2017 International Conference on Computing, Networking and Communications (ICNC): Communications QoS and System Modeling

Delay and Quality Metrics in Voice over LTE (VoLTE) Networks: an End-Terminal Perspective

Najmeddine Majed[1,2], Stéphane Ragot[1], Xavier Lagrange[2], Alberto Blanc[2]
[1]Orange Labs, Lannion, France [2]Telecom Bretagne, Rennes, France

E-mail:{najmeddine.majed, stephane.ragot}@orange.com
E-mail:{xavier.lagrange, alberto.blanc}@telecom-bretagne.eu

***Abstract*—In this paper we evaluate metrics specified in 3GPP to characterize trade-offs between delay and quality of Voice over LTE (VoLTE) mobile phones in various network conditions. We report test results on clock accuracy, terminal delay in uplink and downlink under error-free conditions, as well as delay and quality in the presence of packet losses and network jitter. We discuss how the underlying methodology intended for delay testing can be extended to evaluate de-jitter buffer performance using a black-box approach, and how to model VoLTE packet delay/loss characteristics in a realistic way.**

***Index Terms*—VoLTE, LTE, network model, QoE, packet delay.**

I. Introduction

Mobile operators have recently deployed Voice over LTE (VoLTE) to add the support of telephony services in Long-Term Evolution (LTE) and Evolved Packet Core (EPC) networks. Unlike mobile networks from previous generations that used circuit-switching (CS) for voice and packet-switched for data, fourth Generation (4G) networks are "all-IP" networks. VoLTE is a form of mobile voice over IP (VoIP) with specific network optimizations and using the IP Multimedia Sub-System (IMS) to provide network quality of service (QoS) [1]. Several key elements of VoLTE are left unspecified and proprietary (e.g., scheduling in eNodeB, de-jitter buffer in mobile phones). These aspects can have a strong influence on performance. Moreover, mobile operators face an increased competition from Over The Top (OTT) players, and quality of experience (QoE) has become a major issue to ensure end-users get the best possible call quality in this competitive environment.

QoE is affected by many factors [2], including audio quality, service availability, cost, security, and we limit ourselves to the audio quality dimension, which can be characterized by various metrics, such as Mean Opinion Score (MOS) [3], [4], mouth-to-ear delay, perceived loudness and frequency spectrum, interruptions (audio gaps). QoE represents quality as perceived by the end user; however, one may split contributions from end terminals and network to derive QoE models in a more tractable and modular way [5]. In this paper, we follow an approach where mobiles phones are characterized with well-defined input/output reference points; the complete audio chain is modeled by uplink and downlink metrics (characterizing end terminals) combined with end-to-end network parameters (e.g., delay/loss packet traces).

A key difference between CS and VoIP mobile phones is that VoIP terminals have to compensate for independent clocks in end points [6] and asynchronous transport over IP. A de-jitter buffer is used in the VoIP receiver to smooth out packet delay variations; other parts in the audio path (e.g., codecs, noise reduction, gain control, echo cancellation) may be considered unchanged or similar in performance for a given phone operated in CS or VoLTE. For this reason, we focus on metrics characterizing quality impairments resulting from clock skew and packet delay variations, including the effect of de-jitter buffers.

A comprehensive review of quality metrics in VoIP can be found in ITU-T G.1020 [5] and G.1021 [7], including network, terminal and overall metrics. These specifications also provide an example of de-jitter buffer model, with an analysis of de-jitter buffer types and metrics. Test methods and performance targets on the quality of de-jitter buffer adjustment and the efficiency of delay variation removal in VoIP terminals are defined in [8]. De-jitter buffer size estimation and optimization with respect to QoE is discussed in [9]. The impact of de-jitter buffer playout delay adjustments has been studied for instance in [10], [11].

Several methods have been proposed to measure delay and characterize de-jitter buffer performance in VoIP. In many cases, delay is measured at the IP level and not at the acoustic level [12], [13]. Black-box delay measurements at the acoustic level have been reported for instance in [14], where mouth-to-ear delay was estimated by cross-correlation between the recorded original and output audio of VoIP phones; delay was reported in terms of average delay. It may be noted that clock synchronization of end points was not used and network conditions were not time synchronized with audio signals to ensure repeatable measurements in separate calls.

