay spread of 45 ns and a maximum tap delay of 410 ns. The EPA5 profile has an associated maximum Doppler frequency of 5 Hz, and the associated tap delay and relative power is shown in Table 2.

Table 1: Resource scheduling and radio frequency configuration in the LTE test base station.

Parameter	Configuration
MIMO configuration	2 × 2
Channel bandwidth	10 MHz
Reference signal power	−60 dBm/15 kHz
Noise power	−67 to −73 dBm/15 KHz
Max HARQ retransmissions	3
RLC transmission mode	UM
RLC sequence number size	5
PDCP discard policy	No discard
PDCP sequence number size	7
Peak PDSCH bandwidth	120 Kbps
Resource allocation	Periodic, 5 ms
Modulation	16QAM
MCS (mod.& coding index)	13
PHICH duration	Normal
PHICH resources	1/6
Number of PDCCH symbols	1
Specific aggregation level	2
Fading profile	EPA5

Table 2: Extended Pedestrian A channel model.

Excess tap delay (ns)	Relative power (dB)
0	0.0
30	−1.0
70	−2.0
90	−3.0
110	−8.0
190	−17.2
410	−20.8

As demonstrated in the table, the signal to noise ratio (SNR) has been swept in a range of 7 to 13 dB to analyze the results under moderate packet loss conditions. As the VoIP service has real-time requirements, the RLC layer is configured to operate in unacknowledged mode (UM), that does not retransmit. However, although the RLC UM does not retransmit unconfirmed data, the hybrid automatic repeat request (HARQ) at medium access control (MAC) level provides convenient fast retransmission with incremental redundancy. Thus, even in the presence of a moderate Physical downlink shared channel (PDSCH) bLock error rate (BLER), a higher layer protocol data unit (PDU), will only be lost if the maximum number of HARQ retransmissions is reached at MAC level.

4. Analysis of IP Performance

The aim of this paper is not only to introduce a reference framework for measurements of IP services deployed over LTE but also the importance of cross-layer results in the

context of the VoIP service. Using the testbed described in the previous section, a campaign of experiments have been carried out to obtain the results referred to. Also cross-layer correlations between LTE mechanism and IP performance are explained.

4.1. IP Parameters. As we have stated in previous sections, some variable bit rate codecs are expected to be used in VoIP over LTE however, in this approach we have chosen the G.711 codec to compare because it is a standard and widely studied codec with constant bit rate (CBR). G.711 codec has a 64 kbps voice bandwidth. The constant sampled rate of 20 ms and the fixed 160 bytes of the payload plus 40 bytes of IP/UDP/RTP header produce a flow with a bandwidth of 80 kbps at the IP level. At the radio access a peak PDSCH bit rate of 120 kbps has been scheduled to provide enough throughput headroom. We have analyzed the IP bandwidth of the flow received by the mobile device during 10 consecutive VoIP calls in the worst-case scenario, that is, the scenario with 7 dB of SNR. The call length is 30 seconds. Results are depicted in Figure 3(a). The instantaneous evolution of the IP bandwidth for the 10 calls is compared with the nominal 80 kbps constant bit rate generated by the source. The IP bandwidth fluctuations obtained at the destination are caused by lost and delayed packets, which cause instantaneous decrements of the received bit rate. Successful retransmissions generate bandwidth peaks one second after the decrement, because the calculation is made averaging the received packets during the last second. Although we have used the LTE same fading profile and nominal signal to noise ratio in all the calls, different instantaneous results have been obtained because of the random nature of the fading and noise generators.

A packet loss rate close to 0% and a jitter lower than 2 ms can provide good quality VoIP calls, even comparable with a public switched telephone network (PSTN) call. However most codecs used in the VoIP service are not tolerant of higher packet losses. For the "standard" G.711 codec or the G.729 codec, a 1% packet loss rate significantly degrades a call [26]. In Figure 3(b), we depict packet losses obtained during sessions where different SNR were configured. We can see that packet losses are higher than 1% only for VoIP call with the lower configured SNR, 7 dB.

The interarrival jitter is calculated as defined in [27] using the IP traces captured at the mobile subscriber terminal. Each RTP packet contains a timestamp which reflects the sampling instant of the first octet in the RTP data packet. The instantaneous variation of the delay is obtained by comparing the elapsed time between two received packets with the difference between their timestamp. The jitter is then derived applying a filter to the instantaneous delay variation. In Figure 3(c), we observe the temporal evolution of instant jitter during VoIP calls conducted in the scenario configured with 7 dB of SNR, while Figure 3(d) shows the mean jitter obtained in all the scenarios. This is a traditional analysis based on only IP parameters, which is very useful to characterize the performance of the service under study. However, it is not enough for the adaptation and optimization of the service to the underlying transport technology. This can be better appreciated by observing Figure 3(d). Concretely, it can be seen that despite the delay variations introduced by HARQ retransmissions, the average jitter is kept in only a few ms, although eventually the instantaneous delay may vary in the order of tens of ms. However, to obtain a better comprehension of VoIP performance over LTE, it is necessary to monitor low level parameters and correlate them with IP parameters. In the following section, we will analyze the correlations between parameters monitored at different LTE layers.

5. Cross-Layer Measurement Analysis

In this section, a further analysis of the results obtained in previous experiments is provided. Specifically, we will present a correlation of the SNR configured in the experiments with different IP and RF measured parameters. In addition, we will also depict the mapping of voice quality measurements. Known functions (linear, polynomial, exponential, and logarithmic functions) have been applied to obtain the correlation between the parameters. To retrieve the degree of correlation, the coefficient of determination R^2 has been calculated. R^2 ranges from 0 (indicating the absence of a systematic correlation) to 1 (indicating a perfect correlation).

5.1. HARQ and Packet Losses. Figure 4(a) shows the effect of the SNR on the packet loss rate. We have represented the mean value, as well as the minimum and the maximum packet loss rate to illustrate the maximum variability for a given SNR value. It must be noted that although in some points the slope seems to change, these effects may appear because of the randomness of the propagation conditions, these magnitudes require large statistical analysis and in a limited set of experiments small deviations may appear in the results.

We have also compared the PDSCH BLER with the IP packet loss rate. The PDSCH BLER represents the ratio of correctly acknowledged transport blocks to the total number of transmitted transport blocks. As the SNR decreases, the effect of the noise makes the PDSCH BLER increase. In Figure 4(b), we correlate the PDSCH BLER with the packet loss rate. In absence of HARQ, the PDSCH BLER should match the packet loss rate, but HARQ reduces the rate of packet losses at the cost of additional use of PDSCH resources to allocate retransmissions. As we are operating in relatively ideal conditions, the relation between the BLER and the effective bandwidth reduction is approximately linear as a single retransmission will succeed typically. In worse SNR conditions, the ratio of lost packets rate to PDSCH BLER would be even further reduced at the expense of a more noticeable impact on the bandwidth.

Other useful magnitudes are also related with SNR and parameters. The CQI is a magnitude reported by a mobile device, with a configurable periodicity, that provides an estimation of the instantaneous quality of the channel. The larger the reported CQI, the higher the coding rate (lower redundancy) that can be used for transmission. We have verified that the reported average CQI decreases consistently as the noise increases, and in future work, we will provide

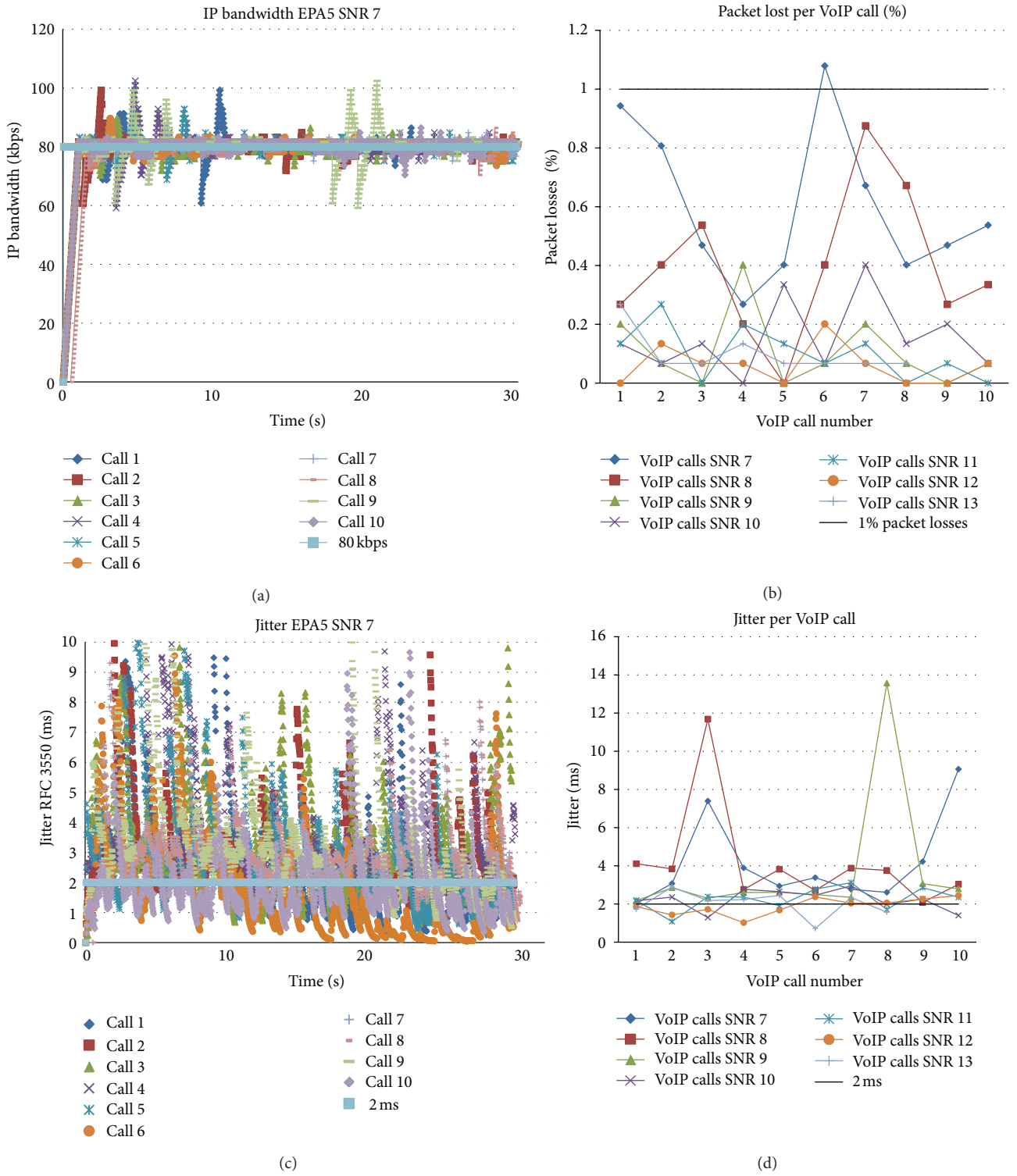


Figure 3: IP performance parameters analysis. (a) Instant IP bandwidth measurements during 10 VoIP calls in an EPA5 LTE scenario with a 7 dB of SNR. (b) Packet losses per VoIP call for different levels of SNR in an EPA5 scenario. (c) Instant jitter measurements during 10 VoIP calls in a EPA5 LTE scenario with a 7 dB of SNR. (d) Mean Jitter per VoIP call for different levels of SNR.

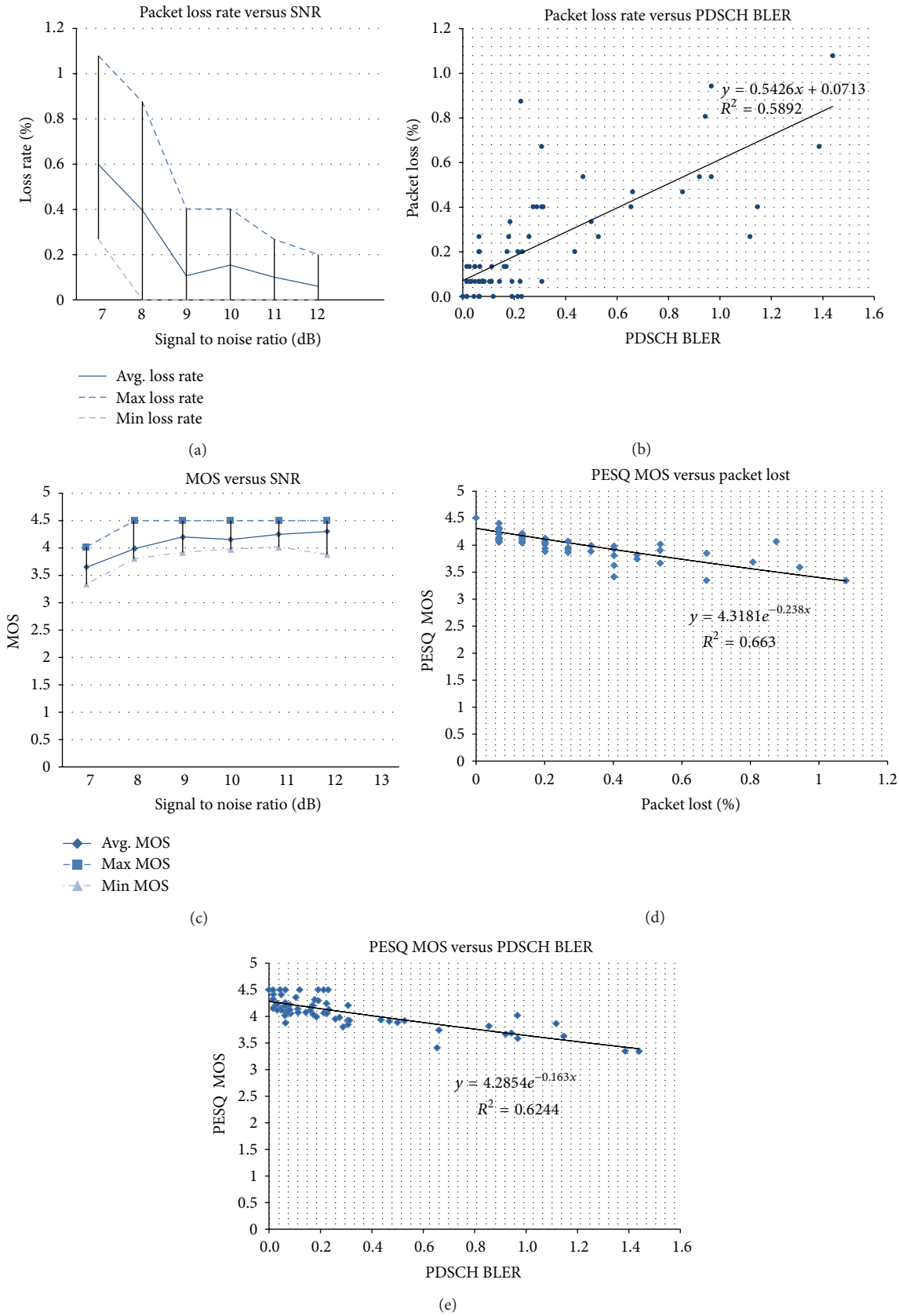


Figure 4: Cross-layer measurements and correlations. (a) Packet Loss versus SNR. (b) Packet Loss versus PDSCH BLER. (c) MOS versus SNR. (d) PESQ MOS versus packet lost. (e) PESQ MOS versus PDSCH BLER.

detailed results of the mapping of PDSCH BLER to reported CQI for different cell configurations and propagation conditions. Furthermore, we will also analyze different CQI adaptive schedulers, that will react to the instantaneous received CQI to provide appropriate resource allocations.

In general, low R^2 values have been obtained (0.7 or less) with the equations used, which exposes the complexity of the relationships between the cross-layer parameters analyzed. In future work, we will apply objective-driven simulations [28] to obtain more accurate patterns between cross-layer parameters under study.

5.2. Voice Quality Measurements. In this section, we provide the voice quality results associated to the former experiments. VoIP voice quality has been calculated using the objective PESQ algorithm standardized by ITU-T. The PESQ algorithm outcome provides a quality metric mapped to the MOS. The algorithm requires injecting a known speech signal into the system under test, and the degraded output signal is compared with the original (reference) input. Values between 4 and 4.5 represent toll quality, which is the typical quality offered by the PSTN, while values below 3.5 are often considered by some users. In Figure 4(c), we can see that an SNR of 8 dB or higher produces a an average MOS higher than 4, whereas an SNR of 7 dB results in an MOS of below 3.5 for some calls.

Although the mapping of MOS to packet losses has been analyzed in the literature, it is also represented in Figure 4(d) to verify that a linear estimation can be derived. Particularly, it can be verified that for packet losses close to 0%, the maximum quality is reported by PESQ, whose maximum output is 4.5. It must be noted that the PESQ algorithm does not consider factors such as the end to end delay that affect the subjective quality and are considered in other methods such as the E-model.

In Figure 4(e), the mapping of MOS to the PDSCH BLER is represented. Although it will depend on the cell configuration, in future work, we will consider using delay aware quality algorithms such as E-model to review the relation between MOS and PDSCH BLER, as the delay introduced by HARQ retransmissions could have also an impact on quality depending on the configuration and the end to end delay budget.

6. Conclusions

In this paper, we have provided a detailed state of the art of LTE improvements for enhanced VoIP support. It has also been identified that standardization organizations, alliances, and network operators demand more realistic measurements to improve the QoS and QoE of LTE deployments in general and VoIP over LTE in particular. To that end, we have proposed an experimental testing setup based on a real-time implementation of the LTE radio access where it is possible to test LTE protocol settings, commercial devices, and real applications.

We have also shown the results obtained by the provision of relevant reference performance measurements carried out for a specific configuration at different levels of the communication, covering LTE physical level, IP performance and voice quality evaluation. Moreover, the measurements are correlated to check the consistency of the results and extract the relationships between them. For example, the number of configured HARQ retransmissions has impact on the tradeoff between maximum delay and packet losses. These results will also complement the limited nonsimulated results currently available in the scientific literature.

During the experiments, it has been identified a noticeable variability of the results because of the use of realistic impairments such as fading and noise. Real-time test environment where the experiments can be largely repeated, as the proposed one, will definitely contribute to improve the statistical relevance of the results with shorter test times.

For future work we plan to extend the number of fading profiles, increase the parameters logged at the emulated environment, and automate the execution of the tests to increase the number of VoIP calls under study, which will improve statistical validity of the results. With these automation features, we will apply the presented methodology to validate the proposals identified in the summarized state of the art. The testbed will be also extended with a core network which will enable to evaluate the impact of the transport in the evolved packet core (EPC) and study the performance of mobility procedures such as S1- and X2-based handovers.

Acknowledgments

This work has been funded by the Government of Andalusia under Grant ITC-20111046, the Spanish Ministry of Innovation and Science under Grant IPT-2011-1034-370000 and FEDER from the European Commission.

References

[1] C. S. Bontu and E. Illidge, "DRX mechanism for power saving in LTE," *IEEE Communications Magazine*, vol. 47, no. 6, pp. 48–55, 2009.

[2] D. Jiang, H. Wang, E. Malkamaki, and E. Tuomaala, "Principle and performance of semi-persistent scheduling for VoIP in LTE system," in *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM '07)*, pp. 2861–2864, September 2007.

[3] X. Mang, Y. Levy, C. Johnson, and D. Hoeflin, "Admission control for VoIP calls with heterogeneous codecs," in *Proceedings of the 18th Annual Wireless and Optical Communications Conference (WOCC '09)*, pp. 1–4, May 2009.

[4] O. Fresan, T. Chen, K. Ranta-aho, and T. Ristaniemi, "Dynamic packet bundling for VoIP transmission over Rel'7 HSUPA with 10ms TTI length," in *Proceedings of the 4th IEEE International Symposium on Wireless Communication Systems (ISWCS '07)*, pp. 508–512, October 2007.

[5] J. Puttonen, T. Henttonen, N. Kolehmainen, K. Aschan, M. Moisio, and P. Kela, "Voice-over-IP performance in UTRA long term evolution Downlink," in *Proceedings of the IEEE 67th Vehicular Technology Conference (VTC Spring '08)*, pp. 2502–2506, May 2008.

[6] R. Susitaival and M. Meyer, "LTE coverage improvement by TTI bundling," in *Proceedings of the IEEE 69th Vehicular Technology Conference (VTC Spring '09)*, esp, April 2009.

[7] NGMN, "NGMN top OPE recommendations version 1.0.21," Tech. Rep., NGMN Alliance, 2010.

[8] A. Diaz, P. Merino, and F. J. Rivas, "Test environment for QoS testing of VoIP over LTE," in *Proceedings of the IEEE Network Operations and Management Symposium (NOMS '12)*, pp. 780–794, 2012.

[9] 3GPP, "Digital cellular telecommunications system (Phase 2+), Universal Mobile Telecommunications System (UMTS), Circuit Switched (CS) fallback in Evolved Packet System (EPS), stage 2," TS 23.272.

[10] GSMA, "Permanent Reference Document (PRD) IP Multimedia Subsystem (IMS) profile for voice and SMS," GSMA IR92, 2012.

[11] V. Forum, "Voice over LTE via generic access, requirements specification, phase1," TS 1.4.0, 2010.

[12] ETSI, "Digital cellular telecommunications system (Phase 2+), Generic Access Network (GAN), stage 2," TS 43.318, 2012.

[13] M. Forum, "MSF VoLTE interoperability event 2011. Multivendor testing in global LTE and IMS Networks," WhitePaper, 2011.

[14] Epitiro, "LTE "Real World" Performance Study. Broadband and voice over LTE, (VoLTE) quality analysis: TeliaSonera," Tech. Rep., Epitiro, 2011.

[15] 3GPP, "Digital cellular telecommunications system (Phase 2+), Universal Mobile Telecommunications System (UMTS), LTE, Services and service capabilities," TS 22.105.

[16] 3GPP, "Universal Mobile Telecommunications System (UMTS); LTE; IP Multimedia Subsystem (IMS); Multimedia telephony; Media handling and interaction," TS 26.114.

[17] A. T. Harri Holma, *LTE for UMTS: Evolution to LTEAdvanced*, chapter 8, Wiley, 2011.

[18] 3GPP, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification," TS 36.331.

[19] 3GPP, "LTE; General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access," TS 23.401.

[20] 3GPP, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; stage 2," TS 36.300.

[21] 3GPP, "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Policy and charging control architecture," TS 23.203.

[22] A. Díaz, P. Merino Gomez, and F. Rivas Tocado, "Mobile application profiling for connected mobile devices," *IEEE Pervasive Computing*, vol. 9, no. 1, pp. 54–61, 2010.

[23] A. Alvarez, A. Diaz, P. Merino, and F. J. Rivas, "Field measurements of mobile services with android smartphones," in *Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC '12)*, pp. 105–109, 2012.

[24] I. T. U. -T, "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Recommendation P.862.

[25] 3GPP, "Evolved Universal Terrestrial Radio Access (EUTRA); User Equipment (UE) conformance specification; radio transmission and reception; part 1: conformance testing," Tech. Rep. 36.521-1, TS.

[26] O. Heckmann, *The Competitive Internet Service Provider: Network Architecture, Interconnection, Traffic Engineering and Network Design*, Wiley, 2007, Edited by, D. Hutchison.

[27] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, *RTP: A Transport Protocol For Real-Time Applications*, Internet Engineering Task Force, RFC.

[28] A. Díaz, P. Merino, and A. Salmerón, "Obtaining models for realistic mobile network simulations using real traces," *IEEE Communications Letters*, vol. 15, no. 7, pp. 782–784, 2011.

Hindawi Publishing Corporation
Journal of Computer Networks and Communications
Volume 2013, Article ID 412491, 12 pages
http://dx.doi.org/10.1155/2013/412491

Research Article

Dealing with Energy-QoE Trade-Offs in Mobile Video

Fidel Liberal, Ianire Taboada, and Jose-Oscar Fajardo

University of the Basque Country UPV/EHU, Faculty of Engineering of Bilbao, Alameda Urquijo s/n, 48013 Bilbao, Spain

Correspondence should be addressed to Fidel Liberal; fidel.liberal@ehu.es

Received 7 September 2012; Accepted 11 March 2013

Academic Editor: Maria G. Martini

Scalable video coding allows an efficient provision of video services at different quality levels with different energy demands. According to the specific type of service and network scenario, end users and/or operators may decide to choose among different energy versus quality combinations. In order to deal with the resulting trade-off, in this paper we analyze the number of video layers that are worth to be received taking into account the energy constraints. A single-objective optimization is proposed based on dynamically selecting the number of layers, which is able to minimize the energy consumption with the constraint of a minimal quality threshold to be reached. However, this approach cannot reflect the fact that the same increment of energy consumption may result in different increments of visual quality. Thus, a multiobjective optimization is proposed and a utility function is defined in order to weight the energy consumption and the visual quality criteria. Finally, since the optimization solving mechanism is computationally expensive to be implemented in mobile devices, a heuristic algorithm is proposed. This way, significant energy consumption reduction will be achieved while keeping reasonable quality levels.

1. Introduction

The evolution of multimedia encoding techniques allows efficiently provisioning video services at different quality levels. However, resulting streams lead also to different energy consumptions making it difficult to simultaneously satisfy both energy consumption and quality requirements. Therefore, an energy versus quality compromise solution is commonly required. In commercial cellular networks, users are used to dealing with these trade-offs either manually or automatically (i.e., using small widgets to reduce display brightness, disable radio interfaces, etc.) and normally maintaining the same play-out quality. However, reduced energy consumption becomes a truly severe constraint in specific communication scenarios such as mobile emergency networks or distributed sensors. Additionally, any solution will also depend on the characteristics of the video players although higher resolution video could improve visual quality for high-end mobile devices, for others no visible quality improvement is achieved due to available screen resolution, codecs, or CPU power. So, additional energy consumption, higher data bandwidth, and spectrum use would have no real impact on users satisfaction. Energy- and visual quality-aware video dynamic transmission schemes would allow network operators and users to avoid such waste of resources.

In order to cope with the heterogeneity of mobile devices and user requirements for efficient mobile video delivery, a multilayer scheme is broadly considered as the best solution. In this paradigm, each video is encoded into a single stream with multiple layers, where each layer is only transmitted once. Scalable Video Coding (SVC) standard, an extension for H.264/AVC standard, makes this multilayering possible, becoming the most promising encoding technology for solving the problem of multiuser video streaming in most mobile environments (see [1, 2]). This mechanism for content delivery provides quality differentiation, so that the same content is sent simultaneously in different qualities without replicating the original information. This way, users that demand lower energy consumption can maintain the reproduction but accepting lower quality level.

For example, [3–5] have proposed broadcasting schemes that would allow mobile devices to receive and decode the most suitable number of layers, maintaining the perceived video quality proportional to the consumed energy in a DVB-H scenario. Under typical system parameters of mobile TV networks, the proposed schemes allow mobile devices to

achieve energy savings between 60% and 95% depending on how many layers they receive. In [5], not only does the proposed scheme enable each device to achieve energy saving proportional to perceived quality but also low channel switching delays are guaranteed, which is also important to user experience.

However, in these works the expected visual quality level as perceived by the user is not quantified, which is fundamental to find a relation between energy consumption and user Quality of Experience (QoE).

In [6], authors follow a similar approach but in a 802.11e environment by using sleep cycles of wireless adapter for energy saving. In this case, a simple QoE estimation algorithm is used to trigger specific power save protocol operations. However, they only modify the behaviour of the receiver with a single version of the content. Furthermore, no analysis of the optimality of their approach is included.

Such kind of optimization in 802.11 is carried out in [7], including both linear programming techniques and heuristics. Unfortunately, only energy is considered as optimization criteria and no effect into quality is analyzed.

Finally, [8] proposes a method for statically selecting the best number of layers for the whole duration of a video with energy constraints. However, no compromise solution is provided and DVB-H scenarios only are considered.

In order to cover these lacks, in this paper we analyze the optimal strategy to be applied if we can modify dynamically the number of layers during the video reproduction considering energy and quality constraints. Therefore, the main contributions of the paper are as follows. (1) We analyze the energy versus QoE trade-offs considering either energy or quality as constraints. (2) We formulate the aforementioned scenarios in terms of single-objective linear programming (SOLP) problems and analyze the optimal sets in both the input and the objective space. (3) We study the multiobjective problem therefore allowing deployers to fine tune the relative weights of green consciousness and quality criteria. (4) We propose lightweight heuristics for energy- and QoE-aware optimization that ensure a feasible implementation on the end user while providing near-optimal solutions.

Therefore, the rest of this paper is structured as follows. In Section 2, we analyze the dual energy minimizing and QoE maximizing problem while alternating the number of reproduced layers during a video session. Section 3 considers users that have both energy and QoE related criteria, and Section 4 summarizes achieved results.

2. Energy/QoE Single Optimization

In this section, we will express the Energy versus QoE trade-off in terms of a typical optimization problem with different objectives and constraints, analyze the optimal strategy, and compare it with the traditional ones.

Traditional QoE-aware energy-constrained video reproduction strategies select the maximum number of layers that the available battery [8] and/or CPU load [9] would allow for the full video playout in an static way, so that the decision is taken just once. We will generalize and refer to this kind of strategies as basic strategy.

Considering (1) that each additional layer provides different QoE level (see, e.g., [10, 11]) and energy consumption and (2) that SVC players [12] support switching from a layer to another, we will instead propose a dynamic method for triggering layer switching considering energy constraints.

We will therefore focus on selecting the best set of different time periods t_i so that during $t_i(s)i$ layers will be reproduced.

For simplicity purposes, we will consider 4 layers to illustrate the method. For a video of duration $T(s)$ it is clear then that $t_1 + t_2 + t_3 + t_4 = T$. In our optimization problem $\mathbf{t} = [t_1 \;\; t_2 \;\; t_3 \;\; t_4]$ will be the input variables.

In order to define our optimization problem completely, we will consider that the average of the satisfaction over the whole video reproduction $\widehat{\mathrm{QoE}}$ is a good estimator of video quality. Then, we have an optimization problem consisting of the following.

(A2.1) Minimizing energy consumption for a given video and certain minimum acceptable visual quality.

(A2.2) Maximizing user QoE for a given video and certain maximum energy constraint.

Any single-objective optimization problem (SOP) [13] like (A2.1) and (A2.2) aims at choosing the "best" possible combination of input parameters in order to optimize the one considered criterion.

Let $\mathbf{x} \in \mathbb{R}^M$ be a vector of M input variables of the optimization problem.

The SOP can be stated as follows:

$$\begin{aligned} \max \quad & \{f(\mathbf{x}) = z\} \\ \text{s.t.} \quad & \mathbf{x} \in S, \quad z \in \mathbb{R}, \end{aligned} \tag{1}$$

where S is the set of feasible points in the input space delimited by i inequalities and j equalities such as

$$\mathbf{x} \in S \Longleftrightarrow \begin{cases} g(\mathbf{x}) \leq \mathbf{b}, & (b_1, b_2, \ldots, b_i) \\ h(\mathbf{x}) = \mathbf{c}, & (c_1, c_2, \ldots, c_j). \end{cases} \tag{2}$$

Thus, the SOP could be summarized as "finding the set of input variables $\mathbf{x_{opt}}$ belonging to S region so that $f(\mathbf{x_{opt}}) = z_{\mathrm{opt}}$ is max," where S is constrained by inequalities $g(\mathbf{x}) \leq \mathbf{b}$ and equalities $h(\mathbf{x}) = \mathbf{c}$.

If both $f(\mathbf{x})$ is linear and S is defined by linear conditions (therefore both $g(\mathbf{x})$ and $h(\mathbf{x})$ are linear functions as well), then the SOP is called single objective linear programming problem (SOLP) [14].

The basic metric that we will use for estimating the whole user satisfaction regarding video quality will be the average value of the Mean Opinion Score (MOS) according to the number of layers reproduced along each time period. Therefore,

$$\widehat{\mathrm{QoE}} = \frac{1}{T} \sum_{i=1}^{4} \int_{t_i} \mathrm{MOS}_i(t) \cdot dt, \tag{3}$$

where $\mathrm{MOS}_i(t)$ is the evolution of the QoE measured in the MOS scale along the ith period.

If the visual quality is considered roughly constant for a certain bitrate or number of layers and we denote $1 \leq \mathrm{MOS}_i \leq 5$ as the MOS estimated for a certain number of layers according to subjective tests, then $\mathrm{MOS}_i(t) \approx \mathrm{MOS}_i$ for the whole period. Therefore,

$$\mathbf{Q} = [Q_1 \;\; Q_2 \;\; Q_3 \;\; Q_4], \qquad Q_i = \frac{\mathrm{MOS}_i}{T} \tag{4}$$

are the normalized quality coefficients, and (3) can be expressed as follows:

$$\widehat{\mathrm{QoE}} = \mathbf{Q} \cdot \mathbf{t}^{\mathsf{T}}. \tag{5}$$

Similarly, according to [8, 15], the battery consumption in mobile video players shows some kind of dependence on the number of layers received. Therefore, the total energy consumption will depend on the number of layers that will be received so that total consumption P along the whole reproduction time can be expressed as follows:

$$P = \mathbf{B} \cdot \mathbf{t}^{\mathsf{T}}, \tag{6}$$

where $\mathbf{B} = [B_1 \;\; B_2 \;\; B_3 \;\; B_4]$ is the vector of normalized coefficients of battery consumption.

This way, problem (1) can be expressed as a SOLP problem for both (A2.1) and (A2.2).

2.1. Minimizing Energy Consumption Assuring a QoE Threshold (A2.1). In this case, the objective function to be minimized is the energy consumption so that $f(\mathbf{x}) = P$. Minimum energy consumption will be constrained by the QoE threshold, the user will stand (E) in terms of $\widehat{\mathrm{QoE}} \geq E$. This constraint will lead to the associated inequality in (7). Similarly, the sum of the time period for all the layer numbers must be the total duration of the video T leading to an equality constraint expressed with the identity vector $\mathbf{i} = [1 \;\; 1 \;\; 1 \;\; 1]$. Consider the following:

$$\begin{aligned} \max \quad & \{-\mathbf{B} \cdot \mathbf{t}^{\mathsf{T}}\} \\ \text{s.t.} \quad & \begin{bmatrix} [\mathbf{I}] \\ \mathbf{Q} \end{bmatrix} \cdot \mathbf{t}^{\mathsf{T}} \geq \begin{bmatrix} [\mathbf{0}] \\ E \end{bmatrix} \\ & \mathbf{i} \cdot \mathbf{t}^{\mathsf{T}} = T, \end{aligned} \tag{7}$$

where the natural condition $t_i \geq 0$ for all i is included in the matrix notation in (7) using the identity 4×4 matrix $[\mathbf{I}]$.

Then, the SOLP problem in (7) can be solved by typical well-known optimization problem resolution mechanisms such as the simplex method [16]. Therefore, we can easily calculate the solution of the SOLP for a wide range of constraints and $\mathbf{Q}$ parameters. In fact, considering the proposed transmission scheme in [8], we have carried out the optimization for the parameters collected in Table 1 and obtained Figures 1 and 2. Note that the normalized available battery is measured in seconds to avoid specific details about battery characteristics (like voltage and capacity in mAh).

For mobile videos with different motion levels or Content Types (CTs) (referred to as Low Motion (LM), Medium Motion (MM), and High Motion (HM)) and Spatial Resolutions (SRs) we have calculated the QoE-optimal layer selection strategy with the energy constraint. Note that, although we have considered a specific scenario in our study, the method can be applied to any multilayer broadcasting technology just by considering related $\mathbf{Q}$ and MOS_i coefficients.

In order to calculate realistic QoE constraints, we must consider $E_{\min}$ and $E_{\max}$: the minimum and maximum achievable $\widehat{\mathrm{QoE}}$ considering $t_1 = T$ and $t_4 = T$, respectively (i.e., receiving only 1 layer—worst quality—or 4 layers—best quality—during the whole video length). According to the metric used $E_{\min} = \mathrm{MOS}_1$ and $E_{\max} = \mathrm{MOS}_4$. Finally, since

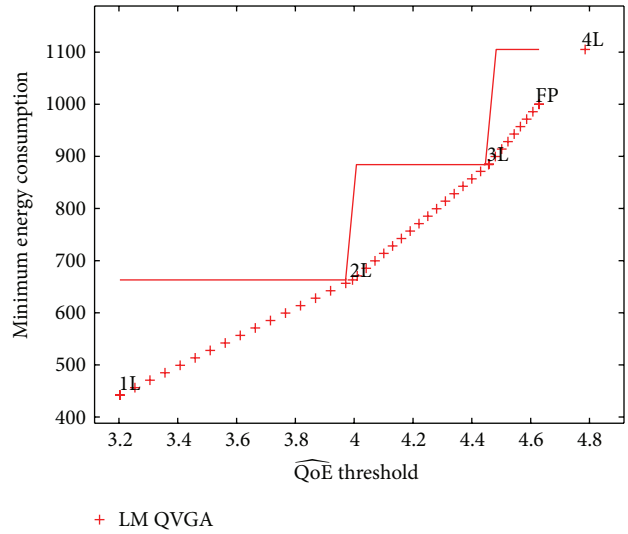


Figure 1: Comparison between basic strategy and energy optimization.

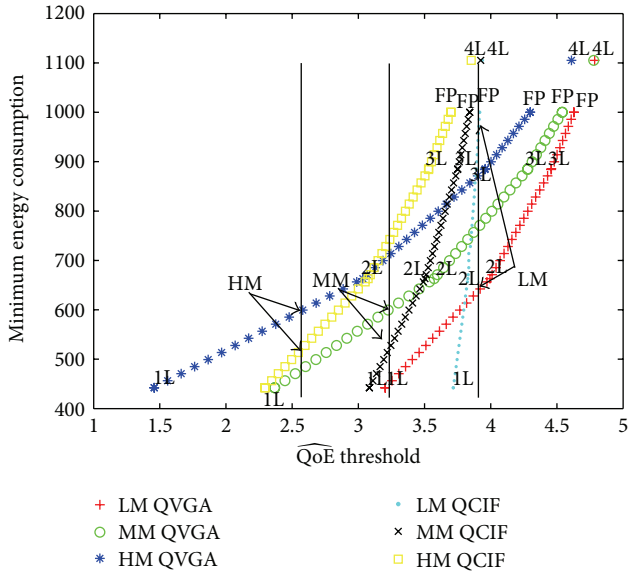


Figure 2: Static versus dynamic energy optimization for different CTs and SRs.

Table 1: Values for coefficients for battery minimizing SOLP obtained from [8].

Concept			Value
Video length T (s)			7200
QoE threshold E			$E_{\min} \le E \le E_{\max}$
Available battery (s)			1000
	Content Type	SR	Values per layer
MOS_i	LM	QVGA	$[3.20 \ 3.99 \ 4.45 \ 4.78]$
	MM	QVGA	$[2.36 \ 3.57 \ 4.27 \ 4.77]$
	HM	QVGA	$[1.45 \ 3.03 \ 3.95 \ 4.60]$
	LM	QCIF	$[3.71 \ 3.82 \ 3.89 \ 3.93]$
	MM	QCIF	$[3.08 \ 3.50 \ 3.74 \ 3.92]$
	HM	QCIF	$[2.29 \ 3.07 \ 3.52 \ 3.85]$
B			$K \cdot [2 \ 3 \ 4 \ 5]$ where the constant K depends on the transmission scheme only regardless of the CT and resulting spatial resolution

the additional constraint of the total energy consumption must be less than the available battery (if a handheld device) or the energy budget, not all points will be feasible.

In Figure 1, the objective space (namely, minimum battery consumption versus QoE threshold) and the finite decision points associated to the traditional selection of a fixed number of layers for the whole video are shown for a certain CT/SR combination. Considered remaining battery constraint results in a Feasible Point (FP) that represents the maximum achievable QoE level. Both this FP and discrete no. L points are shown (the points related to reproducing a certain number of layers during the whole play-out time, T).

We can see how, by changing the number of layers reproduced along the video duration, we can optimize the energy consumption with finer grain QoE constraints. This way, for $MOS_1 \le \widehat{QoE} < MOS_2$ traditional static strategies would result in selecting 2 layers (2L) for the whole duration of the video and consuming associated energy. Our dynamic approach would instead allow the selection of different (energy, QoE) points. On the other hand, if we compare in detail both approaches for this concrete CT and SR (LM and QVGA), obtained optimum strategy does not match the simplest one in 2L point. This clearly reflects the energy saving achieved as for the same MOS less energy is consumed.

Additional results are depicted in Figure 2 for every combination of video types and encoding resolutions considered. Examining in detail the figure we can see how optimization strategy does not match the traditional static approaches (fixed no. L points). Furthermore, vertical lines show the result of ensuring certain QoE levels for different CTs. For example, for selected $\widehat{QoE}$ thresholds, the QCIF versions are better in terms of energy consumption for HM and MM videos, while QVGA is better for the threshold considered in the LM case.

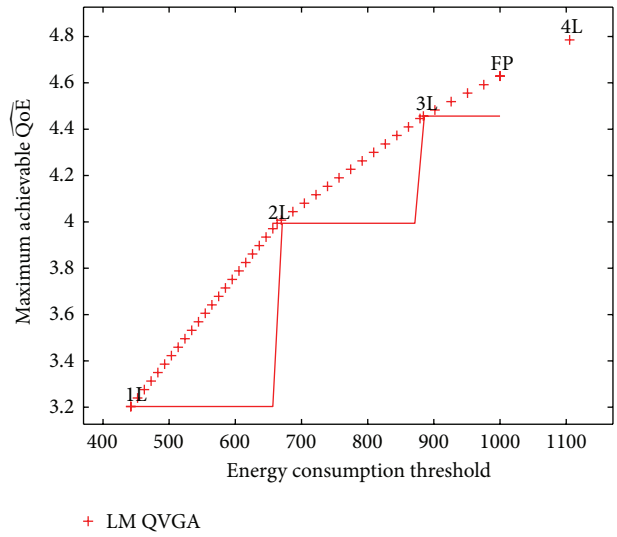


Figure 3: Comparison between basic strategy and QoE optimization.

