Measuring and improving the quality of experience of mobile voice over IP

Najmeddine Majed

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

THESE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPERIEURE MINES-TELECOM ATLANTIQUE
BRETAGNE PAYS DE LA LOIRE - IMT ATLANTIQUE
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication

Spécialité : *Informatique*

Par

Najmeddine MAJED

« Measuring and Improving the Quality of Experience of Mobile Voice over IP»

Thèse présentée et soutenue à « IMT-Atlantique », le « 03/10/2018
Unité de recherche : IRISA
Thèse N° : 2018IMTA0099

Rapporteurs avant soutenance :

Philippe MARTINS	Professeur, Télécom ParisTech
Toufik AHMED	Professeur, Bordeaux-INP ENSEIRB-MATMECA, LaBRI

Composition du Jury :

Président :	Ken CHEN	Professeur, Université Paris XIII, L2TI
Examinateurs :	Véroniques VEQUE	Professeur, Université Paris-Sud
	Philippe MARTINS	Professeur, Télécom ParisTech
	Stéphane RAGOT	Ingénieur de recherche, Orange Labs - Lannion
	Toufik AHMED	Professeur, Bordeaux-INP - Enseirb-Matmeca, LaBRI
	Alberto BLANC	Maître de conférences, IMT Atlantique
Directeur de thèse :	Xavier LAGRANGE	Professeur, IMT Atlantique

Abstract: Fourth-generation mobile networks, based on the Long Term Evolution (LTE) standard, are all- IP networks. Thus, mobile telephony providers are facing new types of quality degradations related to the voice packet transport over IP network such as delay, jitter and packet loss. These factors can heavily degrade voice communications quality. The real-time constraint of such services makes them highly sensitive to delay and loss. Network providers have implemented several network optimizations for voice transport to enhance perceived quality. However, the proprietary quality management algorithms implemented in terminals are left unspecified in the standards. In this context, we are interested in media adaptation mechanisms integrated in terminals to enhance the overall Quality of Experience (QoE). In particular, we experimentally evaluate Voice over LTE (VoLTE) QoE metrics such as delay and Mean Opinion Score (MOS) sing a standardized test method. We propose some enhancements to the actual test method and discuss how this method can be extended to evaluate de-jitter buffer performance. We also experimentally evaluate WebRTC voice quality in different radio conditions using a real LTE test network. We evaluate the impact of jitter buffer and bit rate variations on the measured quality. To enhance voice codec robustness against packet loss, we propose a simple application-layer redundancy. We implemented it for the Enhanced Voice Service (EVS) codec and evaluate it. Finally, we propose a signaling protocol that allows sending redundancy requests during a call to dynamically activate or deactivate the redundancy mechanism.

Résumé: Les réseaux mobiles 4G basés sur la norme LTE (Long Term Evolution), sont des réseaux tout IP. Les différents problèmes de transport IP comme le retard, la gigue et la perte des paquets peuvent fortement dégrader la qualité des communications temps réel telles que la téléphonie. Les opérateurs ont mis en œuvre des mécanismes d'optimisation du transport de la voix dans le réseau afin d'améliorer la qualité perçue. Cependant, les algorithmes propriétaires de gestion de la qualité dans les terminaux ne sont pas spécifiés dans les standards. Dans ce contexte, nous nous intéressons aux mécanismes d'adaptation de média, intégrés dans les terminaux afin d'améliorer la qualité d'expérience (QoE). En particulier, nous évaluons de manière expérimentale des métriques QoE de la voix sur LTE (VoLTE) en utilisant une méthode de test standardisée. Nous proposons d'améliorer la méthode de test et discutons la manière dont cette méthode peut être étendue pour évaluer les performances du buffer de gigue. Nous évaluons également de manière expérimentale la qualité de WebRTC dans différentes conditions radios en utilisant un réseau réel. Nous évaluons l'impact du buffer de gigue et de la variation du débit sur la qualité mesurée. Pour améliorer la robustesse des codecs contre la perte de paquets, nous proposons d'utiliser une redondance simple au niveau applicatif. Nous implémentons cette redondance pour le codec EVS (Enhanced Voice Service) et nous évaluons ses performances. Enfin, nous proposons un protocole de signalisation qui permet d'envoyer des requêtes de redondance au cours d'une communication afin d'activer ou désactiver celle-ci dynamiquement.

Acknowledgment:

At the end of this thesis, I would like to thank all the people without whom this project would never have been possible. Many people have contributed to this research in their own particular way and for that I want to give them special thanks.

I would like to thank may supervisors, Pr. Xavier LAGRANGE, Dr. Alberto BLANC and Dr. Stéphane RAGOT for their precious time, effort, advises and encouragements in moments of doubts.

I would like to thank all my colleagues in Orange Labs for their support and help in my research.

I would like also to thank my jury members, Pr. Véronique Vèque, Pr. Philippe Martins, Pr. Toufik Ahmed and Pr. Ken Chen for making my thesis defense an enjoyable moment.

Finally, I would like to thank my family and friends for believing in me and supporting me during all these years.

