Implementation of VoLTE considering different QoS and QoE overall approaches

João Cardoso
Instituto Superior Técnico
University of Lisbon
Lisbon, Portugal
joao.rodrigues.cardoso@tecnico.ulisboa.pt

Luis M. Correia
Instituto Superior Técnico / INOV-INESC
University of Lisbon
Lisbon, Portugal
luis.m.correia@tecnico.ulisboa.pt

***Abstract*— The main goal of this thesis is to study the impact in other services of deploying VoLTE over an existing LTE network while monitoring VoLTE call quality. The QoS performance of seven distinct services is analysed in terms of allocation delay, packet failure and throughput satisfaction. QoE for VoLTE users is assessed using the E-model. For that purpose, a single-cell model for the DL resource allocation in LTE was proposed and implemented on a time based simulator. Several parameters were analysed, namely number of users, type of environment, bandwidth and other service-related parameters. As VoLTE is the highest priority service, call quality is barely affected for realistic numbers of users. For the reference urban scenario being considered, one concludes that up to 60 and 36 simultaneous VoLTE users are supported for the AMR-WB and EVS codecs, respectively, without significant impact in other services. Regarding system bandwidth, 10 and 20 MHz bandwidths were compared, being observed that the offered throughput scales linearly with the system bandwidth. The cell throughput for the urban scenario is 38% above the rural one and no significant difference is observed between urban and suburban ones. Finally, the service penetration influence was analysed by defining a VoLTE centric scenario and a Video centric one. In the Video centric scenario, the percentages of satisfied users for the services with a lower priority than video streaming stay below 50%. For the VoLTE centric scenario, user satisfaction levels are above 90%.**

Keywords- LTE, VoLTE, resource allocation, QoS, delay, satisfied users

I. INTRODUCTION

Mobile communications systems were designed from the very beginning, since the analogue First Generation (1G), with the goal of providing voice services. With the arrival of the Second Generation (2G) using digital data signalling, a wider range of voice services became available to a big number of users, with Global System for Mobile Communications (GSM). In a later stage, data services were added but voice remained as the main traffic source. The Third Generation (3G) emerged with the intent of providing high-speed packet switched data transfer following the abrupt growth of popularity of the Internet in the fixed networks, bringing Universal Mobile Telecommunications System (UMTS) as the evolution of GSM according to the specifications of the Third Generation Partnership Project (3GPP).

Long Term Evolution (LTE) appears as 3GPP's Fourth Generation (4G) solution, based on an entirely packet switched oriented architecture aimed at fulfilling the exponential demand for Internet Protocol (IP) data services. In 2000, 3GPP standardised the IP Multimedia Subsystem (IMS), a framework aimed at providing multimedia services like voice, over 3GPP systems. However, implementation aspects like session setup, authentication or bearer setup were left for the service providers and vendors to decide upon. In November 2009, the One Voice initiative was established, proposing a standard solution for an IMS-based Voice over LTE (VoLTE) service [1]. In March 2010, the GSM Association (GSMA) specified a set of requirements to provide Voice over IP (VoIP) calls over LTE [2].

VoLTE constitutes an operator and user friendly solution when compared to third-party VoIP which poses several disadvantages. Traffic generated by applications like, for example, Skype does not differ from any other IP-based application, turning the service performance dependent of what the Internet can provide. This of course has an impact in the Quality of Experience (QoE) perceived by the end user as the network is not able to ensure a minimum guaranteed bit rate as well as a maximum value of end to-end delay. In the other hand, VoLTE using the IMS allows Quality of Service (QoS) control as the User Equipment (UE) is able to specify multiple performance requirements. With this functionality, the network can prioritise voice packets over data ones, which are not time critical. VoLTE should be considered as not only just a migration from the traditional voice services to an all-IP technology but as way to integrate several services taking advantage of the IMS capabilities. These services can be: High Definition (HD) voice, video communications, IP messaging, content sharing between calls, among others [3].

Operators face several challenges in the time of replacing the current voice service delivered through the circuit-switched technologies of systems prior to LTE as they have proven to be successful in the past. However, this replacement is somehow inevitable as the demand for data services is growing at a high pace and, at a long term perspective, merging all the services to packet-switched technologies will result in a more cost effective solution. At least, the implementation of VoLTE is supposed to deliver the same levels of performance when compared to the existing technologies. This work focuses on evaluating the

impact in other services of deploying VoLTE over an existing LTE network and monitoring the performance of VoLTE calls. The method to evaluate these parameters is through a model of the downlink (DL) resource allocation in LTE considering a set of seven different services including VoLTE. The proposed model takes into account several input variables which are related to the cell environment, the cell bandwidths and parameters related to the users' behaviour like the rate of arrival of users to the cell and the usage pattern of the considered set of services.

The paper is organised as follows: Section I – Introduction; Section II - State of the art; Section III – Models and simulator description; Section IV – Results analysis; Section V – Conclusions.

II. State of the art

The implementation of VoLTE in the context of already deployed LTE networks increased the interest of the scientific community in aspects related to DL scheduling algorithms as these heavily influence how resources are distributed among the network users. In [4], the authors show the importance of service prioritisation for the performance of delay-critical services like VoIP in the presence of concurrent Best Effort traffic like web browsing. The study concludes that capacity gains of 105 to 710% can be obtained in terms of VoIP capacity at the cost of small capacity losses on the other services. The gain variation is essentially justified by the different channel conditions of the users.

In [5], a Channel and QoS Aware (CQA) scheduler in time and frequency domains is proposed to enhance the capacity of VoLTE systems and a comparison with schedulers of the same nature is presented. In the time domain, a metric that prioritises users with the highest value of Head of Line (HOL) delay. In the frequency domain, the proposed metric has the purpose of providing to all the data flows the same level of QoS in terms of delay and bit rate by prioritising flows with higher HOL delay and a larger ratio of achievable throughput and past throughput. The authors compare this scheduler with the Priority Set Scheduler (PSS) which is similar to the CQA scheduler but does not consider HOL delay and the HOL scheduler which only considers the HOL delay as the scheduling metric. They conclude that the proposed scheduler outperforms the PSS and HOL schedulers in 20 and 100% in terms of VoLTE capacity, for a realistic pedestrian scenario.

In what refers to VoLTE performance, in [6], the author performs simulations to measure several KPIs for four different scenarios with different network congestion conditions which reflects in different values for the available link bit rate. It is concluded that for link utilisations above 75% the performance is still acceptable as the obtained MOS reflects a speech quality perceived by the end user which is "not annoying". If this percentage lies under 50% speech quality becomes "slightly annoying", with an increase of the end-to-end delay.

