Lipovac *et al. J Wireless Com Network* (2021) 2021:83
https://doi.org/10.1186/s13638-021-01970-y

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Practical cross-layer testing of HARQ-induced delay variation on IP/RTP QoS and VoLTE QoE

Adriana Lipovac[*], Vlatko Lipovac, Ivan Grbavac and Ines Obradović

*Correspondence:
adriana.lipovac@unidu.hr
University of Dubrovnik,
Dubrovnik, Croatia

Abstract

As the PHY/MAC-layer IR-HARQ and RLC-layer ARQ error recovery procedures, adopted in LTE, may impose additional delay when their code-block retransmissions occur, the arising question is whether these significantly contribute to IP and consequently RTP packet delays, and finally degrade the overall application-layer end-to-end QoE, especially when voice is transmitted over LTE? With this regard, we propose and demonstrate a VoLTE QoS and QoE test procedure based on PHY/MAC/RLC/IP/TCP-UDP/RTP cross-layer protocol analysis and perceptual speech quality QoE measurements. We identified monotonic relationship between the paired observations: QoE and HARQ RTT, i.e. between the PESQ voice quality rating and the IP/RTP packet latency, for given BLER of the received MAC/RLC code-blocks. Specifically, we found out that, for the HARQ RTT value of about 8 ms, only up to 2 HARQ retransmissions (and consequently no RLC-ARQ one) is appropriate during any voice packet, otherwise delay accumulation might not be accordingly "smoothed out" by jitter/playback buffers along the propagation path.

Keywords: VoLTE, Voice QoE testing, Protocol analysis

1 Introduction

Voice-over-IP (VoIP) has been deployed for quite a long time in wireline networks paving the way to IP telephony [1]. In mobile and wireless environment, the fourth generation (4G) networks—namely the Long Term Evolution (LTE) in particular, has provided generally very good Quality-of-Service (QoS), which includes Voice-over-LTE (VoLTE) as well, and enables its widespread use. Still, a need can arise to practically explore the impact of inherent drawback of LTE for voice transport—additional delay due to retransmissions at lower layers, which, under some circumstances, could degrade the end-to-end VoLTE Quality-of-Experience (OoE).

So, targeting either reactive troubleshooting in such situation, or proactive actions during network installation and commissioning, we propose testing through the layers ("bottom-up", "top-down", or "middle split") by applying the test procedures proposed here.

With this regard, let us recall that, in addition to the Transmission Control Protocol (TCP) retransmissions, in LTE protocol stack, there are two more retransmitting layers: the Physical/Medium Access Control (PHY/MAC) and the Radio Link Control (RLC), which may impose additional delay onto their data frames, and thus onto Internet Protocol (IP) and, finally, Real-Time transport Protocol (RTP) packets, which, further on, may degrade the end-to-end Quality-of-Experience (QoE), especially when voice is transmitted over LTE [2].

Accordingly, as the LTE protocol stack PHY and MAC layers are both with a number of variable parameters and fairly complex, with deployment of VoLTE in particular, providing their services to the network and transport layers through cross-layer design and management, has got new challenges in the everlasting goal to achieve high service throughput and low delay [3].

More specifically, as the LTE PHY/MAC layer Incremental Redundancy (IR) Hybrid Automatic Repeat-reQuest (HARQ) protocol and the RLC layer Automatic Repeat-reQuest (ARQ) protocol, adopted in LTE, may create sudden delay ramping when retransmissions occur, the question arising here is whether this significantly and dominantly contributes to the overall RTP packet delay and so degrades the end-to-end voice perceptual QoE?

With this regard, there have been numerous mathematical models and simulation studies [3–5] addressing the VoLTE QoS and perceptual QoE, but in the following, we propose a practical means for testing real-life VoLTE traffic, which we set up in our laboratory, and present some preliminary results that we obtained this way.

In Section II, the packet-level QoS and the end-to-end perceptual QoE of the legacy VoIP is considered, whereas in Section III, the VoLTE protocol stack is reviewed with accent on added delay by PHY/MAC IR-HARQ and RLC ARQ procedures. The test tools and preliminary results are presented in Section IV, while conclusions are drawn in Section V.

2 Methods

VoIP QoS enhancements have led to IP telephony, whose QoS has approached the one of legacy circuit-switched networks. This has made VoIP integrated in all-IP packet-switched state-of-the-art networks, according to the protocol stack presented in Fig. 1 [6].

As it can be seen, the utmost sensitive signaling information is transported via the connection-oriented TCP, so any corrupted TCP segment must be corrected by the TCP retransmission mechanism. However, the voice service itself that is less sensitive to transmission errors, but (as real-time) very sensitive to delay, is handled by the connectionless User Datagram Protocol (UDP), which introduces far lesser processing time at the transmission layer than TCP does.

2.1 IP/RTP packet-level QoS

VoIP transmission adds substantial delay to a signal that traverses the network. Specifically, the end-to-end one-way delay below 100 ms can be perceptible, whereas one-way values already above 150 ms are annoying. The problem is, however, that due to various influencing factors such as terminal/phone voice signal processing in codecs, queuing,

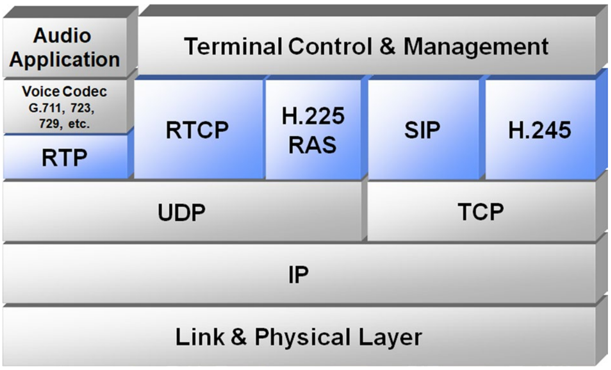


Fig. 1 VoIP protocol stack [6]

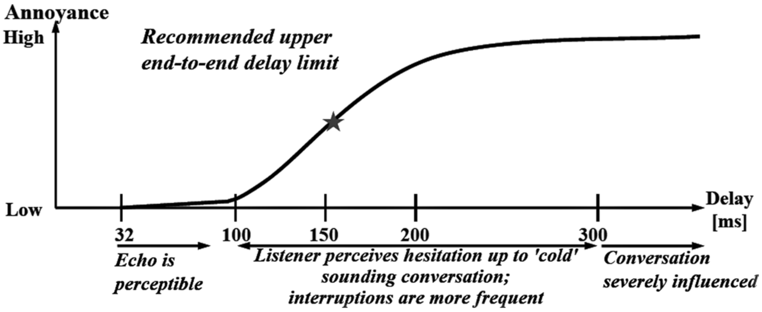


Fig.2 VoIP personal annoyance by end-to-end ("mouth-to-ear") signal delay

switching/routing, receive jitter buffering, transmit packetization, number of hops, etc. [7], the cumulative VoIP delay can easily exceed 200 ms [8], Fig. 2, so it is difficult to achieve the preferred mouth-to-ear delay maximal value of 150 ms [9].

Moreover, not only is the IP packet delay important, but also the delay variation, which is commonly referred to as jitter that significantly deteriorates the QoS and the QoE, as it propagates upwards the protocol stack and finally ends up with abrupt RTP PDU jitter increase, Fig. 3, which can turn itself into major degradation of perceptual speech quality observed as the end-to-end QoE [1].

On the other hand, the IP packet loss, too (and the consequent RTP packet loss) can cause much harm to voice quality, Fig. 4.

Moreover, in case of IP packets carrying signaling information and thus being transported via TCP, the packet loss will cause TCP segments' retransmissions, which will delay the connection establishment and postpone the voice transport, i.e. incur additional delay.