Several test methods have been standardized to measure terminal delay at the acoustic level. In CS voice services, terminal delay is independent from network conditions, due to the synchronous transmission of speech data. For 2G, terminal delay test methods and requirements have been defined in [15, clause 32] under error-free conditions. For 3G, terminal delay tests have been defined in Release 11 of TS 26.131 [16] and TS 26.132 [17]; these tests under error-free conditions

978-1-5090-4588-4/17/$31.00

have been extended to LTE terminals in Release 12 of these specifications.

This work is based on LTE terminal delay testing specified in 3GPP [16], [17]. The main contribution of this paper is to analyze in details the existing test method, with results illustrating the associated metrics, and to investigate how this methodology can be extended to evaluate the performance of de-jitter buffers in realistic conditions. In particular, we propose enhancements to delay/loss packet trace simulations to better represent the actual delay and quality that can be experienced in VoLTE.

The paper is organized as follows. In Section II, we detail the experimental set-up. In Section III, we consider the clock accuracy metric. In Section IV, we present test results for delay metrics for the uplink and downlink under error-free conditions. In Section V, we present test results for delay and quality metrics for the downlink for network conditions simulated with delay/loss profiles. In Section VI, we discuss the issue of generating delay/loss packet traces that model realistic VoLTE scenarios, before concluding in Section VII.

II. Experimental Set-up for VoLTE Delay Testing

The test set-up used in this work follows 3GPP acoustic tests defined in [17]. We used an example of implementation based on test equipments from different vendors, as detailed in [18]. The test set-up is specific to the case of a mobile phone in handset mode – the headset and hands-free modes are not considered here. Testing is conducted separately for the uplink (sending direction) and downlink (receiving direction). Note that the de-jitter buffer is located on the receiver side of the mobile phone, hence most tests focused on downlink tests.

The mobile phone is mounted in handset mode on a manikin – also called head and torso simulator (HATS) [19] – with built-in artificial ear [20] (Type 3.3) and artificial mouth [19]. A specific SIM card with proper operator settings is inserted in the mobile phone for testing purposes. A Voice over LTE (VoLTE) call is established with an LTE/EPC network emulator (Rhode & Schwarz CMW500) using its internal IMS server to setup a dedicated bearer with $QCI = 1$ [1]. All tests have been conducted with mobile originated calls with the AMR-WB codec at 12.65 kbit/s and discontinuous transmission (DTX) deactivated. A computer operating a measurement system (Head Acoustics ACQUA) compliant with [17] is used to conduct testing, collect and analyze test data. The measurement system is connected to an acoustic front-end, called *reference client* (Head Acoustics MFE VIII.1), which implements codecs (AMR-WB), a de-jitter buffer for uplink tests, and an IP network emulator to inject delay/loss conditions for downlink tests. The LTE/EPC network emulator is set in forwarding mode, hence testing take places as if the VoLTE call was between the mobile phone and the acoustic front-end. Moreover another front-end (Head Acoustics MFE VI.1) is used as an audio interface performing A/D and D/A conversion between the manikin and the reference client.

For delay measurements, it is essential to estimate and compensate for clock skews between end points. Based on [21], we define here relative *clock skew* as the difference between the frequencies of two clocks. Clock skew is estimated at the acoustic level by regularly repeating the same delay measurement using a CSS test signal [17] in the same call and by calculating the slope of the resulting delay curve after rectifying delay jumps [22], [17, Annex D]. Note that in this work the reference client (MFE VIII.1) had a clock frequency of 48000 Hz; this clock frequency was reset prior to measuring clock skew, and it was then adjusted to compensate the estimated clock skew. All test results presented in this paper have been performed after synchronizing the clocks of the mobile phone and reference client.

III. Clock Accuracy of VoLTE Mobile Phones

Table I lists the relative clock skew obtained as a by-product of clock synchronization in sending and receiving for delay measurements for different VoLTE mobile phones. It can be noted that the absolute value was below 3 ppm for the tested mobile phones and repeating clock skew estimation produced results of the same order. This may be explained by the fact that typically the underlying audio clock in the mobile phone (chipset) can compensate for temperature and be adjusted based on a network clock. Based on these results, one might conclude that clock skew may not have a significant impact on QoE for VoLTE and could be neglected in a first approximation. However, VoLTE calls with a software client running at the application level (e.g., a laptop with an LTE modem, or mobile phone with voice processing outside the chipset) may not have such an accurate clock.