In [7], the performance of VoLTE is evaluated in conditions where the transport network is congested with data traffic and no QoS prioritisation is considered. The considered scenario divides into a first case where only voice traffic is generated that is compared to a second case where both voice and File Transfer Protocol (FTP) traffic are generated. End-to-end delay and jitter remain constant in the first case as the peak traffic is never reached, even for a worst case with 100 users. The same is not true for the second case where the network becomes congested and the VoLTE packets are queued while FTP packets are processed. In this case, delay can reach values around 350 ms in the worst conditions, which is not tolerable.

In [8] the authors intend to propose a more realistic approach to the analysis of QoS and QoE in LTE networks, using VoLTE as a use case, by experimenting in a real context with real devices, services, and radio configurations. Moreover, they perform cross-layer measurements by correlating the radio configurations with the QoS parameters measured at the application layer. The authors presented an experimental testbed to be used for multiple scenarios that included a LTE test base station able to emulate features like for example channel propagation and which supported connection to real LTE devices. The use of realistic impairments like fading and noise produces noticeably variable results which cannot be obtained through non simulated results like many of those that are available in the literature. Correlations below 0.8 were obtained when low level parameters and IP parameters are compared.

Among many other parameters already mentioned, the authors show the relationship between the received Signal-to-Noise Ratio (SNR) and the packet loss rate, concluding that the radio conditions in the receiver's side can impact the amount of losses as for a variation of 5 dB in the SNR can result in a variation of the packet loss rate of approximately 1%. The authors also show that a linear estimation can be obtained to relate MOS to packet losses and that for packet losses close to 0% the PESQ algorithm outputs the maximum quality which corresponds to a MOS of 4.5.

III. Models and simulator description

A. Model description

In order to evaluate the impact of the implementation of VoLTE over other services, one considers a single-cell model where all users are connected to a single base station, each of them performing a given service. This model is based on the resource allocation of the DL traffic generated in the network, describing its time evolution for every TTI at the packet level for different types and classes of services. Interference and handover issues are not taken into account for simplification purposes as they do not constitute the main focus of this thesis which is aimed at the cell level and not at the cellular network one. From a high-level perspective, the model is composed of three core parts as illustrated in Figure 1.

Input parameters are classified as Scenario, Network and User ones. Scenario parameters refer to the characterisation of the cell environment namely the type of environment, the associated SNR parameters and the percentage of indoor users. For this work, three types of environment are considered: rural, suburban and urban. Network parameters refer in this context to the configured radio channel bandwidth as this is one of the main aspect that determines the system's capacity together with SNR. User parameters refer to all the aspects involved in traffic generation. The number of users determines the average rate of user arrivals. Each user is associated with the service he is

requesting based on the service penetration which defines the number of users performing each service. The service parameters characterise the behaviour of the network traffic generated by the different services.

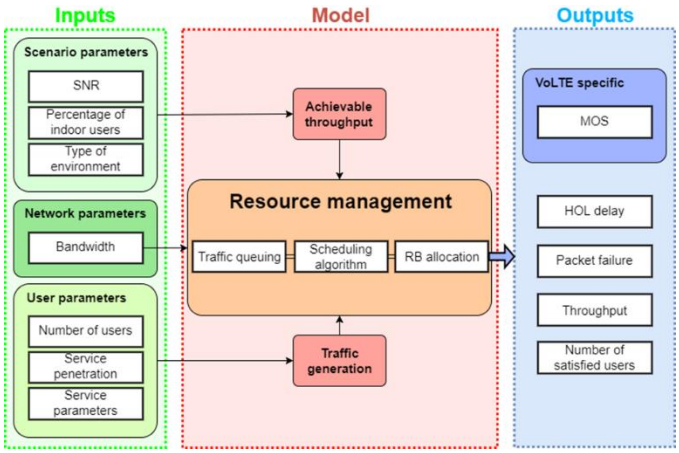


Figure 1. Model architecture.

The proposed model consists of three main steps: calculation of the achievable throughput, traffic generation and resource management. The achievable throughput per RB in every TTI is estimated based on each user's experienced SNR. It is assumed that the SNR is described by a Log-Normal distribution that depends on the considered type of environment, e.g., rural, suburban or urban. The estimated value of SNR is used to compute the achievable transmission rate considering the mathematical formulation developed by [9] to approximate the throughput of a single RB. This formulation considers a network using MIMO 2x2 for QPSK, 16-QAM and 64-QAM modulations.

For traffic generation, users are associated to the network according to a Poisson process with a specified average rate of users. Traffic for each user and the corresponding service is modelled through appropriate traffic source models, based on [10]. For the purpose of this work, seven services are considered, namely: VoLTE, video calling, video streaming, music streaming, web browsing, file transfer and e-mail. One considers an ON-OFF model for VoLTE calls. Streaming services, namely video and music streaming, are also implemented using an adaption of this model for convenience at the implementation level as they share common characteristics with the behaviour of voice traffic, especially the fact that both use a fixed framerate. For the video calling service, the Gamma Beta Auto-Regressive (GBAR) [11]. For the non-conversational applications, namely web browsing, file transfer and e-mail, the model described by [10] for WWW sessions is used.

The core step of the model deals with resource management in order to schedule users at the radio interface level. In a first instance, network generated traffic is queued for each user through an individual First-In-First-Out (FIFO) queuing system. The model for queue management at the eNodeB level consists of one queue buffer per user, with $Q_k[n]$ representing the total number of bits for all the packets in queue for user *k* at the time instant *n*. Each queue is updated whenever a packet arrives at the eNodeB. In parallel with this process, a timer is started for each packet received in the eNodeB. This timer is updated every TTI *n*, allowing the computation of the HOL delay $\tau_{HOL}[n]$, which represents the amount of time spent by the first packet to be transmitted. Figure 2 illustrates the algorithm to manage user queues.

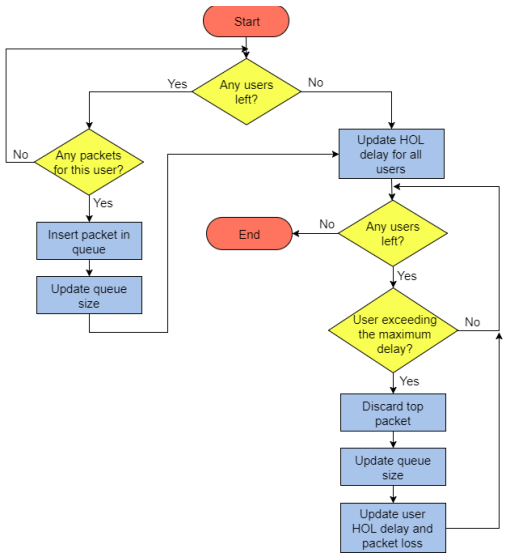


Figure 2. Queue management algorithm.