Therefore, although signaling delay is not directly a part of the voice delay budget (that is of primary concern here), without reliable transport of signaling information, even instant transmission of voice service is of no value. That is why, let us consider some common TCP/IP troubleshooting issues as it follows:

The retransmissions can be qualified as excessive with respect to certain long-term-observations based thresholds, such as e.g. 3 TCP retransmissions out of a total of 5,

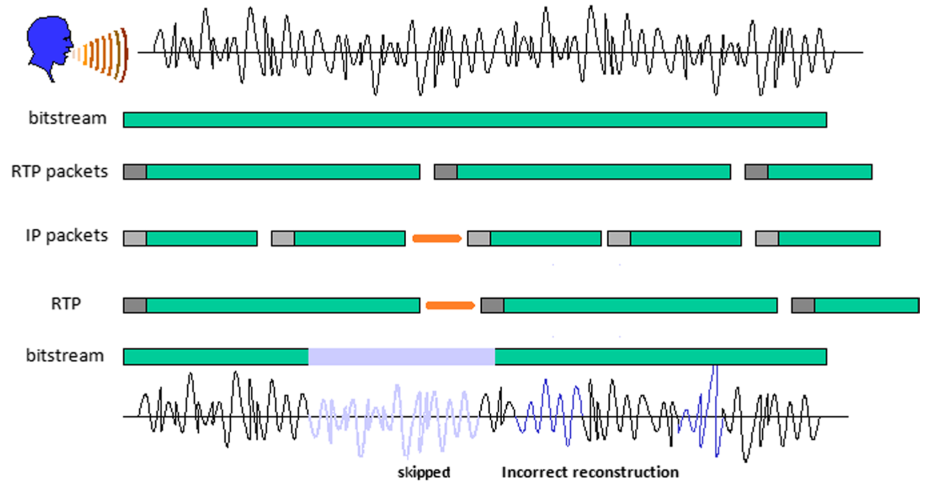


Fig. 3 IP/RTP packet delay variation [1]

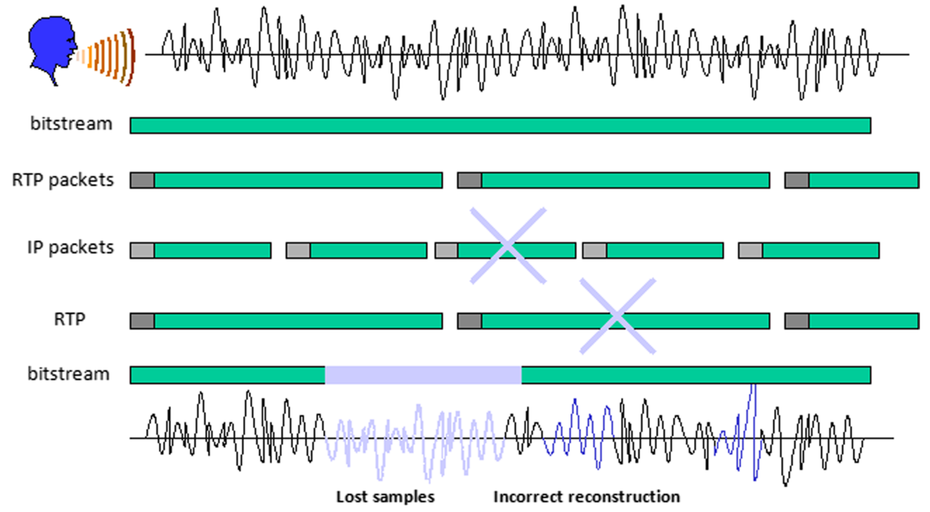


Fig. 4 IP/RTP packet loss [1]

6, or 7 transmitted segments. Accordingly, the TCP excessive retransmissions percentage can be calculated and reported by a protocol analyzer.

With this regard, troubleshooting requires several things to be checked, such as:

whether the network has a drop in performance that prevents positive acknowledgements (ACK) from responding in time, so that TCP timeouts occur, or

whether it is always the client sending the retransmissions because of overdue ACKs, pinpointing to the server being overloaded by supporting too many sessions, or

whether it is always the server sending the retransmissions, so it might be required to check if the client's cache memory is set to the optimal size, or

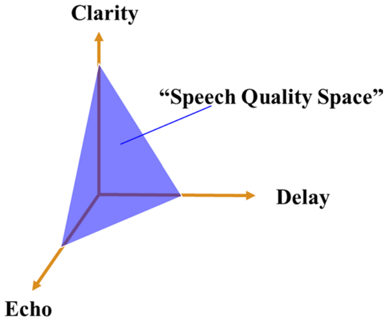


Fig. 5 Speech quality factors

whether the network has devices with slow performance or the devices are unable to gain access to the network, e.g. Ethernet, because of high network utilization, so it might be useful to check if a bridge/switch, or a router is discarding frames, or

whether the standard IP reassembly timeout (in case of fragmentation at the IP layer) is set too low, implying too slow network response, or

whether noise in the physical media is excessive to cause TCP retransmissions, or

whether the IP packets with low time-to-live (TTL) appear, indicating that the destination address is not receiving the segments before they time out and, because of this, is requesting retransmissions, so pointing to the need to check the sending station's network parameters configuration. (The TCP connection setup packets may not be reaching the server because of low TTL, or a connection can be slowed due to dropped packets caused by low TTL which consequently cause retransmissions). TTL in IP can also incorporate actual latency of a router, but this is routing-protocol-dependent (i.e. RIP, OSPF, etc.), and complies to the routing protocol definition of details on its use of TTL.

So, having summarized some TCP/IP concerns affecting the call setup time, we proceed in the media plane.

Accordingly, the RTP-coded voice is handled by the connectionless UDP protocol that makes no error control at the transport layer (and so introduces less latency), as speech intelligibility is not so vulnerable with regard to sporadic corrupted signal sections, whereas the delay remains dominant impairment.

Based on statistical properties of speech signal—namely, its quasi-stationarity within about 20 ms on average, equal RTP voice packet duration is adopted [6].

2.2 End-to-end VoIP perceptual quality

The end-to-end speech quality is not simply related to the IP/RTP QoS. In fact, the speech quality constituents are, namely, clarity (correlation of the information that can be extracted out of a conversation versus the information sent), echo (reflection of the transmitted signal from the far end with enough strength to be perceptible to a human), and delay, Fig. 5.

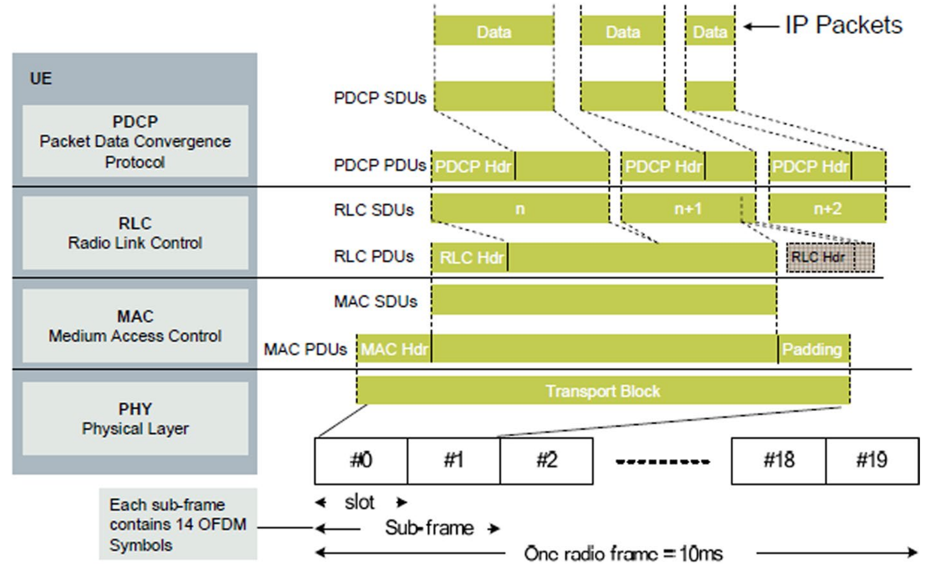


Fig. 6 LTE downlink protocol stack [2]

2.3 (Vo)LTE QoS

Let us consider the VoLTE transmission with the LTE downlink protocol stack in time domain, Fig. 6.

The frame lasts 10 ms, consisting of 1-ms-long subframes, each containing a transport block (TB) encapsulating the MAC PDU (and additional padding bits), which carries the RLC PDU containing the Packet Data Convergence Protocol (PDCP) PDU that finally encapsulates the IP packet—all the way up at the network layer.

The IR-HARQ protocol is implemented at both MAC and PHY layers, where the former performs error recovery management and signalling, while the latter executes it. (For this purpose, each TB is complemented by its Cyclic Redundancy Check (CRC) addendum and thus becomes a code-block.) Consequently, the IR-HARQ procedure sends negative acknowledgement (NACK) for each code-block with failed CRC, thus initiating the PHY layer to retransmit the code-block with more redundancy, while also preserving the previous (failed) code-block(s) to be combined with the current one, and then subject to CRC algorithm again. The process goes on until an errorless code-block is identified by CRC, initiating MAC layer to send ACK.