2.2. Maximizing User QoE Constrained by Maximum Energy Consumption (A2.2). For (A2.2), in an analogous way to Section 2.1, the objective function to be maximized is the MOS so that $f = \widehat{QoE}$. Maximum achievable QoE will be constrained by the maximum battery consumption C. This constraint will lead to the associated inequality in the following:

$$\begin{aligned} \max \quad & \{\mathbf{Q} \cdot \mathbf{t}^{\intercal}\} \\ \text{s.t.} \quad & \begin{bmatrix} [\mathbf{I}] \\ -\mathbf{B} \end{bmatrix} \cdot \mathbf{t}^{\intercal} \ge \begin{bmatrix} [\mathbf{0}] \\ -C \end{bmatrix} \\ & \mathbf{I} \cdot \mathbf{t}^{\intercal} = T. \end{aligned} \tag{8}$$

Figure 3 depicts the evolution in the objective space of the proposed optimal strategy in comparison with the basic one considering #L points only. Once more, we can see how the proposed scheme allows us to set a continuous range of energy constraints resulting in different values of $\widehat{QoE}$. In this case, an available energy budget in the $[E_1, E_2)$ range would collapse into the same 1L point, therefore allowing only minimum $\widehat{QoE}$.

When we compare the obtained results for different CT/SR combinations (in Figure 4), we can see, again, how the QoE maximizing quality selection mechanism would choose QCIF (for the particular energy saving constraints marked with a vertical line in the figure) for LM and MM videos, while QVGA for the HM one.

2.3. Comparison between SOLP Approaches. In Sections 2.1 and 2.2, we have used the simplex optimization method for SOLPs in order to optimize either energy for a given QoE

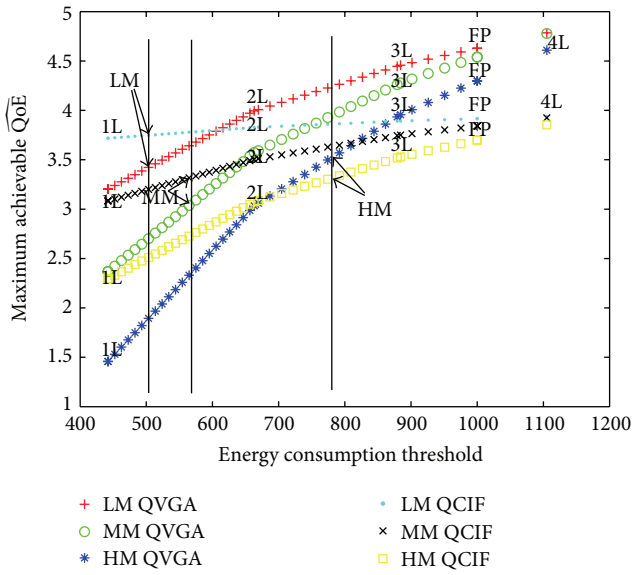


Figure 4: Static versus dynamic QoE optimization for different CTs and SRs.

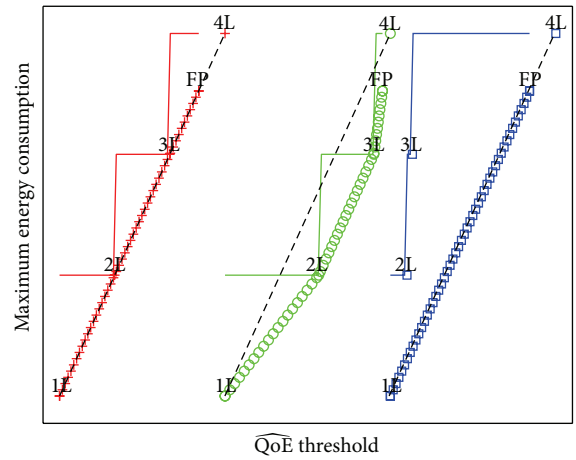


Figure 5: Pareto fronts for 3 different MOS_i and **B**.

threshold or QoE for a given maximum battery consumption. Due to the nature of the linear optimization problem, if we compare both optimization approaches, we can see in Figure 6 that both lead to the same shape in the objective space. This conclusion is consistent with the nature of the dual problem itself, since energy and QoE are opposite objectives, and the constraint in one problem becomes the objective function in the other.

Since carrying out the simplex method for the optimization of every video reproduction would be inviable for handheld devices due to the high CPU power needed, we focus on developing heuristics capable of providing near-optimal solutions. This way, both video operators and end users themselves would be able to reduce their energy consumption while maintaining QoE levels. In order to do so, we have analyzed more deeply the shape of the optimal set (namely, Pareto front in the objective space) for three different situations and compared it with the #L points. We can see (Figure 5) how the optimal strategy for energy versus QoE, depending on MOS_i and **B** evolution:

(1) follows exactly the line between 1L and 4L (including 2L and 3L), constrained by the FP;

(2) follows the polygon (the 2D polytope in the objective space) 1L-2L-3L-4L constrained by the FP;

(3) follows the 1L and 4L but without going through 2L and 3L points.

We conclude that, regardless of the specific parameters of the SOLP problem to be solved, the optimal strategy always includes 1L and 4L points (or associated FP point if 4L is not feasible due to the battery constraint). This result is quite evident since reproducing just 1 or all the 4 layers gives the minimum and maximum QoE and battery. However, both 2L ($[0 \;\; T \;\; 0 \;\; 0]$) and 3L ($[0 \;\; 0 \;\; T \;\; 0]$) points are not always optimal points (i.e., they do not belong to the Pareto front). As a result, the simple strategy of following the 1L-2L-3L-4L path must be carefully reviewed.

Therefore, in order to propose an optimization strategy, we must evaluate the evolution of the Pareto front in both the objective and the input space for every possible situation (1, 2, and 3).

In Figures 6 and 7, we can see such evolution for the objective and input space in the aforementioned Situations 1 and 2. Figure 6 confirms the superposition of both (A2.1) and (A2.2) Pareto fronts in the objective space for both situations. In Figure 7, we depict the 4D input space with the points that belong to the Pareto front in both situations. In order to do so, (x, y, z) space coordinates correspond to t_1, t_2, and t_3, respectively. The fourth dimension (t_4) is represented by the area of the sphere in the (t_1, t_2, t_3) point. Finally, the result of the objective function is represented by the color of the sphere according to the colormap shown. $(T, 0, 0, 0)$, $(0, T, 0, 0)$, $(0, 0, T, 0)$, and $(0, 0, 0, T)$ points are also depicted with spheres.

When comparing Situations 1 and 2 in the input space, we can see how, for the former, the optimal path does not follow any simple strategy. However, the latter follows a linear path in a set of consecutive planes, which leads to easy-to-implement heuristics.

After comparing the figures for MOS_i for every considered video CTs and SRs and the evolution of the energy consumption versus the number of layers, we conclude that all of them follow the Situation 2. In fact, if we analyze the sufficient conditions leading to Situation 2 it is clear that, if the function $\max \text{MOS} = f(\min \text{Energy})$ in the objective space is convex, the resulting shape would belong to this group. Most QoE studies aiming at mapping satisfaction versus network performance parameters use logarithmic expressions (see [17]), so that they are convex. In our case, due to this convexity of the MOS expression and the proportional $K \cdot (i + 1)$ dependency of the energy with $\mathbf{t}(i)$ for all i, an increment achieved by the simplex algorithm in terms of

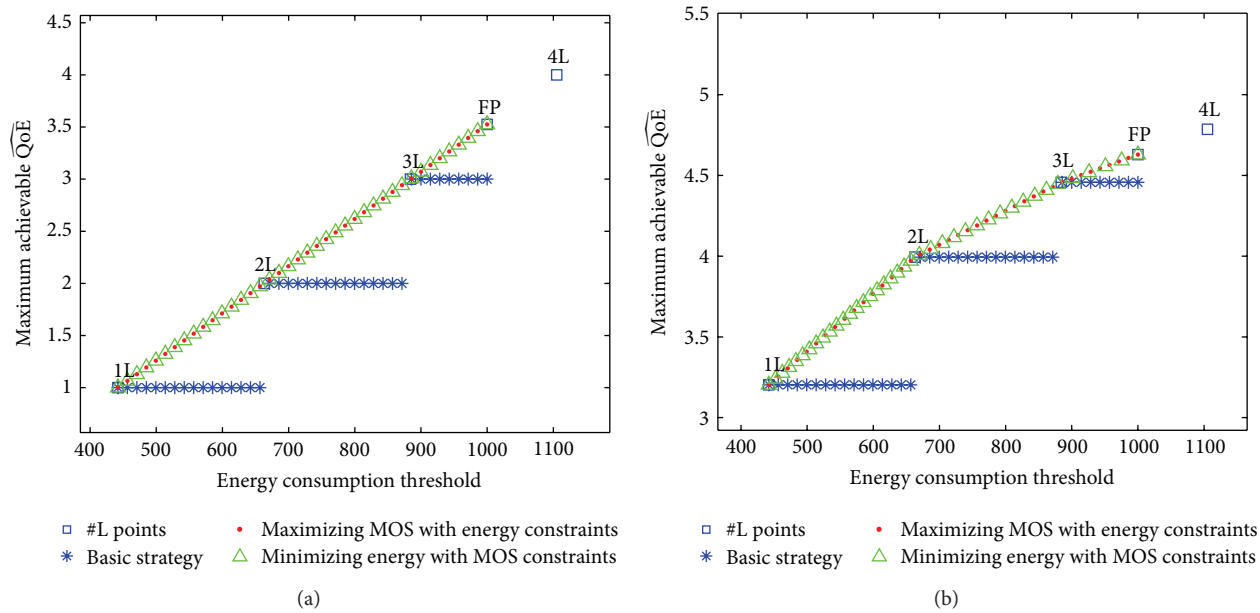


Figure 6: Situations 1 and 2 in the objective space.

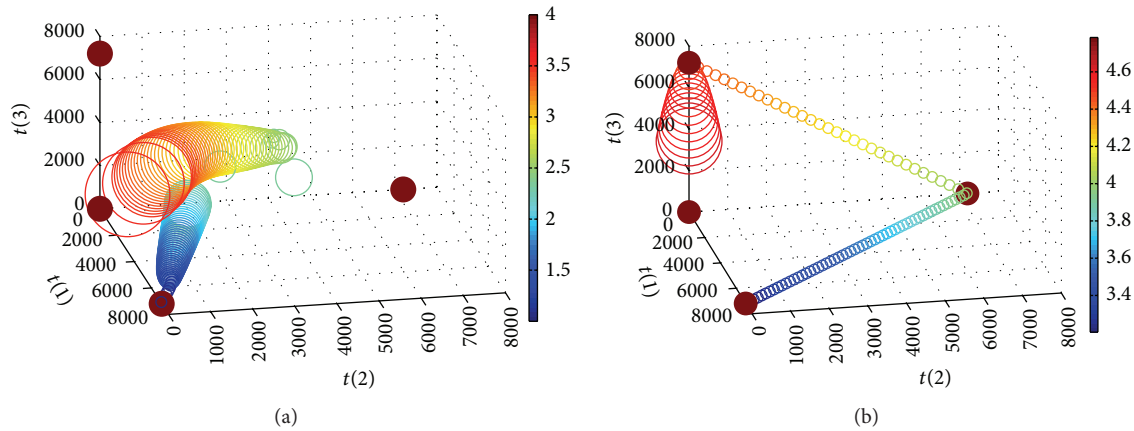


Figure 7: Situations 1 and 2 in the input space.

$\text{MOS}' = \text{MOS} + \Delta\text{MOS}$ would be caused by a displacement vector $\mathbf{t}' = \mathbf{t} + \Delta\mathbf{t}$ leading to a (P', MOS') point above the 1L–4L line in the objective space. An equivalent conclusion will be obtained if we calculate the gradient of both MOS and energy, or if we consider that many optimization methods indeed provide the convex part of the Pareto front.

Therefore, in this case the associated optimization heuristic is straightforward, regardless either optimization approach 1 (minimize energy for a certain QoE) or 2 (maximize QoE for a certain energy budget) is applied. Algorithm 1 shows the complete optimization heuristic procedure. The algorithm is based on the fact that the optimal polytope will follow the 1L-2L-3L-4L path (constrained by the FP). Additionally, we take into account that, due to the linear constraints, the polytope will be formed by the intersection of planes leading to lines in the different 2D planes formed by the successive input parameters. So, since $\mathbf{t} = [t_1 \ t_2 \ t_3 \ t_4]$ is the 4D input space, the optimization heuristic will consider the polytope delimited by 1L, 2L, 3L, and 4L/FP points and planes $[t_1 \ t_2 \ 0 \ 0]$, $[0 \ t_2 \ t_3 \ 0]$, and $[0 \ 0 \ t_3 \ t_4]$. Note here that along the different figures in this paper, we have set a maximum energy constraint leading to an FP between 3L and 4L for illustration purposes. In a real scenario, the FP could be anywhere in the 1L–4L Pareto Front.

Regardless our aim is minimizing energy for a certain QoE or maximizing QoE with a certain energy budget, we will calculate the 2D plane where the target point in the input space will be located at. Since both energy and QoE are monotonically growing with $\mathbf{t}(i)$ for all i either the QoE or the battery constraint will allow us to select the plane by evaluating the $\text{BAT}_i - \text{BAT}_{i+1}$ or $\text{QoE}_i - \text{QoE}_{i+1}$ intervals where $i = 1 \cdots \text{FP}$. Later, the linear dependence will allow us to express $\mathbf{t}_{\text{opt}}(i)$ as a function of provided constraints (see lines 18–21 and 28–30).

In Figure 8, the results obtained with the heuristics for both Situations 1 and 2 in the objective space are depicted.

(1) $T \leftarrow$ Video Length {Initialize total time to play the whole video}
(2) $\mathbf{B} \leftarrow$ obtain B from Transmission Scheme() {Obtain battery related coefficients}
(3) $\mathbf{Q} \leftarrow$ obtain Q from Subjective Tests (CT, SR) {Obtain QoE related coefficients}
(4) $E_{\max} \leftarrow$ maximum Battery Consumption
(5) **for** $i = 1 \rightarrow 4$ **do**
(6) $\quad \mathbf{t}_i \leftarrow [0 \;\; 0 \;\; 0 \;\; 0]$
(7) $\quad \mathbf{t}_i(i) = L$
(8) $\quad \mathrm{QoE}_i \leftarrow \mathbf{Q} \cdot \mathbf{t}_i^{\mathrm{T}}$ {Obtain QoE for #L point}
(9) $\quad \mathrm{BAT}_i \leftarrow \mathbf{B} \cdot \mathbf{t}_i^{\mathrm{T}}$ {Obtain Battery consumption for #L point}
(10) **end for**
(11) $\mathrm{BAT}_{\mathrm{FP}} \leftarrow E_{\max}$ {Maximum battery sets the FP}
(12) $\mathrm{QoE}_{\mathrm{FP}} \leftarrow$ calculate FP($\mathbf{t_1}, \mathbf{t_2}, \mathbf{t_3}, \mathbf{t_4}, \mathrm{BAT}_{\mathrm{FP}}$) {Obtain QoE for FP}
(13) **If** minimize Energy **then** {minimize energy for a given QoE constraint}
(14) $\quad E \leftarrow$ minimum Acceptable QoE
(15) $\quad i \leftarrow$ find Plane($E, \mathrm{QoE}_i, \mathrm{BAT}_i, \mathrm{BAT}_{\mathrm{FP}}, \mathrm{QoE}_{\mathrm{FP}}$)

$$(16) \quad a \leftarrow \frac{\mathrm{BAT}_{i+1} - \mathrm{BAT}_i}{\mathrm{QoE}_{i+1} - \mathrm{QoE}_i}$$

$$(17) \quad b \leftarrow \frac{\mathrm{BAT}_{i+1} - \mathrm{BAT}_i \cdot \mathrm{QoE}_{i+1}/\mathrm{QoE}_i}{1 - \mathrm{QoE}_{i+1}/\mathrm{QoE}_i}$$

(18) $\quad \mathrm{BAT}_{\mathrm{opt}} \leftarrow a \cdot E + b$
(19) $\quad \mathbf{t}_{\mathrm{opt}} \leftarrow [0 \;\; 0 \;\; 0 \;\; 0]$

$$(20) \quad \mathbf{t}_{\mathrm{opt}}(i) \leftarrow \frac{E - T \cdot \mathbf{Q}(i+1)}{\mathbf{Q}(i) - \mathbf{Q}(i+1)}$$

(21) $\quad \mathbf{t}_{\mathrm{opt}}(i+1) \leftarrow T - \mathbf{t}_{\mathrm{opt}}(i)$
(22) **else** {maximize QoE for a given energy constraint}
(23) $\quad C \leftarrow$ maximum Energy Consumption
(24) $\quad i \leftarrow$ find Plane($C, \mathrm{QoE}_i, \mathrm{BAT}_i, \mathrm{BAT}_{\mathrm{FP}}, \mathrm{QoE}_{\mathrm{FP}}$)

$$(25) \quad a \leftarrow \frac{\mathrm{BAT}_{i+1} - \mathrm{BAT}_i}{\mathrm{QoE}_{i+1} - \mathrm{QoE}_i}$$

$$(26) \quad b \leftarrow \frac{\mathrm{BAT}_{i+1} - \mathrm{BAT}_i \cdot \mathrm{QoE}_{i+1}/\mathrm{QoE}_i}{1 - \mathrm{QoE}_{i+1}/\mathrm{QoE}_i}$$

$$(27) \quad \mathrm{QoE}_{\mathrm{opt}} \leftarrow \frac{C - b}{a}$$

(28) $\quad \mathbf{t}_{\mathrm{opt}} \leftarrow [0 \;\; 0 \;\; 0 \;\; 0]$

$$(29) \quad \mathbf{t}_{\mathrm{opt}}(i) \leftarrow \frac{C - T \cdot \mathbf{B}(i+1)}{\mathbf{B}(i) - \mathbf{B}(i+1)}$$

(30) $\quad \mathbf{t}_{\mathrm{opt}}(i+1) \leftarrow T - \mathbf{t}_{\mathrm{opt}}(i)$
(31) **end if**

Algorithm 1: Heuristic for both SOLP optimization problems.

Note that, even for Situation 1 the heuristic provides values very close to the Pareto set. The reason is that 2L, 3L and the points in the 1L–4L path belong to the Pareto front. At the same time, when approaching these points from $[t_1 \;\; t_2^{-} \;\; t_3 \;\; t_4]$ and $[t_1 \;\; t_2^{+} \;\; t_3 \;\; t_4]$, there exist different combinations of $\mathbf{t}(i)$ that result in the equivalent point in the objective space. Therefore, depending on the path followed by the simplex algorithm and the stopping thresholds, the simulation would provide different points in the input space, but the heuristic is still capable of providing equivalent optimal points in the objective space.

Figure 9 shows the effect of applying developed algorithm for Situation 3. As already mentioned, although each case should be evaluated in terms of MOS_i and $\mathbf{B}$ in order to estimate associated situation, the convex shape of the MOS function makes this situation highly improbable. Anyway, the equivalent heuristic is again quite simple, since the Pareto front follows the 1L–4L path.

3. Hybrid Approach: Optimizing Energy Consumption and QoE

In the previous section, we have focused on analysing the problem of optimizing one single objective (i.e., either energy or QoE) in our mobile video scenario. Therefore, we have

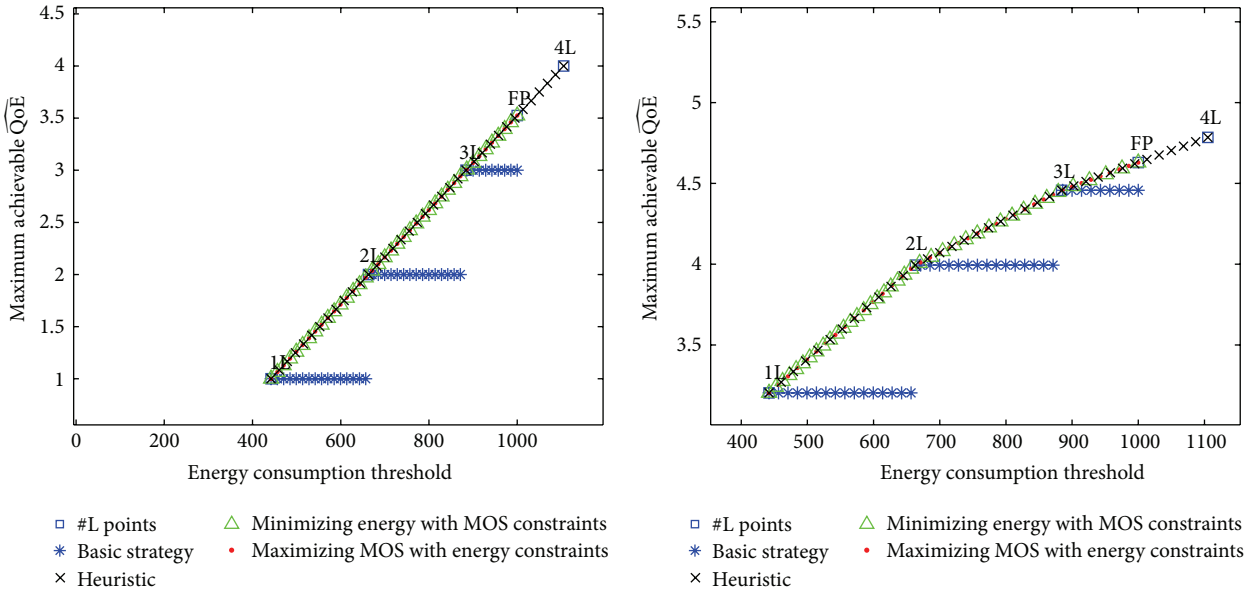


Figure 8: Situations 1 and 2 in the objective space including heuristics.

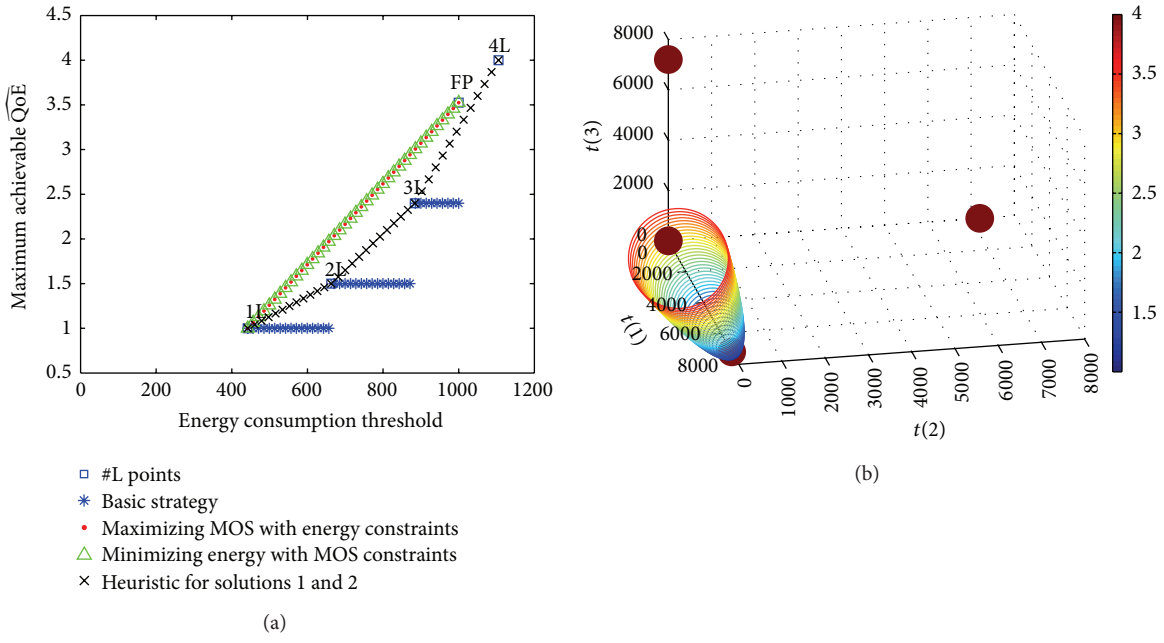


Figure 9: Situation 3 in the objective and input space including original heuristics.

expressed both optimization problems in terms of SOP. The considered $\widehat{\text{QoE}}$ metric allowed us to reduce such problems to SOLP ones and describe them with the matrix notation in (7) and (8).

However, real users do not usually consider a single criterion while evaluating a product or a service [18]. Generally speaking, most users will not care about energy consumption or quality in an isolated way but will take into account both criteria. On one hand, commercial users would probably prefer assuring higher quality for playing movies. In other scenarios, such as aforementioned emergency networks, they would instead put the emphasis on preserving battery. However, minimum image quality levels should be also provided to keep quality of information so that, for example, first responders would be still capable of evaluating the risks of an emergency. Therefore, the two original SOPs merge into

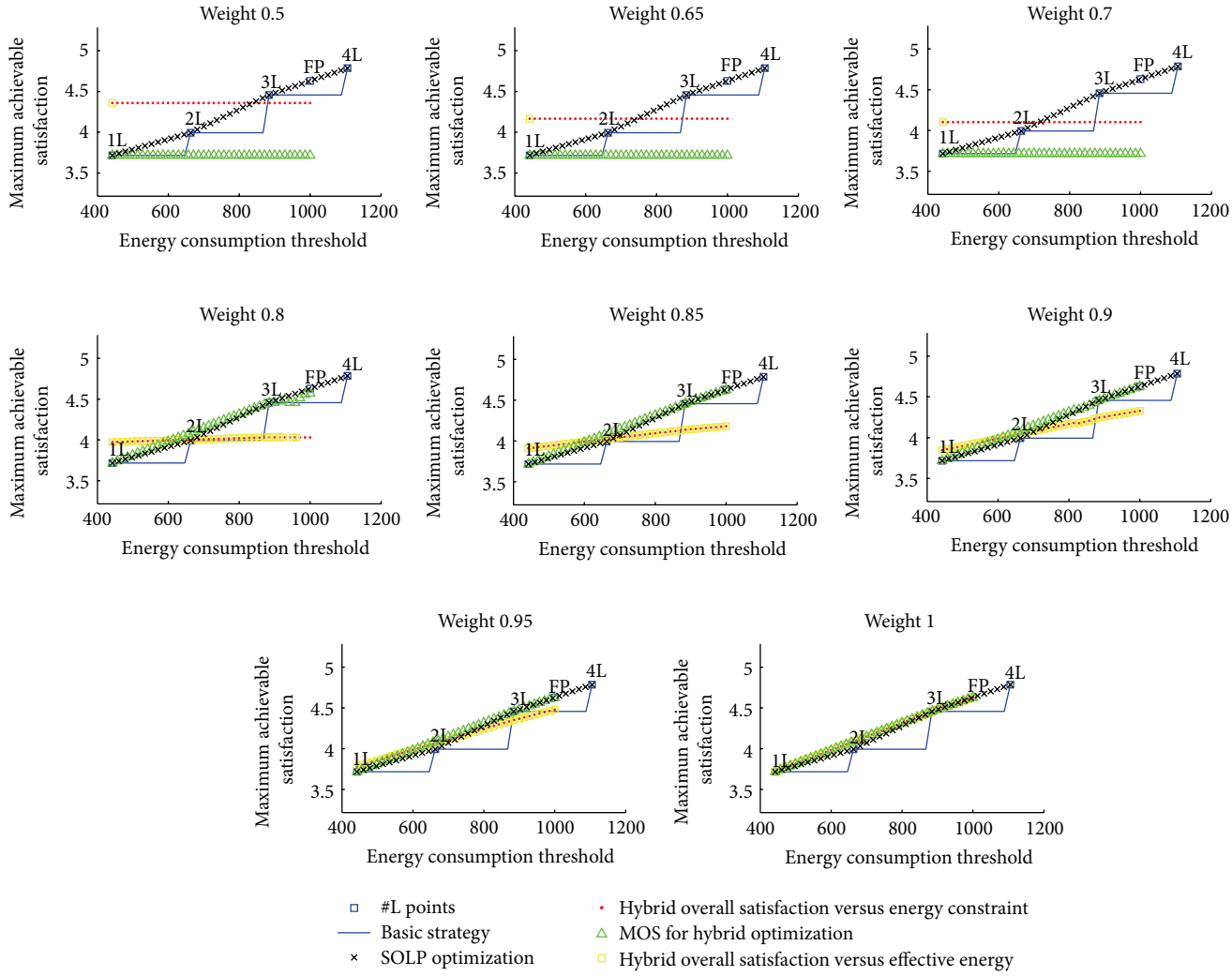


Figure 10: Evolution of optimization strategies with $0.5 \leq w \leq 1$ in the objective space.

a more complex Multiple Objective Optimization Problem (MOOP) including two objectives: minimizing energy consumption (i.e., $f_1(\mathbf{t})$) and maximizing QoE (i.e., $f_2(\mathbf{t})$).

Then, the MOOP is an extension of the SOP, which can be defined as follows:

$$\begin{aligned} \max \quad & \{f_1(\mathbf{x}) = z_1\} \\ \max \quad & \{f_2(\mathbf{x}) = z_2\} \\ & \vdots \\ \max \quad & \{f_k(\mathbf{x}) = z_k\} \\ \text{s.t.} \quad & \mathbf{x} \in S, \end{aligned} \tag{9}$$

where f_i is the ith criterion function.

The simplest solution for the MOOP problem consists of finding the input vector $\mathbf{x_{opt}}$ so that

$$\exists \mathbf{x_{opt}} \in S \mid \max \left\{ f_i\left(\mathbf{x_{opt}}\right) = z_{i_{\text{opt}}} \right\} \quad \forall i = 1, 2, \ldots, k. \tag{10}$$

In most of the cases, there will not exist such $\mathbf{x_{opt}}$ which maximizes all the criteria simultaneously. So, we will have to redefine the nature of the problem by introducing the concept of utility function, U. Then, the real formulation of the MOOP can be expressed mathematically as follows:

$$\max \{U(z_1, z_2, \ldots, z_k)\}. \tag{11}$$

Then, any MOOP requires the definition of a utility function that collects users' preferences regarding different considered criteria as in (11). Many authors have considered the linear composition of preferences with different weights to express the articulation of preferences in communications systems with methods such as the Analytic Hierarchy Process (AHP; see, e.g., [19]) to infer the weight of each criterion out of user surveys. If we consider that our utility function follows this linear approach:

$$U(\mathbf{x}) = w_1 \cdot f_1(\mathbf{x}) + w_2 \cdot f_2(\mathbf{x}), \tag{12}$$

where, if we normalize f_1 and f_2 between the same value ranges (i.e., $1 \leq f_i \leq 5$ as in the MOS scale), associated

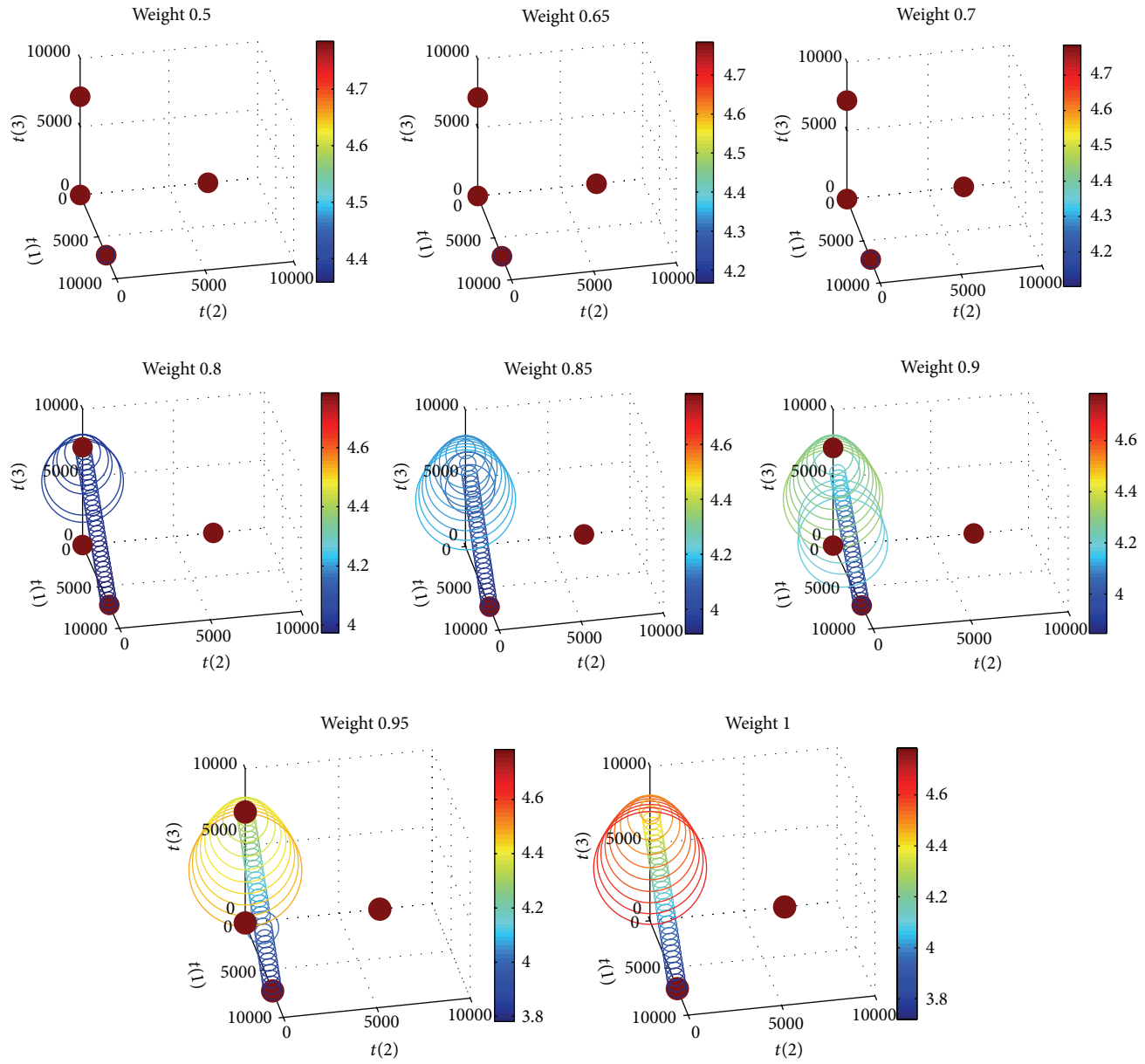


Figure 11: Evolution of optimization strategies with $0.5 \leq w \leq 1$ in the input space.

weights will follow $w_1 + w_2 = 1$ so that we could express the utility function in a linear way as in the following:

$$U(\mathbf{x}) = w \cdot f_1(\mathbf{x}) + (1 - w) \cdot f_2'(\mathbf{x}). \quad (13)$$

Since both f_1 and f_2' (normalized f_2) are related to QoE and battery criteria, respectively, if we consider similar expressions as those in (7) and (8) for the objective functions, we could expand expression (13) as follows:

$$U(\mathbf{t}) = w \cdot \sum_{i=1}^{4} \mathbf{Q}(i) \cdot \mathbf{t}(i) + (1 - w) \cdot \sum_{i=1}^{4} \mathbf{B}'(i) \cdot \mathbf{t}(i), \quad (14)$$

where $\mathbf{B}'$ is the vector of normalized battery parameters so that $1 \leq f_2' \leq 5$. Note that, contrary to $\mathbf{Q}$, $\mathbf{B}'$ coefficients will decrease with $i\,\mathbf{B}'(i) > \mathbf{B}'(i+1)$ for all i since users satisfaction regarding battery will decrease when the number of layers and therefore the battery consumption grows. If we manipulate this expression for the utility function, we can express the complex MOOP into a simplified SOLP similar to (7) and (8) as follows:

$$\begin{aligned} \max \quad & \{\mathbf{M} \cdot \mathbf{t}^{\mathsf{T}}\} \\ \text{s.t.} \quad & \begin{bmatrix} [\mathbf{I}] \\ -\mathbf{B} \\ \mathbf{Q} \end{bmatrix} \cdot \mathbf{t}^{\mathsf{T}} \geq \begin{bmatrix} [\mathbf{0}] \\ -C \\ E \end{bmatrix} \\ & \mathbf{I} \cdot \mathbf{t}^{\mathsf{T}} = T, \end{aligned} \quad (15)$$

where $\mathbf{M}(i) = w \cdot \mathbf{Q} + (1 - w) \cdot \mathbf{B}'$. In order to calculate $\mathbf{B}'$, we carry out a mapping between the actual battery consumption and associated satisfaction, where $\mathbf{B}'(1)$ will be associated with the maximum satisfaction and $\mathbf{B}'(4)$ with the lowest one (maximum consumption).

Once $\mathbf{M}$ is calculated and maximum battery and minimum QoE are set, we can use simplex again in order to optimize users' satisfaction considering both battery and QoE. Figure 10 depicts the evolution of the optimal strategy in the objective space for different values of w, the relative weight of both criteria. Both hybrid optimization (with proposed utility function) together with maximizing QoE only policy, static assignment and proposed heuristics are shown. The different values for w represent how important is the "Green characteristic" of the device for the user and/or video provider when compared with video quality (with $w = 1$ completely important and $w = 0$ not important at all).

We have computed only the range $0.5 \leq w \leq 1$, since the same figures were obtained for lower values of w. We can see how, for all $w < 0.7$, the optimal shape is restricted to a single point (corresponding to 1L). The reason is that, considering the low importance of QoE, the optimization algorithm considers that it is always more convenient to restrict the energy consumption rather than receive more layers and obtain better video quality. The higher the w value, the closer the Pareto fronts gets to our original single-objective optimization.

In Figure 11, we carry out the equivalent analysis in the input space. Once again, the Pareto front for considered $\mathbf{Q}$ and $\mathbf{B}'$ leads to a simplified optimization strategy for $\mathbf{t}$. In this case, however, the evolution is from plane $[t_1 \;\; 0 \;\; t_3 \;\; 0]$ to plane $[0 \;\; 0 \;\; t_3 \;\; t_4]$.

In order to get the exact $\mathbf{t}_{\text{opt}}$ point for every E and C constraint pair, we carry out the same equations as in lines 18–21 and 28–30 of Algorithm 1.

Therefore, in order to solve the MOLP problem, we could use any LP optimization technique (i.e., simplex or Interior Points Method) or analyze the problem in a case per case basis in order to infer whether a simple per-plane heuristic could be applied. In any case, the complexity of the solution is low. For example, although simplex's complexity analysis is rather a problem dependent leading even to worst-case exponential, due to the small size of the input matrix, results are obtained in less than 20 iterations, and empirical tests have led to less than 30 ms of CPU consumption.

4. Conclusions

In this paper, we analyze the trade-off between energy consumption and visual quality for mobile video systems.

Different optimization approaches have been evaluated. The simplest static strategy comprises receiving the highest number of video layers while coping with the video duration requirements. Thus, taking as inputs the video length and the amount of remaining battery, we always select the best possible visual quality. The main drawback of this approach is that all the battery is available to be used in the video playout.

In Section 2, we introduce a single-objective optimization problem as a way to provide an automated decision making process to the mobile device. Two approaches have been defined and solved by linear programming: energy consumption minimization constrained to a minimal QoE threshold, and QoE maximization constrained to a maximum level of battery consumption. Contrary to the previous case, the decision maker may provide a noninteger number of layers, providing a finer grain resolution for quality optimization. Yet although we are able to introduce additional energy constraints to the automated decision making, the desired remaining battery level must be a priori computed without further information of the achievable QoE level.

Additionally, once the energy or quality constraints are assured, the single-objective optimization will lead to the feasible point of maximum quality or minimum energy consumption. Therefore, this approach does not explore the intermediate points as possible optimum solutions, where we can make use of the different relations between increased energy consumption and enhanced visual quality.

In order to overcome this drawback, we analyze in Section 3 the problem from a multiobjective optimization standpoint. Both energy and quality are considered as objective functions by means of a weighted utility function, which allows us to solve the problem as a single-objective linear programming problem. Different weights have been evaluated, which entail different priority to energy saving or required quality. These weights could be used to define different user profiles, different device energy saving modes, or dynamically adapted based on the status of the device battery.

Since the implementation of the optimization algorithm in a mobile handset may result on a resource-consuming process, we propose the use of a heuristic algorithm. From the analysis of the evolution of both the objective function and input variable spaces, different alternatives are found for the shape of the Pareto front. However, considering the logarithmic shape of the evolution of most MOS-related utility functions, a simple heuristic has been proposed. As a result, the proposed algorithm can be run on a mobile device as a decision making process to trigger the switching between layers.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under Grant agreement 284863 (FP7 SEC GERYON) and from the Basque Country under BASAJAUN Project.

References

[1] X. Ji, J. Huang, M. Chiang, G. Lafruit, and F. Catthoor, "Scheduling and resource allocation for SVC streaming over OFDM downlink systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 10, pp. 1549–1555, 2009.

[2] M. Brandas, M. Martini, M. Uitto, and J. Vehkapera, "Quality assessment and error concealment for svc transmission over unreliable channels," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '11)*, pp. 1–6, 2011.