Generated traffic is allocated according to a scheduling algorithm that manages RB allocation according to the system's capacity. RBs are assigned at every time instant *n* in order with the maximum number of available RBs determined by the radio channel bandwidth. In a first approach, the total number of RBs required to allocate all the traffic in queue for each user is computed. The number of RBs needed to accommodate all the queued traffic for each user *k* is estimated as:

$$\widetilde{N}_{RB,k}[n] = \left\lceil \frac{Q_k\,[\text{bits}][n]}{R_{b,RB}\,[\text{kbps}][n] \times T_{TTI}\,[\text{ms}]} \right\rceil \quad (1)$$

where:

- $R_{b,RB}$: Achievable RB throughput.
- T_{TTI}: TTI (equal to 1 ms).

After that, system capacity is checked as the number of RBs requested by all users cannot exceed the maximum number of RBs available at each TTI. If the cell load is above 100%, the total number of requested RBs is larger than the system capacity and a reduction must occur. To manage this need for optimising RBs allocation, a scheduling algorithm composed of two levels of optimisation runs iteratively until system capacity is not exceeded. At a higher level, users requesting for RBs are distinguished according to the priority of the service they are performing. This means that the set of users performing a service which is associated to the lowest priority of all the existing services are the first to be reduced. Then, for each of these sets of users performing a given service, further optimisation must occur in order to distinguish among these users. At this point, a concept inspired by the Proportional Fair algorithm is introduced [11]. The main premise is that the number of RBs assigned to each user is proportionally reduced among each set of users.

Once it is guaranteed that the system capacity is not exceeded, RBs are assigned to the requesting users. After this assignment, the state of the traffic queues is updated coherently. For that purpose, for every TTI n that a user k is successfully scheduled, the total number of bits in the queue of user k is updated as:

$$Q_k[n]_{[\text{bits}]} = Q_k[n-1]_{[\text{bits}]} - (N_{RB,k}[n] \times R_{b,RB\ [\text{kbps}]}[n] \times T_{TTI\ [\text{ms}]}) \quad (2)$$

where:

- $N_{RB,k}$: Number of RBs allocated to user k.

Figure 3 shows the developed algorithm to allocate RBs in every TTI. The variables involved in this algorithm that where not defined up to this point have the following notation:

- P_{max}: Higher Priority level.
- $N_{RB,reduce}$: Sum of all the RBs that must be reduced.
- $N_{RB,usersub}$: Number of RBs that must be reduced for each user.
- $N_{RB,max}$: Maximum number of RBs per TTI.

Finally, the model outputs are evaluated by using several metrics for the performance of all the services and VoLTE specifically. The performance of each service is assessed based on the experienced throughput, queuing delay, packet failure associated to the DL resource allocation and the number of satisfied users. Besides these parameters, QoE for VoLTE is evaluated based on the calculation of the MOS using the E-model with estimations of the end-to-end delay and packet loss.

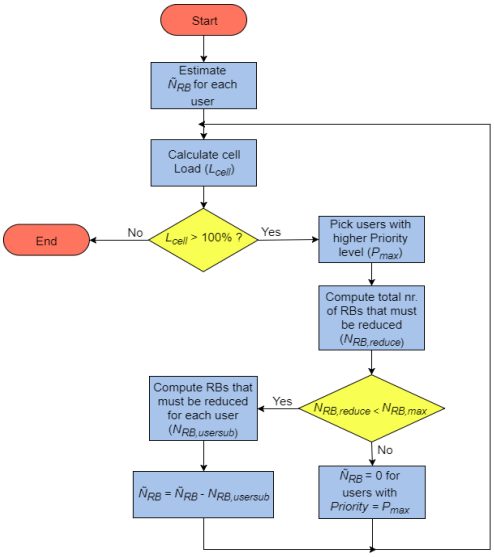


Figure 3. RB allocation algorithm.

B. QoS and QoE metrics

To approach the measurement of QoS and QoE, a distinction between VoLTE and other services is made in the sense that the considered metrics are divided among those which allow a qualitative characterisation of VoLTE and those which provides a measure of the overall performance of any service and the corresponding degree of satisfaction from the user viewpoint. As the quality of VoLTE as a real time application is deeply affected by network delay, one considers the HOL delay of each voice packet to assess the queuing delay due to the DL resource allocation. The average HOL delay for all users performing a service s, is evaluated as:

$$\overline{\tau_{HOL,s}}_{\ [\text{ms}]} = \frac{1}{N_{u,s}} \sum_{k=1}^{N_{u,s}} \overline{\tau_{HOL,k}}_{\ [\text{ms}]} \quad (3)$$

where:

- $N_{u,s}$: Number of users performing service s.
- $\overline{\tau_{HOL,k}}$: Average HOL delay of the packets received by user k.

Using the delay budget analysis considered in [12], an estimation of the end-to-end delay of a user k for a VoLTE call using the AMR-WB codec is given by:

$$\tau_{k\ [ms]} = \overline{\tau_{HOL,k}}_{\ [\text{ms}]} + 105 \quad (4)$$

Packet failure is used to measure the percentage of transmitted packets which had transmission problems due to excessive delay and would require a retransmission, being defined as:

$$\Gamma_{failure\ [\%]} = \frac{N_{sent} - N_{received}}{N_{sent}} \times 100 \quad (5)$$

where:

- N_{sent}: Number of packets sent.
- $N_{received}$: Number of packets received.

To achieve an estimation of the QoE at the user end, one adopts the E-model defined in [13] by converting its output, the transmission rating factor R, to a MOS which reflects the level of user satisfaction in terms of the perceived voice call quality. A simplified model can be applied to voice calls that use the AMR-WB codec, as it is the case for VoLTE, where R is simplified as [14]:

$$R = 129 - I_{d,wb} - I_{e,eff,wb} \quad (6)$$

where:

- $I_{d,wb}$: Delay impairment factor.
- $I_{e,eff,wb}$: Equipment impairment factor.

The value of $I_{d,wb}$ depends only on the experienced end-to-end delay and is obtained as [14]:

$$I_{d,wb} = \begin{cases} 0.024\tau_{[\text{ms}]} & ,\tau < 177.3 \text{ ms} \\ 0.024\tau_{[\text{ms}]} + 0.11 \times (\tau_{[\text{ms}]} - 177.3) & ,\tau \geq 177.3 \text{ ms} \end{cases} \quad (7)$$

The value of $I_{e,eff,wb}$ is given by [14]:

$$I_{e,eff,wb} = I_{e,wb} + (129 - I_{e,wb}) \frac{\Gamma_{failure\ [\%]}}{\Gamma_{failure\ [\%]} + B_{pl}} \quad (8)$$

where:

- $I_{e,wb}$: Equipment impairment factor without any packet loss (11.0 as proposed by [15]).
- $I_{e,eff,wb}$: A codec-specific factor which characterises its robustness against packet loss (13.0 as proposed by [15]).