Specifically, with the IR-HARQ algorithm, the first transmission comes with high code rate, while the following ones contain gradually increasing redundancy (parity) bits at the expense of systematic ones, so that, in the LTE downlink, up to 4 codeword redundancy versions (rv0 to rv3) with increasing coding gains at the receiver can be sent, enabling throughput adaptation to channel conditions, Fig. 7 [9].

Moreover, for each TB, there is another retransmission mechanism—ARQ at the RLC layer, taking over the error recovery process if the complete HARQ procedure fails (i.e. if even after rv3 the transmission remains erroneous).

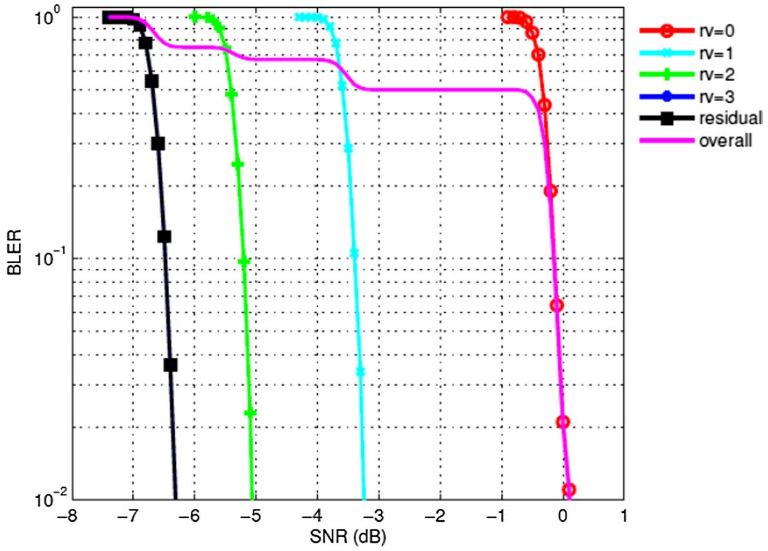


Fig. 7 BLER versus SNR for MCS index 6, AWGN channel, 25 RBs HARQ rv = 0, 1, 2, 3 [7]

So, back to Fig. 6, we summarize that, in addition to TCP retransmissions, with the LTE protocol stack, there are two more (possibly) retransmitting layers—PHY/MAC and RLC, whose respective IR HARQ and ARQ processes may create sudden ramping of the IP/RTP packet delay when retransmissions occur, with the final outcome of degrading perceptual speech quality.

Now recall that out of the overall end-to-end delay budget for high-quality voice transfer presumed in LTE [2], reduced by the core network allowance, up to 50 ms is left to tolerable air interface delay comprising MAC buffering/scheduling and detection. So, e.g. in downlink, the eNodeB transmits data whereas the UE responds (with ACK or NACK) after 4 ms. Then it takes another 4 ms for the eNodeB (having received the ACK or NACK), to send a new transmission or re-transmission. Thereby, with LTE Frequency Division Duplex (FDD), this allows the HARQ round trip time (RTT) of 8 ms for a single code-block transmission, and theoretically up to 6 HARQ transmissions for a VoIP packet [2].

On the other hand, the main VoIP quality criterion for LTE [2] is that an outage is identified and counted if more than 2% of the packets are not received within the delay budget, when monitored over the whole call.

With this regard, it is the task in practice to find out whether and to what extent the lower-layer impairments contribute to the overall packet delay and end-to-end voice QoE.

In the following, we propose a practical means for testing real-life VoLTE traffic, and present some preliminary results that we have achieved so far in this way.

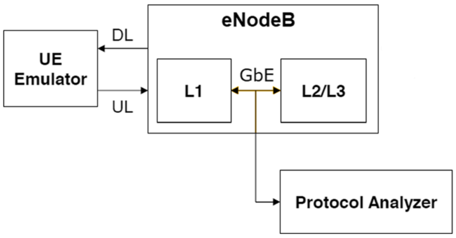
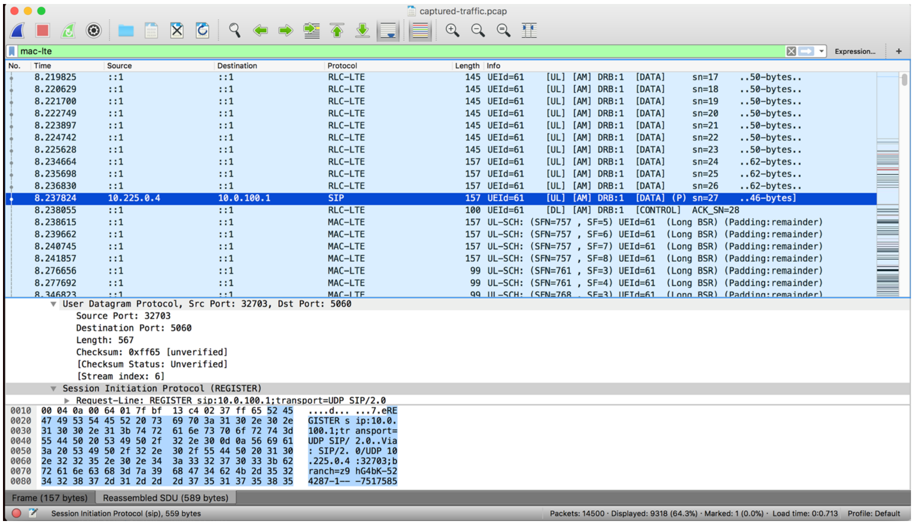


Fig. 8 Test system



```
No.  Time      Source      Destination  Protocol  Length Info
     8.219825  ::1         ::1          RLC-LTE    145 UEId=61   [UL] [AM] DRB:1  [DATA]     sn=17   ..50-bytes..
     8.220629  ::1         ::1          RLC-LTE    145 UEId=61   [UL] [AM] DRB:1  [DATA]     sn=18   ..50-bytes..
     8.221700  ::1         ::1          RLC-LTE    145 UEId=61   [UL] [AM] DRB:1  [DATA]     sn=19   ..50-bytes..
     8.222749  ::1         ::1          RLC-LTE    145 UEId=61   [UL] [AM] DRB:1  [DATA]     sn=20   ..50-bytes..
     8.223897  ::1         ::1          RLC-LTE    145 UEId=61   [UL] [AM] DRB:1  [DATA]     sn=21   ..50-bytes..
     8.224742  ::1         ::1          RLC-LTE    145 UEId=61   [UL] [AM] DRB:1  [DATA]     sn=22   ..50-bytes..
     8.225628  ::1         ::1          RLC-LTE    145 UEId=61   [UL] [AM] DRB:1  [DATA]     sn=23   ..50-bytes..
     8.234664  ::1         ::1          RLC-LTE    157 UEId=61   [UL] [AM] DRB:1  [DATA]     sn=24   ..62-bytes..
     8.235698  ::1         ::1          RLC-LTE    157 UEId=61   [UL] [AM] DRB:1  [DATA]     sn=25   ..62-bytes..
     8.236830  ::1         ::1          RLC-LTE    157 UEId=61   [UL] [AM] DRB:1  [DATA]     sn=26   ..62-bytes..
     8.237824  10.225.0.4  10.0.100.1   SIP        157 UEId=61   [UL] [AM] DRB:1  [DATA] (P) sn=27   ..46-bytes]
     8.238055  ::1         ::1          RLC-LTE    100 UEId=61   [DL] [AM] DRB:1  [CONTROL]  ACK_SN=28
     8.238615  ::1         ::1          MAC-LTE    157 UL-SCH: (SFN=757 , SF=5) UEId=61  (Long BSR) (Padding:remainder)
     8.239662  ::1         ::1          MAC-LTE    157 UL-SCH: (SFN=757 , SF=6) UEId=61  (Long BSR) (Padding:remainder)
     8.240745  ::1         ::1          MAC-LTE    157 UL-SCH: (SFN=757 , SF=7) UEId=61  (Long BSR) (Padding:remainder)
     8.241857  ::1         ::1          MAC-LTE    157 UL-SCH: (SFN=757 , SF=8) UEId=61  (Long BSR) (Padding:remainder)
     8.276656  ::1         ::1          MAC-LTE     99 UL-SCH: (SFN=761 , SF=3) UEId=61  (Long BSR) (Padding:remainder)
     8.277692  ::1         ::1          MAC-LTE     99 UL-SCH: (SFN=761 , SF=4) UEId=61  (Long BSR) (Padding:remainder)
     8.346823  ::1         ::1          MAC-LTE     99 UL-SCH: (SFN=768 , SF=3) UEId=61  (Long BSR) (Padding:remainder)

  ▼ User Datagram Protocol, Src Port: 32703, Dst Port: 5060
      Source Port: 32703
      Destination Port: 5060
      Length: 567
      Checksum: 0xff65 [unverified]
      [Checksum Status: Unverified]
      [Stream index: 6]
  ▼ Session Initiation Protocol (REGISTER)
    ▶ Request-Line: REGISTER sip:10.0.100.1;transport=UDP SIP/2.0

0010  00 04 0a 00 64 01 7f bf  13 c4 02 37 ff 65 52 45   ....d... ...7.eRE
0020  47 49 53 54 45 52 20 73  69 70 3a 31 30 2e 30 2e   GISTER s ip:10.0.
0030  31 30 30 2e 31 3b 74 72  61 6e 73 70 6f 72 74 3d   100.1;tr ansport=
0040  55 44 50 20 53 49 50 2f  32 2e 30 0d 0a 56 69 61   UDP SIP/ 2.0..Via
0050  3a 20 53 49 50 2f 32 2e  30 2f 55 44 50 20 31 30   : SIP/2. 0/UDP 10
0060  2e 32 32 35 2e 30 2e 34  3a 33 32 37 30 33 3b 62   .225.0.4 :32703;b
0070  72 61 6e 63 68 3d 7a 39  68 47 34 62 4b 2d 35 32   ranch=z9 hG4bK-52
0080  34 32 38 37 2d 31 2d 2d  2d 37 35 31 37 35 38 35   4287-1-- -7517585

Session Initiation Protocol (sip), 559 bytes     Packets: 14500 · Displayed: 9318 (64.3%) · Marked: 1 (0.0%) ·  Load time: 0:0.713   Profile: Default
```