TABLE I
Clock skew in sending and receiving.

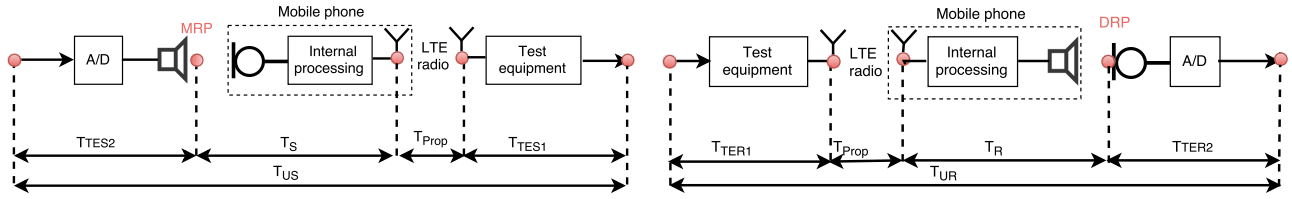
Phone	Min. clock skew (ppm)	Max. clock skew (ppm)
A	-2.7	-0.3
B	0.2	1.3
C	0.4	0.7
D	0.4	0.6

IV. Uplink and Downlink Delay (Error-Free Case)

Terminal delay was measured under error-free conditions, i.e., no packet loss and nearly no jitter ($<$ 1 ms) from the network emulator. A Composite Source Signal (CSS) from [23] of 32000 samples (at 48 kHz sampling rate) was used as a test signal.

A. Terminal Delay in the Uplink: Definition and Measurement Methodology

The sending delay T_S of the mobile phone is defined by the delay between the first acoustic event at the Mouth Reference Point (MRP) of the artifical mouth and the last bit of the corresponding speech frame at the phone antenna [16], as illustrated in Fig. 1.a. To calculate the sending delay T_S, the uncompensated sending delay T_{US} is measured by cross-correlation between two *measurement points*, that is, between the output of the test equipment (reference client) and the original signal played at MRP. Then, the delay caused by the



(a) Sending (Uplink) (b) Receiving (Downlink)

Fig. 1. Sending and receiving delay measurement.

test equipment is subtracted; this includes the delay T_{TES1} of A/D conversion and the delay T_{TES2} of the other test equipment units (reference client, network simulator). The overall test equipment delay is $T_{TES} = T_{TES1} + T_{TES2}$. Note that the propagation time on the LTE interface is assumed to be negligible, i.e., $T_{Prop} = 0$ ms. Consequently, the sending delay can be evaluated as follows:

$$T_S = T_{US} - (T_{TES1} + T_{TES2} + T_{Prop}) = T_{US} - T_{TES} \quad (1)$$

For the test setup used in this work, we have $T_{TES} = 192.37$ ms which includes the following components:

- Reference client delay: 42.5 ms (including decoding and resampling operations)
- Reference client jitter buffer depth: 140 ms (7 frames of 20 ms) – note that this value is actually a user-defined parameter of the MFE VIII.1 equipment
- Network emulator delay: 9.47 ms [24]
- Acoustic front end delay (A/D): 0.4 ms

B. Terminal Delay in the Downlink: Definition and Measurement Methodology

The receiving delay T_R of the mobile phone is defined by the delay between the first bit of a speech frame at the phone antenna and the first acoustic event corresponding to that speech frame at the Drum Reference Point (DRP) of the artificial ear [16], as shown in Fig. 1.b. To calculate T_R, we measure the uncompensated delay T_{UR} by cross-correlation analysis between the measured signal at DRP and the original signal at test equipment input (reference client). The calculation of T_R is similar to the sending delay case and we have the following equation:

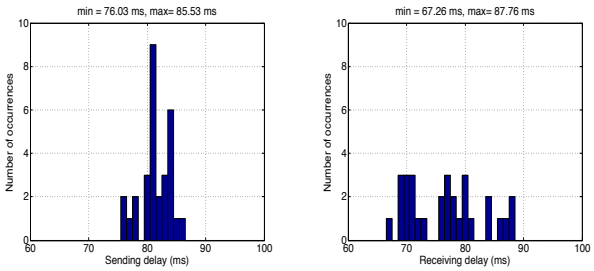
$$T_R = T_{UR} - T_{TER} \quad (2)$$

with $T_{TER} = 78.04$ ms which consists of:

- Reference client delay: 68.5 ms (including resampling and encoding operations)
- Network simulator delay: 8.73 ms [24]
- Acoustic front end delay (A/D): 1.31 ms

C. Test Results vs. Delay Targets

We repeated 30 times the same delay measurement (T_S and T_R) with separate calls using phone A. Establishing a new call was required to put the phone in a pre-determined state. The histogram of measured delays is shown in Fig. 2. One can verify that the histogram covers an interval of at most 20 ms which corresponds to the codec frame length. This measurement uncertainty has been attributed in [24] to random phase shifts between sending and receiving frames in the VoIP end points. This may be interpreted as the result of the clock offset between synchronized VoIP sender and receiver, where the offset is random in each separate call.



(a) Sending (Uplink) (b) Receiving (Downlink)

Fig. 2. Sending and receiving delay in error free condition.

In [17], it is required to repeat five times the sending and receiving delay measurement and to take the maximum value as the measured delay value. The histograms of T_S and T_R in Fig. 2 show that repeating the delay test only five times may not be sufficient to cover the expected delay variability. The five repetitions have been chosen to be a compromise between testing time and accuracy/repeatability.

Note that VoLTE terminal delay targets are specified in [16] as overall (send+receive) delay: $T_S + T_R \leq 190$ ms (mandatory) and ≤ 150 ms (recommended) for voice calls. We observe that the specific mobile phone considered here $T_S + T_R = 173.29$ ms which would pass the required limit of 190 ms. These targets consist of a *vendor specific implementation* part (≤ 83 ms recommended and $\leq$123 ms mandatory) and a fixed *implementation independent part* (67 ms) split into speech frame buffering (25 ms), LTE transmission time (1+1 ms) and a default jitter buffer depth of 40 ms (2 codec frames); these requirements have been defined with the idea of keeping the *vendor specific implementation* part identical to the CS case but replacing the part related to CS *implementation independent part* by its VoLTE counterpart.

V. Delay/Quality under Delay/Loss Conditions

The delay tests specified in [17] include conditions with simplified network impairments to verify that a mobile phone has a de-jitter buffer that can adapt to network conditions.

Since the design of a de-jitter buffer is a compromise between buffering/playout delay and quality [1] [25, chap. 8], in such non-ideal conditions, both delay and quality are measured. Quality is evaluated using P.863 [26] (also known as POLQA). The test signal is in this case a real speech signal [17] consisting of four 8-second English sentence pairs according to [23, Annex B.3.3] (two male/female speakers), which are repeated 5 times, resulting in an overall duration of $5 \times 8 \times 4$ = 160 seconds, that is, 8000 frames of 20 ms. Each speech sentence is centered within a 4-second time window.

A. Delay/Loss Packet Traces (Profiles) at the IP level

To emulate network impairments, delay/loss degradations at the IP level have been proposed in [27] for delay testing. 3GPP adopted three end-to-end profiles that have been generated by simulation, with a MATLAB source code given in [17, Annex E]. Note that for wideband calls only two profiles are applicable, and the third profile is intended for super-wideband voice call testing; in this work we use this extra profile to obtain additional data. Each profile consists of 8000 delay/loss entries corresponding to 160 seconds of speech transported in 20 ms RTP packets. The associated characteristics are summarized in Table II. The profiles simulate RTP packet impairments between the IP network interfaces of two VoLTE mobile phones, at the antenna reference points shown in Fig. 1. They model static jitter conditions (i.e., no mobility, no varying cell load) with a simplified handling of a dedicated bearer with $QCI = 1$ [1] and Discontinuous Reception (DRX) [28].

In this work, profiles were applied at the IP level by the reference client which combines a VoIP client and a network emulator (similar to *netem*) in the downlink of the mobile phone under test. Note that profiles were synchronized with audio packets, so that voice packets experienced the same network degradations in separate calls for a given condition; this audio/network impairment synchronization was implemented directly in the reference client. The receiving delay with network impairments, T_R^{imp}, is measured as in the error-free case (see Eq. 2), except that the minimum delay added by profiles (30 ms for the 3 profiles) is also subtracted.