The MOS is then given by [14]:

$$\text{MOS} = \begin{cases} 1 & ,R_x < 0 \\ 1 + 0.035R_x + R_x(R_x - 60)(100 - R_x) \times 7 \times 10^{-6} & ,0 \le R_x \le 100 \\ 4.5 & ,R_x > 100 \end{cases} \quad (9)$$

where:

$$R_x = \frac{R}{1.29} \quad (10)$$

In what refers to all the other services besides VoLTE, performance is measured by considering the user throughputs and the level of satisfaction associated to them. To obtain the total cell throughput during a given period of time one computes:

$$R_{b,eNodeB\ [\text{Mbps}]} = \frac{N_{bits}}{T_{sim\ [\text{s}]} \times 10^6} \quad (11)$$

where:

- N_{bits}: Total number of bits transmitted.
- T_{sim}: Simulation time.

Finally, the number of satisfied users served by the cell is also a relevant metric. A user is considered to be satisfied if its minimum throughput can be guaranteed.

C. Model implementation

A time based simulator was developed to allow the analysis of the network during a given period of time. This simulator was implemented using MATLAB and it operates with a resolution of 1 ms, corresponding to the TTI in LTE.

The first step in the simulation loop corresponds to the update of all the variables related to the environment. For every second, the speed of each user is updated and the corresponding coherence time is estimated. The coherence time will determine how fast the SNR will change and consequently the achievable throughput per RB. After all the scenario variables are updated, traffic is generated by using the three models used to implement the services for this study. The ON-OFF model consists of intermittent sequences of active states where packets are generated at fixed framerates and silent states where no packets are generated (for the streaming case) or SID frames are generated (for the VoLTE case). Each session is limited by the time duration T_{call} which is counted since the arrival of the user to the network. The GBAR model used for the video calling service is simply implemented by generating for each user, at a fixed framerate, a packet with a size given by the stochastic process described in [11]. The non-conversational algorithm consists of generating sequences of Hypertext Transfer Protocol (HTTP) objects in the web browsing case and files in the file transfer and e-mail cases.

A queue management block is responsible for updating the state of the user queues. Generated packets are inserted in the end of each queue and the HOL delay is updated according to the current simulation time. At the end, packet failure statistics are updated to account for the number of lost packets.

Once all the generated traffic is queued it is ready to be scheduled through RB allocation. In a first stage the cell load is estimated based on the number of RBs needed to allocate all the queued traffic. When the capacity is exceeded, the algorithm performs RB reduction which corresponds to the strategy described in Section III.A. After the number of RBs was determined taking the cell capacity into consideration, the RBs are allocated to the corresponding users and all the output statistics are updated, namely the user throughput, average delay and the number of allocated packets.

D. Simulator assessment

Several tests were done to assess the simulator. Aspects related to the user environment and the calculation of the achievable throughput were assessed. Globally it was observed One checked that the SNR follows the desired Log-Normal distributions and that the achievable RB throughput in each TTI is correctly computed. Traffic generators were also assessed to ensure that the three traffic source models are working properly. Queues for DL traffic were tested by generating traffic and defining a fixed value for the global throughput of the queues.

In terms of resource allocation, one verified that the radio channel bandwidth allows the allocation of the expected number of RBs per TTI. The LTE peak data rate was also tested by considering a scenario with a single static user in very good radio conditions with a fixed SNR of 40 dB and 100 RBs during the entire simulation period. As users are considered to be arriving at an average arrival rate described by a Poisson process, the number of simultaneous active users is monitored to check its behaviour. Figure 4 shows the behaviour of the system in terms of the simultaneous number of users during a 75 minutes period.

The minimum simulation time is set to one hour. This value allows enough time to achieve a statistically relevant sample of the time evolution of the system. According to the observed results, one decides to neglect the first 15 minutes of simulation to fully filter all the transient effects. Due to the variation observed in terms of the global throughput one verifies that three simulations are enough to correctly describe a simulation scenario as they ensure an error lower than 10% relative to the average value.

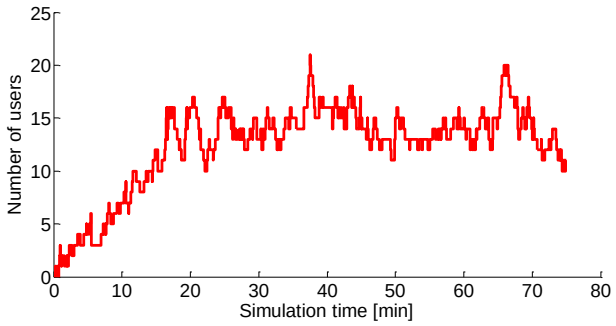


Figure 4. Time evolution of the number of simultaneous active users.

IV. RESULTS ANALYSIS

A. Reference scenario

The reference scenario for this study is a cell operating in an urban environment with 80% indoor users. The input parameters used for the reference scenario simulation are detailed in Table 1.

Table 1. Input configuration for the reference scenario.

Input parameter	Description/Value
Average number of users	150/hour
Type of environment	Urban
Percentage of indoor users	80%
Bandwidth	20 MHz
VoLTE codec	AMR-WB 12.65

The analysis focuses on the impact of the number of users and bandwidth on the level of satisfaction of VoLTE users. Additionally, one analyses the impact of the variation of the number of users on the performance of the multiple services by comparing the reference scenario with an additional scenario where VoLTE is the highest priority service and there is no distinction among the remaining data services. One also performs a variation of the number of VoLTE users while fixing the number of data users in the reference scenario to evaluate the degradation of the other services. The influence of the two voice codecs AMR-WB and EVS is also analysed.

An analysis of the reference scenario configuration is carried through rural, suburban and urban environments. The influence of the number of indoor users is also evaluated for three situations: 20, 50 and 80% of the number of users in the reference scenario. The impact of the service parameters on the performance of the system is analysed. Finally, different scenarios in terms of service penetration are also tested, considering two additional scenarios: a Video centric and a VoLTE centric one.