[3] C. Hsu, M. Hefeeda, and Video broadcasting to heterogeneous mobile devices, in *Proceedings of the 8th International IFIP-TC 6 Networking Conference (NETWORKING '09)*, pp. 600–613, 2009.

[4] C.-H. Hsu and M. Hefeeda, "Flexible broadcasting of scalable video streams to heterogeneous mobile devices," *IEEE Transactions on Mobile Computing*, vol. 10, no. 3, pp. 406–418, 2011.

[5] C. Hsu and M. Hefeeda, "Multi-layer video broadcasting with low channel switching delays," in *Proceedings of the 17th International Packet Video Workshop (PV '09)*, May 2009.

[6] M. Csernai and A. Gulyas, "Wireless adapter sleep scheduling based on video qoe: how to improve battery life when watching streaming video?" in *Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN '11)*, pp. 1–16.

[7] J. Lorincz, M. Bogarelli, A. Capone, and D. Begušić, "Heuristic approach for optimized energy savings in wireless access networks," in *Proceedings of the 18th International Conference on Software, Telecommunications and Computer Networks (SoftCOM '10)*, pp. 60–65, September 2010.

[8] I. Taboada, J. Fajardo, and F. Liberal, "Qoe and energy-awareness for multilayer video broadcasting," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '11)*, pp. 1–6.

[9] A. M. Alt and D. Simon, "Control strategies for H.264 video decoding under resources constraints," in *Proceedings of the 5th International Workshop on Feedback Control Implementation and Design in Computing Systems and Networks (FeBiD '10)*, pp. 13–18, ACM, April 2010.

[10] D. WeiSong and S. Azad, "User-centered video quality assessment for scalable video coding of H. 264/AVC standard," in *Proceedings of the 16th International Multimedia Modeling Conference (MMM '10)*, Advances in Multimedia Modeling, p. 55, Springer, Chongqing, China, January 2010.

[11] D. Migliorini, E. Mingozzi, and C. Vallati, "QoE-oriented performance evaluation of video streaming over WiMAX," in *Proceedings of the 8th International Conference on Wired/Wireless Internet Communications (WWIC '10)*, pp. 240–251, 2010.

[12] M. Blestel and M. Raulet, "Open SVC decoder: a flexible SVC library," in *Proceedings of the 18th ACM International Conference on Multimedia ACM Multimedia (MM '10)*, pp. 1463–1466, October 2010.

[13] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, 1999.

[14] G. Dantzig, *Linear Programming and Extensions*, Princeton University Press, 1998.

[15] F. Pescador, E. Juarez, D. Samper, C. Sanz, and M. Raulet, "A test bench for distortion-energy optimization of a DSP-based H.264/SVC decoder," in *Proceedings of the 13th Euromicro Conference on Digital System Design: Architectures, Methods and Tools (DSD '10)*, pp. 123–129, September 2010.

[16] R. Bartels and G. Golub, "The simplex method of linear programming using LU decomposition," *Communications of the ACM*, vol. 12, pp. 266–268, 1969.

[17] R. M. Salles and J. A. Barria, "Fair and efficient dynamic bandwidth allocation for multi-application networks," *Computer Networks*, vol. 49, no. 6, pp. 856–877, 2005.

[18] B. Erman and E. P. Matthews, "Analysis and realization of IPTV service quality," *Bell Labs Technical Journal*, vol. 12, no. 4, pp. 195–212, 2008.

[19] A. M. S. Alkahtani, M. E. Woodward, and K. Al-Begain, "Prioritised best effort routing with four quality of service metrics applying the concept of the analytic hierarchy process," *Computers and Operations Research*, vol. 33, no. 3, pp. 559–580, 2006.

Hindawi Publishing Corporation
Journal of Computer Networks and Communications
Volume 2013, Article ID 176217, 13 pages
http://dx.doi.org/10.1155/2013/176217

Research Article

Angry Apps: The Impact of Network Timer Selection on Power Consumption, Signalling Load, and Web QoE

Christian Schwartz, Tobias Hoßfeld, Frank Lehrieder, and Phuoc Tran-Gia

Institute of Computer Science, University of Würzburg, Chair of Communication Networks, Am Hubland, 97074 Würzburg, Germany

Correspondence should be addressed to Tobias Hoßfeld; hossfeld@informatik.uni-wuerzburg.de

Received 2 November 2012; Revised 23 January 2013; Accepted 13 February 2013

Academic Editor: Pedro Merino

The popularity of smartphones and mobile applications has experienced a considerable growth during the recent years, and this growth is expected to continue in the future. Since smartphones have only very limited energy resources, battery efficiency is one of the determining factors for a good user experience. Therefore, some smartphones tear down connections to the mobile network soon after a completed data transmission to reduce the power consumption of their transmission unit. However, frequent connection reestablishments caused by apps which send or receive small amounts of data often lead to a heavy signalling load within the mobile network. One of the major contributions of this paper is the investigation of the resulting tradeoff between energy consumption at the smartphone and the generated signalling traffic in the mobile network. We explain that this tradeoff can be controlled by the connection release timeout and study the impact of this parameter for a number of popular apps that cover a wide range of traffic characteristics in terms of bandwidth requirements and resulting signalling traffic. Finally, we study the impact of the timer settings on Quality of Experience (QoE) for web traffic. This is an important aspect since connection establishments not only lead to signalling traffic but also increase the load time of web pages.

1. Introduction

Together with the wide-spread usage of smartphones in today's UMTS networks, the popularity of smartphone apps has seen a tremendous growth during the last years [1]. The resulting traffic is expected to exceed half of the global mobile data traffic in the next years [2]. One of the major reasons for this phenomenon is that smartphones are very convenient for users to stay always connected to the Internet. In turn, this has led developers of smartphone apps to the assumption of continuous Internet connectivity. Therefore, many apps such as social network clients, weather forecasts, or instant messengers update their status frequently, which raises a number of problems—in the mobile network as well as on the smartphones.

In contrast to desktops or laptops, smartphones are equipped only with limited battery. Since established connections from the smartphone to the mobile network consume a large amount of energy, some smartphones close these connections soon after the data transmission is finished, that is, after a very short period of no traffic activity, which is controlled by an inactivity timer. This saves energy and prevents battery drain caused by established but unused connections. However, it might also degrade the user experience since the connection start-up delay is in the order of a few seconds. Therefore, users might get annoyed, for example, if the load time of every web page in their browser is increased. Hence, two oppositional effects impact the quality of experience (QoE) perceived by the user: lower energy consumption and short page load times—a tradeoff that we investigate thoroughly in this paper.

From a mobile operator's point of view, each change between connected and disconnected states of the smartphone causes signalling in the network, and excessive signalling load has led to severe problems in the recent past. For example, a large mobile network in Japan suffered from an outage of several hours in January 2012. The operator of this network mentioned a large number of small control

messages of certain popular apps (such as keep-alive messages or buddy list updates in VoIP apps) as a probable cause for this outage (Penn-Olson, "Finding a connection: Android, Line, and Docomo's network outage," 2012, available at http://www.penn-olson.com/2012/01/30/docomo-outage-line/). Therefore, we argue that not an overload in data traffic but the high number of establishments and teardowns of wireless connections has brought down the network. As a consequence, the network operator may adjust the timer setting accordingly. A longer timer setting can reduce signalling load and resource costs of the mobile operator, however, at the cost of energy consumption of the user device and an increased consumption of radio resources. Hence, an additional tradeoff exists between battery efficiency on smartphones and signalling load in the mobile network. The proprietary fast dormancy mode as implemented by smartphones (i.e., smartphones tear down the connection earlier than advised by the network) additionally affects the signalling load in the network and is also investigated in this paper. The influence of traffic generated by applications on radio resources is another point of interest of mobile operators which is already well investigated in the literature, for example [5], and out of scope in this paper. In light of current troubles with network outages induced by signalling storms, the interest of mobile operators and equipment vendors has shifted towards signalling generated by the application traffic [6, 7] on which we focus here.

A reason for this signalling storm is the following. UMTS networks are designed to provide wireless, high bandwidth Internet access for mobile users, for example, video telephony. Therefore, high load in terms of data traffic was carefully considered during the design of such networks. However, some currently popular apps such as Aupeo radio streaming or Skype VoIP apps load the mobile network all the time, even if they are only running in the background. They send update and keep-alive messages every few seconds, which is problematic for today's UMTS networks. The reason is that the mobile network usually tears down wireless connections to a user equipment (UE) after an inactivity timeout of a few seconds, that is, when no data was transmitted for this short time. If the time between two such keep-alive messages is slightly above the connection timeout of the mobile network, the wireless connections between the UE and the mobile network are established and torn down every few seconds. This is an issue that UMTS networks have not been designed for.

The contributions of this paper are the following. First, we study the tradeoff between energy consumption at the smartphone and the generated signalling traffic in the mobile network. From the user's point of view, battery efficiency is one of the determining factors for a good QoE due to a very limited energy capacity of smartphones [8]. Thereby, we analyse the impact of the inactivity timer as well as of the fast dormancy mode for exemplary smartphone apps based on measurements in a public 3G network. These apps are selected to cover a wide range of traffic characteristics in terms of bandwidth requirements and resulting signalling traffic. In particular, we consider background applications (like Twitter or Skype in passive mode), bandwidth intensive (like Aupeo radio streaming), and interactive applications (like Angry Birds). Then, we answer the question whether a mobile operator is able to find optimal values for the inactivity timer for a given application or network configuration. Further, we see which apps (and which traffic characteristics) make the smartphone users and the mobile operators angry, respectively. We compare the results from these different applications and study the impact of choosing one such optimum for one specific application on other applications. This comparison sheds new light on the practice of using network parameters to optimise for power consumption or generated amount of signalling frequency. Furthermore, we discuss new paradigms for the design of mechanisms which try to resolve such conflicts, namely, economic traffic management [9] and design for tussle [10, 11]. Finally, we study the impact of the timer settings on QoE for web browsing based on existing models for web browsing [12]. This is an important aspect since connection establishments do not only lead to signalling traffic but also increase the load time of web pages which is the key influence factor on web QoE.

The remainder of this paper is structured as follows. Section 2 provides the background on UMTS networks and the radio resource control protocol used for connection establishment and teardown. Related work on measurement studies of relevant RRC parameters like the inactivity timer is reviewed. Further, existing optimisation approaches of the resource consumption are revisited. Then, energy consumption as key QoE influence factor of smartphone users is considered. Section 3 describes the measurement setup and the algorithm to infer (connectivity and energy) state transitions of the smartphone from measured IP packets. Then, we calculate signalling frequency and power consumption based on the state transitions. Section 4 presents the numerical results of our analysis. The traffic of the four popular smartphone apps are characterized. Afterwards, we compare the signalling frequency and power consumption depending on the network configuration. Finally, we study the influence of network parameters on QoE for web browsing based on page load times. Section 5 concludes this work with an outlook on open challenges of how to optimise mobile networks without annoying users or operators.

2. Background and Related Work

This section provides background information on the structure of UMTS networks and their components. In addition, it explains the radio resource control (RRC) protocol, which is used for the allocation of wireless transmission resources. Afterwards, it reviews related work on measurement studies of RRC parameters in real networks and on optimisation approaches of such resources, both in the energy and the wireless domain. Finally, we discuss the impact of energy consumption on mobile devices such as smartphones on the QoE of the end user.

2.1. UMTS Networks and the RRC Protocol. UMTS networks consist of the UEs, the radio access network (RAN), and the core network (CN). Their basic structure for packet-switched

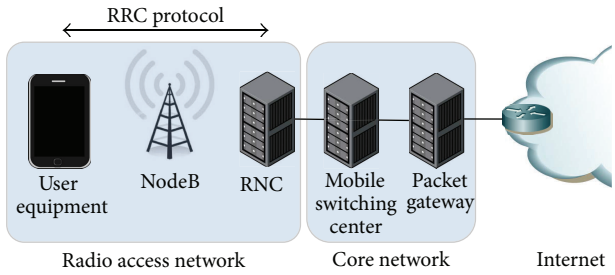


Figure 1: Basic structure of a UMTS network for data calls.

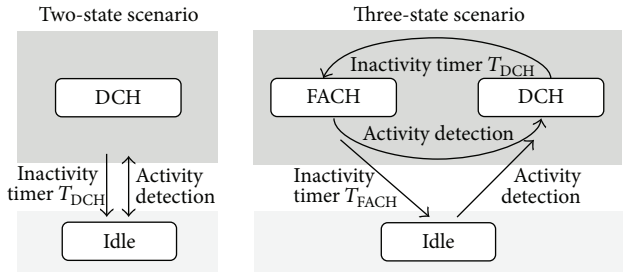


Figure 2: RRC state transition diagrams for the two-state and three-state models.

connections is illustrated in Figure 1. UEs are usually cell phones or computers connected using datacards. The RAN connects the UEs to the CN, which is connected to the Internet. The RAN contains the NodeBs in which the UEs are connected to using the air interface. The NodeBs in turn are connected to radio network controllers (RNCs), which among other tasks are responsible for radio resource control using the RRC protocol [3].

For resource management reasons, a UE may be in one of several RRC states, where each state allows the UE a different level of connectivity to the RAN. The RRC protocol specifies five states for the connection between the UE and the RAN: Idle, URA-PCH, CELL-PCH forward access channel (FACH), and dedicated transport channel (DCH).

In the Idle state, the cell is notified of the UE presence, but the UE can neither send nor receive data. The FACH and DCH states allow communication with the RAN, where the DCH state allows for larger bandwidth at the cost of a higher power consumption. For simplicity reasons, we neglect URA-PCH and CELL-PCH in this study. While URA-PCH plays only a role in scenarios of high mobility, CELL-PCH is not yet widely implemented. Our results are still of general nature and do not depend on the limited number of considered RRC states.

State transitions are controlled by the RNC using rules specified in the RRC protocol and timer values set by the network operator [5, 13]. A brief description is given in the following. We consider two different state transition models, depicted in Figure 2. The first model includes the Idle, FACH, and DCH states (cf. the right part of the figure). Therefore, we call it a three-state model. If the UE is in the Idle state and activity is detected (i.e., a packet is sent or received), the connection transits to the DCH state. After each transmission, a timer T_{DCH} is started, and it is reset whenever a new packet is sent or received. If the timer expires, the connection transits to the FACH state. Upon entering, the T_{FACH} timer is started. If a new transmission occurs, the connection again transits to the DCH state. If T_{FACH} expires, the connection transits to the Idle state.

The second model, denoted as the two-state model, only includes the Idle and DCH states. If the UE is in the Idle mode and a packet is sent or received, the connection transits to the DCH state. Once in the DCH mode, the T_{DCH} timer is started, and it is reset whenever a new packet is sent or received. If the timer expires, the UE transits back to the Idle state.

While the three-state model is closer to the specified RRC protocol, the two-state model is similar to some proprietary fast dormancy implementations used by UE vendors. In these fast dormancy implementations, the UE tears down the connection to the network state as soon as no data is ready to be sent for a certain time, that is, it forces the network to transit to the Idle state. In contrast to the three-state model, there is no transition to the FACH state. If a device disconnects from the network by transitioning to the Idle state, it has to be reauthenticated before another transition to the DCH state occurs. This results in additional signalling traffic and causes more load on the network [6] due to frequent reestablishments of the RRC connection. These proprietary fast dormancy algorithms do not adhere to the RRC specification [14] but, nonetheless, exist in the real world and have been identified as possible causes for signalling storms. The major reason for fast dormancy implementations is the decrease in power consumption on the UE, since the transmission unit of the UE consumes only 1%–2% of the energy in the Idle state compared to the DCH state. Thus, both models warrant further investigation.

2.2. Related Work on Measurements of RRC Parameters and Optimisation of Resource Consumption. The increased use of mobile broadband networks has caused radio resource management to become a hot research topic. In particular, a considerable research effort is spent on measurements of RRC configurations observed in real UMTS networks. The authors of [13] present a measurement tool exactly for that purpose. They use round-trip times of data packets to infer the RRC state of the connection. This is possible since the latency in the FACH is significantly higher than in the DCH state. This allows the authors to measure RRC state transition parameters such as T_{DCH} and T_{FACH}, channel setup delays, and paging delays in different networks. They find that the settings vary by network and give values of 1.2 seconds for the DCH release timer and values of more than one minute for the Idle timer. To validate their method of inferring the RRC states from the round-trip times, they compare the actual energy consumption of the device with the expected energy consumption in a specific RRC.

A similar approach is applied by the authors of [5]. They reach comparable conclusions but report timer values of 5 seconds for T_{DCH} and 12 seconds for T_{FACH}. Furthermore, the authors identify sets of RRC state transitions from two network providers. We used these results to define the state models in Section 2.1.

In [15], the authors propose the tail optimisation protocol. The main idea is that the applications on the UE can accurately predict whether traffic will soon be transmitted or not. This knowledge permits to avoid the unused tail of DCH periods if no further data has to be sent. If traffic activity is expected within a short time frame, the UE stays in the DCH state to avoid frequent connection reestablishments and the associated signalling load. The same authors extend this idea in [4] and propose ARO, an application resource optimiser, together with an implementation for Android 2.2. This tool also includes additional features such as batching up data or increasing the update rate of application to optimise their resource consumptions.

Finally, the 3GPP has released a technical report [16] about the adverse impact of mobile data applications. This report states that frequent connection reestablishments due to small data packets caused, for example, by status updates of social network or instant messaging apps, can lead to problems of increased signalling load. This highlights the importance of this topic.

2.3. Smartphone Energy Consumption and Quality of Experience. In [8], the authors performed a 4-week long study with 29 participants to identify factors influencing QoE of mobile applications. The study comprises (1) data from context sensing software, (2) user feedback using an experience sampling method several times per day, and (3) weekly interviews of the participants. To determine the factors of influence, the authors analyse the frequency of specific keywords in the interviews and the surveys. It turns out that the term *battery* has the highest frequency. According to [8], this is reasonable since the battery efficiency has a strong impact on the user perceived quality, in particular, when it is nearly discharged.

As a consequence of this finding, we investigate the energy consumption of a smartphone as one of the main indicators for QoE. In addition, we show that the energy consumption of the transmission unit of a smartphone can be reduced if the state of wireless connection between the UE and the RNC is set to Idle shortly after data transmission. However, this can lead to frequent connection reestablishments since some apps send or receive small status updates very often. The signalling load produced by such issues is one of the major problem in today's UMTS networks. Hence, this is clearly a tradeoff between battery efficiency on the side of the end user and the network load for the network operators.

There are different approaches to cope with such tradeoffs in the design of specific mechanisms. The two most prominent ones are economic traffic management (ETM) [9] and design for tussle [10, 11]. The ETM paradigm suggests that the different stakeholders collaborate and exchange information so as to permit a joint optimisations of the tradeoff. This can achieve better results than separate optimisation since these tend to work in opposite directions. Design for tussle focuses more on conflicting interests than on cooperation. It means that mechanisms should be able to adapt the outcome of a certain *tussle* at run time and not at design time, for example, by giving the sole control of essential entities to a certain stakeholder.

3. Inferring Signalling Frequency and Power Consumption from Network Trace

However, RRC state transitions are triggered by the UE's firmware. While solutions exist to capture RRC state transitions on specific hardware [17], they are not available for all modern smartphone platforms. Other options to measure the required information include using costly hardware and specific UEs, usually not available to researchers and developers. This prevents the application developers from evaluating the effect their applications have on the overall health on the network. Consequently, they cannot take measures to prevent the harmful behaviour of their applications. However, it is possible to infer the RRC state transitions for a given packet trace if the network model is known.

In this section, we first describe the setup used to capture network packet traces for arbitrary apps. Then, we give an algorithm to infer the RRC state transitions for a given packet trace. Based on these state transitions, we can calculate the number of signalling messages generated by the packet trace. Finally, we use the information of the RRC state of the UE at every given point in time to calculate the power consumption of the UE's radio interface.

3.1. Measurement Procedure and Setup. To investigate the behaviour of the application under study, we capture traffic during a typical use of the application on a smartphone. The smartphone runs the Android operating system and is connected to the 3G network of a major German network operator. To obtain the network packet traces, we use the *tcpdump* application. This application requires *root* privileges which are obtained by rooting the device and installing the custom *cyanogenMod* ROM (http://www.cyanogenmod.org/). Once *tcpdump* is installed and running, we start the application under study and capture packet traces while the application is running. Then, the *android debugging bridge* is used to copy the traces to a workstation. The traces contain Internet Protocol (IP) packets as well as Linux Cooked Captures. We only require the IP packets; thus, we filtered the traces for IP packets which are used during the following analysis.

3.2. Inferring Network State. In this section, we study the influence of the application traffic on RRC state transitions and signalling messages. Since RRC state transitions cannot be captured using commonly available tools, we introduce an algorithm to infer RRC state transitions from IP packet traces. Using this algorithm, we analyse the RRC state transition frequency and signalling message load for the two-state model and the three-state model.

Traffic below the network layer cannot be measured without specific equipment which is often out of reach for developers interested in assessing the impact of their applications on the network. Based on the two-state and the three-state models introduced in Section 2.1, we process *tcpdump* captures of the application traffic. However, it should be noted that this method is not restricted to a specific network model but can be extended to any other network model as well. Using these captures, we extract the timestamps when IP packets are sent or received. Furthermore,

Input: Packet arrival timestamps ts
DCH to FACH timer T_{DCH}
FACH to Idle timer T_{FACH}
Output: Times of state transition *state_time*
New states after state transitions state
`interarrival(i)` ← ts($i+1$) − ts(i)
`index` ← 0
for all ts(i) **do**
  if `state(index)` = Idle **then**
    `index` ← `index` + 1
    `state(index)` ← DCH
    `state_time (index)` ← ts(i)
  end if
  if `interarrival`($i-1$)> T_{DCH} **then**
    `index` ← `index` + 1
    `state (index)` ← FACH
    `state_time (index)` ← ts(i) +T_{DCH}
  end if
  if `interarrival`($i-1$)> $T_{\text{DCH}} + T_{\text{FACH}}$ **then**
    `index` ← `index` + 1
    `state(index)` ← Idle
    `state_time (index)` ← ts(i) + $T_{\text{DCH}} + T_{\text{FACH}}$
  end if
end for

Algorithm 1: Inferring RRC state transitions based on IP timestamps.

we require the timer values of the transition from the DCH state to the FACH state, T_{DCH}, and the timer for the transition between the FACH and the Idle states, T_{FACH}. Based on this information, Algorithm 1 infers the timestamps of state transitions according to the 3GPP specification [3] for the three-state model. This algorithm can be simplified to also work for the two-state model. Alternatively, a way to postprocess the results of the algorithm to obtain results for the two-state model is given at the end of this section. The algorithm first computes the interarrival times for all packets. Then, each timestamp is considered. If the UE is currently in the Idle state, a state transition to DCH occurs at the moment the packet is sent or received. If the interarrival time exceeds the T_{DCH} timer, the UE transits to the FACH, T_{DCH}, seconds after the packet was sent or received. Similarly, if the interarrival time exceeds both the T_{DCH} and T_{FACH} timers, a state transition to Idle occurs, T_{FACH}, seconds after the state transits to the FACH.

UE vendors always search for ways to decrease energy consumption of their devices. A straightforward way to achieve this, if only the wellbeing of the UE is considered, is to transit from the DCH to Idle states as soon as no additional data is ready for sending. While this transition is not directly available in the 3GPP specification for the RRC protocol [3], a UE may reset the connection, effectively transitioning from any state to Idle. This behaviour can be modelled using the two-state model introduced in Section 2.1.

State transitions for the two-state model can be calculated using a similar algorithm. Alternatively, the behaviour of the two-state model can be emulated using Algorithm 1 if T_{FACH} is set to 0 seconds and all state transitions to FACH are removed in a postprocessing step.

Table 1: Number of signalling messages per RRC state transition perceived at the RNC (taken from [3]).

From/to	Idle	FACH	DCH
Idle	—	28	32
FACH	22	—	6
DCH	25	5	—

Table 2: Power consumption of the UE radio interface depending on current RRC state (taken from [4]).

RRC state	Power consumption (mW)
Idle	0
FACH	650
DCH	800

3.3. Calculating Signalling Frequency and Power Consumption. In reality, the number of state transitions is not the metric of most importance if network load should be evaluated. Each state transition results in a number of RRC messages between the UE and different network components. For this study, we consider, the number of messages perceived at the RNC, which can be found in [3] and is summarized in Table 1. It can be seen that transitions from or to the Idle state are especially expensive in terms of number of messages sent or received. This is due to the fact that upon entering or leaving, the Idle state authentication has to be performed. Note that for the two-state model, only transitions from or to the Idle state occur. This results in the fact that for the same network packet trace, the number of signalling messages occurring in the two-state model is generally higher than in the three-state model. To obtain the total number of signalling messages, we weight the number of state transitions with the number of messages sent per state transitions. Then, we average the number of state transitions over the measurement duration to obtain a metric for the signalling load at the RNC. The inference algorithm does not differentiate between state changes caused by upstream or downstream traffic. State changes caused by downstream traffic usually generate some additional signalling messages, as paging is involved. The inference algorithm can be easily enhanced to support this behaviour. However, the results discussed in the next section would only change quantitatively. Furthermore, the inference of signalling messages can be easily adapted to new networking models or signalling numbers.

From users' point of view, the signalling message frequency is not important. The user is interested in a low power consumption because this increases the battery time of the device. To calculate the battery time, we use the time when state transitions occurred, and the new RRC state of the UE to calculate the relative amount of time that was spent in each state. Given the relative time spent in each state, we use Table 2 (taken from [4]) to compute the power consumption of the radio interface during the measurement phase. To obtain the energy consumption, the power consumption can

Table 3: Qualitative characterization of applications under study.

Application	Traffic characteristic	Application use	Required bandwidth
Angry Birds	Interactive	Foreground	Low bandwidth
Aupeo	Interactive	Background	High bandwidth
Twitter	Periodic, low frequency	Background	Low bandwidth
Skype	Periodic, high frequency	Background	Low bandwidth

be multiplied with the duration of the measured network packet trace. We only focus on the power consumption of the radio interface, as it is possible to measure the aggregated power consumption using out-of-the-box instrumentation techniques provided by the hardware vendor.

4. Numerical Results of Measurement Study

In the measurement study, we apply the methods introduced in Section 3 to four popular smartphone applications to infer signalling traffic and energy consumption. In Section 4.1, we characterize the applications in terms of traffic patterns, application usage, and bandwidth requirements. In Section 4.2, we study the signalling frequency and power consumption caused by these applications, if inactivity timers such as T_{DCH} or T_{FACH} are modified. Finally, we analyse in Section 4.3 the influence of network parameters on web QoE in terms of mean opinion score (MOS) depending on page load times which are influenced by the network settings.

4.1. Characterization of Traffic Patterns for Selected Applications. For this study, we chose four specific applications in order to cover a broad spectrum of traffic characteristics, as described in Table 3. First, we discuss said characteristics for these applications. We differentiate between applications, where the user interaction causes the generation of traffic, and such where the application periodically sends or receives traffic. Finally, we consider the amount of bandwidth used by the application.

Angry Birds for Android is a popular *interactive* free-to-play game and runs in the *foreground*. To finance the game, an advertisement is shown once the player starts or restarts a level. Advertisements are downloaded on demand by the application but require *low bandwidth*. Thus, the time between two advertisements depends on the frequency of the player advancing to the next level or deciding to restart the current one.

Aupeo is an Internet radio application, allowing a user to listen to content from personalised radio stations while running in the *background*. Content is not streamed but downloaded at the beginning of the track. The exact duration depends on the radio stations chosen by the user and is thus *interactive*. This results in large times of inactivity during the playback of the track itself. Due to the fact that audio files are downloaded, there is a *high bandwidth* requirement.

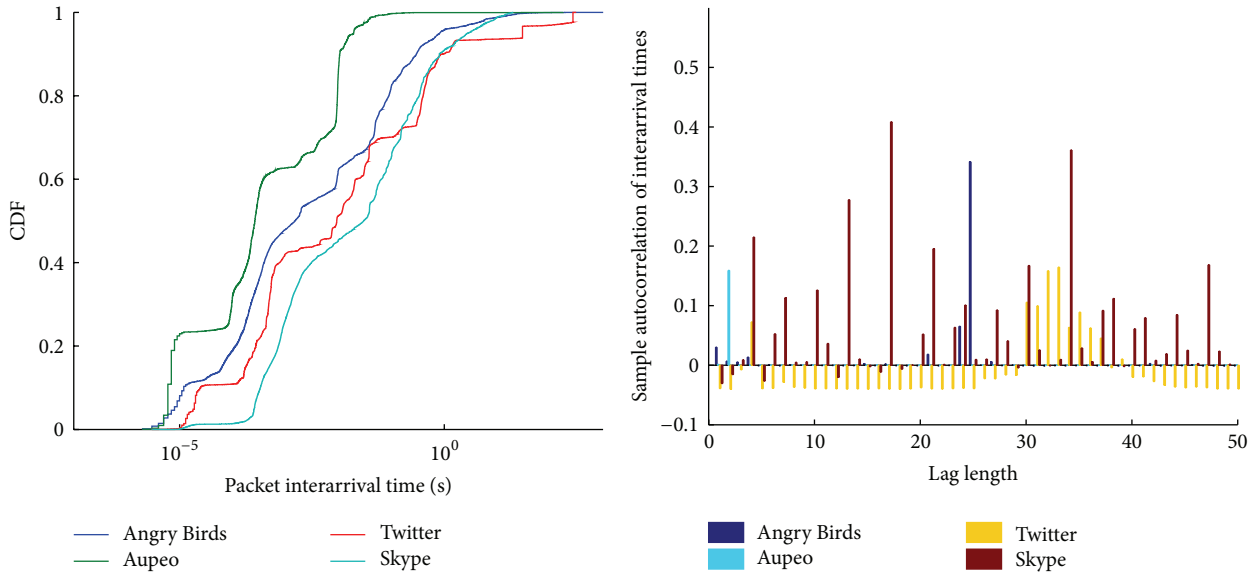
The Twitter client is used to send and receive new short messages from the user's Twitter account. Transferring these messages requires relatively *low bandwidth*. To this end, the user can specify an update frequency when to pull new messages in the *background*. Thus, the downloads occur with a *periodic behaviour of low frequency*, where the client sends an HTTPS request to the Twitter server and in return receives new Tweets for the user's account. We do not consider an active user who is publishing new Tweets. Such behaviour would manifest as additional traffic to the periodic one generated by the status updates. Due to the fact that publishing updates occurs relatively infrequently and updating the feed occurs more often, the traffic generated by publishing updates is dominated by that occurring due to updates and thus can be neglected.

Finally, we consider the Skype application. We do not consider any Voice over IP (VoIP) calls but the application's Idle behaviour, that is, when the application is running in the *background*. During this time, the application sends keep-alive messages to the network. These keep-alive messages are sent with *high frequency* and require *low bandwidth*.

In addition to the applications considered, there exist other categories of applications which are running in the *foreground* and *interactively* require a *high bandwidth*. One example for such an application is Skype while taking a VoIP call. These applications are not considered in this study because this kind of behaviour causes the UE to be always online. This results the minimal amount of signalling messages to be sent and a maximal power consumption at the UE, independent of network model, or used parameters. Other combinations of traffic criteria also exist. However, from both a signalling load as well as a power consumption point of view, they can be mapped to one of the discussed cases. For example, if an application is sending periodic updates with low bandwidth without user interaction, then the fact that the application is running in the foreground or the background is without consequence for the generated signalling load or power consumption. However, these cases should be considered when the optimisation strategies for message sending are under study. For example, background applications could allow for the batching of messages because the transmission is usually not urgent, while foreground applications do not allow for such behaviour because it would decrease QoE.

Next, we describe the applications under study in more detail. For each application, we show the cumulative distribution function (CDF) of the interarrival times in Figure 3(a) and give information about the mean values and standard deviation of both interarrival times and bandwidth in Table 4, respectively.

Let us again begin with the Angry Birds application. We see that there are no distinct peaks in interarrival time, which would hint at periodic behaviour. Furthermore, we see that 5% of all interarrival times are greater than 1 second. As we consider only T_{DCH} values above 1 second, those are candidates for triggering state transitions. The mean interarrival time is 0.66 seconds, with a relatively high standard deviation



(a) CDF of interarrival times

(b) Autocorrelation of interarrival times

Figure 3: Characteristics of the four different smartphone applications in our measurement study.

Table 4: Mean and standard deviation of interarrival time and bandwidth for considered apps.

Application	Mean interarrival time (seconds)	Standard deviation of interarrival time (seconds)	Mean bandwidth (kilobit/second)	Standard deviation of bandwidth (kilobit/second)
Angry Birds	0.66	15.90	4.42	4.50
Aupeo	0.06	3.06	129.76	482.63
Twitter	8.91	44.09	0.27	0.04
Skype	0.55	1.95	1.30	1.84

of 15.90 seconds. This is caused by the low interarrival times in one advertisement request and the relatively large interarrival times between two advertisements. Mean bandwidth is relatively low with 4.42 kbps and a high standard deviation of 4.5 kbps. These differences can be explained by considering the behaviour of the application. During long phases of use, no traffic is sent, and, after a level is restarted, a new advertisement has to be obtained, causing the transmission of data.

Next, we study the behaviour of the Aupeo application. We see that the application generates packets with relatively small interarrival times. This finding is backed up by the small mean interarrival time of 0.06 seconds. The high standard deviation of 3.06 seconds is caused by the wait between two tracks. Furthermore, we see a high mean bandwidth of 129.76 kbps and a standard deviation of 482.63 kbps. This is caused by the difference in traffic activity between times when tracks are either downloaded or not.

For Twitter, we see that 90% of all transmissions occur with an interarrival time of 1 seconds. Also, we can observe a high mean interarrival time of 8.91 seconds and a high standard deviation of 44.49 seconds. Additionally, the mean bandwidth is low with only 0.27 kbps and a low standard deviation of 0.04 kbps due to the fact that Twitter text messages are only 140 characters in length, and thus only a low volume of traffic needs to be transmitted.

Finally, we consider the Skype application. Similar to the Twitter application, we see that 90% of all packets occur with an interarrival time of less than 1 second. However, in contrast to Twitter, we see a low mean interarrival time of 0.55 seconds with a standard deviation of 1.95 seconds. Further, we observe a relatively low mean bandwidth of 1.30 kbps and a standard deviation of 1.8 kbps.

To further study the traffic patterns of the applications, we study the autocorrelation of the packet interarrival time with regard to the lag length in Figure 3(b). We note that all studied applications present completely different autocorrelations for the interarrival times. This is one of the reasons that the applications under consideration will display different signalling behaviour in the next section.

4.2. Influence of Application Characteristics on Optimisation with Network Timers. This section studies the impact of traffic generated by applications on both the network and the QoE of the user. We consider two metrics. First, we consider the *frequency of signalling messages* perceived at network components such as an RNC. In light of network outages caused by the so called *signalling storms* (a large

number of signalling messages leading to overload at network equipment), it is in the interest of a network operator to reduce the number of signalling messages arriving at the RNC. One possible way to reduce the signalling frequency is to modify network timer values like T_{DCH} and T_{FACH}. As discussed in Section 2.3, the QoE a user perceives while using his device is influenced by the battery life of the UE. Thus, the second metric considered is the influence of the used network model and associated timer settings on the device's *power consumption*. As described in Section 3.3, based on a measurement trace for an application, we use Algorithm 1 to infer the state transitions occurring during the use of the application. Then, we calculate the relative time spent in each state and use Table 2 to compute the mean power consumption of the radio interface during the measurement. We study both metrics: first on its one and then aggregated for both network models introduced in Section 2.1. First, we consider the three-state model in Section 4.2.1, which describes the default behaviour in 3G networks. Then, we describe the influence of the two-state model in Section 4.2.2. Here, we model a network behaviour similar to that if proprietary fast dormancy algorithms are used. These algorithms have been identified as one of the causes of a signalling storm [6]. Finally, we summarize the results and discuss the possible ramifications of using network timer values to reduce the signalling frequency in Section 4.2.3. More numerical results and details can be found in [18].

4.2.1. Three-State Model: Signalling Frequency versus Power Consumption. First, we investigate the signalling frequency generated by the studied applications for the three-state network model. Figure 4(a) shows the signalling frequency with regard to the T_{DCH} timer. For all studies of the three-state model, the FACH timeout is set to $T_{\mathrm{FACH}} = 2 \cdot T_{\mathrm{DCH}}$, a realistic value, as shown in [4]. We see that for T_{DCH} timers shorter than 6 seconds, the Skype application in Idle mode generates the highest signalling message frequency. The Angry Birds application generates the second highest frequency of signalling messages, followed by the Aupeo application. The Twitter application generates the smallest signalling load. If the T_{DCH} value is longer than 15 seconds, this order changes. However, in general, the signalling message frequency for higher T_{DCH} timeouts is lower than for shorter T_{DCH} timeouts. Now, the Aupeo application has the highest signalling frequency, followed by the Twitter application. The signalling message frequency for the Angry Birds application takes the third place. The application which generated the highest signalling message frequency generates the lowest frequency for higher timeout values. This behaviour can be explained by the fact that the Skype application sends keep-alive messages with an interval of less than 20 seconds. If the timer is greater than the interval time of the keep-alive messages, the UE stays always connected and thus generates almost no signalling.

These results show that the traffic patterns of the application have a large influence on the generated signalling load. Signalling is generated for every pause in sending or receiving larger than the configured timeouts. If such pauses occur frequently, this increases the signalling load as shown on the examples of Skype and Angry Birds. Applications with more time between the sending or receiving of data cause less signalling, as shown by Aupeo and Twitter. Furthermore, we can observe that the signalling load can be reduced by increasing the DCH timeout, with the minimum being reached as T_{DCH} approaches infinity. From a signalling load perspective, a value of 20 seconds would probably be sufficient; however, if other metrics such as radio resource consumption are considered, 10 seconds would be acceptable for a network operator.

Based on this finding, we see that increasing the T_{DCH} timer decreases the signalling frequency at the RNC. However, the actual signalling frequency depends on the application running at the UE. From a network operator's point of view, the three-state model should always be preferred to the two-state model because it generates less signalling messages per second, thus decreasing the load at the RNC. This view however does not consider the additional radio resources which are kept in use for a longer time if larger T_{DCH} values are used. Additionally, it should be noted that the choice of the network model is sometimes outside of the domain of the network operator. Proprietary fast dormancy algorithms, as the considered two-state model, are enabled on the UE by the user.

In Figure 5(a), we consider the power consumption if the network uses the three-state model, that is, if the fast dormancy mode of the UE is disabled. The figure shows the mean power consumption of the device with regard to the T_{DCH} timeout. Possible values range between 0 mW if the UE was in Idle state during the whole measurement and 800 mW if the UE was in DCH state during the complete measurement. We see that the least power over all considered T_{DCH} values is consumed by the Twitter application. The second least power consumption is required by Aupeo, followed by Angry Birds. Finally, the most power is consumed by Skype. Here, we see that the maximum value of 800 mW is reached at a T_{DCH} timeout of 20 seconds. This is because, due to the periodic traffic behaviour of Skype, the device is always in the DCH state. Again, we see that the traffic characteristics of the applications impact the power consumption. Applications with more network activity are forced to stay in more power consuming states for a longer time. We see that for very small network timers, the power consumption is minimal. However, as seen in the last section, small timers increase the signalling load at the RNC. Again, a choice of 10 seconds for the T_{DCH} timer can be seen as a compromise between signalling load and power consumption.

Finally, we aggregate both metrics in Figure 6(a). The x-axis of the figure gives the signalling message frequency. On the y-axis, we show the power consumption. Different T_{DCH} values are shown by different colors as specified by the colorbar. First, we consider Angry Birds. We observe that as the signalling frequency approaches zero, the power consumption rapidly increases, even if only small gains in signalling frequency reduction can be achieved. The Aupeo application presents a completely different picture. Here, we can see multiple almost horizontal lines of markers. If T_{DCH} is chosen in this range, each increase of T_{DCH} brings a small

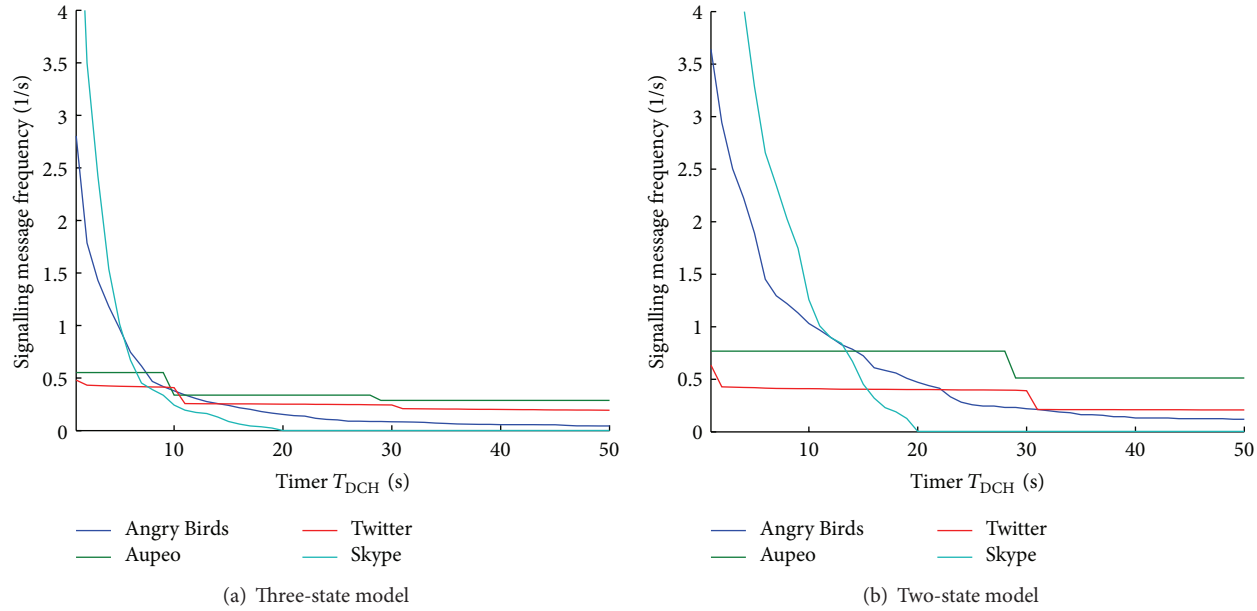


FIGURE 4: Signalling messages frequency for varying T_{DCH} timers.