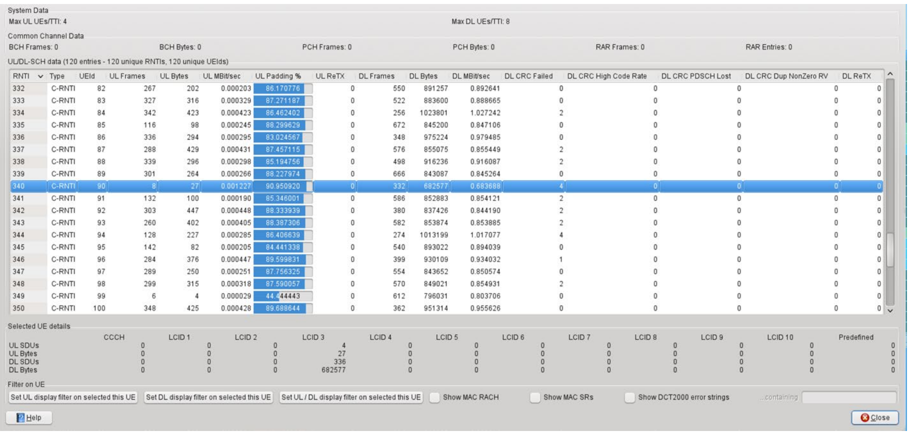
Fig. 9 Protocol analyzer decodes screenshot

3 Discussion and results

We intended to run our tests under live VoLTE network conditions, collecting and analyzing the real-life data. However, it is not always possible to get enough characteristic traffic for testing the target edge performance conditions. So, our intention is to “enhance” the real-time traffic with this regard by operating the monitoring test equipment in “data through” connection mode, which can enable adding various impairments such as delay, jitter and loss (or missinsertion) to the packets outgoing from the selected network interface, and thus enable stress-testing of the network or its particular components etc. [2, 7, 10].

Moreover, even with such modifications, it is not possible to accomplish some tests with statistically significant results as coming out of enough samples, so under such circumstances, software emulation tools are used as well, such as the one emulating the UE, presented in Fig. 8.

Legacy protocol analyzers decode PDUs and provide statistical analysis for various wireline and wireless protocols. So, for example, the screen shot in Fig. 9 presents summary decodes for the observed PDU series, with a selected detailed decode (of SIP PDU in this example) also displayed in hexadecimal and ASCII forms.



System Data

Max UL UEs/TTI: 4 Max DL UEs/TTI: 8

Common Channel Data

BCH Frames: 0 BCH Bytes: 0 PCH Frames: 0 PCH Bytes: 0 RAR Frames: 0 RAR Entries: 0

UL/DL-SCH data (120 entries - 120 unique RNTIs, 120 unique UEIds)

RNTI	Type	UEId	UL Frames	UL Bytes	UL MBit/sec	UL Padding %	UL ReTX	DL Frames	DL Bytes	DL MBit/sec	DL CRC Failed	DL CRC High Code Rate	DL CRC PDSCH Lost	DL CRC Dup NonZero RV	DL ReTX
332	C-RNTI	82	267	202	0.000203	86.170776	0	550	891257	0.892641	0	0	0	0	0
333	C-RNTI	83	327	316	0.000329	87.271187	0	522	883600	0.888665	0	0	0	0	0
334	C-RNTI	84	342	423	0.000423	86.462402	0	256	1023801	1.027242	2	0	0	0	0
335	C-RNTI	85	116	98	0.000245	88.299629	0	672	845200	0.847106	0	0	0	0	0
336	C-RNTI	86	336	294	0.000295	83.024567	0	348	975224	0.979485	0	0	0	0	0
337	C-RNTI	87	288	429	0.000431	87.457115	0	576	855075	0.855449	2	0	0	0	0
338	C-RNTI	88	339	296	0.000298	85.194756	0	498	916236	0.916087	2	0	0	0	0
339	C-RNTI	89	301	264	0.000266	88.227974	0	666	843087	0.845264	0	0	0	0	0
340	C-RNTI	90	8	27	0.001227	90.950920	0	332	682577	0.683688	4	0	0	0	0
341	C-RNTI	91	132	100	0.000190	85.346001	0	586	852883	0.854121	2	0	0	0	0
342	C-RNTI	92	303	447	0.000448	88.333939	0	380	837426	0.844190	2	0	0	0	0
343	C-RNTI	93	260	402	0.000405	88.387306	0	582	853874	0.853885	2	0	0	0	0
344	C-RNTI	94	128	227	0.000285	86.406639	0	274	1013199	1.017077	4	0	0	0	0
345	C-RNTI	95	142	82	0.000205	84.441338	0	540	893022	0.894039	0	0	0	0	0
346	C-RNTI	96	284	376	0.000447	89.599831	0	399	930109	0.934032	1	0	0	0	0
347	C-RNTI	97	289	250	0.000251	87.756325	0	554	843652	0.850574	0	0	0	0	0
348	C-RNTI	98	299	315	0.000318	87.590057	0	570	849021	0.854931	2	0	0	0	0
349	C-RNTI	99	6	4	0.000029	44.444443	0	612	796031	0.803706	0	0	0	0	0
350	C-RNTI	100	348	425	0.000428	89.688644	0	362	951314	0.955626	0	0	0	0	0

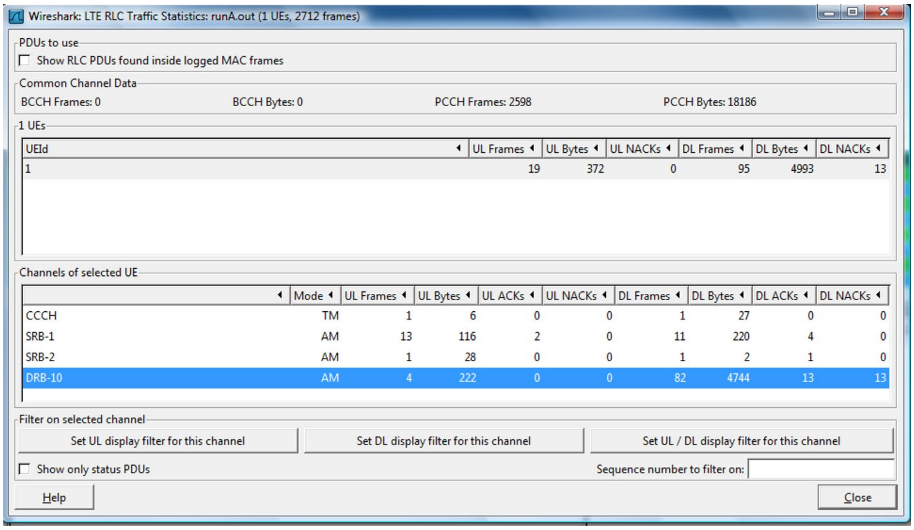
Selected UE details

	CCCH	LCID 1	LCID 2	LCID 3	LCID 4	LCID 5	LCID 6	LCID 7	LCID 8	LCID 9	LCID 10	Predefined
UL SDUs	0	0	0	4	0	0	0	0	0	0	0	0
UL Bytes	0	0	0	27	0	0	0	0	0	0	0	0
DL SDUs	0	0	0	336	0	0	0	0	0	0	0	0
DL Bytes	0	0	0	682577	0	0	0	0	0	0	0	0