B. VoLTE quality

VoLTE requires a low throughput and as it is considered as the highest priority service for scheduling purposes, it is beforehand expected that capacity for VoLTE users is not an issue under usual conditions. For evaluating it one analyses a scenario with 100% penetration of VoLTE users. Figure 5 shows the observed variation of the MOS. One concludes that for the 10 MHz bandwidth, the MOS is abruptly reduced for a number of users per hour above 10 000 which, according to the simulation results, corresponds to an average number of simultaneous users of 164. For the 20 MHz bandwidth the value of MOS is practically independent on the number of users for the case being studied. For a matter of a better understanding, a variation on the number of users from 10 000 to 13 750 users implies that the average estimated end-to-end delay goes from 108.8 to 164.3 ms which gets significantly closer to the advisable maximum value of 200 ms for voice communications.

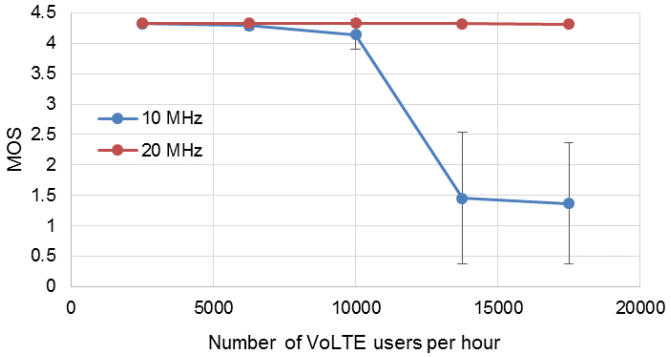


Figure 5. Average user MOS.

C. Number of users

One starts by characterising the overall behaviour of the reference scenario. Figure 6 shows the variation of the total cell throughput as a function of the total number of users. One studies the throughput evolution for both 10 and 20 MHz bandwidths.

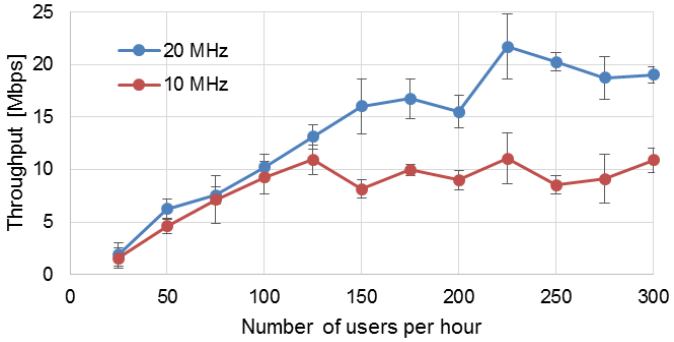
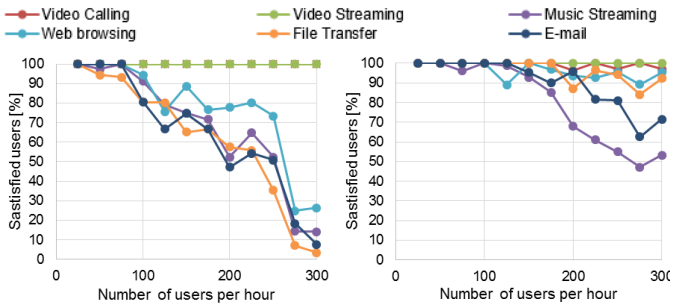


Figure 6. Total cell throughput for different numbers of users.

The results suggest that the reference scenario has a similar behaviour for both bandwidths for a number of users up to 125. At this point, the throughput stabilises for the 10 MHz bandwidth while the 20 MHz only stabilises at approximately 225 users. The stabilisation of the offered throughput for a scenario with a fixed configuration is an expected behaviour as the system's capacity is finite. This implies that as users are arriving at a constant rate at the cell, performing the same set of services according to a fixed service penetration, the maximum throughput the cell can offer is a constant value. It is interesting to notice that this maximum throughput for the 10 and 20 MHz bandwidths is about 10 and 20 Mbps, respectively, corresponding to a spectral efficiency of one. That is not surprising since 80% of users are indoor ones, hence, experiencing low SNR values.

The analysis is further taken to the performance of the various services. Figure 7 shows the variation of the percentage of satisfied users for each service comparing the reference scenario with a scenario where no priorities are defined among data services. As one has concluded that the performance of VoLTE is barely affected for the considered range of the number of users in the cell, one discards its analysis and focuses on its impact on the performance of other services.



(a) With priorities among data services.

(b) Without priorities among data services.

Figure 7. Satisfied users per service for different numbers of users.

The results show that guaranteeing a satisfaction of the throughout requirements for all the services of at least 90% is only possible for an arrival rate of 75 users per hour when priorities are considered. After that, the user satisfaction degrades with the lowest priority services being the first to suffer in terms of throughput. The exception is music streaming where the results show a more emphasised decrease on user satisfaction than in web browsing which has a lower priority. This is basically justified by the fact that the music service has much stricter throughput requirements when compared to web browsing.

For the scenario with no priorities, the results reflect the low levels of packet failure for higher loads. It is possible to ensure a 90% throughput satisfaction in all services up to an arrival rate of 150 users per hour. In this scenario, capacity for higher priority services is essentially taken by lower priority ones. Because of this, services like video calling and video streaming suffer a bigger impact in terms of throughput. However, results show that for the simulated range of users this is not enough to significantly degrade the number of satisfied users for these services. Some impact is verified in the case of video calling but the percentage of satisfied users never drops below 90%. Video streaming is not affected due to the large difference between its average and minimum throughput values.

Another approach to analyse the influence of the number of users is to fix the number of users performing all the services in the reference scenario and to change only the number of VoLTE users to stress the effect of introducing VoLTE. Figure 8 shows the percentage of the total generated traffic during one hour that corresponds to VoLTE traffic. One compares results using both AMR-WB and EVS codecs. As the considered mode for the EVS codec has a source bit rate that is roughly two times the one from the AMR-WB codec, network generated traffic for EVS is also approximately the double when compared to AMR-WB.

The most important aspect to observe is that even for a very high number of users and with a high quality voice codec, the percentage of voice traffic does not exceed 6% of the total network traffic under a regular scenario in terms of data services penetration. Assuming a situation where VoLTE is deployed over an existing LTE network, one does not expect a significant performance impact if the service usage pattern remains similar to the one observed in older voice technologies.