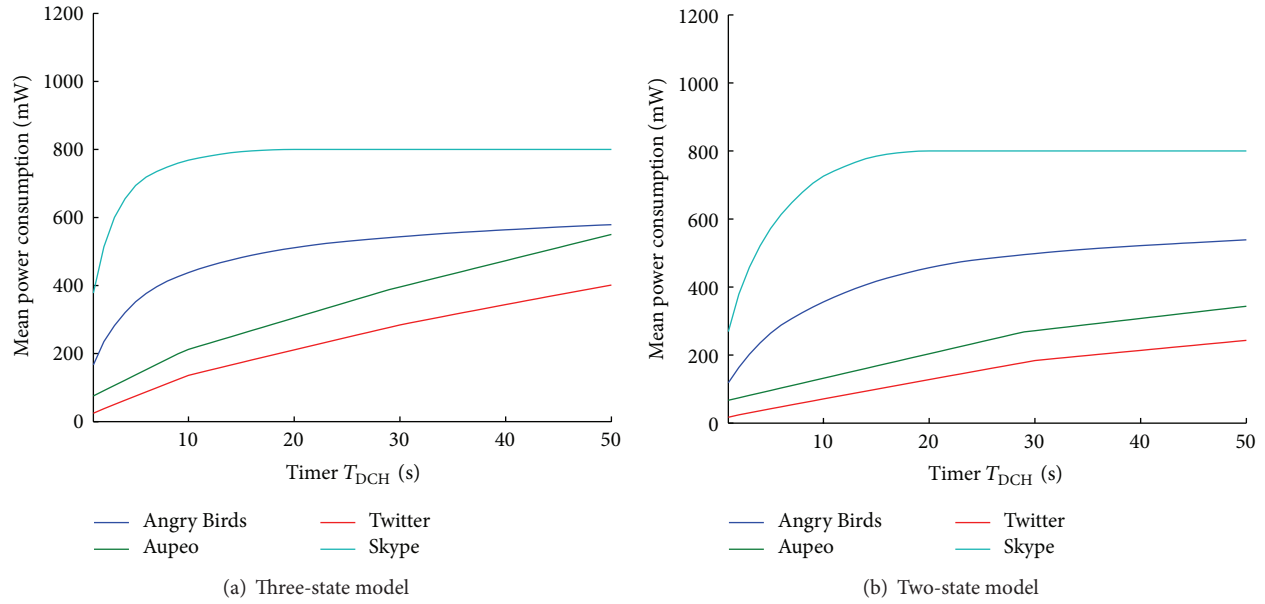
decrease in signalling frequency for an increase in power consumption. However, some points of discontinuity exist. If, for example, the DCH timer is increased from 10 seconds to 11 seconds, a decrease in power consumption of 40% can be achieved by only suffering from a small increase of power consumption. These points of discontinuity would present themselves to be suitable targets of optimisation. Next, we consider the Twitter application. It displays a similar behaviour as the Aupeo application, with multiple points of discontinuity. Note that Twitter exhibits a different point of discontinuity, and the T_{DCH} value of 10 seconds, which provided good results for Aupeo, is not optimal for Twitter. Finally, Skype shows a completely different picture. First, note that due to the large signalling frequency of Skype for small values of T_{DCH}, $T_{DCH} = 1$ second is not displayed in the figure. Furthermore, as the T_{DCH} timer increases above 20 seconds, the signalling frequency does not decrease any further, and the power consumption remains at the maximum value. We observe that there is no common optimal value for all applications which would result in an acceptable tradoff.

4.2.2. Two-State Model: Signalling Frequency versus Power Consumption. Now, we study the consequences of the application traffic in a network using the two-state model. The two-state model occurs in reality if fast dormancy implementations are considered. Here, the UE disconnects from the network if for a certain time no traffic is sent or received in order to reduce power consumption. As for the three-state model, Figure 4(b) shows the signalling frequency with regard to the setting of the T_{DCH} timer. We see the same general behaviour as with the three-state model; however, the signalling frequency generated by each of the applications for the two-state model is usually higher. For example, even for relatively high T_{DCH} timeout values of 10 seconds, the Angry Birds application causes 270% of the signalling frequency as in a network using the three-state model.

Next, we consider the changes in the power consumption of the UE if the user decides to enable fast dormancy, that is, switch to a two-state model, in Figure 5(b). As with the signalling frequency, we only see a quantitative differences to the three-state model. Again, we compare the differences between the two-state model and the three-state model on the example of the Angry Birds application. For the same considered T_{DCH} timeout of 10 seconds, we see a decrease of 81% in power consumption when compared with the 3-state model.

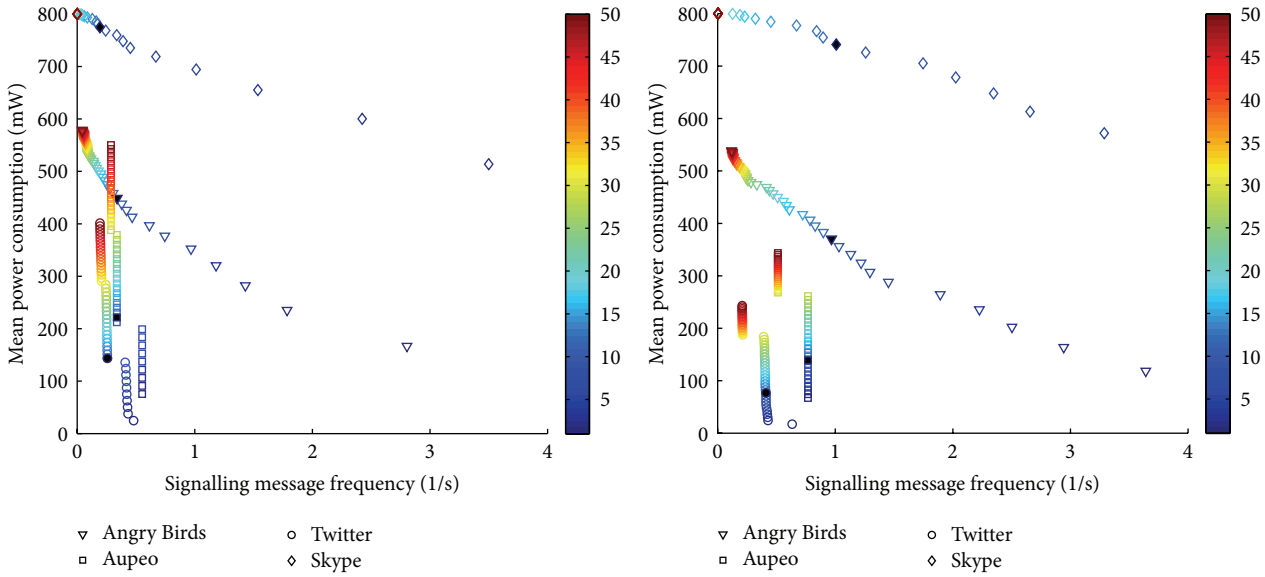
Finally, we compare the influence of changes of the T_{DCH} timeout on both signalling frequency and power consumption for the two-state model in Figure 6(b). As for the three-state model, we see that there is no tradeoff between power consumption and signalling frequency that would be acceptable for all application. Even for single applications, T_{DCH} values such as 11 seconds which was an acceptable tradeoff for Angry Birds is no longer a good choice in the two-state model.

4.2.3. Consequences of Tradeoff: Signalling Frequency versus Power Consumption. To illustrate the ramifications of the behaviour discussed in the previous section, we compare the influence of the T_{DCH} timer on an application with different traffic characteristics, for example, the Aupeo application as shown in Figure 7(b). The signalling load before the increase of the DCH timer was 0.55 messages per second; after the change to $T_{DCH} = 8$ seconds, the load remains unchanged. Thus, the policy change based on one application brings no significant gain to other applications. However, from a user's point of view, the power consumption increased



(a) Three-state model (b) Two-state model

FIGURE 5: Power drain for varying T_{DCH} timers.

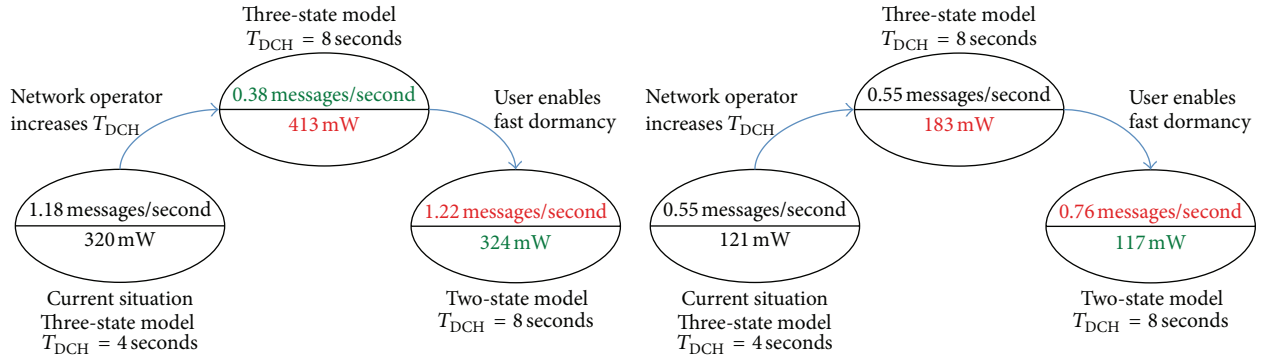


(a) Three-state network model (b) Two-state network model

FIGURE 6: Influence of manipulating T_{DCH} timer on signalling message frequency and power consumption. Filled marker highlights T_{DCH} = 11 seconds.

from 121 mW to 183 mW. Again, we assume that the user activates fast dormancy to deal with the increase in power consumption of more than 50%. This results in a decrease of power consumption to 117 mW and an increase of overall signalling frequency to 0.76 signalling messages per second. By changing the value without considering all applications, the network operator has decreased the QoE for other users and worsened their overall situation. Thus, due to the large number of applications, it seems impossible to optimise the DCH timeout to reduce the signalling message frequency without negatively impacting the users QoE in unexpected ways.

There exist applications, like Twitter and Aupeo, where optimisation by modifying the T_{DCH} values can provide acceptable results. However, these optimisations are only successful if a single application or network model is considered. For other applications, like Angry Birds or Skype, this optimisation approach does not seem to be successful. A reduction of signalling load and power consumption is possible, if the application developers are incentivised to



(a) Results of the optimisation on target application Angry Birds

(b) Influence of the optimisation for Angry Birds on Aupeo

FIGURE 7: Influence of manipulating T_{DCH} timer on different applications.

optimise their applications in these regards. In [4], the authors suggest methods to achieve this optimisation, for example, batch transfer of advertisements for applications like Angry Birds or decreasing the refresh rate in applications like Skype. However, at the moment, neither application developers are receiving incentives to optimise applications in this way, nor hardware vendors do provide interfaces to facilitate such optimisation. Such interfaces would allow application developers to schedule their data transmissions in such a way that both signalling and battery drain would be reduced. Additionally, these interfaces would need to allow the application developer to specify if sending the transmission is urgent because the application is being actively used by the user and requires the feedback of the transmission or if the data is being sent as a regular update, while the application is running in the background and can be scheduled for later transmission as suggested by [19, 20].

To bring network operators, hardware vendors, and application developers together and allow for global optimisation of all relevant metrics, holistic approaches like Economic Traffic Management or Design for Tussle, as described in Section 2.3, are required.

4.3. Impact of Network Configuration and Background Traffic on Web QoE. So far, we have discussed only power consumption as a QoE influence factor. For applications like web browsing, one relevant QoE influence factor is page load times. Therefore, we consider a web QoE model which quantifies the impact of page load times on mean opinion scores [12]. We distinguish here between web QoE and QoE as no QoE models are currently existing which consider page load times as well as power consumption. In this section, we study the impact of background traffic as well as network timer settings on the page load time of an image and the resulting MOS. For this study, we only consider the three-state network model, but the results can be applied to the two-state model as well.

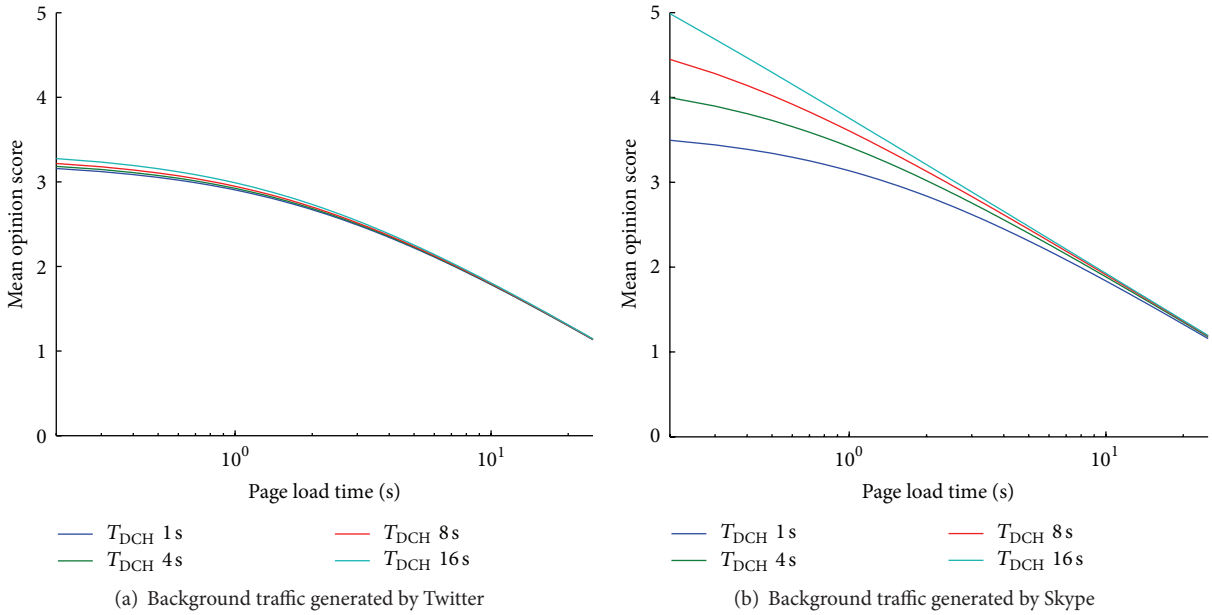
We assume a scenario where a user is running a background application like Twitter or Skype. Then, while the application is in the background, the user begins to download an image from a website. Due to the background traffic, and depending on the network model and associated timer values, the UE may be currently either in Idle, FACH, or DCH state. We give the probability of a random observer encountering the system in FACH state by p_{FACH} and the probability of a random observer encountering in Idle state by p_{Idle}. If the device is currently not in DCH state, it takes some time to connect. This promotion time depends on the current state and is according to [15] 2 seconds if the UE is in Idle state and 1.5 seconds if the device is in the FACH state. For this study, we assume that the user randomly chooses a time to begin downloading an image. The time until the image is displayed consists of the time to load the page t_p as well as the time to go online t_o, where t_o is the mean time to go online, given as

$$t_o = p_{Idle} \cdot 2.5\,\text{s} + p_{FACH} \cdot 1.5\,\text{s}. \quad (1)$$

Thus, the total time t that is required to download the image is given by $t = t_o + t_p$.

The authors of [12] give a function to calculate the MOS based on the required page load time as $\text{QoE}(t) = a \cdot \ln t + b$, where a and b depend on the type of content being downloaded. For our scenario, picture download and values of $a = -0.8$ and $b = 3.77$ are suggested. It has to be noted that for different websites, the logarithmic function was still observed, but different values for a and b were obtained as given in [12]. These values depend for example on the type of web page as well as the size of the content. Nevertheless, the results presented in this section are therefore generalizable for web browsing to various pages. This allows us to give an expected MOS for downloading pictures while a background application is influencing the probability of a device already being in DCH state or still having to be promoted to DCH state.

Using this methodology, we study the influence of background traffic on the QoE for two background applications with different traffic characteristics. In Figure 8(a), we assume that the user is running the Twitter application as a background process. The application is set to update the users' status feed every 5 minutes. In Figure 8(b), the user is running the Skype application as a background application. This application sends keep-alive messages every 20 seconds. For each application, we assume that the three-state network model with T_{DCH} settings of 1, 4, 8, and 16 seconds is used. We always set $T_{FACH} = 2 \cdot T_{DCH}$. In both figures, we show the



(a) Background traffic generated by Twitter

(b) Background traffic generated by Skype

FIGURE 8: Perceived QoE for loading a page with existing background traffic.

assumed page load time, as provided by the network, on the x-axis for values from 0.2 seconds to 25 seconds. We assume that 0.1 seconds is a lower bound because page load times lower than 0.1 seconds are not distinguishable [21] by humans. The calculated MOS values are given on the y-axis.

The picture downloads with the background traffic generated by the Twitter application result in MOS values beginning at 3.15 for $T_{DCH} = 1$ second, 3.18 for $T_{DCH} = 4$ seconds, 3.21 for $T_{DCH} = 8$ seconds, and 3.27 for $T_{DCH} = 16$ seconds, respectively. With increasing page load time, the MOS again decreases. This behaviour is due to the fact that the Twitter application periodically sends traffic every 5 minutes. Then, no further activity occurs until the next refresh occurs. In this time, the UE transits to Idle state. This traffic characteristic causes an high probability of a user encountering the device in an Idle state. In contrast, downloading pictures with the Skype application generating background traffic causes different MOS values. For a page load time of 0.2 seconds, the MOS value with $T_{DCH} = 1$ second is 3.49, with $T_{DCH} = 4$ seconds is 3.99, with $T_{DCH} = 8$ seconds is 4.44, and finally with $T_{DCH} = 16$ seconds is 4.99, respectively. For increasing page load times, the MOS decreases. This increased MOS values occur because of the high frequency of traffic sent by the Skype application. Here, every 20 seconds, traffic is sent. This means that even for relatively low values of T_{DCH}, the user has a high probability of encountering a state where no promotion delay is required before the actual page load time can begin.

From these studies, we can conclude that, when considering QoE on mobile devices, not only the page load time caused by the network but also additional delays caused by the state of the device should be considered. As shown on two examples, this state can be affected by other applications which are running in the background and generate traffic.

5. Conclusion

In this study, we investigate the influence of both application traffic characteristics and network configuration on the signalling load at the RNC as well as the QoE for the user. We consider two different influence factors on QoE: power consumption and page load times for web browsing.

First, we consider the power consumption at the UE. The network operator can reduce the signalling load at the RNC by increasing a network parameter, the T_{DCH} timer, which determines the time a UE remains connected to the network. However, increasing this timer also increases the battery drain at the UE. While it is possible to find a tradeoff between signalling load and battery drain for any single application, we show that for a set of applications, this is not possible. A network parameter, which might be optimal for one application, may cause another application to generate either a high signalling load or an increased battery drain. Additionally, we consider the influence of background applications on the web QoE. For mobile applications, the page load time may be increased if the UE is currently disconnected and has to connect to the network before obtaining the requested content. If another application is already running, the device may, depending on the traffic pattern of the application, already be online, that is, connected to the radio access network. We suggest to consider this when performing QoE studies for mobile users.

When considering ways to improve mobile networks, many different players, metrics, and tradeoffs exist. We highlighted one examples of such a tradeoff, that is, signalling frequency versus power consumption, and discussed the influence of the current optimisation parameters and the network timers on another. However, many additional tradeoffs exist. For example, the mobile operator has to balance the use of radio resources with the number of

generated signalling frequencies. Furthermore, application providers seek to improve the user experience which usually results in a higher frequency of network polls, creating additional signalling traffic. The high number of tradeoffs and involved actors in this optimisation problem indicates that the current optimisation technique used by operators is no longer sufficient.

Approaches like *Economic Traffic Management* or *Design for Tussle* could be applied to find an acceptable tradeoff for all parties. In economic traffic management all, participating entities share information in order to enable collaboration. This collaboration allows for a joint optimisation of the tradeoff. Design for tussle aims to resolve tussles at run time, instead of design time. This prevents the case that one actor has full control over the optimisation problem, which would likely result in the actor choosing a tradeoff only in its favour, ignoring all other participants.

One example of an actor providing information for another in order to optimise the total system would be UE vendor could provide interfaces for application developers to use when sending data. These interfaces would schedule data to be transmitted in such a way that signalling load and battery drain would be reduced, if the application's requirements allow for it. Until such interfaces exist, application developers could consider the effect of the traffic their applications produce both on the UE and the network. To this end, we will provide a web application as future work, which will allow application developers to upload and analyse their applications' traffic according to different network models.

Acknowledgments

This work was partly funded by the Deutsche Forschungsgemeinschaft (DFG) under Grants HO 4770/1-1 and TR257/31-1. The authors would like to thank Florian Wamser for the fruitful discussions. The authors alone are responsible for the content.

References

[1] F. Cuadrado and J. Dueñas, "Mobile application stores: success factors, existing approaches, and future developments," *IEEE Communications Magazine*, vol. 50, no. 11, pp. 160–167, 2012.

[2] Cisco, "Cisco visual networking index: global mobile data traffic forecast update, 2011–2016," White Paper, 2012.

[3] TS 25.331, "Radio Resource Control (RRC); Protocol specification," 3GPP Std., 2012.

[4] F. Qian, Z. Wang, A. Gerber, Z. Mao, S. Sen, and O. Spatscheck, "Profiling resource usage for mobile applications: a cross-layer approach," in *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services (MobiSys '11)*, pp. 321–334, July 2011.

[5] F. Qian, Z. Wang, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "Characterizing radio resource allocation for 3G networks," in *Proceedings of the 10th Internet Measurement Conference (IMC '10)*, pp. 137–150, Melbourne, Australia, November 2010.

[6] Nokia Siemens Networks, "Understanding smartphone behavior in the network," White Paper, 2011.

[7] C. Yang, "Huawei communicate: weather the signalling storm," White Paper, 2011.

[8] S. Ickin, K. Wac, M. Fiedler, L. Janowski, J. Hong, and A. Dey, "Factors influencing quality of experience of commonly used mobile applications," *IEEE Communications*, vol. 50, no. 4, pp. 48–56, 2012.

[9] T. Hoßfeld, D. Hausheer, F. Hecht et al., "An economic traffic management approach to enable the TripleWin for users, ISPs, and overlay providers," in *Towards the Future Internet—A European Research Perspective*, G. Tselentis, J. Domingue, A. Galis et al., Eds., pp. 24–34, IOS Press Books Online, May 2009.

[10] D. D. Clark, J. Wroclawski, K. R. Sollins, and R. Braden, "Tussle in cyberspace: defining tomorrow's internet," *IEEE/ACM Transactions on Networking*, vol. 13, no. 3, pp. 462–475, 2005.

[11] Trilogy Project, "D2—lessons in "designing for tussle" from case studies," Deliverable of the Trilogy Project ICT-216372, 2008.

[12] S. Egger, P. Reichl, T. Hoßfeld, and R. Schatz, "Time is bandwidth? Narrowing the gap between subjective time perception and quality of experience," in *Proceedings of the International Conference on Communications (ICC '12)*, p. 1325, Ottawa, Canada, June 2012.

[13] P. H. J. Perälä, A. Barbuzzi, G. Boggia, and K. Pentikousis, "Theory and practice of RRC state transitions in UMTS networks," in *Proceedings of the Global Communication Conference (GLOBECOM '09)*, pp. 1–6, Honolulu, Hawaii, USA, December 2009.

[14] GSMAssociation and others, "Network efficiency task force fast dormancy best practices," White Paper, May 2010.

[15] F. Qian, Z. Wang, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "TOP: tail optimization protocol for cellular radio resource allocation," in *Proceedings of the 18th IEEE International Conference on Network Protocols (ICNP '10)*, pp. 285–294, Kyoto, Japan, October 2010.

[16] TR 22.801, "Study on non-MTC Mobile Data Applications impacts (Release 11)," 3GPP Std., 2011.

[17] A. Díaz Zayas and P. Merino Gómez, "A testbed for energy profile characterization of IP services in smartphones over live networks," *Mobile Networks and Applications*, vol. 15, no. 3, pp. 330–343, 2010.

[18] C. Schwartz, T. Hoßfeld, F. Lehrieder, and P. Tran-Gia, "Performance analysis of the trade-off between signalling load and power consumption for popular smartphone apps in 3G networks," Tech. Rep. 485, University of Würzburg, January 2013.

[19] M. Calder and M. K. Marina, "Batch scheduling of recurrent applications for energy savings on mobile phones," in *Proceedings of the 7th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON '10)*, pp. 1–3, Boston, Mass, USA, June 2010.

[20] E. Vergara and S. Nadjm-Tehrani, ""Energy-aware cross-layer burst buffering for wireless communication," in *Proceedings of the Conference on Future Energy Systems: Where Energy, Computing and Communication Meet*, p. 24, Madrid, Spain, May 2012.

[21] S. Egger, T. Hoßfeld, R. Schatz, and M. Fiedler, "Waiting times in quality of experience for web based services," in *Proceedings of the Workshop on Quality of Multimedia Experience (QoMEX '12)*, pp. 86–96, Yarra Valley, Australia, July 2012.

Hindawi Publishing Corporation
Journal of Computer Networks and Communications
Volume 2013, Article ID 165146, 28 pages
http://dx.doi.org/10.1155/2013/165146

Review Article

Survey and Challenges of QoE Management Issues in Wireless Networks

Sabina Baraković[1] and Lea Skorin-Kapov[2]

[1] *Ministry of Security of Bosnia and Herzegovina, Trg BiH 1, 71000 Sarajevo, Bosnia and Herzegovina*
[2] *Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 10000 Zagreb, Croatia*

Correspondence should be addressed to Lea Skorin-Kapov; lea.skorin-kapov@fer.hr

Received 7 September 2012; Revised 8 December 2012; Accepted 12 December 2012

Academic Editor: Raimund Schatz

With the move towards converged all-IP wireless network environments, managing end-user Quality of Experience (QoE) poses a challenging task, aimed at meeting high user expectations and requirements regarding reliable and cost-effective communication, access to any service, anytime and anywhere, and across multiple operator domains. In this paper, we give a survey of state-of-the-art research activities addressing the field of QoE management, focusing in particular on the domain of wireless networks and addressing three management aspects: QoE modeling, monitoring and measurement, and adaptation and optimization. Furthermore, we identify and discuss the key aspects and challenges that need to be considered when conducting research in this area.

1. Introduction

Wireless mobile communications have experienced phenomenal growth throughout the last decades, going from support for circuit-switched voice services and messaging services to IP-based mobile broadband services using High Speed Packet Access (HSPA), Worldwide Interoperability for Microwave Access (WiMAX), and Long-Term Evolution (LTE) Radio Access networks [1]. Increasingly, mobile applications and services are being used in daily life activities in order to support the needs for information, communication, or leisure [2]. Mobile users are requiring access to a wide spectrum of various multimedia applications/services without being limited by constraints such as time, location, technology, device, and mobility restrictions. This represents the outcome of the currently leading trend and future aim in the telecommunications domain: the convergence between fixed and mobile networks, and the integration of existing and new wireless technologies. Such integrations aim to satisfy mobile users' requirements in terms of providing access to any service, along with reliable and cost-effective communication, anytime and anywhere, over any medium and networking technology, and across multiple operator domains [3].

The ITU has specified the Next Generation Network (NGN) as a generic framework for enabling network convergence and realizing the aforementioned requirements [4]. The NGN concept is centered around a heterogeneous infrastructure of various access, transport, control, and service solutions, merged into a single multimedia-rich service provisioning environment. Today, an increasing number of mobile operators are migrating their networks in line with the 3GPP specified Evolved Packet System (EPS), consisting of a multiaccess IP-based core network referred to as the Evolved Packet Core (EPC), and a new LTE 4G radio access network based on Orthogonal Frequency Division Multiplexing (OFDM) [1, 5, 6]. While the network controlled and class-based Quality of Service (QoS) concept of the EPC are based on the 3GPP policy and charging control (PCC) framework [7–9] (discussed further in Section 4), intense recent research in the area of Quality of Experience (QoE) has shown that such QoS mechanisms may need to be complemented with more user-centric approaches in order to truly meet end-user requirements and expectations.

Today, humans are quality meters, and their expectations, perceptions, and needs with respect to a particular product, service, or application carry a great value [10]. While the ITU-T has defined QoE as the "*overall acceptability of an application or service, as perceived subjectively by the end user*" [11], ETSI defines QoE as "*a measure of user performance based on both objective and subjective psychological measures of using an ICT service or product*" [12] and extends QoE beyond subjective to include objective psychological measures.

QoE is therefore considered to be a multidimensional construct, encompassing both objective (e.g., performance related) and subjective (e.g., user related) aspects [13]. As such, QoE has been considered in relation to both QoS, which is primarily a technical, objective, and technology-oriented concept, as well as to User Experience (UX) [14] which is generally considered as a more user-oriented concept. The former focuses on the impact of network and application performance on user quality perception, while the latter primarily deals with the individual users' experiences derived from encounters with systems, impacted by expectations, prior experiences, feelings, thoughts, context, and so forth.

Various approaches such as [15–19] provide definitions of QoE that are closely related to technology-centric logic, not accounting for the subjective character of human experience, and lacking consideration of a broader definition of QoE [20]; that is, the consequence of the assumption that the optimization of QoS-related parameters will automatically result in increasing the overall QoE, leading to swift adoption of products and services on the consumption side. However, QoS is only a subset of the overall QoE scope. Higher QoS would probably result in higher QoE in many cases, but fulfilling all traffic-related QoS requirements will not necessarily guarantee high user QoE. Moreover, it is assumed that products and services that meet users' requirements and expectations and that allow them to have high QoE in their personal context will probably be more successful than products and services that have higher QoS but fail to meet users' high demands and expectations [21].

A recent definition that has emerged from the EU Qualinet community (COST Action IC1003: "European Network on Quality of Experience in Multimedia Systems and Services") encompasses the discussed aspects and defines QoE as "*the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state. In the context of communication services, QoE is influenced by service, content, device, application, and context of use*" [22].

In the context of converged all-IP wireless networks, an important consideration is the impact of mobility on user QoE, further related to mechanisms for assuring session and connection continuity. Session establishment delays are impacted by authorization and authentication procedures, as well as session establishment signaling. Handover may impose additional delays and packet losses, resulting in potential loss of session-related content and awkward communication. Hence, mechanisms are necessary for achieving seamlessly session continuity and minimizing disruption time [21]. Besides the challenges that various types of mobility (terminal, session, user, service) and seamless handover between networks that use the same technology (horizontal handover) or networks that use different technologies (vertical handover) represent, the fast development of new and complex mobile multimedia services that can be delivered via various new mobile devices (smartphones, tablets, etc.) poses an additional challenge in the QoE provisioning process [3]. In the context of wireless systems, limitations arising from both device and transmission channel characteristics have a clear impact on user quality perception [23].

The overall goal of QoE management may be related to optimizing end-user QoE (end-user perspective), while making efficient (current and future) use of network resources and maintaining a satisfied customer base (provider perspective). In order to successfully manage QoE for a specific application, it is necessary to understand and identify multiple factors affecting it (subjective and objective) from the point of view of various actors in the service provisioning chain, and how they impact QoE. Resulting QoE models dictate the parameters to be monitored and measured, with the ultimate goal being effective QoE optimization strategies. Therefore, the overall process of QoE management may be broken down into three general steps: (1) QoE modeling, (2) QoE monitoring and measurements, and (3) QoE optimization and control [24].

With the implementation of successful QoE management, users will benefit with satisfied requirements/expectations and may be further inclined to adopt new complex services and support further technology development. Furthermore, QoE management is very important for all actors and stakeholders involved in the service provisioning chain: device manufacturers, network providers, service providers, content providers, cloud providers, and so forth. In today's highly competitive environment, where providers' price levels are decreasing and pricing schemes are becoming more similar [25], it is not enough to simply make services available to users, who further have the option of choosing from a spectrum of various providers. Actors involved in the process of service provisioning have identified the need to work towards meeting users' requirements and expectations by maximizing users' satisfaction with the overall perceived service quality, while at the same time minimizing their costs. Understanding and managing QoE is needed in order to react quickly to quality problems (preferably) before customers perceive them. Hence, successful QoE management offers stakeholders a competitive advantage in the fight to prevent customer churn and attract new customers.

Based on the previously mentioned, it may be concluded that QoE is a fast emerging multidisciplinary field based on social psychology, cognitive science, economics, and engineering science, focused on understanding overall human quality requirements [10]. Consequently, management of QoE as a highly complex issue requires an interdisciplinary view from user, technology, context, and business aspects, with flexible cooperation between all players and stakeholders involved in the service providing chain.

In this paper we give a survey of approaches and solutions related to QoE management, focusing in particular on wireless network environments. It is important to note

that different QoE models and assessment methodologies are applicable for different types of services (e.g., conversational voice services, streaming audio-visual services, interactive data services, collaborative services). We do not focus on a particular service but rather give a general survey of approaches. The paper is organized as follows. Section 2 surveys the modeling of QoE by discussing the classification of a wide range of QoE influence factors into certain dimensions and describing existing general QoE models. The monitoring and measurement of QoE is described in Section 3, while Section 4 discusses the topic of QoE optimization and control. Finally, Section 5 concludes the paper by pointing out the challenges and open research issues in the field of QoE management.

2. QoE Modeling

As a prerequisite to successful QoE management, there is a need for a deep and comprehensive understanding of the influencing factors (IF) and multiple dimensions of human quality perception. QoE modeling aims to model the relationship between different measurable QoE IFs and quantifiable QoE dimensions (or features) for a given service scenario. Such models serve the purpose of making QoE estimations, given a set of conditions, corresponding as closely as possible to the QoE as perceived by end users. Based on a given QoE model specifying a weighted combination of QoE dimensions and a further mapping to IFs, a QoE management approach will then aim to derive Key Quality Indicators (KQIs) and their relation with measureable parameters, along with quality thresholds, for the purpose of fulfilling a set optimization goal (e.g., maximizing QoE to maximize profit, maximizing number of "satisfied" customers). An important issue to note is that different actors involved in the service provisioning chain will use a QoE model in different ways, focusing on those parameters over which a given actor has control (e.g., a network provider will consider how QoS-related performance parameters will impact QoE, while a content or service provider will be interested in how the service design or usability will impact QoE).

In this section we first discuss QoE IFs in general, and give an overview and comparison of general QoE modeling approaches, discussing in turn their applicability with respect to QoE management strategies. We then further consider more concretely QoE models targeted specifically towards wireless networks, highlighting the differences with respect to fixed networks. We end the section with a summary of QoE modeling challenges.

2.1. QoE Influence Factors. A QoE IF has been defined as "*any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user*" [22]. Figure 1 illustrates a multitude of different factors which may be considered in relation to QoE, making it clear that their grouping into categories aids in identifying such factors in a systematic way. Several existing approaches have addressed this issue and proposed classifications of QoE IFs into multiple dimensions. It should be noted that specific IFs are relevant for different types of services and applications.

Stankiewicz and Jajszczyk [3] have classified the technology-oriented factors that impact QoE into three groups: QoS factors, Grade of Service (GoS) factors, and Quality of Resilience (QoR) factors, believing that provisioning of those at the appropriate level is crucial for achieving high QoE. Also, they take into consideration a number of additional factors (mostly nontechnology related) such as emotions, user profile, pricing policy, application specific features, terminals, codecs, type of content, and environmental, psychological, and sociological aspects, but they do not further group them. On the other hand, Baraković et al. [21] have categorized QoE influence factors into five dimensions: (1) technology performance on four levels: application/service, server, network, and device; (2) usability, referring to users' behavior when using the technology; (3) subjective evaluation; (4) expectations; and (5) context. Recently, Skorin-Kapov and Varela [26] have proposed the ARCU model that groups QoE factors into four multidimensional IF spaces: *Application* (application configuration-related factors), *Resource* (network/system related factors), *Context*, and *User* spaces.

Finally, a recent classification that has emerged from the EU Qualinet community in the form of a White Paper [22] groups QoE IFs into the following three categories (which are additionally divided into several subcategories as described in the referenced whitepaper).

(i) "Human IFs present any variant or invariant property or characteristic of a human user. The characteristic can describe the demographic and socioeconomic background, the physical and mental constitution, or the user's emotional state" (e.g., user's visual and auditory acuity, gender, age, motivation, education background, emotions).

(ii) "System IFs refer to properties and characteristics that determine the technically produced quality of an application or service. They are related to media capture, coding, transmission, storage, rendering, and reproduction/display, as well as to the communication of information itself from content production to user" (e.g., bandwidth, delay, jitter, loss, throughput, security, display size, resolution).

(iii) "Context IFs are defined as factors that embrace any situational property to describe the user's environment in terms of physical, temporal, social, economic, task, and technical characteristics" (e.g., location, movements, time of day, costs, subscription type, privacy).

2.1.1. Relationships between QoS and QoE. Among the wide scope of discussed IFs, a great deal of research has focused in particular on identifying the relationships between QoS parameters and QoE, whereby in many cases a user's perceived quality has been argued to mostly depend on QoS [27–29]. In studies focused on the mathematical interdependency of QoE and QoS, Reichl et al. [28, 30] have identified a

Figure 1: Different factors to be considered in relation to QoE.

logarithmic relationship between QoE and QoS. They argue that it can be explained on behalf of the Weber-Fechner Law (WFL) [31], which studies the perceptive abilities of the human sensory system to a physical stimulus in a quantitative fashion, and states that the *just noticeable difference* between two levels of certain stimulus is proportional to the magnitude of the stimuli. A logarithmic relationship formulates the sensitivity of QoE as a reciprocal function of QoS.

On the other hand, the IQX hypothesis presented and evaluated by Fiedler et al. in [29] formulates the sensitivity of QoE as an exponential function of a single QoS impairment factor. The underlying assumption is that the change of QoE depends on the actual level of QoE. Understanding the relationship between network-based QoS parameters and user-perceived QoE provides important input for the QoE management process, in particular to network providers with control over network resource planning and provisioning mechanisms.

2.2. From Subjective Quality Assessment to Objective Quality Estimation Models. When building a QoE model, quality assessment methodologies must be employed. While actual "ground-truth" user perceived quality may be obtained only via subjective assessment methodologies, the goal is to use subjective tests as a basis for building objective QoE models capable of estimating QoE based solely on objective quality measurements. Hence, we shortly describe subjective quality assessment methodologies and different types of objective quality assessment methods and models.

2.2.1. Subjective QoE Assessment. Subjective quality assessments are based on psychoacoustic/visual experiments which represent the fundamental and most reliable way to assess users' QoE, although they are complex and costly. These methods have been investigated for many years and have enabled researchers to gain a deeper understanding of the subjective dimensions of QoE. Most commonly, the outcomes of any subjective experiment are quality ratings from users obtained during use of the service (in-service) or after service use (out-of-service), which are then averaged into Mean Opinion Scores (MOSs). This approach has been specified in ITU-T Recommendation P.800.1 [32] and expresses an average quality rating of a group of users by means of standardized scales. For a number of reasons, the use of MOS has been criticized [33] and extended to other ITU recommended subjective assessment procedures classified by type of application and media [34]. An interested reader is referred to the large number of standards which are referenced in [34].

In addition to standardized subjective QoE assessment methods, additional (sometimes complementary) relevant methods used for long-term user experience assessment have

been used. In their studies involving QoE evaluations of mobile applications, Wac et al. [35] have collected users' QoE ratings on their mobile phones via an Experience Sampling Method (ESM) [36] several times per day, while a Day Reconstruction Method (DRM) [37] has been used to interview users on a weekly basis regarding their usage patterns and experiences towards the mobile applications. These methods have served to analyze possible relations and causalities between QoE ratings, QoS, and context.

With regards to collection of data and running of QoE experiments, assessments may be conducted in a laboratory setting [38], in a living labs environment [20], or in an actual real world environment [27, 35]. Some performance criteria are modified in a given range in a controlled fashion and subsequently users' opinions regarding the service performance are quantified. As an emerging and very prospective solution focusing on obtaining a large number of ratings in a real world environment, crowdsourcing methodology [39, 40] has been studied and utilized.

2.2.2. Objective QoE Models. Objective QoE models are defined as the means for estimating subjective quality solely from objective quality measurement or indices [41]. In other words, these models are expected to provide an indication which approximates the rating that would be obtained from subjective assessment methods. Different types of objective quality estimation and prediction models have been developed. Each model has its proper domain of application and range of system or service conditions it has been designed for. Since there exists no universal objective quality assessment approach, proposed ones can be categorized by various criteria in order to determine their application area [34]: (1) application scope; (2) quality features being predicted; (3) considered network components and configurations; (4) model input parameters; (5) measurement approach; (6) level of service interactivity that can be assessed; and (7) level to which psychophysical knowledge or empirical data have been incorporated.