TABLE II
JITTER/LOSS PROFILE PARAMETERS [17].

Model parameters/statistics	Cond1	Cond2	Cond3
Target BLER (%)	10	10	22
Max. of HARQ retrans.	2	2	2
Duration of DRX cycle (ms)	20	40	40
EPC jitter (ms)	6	6	8
Packet loss rate (%)	0.2375	0.2625	2.6375
Out of sequence packets (%)	0	10.575	9.1125
Avg. packet delay (ms)	43.48	66.03	67.84
Avg. jitter – RFC 3550 (ms)	10.14	27.63	27.04

B. Test Results and Comparison with Delay/Quality Targets

Fig. 3 shows measurement results obtained for the three profiles using the same VoLTE phone (phone A) as in error-free conditions, with ten repeats in separate calls. In [17], the measured receiving delay T_R^{imp} for each condition (profile) is defined as the $95th$ *percentile* of the delay values obtained per 4-second speech sentence, where the first two delay values are discarded to allow some convergence time for the de-jitter buffer. For each profile a quality score is computed using P.863 (in super-wideband mode) for each 8-second speech sentence pair (except the first one, for the same reason of de-jitter buffer convergence) and the resulting 19 scores are averaged to produce a mean MOS-LQO$_S$ (Mean Opinion Score-Listening Quality Only, super-wideband) value. The 10 repeats of each condition (1 to 3) show the same delay variability as in error-frame with an interval of at most 20 ms. Similarly, the quality score variability (around 0.1 MOS) is in the expected range for P.863. In [17], the test in delay/loss condition is performed only once for each profile. One can verify that delay increases when network condition gets worse, as the de-jitter buffer normally adapts its depth to compensate for network jitter. Unsurprisingly, quality is degraded when the packet loss rate is increased. Note that the delay range for Cond. 1 is actually lower than the delay range in ideal case (see Fig. 2 (b)); this can be explained by the fact that the initial delay of the de-jitter buffer is higher at call startup for the tested device and the CSS test signal used for receiving delay in ideal case is triggered when delay adaptation has not occured yet. These test results show that it would be better to repeat delay measurements in separate calls to capture variability. To avoid repeating this measurement, one could concatenate a CSS signal and a 160s speech signal to run both a short receiving delay measurement (T_R) in error-free case and a receiving delay (T_R^{imp}) with network impairments in the same call, and to adjust the value of T_R^{imp} by an offet based on the maximum value of T_R previously obtained after several repeats in error-free case.

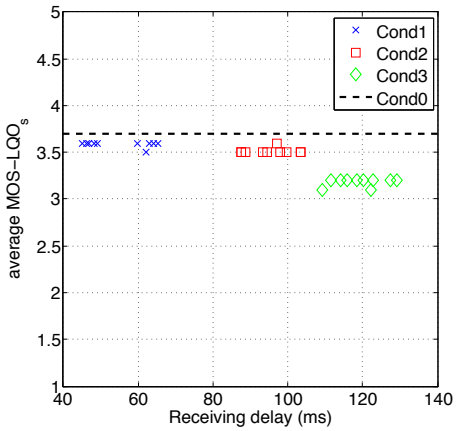


Fig. 3. Delay vs. speech quality in jitter/loss conditions (Cond0 corresponds to the error-free case and gives the reference P.863 score).

VoLTE terminal delay targets for impaired conditions are specified in [16] in terms of overall delay (send in error-free+receive in delay/loss condition) and quality degradation with respect to the error-free case (denoted here Cond0) as shown in Table III. The tested phone meets all requirements.

Fig. 3 illustrates a particular design choice with a certain trade-off between late loss rate (caused by late packets) and delay.

TABLE III
DELAY/QUALITY TARGETS FOR WIDEBAND CALLS [16] (WITH AN EXTRA CONDITION TAKEN FROM SUPER-WIDEBAND REQUIREMENTS).