Filter on UE

Set UL display filter on selected this UE
Set DL display filter on selected this UE
Set UL / DL display filter on selected this UE
□ Show MAC RACH
□ Show MAC SRs
□ Show DCT2000 error strings
...containing
Help
Close

Fig. 10 Protocol analyzer MAC-LTE statistics screenshot



Wireshark: LTE RLC Traffic Statistics: runA.out (1 UEs, 2712 frames)

PDUs to use

□ Show RLC PDUs found inside logged MAC frames

Common Channel Data

BCCH Frames: 0 BCCH Bytes: 0 PCCH Frames: 2598 PCCH Bytes: 18186

1 UEs

UEId	UL Frames	UL Bytes	UL NACKs	DL Frames	DL Bytes	DL NACKs
1	19	372	0	95	4993	13

Channels of selected UE

	Mode	UL Frames	UL Bytes	UL ACKs	UL NACKs	DL Frames	DL Bytes	DL ACKs	DL NACKs
CCCH	TM	1	6	0	0	1	27	0	0
SRB-1	AM	13	116	2	0	11	220	4	0
SRB-2	AM	1	28	0	0	1	2	1	0
DRB-10	AM	4	222	0	0	82	4744	13	13

Filter on selected channel

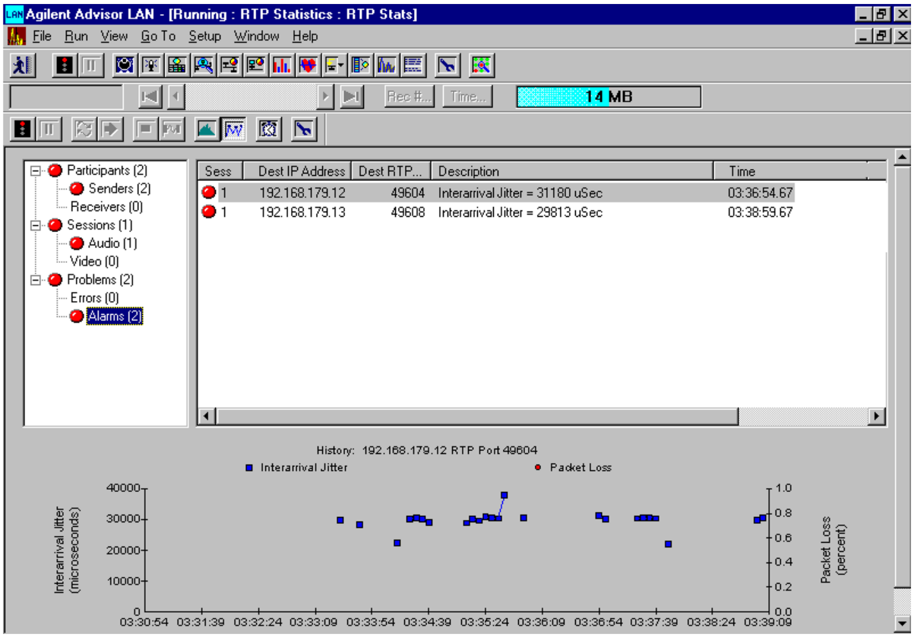
Set UL display filter for this channel
Set DL display filter for this channel
Set UL / DL display filter for this channel
□ Show only status PDUs
Sequence number to filter on:
Help
Close

Fig. 11 Protocol analyzer RLC-LTE statistics screenshot

However, in accordance with our testing goal of focusing the LTE protocol stack—the MAC-layer and the RLC-layer PDUs in particular, mere protocol decodes of individual retransmitted IR HARQ code-block redundancy versions rv1, rv2 and rv3, as well as of the ARQ NACKs, will not be of much value, so we used them in the protocol analyzer statistical analysis as trigger events to count uplink (UL) and downlink (DL) retransmissions ULReTX and DLReTX for MAC-HARQ, and negative acknowledgements UL-NACKs and DL-NACKs for RLC-ARQ.

The exemplar screenshots for the MAC-LTE and the RLC-LTE statistics of retransmissions are shown in Figs. 10 and 11 (the mid and the most right columns), respectively.

Such test results can be used to provide the time-variant statistics of retransmissions, so that its impact upwards the VoLTE protocol stack can be identified and traced.



Agilent Advisor LAN - [Running : RTP Statistics : RTP Stats]

Participants (2)
Senders (2)
Receivers (0)
Sessions (1)
Audio (1)
Video (0)
Problems (2)
Errors (0)
Alarms (2)

Sess	Dest IP Address	Dest RTP...	Description	Time
1	192.168.179.12	49604	Interarrival Jitter = 31180 uSec	03:36:54.67
1	192.168.179.13	49608	Interarrival Jitter = 29813 uSec	03:38:59.67

Fig. 12 RTP packet jitter and loss

In paralel with the HARQ-retransmissions-triggered statistics and decodes, we track the statistical analysis of mutually time-correlated RTP packets jitter and loss, Fig. 12.

Accordingly, from the example in Fig. 12, it is evident that the packet jitter produces no significant simultaneous packet loss in this case, so some randomly scattered packet loss bursts (definitely not desired for some applications) that were detected, do not seem to be related to packet delay variations, but to other impairments that may have occurred and manifest themselves at various layers of the protocol stack.

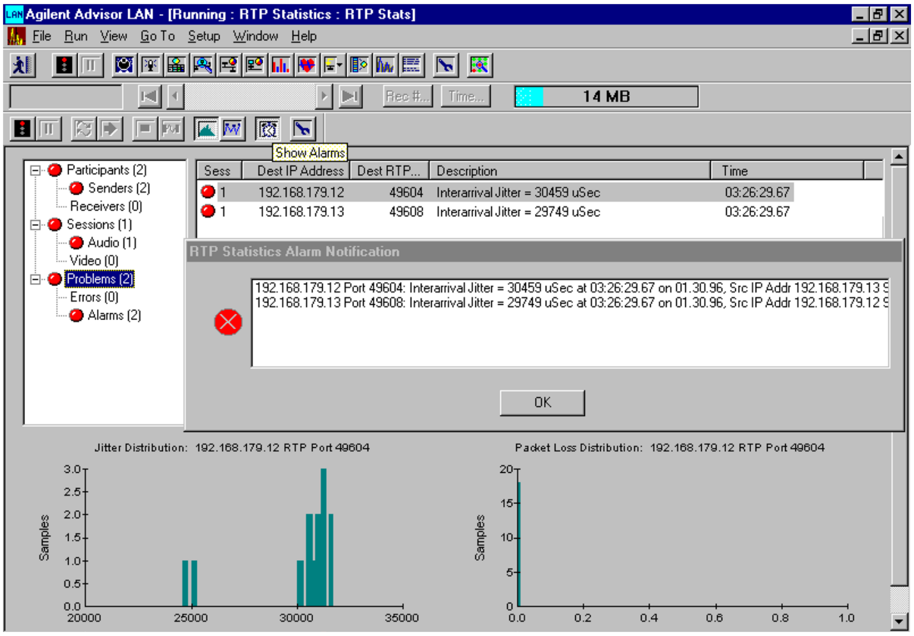
Furthermore, it is of interest here to present the histogram of packet jitter values, which approximates the jitter probability distribution, and thus helps in identification of dominant jitter accumulation values.

Accordingly, as coming out of the example related to Fig. 13, the presented RTP jitter does not seem to be uniformly distributed within the occupied range, but accumulates at its certain values, whose detection and isolation is important in the system design, specifically with regard to VoLTE delay budget limitation.

This implies the need to test the hypothesis that the RTP jitter probability distribution follows a specific model, such as the uniform or the normal one.

Furthermore, the task to follow is checking how the IP/RTP delay variations influence the end-to-end QoE. So, the application-layer average delay is measured either directly, Fig. 14, or assessed indirectly from the measured impulse response, Fig. 15.