According to the level at which the input information is extracted, there are five types of objective QoE models [34]: (1) media-layer; (2) packet-layer; (3) bitstream-layer; (4) hybrid; and (5) planning models. Media-layer models [42–45] estimate the audio/video QoE by using the actual media signals as their input. In dependence of utilization of the source signal, one can use three different approaches [46, 47]: (1) no-reference (NR) model; (2) reduced-reference (RR) model; and (3) full-reference (FR) model. On the other hand, packet-layer models [48] utilize only the packet header information for QoE estimation, which describes them as in-service nonintrusive quality monitoring approaches. A bitstream-layer model is the combination of the two previously mentioned models, since it utilizes bitstream information as well as packet header information. Similarly, hybrid models [49–51] are conceived as a combination of the previously described three models. Finally, planning models do not acquire the input information from an existing service, but estimate it based on service information available during the planning phase.

In addition to ITU-T standards, ETSI gives a comprehensive guide with generic definitions and test methods for most of the key telecommunication services [52]. There are a number of other standardization bodies that deal with QoE assessment, including VQEG, MPEG, and JPEG.

While most of the current literature considers objective measures in relation to technology oriented collections of data, it is important to note that objective measurements may also refer to objective estimations of user's behavior (e.g., task duration, number of mouse clicks) which is commonly considered only as subjective [53].

2.3. A Survey of General QoE Modeling Approaches. Besides briefly described subjective assessment methods that may contribute to building objective QoE models, also addressed in the previous subsection, we survey a number of general QoE modeling approaches in order to obtain a "broader picture" on this topic. Therefore, this subsection provides an analysis of several general QoE models that were validated by various types of services. While the listed models are not all strictly limited to wireless environments, they aim to identify numerous QoE IFs and provide mechanisms for relating them to QoE.

Table 1 summarizes these QoE models based on the following comparison parameters: IFs according to categorization in [22], type of service, consideration of wireless aspects, provisioning of the concrete QoE model, and applicability with respect to QoE management. This comparison provides an extension and modification of the analysis given by Laghari et al. in [10]. We consider more concretely QoE models targeted specifically for wireless environments in the following subsection.

Perkis et al. in [54] present a model for measuring the QoE of multimedia services, distinguishing between measurable and nonmeasurable parameters. In other words, the approach does not provide any QoE metric relationship formulae but rather addresses the factors that influence user's QoE. The measurable model parameters are closely related to the technology aspects of the terminal and service, and nonmeasurable entities are closely related to user's perception of a service, his/her expectations, and behavior. Additionally, a framework for quantifying the model parameters is described and validated with Voice on Demand (VoD) and mobile TV services in a mobile 3G environment. Although the authors categorize the parameters by accounting for QoS, QoE, and business aspects, and thereby encompass human and system parameters, the model does not fully include the context dimension. However, the proposed modeling framework gives the input information for the measurement process and thereby contributes to the overall QoE management by aiding the various parties in improving their performance.

A model that does not clearly encompass all QoE dimensions, but rather considers them in a limited manner, is introduced by Kim et al. [55]. The In-service Feedback QoE Framework (IFQF) is a user-triggering scheme aimed at investigating the main reasons of quality deterioration and thereby contributing to the overall QoE management process. The architecture consists of four agents: server, network, user,

Table 1: Comparison of QoE modeling approaches.

Modeling approach	Human influence factors		System influence factors				Context influence factors	Type of service	Wireless aspects	Concrete QoE model	QoE management applicability
	Low-level processing	High-level processing	Content	Media	Network	Device					
Perkis et al. [54]	Limited	No	Yes	No	Yes	Yes	No	Voice on demand, mobile TV	Yes	No	(i) Provides the input info for measurement purposes (ii) Most applicable in the context of service optimization
Kim et al. [55]	Limited	Limited	Limited	Not explicitly considered	Yes	Yes	Limited	IP television	Not specified	No	(i) Investigates the reasons for quality deterioration (ii) Most applicable in the context of service optimization
Kilkki [56]	Not explicitly considered	Not explicitly considered	Not explicitly considered	Not explicitly considered	Yes	No	Not explicitly considered	Multimedia	Not specified	No	(i) Proposes a holistic approach to QoE (ii) Most applicable in the context of development of a general QoE management strategy
Möller et al. [57]	Yes	Yes	Not explicitly considered	Not explicitly considered	Yes	Yes	Yes	Multimedia	Not specified	No	(i) Provides a detailed taxonomy of QoE aspects (ii) Most applicable in the context of target-oriented design and system optimization
Geerts et al. [58]	Yes	Yes	Not explicitly considered	No	Yes	Yes	Yes	Not specified	Yes	No	(i) Offers info on measuring different components of QoE (ii) Most applicable in the context of system (service and network) optimization
Laghari et al. [10]	Yes	Yes	Yes	Not explicitly considered	Yes	Yes	Yes	Voice on demand	Not specified	Yes	(i) Establishes the causal relationship between different QoE domains (ii) Most applicable in the context of system optimization

Table 1: Continued.

Modeling approach	Human influence factors		System influence factors				Context influence factors	Type of service	Wireless aspects	Concrete QoE model	QoE management applicability
	Low-level processing	High-level processing	Content	Media	Network	Device					
Volk et al. [59]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Multimedia	Not specified	Yes	(i) Provides the basis for QoE estimation (ii) Most applicable in the context of service estimation and optimization (iii) Most applicable in the context of optimized network resource allocation
De Moor et al. [20]	Limited	Yes	Yes	Yes	Yes	Yes	Yes	Multimedia	Yes	No	(i) Proposes an architecture for QoE monitoring and measurement (ii) Most applicable in the context of service estimation and optimization (iii) Most applicable in the context of optimized network resource allocation
Song et al. [60]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Video	Yes	No	(i) Enables the development of strategies for QoE improvement (ii) Most applicable in terms of optimizing user experience in a mobile context
Reichl et al. [28]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Broadband and Internet	Yes	Yes	(i) Provides QoE prediction and root cause analysis—applicable for network planning (ii) Most applicable in the context of service optimization (iii) Most applicable in the context of development of a general QoE management strategy

and management agent that gather information and form a feedback loop to find out the reason and location of faults, and thereby to minimize the difference between QoE value estimated by operators and the real QoE (as subjectively perceived by the user).

In [56], Kilkki proposed a framework that identifies the relationship between QoS and QoE, but does not explicitly consider QoE components in detail. The framework connects different research communities, including engineers, economists, and behavioral scientists. The author makes a strong case for a holistic approach to QoE and suggests the establishment of a multidisciplinary research group which would address the complexity of QoE. Additionally, key terms in the communication ecosystem are stated, but no classification of QoE factors or any details on the taxonomy are provided. However, the framework introduces new concepts such as Quality of User Experience (QoUE) and Quality of Customer Experience (QoCE).

On the contrary to the previous approaches which considered various QoE factors only partially or in a more abstract fashion, Möller et al. [57] have developed a detailed taxonomy of the most relevant QoS and QoE aspects focusing on multimodal human-machine interactions, as well as factors influencing its QoS. The taxonomy consists of three layers: (1) the QoS-influencing factors related to the user, the system, and the context of usage; (2) the QoS interaction performance aspects describing user and system behavior and performance; and (3) the QoE aspects related to the quality perception and judgment processes taking place inside the user. In addition to previously described approaches, this one also does not provide any concrete formulation of QoE metrics relationship but recognizes the need for one with corresponding weights given to QoE IFs in order to contribute to target-oriented design and QoE optimization in future systems. However, it is believed that the developed detailed taxonomy provided in this paper will aid in producing concrete formulation.

As in [56], Geerts et al. [58] have taken a multidisciplinary approach and included researchers from backgrounds such as sociology, communication science, psychology, software development, and computer science in order to create a comprehensive framework. The proposed model consists of four components: user, ICT product, use process, and context. Each component is divided into several subcategories, which then encompass all three aforementioned QoE IF categories. Although the proposed approach does not introduce a weighted QoE formulae, it aims to provide a detailed look at the different components of QoE offering concrete information on how they can be measured.

Another approach accounting for all QoE IF categories is proposed by Laghari et al. [10]. The authors have proposed a high-level QoE model that can be adapted to many specific contexts. It consists of four domains, that is, sets of knowledge, activity, or influence in the proposed model: human, context, technology, and business. Therefore, the model addresses QoE from multiple aspects, while it can be noted that it is more subjectively oriented towards the human domain. Also, the framework defines the main interactions of the domains: human-context, human-technology, human-business, technology-business, and context-technobusiness, as well as presenting causal relationships between domain characteristics. In other words, the presented formulation relates QoE (set of outcome factors) with a "cause-effect" relationship directly affected with the prediction factors (e.g., technological, business, or contextual characteristics) and indirectly with mediating factors (e.g., associations between aforementioned factors). Additionally, by providing a well-structured detailed taxonomy of QoE relevant variables and formulating the causal relationship between them, this approach aids various interested parties in comprehending and managing QoE in a broader manner.

Volk et al. [59] present a novel approach to QoE modeling and assurance in an NGN Service Delivery Environment (SDE). The proposed model is context aware and comprises a comprehensive set of quality-related parameters available throughout various information factories of the NGN and accessible by employing standardized procedures within the NGN SDE. QoS and various human perception components are addressed. Furthermore, parameter selection and mapping definitions are established vertically from the transport layer through the application layer to the end-user layer, and horizontally with concatenation of point-to-point QoS and end-to-end QoE.

De Moor et al. [20] propose a framework that enables the evaluation of multidimensional QoE in a mobile testbed-oriented living lab setting. The model consists of a distributed architecture for monitoring the network QoS, context information, and subjective user experience based on the functional requirement related to real-time experience measurements in real-life settings. The architecture allows the study and understanding of cross-contextual effects, the assessment of the relative importance of parameters, and the development of a basic algorithmic QoE model.

Although Song et al. [60] have not proposed exact formulae for QoE calculation, they have organized QoE IFs into three components: user, system, and context and mapped their impacts upon four elements of the mobile video delivery framework, namely, mobile user, mobile device, mobile network, and mobile video service, since the model is created for a mobile video environment. User-centered design of mobile video may benefit from this model, as well as mobile video vendors that may develop effective strategies to improve user's experience.

Finally, the framework discussed by Reichl et al. [28] is aimed at improved modeling, measurement, and management of QoE for mobile broadband services (e.g., mobile Web browsing, file download). The authors have developed a model for predicting QoE of network services, based on a layered approach distinguishing between the network, application and the user layers. The layered approach derives the most relevant performance indicators (e.g., network performance indicators, user-experience characteristics, and specific application/service related performance indicator) and aims to builds accurate QoE models by combining user studies providing direct ratings, with logged data at the application layer and traffic measurements at the network layer. The employed laboratory setup includes the user's device and two network emulators which model the behavior

of a variable UMTS/HSPA network. Participants' network traffic is captured using the METAWIN [61] passive monitoring system that monitors traffic on all the interfaces of the packet-switched core network. Passive network measurements combined with obtained subjective user ratings serve to build reliable QoE models for future QoE estimation. The proposed model provides an interdisciplinary perspective including aspects such as device and application usability, usage context, user personality, emotional issues, and user roles. Thereby, it can be used not only for QoE prediction and management, but also for uncovering functional dependencies between causally relevant performance indicators and the resulting perceived quality.

Based on the comparison given in Table 1, it may be summarized that the majority of discussed approaches [10, 20, 28, 57–60], differing in considered type of service and wireless aspects, address the human IFs at both low-level (i.e., physical, emotional, and mental constitution of the user) and high-level processing (i.e., cognitive ability, interpretation, or judgment). These models also consider system IFs classified into content, media, network, and device factors, as well as context IFs, while a set of them such as [55, 56] do not explicitly address IFs in that fashion. However, although the analyzed approaches address different IFs, most provide a QoE modeling framework, while only a few of them provide a concrete model [10, 28, 59].

In the context of QoE management, all addressed QoE modeling approaches, each in its own way, contribute to this process and may be applied in various contexts such as system or service/application optimization, as well as network resource allocation improvement. Thus several of them have been built to aid monitoring, measurement, and estimation of QoE [20, 54, 58, 59], while others contribute to QoE improvement by diagnosing the main reasons for quality deterioration [55], enabling development of strategies [60], and providing detailed taxonomy [56, 57] and causal relationships [10], as well as QoE prediction mechanisms [28].

2.4. QoE Modeling in the Context of Wireless Networks. QoE modeling becomes even more challenging in the context of wireless and mobile networks due to additional issues posed by this variable environment. Previously analyzed QoE modeling approaches have addressed the points common to both fixed and wireless-mobile environments in terms of system, user, and context IFs. However, in order to gain better understanding of QoE modeling in wireless environments, additional aspects that need to be considered and stressed in addition to common ones have been listed and classified according to the categorization in [22] in Table 2.

Beginning with the environment itself, we address reliability and variability. Wireless channels are more prone to errors than fixed networks because of exposure to various physical phenomena such as noise, fading, or interference. This leads to packet losses, as well as to excessive and variable delays, which consequently affect metrics such as Round Trip Time (RTT), Server Response Time (SRT) or throughput, and integrity and fluency of transmitted data. The wireless infrastructure has been marked as an air bottleneck in data transmission between the user device and the gateway due to several other features such as wireless capacity in terms of speed, coverage radius, or limited bandwidth, and channel sharing with other users or signal strength which may be affected by temperature, humidity, distance from the antenna, and so forth. Also, regarding the wireless channels' reliability, one must consider security issues and interceptions. In addition, the usage of the multitude of wireless access technologies differing in their characteristics (e.g., bandwidth, capacity and coverage constraints, congestion mechanism) also influences the wireless network performance in many ways and thereby the QoE as well.

In contrast to a fixed environment, mobility (horizontal handover) as well as the freedom of switching between various available wireless access technologies, that is, migration of communication from one network to another (vertical handover), leads to another factor affecting QoE—session establishment delay. Namely, during session establishment, a mobile user passes several steps. Firstly, the user has to wait for the security procedures to be performed in order to be granted access to the network. The user then additionally waits while the signaling procedures are completed in order to establish the session. Therefore, in order to initially establish the session or reestablish one due to interruption caused by handover, these procedures have to be performed [21]. With high users' mobility rate, signaling procedures are performed more frequently, increasing the amount of the signaling traffic. This affects the overall usage of wireless resources and increases the session establishment delay, leading to a negative impact on QoE.

However, the increased amount of signaling traffic exchanged in session set up, modification, or tear down procedure does not only affect radio and signaling resources, but also affects device performance. For example, modern mobile phones have an impressive repertoire of functions and features and support applications that require constant connection with the network [62]. The connectivity is maintained by frequently exchanging signaling traffic, which may dominate in comparison with data traffic. The consequence is the overload of computational resources on the mobile device and faster battery consumption. These factors have been considered neither in the case of laptops where the signaling traffic is not generated that frequently and batteries are bigger and allow longer connection maintenance, nor in the QoE modeling for fixed environments where battery consumption is not an issue. Additionally, battery consumption is not only linked with mobile-device interaction with the network, but also to user-device interaction. Therefore, one may conclude that battery lifetime and its consumption are major factors that need to be considered when modeling QoE in the wireless context.

In addition to battery consumption, a number of mobile device features impact QoE. The size of the mobile device screen, as well as position and location of the keys on the screen, may cause difficulties with resizing or scrolling. The small keyboard can impact overall usability and lead to aggravation when typing. Furthermore, as stated in [2, 35], end-user perceived quality may be affected by a lack of "features," such as flash player, personalized alarm clock, features for

Table 2: Aspects to be considered and stressed when modeling QoE in a wireless context.

Category	Aspects impacting QoE to be considered and stressed in the wireless context	
	Network	Device
System	(i) Physical phenomena of wireless channel—variability (affected by noise, fading, and interference)	(i) Signal strength (terminal antenna gain, terminal receiver sensitivity)
	(ii) Wireless channel reliability (interception, security issues)	(ii) Battery lifetime—energy consumption
	(iii) Wireless capacity (speed, coverage, limited bandwidth, shared resources)	(iii) Computational power/resources, storage capacity, processor capability
	(iv) Channel sharing among users	(iv) Screen and keyboard size
	(v) Signal strength (affected by the temperature, humidity, distance from the antenna, base station antenna gain)	(v) Signaling traffic overload
	(vi) Signaling traffic overload	
	(vii) Handover delay	
	Media and Content	
	(i) Adaptation capabilities (e.g., capability to adapt various application parameters to fit the device, network, usage context constraints)	
	(ii) Usability of mobile device	
	(iii) Adjustment to the device power consumption	
	(iv) Data access	
	(v) Offline capabilities	
	(vi) Transparent synchronization with backend systems	
	(vii) Security issues	
	(viii) Lack of add-ons	
User and context	User routine and lifestyle	
	The impact of multiple contexts on user's perception (e.g., mobility, time of day, noisy environment, prior experience)	

privacy settings, Global Positioning System (GPS), and built-in dictionary.

In the context of achieving high QoE, mobile application developers should consider all usage scenarios and address various challenges. Application adaptation capabilities (dynamic or static) are important, in terms of adapting service/application content to fit the device and access network capabilities. Besides usability which is mostly considered in all QoE models, applications should be adjusted to device computational power [63]. Various means of data access, security issues, and offline capabilities as well as transparent synchronization with the backend systems also must be addressed when considering mobile applications.

Finally, user behavior in the wireless context is different as compared to fixed environments. Users are able to access services via various available wireless technologies and different mobile devices, which expose them to dynamic environments. Although addressed in most existing QoE modeling approaches, it is particularly important to address the various usage contexts in wireless environments, since they change the users' perceived quality greatly. The authors in [2] have recognized these important user- and context-related aspects in mobile environments and summarized them in the user routine and lifestyle.

2.5. Summary of QoE Modeling Challenges. There are a number of challenges related to the topic of QoE modeling. Firstly, there is a need to identify a long list of various factors affecting QoE for a given type of service. Secondly, well-planned extensive subjective studies need to be conducted involving human quality perception (including both cognitive and behavioral modeling) in order to model the relationship between identified IFs and (multiple dimensions of) QoE. Some of the main aspects to be considered when planning subjective tests include specification of the methodology to be used, identification of the dependant and independent variables to be considered, user test subjects, testing scenarios, testing environment, and rating scales. Test results analysis leads to identification of the IFs with the most significant impact on QoE and enables the derivation of key QoE IFs. The identification of key QoE IFs and their quality thresholds provide input for relevant QoE optimization strategies. Thirdly, general QoE models should be generic and designed in an elastic way so as to account for fast technology and service advances in converged wireless networks.

While standards specify subjective testing methodologies for multimedia services such as audio, video, and audio-visual services [49, 50, 64, 65], new methodologies are currently being studied for emerging services such as Web

and cloud-based services [66–70]. In addition, this section contributes by summarizing the QoE modeling challenges in the wireless context.

3. QoE Measurement and Modeling

As previously stated, QoE is a multidimensional concept which is difficult not only to define in a simple and unified manner, but also to monitor and measure, considering the large number of QoE IFs to be considered. In order to provide accurate QoE assessment, consideration of only one or two QoE IFs is generally not sufficient. On the contrary, QoE should be considered in all its dimensions taking into account as many IFs as possible (and relevant). Knowledge of the key IFs related to a given type of service drawn from QoE models provides input for QoE monitoring purposes.

The QoE monitoring and measurement process encompasses the acquisition of data related to the network environment and conditions, terminal capabilities, user, context, and application/service specific information and its quantification [24]. The parameters can be gathered via probes at different points in the communication system, at different moments, as well as by various methods. A diversity of QoE monitoring and measurement points, moments, and methods together with the selection of the key QoE IFs for a given service additionally increases the complexity of this process.

In order to be able to manage and optimize QoE, knowledge regarding the root cause of unsatisfactory QoE levels or QoE degradations is necessary. As noted by Batteram et al. [71] and also by Reichl. et al. [30], a layered approach relates network-level Key Performance Indicators (KPIs, for example, delay, loss, throughput, etc.) with user-level application specific Key Quality Indicators (KQIs, for example, service availability, usability, reliability, etc.), which then provide input for a QoE estimation model. Additional input to a QoE estimation model may then be provided by user-, context-, and device-related IFs. Knowledge regarding this mapping between KPIs and KQIs (or what we have referred to as quality dimensions) will provide valuable input regarding the analysis of the root causes of QoE degradation. Hence, monitoring probes inserted at different points along the service delivery chain to collect data regarding relevant KPIs are necessary.

When discussing monitoring points, we may roughly distinguish between *network-based* probes and *client side* probes (note that measurements in both cases may be conducted at different layers of the protocol stack). At the client side, we may further distinguish between probes that collect end-user-related data (e.g., objective measures such as user mouse clicks, or data such as user demographics, user motivation, etc.), context data, device-related data, application data, and network traffic data. While monitoring at the client side provides the best insight into the service quality that users actually perceive, a challenge lies in providing QoE information feedback to the network, service/application, content, or cloud provider to adapt, control, and optimize the QoE. As noted by Hoßfeld et al. [24], this client side monitoring point poses the issues of users' privacy, trust, and integrity, since users may cheat in order to receive better performance. Consequently, collecting data from within the network without conducting client side monitoring (in an either objective or subjective manner), and vice versa, will not generally provide sufficient insight into QoE. Hence, accurate monitoring of QoE needs to employ both: monitoring from within the network and at the client side.

Soldani et al. [72] have used the same conclusions for the monitoring and measurement of QoE, specifically in mobile networks, that is, the need for complementary application of QoE monitoring and measurement methods. Two approaches were proposed: (1) a service-level approach using statistical samples; and (2) a network management system approach using QoS parameters. The former one uses application-level performance indicators and provides the real user opinion towards the used service, while the latter maps QoS performance metrics from various parts of the network onto user-perceptible QoE performance targets. Several similar QoE measurement approaches were standardized by 3GPP in particularly for Real-Time Protocol (RTP) based streaming, Hypertext Transfer Protocol (HTTP) streaming [73], Dynamic Adaptive Streaming over HTTP (DASH), progressive download [74], and Multimedia Telephony (MMtel) [75] for 3GPP devices. Namely, the quality of mobile media is usually degraded due to issues that arise in the last wireless hop, and, consequently, network oriented QoE assessment may not be very reliable. Therefore, it has been reported that the best way to obtain an accurate QoE assessment is to monitor and measure it in the mobile device and report it back to the system [76]. Reported QoE data is combined with other network collected measurements and facilitates the identification of the root causes of quality degradation.

Regarding timing, QoE measurements may be conducted (1) before the service is developed, which includes the consideration of individual quality factors as well as quality planning; (2) after the service is developed, but not delivered; (3) during/after service delivery, which comprises quality monitoring within the network and at the end user side during/after service usage. It has been noted that in the context of closed loop adaptation, there is a growing demand for suitable objective (rather than subjective) QoE evaluation techniques to facilitate optimal use of available wireless resources [23].

As discussed in [71], there are primarily three techniques prevalent in the market today for measuring performance: (1) using test packets; (2) using probes in network elements and user equipment; and (3) using the measurement combination from several network elements. Various approaches involving passive measurements have been reported, based on analyzing the correlations between traffic characteristics and performance criteria [28, 77]. Conducting passive measurements is often cheap and may be used for the evaluation of new applications. However, using network QoS measures for QoE estimation generally implies discerning individual media streams, hence putting additional effort on the monitoring process (involving packet filtering and stream reconstruction) [78].

It is a great challenge today to find a consensus regarding QoE measuring practices. On one hand, QoE has been

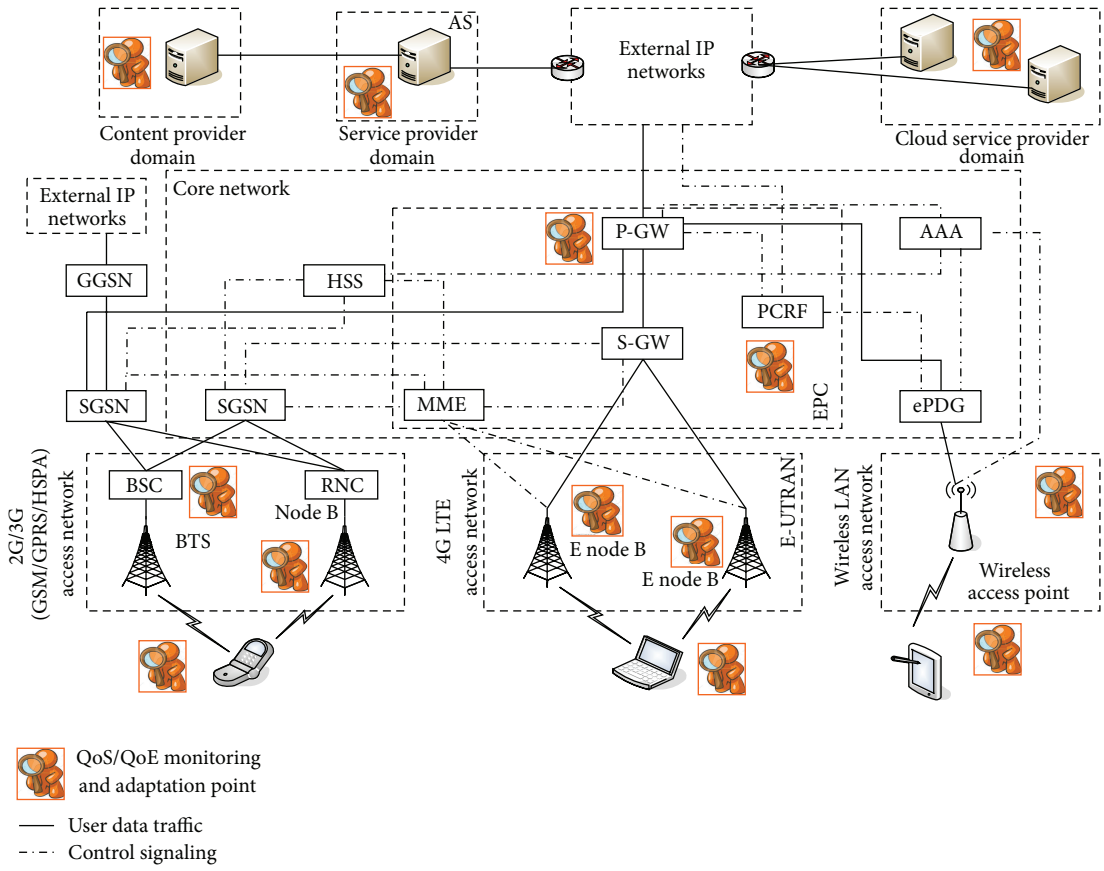


Figure 2: QoE/QoS monitoring and adaptation points in a wireless network environment (see Abbreviations).

mainly measured in terms of technical metrics, since it is often interpreted in terms of QoS. This measurement and assessment approach is criticized when stressing the multidimensional character of QoE [20, 54, 56, 79–81]. On the other hand, measuring the subjective dimensions of the experience is often skipped or neglected because of shorter product/service life cycles, time pressure, budgetary reasons, or simply because they are ignored.

3.1. A Survey of QoE Monitoring and Measurement Approaches in Wireless Networks. Figure 2 illustrates possible QoE/QoS monitoring and adaptation points in a wireless network environment in the context of the 3GPP EPS. The EPS supports multiple access networks and mobility between them via a converged all-IP core network referred to as the previously mentioned EPC [1]. The figure portrays a simplified architecture combining 2G/3G access networks, non-3GPP radio access networks, and the 3GPP LTE access network [6, 82, 83]. As shown, QoE/QoS-related data may be collected from within the network, at the client side, or both. The QoE monitoring and measurement within the network may include data collection at different points such as the base stations within the various access networks, the gateways or routers within the core network, or the servers in the service/application, content, or cloud domains. The acquired parameters may be derived from application level (e.g., content resolution, frame rate, codec type, media type), network level (e.g., packet loss, delay, jitter, throughput), or a combination thereof, that is, in the cross-layer fashion. This approach of enabling QoS at different layers is required due to the fact that existing QoS support in wireless access technologies (e.g., WiMAX or LTE) focuses only on the access network [84]. Traffic collection closer to the end user will provide input for a more accurate estimate of QoE, as discussed also in [85]. Furthermore, the amount of data to process is greatly reduced as compared to data collected in the core network. While QoS solutions commonly use network egress routers for conducting traffic analysis, there is a need to consider computational load.

The remainder of this subsection provides a discussion of several QoE measurement and monitoring studies that have been conducted with various types of services. The approaches are categorized as focusing on client side or network measurements, or their combination. A summary given in Table 3 compares these approaches based on the QoE monitoring point, QoE estimation point, method of conducting the QoE measurement, metrics (subjective or objective), type of service, and QoE management applicability, as well as deployment challenges in the context of QoE management. We note that while subjective measurements are generally more applicable in the context of building QoE models (or validating monitoring approaches), objective measurements are generally employed for QoE estimation and subsequently optimization purposes.

Table 3: Comparison of QoE monitoring and measurement approaches.

Approach	QoE monitoring point	QoE estimation point	QoE management applicability	Measurements environment	Subjective/Objective	Type of service	Deployment challenges
3GPP 26.247 [74]	End user device (technical data)	Network	(i) Dynamic adaptation of service delivery to meet access network capabilities (ii) Applicable in the context of optimization of limited network resource management	Real	Objective	Dynamic adaptive streaming over HTTP	(i) QoE service differentiation and prioritization (ii) Battery consumption (iii) Scalability (iv) Computational complexity (v) Data integrity (vi) User's privacy issues
Ketykó et al. [88]	End user device (technical and user data)	End user device	Applicable in the context of application design improvement (better understanding of content effect)	Semi-real life	Both	Mobile YouTube video streaming	(i) Scalability (ii) User's fairness/correctness (iii) Feedback the estimation output to optimize the network (iv) Data integrity
Wac et al. [35]	End user device (technical, user, and context data)	End user device	Applicable in the context of application design improvement	Real	Both	Wide range of mobile applications	(i) User's fairness/correctness (ii) User's privacy issues (iii) Computational complexity (iv) Feedback the estimation output to optimize the network
Volk et al. [59]	Network	Network	(i) Applicable in the context of service estimation and optimization (ii) Applicable in the context of the network resource allocation optimization	Real	Objective	SDE	(i) User subjectivity (ii) Computational complexity
Varela and Laulajainen [91]	End user device	End user device	(i) Applicable in the context of service estimation and optimization (ii) Applicable in the context of QoE-driven access network selection	Laboratory	Both	Mobile VoIP	(i) Tighter integration with the MIP software (ii) Cost and battery consumption (iii) User's fairness
Hoßfeld et al. [96]	End user device and network	Network	Applicable in the context of network resource optimization	Laboratory	Objective	YouTube video streaming	(i) User's privacy and subjectivity (ii) Limitation of scalability (iii) Additional costs due to DPI

Table 3: Continued.

Approach	QoE monitoring point	QoE estimation point	QoE management applicability	Measurements environment	Subjective/Objective	Type of service	Deployment challenges
Menkovski et al. [38]	Network	End user device	(i) Applicable in the context of network resource management (ii) Applicable in the context of content encoding management	Laboratory	Both	Commercial mobile TV	(i) Inability to give information on service perception (ii) Computational complexity
Staehle et al. [97]	End user device (technical data) and network	Network	(i) Applicable in the context of radio resource management (ii) Applicable in the context of service degradation	Laboratory	Objective	YouTube video streaming	(i) Cross-layer information extraction (ii) User's privacy issues (iii) Battery consumption
Ketykó et al. [98]	End user device (technical and user data) and network	End user device	Applicable in the context of service estimation and optimization	Semi-real life	Both	Mobile video streaming	(i) Scalability (ii) User's fairness (iii) Feedback the estimation output to the network optimization (iv) Battery consumption

3GPP: 3rd Generation Partnership Project, DPI: Deep Packet Inspection, HTTP: Hypertext Transfer Protocol, MIP: Mobile Internet Protocol, MMORPG: Massively Multiplayer Online Role-Playing Game, QoE: Quality of Experience, QoS: Quality of service, RRM: Radio Resource Management, SDE: Service Delivery Environment, VoIP: Voice over Internet Protocol.

3.1.1. QoE Monitoring at the Client Side. While 3GPP policy and QoS mechanisms are based on centralized control, there have been complementary efforts to move certain intelligence from the network to the client. In the context of DASH services (Dynamic Adaptive Streaming over HTTP), 3GPP and MPEG standardization bodies have standardized mechanisms for activating QoE measurements at the client device, as well as the protocols and formats for the delivery of QoE reports to the network servers [74]. It is important to note that HTTP adaptive streaming in general provides the client with the ability to fully control the streaming session. This methodology is mostly suitable for mobile wireless environments and proposed as an optional feature on client devices. The QoE monitoring and reporting framework as standardized by 3GPP is composed of the following phases: (1) a server activates QoE reporting, requests a set of QoE metrics to be reported, and configures the QoE reporting framework; (2) a client monitors or measures the requested QoE metrics according to the QoE configuration; and (3) the client sends the QoE report to the network server in Extensible Markup Language (XML) format by using HTTP [86]. In the context of 3GPP LTE systems, it is important to devise and adopt new QoS delivery and service adaption methods targeting DASH services, since they are beneficial in the sense of optimal management of limited network resources and improved QoE provisioning to the end user [87]. The benefits of adaptive streaming have in particular been recognized in the case of high bandwidth-consuming mobile video communications. Figure 3 depicts a possible example of PCC architecture performing end-to-end QoS/QoE delivery for DASH services. As noted in [87], the current 3GPP PCC architecture supports only QoS delivery and service adaptation for RTSP-based adaptive streaming services, with the need for new methods for HTTP adaptive streaming services.

Other work has addressed concrete cases of collecting QoE relevant metrics at the client side, albeit primarily for QoE modeling purposes and not considering QoE reporting mechanisms providing feedback to the network. In their studies of mobile video streaming, Ketykó et al. [88] have introduced an implementation of a QoE measurement approach on the Android platform based on the collection of both objective and subjective parameters. Observed objective parameters are logged by a QoS and context monitor component deployed on an Android device node and include audio and video jitter and packet loss rate, as well as percentage of duration of connection to a specific data network type in relation to the total duration of a video watching session. In addition, observed end user subjective parameters are logged by an *Experience Monitor* component also deployed on the Android device and include test users' ratings of content, picture and sound quality, fluidness, matching to interests, and loading speed. Similar to the procedures in [88], Verdejo et al. [89] discuss the Android-based QoE measurement framework as applied in the context of playing a mobile location-based real-time massively multiplayer online role-playing game (MMORPG). Users' evaluations regarding the feelings of amusement, absorption, or engagement experienced while playing the game are taken into account and related to a set of objective QoS-related parameters, contextual data, and physiological data obtained from an on-body sensor.

Previously described measurement approaches may contribute to the overall QoE management process in the context of improving the application design, that is, better understanding of content and physical effects. However, if applying such measurement techniques, for example, for optimizing network performance, the challenge lies in reporting QoE feedback obtained at the client side back to the network. Other potential deployment challenges are user related. As previously mentioned, if a user is providing QoE related feedback, they may cheat to improve their performance. Finally, user's privacy may be an issue when it comes to behavioral monitoring.

Besides the previously described subjective data collection methods ESM and DRM, Wac et al. [35] have also addressed technical aspects of QoE in their measurement approach involving a real-life four-week-long study. They have developed an Android Context Sensing Software (CSS) application that unobtrusively collects context and QoS data from users' Android phones. Gathered context data includes current time and user's geographical location, wireless access network technology, cell-ID or an access point name, Received Signal Strength Indication (RSSI), and current used applications with total amounts of application throughput) while the measured network parameter is Round Trip Time (RTT) for an application-level control message sent every minute from a mobile device through the available wireless access network to a dedicated server. As a result of the study, the authors identified a number of QoE IFs for mobile applications, such as user's routine, prior experience, and the possibility to choose between a PC and their mobile device. This approach contributes to the QoE management process in the context of mobile applications design improvement but might experience the same deployment challenges as the two previously described approaches.

With regards to collecting user feedback, a framework proposed by Chen et al. [90] quantifies the users' quality perceptions by having users click a dedicated button whenever he/she feels dissatisfied with the quality of the used application. Hence, the framework is called OneClick and has been demonstrated through user evaluations of different multimedia content in variable network conditions.

Finally, we mention the applicability of client side monitoring in the context of QoE-driven mobility management. Focusing on voice services, Varela and Laulajainen [91] describe QoE estimations for VoIP to improve the existing network-level IP mobility management solutions. The proposed solution performs QoE estimations by passive network QoS monitoring for VoIP traffic, feeding the network's QoS information to a Pseudo-Subjective Quality Assessment (PSQA) tool [92]. The aim is to aid in making access network handover decisions. A presented prototype implementation is tested in scenarios representing real VoIP service usage. We note that VoIP QoE estimation and prediction based on passive probing mechanisms and integrated directly into a mobility management protocol is further addressed by Mitra et al. [93]. Furthermore, a user-centric approach to

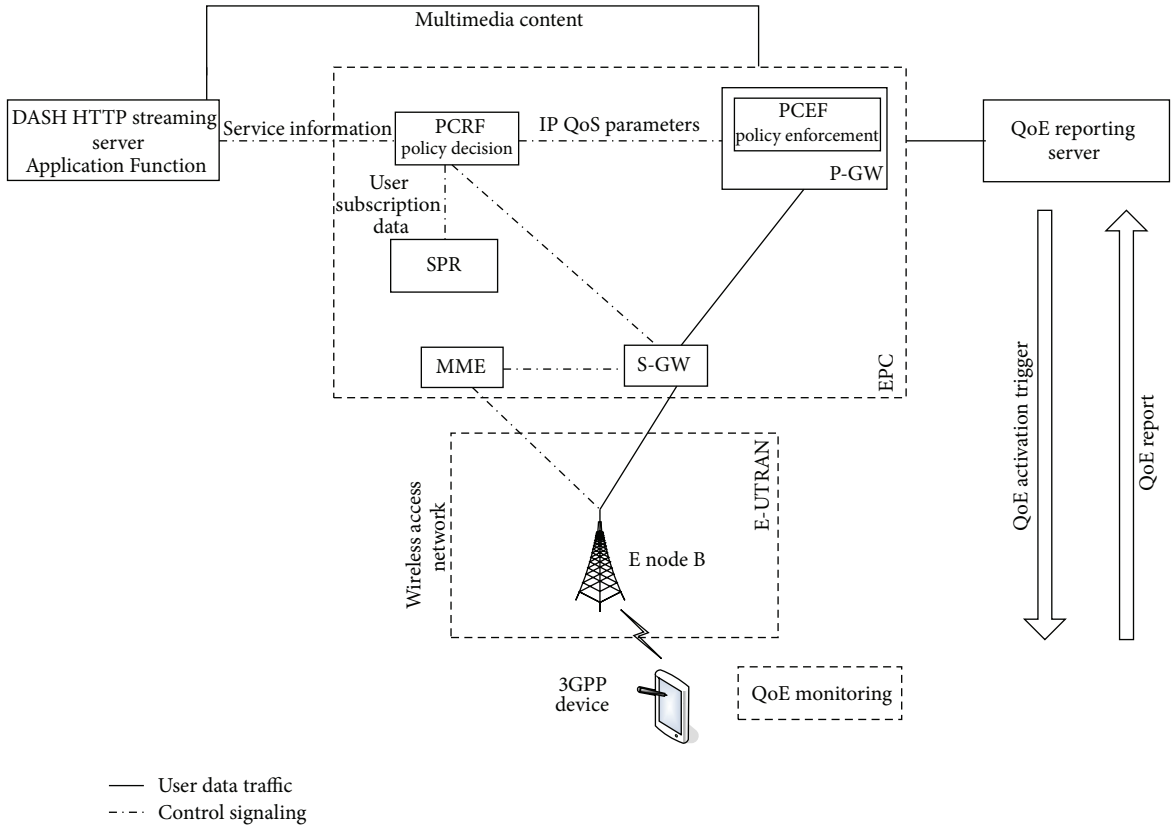


Figure 3: Example of possible PCC architecture performing end-to-end QoS/QoE delivery for DASH services (see Abbreviations).

proposing seamless mobility management solutions was one of the key focus areas of the EU FP7 PERIMITER project [94] (available deliverables provide detailed insight into user-centric mobility design).

3.1.2. QoE Monitoring in the Network. Volk et al. [59], whose approach is described in Section 2 in the context of QoE modeling, have focused on an NGN and IP Multimedia Subsystem (IMS) [95] based environment and proposed a solution for QoE assurance which employs an automated proactive in-service algorithm for the user's QoE estimation, rather than relying on regular QoS techniques or acquiring the subjective end-user's feedback. The QoE estimation algorithm is based on context-aware objective end-to-end QoE modeling and is run by a dedicated application server, implemented as a value-added service enabler. The authors argue reasons for conducting centralized QoE estimations in the network as being availability of and access to a wide range of quality-related information, the possibility of non-intrusive in-service QoE estimation, and the potential of proactive in-service quality assurance functionality (e.g., the application server may invoke adaptation/modification of certain quality affecting parameters). While the authors argue that this approach ensures fair interpretation of subjectively perceived quality towards any end user or service, and operational efficiency in terms of no end-user involvement (guaranteeing universality of the QoE estimation), the notion of user singularity may be considered neglected.

Hoßfeld et al. [96] compare two YouTube QoE monitoring approaches operating at the end user level (described in the following heading) and within the network. A novel YouTube in-network (YiN) monitoring tool is proposed as a passive network monitoring tool. This approach aims at detecting and measuring stalling of the video playback by approximating the video buffer status by comparing the playback times of video frames and the time stamps of received packets. The challenges of this approach are related to the accurate reconstruction of the stalling events that arrive at the application layer which requires additional costs and limits the scalability in terms of the number of YouTube video streaming flows that can be actually monitored by a probe.