	$T_S + T_R^{imp}$ (recommended)	$T_S + T_R^{imp}$ (mandatory)	MOS-LQO (mandatory)
Cond1	$\leq$ 150 ms	$\leq$ 190 ms	$\geq$ MOS-LQO$_{Cond0} - 0.3$
Cond2	$\leq$ 190 ms	$\leq$ 230 ms	$\geq$ MOS-LQO$_{Cond0} - 0.3$
Cond3	$\leq$ 190 ms	$\leq$ 230 ms	$\geq$ MOS-LQO$_{Cond0} - 1$

VI. TOWARDS DE-JITTER BUFFER PERFORMANCE METRICS USING REALISTIC VOLTE NETWORK MODELS

We investigate here how the 3GPP delay test methodology with delay/loss packet traces can be extended to evaluate de-jitter buffer performance in a black-box approach. The main problem is to define appropriate profiles and associated QoE metrics. We do not address here QoE metrics in details, however one may measure receiving delay and quality (P.863) with some caution to ensure convergence of de-jitter buffers, and evaluate parameters from ITU-T G.1020 [5] and G.1021 [7].

The VoLTE delay tests specified in 3GPP have not been designed to evaluate the performance of de-jitter buffers; they only verify the basic capability of mobile phones to adapt delay according to network conditions. The three delay/loss profiles were generated using a MATLAB simulation in [17, Annex E] with strong simplifications: eNodeB scheduling is perfectly periodic, random block errors on the LTE radio interface are independent for each speech frame, EPC jitter is modeled with a uniform distribution in an interval of $(27, 33)$ ms or $(24, 36)$ ms; the handling at different protocol layers (PDCP/RLC/MAC/PHY) is not taken into account, optimizations like TTI bundling and intra-LTE handovers are not modelled. Due to the simplified EPC jitter model, the ratio of out-of-sequence packets is quite high in DRX 40 ms profiles, which is typically not observed in practice. Note that packet delay variations in the three profiles of [17] are stationary and well-bounded by design, to be able to define the associated jitter buffer depth and terminal delay target with no ambiguity.

A simple method to obtain profiles would be to capture RTP packets from real VoLTE calls, and to convert them into delay/loss traces. This method has two drawbacks. First, it depends on a specific VoLTE network, recalling that eNodeB scheduling algorithms are proprietary and LTE/EPC/IMS network settings are specific to each mobile operator (e.g. radio signal levels to trigger handovers or activation of TTI bundling). Second, when Discontinuous Transmission (DTX) [1] is used, RTP streams depend on the speech signal used in the uplink and delay/loss profiles would be tied to a specific input. To avoid these issues, we propose to modify the generic simulation model from [17, Annex E] to obtain packet traces that are more representative of real VoLTE networks.

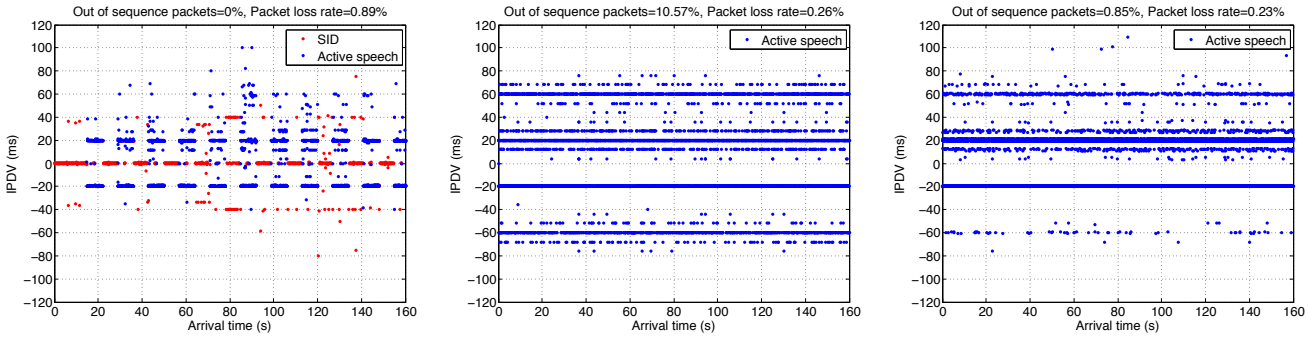
Figure 4 (a) shows an example of instantaneous inter-packet delay variation (IPDV) metrics for a commercial VoLTE network using DRX 40 ms and semi-persistent scheduling (SPS). The instantaneous IPDV is defined as: $IPDV(i) = D(i) - D(i-1)$, where $D(i)$ denotes the one-way delay of the ith packet. Measurements were made with one static mobile phone and another mobile phone in a car during a drive test. Fig. 4 (a) shows an excerpt (160 seconds) captured in the downlink of the mobile phone in the car which experienced several (local) handovers while the other mobile phone had good conditions. The call used the AMR-WB speech codec with DTX on, silence periods were coded by SID (Silence Insertion Description) frames sent every 160 ms on average. Different colors (blue and red, respectively) are used for IPDV in active speech and SID frames. The packet loss rate was around 0.89% and there was no out-sequence packet. One can observe that IPDV is mainly around 0 for SID frames and $\pm$ 20ms for active speech, with other IPDV values at relative offsets of 8 or 16 ms due to HARQ retransmissions and multiples of 40 ms due to missed DRX cycles.