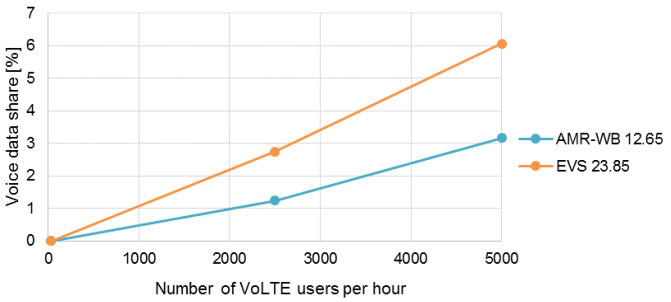
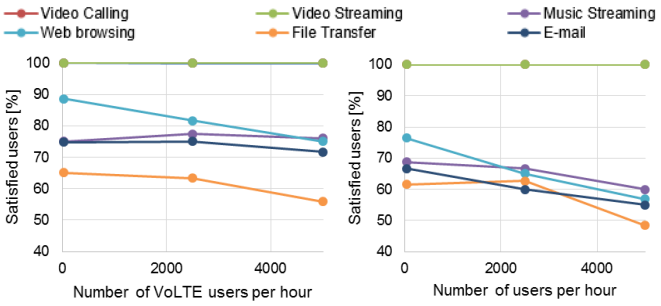


Figure 8. Percentage of voice traffic for different numbers of VoLTE users.

The user throughput satisfaction is analysed in order to assess the real impact that voice traffic has in the services' throughputs. Figure 9 shows the behaviour of the system in terms of satisfied users for different numbers of VoLTE users considering the two codecs under analysis in this study. Video calling appears superimposed with video streaming as both these services are not affected in these conditions. One observes for this metric that significant changes only arise between 2 500 and 5 000 VoLTE users per hour, with the exception of web browsing. The high number of generated packets contributes to a high packet failure ratio and, consequently, to a reduction of the experienced throughput.



(a) AMR-WB codec.

(b) EVS codec.

Figure 9. Satisfied users for different numbers of VoLTE users.

Video calling and video streaming are not affected at all as they are the highest priority services. The percentage of satisfied users decreases in general for all the other services. Comparing the results between the 33 VoLTE users' case and the 5 000 one, web browsing shows the worst impact with a 13% reduction in terms of satisfied users for the AMR-WB code and close to 20% for the EVS codec. This means that the impact on service performance is not very significant even when a high number of VoLTE users is assumed.

Assuming a target of 10% as the maximum reduction on the percentage of satisfied users for all services, one can estimate a maximum number of simultaneous VoLTE users that the system can support. Results have shown that for the AMR-WB codec, the web browsing service suffers the higher degradation. In this case, one verifies that a 10% reduction of the satisfied users occurs for 3 800 VoLTE users per hour. Therefore, according to Figure 4.11, an approximate maximum of 60 simultaneous VoLTE users can be supported. For the EVS codec, web browsing is also the critical service. In this case, the 10% reduction in terms of satisfied users occurs approximately for

2 200 VoLTE users per hour, which corresponds to an approximate maximum of 36 simultaneous VoLTE users.

For a matter of comparison, if a higher margin, e.g., 20% for the reduction on the number of satisfied users was considered, more than 82 simultaneous VoLTE users could be supported for both codecs. It is important to mention that these values correspond to a scenario with approximately 12 users performing data services simultaneously and should not be interpreted as maximum theoretical values but rather a typical urban scenario.

D. Type of environment

The scenarios under analysis are based on a single cell operating on rural, suburban and urban single-cell environments whose SNR is characterised by a Log-Normal distribution with a standard deviation of 6.0 dB and mean values of 13, 16 and 18 dB, respectively [17]. Besides the type of environment, each user in a given scenario can be characterised as indoor or outdoor. As the default values considered for the SNR are for outdoor environments, users which are indoor are assigned an additional attenuation with a value between 12 and 20 dB.

It is observed that the difference between urban and suburban environments in terms of total throughput offered by the cell has only a decrease close to 11%, between urban and rural one observes a difference of approximately 6 Mbps which corresponds to a 38% decrease. To get further detail into the performance of each of the services, Figure 10 shows the behaviour of the system in terms of queuing delay for the three types of environment.

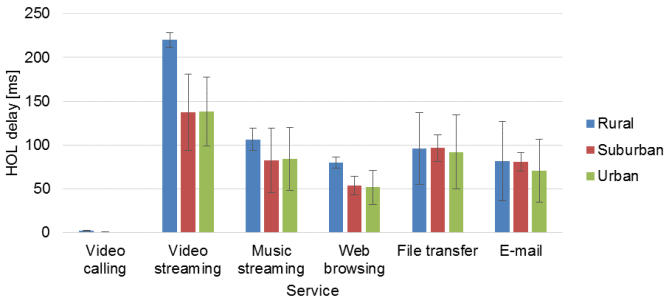


Figure 10. Average HOL delay per service for the different environments.

Even with a significant decrease of the achievable throughputs, non-GBR services, namely web browsing, file transfer and e-mail, are not deeply affected in terms of delay and all of them have HOL delays below 100 ms for the three environments. For GBR services, namely video calling, video streaming and music streaming, results do not show any significant distinctions between suburban and urban environments. Differences are more noticeable between the video and music streaming services on the rural case. When one compares the rural environment with the suburban and urban ones, an increase of about 80 ms on the average HOL delay of video streaming is verified. In terms of user satisfaction, no relevant differences exist between urban and suburban environments unless for music streaming where the reduction of satisfied users reaches about 13%. For rural environments, the level of satisfaction is significantly lower in all cases, with the exception of the video calling and video streaming services.

Similarly, to the type of environment, one analyses the results for different percentages of indoor users. Figure 11 shows the behaviour of the system in terms of the satisfied users for three different values of the percentage of indoor users.

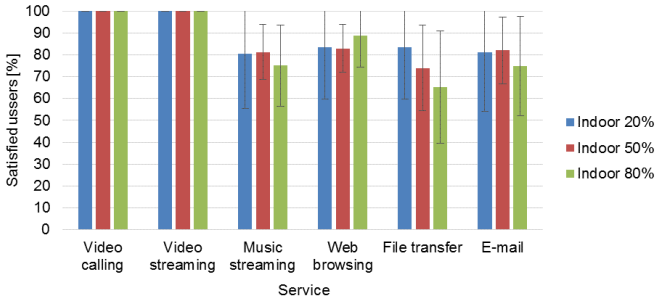


Figure 11. Satisfied users for different percentages of indoor users.

As one can observe, the number of indoor users does not significantly impact the services' performance. This improvement is noticeable when one compares the 80% and 50% scenarios but becomes irrelevant between the 50% and 20% ones.