Menkovski et al. [38] have presented a method for assessing QoE and developed a platform for conducting the QoE estimation for a service provider. The designed platform estimates QoE of the mobile TV services based on existing QoS monitoring data together with QoE prediction models. The prediction models are built using Machine Learning (ML) techniques from subjective data acquired by limited initial subjective user measurements. Thereby, subjective measurement associated complexities are minimized, while the accuracy of the method is maintained. This platform is

currently used for mobile TV systems where it estimates the QoE of the streaming media content and is further used to manage the services and resources. However, the deployed system cannot give any information as to how the service is perceived by the end user.

3.1.3. QoE Monitoring Combining Client Side and Network Measurements. Further focusing on YouTube as one of the most common and traffic intensive mobile applications, Staehle et al. [97] have proposed a previously mentioned YouTube Application Comfort (AC) monitoring tool—*YoMo*, which monitors the QoE at the client's side. The tool detects the YouTube video and determines its buffered playtime. Thereby, the YoMo tool is able to detect an imminent QoE degradation, that is, stalling of the video. The interruption of the video playback is the only considered factor influencing YouTube QoE, as it has been argued that this is the key IF in the given case. Additionally, the tool communicates the stalling information to the network advisor and raises an alarm if the AC becomes bad. What makes YoMo particularly suitable for QoE monitoring and measurement is its ability to predict the time of stalling in advance. Thus, it allows the network operator to react prior to the QoE degradation and to avoid unsatisfied customers.

Ketykó et al. [98] have introduced a measurement concept of QoE related to mobile video streaming in a 3G network environment and semi-real-life context. The data collection, which is based on the Experience Sampling Method, combines objective and subjective data for evaluating user experiences. Observed technical-quality-related QoE parameters, audio and video packet loss and jitter, are obtained at the server-side from the Real-time Transport Control Protocol (RTCP) Receiver Records (RR), while the observed RSSI parameter is obtained from the *MyExperience* in situ measurement tool used at the client side. The subjective assessments parameters have been conducted in two phases: (1) preusage questionnaire (obtaining users' experiences towards mobile applications) and (2) usage phase where users are asked to use the mobile application in six different usage contexts: indoor and outdoor, at home, at work, and on a train/bus. This study has shown that QoE of mobile video streaming is influenced by the QoS and by the context. Additionally, the authors have proposed linear functions for modeling the technical-quality-related QoE aspects and argued that spatial quality and emotional satisfaction are the most relevant QoE aspects for the tested users.

Having surveyed several chosen QoE measurement approaches, we have classified them into ones that aim to perform QoE monitoring by acquiring data only at the client side [35, 74, 88–90], only within the network [28, 38, 59, 91, 96], or by collecting data at both, the client side and within the network [97, 98]. As stated previously, in order to assure accurate QoE estimation and identification of the causes of QoE degradation, measurements collected along the end-to-end service delivery path are needed. The majority of approaches comprise both subjective and objective parameters with end users estimating QoE. While the collection of subjective assessments is generally conducted in the scope of empirical QoE studies targeted towards building accurate QoE models, objective measurements provide input for QoE prediction mechanisms and are commonly employed for QoE optimization and control purposes. Furthermore, in terms of the measurements environments that have been discussed, several approaches have been illustrated in a laboratory testbed [28, 38, 91, 96, 97], while others were demonstrated in real-life [35, 59, 74, 90] or semi-real-life environments [88, 89, 98].

As it can be observed from the previous analysis, QoE monitoring approaches in the wireless context often measure parameters such as packet loss rate, bandwidth, throughput, delay, and jitter and do not in general differ from ones addressing a fixed environment. However, in order to gain a deeper understanding of QoE IFs in wireless and mobile environments, there is a need to monitor parameters characteristic for such environments (Table 2). For example, RSSI measurements can be utilized for addressing wireless factors such as channel exposure to physical phenomenon, its capacity and sharing among the users, signal strength, terminal antenna gain. Additionally, in order to gain the accuracy regarding the wireless channels, this measurement can be combined with measurements of base station and terminal antenna gain, distance from the antenna, temperature, and so forth. The information obtained by combining the aforementioned measurements can give a clearer picture of which factors impact QoE and to what extent. Another example is the measurement of the radio cell reselection frequency or signaling update frequency which may reveal how these wireless and mobile specific issues can be optimized. Additionally, although not considered in the analysis, it is recommended to measure the mobile device power consumption while using different applications, processor capability, and storage capacity.

3.2. Summary of QoE Monitoring and Measurement Challenges. The previous discussion has shown that the QoE monitoring and measurement process is complex due to the diversity of factors affecting QoE, data acquisition points, and timings, as well as methods of collecting data, and the lack of consensus regarding these issues. The main challenge in this process is to answer the following four questions: (1) *What* to collect?; (2) *Where* to collect?; (3) *When* to collect?; and (4) *How* to collect?

Firstly, one needs to determine *which* data to acquire. The *what/which* clause is specified by the QoE metrics selection which depends on the service type and context. The decision regarding data that should be acquired considering the wide spectrum of QoE IFs is challenging, but it is the prerequisite for any QoE monitoring and measurement approach. Secondly, choosing a location *where* to collect data is another critical issue in the QoE assessment process, that is, determine the location of monitoring probes. As previously mentioned, data can be collected within the network, at the client side, or both (depending also on whether measurements are conducted for QoE modeling purposes or for QoE control purposes). The QoE monitoring and measurement within the network may include data collection at different points such as the base stations within the various access networks, the gateways or routers within the core network, or the

servers in the service/application, content, or cloud domains. Additionally, the acquired parameters may be derived from application level, network level, or a combination thereof. Each acquisition location addresses the specific challenges discussed previously. Furthermore, if performing in-service QoE management (e.g., QoE-driven dynamic (re)allocation of network resources), collected data generally needs to be communicated to an entity performing QoE optimization decisions. Hence, the passing of data to a control entity needs to be addressed. Thirdly, one should determine *when* to collect data: (1) before the service is developed; (2) after the service is developed, but not delivered; and (3) after the service is delivered. Additionally, *how often* data should be monitored and measured needs to be considered. Finally, *how* to perform the data acquisition is determined by the *where* and *when* clauses. The QoE monitoring process implies computational operations, hence computational complexity and battery life of mobile devices need to be considered.

It may be concluded, as in the QoE modeling process, that different actors involved in the service provisioning chain will monitor and measure QoE in different ways, focusing on those parameters over which a given actor has control (e.g., a network provider will monitor how QoS-related performance parameters will impact QoE, a device manufacturer will monitor device-related performance issues, while application developers will be interested in how the service design or usability will affect QoE).

Having chosen the proper QoE metrics and monitoring and measurements approach, it is important to provide mechanisms utilizing this information for improving service performance, network planning, optimization of network resources, specification of service level agreements (SLAs) among operators, and so forth. Such issues are addressed as the "final step" in the QoE management process, discussed in the following section.

4. QoE Optimization and Control

Following QoE modeling, monitoring, and measurements, the ultimate goal of QoE management is to control QoE via QoE optimization and control mechanisms. Such mechanisms yield optimized service delivery with (potentially) continuous and dynamic delivery control in order to maximize the end-user's satisfaction and optimally utilize limited system resources. From an operator point of view, the goal would be to maintain satisfied end users (in terms of their achieved QoE) in order to limit customer churn, while efficiently allocating available wireless network resources. QoE optimization as such may be considered a very challenging task due to a number of issues characteristic for converged all-IP wireless environments, including limited bandwidth and its variability, the growth of mobile data, the heterogeneity of mobile devices and services, the diversity of usage contexts, and challenging users' requirements and expectations, as well as the strive to achieve cost efficiency.

4.1. An Overview of QoE Optimization Approaches in Wireless Environment. A number of strategies for optimizing QoE in a wireless environment that have been proposed differ in the applied approach (network/user oriented), parameters chosen to be adjusted, control location(s) and timing(s), and so forth. Therefore, in this section, Table 4 will provide an overview of the state-of-the-art approaches in terms of QoE optimization point, optimization strategy, considered wireless technologies, and deployment challenges.

Since QoS (and ultimately QoE) provisioning is a key issue in the context of wireless networks, 3GPP has proposed a set of comprehensive QoS concepts and architectures for UMTS [7, 8]. A Policy and Charging Rules Function (PCRF) included in the EPS as part of the 3GPP PCC architecture [9], which is shown in Figure 4, impacts end-user QoE for a particular subscription and service type by providing service-aware network-based QoS. Service requirements may be extracted from the application-level signaling (e.g., based on the Session Initiation Protocol (SIP)) and passed down to the PCRF, responsible for executing policy rules. Execution of policy rules and their enforcement at the network level serves to manage network congestion, provide differentiated service quality based on heterogeneous service requirements, and create a framework for new business models. The bearer- and class-based QoS concept introduces a QoS Class Identifier (QCI) which specifies standardized packet forwarding treatment for a given traffic flow. The standardized QCI characteristics are specified in terms of bearer type (Guaranteed Bit-Rate (GBR) or non-GBR), priority, Packet Delay Budget (PDB), and Packet Error Loss Rate (PELR) [99]. Apart from nine QCIs, QoS parameters defined in the EPS include Allocation and Retention Priority (ARP), Maximum Bit-Rate (MBR), and GBR.

Skorin-Kapov and Matijašević [100] have proposed QoE-driven service adaptation and optimized network resource allocation mechanisms in the context of the 3GPP IMS and PCC architectures (Figure 5(a)). Service requirements and user preferences are signaled in the form of utility functions and serve as input for an optimization process (conducted by a proposed QoS Matching and Optimization Function) aimed at calculating the optimal service configuration and network resource allocation, given network resource, service, and operator policy constraints. The calculation results are passed to the PCRF node and serve as input for resource allocation mechanisms. Ivešić et al. [101] have further built on this approach by focusing on QoE-driven domain-wide optimal resource allocation among multiple sessions. The resource allocation has been formulated as a multiobjective optimization problem with the objectives of maximizing the total utility of all active sessions along with operator profit in the context of the 3GPP EPS.

Further considering a 3GPP environment, an in-service QoE control mechanism has been proposed by Volk et al. [59] and further studied by Sterle et al. [102] (Figure 5(b)). The proposed application-level QoE estimation function running at the application server in the NGN service stratum is based on collection of a comprehensive set of QoE IFs. The authors attempt to maximize QoE by making the adjustments to identified quality performance indicators. As previously discussed in terms of QoE monitoring and measurement, this approach's benefits include the wide range of quality-related

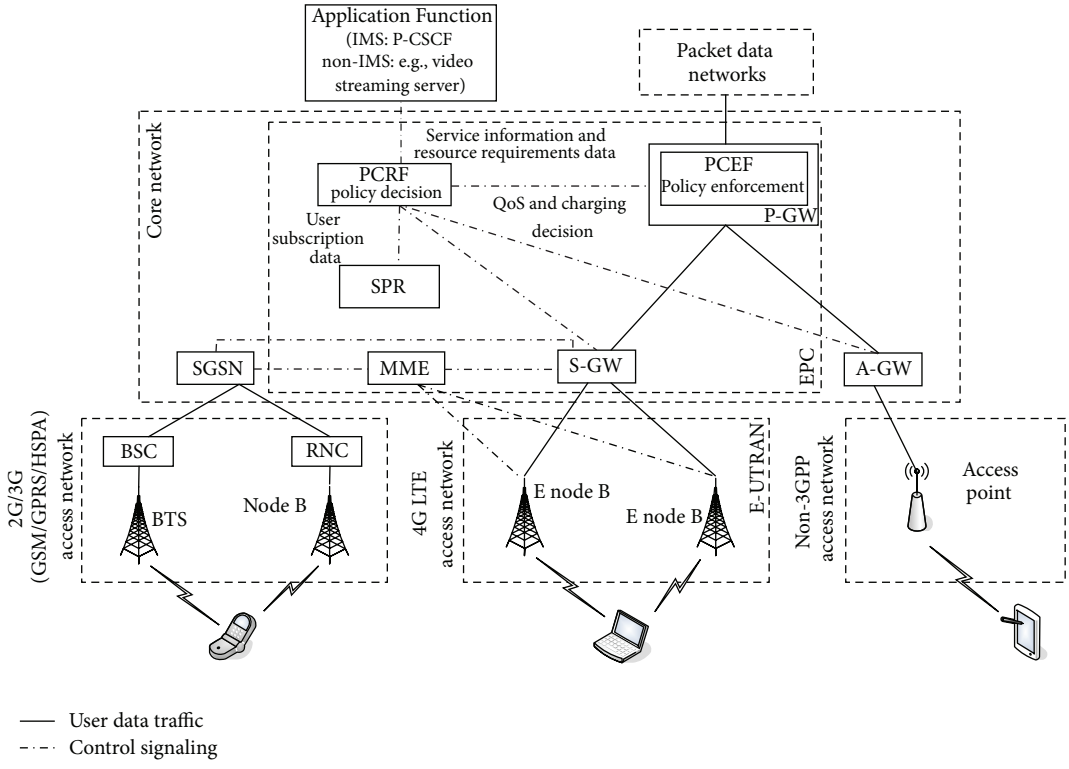


Figure 4: 3GPP PCC architecture for EPS (see Abbreviations).

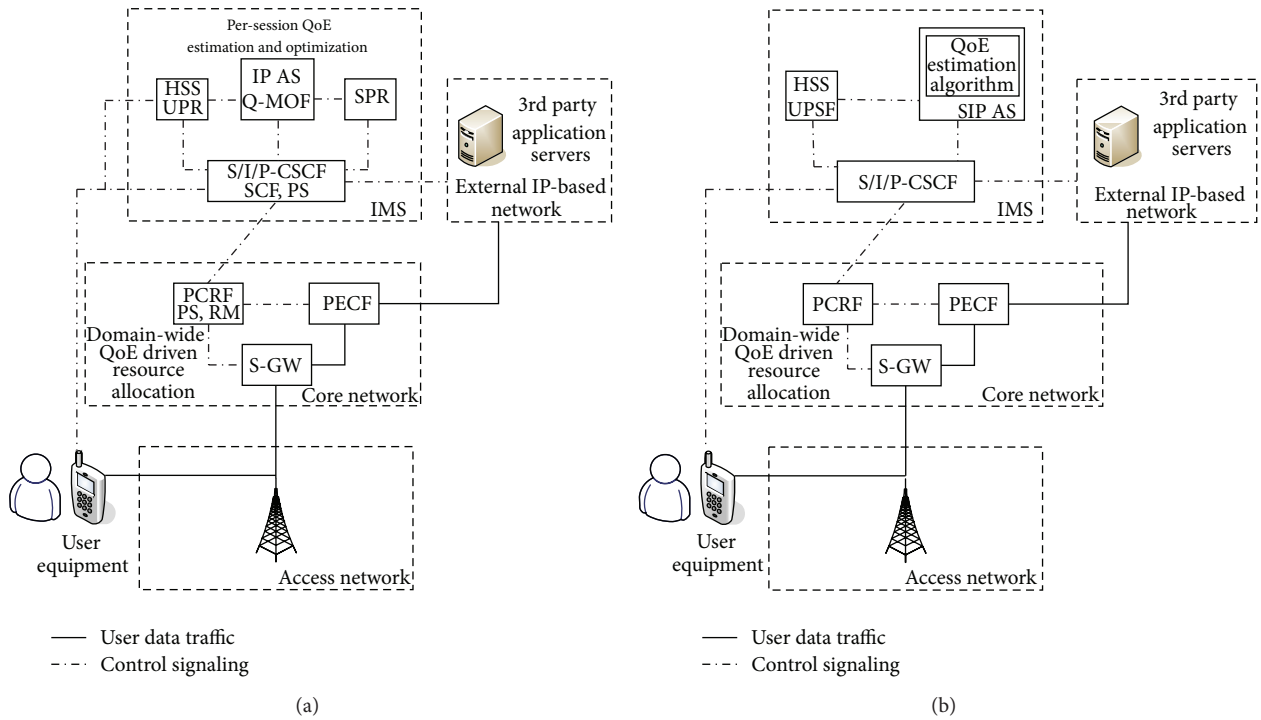


Figure 5: QoE control mechanisms: (a) Skorin-Kapov and Matijašević [100] and Ivešić et al. [101]; (b) Sterle et al. [102] (see Abbreviations).

Table 4: An overview of QoE optimization approaches in wireless environments.

QoE optimization approach	QoE optimization point	QoE optimization strategy	Wireless technologies considered	Deployment challenges
Skorin-Kapov and Matijašević [100]	SIP application server in IMS domain	(i) Mechanisms for service adaptation (ii) Mechanisms for optimal network resource allocation	3GPP access	(i) Computational complexity (ii) Users subjectivity (iii) Signaling overhead (iv) Scalability
Ivešić et al. [101]	SIP Application server in IMS domain and policy engine in core network (e.g., PCRF)	Mechanisms for domain-wide optimal network resource allocation	3GPP access	(i) Computational complexity (ii) Scalability (iii) Users subjectivity
Sterle et al. [102]	SIP application server in IMS domain	Mechanisms for service optimization	WiMAX and UMTS	(i) Computational complexity (ii) Users subjectivity (iii) Scalability (iv) Extendibility
Thakolsri et al. [103]	Core network (cross-layer based optimization)	(i) Mechanisms for optimal radio resource allocation (ii) Mechanisms for rate adaptation	UMTS (HSDPA)	(i) Users subjectivity (ii) Extendibility
Staehle et al. [97]	Core network (cross-layer based optimization)	(i) Mechanisms for optimal radio resource allocation (ii) Mechanisms for service optimization	WLAN	(i) Computational complexity (ii) Users subjectivity (iii) Scalability
Shehada et al. [112]	Core network (cross-layer based optimization)	Mechanisms for optimal network resource allocation	LTE	(i) Extendibility (ii) Users subjectivity
Amram et al. [113]	Core network (cross-layer based optimization)	(i) Mechanisms for optimal network resource allocation (ii) Mechanisms for optimal handover decision	LTE and WLAN	(i) Extendibility (ii) Cost limitations (iii) Optimal CDN node selection (iv) Users subjectivity (v) Scalability (vi) Modification of scheduling algorithm
Aristomenopoulos et al. [109]	Access network (cross-layer based optimization)	(i) Mechanisms for optimal radio resource allocation (ii) Mechanisms for integration of user's subjectivity	CDMA	(i) User's fairness (ii) Extendibility (iii) Scalability
Wamser et al. [110]	Access network (eNodeB)	Mechanisms for service optimization (prioritized traffic scheduling)	LTE	(i) User's fairness (ii) Signaling overhead (iii) Modification of scheduling algorithm
Piamrat et al. [111]	Access network (user- and network-data collection)	Mechanisms for optimized access network selection	WLAN	(i) Computational complexity (ii) Importing intelligence into base stations (iii) Possible signaling overhead (iv) Extendibility (v) User's fairness
Khan et al. [114]	Sender side (preencoding stage over access network)	(i) Mechanisms for service optimization (ii) Mechanisms for rate adaptation	UMTS	(i) Extendibility (ii) Users subjectivity (iii) Modification of SBR

Table 4: Continued.

QoE optimization approach	QoE optimization point	QoE optimization strategy	Wireless technologies considered	Deployment challenges
El Essaili et al. [107]	Distributed on the end user terminal and access network (base station)	(i) Mechanisms for optimal radio resource allocation (ii) Mechanisms for service optimization	LTE	(i) Signaling overhead (ii) Computational complexity (iii) User's privacy
Csernai and Gulyas [115]	End user mobile device	Mechanisms for optimized battery consumption	WLAN	(i) Extendibility (ii) Cost requirements (iii) Modification of scheduling algorithm
Latré et al. [104]	Access network	(i) Mechanism for monitoring the network and building knowledge about it (ii) Mechanisms for analyzing the knowledge and determining QoE actions and enforcing them (iii) Mechanisms for reducing packet loss and switching to different video bit rate to obtain better quality	Not explicitly stated	(i) Scalability (ii) Users subjectivity (iii) Computational complexity
Hassan et al. [106]	Access network	(i) Mechanisms for optimized resource allocation and provider revenue	WLAN (applicable to others)	(i) User's fairness (ii) Scalability (iii) Speech processing

CDMA: Code Division Multiple Access, GPRS: General Packet Radio Service, HSDPA: High-Speed Downlink Packet Access, IMS: IP Multimedia Subsystem, LTE: Long-Term Evolution, QoE: Quality of Experience, RACS: Resource and Admission Control System, SIP: Session Initiation Protocol, UMTS: Universal Mobile Telecommunications System, WiMAX: Worldwide Interoperability for Microwave Access, WLAN: Wireless Access Network.

information sources available in the network and non-intrusive in-service quality assurance and control.

Network-, that is, operator-driven, QoE optimization approaches are primarily concerned with the optimal utilization of available network resources, and in order to maximize QoE, they propose various network resource management mechanisms which rely on the information obtained from the monitoring and measurement process. Therefore, Thakolsri et al. [103] (Figure 6(a)) have applied utility maximization in the context of QoE-driven resource allocation across multiple users accessing different video contents in a wireless network. The proposed scheme allocates network resources and performs rate adaptation such that perceivable quality fluctuations lie within the range of unperceivable changes. Also, the *Aquarema* concept proposed by Staehle et al. [97] (Figure 6(b)) enables application specific network resource management and thereby improves the user QoE in all kinds of networks for all kinds of applications. The authors have achieved the improvement by the interaction of the previously described application comfort monitoring tool—*YoMo*, running at the client side, and a network advisor which may trigger different resource management tools. The tool quantifies how well an application is running and enables prediction of the user experience, thereby allowing the network advisor to act upon an imminent QoE degradation. The principles of these approaches that have placed the resource allocation mechanisms in the core network may be combined. For example, the former one which manages the network resources by prioritizing the users that have better channel conditions and which thereby indirectly assumes the improvement of the overall QoE can be supplemented with the client information obtained from the monitoring tool. Thereby it would gain more information for the fair prioritizing, that is, network resource allocation, and improvement of individual user QoE.

Additionally, Latré et al. [104] (Figure 6(c)) have defined an autonomic management architecture to optimize the QoE in multimedia access networks using a three-plane approach consisting of (1) a *Monitor Plane* which monitors the network and builds up knowledge about it; (2) a *Knowledge Plane* which analyzes the knowledge and determines the ideal QoE actions; and (3) an *Action Plane* that enforces these actions into the network. The authors have focused on the "smart" Knowledge Plane which consists of two reasoners: an analytical one based on a set of equations and the other one based on neural networks. These reasoners can optimize the QoE of video services with two optimizing actions: applying Forward Error Correction (FEC) to reduce the packet loss caused by errors on a link and switching to different video bit rate to avoid congestion or to obtain a better video quality.

In order to efficiently use limited wireless resources and distribute them among users whose perceived quality should be maximized, QoE-driven resource allocation and scheduling mechanisms should incorporate the sensitivity of the human perceived quality [105] which requires a strategy to include the mapping of users' opinions into resource allocation and scheduling algorithms in various wireless access technologies such as Code Division Multiple Access (CDMA), LTE, UMTS, WiMAX, or Wireless Local Area Network (WLAN). For example, Hassan et al. [106] model the

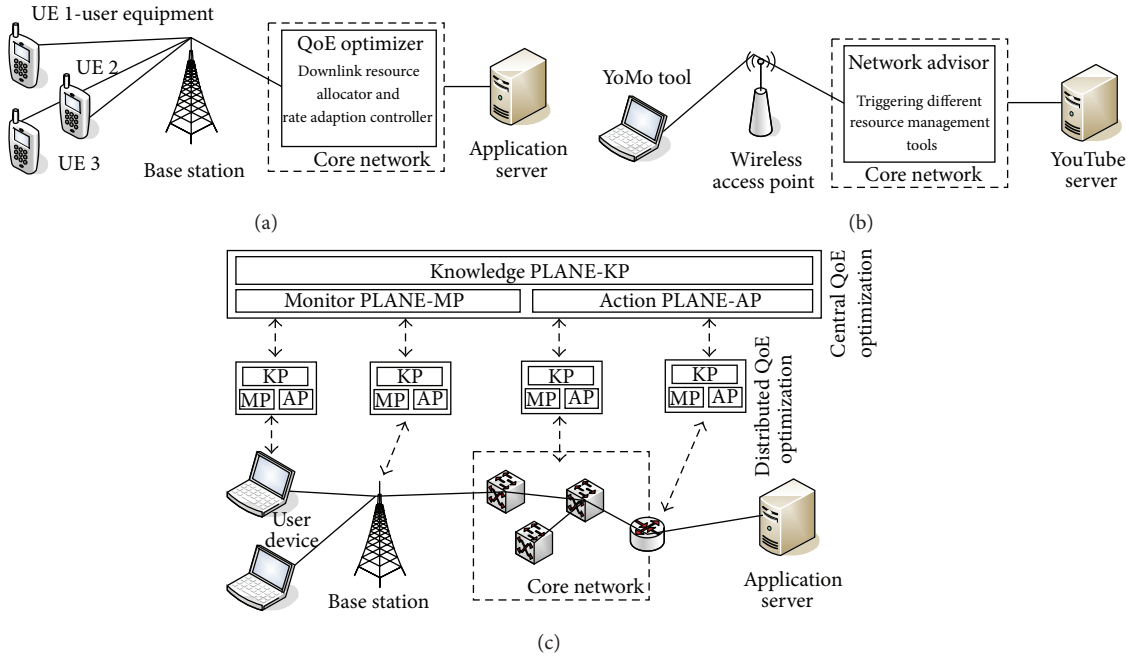


Figure 6: QoE optimization mechanisms: (a) Thakolsri et al. [103]; (b) Staehle et al. [97]; (c) Latré et al. [104].

QoE of mobile VoIP services and use this as input for a QoE management scheme employing a game theoretic approach to resource management.

In this context, it is challenging for network elements responsible for resource management to adapt the constrained uplink and downlink wireless resources by assigning or periodically reassigning them to different service providers and users such that all resource competitors are satisfied. Also, QoE-driven resource management that would result in higher users' satisfaction may be performed by implementing QoE-aware routing and packet controllers which give preferential treatment to certain types of packets, according to priority-based policies that may differ depending on operator's interests. However, in resource variable and constrained systems, such as wireless networks, the priority to gain resources is primarily given to users having good channel condition and accessing low-demand applications that result in his/her satisfaction for a small amount of limited resources [103]. Furthermore, one has to account for users that may be given priority for paying more, although they may not have the above mentioned communication conditions. QoE data may be utilized in order to aid and improve decision making in the context of resource allocation and scheduling [103, 107], mobility management [91], and so forth.

Further focusing on QoE-driven resource allocation, it can be noted that this process may either be adapted to meet different service requirements (e.g., via 3GPP QoS provisioning mechanisms), or services may be adapted to meet dynamic network resource availability (e.g., adaptive streaming based on MPEG DASH). Additionally, it can be considered in terms of QoE optimization for a single user [91] or multiple users [103] differing in maximization of QoE for a given user or total/average QoE for users, as well as for single media flow [108] or multiple media flows [100, 101] with different utility functions corresponding to each media component.

However, comprehensive consideration of the various QoE influence factors and their correlations is needed in order to improve and optimize QoE. Depending on the determination of the appropriate QoE IFs which may need to be adjusted for a given service, QoE optimization may be performed at different locations, as depicted in Figure 2. Thus, as summarized in Table 4, the optimization can be conducted by applying various control mechanisms at the base stations within various access networks [109–111] by applying policy management rules at the gateways or routers within the core network [97, 103, 112, 113], by conducting adjustments at the servers in the service/application [104, 113, 114], content or cloud domains, or the combination thereof [107], as well as on end-user device [115]. Since QoE control relies on QoE monitoring and measurement information, usually the control locations are the same as monitoring and measurement points. This does not necessarily mean that QoE optimization is conducted at the location where the data for its activation is collected, since the data gathered at one point in the system may trigger the optimization at another point. Additionally, the optimization may be performed at levels ranging from link- to application-layer [59, 102], as well as in a cross-layer fashion which is most common [97, 103, 108, 112, 113, 116].

The addressed approaches propose various QoE optimization strategies that are aimed at optimal network resource allocation [100, 101, 112, 113]; optimal radio resource allocation [97, 103, 107, 109]; service optimization [97, 100, 102, 107, 110, 114]; optimized handover decision [113]; access network selection [111]; or battery consumption [115].

However, the overall automatic optimization strategies may successfully manage the finite network resources and fulfill general users' requirements, but these may not always be optimal in terms of an individual user's QoE. It has been argued that automated mechanisms may benefit from the user's manual adjustments of the service settings on his/her own device, although it would affect the network resources. For example, Aristomenopoulos et al. [109] have proposed a dynamic utility adaption framework suitable for real-time multimedia services in a wireless environment, which allows users to express their (dis)satisfaction with the service quality and adjust it at their device. This framework provides the seamless integration of users' subjectivity in network utility-based Radio Resource Management (RRM) mechanisms, enabling cross-layering from the application to the Media Access Control (MAC) layer.

With regards to timing when the QoE optimization should be conducted, one may distinguish between in-service and out-of-service approaches. In-service QoE optimization implies the conduction of the process in the system while it is in operation. In other words, it performs the on-line quality prediction/control during service execution. The user-oriented approaches based on quality-related feedback are in-service, since the events that trigger QoE adaptation are coming from the individual user while she/he is using the service [109, 111]. Also, the network-oriented approaches may be in-service, since the resource management functions can be triggered by events happening during the service (e.g., detection of network congestion, operator policy) [59, 97, 102, 103, 110, 113–115]. On the other hand, out-of-service QoE adaptation is performed in an off-line fashion by implementing the control mechanisms for network planning, load balancing, network congestion detection, and so forth [9, 116].

It may be concluded that in most situations the user perceived QoE will depend on the underlying network performance. However, network-oriented QoE optimization processes would clearly benefit from perceived quality feedback data collected at the user's side, since QoE is inherently user-centric. In addition, as previously mentioned, the network decision making process may benefit from the user's adaptations in terms of more efficient resource usage. Therefore, in order to truly optimize and control QoE, both network- and user-oriented issues should be encompassed.

4.2. Summary of QoE Optimization Challenges. As previously discussed, QoE optimization and control is a challenging task considering numerous constraints. Similarly, as related to QoE monitoring and measuring processes, the main challenges that arise with regards to QoE controlling may be summarized in the answers to the following four questions: (1) *what* to control?; (2) *where* to control?; (3) *when* to control?; and (4) *how* to control?

Answering the question of *what* parameters to optimize relates to the issue of determining the key factors whose adjustments would result in improved QoE. Additionally, the impact of those optimizations on other parameters must be considered, since in certain cases improvements of one set of parameters may result in other parameters degradations (e.g., web browsing a high quality media content may prolong a web page response time). The *what* clause determines another critical issue in the QoE optimization process: the location *where* the optimization will be performed. Thus, the QoE may be optimized at various locations as previously discussed. Furthermore, one needs to determine *when* to perform the QoE optimization: during the service, that is, on-line control or in an off-line fashion. Additionally, it needs to be considered *how often* to optimize QoE. Finally, *how* to optimize QoE is determined by all three previously mentioned clauses.

Basically, there are many strategies for QoE optimization and control, and they can be described as being closer to the network or closer to the user. A promising approach appears to optimize QoE by combining the network- and user-oriented approaches by supplementing the drawbacks of one with the advances of the other. Therefore, by combining different approaches in QoE optimization, multiple stakeholders and players involved in the service delivery process would benefit, since the various additional challenges that each of them pose would be addressed. Additionally, the characteristics of wireless and mobile environments that have been discussed in Table 2 (e.g., the constrained and shared network resources, variable and unstable nature of wireless channels, device diversities and capabilities) will pose additional challenges to cope with when allocating limited resources among users.

Therefore, open research issues include applicability across different wireless access technologies (e.g., LTE, WiMAX, WLAN) and implementation on heterogeneous devices; computational issues in terms of complexity of the mechanism or limited computational capacity of a device which may lead to battery consumption; signaling overhead due to increased amount of signaling traffic exchanged between devices and wireless access points (if the optimization mechanism is distributed and combined) and resulting latency; scalability in terms of resulting time-consuming effect and cost when the solution is applied on a large number of users; and so forth. Since a goal of future research is to consider the user's subjectivity in the QoE optimization procedure, user-related issues such as fairness, trust, or privacy must be addressed appropriately.

5. Conclusion

Satisfying user service quality expectations and requirements in today's user centric and upcoming converged all-IP wireless environment implies the challenge of performing successful QoE management. Requirements put forth by standards including 3GPP, ETSI, and ITU include providing support for access to services anywhere and anytime, over any medium and network technology, and across multiple operator domains. In the context of the highly competitive telecom market, the fast development of new and complex mobile multimedia services delivered via various new mobile devices offers a wide scope of choice for the end user, hence increasingly driving operators to put focus on the QoE

domain. Studies have shown that the management of QoE as a highly complex issue requires an interdisciplinary view from user, technology, context, and business aspects and flexible cooperation between all players and stakeholders involved in the service providing chain.

This paper gives a survey of the state of the art and current research activities focusing on steps comprising the QoE management process: QoE modeling, QoE monitoring and measurement, and QoE adaptation and optimization. Based on the overview, we have identified and discussed the key aspects and challenges that need to be considered when conducting research in the area of QoE management, particularly related to the domain of wireless networks. Challenges related to QoE modeling include the consideration of various QoE IFs and identification of key ones, as well as mapping the key IFs to QoE dimensions. The main challenges related to QoE monitoring and measurement can be summarized in the questions: *what* to collect?, and *where*, *when*, and *how* to collect data? Similarly, the following questions *what* to improve?, and *where*, *when*, and *how* to improve QoE? summarize the challenges of the QoE optimization process. In the context of wireless networks, QoE management has mostly been considered in terms of resource scheduling, whereby resource allocation decisions are driven to optimize end-user QoE. QoE-driven resource management has become an important issue for mobile network operators as a result of the continuously increasing demand for complex and faster multimedia applications that are adaptable to various devices and require increased network resources (e.g., more bandwidth, less delays, higher link quality), as well as the need for satisfying high user expectations towards continuous communication, mobility, and so forth. Therefore, QoE-driven mobility management solutions have also been presented, as well as QoE-driven service adaptation solutions.

Thus, by approaching various QoE aspects from a wide interdisciplinary perspective, this paper aims to provide a better understanding of QoE and the process of its management in converged wireless environments.

Abbreviations

AAA: Authentication, authorization, and accounting
A-GW: Access gateway
AS: Application server
BSC: Base station controller
BTS: Base transceiver station
DASH: Dynamic adaptive streaming over HTTP
E Node B: Evolved node B
EPC: Evolved backet core
ePDG: Enhanced packet data gateway
EPS: Evolved packet system
E-UTRAN: Evolved UMTS terrestrial radio access network
GGSN: Gateway gprs support node
GPRS: General packet radio service
GSM: Global system for mobile communication
HSPA: High speed packet access
HSS: Home subscriber service
HTTP: Hyper text transfer protocol
I-CSCF: Interrogating call session control function
IMS: IP multimedia subsystem
IP: Internet protocol
LAN: Local area network
LTE: Long-term evolution
MME: Mobility management entity
PCC: Policy and charging control
PCEF: Policy and charging enforcement function
PCRF: Policy and charging rules function
P-CSCF: Proxy call session control function
P-GW: Packet data network gateway
PS: Policy server
Q-MOF: QoS matching and optimization function
QoE: Quality of experience
QoS: Quality of service
RM: Resource manager
RNC: Radio network controller
SCF: Session control function
SGSN: Serving GPRS support node
S-GW: Serving gateway
SIP AS: Session initiation protocol application server
SPR: Subscription profile repository
S-SCSF: Serving call session control manager
UPR: User profile repository
UPSF: User profile server.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. L. Skorin-Kapov's work was in part supported by the Ministry of Science, Education and Sports of the Republic of Croatia research projects nos. 036-0362027-1639 and 071-0362027-2329.

References

[1] M. lsson, S. Sultana, S. Rommer, L. Frid, and C. Mulligan, *SAE and the Evolved Packet Core: Driving the Mobile Broadband Revolution*, Elsevier, New York, NY, USA, 2009.

[2] S. Ickin, K. Wac, M. Fiedler, L. Jankowski, J. H. Hong, and A. K. Dey, "Factors influencing quality of experience of commonly used mobile applications," *IEEE Communication Magazine*, vol. 50, no. 4, pp. 48–56, 2012.

[3] R. Stankiewicz and A. Jajszczyk, "A survey of QoE assurance in converged networks," *Computer Networks*, vol. 55, no. 7, pp. 1459–1473, 2011.

[4] I. T. U. -T Recommendation Y.2012, "Functional Requirements and Architecture of the NGN Release 1," September 2006.

[5] K. Bogenine, R. Ludwig, P. Mogensen et al., "LTE part II: Radio access," *IEEE Communications Magazine*, vol. 47, no. 4, pp. 40–42, 2009.

[6] 3GPP Technical Report TR 23.882 V8.0.0, "3GPP System Architecture Evolution: Report on Technical Options and Conclusions," December 2008.

[7] 3GPP Technical Specification TS 23.107 V11.0.0, "Quality of Service (QoS) Concept and Architecture," June 2012.

[8] 3GPP Technical Specification TS 23.207 V10.0.0, "End-to-end Quality of Service (QoS) Concept and Architecture," March 2011.

[9] 3GPP Technical Specification TS 23.203 V11.6.0, "Policy and Charging Control Architecture," June 2012.

[10] K. U. R. Laghari and K. Connelly, "Toward total quality of experience: a QoE model in a communication ecosystem," *IEEE Communication Magazine*, vol. 50, no. 4, pp. 58–65, 2012.

[11] ITU-T Recommendation P.10/G.100, "Vocabulary for performance and quality of service. Amendment 2: New definitions for inclusion in Recommendation ITU-T P.10/G.100," July 2008.

[12] ETSI Technical Report 102 643 V1.0.2, "Human Factors (HF); Quality of Experience (QoE) requirements for real-time communication services," October 2010.

[13] K. De Moor, W. Joseph, I. Ketykó et al., "Linking users' subjective QoE evaluation to signal strength in an IEEE 802.11b/g wireless LAN environment," *EURASIP Journal on Wireless Communications and Networking*, vol. 2010, Article ID 541568, 2010.

[14] V. Roto, E. Law, A. Vermeeren, and J. Hoonhout, Eds., "User Experience White Paper. Bringing Clarity to the Concept of User Experience," February 2011, http://www.allaboutux.org/files/UX-WhitePaper.pdf.

[15] A. Van Ewijk, J. De Vriendt, and L. Finizola, *Quality of Service for IMS on Fixed Networks. Business Models and Drivers for Next-Generation IMS Services*, International Engineering Consortium, USA, 2007.

[16] T. M. O'Neill, "Quality of Experience and Quality of Service for IP video conferencing," *Polycom*; 2002.

[17] M. Siller and J. C. Woods, "QoS arbitration for improving the QoE in multimedia transmission," in *Proceedings of the International Conference on Visual Information Engineering*, pp. 238–241, July 2003.

[18] Empirix, "Assuring QoE on Next Generation Networks," Whitepaper, 2001, http://www.whitepapers.org/docs/show/113.

[19] D. Soldani, "Means and methods for collecting and analyzing QoE measurements in wireless networks," in *Proceedings of International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM '06)*, pp. 531–535, Niagara-Falls, Buffalo-NY, USA, June 2006.

[20] K. De Moor, I. Ketykó, W. Joseph et al., "Proposed framework for evaluating quality of experience in a mobile, testbed-oriented living lab setting," *Mobile Networks and Applications*, vol. 15, no. 3, pp. 378–391, 2010.

[21] S. Baraković, J. Baraković, and H. Bajrić, "QoE dimensions and QoE measurement of NGN services," in *Proceedings of the 18th Telecommunications Forum (TELFOR '10)*, Belgrade, Serbia, November 2010.

[22] P. Le Callet, S. Moller, and A. Perkis, Eds., "Qualinet White paper on Definitions of Quality of Experience (QoE)," May 2012, http://www.qualinet.eu/.

[23] M. G. Martini, C. W. Chen, Z. Chen, T. Dagiuklas, L. Sun, and X. Zhu, "Guest editorial QoE-aware wireless multimedia systems," *IEEE Journal on Selected Areas in Telecommunications*, vol. 30, no. 7, pp. 1153–1156, 2012.

[24] T. Hoßfeld, R. Schatz, M. Varela, and C. Timmerer, "Challenges of QoE management for cloud applications," *IEEE Communication Magazine*, vol. 50, no. 4, pp. 28–36, 2012.

[25] D. Durkee, "Why cloud computing will never be free," *Communications of the ACM*, vol. 53, no. 5, pp. 62–69, 2010.

[26] L. Skorin-Kapov and M. Varela, "A multi-dimensional view of QoE: the ARCU model," in *Proceedings of the 35th Jubilee International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO '12)*, Opatija, Croatia, May 2012.

[27] J. Shaikh, M. Fiedler, and D. Collange, "Quality of experience from user and network perspectives," *Annales des Telecommunications/Annals of Telecommunications*, vol. 65, no. 1-2, pp. 47–57, 2010.

[28] P. Reichl, B. Tuffin, and R. Schatz, "Logarithmic laws in service quality perception: where microeconomics meets psychophysics and quality of experience," *Telecommunication Systems Journal*, vol. 55, no. 1, pp. 1–14, 2011.

[29] M. Fiedler, T. Hoßfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, vol. 24, no. 2, pp. 36–41, 2010.