For comparison purposes, IPDV for condition 2 (DRX 40 ms) from simulated profiles is shown in Fig. 4 (b). DTX is assumed deactivated, and all frames are considered as active speech; this has the advantage that Fig. 4 (b) is independent from any specific speech database. Beside DTX, a key difference between Figures 4 (a) and (b) lies in the amount of out-of-sequence packets (10.575 % for condition 2) and packets that missed their normal DRX cycle; the target BLER of 10% in conditon 2 results in quite many HARQ retransmissions.

To better match the example from Figure 4 (a), one can modify the MATLAB routine `VoLTEDelayProfile_vPHY` from [17, Annex E], as follows:

- eNodeB scheduling can be made less periodic, by adding a random jitter of $\{0, 1\}$ ms to scheduling times.
- The uniform distribution for network delay (i.e., delay between two eNodes) can be replaced by a long-tailed distribution, such as a Weibull mixture model with 2 components, keeping the same minimum network delay as in [17, Annex E]. Note that network delay reflects here EPC delay as well as processing and buffering delays.
- In DRX, the scheduling time can be randomly increased by the cycle length (20 or 40 ms), e.g. with a probability of 1%, to simulate missed cycles due to scheduling grants that could not be properly decoded.

It can be verified from Figure 4 (c) that, with the modifications listed above, the resulting IPDV better reflects the real example in Figure 4 (a). The packet loss rate of 0.22 % is close to that of the original condition 2, because the HARQ simulation from [17, Annex E] is not changed. Note that further enhancements to the delay/loss profile generation would be required, for instance to reflect that EPC delay is typically more correlated for packets transmitted in the same DRX cycle. Moreover, it would be interesting to better represent cell edge cases, by simulating handovers or the use of TTI bundling at the IP level.



(a) Real packet trace (DTX on)

(b) Simulated profile (DTX off) – Condition 2 from [17]

(c) Simulated profile (DTX off) – Enhanced condition 2

Fig. 4. Comparison of inter-packet delay variation (IPDV) in real and simulated conditions.

VII. Conclusion

In this paper, we analyzed the existing delay-related metrics for VoLTE terminals. Minor shortcomings have been highlighted in the 3GPP test methods, in particular the measurement variability is not fully captured with a limited number of trials, and network impairments with three simulated profiles do not capture the behavior of de-jitter buffers in real VoLTE conditions. We also proposed improvements to the VoLTE packet delay/loss simulation model used in 3GPP.

Future work will focus on improving VoLTE delay/loss models, and extending the analysis to Voice over Wifi (VoWifi) and other applications (e.g., WebRTC). Note that terminal-side delay and quality metrics combined with delay/loss packet traces can be used to predict voice quality (MOS) with, for instance, the E-model [29], which is a parametric model to predict MOS based on several factors. The E-model could be extended to predict the effect of de-jitter buffers in various network conditions.

Acknowledgments

The authors would like thank G. Le Tourneur, A. Curti, P. Le Tort, J.-P. Thomas, A. Naglé, M. Stenvot F. Payoux and C. Simon (from Orange) for their help, F. Plante (from Intel) for valuable discussions on test results, and Anders Ericsson (Ericsson) for helpful discussions.

References

[1] S. Chakraborty, T. Frankkila, J. Peisa, and P. Synnergren, *IMS Multimedia Telephony over Cellular Systems*. John Wiley & Sons, 2007.

[2] "Qualinet White Paper on Definitions of Quality of Experience," P. Le Callet, S. Möller and A. Perkis, eds, Version 1.2, March 2013.