E. Bandwidth

By reducing bandwidth from 20 to 10 MHz the instantaneous available capacity directly reduces to a half. In terms of queuing delay for the two bandwidths considered in this study, results show that for most services, only slight increases on the HOL delay are verified. The exception in this case is video streaming which has an increase of 94 ms. Even though the service has a high priority, the amount of traffic that it generates in the network due to its high throughput makes it vulnerable to delay issues. However, as the lower priority services do not have significant degradation in terms of HOL delay, this suggests a significant increase on the number of failed packets to make room for the incoming video streaming traffic. Results for packet failure are coherent with the analysis made for the HOL delay values. Lower priority services, specially web browsing and file transfer, have an increase of 33% and 31% in packet failure, respectively, which happens because these are the most demanding non-GBR services, requiring higher amounts of data. Video streaming has an increase of 45% in packet failure due to the high average values of HOL delay which raises the number of packets close to the delay budget specified for streaming packets.

Figure 12 shows the behaviour of the system in terms of user satisfaction. While the 10 MHz bandwidth still does not impact the higher priority services, namely video calling and video streaming, results show that the remaining services suffer severe reductions in terms of throughput satisfaction. In the worst case, file transfer shows a reduction of around 40% in the percentage of satisfied users due to the high packet failure. Music streaming and web browsing are also similarly affected. E-mail is not very affected, even though it is the service with the lowest priority, because it does not require as much data as the other mentioned services.

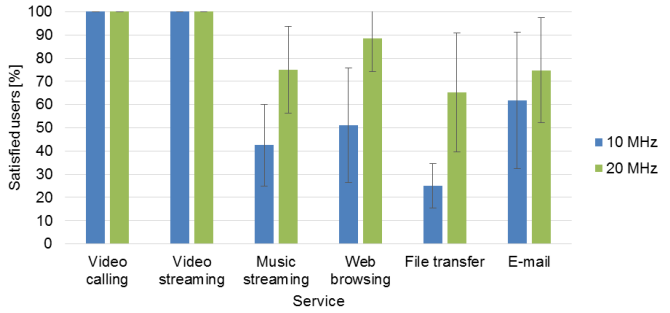


Figure 12. Satisfied users for different bandwidths.

F. Service parameters

This section describes the analysis of the results obtained for scenarios where the service parameters are changed. Figure 13 shows the behaviour of the system in terms of packet failure.

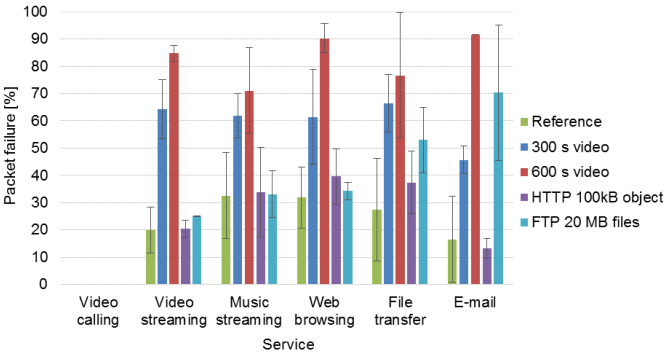


Figure 13. Packet failure per service for scenarios with different service parameters.

It is observed that, increasing the average video duration causes the system to be severely congested and packet failure increases to unacceptable values for video streaming and all services with a lower priority. For the scenario with bigger HTTP objects, results stay similar to the reference scenario. For the scenario with bigger FTP files, results are also similar to the reference with the exception of file transfer and e-mail services. Increasing the average file size for file transfer causes packet failure to increase 26% for file transfer and most noticeably 54% for e-mail.

All the services with a lower priority than video streaming suffer severe degradation in the scenarios with 300 and 600 s video durations as all of them get throughput satisfaction levels below 30%. This is consistent with the results of high HOL delay and packet failure. For the scenario of varying the average size of the web browsing objects it becomes clear that the number of satisfied users follows the results of the reference scenario. Web browsing does not generate great amounts of traffic due to low object sizes and large reading times. Following the increase in HOL delay and packet failure, the percentage of satisfied e-mail users dropped 53% as a consequence of changing the file size in file transfer.

G. Service penetration

This section describes the results obtained when scenarios with different service penetrations are considered. For that purpose, one defined two additional scenarios. The first is Video centric as one assumes that 50% of the users perform video streaming and the other scenario is VoLTE centric with a 50% penetration. From the obtained results one concludes that, in a VoLTE centric scenario the total cell throughput decreases approximately 46%. On the other hand, in the Video centric scenario, the total throughput increases about 14%. Figure 14 shows how this behaviour translates in terms of user satisfaction for the three defined scenarios.

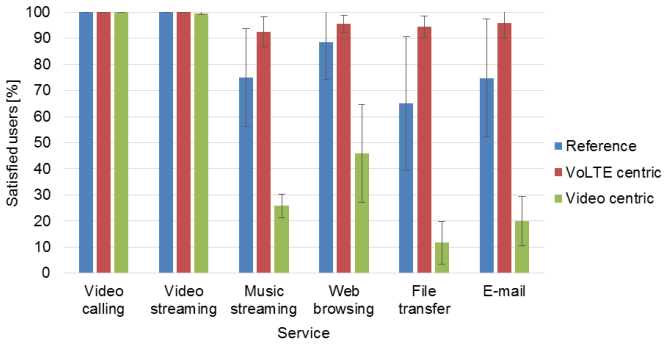


Figure 14. Satisfied users for different service penetrations.

The results for the percentage of satisfied users highlight the impact of increasing the amount of video traffic in the system. For the Video centric scenario, most of the services with a lower priority than video streaming and video calling have unacceptable percentages of satisfied users, all of them below 50%. For the VoLTE centric scenario, results show the improvements due to significantly reducing the cell load. Satisfaction levels are all above 90%, representing the conditions to which the cell is expected to work to provide good levels of QoS to all of its users.

V. CONCLUSIONS

The main goal of this thesis was to evaluate the impact of the implementation of VoLTE in the provided QoS of other services in the context of an already deployed LTE network, while monitoring the performance of the VoLTE service. In order to accomplish this goal, a single-cell model for the DL resource allocation at the radio interface in LTE systems was proposed and implemented into a simulator from which several simulations were performed. For the analysis of VoLTE call quality, a scenario with a full 100% penetration of VoLTE users was tested. As the developed model assumes that VoLTE has the highest priority, call quality is barely affected for realistic numbers of users.

Concerning the variation of the total number of users, both 10 and 20 MHz bandwidths allow a similar cell throughput until a rate of 125 users per hour. After that, throughput stabilises for the 10 MHz bandwidth while the 20 MHz stabilises at approximately 225 users per hour. In what refers to the services' performance, one compares the QoS metrics for two scenarios with and without scheduling priorities among data services. Results show that for a lower number of users per hour, considering no priorities allows a higher number of satisfied users. To guarantee that at least 90% of the users is satisfied for each service, up to a rate of 150 users per hour is supported for the scenario with no priorities while only 75 users per hour are

supported when priorities are considered. For higher cell loads, low priority services like web browsing, file transfer or e-mail, benefit in terms of user satisfaction for the scenario with no priorities among the services.