[30] P. Reichl, S. Egger, R. Schatz, and A. D'Alconzo, "The logarithmic nature of QoE and the role of the Weber-Fechner law in QoE assessment," in *Proceedings of IEEE International Conference on Communications (ICC '10)*, Cape Town, South Africa, May 2010.

[31] E. H. Weber, *Annotationes Anatomicae et Physiologicae: Programmata Collecta: Fasciculi Tres. de Pulsu, Resorptione, Auditu et Tactu*, Koehler, 1834.

[32] ITU-T Recommendation P.800.1, "Mean Opinion Score (MOS) Terminology," July 2006.

[33] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: the MOS is not enough!," in *Proceedings of the 3rd International Workshop on Quality of Multimedia Experience (QoMEX '11)*, Machelen, Belgium, September 2011.

[34] ITU-T Recommendation G.1011, "Reference Guide to Quality of Experience Assessment Methodologies," June 2010.

[35] K. Wac, S. Ickin, J. H. Hong, L. Janowski, M. Fiedler, and A. K. Dey, "Studying the experience of mobile applications used in different contexts of daily life," in *Proceedings of the 1st ACM SIGCOMM Workshop on Measurements up the Stack (W-MUST '11)*, Toronto, Ontario, Canada, August 2011.

[36] J. M. Hektner, J. A. Schmidt, and M. Csikszentmihalyi, *Experience Sampling Method: Measuring the Quality of Everyday Life*, Sage Publications, 2006.

[37] P. A. Dinda, G. Memik, R. P. Dick et al., "The user in experimental computer systems research," in *Proceedings of the Workshop on Experimental Computer Science*, San Diego, Calif, USA, June 2007.

[38] V. Menkovski, G. Exarchakos, A. Liotta, and A. Cuadra-Sanchez, "Managing quality of experience on a commercial mobile TV platform," *International Journal on Advances in Telecommunications*, vol. 4, no. 1-2, pp. 72–81, 2011.

[39] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via Crowdsourcing," in *Proceedings of the IEEE International Symposium on Multimedia (ISM '11)*, December 2011.

[40] K. T. Chen, C. J. Chang, C. C. Wu, Y. C. Chang, and C. L. Lei, "Quadrant of euphoria: a crowdsourcing platform for QoE assessment," *IEEE Network*, vol. 24, no. 2, pp. 28–35, 2010.

[41] A. Takahashi, "Framework and standardization of quality of experience (QoE) design and management for audiovisual communication services," *NTT Technical Review*, vol. 7, no. 4, pp. 1–5, 2009.

[42] ITU-T Recommendation 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.

[43] ITU-T Recommendation P.863, "Perceptual Objective Listening Quality Assessment," January 2011.

[44] ITU-R Recommendation BS.1387, "Method for Objective Measurements of Perceived Audio Quality," November 2001.

[45] ITU-T Recommendation P.563, "Single-ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications," May 2004.

[46] F. Kuipers, R. Kooij, De Vleeschauwer, and K. Brunnstrom, "Techniques for measuring quality of experience," *Wired/Wireless Internet Communications*, pp. 216–227, 2010.

[47] A. Takahashi, D. Hands, and V. Barriac, "Standardization activities in the ITU for a QoE assessment of IPTV," *IEEE Communications Magazine*, vol. 46, no. 2, pp. 78–84, 2008.

[48] ITU-T Recommendation P.564, "Conformance Testing for Voice Over IP Transmission Quality Assessment Models," November 2007.

[49] ITU-T Recommendation G.107, "The E-model: A Computational Model for Use in Transmission Planning," December 2011.

[50] ITU-T Recommendation G.1070, "Opinion Model for Video-Telephony Applications," April 2007.

[51] ITU-T Recommendation G.1030, "Estimating End-to-End Performance in IP Networks for Data Applications," November 2005.

[52] ETSI Guide 202 843 V1.1.2, "Definitions and Methods for Assessing the QoS Parameters of the Customer Relationship Stages Other than Utilization," July 2011.

[53] P. Brooks and B. Hestnes, "User measures of quality of experience: why being objective and quantitative is important," *IEEE Network*, vol. 24, no. 2, pp. 8–13, 2010.

[54] A. Perkis, S. Munkeby, and O. I. Hillestad, "A model for measuring quality of experience," in *Proceedings of the 7th Nordic Signal Processing Symposium (NORSIG '06)*, pp. 198–201, June 2006.

[55] H. J. Kim, K. H. Lee, and J. Zhang, "In-service feedback QoE framework," in *Proceedings of the 3rd International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ '10)*, pp. 135–138, Athens, Greece, June 2010.

[56] K. Kilkki, "Quality of experience in communications ecosystem," *Journal of Universal Computer Science*, vol. 14, no. 5, pp. 615–624, 2008.

[57] S. Möller, K. P. Engelbrecht, C. Kühnel, I. Wechsung, and B. Weiss, "A taxonomy of quality of service and quality of experience of multimodal human-machine interaction," in *Proceedings of the International Workshop on Quality of Multimedia Experience (QoMEx '09)*, pp. 7–12, July 2009.

[58] D. Geerts, K. De Moor, I. Ketykó et al., "Linking an integrated framework with appropriate methods for measuring QoE," in *Proceedings of the 2nd International Workshop on Quality of Multimedia Experience (QoMEX '10)*, pp. 158–163, Trondheim, Norway, June 2010.

[59] M. Volk, J. Sterle, U. Sedlar, and A. Kos, "An approach to modeling and control of QoE in next generation networks," *IEEE Communications Magazine*, vol. 48, no. 8, pp. 126–135, 2010.

[60] W. Song, D. Tjondronegoro, and M. Docherty, "Understanding user experience of mobile video: framework, measurement, and optimization," in *Mobile Multimedia—User and Technology Perspectives*, INTECH Open Access, 2012.

[61] F. Ricciato, "Traffic monitoring and analysis for the optimization of a 3G network," *IEEE Wireless Communications*, vol. 13, no. 6, pp. 42–49, 2006.

[62] D. Kataria, "Mitigating strategies for smart phone signaling overload," December 2011, http://www.ecnmag.com/articles/2011/10/mitigating-strategies-smart-phone-signaling-overload.

[63] V. Menkovski and A. Liotta, *QoE for mobile streaming. Mobile multimedia: user and technology perspectives*, InTech Publishing, 2012.

[64] ITU-T Recommendation J.144, "Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference," March 2004.

[65] A. F. Wattimena, R. E. Kooij, J. M. Van Vugt, and O. K. Ahmed, "Predicting the perceived quality of a first person shooter: the Quake IV G-model," in *Proceedings of the 5th ACM SIGCOMM Workshop on Network and System Support for Games (NetGames '06)*, October 2006.

[66] T. Hoßfeld, S. Biedermann, R. Schatz, A. Plazter, S. Egger, and M. Fiedler, "The memory effect and its implications on web QoE modeling," in *Proceedings of the 23rd International Teletraffic Congress (ITC '11)*, San Francisco, California, USA, September 2011.

[67] S. Egger, P. Reichl, T. Hoßfeld, and R. Schatz, ""Time is bandwidth"? Narrowing the gap between subjective time perception and quality of experience," in *Proceedings of the IEEE International Conference on Communications (ICC '12)*, Ottawa, Canada, June 2012.

[68] T. Ciszkowski, W. Mazurczyk, Z. Kotulski, T. Hoßfeld, M. Fiedler, and D. Collange, "Towards quality of experience-based reputation models for future web service provisioning," *Telecommunication Systems*, vol. 52, no. 2, pp. 1–13, 2013.

[69] M. Jarschel, M. Schlosser, S. Scheuring, and T. Hoßfeld, "An evaluation of QoE in cloud gaming based on subjective tests," in *Proceedings of the 5th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS '11)*, Seoul, Korea, July 2011.

[70] S. Egger, T. Hoßfeld, R. Schatz, and M. Fiedler, "Waiting times in quality of experience for web based services," in *Proceedings of the 4th International Workshop on Quality of Multimedia Experience (QoMEX '12)*, Yarra Valley, Australia, July 2012.

[71] H. Batteram, G. Damm, A. Mukhopadhyay, L. Philippart, R. Odysseos, and C. Urrutia-Valdés, "Delivering quality of experience in multimedia networks," *Bell Labs Technical Journal*, vol. 15, no. 1, pp. 175–194, 2010.

[72] D. Soldani, M. Li, and R. Cuny, *QoS and QoE Management in UMTS Cellular Systems*, John Wiley & Sons, USA, 2006.

[73] 3GPP Technical Specification TS 26.234 V11.0.0, "Transparent end-to-end Packet-switched Streaming Service (PSS); Protocols and Codecs," March 2012.

[74] 3GPP Technical Specification TS 26.247 V10.2.0, "Transparent End-to-End Packet-Switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming Over HTTP (3GP-DASH)," June 2012.

[75] 3GPP Technical Specification TS 26.114 V11.4.0, "IP Multimedia Subsystem (IMS); Multimedia Telephony; Media Handling and Interaction," June 2012.

[76] A. Raake, J. Gustafsson, S. Argyropoulos et al., "IP-based mobile and fixed network audiovisual media services," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 68–79, 2011.

[77] D. Collange and J. L. Costeux, "Passive estimation of quality of experience," *Journal of Universal Computer Science*, vol. 14, no. 5, pp. 625–641, 2008.

[78] D. Collange, M. Hajji, J. Shaikh, M. Fiedler, and P. Arlos, "User impatience and network performance," in *Proceedings of the 8th Euro-NF Conference on Next Generation Internet (NGI '12)*, Karlskrona, Sweden, June 2012.

[79] K. De Moor and L. De Marez, "The challenge of user- and QoE-centric research and product development in today's ICT environment," in *Innovating for and by Users*, pp. 77–90, Office for Official Publications of the European Communities, Luxembourg, UK edition, 2008.

[80] B. Fehnert and A. Kosagowsky, "Measuring user experience: complementing qualitative and quantitative assessment," in *Proceedings of the 10th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '08)*, pp. 383–386, September 2008.

[81] M. Andrews, J. Cao, and J. McGowan, "Measuring human satisfaction in data networks," in *Proceedings of the 25th IEEE International Conference on Computer Communications (INFOCOM '06)*, Barcelona, Spain, April 2006.

[82] 3GPP Technical Specification TS 23.402 V11.3.0, "Architecture Enhancements for non-3GPP Accesses," June 2012.

[83] 3GPP Technical Specification TS 22.278 V12.1.0, "Service Requirements for the Evolved Packet System," June 2012.

[84] P. Rengaraju, C. H. Lung, and F. R. Yu, "On QoE monitoring and E2E service assurance in 4G wireless networks," *IEEE Wireless Communications*, vol. 19, no. 4, pp. 89–96, 2012.

[85] R. Serral-Gracià, E. Cerqueira, M. Curado, M. Yannuzzi, E. Monteiro, and X. Masip-Bruin, "An overview of quality of experience measurement challenges for video applications in IP networks," in *Proceedings of the 8th International Conference on Wired/Wireless Internet Communications (WWIC '10)*, pp. 252–263, Lulea, Sweden, June 2010.

[86] R. Fielding, J. Gettys, J. Mogul et al., "Hypertext Transfer Protocol - HTTP/1.1," IETF Technical Report RFC 2616, Jun 1999.

[87] O. Oyman and S. Singh, "Quality of experience for HTTP adaptive streaming services," *IEEE Communications Magazine*, vol. 50, no. 4, pp. 20–27, 2012.

[88] I. Ketykó, K. De Moor, T. De Pessemier et al., "QoE measurement of mobile youtube video streaming," in *Proceedings of the 3rd Workshop on Mobile Video Delivery (MoViD '10)*, pp. 27–32, October 2010.

[89] A. J. Verdejo, K. De Moor, I. Ketyko et al., "QoE estimation of a location-based mobile game using on-body sensors and QoS-related data," in *Proceedings of the IFIP Wireless Days Conference (WD '10)*, October 2010.

[90] K. T. Chen, C. C. Tu, and W. C. Xiao, "OneClick: A framework for measuring network quality of experience," in *Proceedings of the 28th Conference on Computer Communications (INFOCOM '09)*, pp. 702–710, April 2009.

[91] M. Varela and J. P. Laulajainen, "QoE-driven mobility management—integrating the users' quality perception into network-level decision making," in *Proceedings of the Quality of Multimedia Experience (QoMEX '11)*, Mechelen, Belgium, September 2011.

[92] M. Varela, *Pseudo-subjective quality assessment of multimedia streams and its applications in control [Ph.D. thesis]*, INRIA/IRISA, University Rennes I, Rennes, France, Nov 2005.

[93] K. Mitra, C. Ahlund, and A. Zaslavsky, "QoE estimation and prediction using hidden Markov models in heterogeneous access networks," in *Australasian Telecommunication Networks and Applications Conference*, Brisbane, Australia, 2012.

[94] "EU FP7 PERIMITER project," http://www.ict-perimeter.eu/.

[95] 3GPP Technical Specification TS 23.228 V11.4.0, "IP Multimedia Subsystem (IMS)," March 2012.

[96] T. Hoßfeld, F. Liers, R. Schatz et al., *Quality of Experience Management for YouTube: Clouds, FoG and the AquareYoum*, PIK: Praxis der Informationverarbeitung und - kommnikation (PIK)., 2012.

[97] B. Staehle, M. Hirth, R. Pries, F. Wamser, and D. Staehle, "Aquarema in action: improving the YouTube QoE in wireless mesh networks," in *Proceedings of the Baltic Congress on Future Internet and Communications (BCFIC Riga '11)*, pp. 33–40, February 2011.

[98] I. Ketykó, K. De Moor, W. Joseph, L. Martens, and L. De Marez, "Performing QoE-measurements in an actual 3G network," in *Proceedings of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB '10)*, March 2010.

[99] S. M. Chadchan and C. B. Akki, "3GPP LTE/SAE: an overview," *International Journal of Computer and Electrical Engineering*, vol. 2, no. 5, pp. 806–814, 2010.

[100] L. Skorin-Kapov and M. Matijasevic, "Modeling of a QoS matching and optimization function for multimedia services in the NGN," in *Proceedings of the 12th IFIP/IEEE International Conference on Management of Multimedia and Mobile Networks and Services: Wired-Wireless Multimedia Networks and Services Management (MMNS '09)*, vol. 5842 of *Lecture Notes in Computer Science*, pp. 54–68, October 2009.

[101] K. Ivešić, M. Matijašević, and L. Skorin-Kapov, "Simulation based evaluation of dynamic resource allocation for adaptive multimedia services," in *Proceedings of the 7th International Conference on Network and Service Management (CNSM '11)*, pp. 1–8, October 2011.

[102] J. Sterle, M. Volk, U. Sedlar, J. Bester, and A. Kos, "Application-based NGN QoE controller," *IEEE Communications Magazine*, vol. 49, no. 1, pp. 92–101, 2011.

[103] S. Thakolsri, W. Kellerer, and E. Steinbach, "QoE-based cross-layer optimization of wireless video with unperceivable temporal video quality fluctuation," in *Proceedings of the IEEE International Conference on Communication (ICC '11)*, Kyoto, Japan, Jun 2011.

[104] S. Latré, P. Simoens, B. De Vleeschauwer et al., "An autonomic architecture for optimizing QoE in multimedia access networks," *Computer Networks*, vol. 53, no. 10, pp. 1587–1602, 2009.

[105] P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, P. Mogensen, and J. M. Lopez-Soler, "QoE oriented cross-layer design of a resource allocation algorithm in beyond 3G systems," *Computer Communications*, vol. 33, no. 5, pp. 571–582, 2010.

[106] J. Hassan, M. Hassan, S. K. Das, and A. Ramer, "Managing quality of experience for wireless VoIP using noncooperative games," *IEEE Journal on Selected Areas in Communication*, vol. 30, no. 7, pp. 1193–1204, 2012.

[107] A. El Essaili, L. Zhou, D. Schroeder, E. Steinbach, and W. Kellerer, "QoE-driven live and on-demand LTE uplink video transmission," in *Proceedings of the 13th International Workshop on Multimedia Signal Processing (MMSP '11)*, Hangzhou, China, October 2011.

[108] S. Khan, S. Duhovnikov, E. Steinbach, and W. Kellerer, "MOS-based multiuser multiapplication cross-layer optimization for mobile multimedia communication," *Advances in Multimedia*, vol. 2007, Article ID 94918, 11 pages, 2007.

[109] G. Aristomenopoulos, T. Kastrinogiannis, S. Papavassiliou, V. Kaldanis, and G. Karantonis, "A novel framework for dynamic utility-based QoE provisioning in wireless networks," in *Proceedings of the 53rd IEEE Global Communications Conference (GLOBECOM '10)*, December 2010.

[110] F. Wamser, D. Staehle, J. Prokopec, A. Maeder, and P. Tran-Gia, "Utilizing buffered YouTube playtime for QoE-oriented scheduling in OFDMA networks," in *Proceedings of the 24th International Teletraffic Congress (ITC '12)*, Krakow, Poland, October 2012.

[111] K. Piamrat, A. Ksentini, C. Viho, and J. M. Bonnin, "QoE-based network selection for multimedia users in IEEE 802.11 wireless networks," in *Proceedings of the 33rd IEEE Conference on Local Computer Networks (LCN '08)*, pp. 388–394, Montreal, Canada, October 2008.

[112] M. Shehada, S. Thakolsri, Z. Despotovic, and W. Kellerer, "QoE-based cross-layer optimization for video delivery in long term evolution mobile networks," in *Proceedings of the14th International Symposium on Wireless Personal Multimedia Communications (WPMC '11)*, Le Quartz, Brest, France, October 2011.

[113] N. Amram, B. Fu, G. Kunzmann et al., "QoE-based transport optimization for video delivery over next generation cellular networks," in *Proceedings of the IEEE Symposium on Computers and Communications (ISCC '11)*, Kerkyra, Greece, July 2011.

[114] A. Khan, I. Mkwawa, L. Sun, and E. Ifeachor, "QoE-driven sender bitrate adaptation scheme for video applications over IP multimedia subsystem," in *Proceedings of the IEEE International Conference on Communication (ICC '11)*, Kyoto, Japan, June 2011.

[115] M. Csernai and A. Gulyas, "Wireless Adapter Sleep Scheduling based on video QoE: how to improve battery life when watching streaming video?" in *Proceedings of the 20th International Conference on Computer Communications and Networks (ICCCN '11)*, Maui, Hawaii, USA, August 2011.

[116] A. A. Khalek, C. Caramanis, and R. W. Heath Jr., "A cross-layer design for perceptual optimization of H.264/SVC with unequal error protection," *IEEE Journal on Selected Areas in Communication*, vol. 30, no. 7, pp. 1157–1171, 2012.

Hindawi Publishing Corporation
Journal of Computer Networks and Communications
Volume 2013, Article ID 505910, 15 pages
http://dx.doi.org/10.1155/2013/505910

Research Article

Towards a QoE-Driven Resource Control in LTE and LTE-A Networks

Gerardo Gómez,[1] Javier Lorca,[2] Raquel García,[2] and Quiliano Pérez[2]

[1] *Department of Communications Engineering, University of Malaga, 29071 Malaga, Spain*
[2] *Telefonica I+D, 28006 Madrid, Spain*

Correspondence should be addressed to Gerardo Gómez; ggomez@ic.uma.es

Received 12 September 2012; Revised 9 December 2012; Accepted 19 December 2012

Academic Editor: Lea Skorin-Kapov

We propose a novel architecture for providing quality of experience (QoE) awareness to mobile operator networks. In particular, we describe a possible architecture for QoE-driven resource control for long-term evolution (LTE) and LTE-advanced networks, including a selection of KPIs to be monitored in different network elements. We also provide a description and numerical results of the QoE evaluation process for different data services as well as potential use cases that would benefit from the rollout of the proposed framework.

1. Introduction

The convergence of wireless networks and multimedia communications, linked to the swift development of services and the increasing competition, has caused user expectations of network quality to rise. Network quality has become one of the main targets for the network optimization and maintenance departments.

Traditionally, network measurements such as accessibility, maintainability, and quality were enough to evaluate the user experience of voice services [1]. However, for data services, the correlation between network measurements and user benefits is not as straightforward. Firstly, the data system, due to the use of packet switching, is affected by the performance of individual nodes and protocols through which information travels, and, secondly, radio resources are now shared among different applications. Under these conditions, the performance evaluation of data services is usually carried out by monitoring terminals on the real network.

The end-to-end quality experienced by an end user results from a combination of elements throughout the protocol stack and system components. Thus, the performance evaluation of the service requires a detailed performance analysis of the entire network (from the user equipment up to the application server or remote user equipment).

Quality of experience (QoE) is a subjective measurement of the quality experienced by a user when he uses a telecommunication service. The aim pursued when assessing the quality of service (QoS) may be the desire to optimize the operation of the network from a perspective purely based on objective parameters, or the more recent need of determining the quality that the user is actually achieving, as well as its satisfaction level. However, the QoE goes further and takes into account the satisfaction a user receives in terms of both content and use of applications. In this sense, the introduction of smartphones has been a quantitative leap in user QoE expectations.

Traditionally, QoE has been evaluated through subjective tests carried out on the users in order to assess their satisfaction degree with a mean opinion score (MOS) value. This type of approach is obviously quite expensive, as well as annoying to the user. Additionally, this method cannot be used for making decisions to improve the QoE on the move. That is why in recent years new methods have been proposed to estimate the QoE based on certain performance indicators associated with services. A possible solution to

evaluate instantaneously the QoE is to integrate QoE analyzers in the mobile terminal itself [2]. If mobile terminals are able to report the measurements to a central server, the QoE assessment process is simplified significantly. Other solutions are focused on including new network elements (e.g., network analyzers, deep packet inspectors, etc.) that are responsible for capturing the traffic from a certain service and analyzing its performance [3]. For instance, the work presented in [4] investigates the problem of YouTube quality monitoring from an access provider's perspective, concluding that it is possible to detect application-level stalling events by using network-level passive probing only. In other work, the evaluation of video-streaming quality in mobile terminals is addressed by monitoring objective parameters like packet loss rate or jitter [5].

However, whatever solution intended to estimate the QoE from traffic measurements requires some kind of mapping towards a QoE value. A possible solution to perform this process is to apply a utility function associated to the particular data service in order to map the application level quality of service (QoS) into QoE (in terms of MOS value). Many research works are focused in that direction. For instance, a generic formula that connects QoE and QoS parameters (for different packet data services) is proposed in [6]. The work presented in [7] addresses the perception principles and discusses their applicability towards fundamental relationships between waiting times and QoE for web services. Other work quantifies the impact of initial delays on the user-perceived QoE for different application scenarios by means of subjective laboratory and crowd-sourcing studies [8]. Subjective experiments drawing on the evaluation of objective and subjective QoE aspects by a user panel for quantifying QoE during mobile video are presented in [9, 10].

Previous mentioned studies are mainly focused on QoS and/or QoE evaluation, but no action, procedure, or framework is proposed to enhance the end user quality. Only a few works tackle this issue; for instance, a QoE oriented scheduling algorithm is proposed in [11] to dynamically prioritize YouTube users against other users if a QoE degradation is imminent (based on the buffered playtime of the YouTube video player). Other research work provides a methodology for incorporating QoE into a network's radio resource management (RRM) mechanism by exploiting network utility maximization theory [12]. In [13], a specification and testbed implementation of an application-based QoE controller are presented, proposing a solution for QoE control in next-generation networks although they do not include any QoE modeling or estimation algorithm.

In this paper, we propose a novel architecture that enables LTE operators to be aware of the instantaneous QoE that their subscribers are experiencing. In particular, we propose some additions to existing LTE architecture for QoE-driven resource control purposes. We have also identified a set of key performance indicators (KPIs) at the network and application levels for different data services, as well as method to estimate the QoE for web browsing, video YouTube, and voice over IP. Finally, we describe a set of potential use cases that would benefit from the rollout of the proposed framework.

The remainder of this paper is structured as follows. Section 2 provides an overview of the LTE architecture. The proposed architecture for a QoE-driven control is described in Section 3. Section 4 presents a selection of KPIs to be monitored in different network elements. The QoE evaluation process from lower layers' KPIs is described in Section 5. Different use cases associated to the proposed framework are analyzed in Section 6. Finally, some concluding remarks are discussed in Section 7.

2. Overview of LTE Architecture

A general vision of the LTE architecture is described in this section, focusing on the QoS concepts specified in 3rd generation partnership project (3GPP) specifications.

Figure 1 shows the overall network architecture of the evolved packet system (EPS) including the network elements and the standardized interfaces. The network is comprised of the core network (EPC) and the access network (called Evolved Universal Terrestrial Radio Access Network, E-UTRAN). While the EPC consists of many logical nodes, the E-UTRAN is made up of essentially just one node, the evolved NodeB (eNodeB), which connects to the user Equipments (UEs).

The EPS provides the user with IP connectivity to a packet data network (PDN) for accessing the Internet, as well as for running services such as voice over IP (VoIP). One of the main concepts related to QoS in LTE is the EPS bearer, which is a logical connection (associated with a certain QoS level) between the terminal and the evolved packet core (EPC). Multiple bearers can be established for a user in order to provide different QoS streams or connectivity to different PDNs. For example, a user can be engaged in a voice call (via a VoIP bearer) while at the same time downloading a file (via a best-effort bearer).

The EPS includes a policy and charging control (PCC) subsystem, which provides advanced tools for service-aware QoS and charging control. It provides a way to manage the service-related connections in a consistent and controlled way. It determines how bearer resources are allocated for a given service, including how the service flows are partitioned to bearers, what QoS characteristics those bearers will have, and finally, what kind of accounting and charging will be applied.

The EPC is responsible for the overall control of the UE and establishment of the bearers. The main logical nodes of the EPC are (see Figure 1) as follows.

(i) Policy control and charging rules function (PCRF): it is the policy engine of PCC, and it is responsible for the QoS policy management as well as for controlling the flow-based charging functionalities in the policy control enforcement function (PCEF), which resides in the packet data network gateway (P-GW). The PCRF provides the QoS authorization (QoS class identifier and bit rates) that decides how a certain data flow will be treated in the PCEF and ensures that this is in accordance with the user's subscription profile.

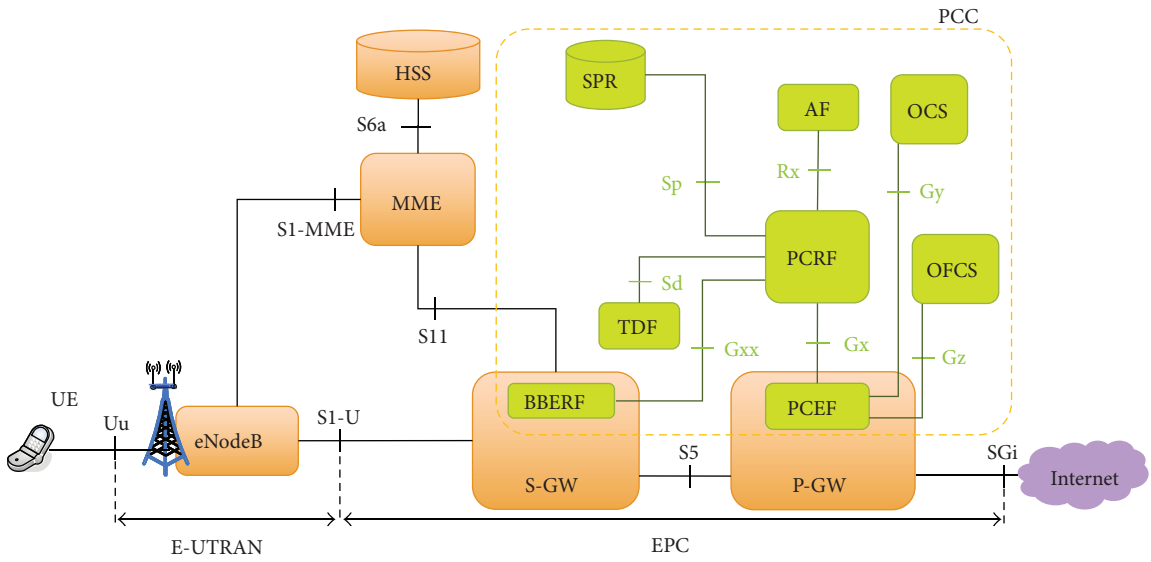


Figure 1: High level architecture for 3GPP LTE and LTE-A EPS.

(ii) Home subscriber server (HSS): it acts as a master repository of all subscriber and service-specific information. It combines the home location register (HLR) and authentication center (AuC) functionality of previous releases. The HSS contains users' subscription data such as the EPS-subscribed QoS profile and any access restrictions for roaming.

(iii) PDN gateway (P-GW): it is responsible for Internet Protocol (IP) address allocation for the UE, as well as QoS enforcement and flow-based charging according to rules from the PCRF. The P-GW is responsible for the filtering of downlink (DL) user IP packets into the different QoS bearers. This is performed based on traffic flow templates (TFTs).

(iv) Serving-GW (S-GW): all IP packets are transferred through the S-GW, which serves as local mobility anchor for data bearers when the UE moves between eNodeBs. It includes a bearer binding and event reporting function (BBERF).

(v) Mobility and management entity (MME): the MME is the control node which processes the signaling between the UE and the EPC. The main functions supported by the MME are related to bearer management (establishment, maintenance and release of the bearers), and connection management.

(vi) Application function (AF): it extracts session information from the application signalling and communicates with the PCRF to transfer this dynamic information, required for PCRF decisions.

(vii) Subscription profile repository (SPR) is the database that stores information related to network usage policies of a subscriber. For example, the SPR can indicate which final services are authorized for a user, the authorized QoS parameters per service, or the user category (e.g., business and consumer). The PCRF may use the subscription information as a basis for the policy and charging control decisions.

(viii) Traffic detection function (TDF): it has been introduced in LTE-A to help the network achieve service awareness by introducing mechanisms for service detection.

(ix) Online charging system (OCS) provides credit management and grants credit to the PCEF based on time, traffic volume, or chargeable events.

(x) Offline charging system (OFCS) receives events from the PCEF and generates charging data records (CDRs) for the billing system.

The access network of LTE, E-UTRAN, simply consists of a network of eNodeBs, which are normally inter connected with each other by means of an interface known as X2, and to the EPC by means of the S1 interface. The eNodeB plays a critical role in the end-to-end QoS. It usually performs the following QoS-related functions: admission control and preemption, rate policing (to protect the network from becoming overloaded and to ensure that the services are sending data in accordance with the specified maximum bit rates), scheduling (to distribute radio resources between the established bearers), and L1/L2 protocol configuration in accordance with the QoS characteristics associated with the bearer.

The UEs in LTE may support multiple applications at the same time, each one having different QoS requirements. This is achieved by establishing different EPS bearers for each QoS flow. EPS bearers can be classified into two categories based on the nature of the QoS they provide: guaranteed bit rate (GBR) bearers in which resources are permanently allocated and non-GBR bearers which do not guarantee any particular bit rate. In the access network, it is the eNodeB's responsibility to ensure that the necessary QoS for a bearer over the radio interface is met. Each bearer has an associated QoS

Table 1: Role of the main EPS interfaces.

Interface	Role
Gx	Used by the PCRF to convey policy enforcement to the P-GW
Gxx	Bearer binding and event reporting function
Rx	Used by application function to convey policy data to the PCRF
Sp	Retrieving per subscriber policy data
Sd	Used to identify/inform about the ongoing session details from the TDF
SGi	It is the reference point between the P-GW and the PDN (e.g., Internet)
S5	Signaling interface for establishing bearers between S-GW and P-GW
S6a	Used by the MME to retrieve subscriber data from HSS
S11	Used by the MME to control path switching and bearer establishment in S-GW
S1-MME	Signaling interface between the eNB and the MME
S1-U	User plane between eNB and S-GW
Uu	This is the air interface between UE and eNB
Gy	Online charging information
Gz	Offline charging information

class identifier (QCI)—characterized by priority, packet delay budget and admissible packet loss rate—and an allocation and retention priority (ARP) used for call admission control. IP packets mapped to the same EPS bearer receive the same bearer level packet forwarding treatment (e.g., scheduling policy, queue management policy, and rate shaping policy). Thus, the UE is not only responsible for requesting the establishment of EPS bearers for each QoS flow, but also for performing packet filtering in the uplink (UL) into different bearers based on TFTs, as P-GW does for the DL.

Table 1 summarizes the role of the main interfaces in the EPS.

3. Proposed Architecture for a QoE-Driven Control

We propose a novel architecture that enables LTE operators to be aware of the instantaneous QoE that their subscribers are experiencing. In the proposed architecture, all the information related to QoS or QoE will be managed in a centralized point that collects performance indicators from different network elements and take potential actions to improve the QoE.

Ideally, the PCRF would be the preferable candidate for this role. However, current PCRF interfaces' specifications do not provide enough flexibility to receive relevant information from any network element. This is why we propose to deploy an ad-hoc QoE-server (with a standardized interface towards the PCRF). A proper dynamic linkage between the QoE server and PCRF is recommended with the aim of achieving a dynamic control of QoS based on customer perception. As defined in the standard, PCRF may receive QoS-related information from different network elements: P-GW, S-GW, AF, and SPR. The goal of including any kind of interaction between the QoE Server and PCRF is to provide a wider vision of the quality perceived by the end users in order to take actions via policy management.

Taking into account that PCRF entity just includes standard interfaces, the communication between both entities could be fulfilled through the following alternatives (see Figure 2).

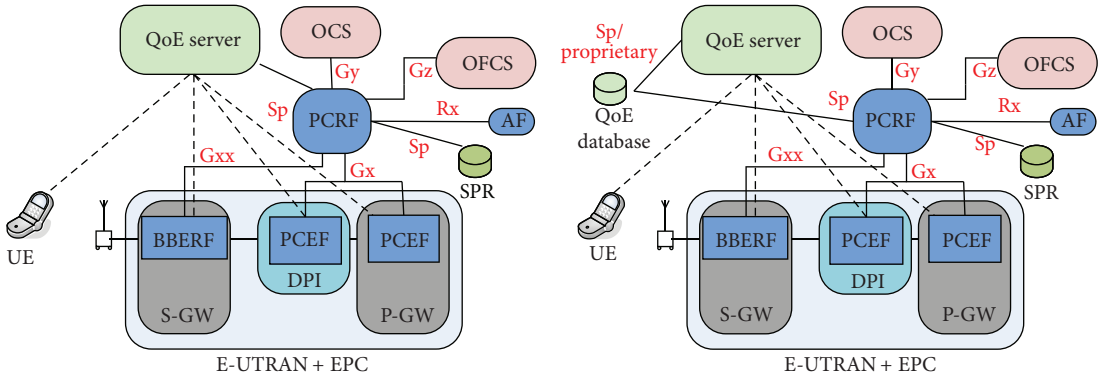
(a) *Via Gx reference point:* this option requires the QoS platform to include the capability of interchanging diameter commands with the PCRF. Note that PCRF manufacturers include Gx interface to manage policy rules between applications and policy enforcement points, such as gateways, DPIs, and so forth. We propose to reuse Gx to connect PCRF to our policy server (acting as a PCEF). It is not required that the QoE engine includes the whole PCEF functionality (like traffic filtering, monitoring, etc.) because these tasks need to be performed at the user plane. Instead, this platform just needs to include the possibility of sending/receiving certain information to/from the PCRF. This option has a higher flexibility to inform the PCRF about particular events related to the QoS.

(b) *Via Sp reference point:* this option relies on storing average QoE/QoS indicators in a proprietary database, which can be accessed via Sp from the PCRF, as it is already done with the standardized SPR. This reference point is used to retrieve subscriber related information like allowed services and preemption, subscriber's usage monitoring-related information, profile configuration, priority level (used to determine the ARP), list of allowed QCIs, and so forth. A possible use of the proprietary database is to provide dynamic subscriber-related information according to their associated performance indicators collected by the QoE engine.

The QoE server will be responsible for the following tasks: (1) collecting performance indicators from different network elements; (2) estimating the QoE for specific data services from previous performance indicators; (3) triggering potential actions (depending on the use case). These tasks are further described along this paper.

3.1. Collection of Performance Indicators. Numerous network elements may contribute to the performance monitoring process. Traditionally, performance-statistics from the operator network management subsystem (NMS) were the main source of feedback information to assess the service quality [14]. However, they are not considered very useful for the evaluation of data service quality, as NMS statistics are averaged for different services and for a long period of time (typically 1 hour).

Instead, mobile network operators usually deploy some kind of monitoring platform based on deep packet inspectors (DPIs). A DPI is a network equipment that potentially allows network providers to monitor, collect, and analyze the data



(a) QoE Server—PCRF connection via Gx

(b) QoE Server—PCRF connection via Sp

Figure 2: Proposed QoE architecture.

communications of millions of users simultaneously. DPIs make it possible to identify the applications being used on the network, which is very valuable information for many purposes such as QoS policy management. If a DPI is available, it will provide very valuable (real-time) information about the QoS being provided to each data flow. The location of such DPI will likely be close to the P-GW (or even within the P-GW as a hardware card). The advantage of having a DPI is that it is able to monitor above IP layers, for example, the transmission control protocol (TCP). Note that at, this location, all EPS bearers are handled by the DPI (or P-GW), each EPS being associated with a particular QoS profile. In fact, the complete QoS profile associated to each EPS bearer (i.e., QCI, ARP, GBR, maximum bitrate (MBR), etc.) is well known. That way, it may be checked whether the provided QoS is in concordance with the negotiated QoS profile; otherwise, EPS bearer renegotiation actions may be triggered via the PCRF. All statistics should be obtained per EPS bearer and calculated at a quick rate (e.g., 1 second) so that the QoE of each data service is (re)estimated every second, and real-time actions may be triggered.

In addition to the DPI monitoring process, the potential complementation of information with other sources of information from different network elements (mobile devices, gateways, etc.) is foreseen. If device-based agents are used within the network, it shall be integrated onto the QoE engine as a manner of enriching overall view of QoE for the network. Agent-based solutions are considered an interesting approach within an overall QoE monitoring strategy as representing a unique solution to access device specific issues. Nevertheless, these solutions are currently seen as a complement to network-based approach that shall help in specific issues related to QoE but are not currently seen as a replacement for network-based approaches due to the following.

(i) It is difficult to think on all-network scalability of these solutions on the medium term as per heavily dependent on handset manufacturer willingness to include them.

(ii) Agent-based solutions provide a very detailed view from end customer perspective, but have strong limitations in terms of exploring root cause for issues beyond pure device and access network-related problems.

(iii) Beyond pure technical aspects around the solutions, there are specific privacy and data protection aspects that need to be considered within the implementation of such solutions.

Based on that, it is recommended that device-based solutions are considered within the overall QoE monitoring strategy, scoping a percentage of the network and with the main function of complementing network-based solutions, specially for these aspects that are not easily seen/estimated from network perspective (device issues, network unavailability, and precise location of events).

4. KPIs Monitoring

As discussed before, our proposed architecture uses KPIs collected from different network elements, being DPIs and mobile terminals the most relevant ones. In this respect, this section describes a set of potential KPIs that might be monitored in such network elements.

4.1. KPIs to Be Monitored at a DPI. Existing DPIs are able to monitor a wide set of parameters and performance indicators at different network layers and associated to different data services. Here we list three basic network performance indicators that are key to characterize the instantaneous network status; they are useful for the estimation of the QoE associated to whatever data service.

(i) *IP level throughput:* it may be used to compare the provided throughput with the GBR and MBR values negotiated during session establishment. In the UL, this KPI provides a good performance indicator associated to the whole EPS bearer, as the statistics are taken at the output proxy of the operator network. However, the incoming DL throughput

measured at the DPI may not be a proper KPI for estimating the QoE when the radio interface is the bottleneck of the network, as, in this case, the measured IP level throughput does not correspond to the IP level throughput experienced by the mobile terminal. However, this problem only occurs when user datagram protocol (UDP) is used as a transport protocol and losses may happen between the DPI and the terminal; in that case, the solution is based on obtaining the IP level throughput directly measured in the terminal, as described later on. When TCP is used, this is not a problem as the IP level throughput is regulated by the TCP congestion control mechanism, and, hence, the measured throughput will be (ideally) similar to the throughput received at the terminal.

(ii) *IP packet loss rate:* the study of packet loss rate is a challenge as packet losses may occur in any network element which data passes through. The procedure to measure the packet loss rate can be based on analyzing upper layer protocols; concretely, this procedure is applicable for services based on TCP (e.g., web browsing, HTTP YouTube progressive downloading) or real-time transport protocol (RTP), like in real time streaming protocol- (RTSP) based video streaming). In case of RTP packets, loss detection shall be based on the sequence number field included in the RTP header, checking for possible missing numbers in the incoming RTP flow. In case of TCP-based services, a simple way to detect TCP losses in the P-GW is to analyze the packet retransmissions from the server, computing the duplicated number of sequence in the DL. If selective acknowledgment (SACK) feature is used in the TCP connection, the number of retransmitted packets will be the same as the lost ones, obtaining an accurate measure of the end-to-end loss rate. On the contrary, if SACK feature is not used, when the server detects a new packet loss, all the packets with higher sequence number will be retransmitted, computing all of them as packet losses. In this case, the estimated loss rate would be higher than the actual loss rate. Note that the estimation of the loss rate has to be averaged for a large number of packets; otherwise, the result might be distorted. Other way to compute the TCP losses would be implementing a part of the TCP protocol in the DPI. Concretely, the TCP control mechanisms could be used to determine the packet losses. It would be necessary to compute the duplicated ACKs from the UL as well as the losses due to the retransmission timeout (RTO). The adjustment of the initial RTO must be set as the value of the initial RTO of the server minus the time spent from the P-GW to the server (easy computed by executing a PING command). But the RTO is a parameter calculated dynamically, so it would also be necessary to implement the corresponding Jacobson algorithm in the P-GW to dynamically estimate the RTO value according to the round trip time (RTT) and RTT variation; the results of this dynamical calculation might not be the same in P-GW and the server due to the delays experienced in the external network. This method has the advantage of detecting packet losses even before the server, but it is very costly computationally, due to the vast number of TCP connections managed by the P-GW.