[3] S. Jelassi, G. Rubino, H. Melvin, H. Youssef, and G. Pujolle, "Quality of experience of VoIP service: a survey of assessment approaches and open issues," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 2, pp. 491–513, 2012.

[4] R. Sanchez-Iborra, M. Cano, and J. Garcia-Haro, "Revisiting VoIP QoE assessment methods: are they suitable for VoLTE?" *Network Protocols and Algorithms, Macrothink Institute*, vol. 8, no. 2, pp. 40–57, 2016.

[5] ITU-T Rec. G.1020, "Performance parameter definitions for quality of speech and other voiceband applications utilizing IP networks," July 2006.

[6] H. Melvin and L. Murphy, "An integrated NTP-RTCP solution to audio skew detection and compensation for VoIP applications," in *Proc. ICME*, vol. 2, July 2003, pp. 537–541.

[7] ITU-T Rec. G.1021, "Buffer models for development of client performance metrics," July 2014.

[8] ETSI TS 202 739, "Transmission requirements for wideband VoIP terminals (handset and headset) from a QoS perspective as perceived by the user, V1.4.1," March 2015.

[9] C. Wu, K. Chen, Y. Chang, and C. Lei, "Evaluation of VoIP Playout Buffer Dimensioning in Skype, Google Talk, and MSN Messenger," in *Proc. ACM NOSSDAV*, 2009, pp. 97–102.

[10] P. Gournay and K. Anderson, "Performance Analysis of a Decoder-Based Time Scaling Algorithm for Variable Jitter Buffering of Speech Over Packet Networks," in *Proc. ICASSP*, May 2006.

[11] P. Pocta, H. Melvin, and A. Hines, "An Analysis of the Impact of Playout Delay Adjustments introduced by VoIP Jitter Buffers on Listening Speech Quality," *Acta Acustica United with Acustica*, vol. 101, no. 3, pp. 616–631, 2015.

[12] J. Bolot, "Characterizing End-to-End Packet Delay and Loss in the Internet," *Journal of High Speed Networks*, vol. 2, pp. 305–323, 1993.

[13] R. G. Cole and J. H. Rosenbluth, "Voice over IP Performance Monitoring," in *Proc. SIGCOMM*, vol. 31, no. 2, 2001, pp. 9–24.

[14] W. Jiang, K. Koguchi, and H. Schulzrinne, "QoS evaluation of VoIP end-points," in *Proc. ICC*, vol. 3, May 2003, pp. 1917–1921.

[15] 3GPP TS 51.010-1, "Mobile Station (MS) conformance specification; Part 1: Conformance specification."

[16] 3GPP TS 26.131, "Terminal acoustic characteristics for telephony; Requirements."

[17] 3GPP TS 26.132, "Speech and video telephony terminal acoustic test specification."

[18] 3GPP Tdoc S4-160457, "On the influence of DTX on UE LTE delay tests with packet delay and loss profiles," Source: ORANGE.

[19] ITU-T Rec. P.58, "Head and torso simulator for telephonometry," Nov. 2013.

[20] ITU-T Rec. P.57, "Artificial ears," Oct. 2012.

[21] V. Paxson, "Measurements and Analysis of End-to-End Internet Dynamics," Ph.D. dissertation, University of California at Berkeley, April 1997.

[22] 3GPP Tdoc S4-AHQ082, "Experimental results and proposals on clock drift measurement," Source: ORANGE.

[23] ITU-T Rec. P.501, "Test signals for use in telephonometry," June 2007.

[24] 3GPP Tdoc S4-140079, "Method for determining one way delays of LTE radio network simulators," Source: HEAD acoustics.

[25] 3GPP TS 26.114, "IP Multimedia Subsystem (IMS); Multimedia telephony; Media handling and interaction."

[26] ITU-T Rec. P.863, "Perceptual objective listening quality assessment," March 2016.

[27] 3GPP Tdoc S4-AHQ077, "Delay profiles for ART-LTE-UED," Source: Qualcomm.

[28] C. Bontu and E. Illidge, "DRX Mechanism for Power Saving in LTE," *IEEE Communications Magazine*, vol. 47, no. 6, pp. 48–55, 2009.

[29] A. Raake, *Speech Quality of VoIP : Assessment and Prediction*. John Wiley & Sons, 2006.