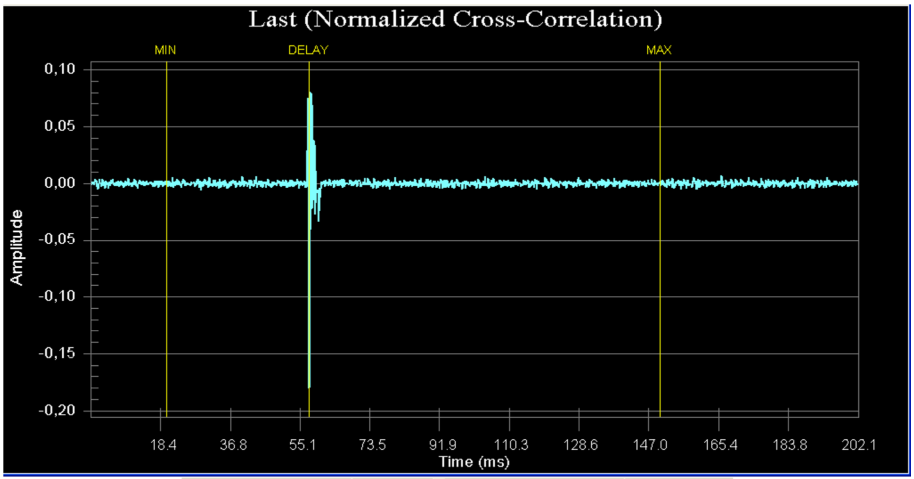
In the first case, the end-to-end normalized time-correlation between the transmitted test sequence and its replica coming back via the channel under test from the loop-backed receiver, is measured out-of-service with 1 ms resolution [1], and the time delay conforming to its maximal value is considered to be the average delay, Fig. 14. To get accurate results this way, it is important that the test sequence lasts at least three times the expected average delay that is to be measured.


Agilent Advisor LAN - [Running : RTP Statistics : RTP Stats]
File Run View Go To Setup Window Help
Rec #... Time... 14 MB
Show Alarms
Participants (2)
Senders (2)
Receivers (0)
Sessions (1)
Audio (1)
Video (0)
Problems (2)
Errors (0)
Alarms (2)
RTP Statistics Alarm Notification
OK
Jitter Distribution: 192.168.179.12 RTP Port 49604
Samples
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20000
25000
30000
35000
Packet Loss Distribution: 192.168.179.12 RTP Port 49604
Samples
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0

Sess	Dest IP Address	Dest RTP...	Description	Time
1	192.168.179.12	49604	Interarrival Jitter = 30459 uSec	03:26:29.67
1	192.168.179.13	49608	Interarrival Jitter = 29749 uSec	03:26:29.67

```
192.168.179.12 Port 49604: Interarrival Jitter = 30459 uSec at 03:26:29.67 on 01.30.96, Src IP Addr 192.168.179.13 S
192.168.179.13 Port 49608: Interarrival Jitter = 29749 uSec at 03:26:29.67 on 01.30.96, Src IP Addr 192.168.179.12 S
```

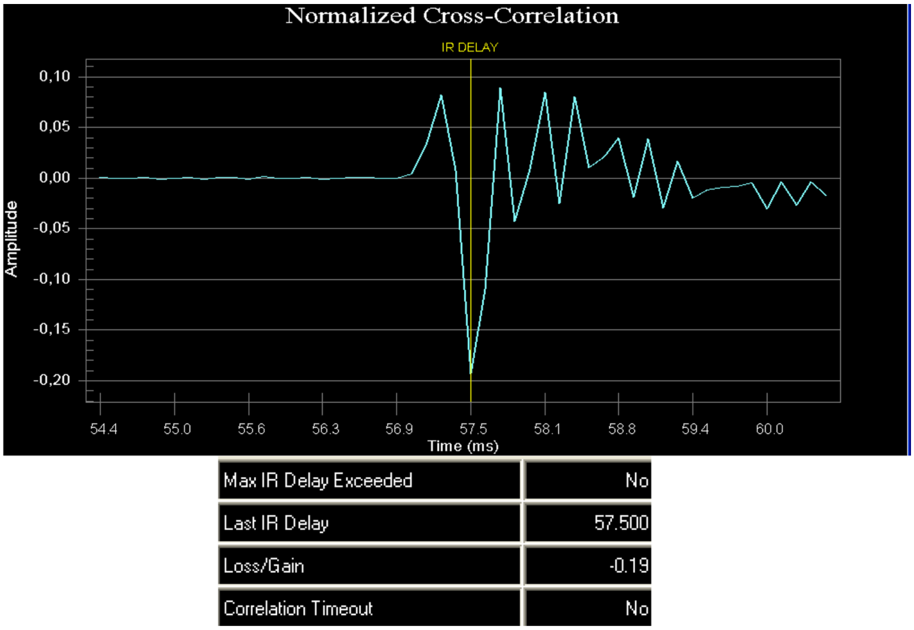
Fig. 13 Histogram of RTP packets jitter and loss



Last Delay (ms)	57.500	Average Delay (ms)	57.500
Maximum Delay (ms)	57.500	Avg Delay Thresh (ms)	100.0
Max Threshold (ms)	150.000	Minimum Delay (ms)	57.500
Max Exceeded Count	0	Min Delay Thresh (ms)	20.000
Std. Deviation (ms)	0.000	Test Duration	00:00:11
Truncated Data	No	Tests Completed	1

Fig. 14 Objective voice delay test results

In the second case, as presented in Fig. 15, the amplitude peak of the measured channel impulse response conforms to maximal energy transfer that occurs at the channel group delay, which is actually the average delay of interest here.



Max IR Delay Exceeded	No
Last IR Delay	57.500
Loss/Gain	-0.19
Correlation Timeout	No

Fig. 15 Average delay from impulse response measurement

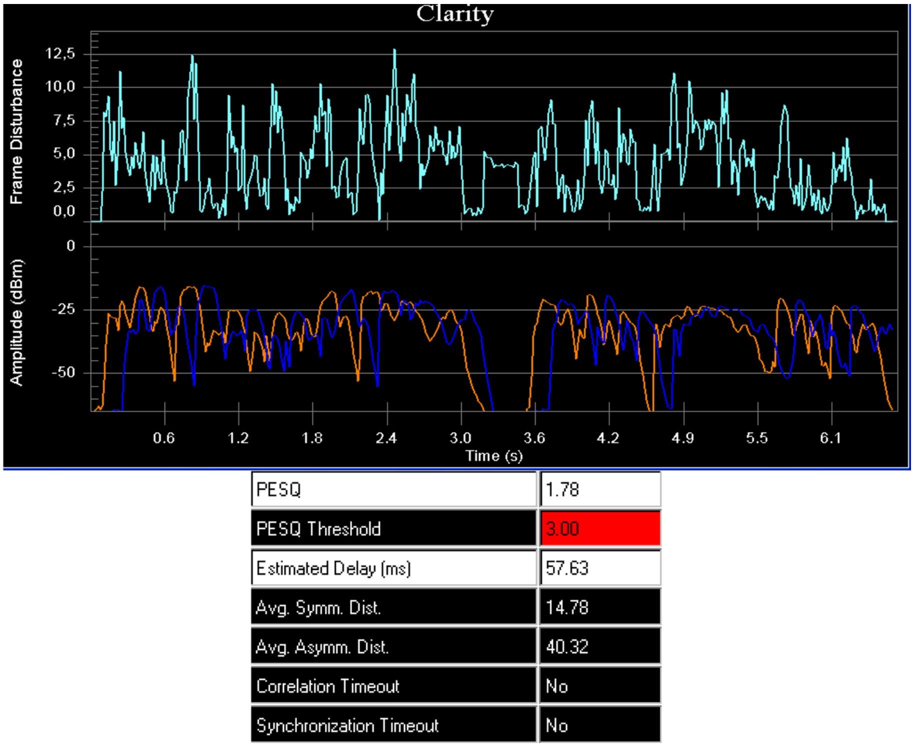
Moreover, as any VoIP uses adaptive jitter buffers, measured delay values may vary in time. Therefore, a series of measurements needs to be done, and average value found and adopted. (Averaging is also useful in case of considerable channel noise.)

Concretely, in Figs. 14 and 15, we can see that the mean delay value, measured in the range between 20 and 150 ms, is 57.5 ms, which is significantly larger than the corresponding IP/RTP jitter accumulation value of slightly above 31 ms identified in Fig. 13. Consequently, it comes out that, in this exemplar case, the IP packet delay is significantly influencing the perceptual one, but other sources of additional delay exist as well.

This implies that the IP/RTP QoS analysis does not solely determine the end-to-end QoE, i.e. how well does the speech actually "sound", as poor RTP QoS for sure leads to poor end-to-end voice quality (VQ), but it is not certain that satisfactory RTP QoS test values will necessarily provide alike VQ ones. Therefore, we need to involve another level of correlation with the HARQ/RLC retransmissions, i.e. make the end-to-end VQ testing.