(iii) *End-to-end IP RTT:* a possible method to measure the RTT is based on analyzing the TCP connection establishment of a particular data service at a particular cell. As it is well known, TCP connection establishment uses a three-way handshake where the bit SYN is active. The following steps should be followed. (1) When the DPI/P-GW receives a TCP/IP packet (from a terminal) with the bit SYN active, it must start a timer computing the RTT; (2) the contribution of the external RTT will be computed after the reception of a new TCP/IP packet with the bit SYN that is active (from the server) acknowledging the previous one. The measurement of this contribution is especially important as the load conditions in the external network are unknown to provide a theoretical estimation; (3) finally, the end-to-end RTT will be completed when a new acknowledgment from the terminal is received at the DPI/P-GW. This measurement should be performed for each TCP connection establishment procedure detected at the DPI/P-GW. The most important statistic related to the RTT is the average RTT, which have a very important impact on upper layers' performance, especially for TCP. In that sense, RTT average value should be given for each QCI, as potential actions for improving the QoE in this scenario will be taken per QCI.

4.2. KPIs to Be Monitored at the Terminals. An important advantage of using real measurements at the terminal side is that they are highly correlated to the real QoE obtained. Thus, collecting statistical data specific for each service and terminal will allow for a better analysis of the performance of each service. Through periodic reporting of these measured values, the QoE assessment process is greatly simplified. In principle, the availability of obtaining certain performance indicators or parameters is dependent on the terminal manufacturer. The focus of this section is to list and describe the main KPIs that should be measured at the terminal side for specific data services.

Such monitoring process shall be carried out by an ad hoc application installed in mobile terminals. The software in charge of collecting terminal KPIs should be low consuming in terms of radio bitrate and processing load in order to not affect the quality of other applications. Such software will be responsible for measuring and reporting a set of KPIs to feed the theoretical model that estimates the QoE. For those mobile terminals that do not include monitoring capabilities, the theoretical model will use default values or average values from terminals located in the same cell.

Potential measurements at the terminal side are divided into the following.

(i) *Signaling KPIs:* associated to signaling delays during service establishment or attach to the EPS network, which just affect the initial service establishment and possible renegotiations of the bearers.

(ii) *Network level KPIs:* although there might be some overlapping with the network level KPIs measured in a DPI, it is always preferable to use KPIs collected by the terminals (if available) as they represent the final QoS received by the user equipment. Example of such KPIs isIP level throughput, IP packet loss rate, IP packet sizes, RTT from terminal-to-server or terminal-to-terminal (which might be periodically measured by sending PING commands from the terminal), and so forth.

(iii) *Transport level KPIs:* main KPIs at this level are TCP parameters, like the TCP advertised window (ADWN) or the maximum segment size (MSS). The ADWN represents the receiver window size, and it is included by the receiver in every ACK segment, indicating the maximum amount of data that it is able to receive. ADWN value should be at least as large as the bandwidth-delay product (BDP); otherwise, the receiver TCP layer will limit the achievable bandwidth. For example, let us consider a LTE terminal with 10 Mbps of transmission capability in DL. Assuming a typical LTE RTT of 10 ms for a 40 bytes packet, BDP can be computed as ADWN $\geq$ BDP = Bandwidth $*$RTT = 10 Mbit/s $*$ 10 ms = 12.5 kbytes. Additionally, during the TCP connection establishment both ends agree on the size of the largest segment that can be used within that connection, known as MSS. The value of the MSS also has an important impact on TCP performance. The larger the MSS the shorter the slow start phase will take to fill the pipe, due to the fact that during slow start the increment of transmitted bytes is in units of segments. In case of bigger segments, the bandwidth utilization during first slow start cycles is higher, and the BDP can be reached quicker. In addition, MSS also has an impact on the total packet overhead introduced by the different protocols. Another drawback of using small TCP segment sizes is the increase of the number of ACKs that are sent back to the transmitter. We recommend obtaining these two parameters values from the terminal in order to adjust the TCP model accordingly.

(iv) *Application level KPIs:* focused on obtaining some parameters that are required to estimate the QoE. One of the key issues when estimating the QoE is a proper identification of the main application performance metrics that affect the service quality, for example, number of rebufferings for streaming or end-to-end delay for VoIP. The knowledge of these application performance metrics may not be straightforward, but it may require to get lower layer QoS metrics in order to estimate them. Note that the process of mapping application QoS into user QoE may require the knowledge of some configuration parameters that cannot be estimated analytically. Such performance indicators/parameters are service-specific (web browsing, YouTube, VoIP) as described in Section 5. The way to measure some of these parameters is explained below.

(a) *Web page downloading time* (D)*:* there are two options to compute this metric. The most accurate option is based on monitoring and parsing HTTP packets. It is important to identify all HTTP transactions belonging to the same web page since the secondary objects contained in a web page might be located in external servers. The information related with the links where these objects are located is included in the main object. So, a possible way to find out all TCP connections related with the actual web page is searching all the links included in the main page HyperText Markup Language (HTML) code. Once all the segments belonging to the same transaction have been identified, the web page downloading time can be estimated as the time spent from the web page request to the last data segment received. Another simpler option is to estimate the web page downloading time from the lower layer throughput and web page size, which could be known a priori by reading the content-length HTTP header response.

(b) *End-to-end delay (d):* in order to compute the end-to-end delay, there are two options. The first option is to analyze the timestamps (if available) included in some packets. This information is included, for example, in real time control protocol (RTCP) packets, which is commonly used in standard VoIP service. In addition, this solution requires sender and receiver to be synchronized via network time protocol (NTP). Another drawback of this solution is that RTCP packet sizes may differ from the size of those packets containing the user data, leading to a difference between the measured delay (using RTCP packets) and the actual delay (corresponding to data packets). The second option is to approximate the end-to-end delay as the half of a RTT, which might only be measured at transport layer during the TCP connection establishment, or at network layer, as described before.

(c) *Loss probability at application level:* for RTP-voice services, loss detection can be based on the sequence number field included in the RTP header, checking for possible missing numbers in the incoming RTP flow.

(d) *Video buffer size:* this parameter is included in the request that the embedded player sends to the multimedia server for the download of the selected video. The "burst" field (within "videoplayback" list of parameters) indicates the

Table 2: Summary of KPIs to be measured at each protocol layer.

Layer	KPI/parameter
Signaling	Attach delay
	EPS bearer establishment delay
Transport	TCP advertised window (ADWN)
	TCP maximum segment size (MSS)
	Average RTT
Network	RTT variance
	IP packet loss rate
	IP packet sizes
Application	
Web browsing	Average throughput at application layer (r)
	Web page downloading time (d)
	DNS resolution time
VoIP (E-model)	End-to-end delay at application level (d)
	Loss probability at application level (e)
Video YouTube (HTTP/TCP)	Video buffer size (B_{full})
	Buffer length when the player paused (B_{empty})
	Video bitrate (λ)
	Video length (l)
	Average TCP goodput (β)
	Number of empty-buffer events (n_{rebuf})

buffer size in seconds, which is multiplied by the video data rate, provides the buffer size in bytes.

(e) *Video bitrate and video length:* these parameters are sent as metadata in the file downloaded from YouTube. This file is Flash Video (FLV) for the majority of non-High Definition clips and MP4 for High Definition clips.

A summary of the main KPIs to be measured at each protocol layer is listed in Table 2.

5. QoE Estimation Process

All KPIs obtained from different network elements are related to different layers below the application. For instance, KPIs measured at the gateways are mostly related to the network level, KPIs measured at a DPI may be associated to the network or transport level, whereas KPIs measured at the terminals can be associated to any level below the application. For that reason, the performance at lower layers received at the QoE server must be mapped onto application performance level, and ultimately, onto a QoE value.

We propose a methodology for estimating the QoS and QoE perceived by the user for different packet data services over wireless networks. The proposed methodology is based on network and protocol models, service-related parameters, and utility functions that map QoS objective metrics into the subjective experienced quality as perceived by the end user.

The modeling methodology follows a bottom-up approach, from the physical up to the application layer, taking into account the effects with a higher impact on the overall QoS. Therefore, layer i provides a set of performance indicators to the layer above ($i + 1$) and successively, up to the application layer. Specific equations that model each layer along the protocol stack is out of the scope of this paper although further details can be found in a previous work from one of the authors [15].

The final goal of this end-to-end model is to evaluate the application level QoS, which will be later mapped into QoE (in terms of MOS value), as shown in Figure 3. This last process is proposed to be performed by means of utility functions associated to each particular service. The goal of the utility functions is to map objective measurements (in terms of QoS) into subjective metrics (in terms of QoE perceived by the user).

Note that utility functions are very service dependent whereas MOS values will be estimated per QCI, which may aggregate different services according to the standard [16]. This may be a problem if the operator decides to aggregate a key data service with other services into the same QCI. If this is the case, it is highly recommended to use proprietary QCIs to keep separated the data services to be optimized.

This mapping process shall consider the specific characteristics of each data service. As an example, we focus on three different services:

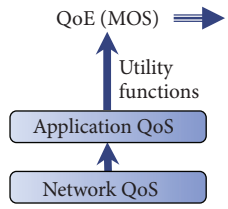
(i) *Web browsing:* the most important objective parameter to estimate the MOS in a web browsing session is the web page downloading time D. The utility function (utility functions are generally obtained through subjective tests to users, by varying the value of the application performance metrics under consideration) that estimates the MOS as a function of D (in seconds) is given by [17]:

$$\text{MOS} = 5 - \frac{578}{1 + (11.77 + 22.61/D)^2}. \tag{1}$$

(ii) *Video YouTube:* among the various works devoted to estimate the MOS for video services [18–20], the analysis presented by [18] provides a utility function for hypertext transfer protocol (HTTP) video streaming as a function of three application performance metrics: initial buffering time T_{init} (time elapsed until certain buffer occupancy threshold has been reached so the playback can start, measured in seconds), mean rebuffering time T_{rebuf} (average duration of a rebuffering event, measured in seconds) and rebuffering frequency f_{rebuf} (frequency of interruption events during the playback, measured in seconds^{-1}). The final MOS expression is given by

$$\begin{aligned}\text{MOS} = {} & 4.23 - 0.0672 T_{\text{init}} \\ & - 0.742 f_{\text{rebuf}} - 0.106 T_{\text{rebuf}}.\end{aligned} \tag{2}$$

Note that these application layer metrics (T_{init}, T_{rebuf}, and f_{rebuf}) can be estimated (at the receiver) from



MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Figure 3: Bottom-up approach to evaluate the QoE.

performance indicators at lower layers (like the TCP throughput) as well as other configuration parameters like video coding rate or buffer size at the receiver (see [18] for further details).

(iii) *VoIP:* in this case the MOS formula just maps the result given by an intermediate model into normalized MOS values. This intermediate model, known as the E model, is specified in [21], and it provides a numerical estimation $R \in [0, 100]$ of the voice quality from a set of network impairment factors related with the signal to noise ratio (SNR) of the transmission channel, delay, distortions introduced by the coding/decoding algorithms, packet losses, and so forth. In [22], a simplification of the E-model is provided, particularizing it for VoIP communications, where the voice quality R is given by the following expression: $R = 94.2 - 0.024 \cdot d - 0.11 \cdot (d - 177.3) \cdot H(d - 177.3) - I_{e\text{-eff}} + A$ being d the end-to-end delay in milliseconds, $I_{e\text{-eff}}$ the effective equipment impairment factor, $H(x)$ the unit step function, and A the correcting factor, which takes into account the environment where the communication takes place. Besides, [22] provides a formula to translate the R value into MOS:

$$\text{MOS} = 1 + 0.035 \cdot R + R \cdot (R - 60) \cdot (100 - R) \cdot 7 \cdot 10^{-6}. \quad (3)$$

The impairment factors, in turn, depend on the specific codec used for the VoIP communication; the values of these factors for a number of codecs are tabulated in [22, 23].

6. Potential Use Cases

There are many use cases that would benefit the rollout of a QoE-monitoring solution as proposed in this paper. The basic utility of the proposed architecture is to monitor the QoE associated to a particular data service and user. Once the QoE server has information about the specific QoE for that service, the mobile network operator may use such information for different purposes, some of them are described next.

6.1. QoE Estimation. The first use case is focused on the pure QoE evaluation process, including average numerical QoE results for the three services described in Section 5. Results have been obtained from simulations assuming a LTE network whose main configuration parameters (at all protocol layers) are summarized in Table 3. A QoE module is responsible for collecting network and application performance indicators and, afterwards, for mapping QoS onto QoE in terms of a MOS value (according to the utility functions) as described in Figure 3.

Regarding the web service analysis, the exchange of information is done via HTTP/TCP, where HTTP version 1.1 has been assumed. This version includes the persistent connection feature, which makes it possible to reuse the same TCP connection for downloading subsequent objects included in the web page. The optional pipelining feature has been also assumed, thus allowing a number of object requests to be simultaneously sent without waiting for the reception of the previous object. Figure 4 on the left shows the MOS results for different network RTTs and different number of secondary objects in the web page (from 2 to 50 objects of 20 kB each). Firstly, long RTTs lead to a worse TCP performance (in terms of throughput) as a consequence of its inherent congestion control mechanisms (both during slow start and steady-state phases). Such throughput reduction has a direct impact on the web page downloading time and MOS. Secondly, a higher number of objects in the web page (assuming equal sizes) leads to longer downloading times, thus degrading the MOS.

In the case of VoIP service, it usually relies on UDP as transport layer with a configurable voice-coding rate from around 6 kbps to 40 kbps. Due to the low data rates that a VoIP flow usually needs, throughput requirements at the network side are not usually an issue over an LTE network. Instead, the network performance indicator mostly affecting the service quality is the end-to-end delay. Taking into account the characteristics of the VoIP traffic, a robust header compression (RoHC) mechanism has been considered at the packet data convergence protocol (PDCP) layer. In addition, the RLC unacknowledged mode (UM) has been selected in order to minimize the end-to-end delay, which is the application layer metric that mostly affects the MOS. Figure 3 on the right shows the MOS results as a function of the one-way end-to-end delay and the voice coding rate. In the QoE computation formulae, a correcting factor (A) value has been set according to a cellular communication inside a building whereas the impairment factor $I_{e\text{-eff}}$ has been obtained from tabulated values [24] for selected voice codecs. It can be observed that the maximum end-to-end delay that

Table 3: Configuration parameters along protocol stack.

Layer	Parameter	Value
PHY	Carrier frequency	2 GHz
	System bandwidth	20 MHz
	Antenna configuration	1-layer MIMO 2 × 2 (beamforming)
	Precoding	LTE 4-words codebook
	Channel type	Extended pedestrian A (at 4 km/h)
	Channel estimation	Zero forcing
	MIMO detection	MMSE
	Target BLER	10%
	Control channel overhead	From 1 to 3 OFDM symbols
	Modulation/coding rate	16 CQI table (4 bits)
	Channel coding scheme	Turbo codes + SOVA
	Average SNR	20 dB
MAC	HARQ model	Incremental Redundancy + Chase Combining
	Scheduling method	Proportional fair
RLC	Maximum number retransmissions	1 for web and YouTube, 0 for VoIP
PDCP	Header compression	Enabled for VoIP, disabled for web and YouTube
IP	End-to-end delay (excluding the radio interface)	Variable (from 0 to 250 ms)
TCP	Maximum segment size	1460 bytes
	initial congestion window	1 segment
	Advertised receiver window	32 kbytes
	#ACKs per segment	1
	SYN timeout	3 s
APP		
Web	Web page size	100 kB text + variable # of objects of 20 kB each
	HTTP version	1.1 (persistent TCP connection)
YouTube	Video length	250 s
	Video coding rate	512 kbps
	Client data buffer necessary to start the playback	32 s
	Buffer threshold that triggers a rebuffering event	2 s
VoIP	Voice coding rate	8.85 kbps and 23.85 kbps

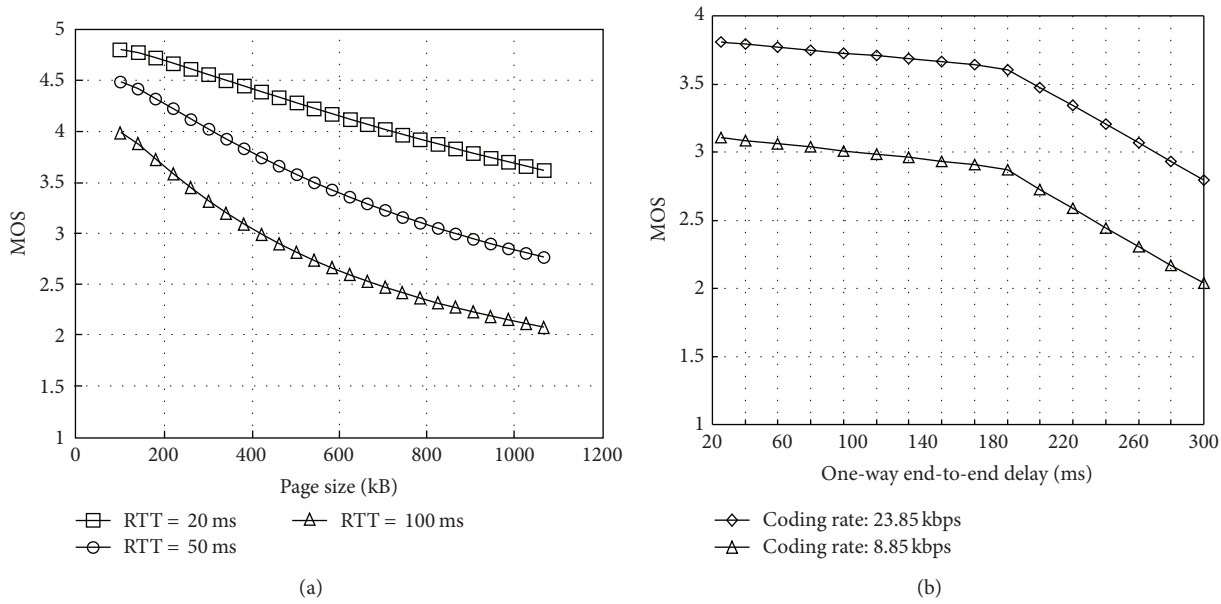


Figure 4: Average MOS results for web browsing (a) and VoIP (b).

makes it possible to obtain a fair quality (i.e., MOS = 3) is around 100 ms (for a coding rate of 8.85 kbps) and 270 ms (for 23.85 kbps).

YouTube service is based on progressive download technique over HTTP/TCP; that is, the client sends an HTTP request and, as a consequence, the YouTube multimedia server delivers the requested video through an HTTP response over TCP. According to (2), the MOS for YouTube depends on three application layer metrics (T_{init}, T_{rebuf}, f_{rebuf}), which can be estimated (at the receiver) from network performance indicators at lower layers (like the TCP throughput, end-to-end RTT, or packet loss rate) as well as other configuration parameter (available at the receiver side): TCP AWND size, video coding rate, video length, play-out buffer size at the receiver, or minimum buffer threshold that triggers a rebuffering event (see [8] for further details).

Figure 5 on the left depicts the results of the three application performance metrics for YouTube as a function of the network RTT. The upper subplot represents the achievable average TCP goodput (computed from [25]). So if the average TCP goodput is higher than the video coding rate (512 kbps), then the probability of rebuffering events will be negligible. As the RTT is increased, TCP goodput is decreased until it becomes lower than the video coding rate at certain RTT value; from this RTT value and above, the parameters related to the rebuffering events (T_{rebuf} and f_{rebuf}) are higher than zero (as shown in the lower subplot). The initial buffering time (T_{init}) is also increased for higher RTTs since lower TCP goodput values lead to longer delays to reach the minimum buffer occupancy (B_{full}). The rebuffering time (T_{rebuf}) has the same behavior although it is null as long as TCP goodput is above the video coding rate (i.e., no rebufferings occur). Besides, it can be seen that $T_{\text{rebuf}} < T_{\text{init}}$ for the same RTT value due to the following reasons: (1) the amount of data needed to be filled (B_{full}) for the computation of T_{init} is greater than the amount of data ($B_{\text{full}} - B_{\text{empty}}$) required for the computation T_{rebuf} and (2) the computation of T_{init} assumes that TCP data transfer starts with a slow-start phase whereas the computation of T_{rebuf} considers the TCP steady state to be reached (being the TCP goodput higher in this second phase). Figure 5 on the right shows the MOS results for different RTTs. As mentioned above, for low RTT values (which achieve TCP goodput values higher than the video coding rate), the initial buffering time is the only metric affecting the MOS (the higher the T_{init}, the lower the MOS). When the rebuffering events start to take effect over the MOS, its value is rapidly decreased since interruptions over the playback are very annoying for the users.

For this specific data service, in which the MOS depends on many configuration parameters available at the mobile terminal, it is required to monitor most of the parameters in the own terminal, and not in the network. However, network performance indicators might be monitored either in the terminal (preferable) or in a DPI.

Although previous results correspond to average MOS values, the estimation of the MOS in a real scenario shall be performed instantaneously in order to have real-time statistics about the user's QoE so that real time actions may be taken (see Figure 6). The evaluation of additional actions from the operator side is not under the scope of this paper although a brief description of potential use cases is given in next subsections.

6.2. QoE/QoS Optimization. The proposed architecture can be applied to estimate the QoE perceived by the end-user for new data services over a specific wireless network. In addition, the knowledge of the instantaneous and average QoE per user may help the operator to perform other actions like for instance the following.

(a) *Modification of subscriber priority:* when a poor performance in a specific location or particular subscriber is detected, the interaction of the QoE Server with the PCRF could be considered in order to prioritize network resource usage for each cell site and/or for each individual subscriber. Such indication could be fulfilled by, for example, modifying priority levels (in the proprietary database) associated to particular subscribers (ARP and/or QCI). In case of using a different QCI, it is recommended to use proprietary QCIs that distinguish between subscriber profiles, not between QCIs associated to different services.

(b) *Flexible bandwidth limits:* it allows operators to set dynamically different bandwidth limits depending on a number of factors like: data service (e.g., streaming, gaming, downloads, and email), usage patterns, subscriber, location, time of day, and so forth. Particular QoS policies may be used to optimize the allocation of available bandwidth across subscribers, increasing fairness in network access and improving the user experience, while still taking into account real-time subscriber preferences and behavior, and network conditions.

(c) *Enforce policy rules on many different enforcement points:* the coordination between QoE server and PCRF makes it possible to enforce policy rules in many different enforcement points including access gateways, DPIs, content optimization servers, and even, subscriber devices (or any other network element with access to the QoE solution). Although PCRF only has a direct communication with P-GW, S-GW, and DPI (acting as a PCEF), it would be also possible to set policy rules on mobile devices through the QoE server (for those mobile devices with an ad hoc application). This procedure would require the QoE server to implement a PCEF entity in charge of receiving the policies from the PCRF and, afterwards, forward them to the mobile devices. This allows service providers to apply policy rules throughout the network and support a diverse set of use cases.

(d) *Send notifications to subscribers:* the QoE server might send notifications triggered from the PCRF based on real-time events, such as exceeding a usage threshold for a specific application, roaming to another network, or qualifying for a customer loyalty program.

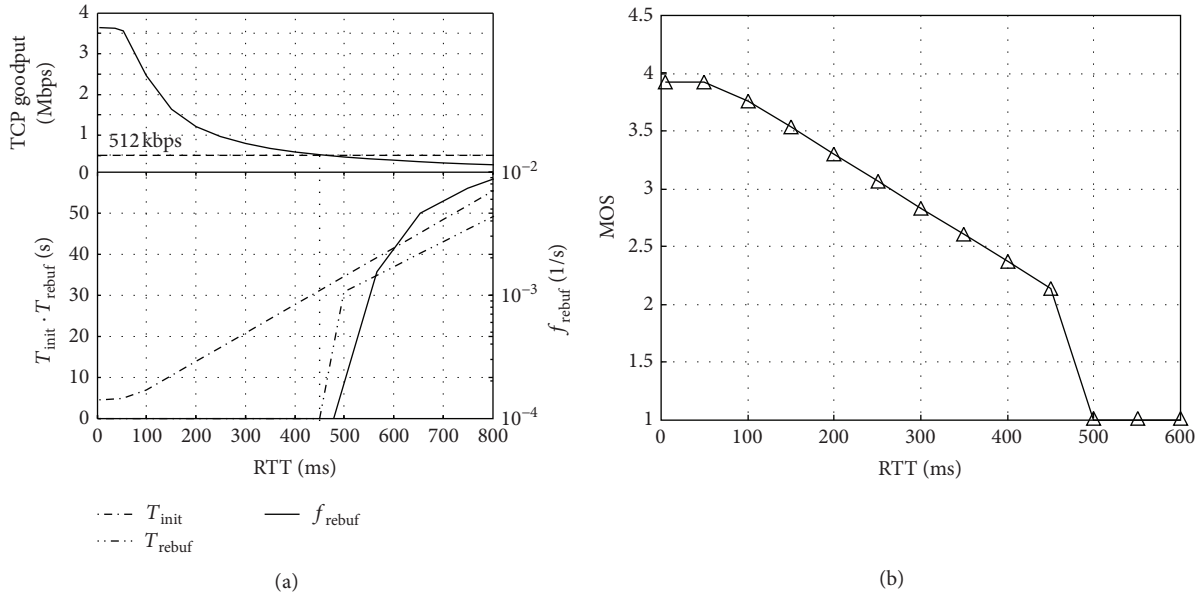


Figure 5: YouTube application performance metrics (a) and MOS (b) for different RTTs.

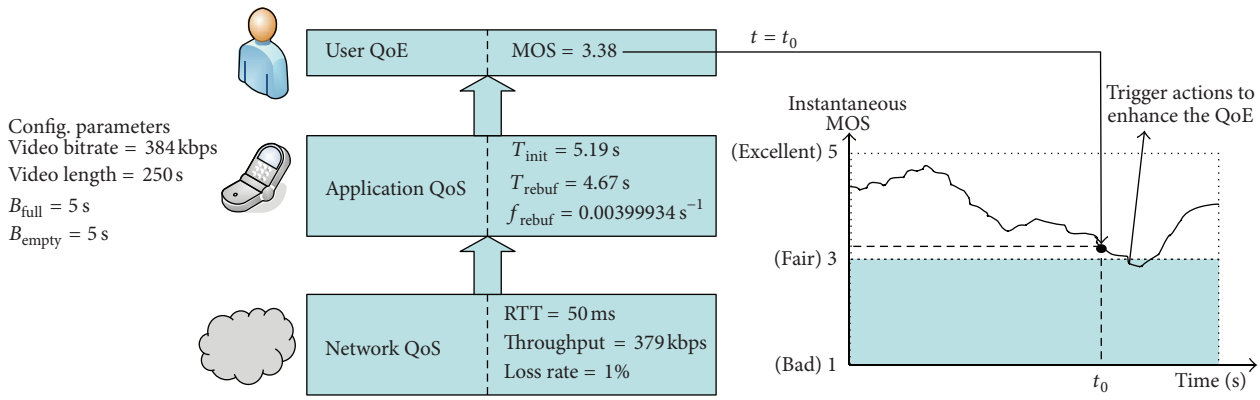


Figure 6: Example of MOS evaluation for YouTube service.

6.3. Network Capacity Planning. Capacity management is based on engineering limits which specify the maximum level of utilization that can be tolerated in order to provide the required quality of experience. For example, in voice legacy networks engineering limits are calculated by the use of Erlang's formula on base of maximum tolerable blocking rates. For mobile broadband networks, however, such a reliable and simple relation connecting quality with capacity does not exist.

The widely accepted processor sharing model only serves as an estimate on perceived throughput but fails in predicting quality for a heterogeneous service mix as it is observed in mobile internet traffic. Therefore, direct quality measurement has to play a more active role than just providing end-to-end control as usually employed for circuit-switched networks. Monitoring of resource utilization should therefore be complemented by monitoring of end-to-end quality, giving a complete view on QoE with regard to network topology and time. This will enable to build a reliable correlation between utilization and quality and thus serve as a basis for economically efficient capacity planning. This is particularly important for real-time data services like VoIP or video streaming.

The proposed QoE solution could help on the identification of the minimum resources required for the radio interface, the E-UTRAN and the EPC to achieve a desired QoE. With the aim of fulfilling the operator's end user quality requirements as well as minimizing CapEx and OpEx, our QoE solution could be used for both budget planning. Its end-to-end approach provides additional benefits by ensuring that all domains involved are consistently dimensioned across the whole network.

Device-based solutions are considered of interest for specific use cases (mostly precise location and device performance impact). Nevertheless, these solutions have significant limitations in terms of scalability and handset manufacturer

dependencies. Based on this, device-based solutions shall be considered as a complement to a network-based solution that may address specific needs on a sample of the network.

Note that network capacity planning is not a real time process, that is, it does not require quick actions as a consequence of certain events in the network. In that sense, a quick availability of performance indicators is not an issue. Additionally, statistics from the NMS database will help in the dimensioning process as it provides both network topology and traffic load information associated to each network element. Taking into account detailed information about customer usage and traffic/usage patterns, our proposed QoE solution would be able to perform for example, the following tasks.

(i) *Identification of network bottlenecks and (re)dimension the network to ensure the targeted QoE*: QoE measurements with full network coverage can improve efficiency of bottleneck identification and extend the capability of existing load monitoring in classifying the grade of congestion according to impact on quality. This analysis makes it possible to (re)dimension those network elements and links with potential problems. This process should provide optimum network configuration for the given requirements, expected traffic mix, and QoS profiles, and it is the previous step to troubleshooting. It includes a (cell-by-cell) dimensioning process of all network interfaces including the radio (both for user plane and control plane) using real network data.

(ii) *Traffic forecasting based on actual and historical data traffic*. The proposed QoE solution could be also used to perform a traffic forecasting process based on historical data traffic stored in its database. Concretely, this solution could implement forecasting algorithms to predict traffic demand in a per-cell basis based on historical data and on the expected global traffic growth. The goal is to estimate the amount of traffic in the future by spreading forecast market data traffic to the sector level, both in terms of total amount of traffic as well as the traffic mixture.

6.4. Handset and Service Performance Benchmarking. With the growing number of mobile handsets and multimedia content launched onto the market, it is becoming increasingly important for operators to benchmark each individual terminal and measure its performance. Detail insight onto how different handsets (smartphones) do perform within the network for different services and applications, as a manner to guide handset selection and certification, and potentially feeding into device commercial negotiations.

This process enables the identification of problematic handsets and analyses of the cause for the faults. By identifying problematic handsets, operators can quickly make the required adjustments to their network to provide support for more handset models, thus improving the customer experience.

Since our QoE solution will receive performance indicators from a set of mobile devices, it will store statistical reports for quality of a wide range of handsets. These reports show the QoS experienced by the handset user over time. In addition, they show handset usage trends enabling operators to optimize the support for various types of handsets.

The main goal of this task would be to provide the following.

(i) To benchmark handset performance: how voice and data are perceived from real handsets and subscribers' perspective from any point of the network. Our QoE solution will analyze handset performance from voice and packet service statistics over time and location.

(ii) To identify, analyze, and resolve problems linked to handsets.

(iii) Handset validation process: benchmark new handsets based on specific criteria (e.g., check that handsets models used by roamers are compatible). This enables a faster handset selection and validation process and contributes to reducing the need for expensive active testing and emulating hundreds of handsets.

(iv) To deliver the best QoE for new applications: new services such as video-streaming applications are an important source of revenue for operators. In order to ensure top quality data services, handset performance monitoring helps to test applications and measure the quality of experience perceived from a handset prospective. It is important to make sure that multimedia applications are fine-tuned for the handsets that use them the most. Furthermore, using this process, marketing team can easily follow up the introduction of new services and handsets and measure their usage.

6.5. Network Troubleshooting. Current network-monitoring tools may not be the best approach for systematic network troubleshooting when issues are detected on customer experience (further than related to pure network issues without clear correlation onto customer impact). The solution shall be able to provide with customable alarming thresholds setting for different indicator functions. Automatic threshold setting, trend-tracking mechanisms, and automatic/self-learning procedures for deviation tracking availability will be positively considered.

The real challenge comes in diagnosing network problems that impact customer experience. These problems may be specific to a particular cell, device, core network element, or application. In a large network with tens of thousands of sites, each using multiple bands and carriers, and linking to hundreds of core network elements and application servers, finding the one issue underlying a problem may require analysis of Terabytes of data.

Quality of service is the most important LTE troubleshooting feature which may give one vendor (or operator) the advantage over the other. To understand the QoS issues in their networks, operators need to have more than basic analysis capabilities of the network performance in any troubleshooting system they implement. Measuring QoS in all IP networks requires an evolved solution. The engineers need

appropriate KPIs (customizable built-in KPIs) to analyze service setup, service quality with dropped sessions, issues, and causes. The network's need to deliver high bandwidth data services, directly influences the capacity required of the monitoring solution and the ability needed to determine which subscribers are using the network and what services are being used.

Several nodes and interfaces are involved in transmitting the subscriber's identifiers which are used for processing the policy and charging control in LTE networks. Any failure to process this information correctly generates mistakes which come to light in the bottom line—customer satisfaction and the billing system. Therefore, it is essential that an operator is able to troubleshoot the relevant nodes preventing both service degradation and loss of income.

Our proposed QoS solution is able to receive real-time KPI alarms, which provide an at-a-glance overview of network and service performance degradations. Automatic notification of problems in the network helps to solve network failure faster, even before the subscribers are aware of such problems.

6.6. Network Monitoring and Reporting. Typically, network monitoring process is mostly based on overall network performance indicators, so that perceived experience by the end customer is not possible to be easily derived on global scale. The proposed QoE solution would allow for a combined network plus customer experience monitoring (based on QoE), being able to anticipate specific customer/service issues and reduce business impact. Detailed information about customer usage and traffic/usage patterns, which is considered of vital importance for both customer business department and for evolving towards increased level of segmentation into multiple dimensions (customer, service, etc.) is based on real trends.

The proposed QoE solution could use passive methods to infer automatically from passive measurements the user perception on the network. The goal would be to automatically derive user perception, from specific indicators being accessed purely from monitoring (eliminating the need for customer surveys) both from the network and terminal sides.

6.7. Customer Care. Currently, customer perception is evaluated mostly via periodical questionnaires and interviews with selected customers that provide views/insights onto perceived experience. The ability of linking perceived (subjective) experience with measured (objective) QoE indicators may lead to significant benefits in terms of achieving a better insight onto customer perceived quality in a much more wide approach than current one based on sampling of specific customers—evolution towards full network customer quality tracking.

QoE-monitoring solutions are linked onto Customer Care centers by means of simplified interfaces and overall status for real-time access to customer specific information, enhancing the response to customer quality and thus satisfaction. Customer care teams can rapidly diagnose problems and identify whether the root cause is linked to a badly performing network, mobile terminal, or application. This makes it possible to identify problems before they affect customers communicating more proactively thereby increasing overall customer satisfaction.

A of this use case, related to the potential active remote handling of devices (e.g., accessing remotely the PC to determine configuration issues), has been identified.

7. Conclusions

In this paper, we have proposed a novel architecture for providing QoE awareness to mobile operator networks. The proposed architecture makes it possible to link QoE engine and PCRF with the aim of achieving a dynamic control of QoS based on customer perception. Combining sophisticated metering capabilities with a highly configurable business rules engine, the PCRF can manage the QoS, optimizing high bandwidth traffic, and enforcing usage quotas. The communication between both entities could be fulfilled through the following alternatives: (a) via Gx reference point or (b) via Sp reference point (through a proprietary database). Several use cases (that take the advantage of such coordination) have been proposed, including the modification of subscriber priority for future bearer establishments, dynamic configuration of bandwidth limits, enforcement of policy rules on many different enforcement points, or the possibility to send notifications to mobile terminals.

Regarding other potential applications of the proposed QoE solution (without requiring interaction with the PCRF), other use cases that would benefit the rollout of a QoE-monitoring platform solution have been described, including network capacity planning, handset and service performance benchmarking, network troubleshooting, network monitoring and reporting based on QoE, and customer care.

Future work will be focused on a feasibility study towards a real-world practice over a LTE testbed. The final goal is to provide practical results and experience on dynamic QoE provisioning in EPS systems.

Acknowledgments

This work has been partially supported by the Junta de Andalucía (Proyecto de Excelencia TIC-06897 and TIC-03226) and by the Spanish Government (TEC2010-18451).

References

[1] G. Gomez and R. Sanchez, *End-To-End Quality of Service Over Cellular Networks: Data Services Performance Optimization in 2G/3G*, John Wiley & Sons, 2005.

[2] A. Díaz, P. Merino, and F. J. Rivas, "Customer-centric measurements on mobile phones," in *Proceedings on 12th IEEE International Symposium on Consumer Electronics (ICSE '08)*, pp. 14–16, 2008.

[3] J. K. Pathak, S. Krishnamurthy, and R. Govindarajan, "System, method, and apparatus for measuring application performance management," Patent number: WO03007115, 2003.

[4] R. Schatz, T. Hossfeld, and P. Casas, "Passive youtube QoE monitoring for ISPs," in *Proceedings of the 6th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS '12)*, pp. 358–364, July 2012.

[5] I. Ketykó, K. De Moor, W. Joseph, L. Martens, and L. De Marez, "Performing QoE-measurements in an actual 3G network," in *Proceedings of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB '10)*, pp. 1–6, March 2010.

[6] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, vol. 24, no. 2, pp. 36–41, 2010.

[7] S. Egger, T. Hossfeld, R. Schatz, and M. Fiedler, "Waiting times in quality of experience for web based services," in *Proceedings of the 4th International Workshop on Quality of Multimedia Experience (QoMEX '12)*, pp. 86–96, July 2012.

[8] T. Hossfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen, "Initial delay vs. interruptions: between the devil and the deep blue sea," in *Proceedings of the 4th International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, July 2012.

[9] T. De Pessemier, K. De Moor, A. Juan, W. Joseph, L. De Marez, and L. Martens, "Quantifying QoE of mobile video consumption in a real-life setting drawing on objective and subjective parameters," in *Proceedings of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB '11)*, pp. 1–6, June 2011.

[10] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens, "Quantifying the influence of rebuffering interruptions on the user's quality of experience during mobile video watching," *IEEE Transactions on Broadcasting*, no. 99, pp. 1–5, 2013.

[11] F. Wamser, D. Staehle, J. Prokopec, A. Maeder, and P. Tran-Gia, "Utilizing buffered YouTube playtime for QoE-oriented scheduling in OFDMA networks," in *Proceedings of the 24th International Teletraffic Congress (ITC 24' 12)*, pp. 1–8, September 2012.

[12] G. Aristomenopoulos, T. Kastrinogiannis, V. Kaldanis, G. Karantonis, and S. Papavassiliou, "A novel framework for dynamic utility-based QoE provisioning in wireless networks," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '10)*, pp. 1–6, 2010.

[13] J. Sterle, M. Volk, U. Sedlar, J. Bester, and A. Kos, "Application-based NGN QoE controller," *IEEE Communications Magazine*, vol. 49, no. 1, pp. 92–101, 2011.

[14] D. Soldani, "Means and methods for collecting and analyzing QoE measurements in wireless networks," in *Proceedings of the International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM '06)*, pp. 531–535, June 2006.

[15] G. Gómez, J. Poncela González, M. C. Aguayo-Torres, and J. T. Entrambasaguas Muñoz, "QoS modeling for end-to-end performance evaluation over networks with wireless access," *EURASIP Journal on Wireless Communications and Networking*, vol. 2010, Article ID 831707, 2010.

[16] "TS 23. 203 V8. 11. 0. 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects," Policy and charging control architecture (Release 8).

[17] P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, P. Mogensen, and J. M. Lopez-Soler, "QoE oriented cross-layer design of a resource allocation algorithm in beyond 3G systems," *Computer Communications*, vol. 33, no. 5, pp. 571–582, 2010.

[18] R. K. P. Mok, E. W. W. Chan, and R. K. C. Chang, "Measuring the quality of experience of HTTP video streaming," in *Proceedings of the 12th IFIP/IEEE International Symposium on Integrated Network Management (IFIP/IEEE IM' 11)*, 2011.

[19] T. Porter and X. Peng, "An objective approach to measuring video playback quality in lossy networks using TCP," *IEEE Communications Letters*, vol. 15, no. 1, pp. 76–78, 2011.

[20] I. Ketykó, K. Moor, T. Pessemier et al., "QoE measurements of mobile Youtube video streaming," in *Proceedings of the 3rd Workshop on Mobile Video Delivery (MoViD' 10)*, 2010.

[21] "ITU-T recommendation G. 107," The E-model, a computational model for use in transmission planning, 2009.

[22] R. G. Cole and J. H. Rosenbluth, "Voice over IP performance monitoring," *ACM SIGCOMM Computer Communication Review*, vol. 31, no. 2, pp. 9–24, 2001.

[23] "ITU-T amendment 1 for recommendation G.113," Amendment 1: revised appendix IV—provisional planning values for the wideband equipment impairment factor and the wideband packet loss robustness factor, 2009.

[24] "ITU-T recommendation G. 113," General recommendations on the transmission quality for an entire international telephone connection, 2007.

[25] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, "Modeling TCP throughput: a simple model and its empirical validation," *ACM SIGCOMM Computer Communication Review*, vol. 28, no. 4, pp. 303–314, 1998.