To analyse the impact of the number of VoLTE users over the performance of the other services, the number of data users in the reference scenario was fixed and the rate of VoLTE user arrivals was varied between 33, corresponding to the reference scenario, and 5000 users per hour. Results were compared for both AMR-WB and EVS codecs. It is observed that VoLTE generated traffic does not exceed 6% of the total cell traffic in the worst case with the EVS codec. To guarantee that the reduction on the number of satisfied users for each service stays below 10%, up to 60 and 36 simultaneous VoLTE users are supported for the AMR-WB and EVS codecs, respectively.

Rural, suburban and urban environments were analysed to assess their influence on the system's performance. The cell throughput for the urban scenario is 38% above the rural one. No significant degradation is caused when comparing the urban with the suburban environment. Packet failure for lower priority services increases up to 40%, in the web browsing case, from the suburban to the rural environment. The percentage of indoor users was also changed to assess how indoor attenuation affects the achievable user throughputs due to a reduction of the average SNR. The total cell throughput in the reference scenario decreases about 50% when the percentage of indoor users increases from 20% to 80%, being more noticeable between the 80% and 50% scenarios. Results show that there is no significant degradation in terms of the number of satisfied users for the load conditions of the reference scenario.

Regarding the radio channel bandwidths, a comparison between the 10 and 20 MHz bandwidths were made. For the defined reference scenario, reducing the available bandwidth results in a scenario where the total cell throughput is reduced from 16 to 8 Mbps. As a consequence of decreasing the available capacity for a half, services with a lower priority than video streaming have a reduction that can go up to 40% in the percentage of satisfied users. Higher priority services, namely video calling and video streaming, do not show any degradation in terms of satisfied users.

Several service parameters were changed in order to assess their influence in the system performance. Increasing the average video duration to 300 and 600 s has a serious impact in all the services with a lower priority, as the percentage of satisfied users for these services drops below 30%. Increasing the web browsing object sizes by a factor of 10 barely had any influence in the observed QoS. Increasing the file sizes in file transfer shows that the biggest impact occurs for e-mail which is the only service with lower priority.

In terms of service penetration, for the VoLTE centric scenario with 50% of VoLTE users, a decrease on the total cell throughput of 46% is verified. For the Video centric scenario, the 120 ms increase in queuing delay for video streaming caused an increase in packet failure for this service and for those with a lower priority. For these services, less than 50% of the users get their throughput satisfied while for the VoLTE centric scenario, user satisfaction levels are above 90% for all services.

Future work includes the study of a real multi-cellular network that would allow to study aspects related to mobility like, for example, the impact of handover on the performance of all services including VoLTE. Other aspects related to the modelling of resource management in LTE could be further studied. It would also be interesting a comparison between several scheduling algorithms proposed in the literature to improve VoLTE capacity. The obtained results for a LTE system could also be compared with the voice capacity of currently deployed GSM and UMTS systems.

REFERENCES

[1] S. Yi, S. Chun, Y. Lee, S. Park and S. Jung, *Radio Protocols for LTE and LTE-Advanced (First Edition)*, John Wiley & Sons, Chichester, UK, Sep. 2012.

[2] GSMA, *GSMA IR.92: IMS Profile for Voice and SMS*, Ver. 10.0, May 2016 (http://www.gsma.com/newsroom/wp-content/uploads/IR.92-v10.0.pdf).

[3] Ericsson, *Ericsson Mobility Report*, Public Consultation, Stockholm, Sweden, June 2017.

[4] I. Siomina and S. Wanstedt, "The impact of QoS support on the end user satisfaction in LTE networks with mixed traffic", in *2008 IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications*, Cannes, France, Sep. 2008.

[5] B. Bojovic and N. Baldo, "A new channel and QoS aware scheduler to enhance the capacity of voice over LTE systems", in *2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14)*, Barcelona, Spain, Feb. 2014.

[6] A. Vizzarri, "Analysis of VoIP Over LTE End-To-End Performances in Congested Scenarios", in *2014 Second International Conference on Artificial Intelligence, Modelling and Simulation*, Madrid, Spain, Nov. 2014.

[7] O. Kadatskaya and S. Saburova, "Research of Requirements to QoS for Voice over LTE", in *2014 First International Scientific-Practical Conference: Problems of Infocommunications, Science and Technology*, Kharkiv, Ukraine, Oct. 2014.

[8] F. J. Rivas, A. Díaz, and P. Merino, "Obtaining More Realistic Cross-Layer QoS Measurements: A VoIP over LTE Use Case", *Journal of Computer Networks and Communications*, Vol. 2013, Article ID 405858, Aug. 2013.

[9] D. Almeida, *Inter-Cell Interference Impact on LTE Performance in Urban Scenarios*, M.Sc. Thesis, Instituto Superior Técnico, Lisbon, Portugal, 2013.

[10] F. Khan, *LTE for 4G Mobile Broadband: Air Interface Technologies and Performance*, Cambridge University Press, Cambridge, UK, 2009.

[11] D. Heyman, "The GBAR Source Model for VBR Video conferences", *IEEE/ACM Transactions on Networking*, Vol. 5, No. 4, Aug. 1997, pp. 554-560.

[12] H. Holma and A. Toskala, *LTE for UMTS: Evolution to LTE Advanced (2nd Edition)*, John Wiley & Sons, Chichester, UK, Mar. 2011.

[13] ITU-T, *The E-model: a computational model for use in transmission planning*, Recommendation G.107, Feb. 2014.

[14] D. Nguyen and H. Nguyen, "A dynamic rate adaptation algorithm using WB E-model for voice traffic over LTE network", in *2016 Wireless Days (WD)*, Toulouse, France, May 2016.

[15] S. Moller, A. Raake, N. Kitawaki, A. Takahashi and M. Waltermann, "Impairment factor framework for wideband speech codecs", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No.6, Nov. 2006, pp. 1969–1976.

[16] Cisco, *Encrypted Traffic Analytics*, White Paper, June 2017 (https://www.cisco.com/c/dam/en/us/solutions/collateral/enterprise-networks/enterprise-network-security/nb-09-encrytd-traf-anlytcs-wp-cte-en.pdf).

[17] P. Carreira, *Data Rate Performance Gains in UMTS Evolution to LTE at the Cellular Level*, M.Sc. Thesis, Instituto Superior Técnico, Lisbon, Portugal, 2011.