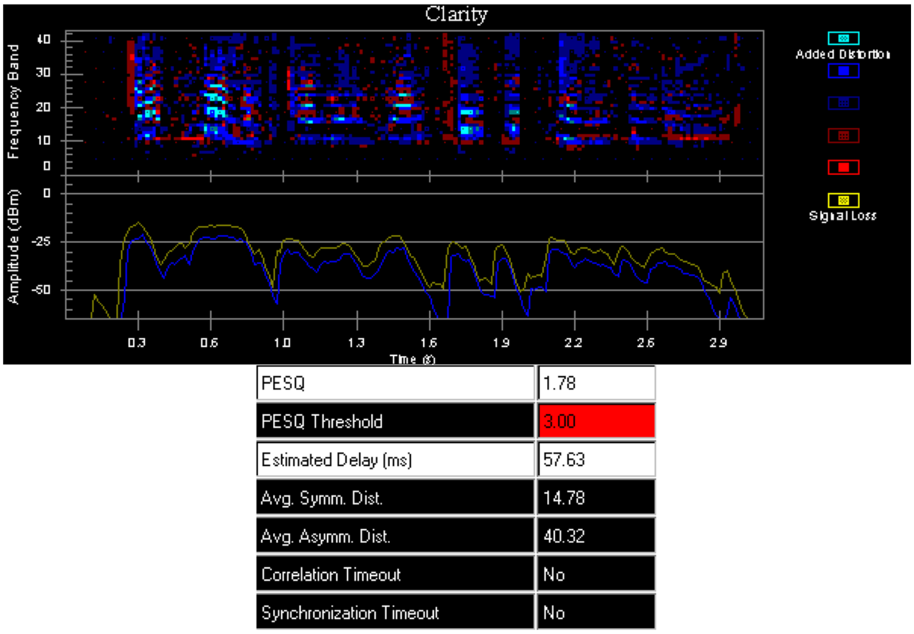
Generally, Mean Opinion Score (MOS) is the traditional metric for evaluation of voice quality in any voice network. However, as it is a subjective method, for quite a while, the objective ones have been in use as well. Specifically, we tested clarity using the VoLTE—recommended and ITU-T Rec. P.862 based Perceptual Evaluation of Speech Quality (PESQ) method, providing raw scores in the range − 0.5 to 4.5, and being linearly related to MOS. Apparently, in Fig. 16, the frame-by-frame disturbance values, correlated with input and output signal graphs, are graphically presented to show quality variances and momentary impairments.

Moreover, the 2-D Error surface shows magnitude of audible errors in the output signal, for both added distortion and subtractive distortion, Fig. 17.



PESQ	1.78
PESQ Threshold	3.00
Estimated Delay (ms)	57.63
Avg. Symm. Dist.	14.78
Avg. Asymm. Dist.	40.32
Correlation Timeout	No
Synchronization Timeout	No

Fig. 16 Clarity test results



PESQ	1.78
PESQ Threshold	3.00
Estimated Delay (ms)	57.63
Avg. Symm. Dist.	14.78
Avg. Asymm. Dist.	40.32
Correlation Timeout	No
Synchronization Timeout	No

Fig. 17 2-D Error surface clarity test results

Finally, in the following, we will deal with the main goal of the presented practical test setup, which is to find out to what extent does the earlier identified (by means of the rv-based protocol analysis triggers) HARQ rv1–rv3 retransmissions worsen the end-to-end application-layer QoE, through the consequent additional RTP packet delay/jitter that they introduce?

3.1 Test setup

At first, following the LTE-recommended maximal IP/RTP delay of 50 ms, we introduced this large average delay with variable jitter of up to 30 ms applied to the IP/RTP voice carrying traffic packets.

Then, on top of that, we kept simulating up to three HARQ retransmissions by accordingly modifying the traffic generator, and tracked how was this reflected at the QoE level—the VQ clarity, in particular. (As at this time we made just the preliminary tests, these could easy be extended to include testing the delay at both IP/RTP and VQ level.)

3.2 Statistical analysis

The collected HARQ RTT data are continuous, while the QoE raw clarity scores can be regarded as ordinal data, measured and expressed according to the usual 5-point Likert scale: Excellent; Good; Fair; Poor; Bad [11].

So, we need to investigate whether there is a relationship between the QoE and the average HARQ RTT. However, common parametric tests are not any good choice for statistical analysis of ordinal data that we were to conduct here [1], so we used the nonparametric Spearman rank-order correlation test to find the correlation coefficient (ρ) that is a good measure of strength and direction of the relationship between the continuous and the ordinal random variables.

So, considering the QoE and HARQ RTT as being represented by paired observations, our preliminary investigation revealed monotonic relationship between them. The null and the alternative hypotheses are as it follows, respectively:

H0: There is no association between the QoE and HARQ RTT.
H1: There is an association between the QoE and HARQ RTT.

The significance of a test result is articulated with the p-value; the smaller it is, the more significant is the result. With this respect, the comparison reference that is commonly referred to as significance (α), is in fact the probability of the "false positive" decision about null hypothesis rejection when it should be accepted [12]. So e.g., if the null hypothesis is rejected ($p < \alpha$ found), then smaller α value implies stronger evidence that the finding is statistically significant [12]. In our analysis, we adopted a moderate value of $\alpha = 1\%$.

Furthermore, having found whether and to what extent (excessive) HARQ retransmissions incur notable QoE degradation, let us now investigate the correlation between delay variations of HARQ and RTP PDUs.

With this regard, as the Kolmogorov–Smirnov (K–S) test depends neither on particular statistics nor on sizes of the observed samples [12], we applied it to check whether the measured RTP jitter values (modeled as the random variable ξ with the continuous cumulative distribution functions (cdf) $F_\xi(x)$), presented in Fig. 13, are associated with the average HARQ RTT values, modeled as the random variable η with the continuous cdf $G_\eta(y)$, due to frequent retransmissions at PHY/MAC layers.

Accordingly, the null hypothesis is [12]:

$$H_0: \; F_\xi(x) = G_\eta(y) \tag{1}$$

If we observe the random variables ξ and η, represented by m and n mutually independent finite samples $\xi_1, \xi_2, \ldots, \xi_m$ and $\eta_1, \eta_2, \ldots, \eta_n$, exhibiting the empirical continuous cdfs $\hat{F}_{m\xi}(x)$ and $\hat{G}_{n\eta}(y)$, respectively, then the according two-sample K–S statistics [12]:

$$D_{m,n} = \sup_{x,y} \left| \hat{F}_{m\xi}(x) - \hat{G}_{n\eta}(y) \right| \tag{2}$$

provides a reliable nonparametric criterion for identifying and qualifying the similarity between the samples, by testing if $D_{m,n}$ converges to zero.

Accordingly, we recall here the Kolmogorov limit distribution theorem, stating that [12]:

$$\lim_{m,n\to\infty} \Pr\left(\sqrt{\frac{mn}{m+n}} D_{m,n} < z\right) = K_\zeta(z), \quad 0 < z < \infty \tag{3}$$

where $K_\zeta(z)$ is the Kolmogorov cdf.

For the particular argument value $z{=}z_\alpha$, which makes $K_\zeta(z)$ equal to the complement 1-α of the significance α:

$$K_\zeta(z) = 1 - \alpha \;\; \Rightarrow \;\; z_\alpha = K_\zeta^{-1}(1-\alpha) \tag{4}$$

the null-hypothesis (1) should be rejected under condition that:

$$\sqrt{\frac{mn}{m+n}} D_{m,n} > z_\alpha \tag{5}$$

whereas, in the opposite case, the null-hypothesis (1) is to be accepted with significance α.

Finally, if the estimated p-value is such that:

$$p = 1 - K_\zeta\left(\sqrt{\frac{mn}{m+n}} D_{m,n}\right) < \alpha \tag{6}$$

then the null hypothesis should be rejected with significance α, otherwise, the decision should be opposite.

3.3 Preliminary test results

Our preliminary tests revealed that direct relationship exists between the PESQ voice quality rating, the IP/RTP packet latency, and the Block Error Rate (BLER) of the received MAC/RLC frames.

Table 1 Correlation PESQ versus HARQ RTT

Spearman ρ PESQ versus HARQ RTT	p value	*BLER*	H_0
− 0.514	0.034	0.2	Rejected
− 0.328	0.087	0.15	Rejected
− 0.107	0.109	0.1	Accepted
− 0.040	0.232	0.05	Accepted

Specifically, In Table 1, the values of the Spearman correlation coefficient between the end-to-end user PESQ QoE score and the average HARQ RTT (including retransmitted rvs as well), are presented for particular *BLER* and p-values.

As it can be seen above, when the actual BLER value exceeds the optimal one of 10% for LTE HARQ [2], the correlation coefficient gets significantly different from zero with small enough p-value (lesser than $\alpha = 0.1$) meaning that the null hypothesis can be rejected, i.e. that there is an association between the QoE and the HARQ RTT values.

This can be explained by increasing the HARQ RTT with more repetitive retransmissions (due to the larger BLER), where the residual erroneous code-blocks conform to the case when all 4 code-word redundancy versions were transferred unsuccessfully, with as few as just 0.1% NACKs erroneously transferred back as ACKs.

On the other hand, as from Table 1 no significant impact of HARQ RTT onto voice QoE, i.e. PESQ rating, is evident for the lower BLER values, this can be explained by fewer erroneous code-blocks, i.e. less time-consuming repetitive retransmissions to make excessive delay that is harmful for the VQ QoE.

This practically implies that only up to 2 HARQ retransmissions are appropriate during any voice packet, otherwise it might be impossible to "smooth out" the delay accumulation by jitter/playback buffers along the propagation path. Essentially, this can be seen in Fig. 7, as the HARQ rv1 coding gain (graphically, horizontal displacement) with respect to rv0 is dominant among the other 3 rv gains, meaning that in majority of cases, only one retransmission is quite enough to preserve the projected optimal BLER value of 10%.

This is good enough for voice, too, while providing no considerable QoE-harming delay. (Had smaller BLER be required, it would have requested more redundancy—parity bits, whereas allowing higher BLER would have led to excessive retransmissions, also reducing the throughput.)

Thereby, following a single retransmission, the post-HARQ BLER equals 10% times 10%, which is just 1%, so that the percentage of error-free voice RTP packets reaches satisfactory level of 99%. Consequently, as the LTE RLC protocol layer's ARQ mechanism is usually left to retransmit just 1% of failed-HARQ data code-blocks, in case of VoLTE, it can be practically neglected. Moreover, as we already elaborated that the final HARQ retransmission (rv3) increases the overall delay in VoLTE, it is even more so with the very final code-word retransmission by RLC ARQ. Fortunately, as the HARQ coding gain is mostly assigned to the first 2 retransmissions, this implies that rv3 occurs very rarely, so the final (post-HARQ) retransmission to be done at the RLC layer, is even much less probable. This practically removes RLC ARQ as considerable threat to VoLTE latency.

Table 2 Correlation PESQ versus HARQ RTT

$H_0: F_\xi(x) = G_\eta(y)$	p-value	*BLER*	H_0
	0.142	0.2	Accepted
	0.111	0.15	Accepted
	0.089	0.1	Rejected
	0.022	0.05	rejected

Furthermore, having found that only really excessive HARQ retransmissions incur notable QoE degradation, we applied the K-S test to check to what extent is the measured RTP delay variation in consistence with the average HARQ RTT due to frequent retransmissions at PHY/MAC layers.

So, the results of practical testing of association between the HARQ RTT and RTP delay values are presented in Table 2.

As it can be seen, for the BLER values above the ideal figure of 10%, the null hypothesis is accepted, as practically it came out that the cdfs of the HARQ RTT and RTP delay are mutually much alike, whereas in the opposite case of BLER < 10%, the p-values are larger than the adopted value for α, meaning that the association does not hold any more.

This implies that, under the condition of BLER > 10%, the IP/RTP delay and its variations are predominantly determined by the HARQ RTT delay.

4 Conclusions

We proposed and demonstrated a VoLTE QoS and QoE test procedure based on PHY/MAC/RLC/IP/TCP-UDP/RTP cross-layer protocol analysis specifically focusing PHY/MAC IR-HARQ and RLC ARQ induced delay, with regard to perceptual speech quality measurements reflecting the end-to-end QoE.

The nonparametric Spearman's rank-order correlation test identified monotonic relationship between the paired observations: QoE and HARQ RTT, i.e. between the PESQ voice quality rating and the IP/RTP packet latency, for given BLER of the received MAC/RLC code-blocks.

As most of IR-HARQ coding gain is achieved between the redundancy versions rv0 and rv1, then often even a single retransmission enables practically error-free code-block transfer, which imposes just a little additional delay burden within the allowed budget, and so exhibits no practical impact on the end-user QoE (expressed as PESQ in this case). Finally, it comes out that up to 2 HARQ retransmissions are appropriate during any voice packet, otherwise it might not be possible to "smooth out" the delay accumulation by jitter/playback buffers along the propagation path.

Moreover, having found that only really excessive HARQ retransmissions incur notable end-to-end QoE degradation, by applying the Kolmogorov–Smirnov tests, we identified strong association between the RTP jitter values and the average HARQ RTT, specifically for non-optimal (higher) BLER values.

This preliminary work is aimed to present and partly verify how this concept can be used for integral QoS and QoE assessment during installation, commissioning and maintenance of VoLTE networks, and so pave the way to according R&D and field tests taking into account design and deployment issues as well, and using sophisticated hardware and industry-standard software tools.

Abbreviations

ARQ: Automatic repeat request; BLER: Block error rate; CRC: Cyclic redundancy check; FDD: Frequency-division duplex; HARQ: Hybrid automatic repeat request; IP: Internet protocol; IR HARQ: Incremental redundancy HARQ; LTE: Long term evolution; MAC: Medium access control; MOS: Mean opinion score; PESQ: Perceptual evaluation of speech quality; PDU: Protocol data unit; PHY: Physical layer; QoE: Quality of experience; QoS: Quality of service; RLC: Radio link control; RTT: Round trip time; RTP: Real-time transport protocol; TB: Transport block; TCP: Transmission control protocol; TTL: Time-to-live; UDP: User datagram protocol; VoIP: Voice over IP; VQ: Voice quality.

Authors' contributions

AL developed the cross-layer VoLTE test procedure through PHY/MAC/RTP layers up to the end-to-end speech quality, applying the according nonparametric statistics to track correlation between voice quality and the RTP packet latency with HARQ RTT. VL set up general guidelines for this work, specifically with regard to the analysis of test results, and their proper interpretation, as well as model verification. IG accordingly accomodated the acquired test data to match specific input requirements of the selected statistical analysis models. IO preliminary tested the selected simulation tools, as well as made according adjustments of the programs to execute the selected statistical analysis. All authors read and approved the final mauscript.

Funding

Not applicable.

Availability of data and materials

Not applicable.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 3 December 2019 Accepted: 30 March 2021
Published online: 07 April 2021

References

1. V. Lipovac, practical consistence between bit-error and block-error performance metrics up to application layer. Wirel. Pers. Commun. **93**(3), 779–793 (2017)
2. M. Rumnay, "LTE and the Evolution of 4G Wireless; Design and Measurements Challenges", 2nd edition, John Wiley & Sons, 2013
3. S. Donthi, N. Mehta, An accurate model for EESM and its application to analysis of CQI feedback schemes and scheduling in LTE. IEEE Trans. Wirel. Commun. **10**(10), 3436–3448 (2011)
4. Elnashar and M. El-Saidny, "Looking at LTE in practice: a performance analysis of the LTE system based on field test results," IEEE Vehicular Technology Magazine, vol. 8, no. 3, pp. 81–92, September 2013
5. V. Buenestado, J. Ruiz-Aviles, M. Toril, S. Luna-Ramirez, A. Mendo, Analysis of Throughput Performance Statistics for Benchmarking LTE Networks. IEEE Commun. Lett. **18**(9), 1607–1610 (2014)
6. http://www.itu.int/rec/T-REC-P.862/en; 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
7. A. Lipovac, V. Lipovac, M. Hamza, Verification of OFDM error floor prediction in time-dispersive LTE FDD DL channel. Wirel. Pers. Commun. **93**(3), 853–875 (2017)
8. ITU-T Recommendation G.1028, "End-to-end quality of service for voice over 4G mobile networks", April 2016
9. S. Pagès, "Link Level Performance Evaluation and Link Abstraction for LTE/LTE-Advanced Downlink" Ph.D. dissertation, Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona, 2013
10. S. Hemminger, "Network Emulation with NetEm", Open Source Development Lab, 2005
11. Z. Govindarajulu, "Rank Correlation Methods (5th ed.)", Taylor & Francis Online, March 12, 2012
12. M. Kendall, A. Stewart, "The Advanced Theory of Statistics", *Charles Griffin* London, 1966

Publisher's Note