Contents

List of Figures

List of Tables

List of Abbreviation

3GPP	Third Generation Project Partnership
ACK	ACKnowledgement
ACR	Absolute Category Rating
AM	Acknowledged Mode
AMR	Adaptive Multi Rate
ARQ	Automated Repeated Request
AVP	Audio Video Profile
AVPF	Audio Video Profile with minimum Feedback
BEI	Bandwidth Estimation Index
CAM	Channel-Aware Mode
CMR	Codec Mode Request
CSRC	Contributing Sources
DCR	Degradation Category Rating
DMOS	Degradation Mean Opinion Score
DRP	Drum Reference Point
DRX	Discontinuous Reception
DTX	Discontinuous Transmission
eNB	evolved Node B
EPC	Evolved Packet Core
EPS	Evolved Packet System
EVS	Enhanced Voice Services
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
FEC	Forward Error Correction
FB	Fullband
GBR	Garanteed Bit Rate
HARQ	Hybrid Automatic Repeat Request
HSS	Home Subscriber Service
IAT	Inter Arrival Time
IETF	Internet Engineering Task Force
IMS	IP Multimedia Subsystem
IP	Internet Protocol
IPDV	Inter Packet Delay Variation

iSAC	internet Speech Audio Codec
JBM	Jitter Buffer Management
JSEP	JavaScript Session Establishment Protocol
LQO	Listening Quality Objective
LPCM	Linear Pulse Code Modulation
LTE	Long-Term Evolution
M	Marker bit
MDC	Multiple Description Coding
MME	Mobility Management Entity
MOS	Mean Opinion Score
MRP	Mouth Reference Point
MTSI	Telephony Service over IMS
NB	Narrowband
NAT	Network Adress Translation
NACK	Non-ACKnowledgement
OTT	Over The Top
PCRF	Policy Charging and Rules Function
PDCP	Packet Data Convergence Protocol
PESQ	Perceptual Evaluation of Subjective Quality
PGW	PDN Gateway
PLC	packet-Loss Concealment
PLR	Packet Loss Rate
POLQA	Perceptual Objective Listening Quality Assessment
PPM	Part Per Million
PSQM	Perceptual Speech Quality Measure
QCI	QoS Class Identifier
QoS	Quality of service
QoE	Quality of experience
ROHC	Robust Header Compression
RR	Receiver Reports
RTP	Real time transport Protocol
RTCP	Real Time Control Protocol
RTT	round-trip times
RX	Receiver side
RLC	Radio Link Control
RSRP	Reference Signal Received Power
SAVPF	Secure AVPF
SCR	Source Controlled Rate operation
SD	Speech Data

SDP	Session Description Protocol
SID	Silence Insertion Descriptor
SIP	Session Initiation
SGW	Serving Gateway
SN	Sequence Number
SR	Sender Reports
SRTP	Secure Real Time Protocol
SSRC	Synchronization Source Identifier
SWB	Super-Wideband
TCP	Transport Control Protocol
TS	Timestamp
TTI	Transmission Time Interval
TMMBN	Temporary Maximum Media Stream Bit Rate Notification
TMMBR	Temporary Maximum Media Stream Bit Rate Request
ToC	Table of Contents
TURN	Traversal Using Relays around NAT
TX	Transmitter Side
UDP	User Datagram Protocol
UE	User Equipment
UM	Unacknowledged Mode
VAD	Voice Activity Detection
VBR	Variable Bit Rate
ViLTE	Video over LTE
VoWiFi	Voice over Wi-Fi
VoLTE	Voice over LTE
VoIP	Voice over IP
VxIMS	Voice over IMS
WB	Wideband
WebRTC	Web Real-Time Communication
WB-PESQ	Wideband Perceptual Evaluation of Speech Quality
WMM	Wireless Multimedia
W3C	World Wide Web Consortium
XR	Extended Report

Chapter 1

Introduction

Contents

1.1 General context and problem statement

The concept of telephony is very old. The first communication that transmitted speech between two persons was in 1873 thanks to the invention usually attributed to Alexander Graham Bell. The telephony service was only offered through wired networks until the commercialization of the first mobile network by AT&T in 1949. After that, the mobile network has been evolving in a continuous way and transmitting an important part of the overall communications. In France, the Regulatory Authority for Electronic Communications and Posts (ARCEP) declares that the number of fixed telephone service subscribers is around 39 millions in 2017 with no important increase from several years, [8]. However, the number of mobile telephone subscribers is around 90 millions. This number is increasing since the deployment of fourth generation network (4G) with an overall mobile calls volume over 40 Billion minutes in 2017 [8]. In the United States, the increase of mobile telephone users is even faster than in

France. On one hand, 90 percent of houses had a fixed telephone service in 2004 against less than 50 percent in 2017 [9]. On the other hand, the number of houses with only wireless connections is constantly increasing to reach the 50 percent in 2017 [9]. These statistics show that an important part of communications are transmitted over mobile networks. Thus, it is important for network operators to provide mobile telephony service with the best possible quality.

4G networks are all IP networks. Voice is transmitted through packet switching operations in contrast to the previous network generations. In 2G and 3G networks, circuit-switching operations are used for voice transport. In circuit-switched networks, resources with dedicated links (circuits) are reserved to transmit the voice. In packet-switched networks, data and voice can share resources. They are both transported within IP packets.

Packet switching provides more efficient bandwidth sharing than circuit switching. It also reduces the network deployment cost by unifying voice and data networks in a single infrastructure. However, in IP networks, packets may have several problems during transmission including delay, delay variation known as jitter, out-of order transmission, packet loss, etc. These imperfections depend on network conditions. The mobile network performance is variable due to multiple factors. The radio propagation channel is prone to fluctuations due to multiple reflections or diffractions due to mobile and fixed obstacles.

Telephony is a real-time service with high sensitivity to delay, jitter and packet loss. In real-time communications, packets must be received at a regular frequency to play-out voice without interruption. During conversations, important delays can limit interaction between the two parties of the call. The ITU-T has defined thresholds for the end-to-end delay to consider that the service has an acceptable quality. Few losses can be tolerated during a communication. A high packet loss rate heavily degrades the communication quality.

Packet switching is used in Voice over LTE (VoLTE). VoLTE is a form of Voice over IP (VoIP) with specific treatment of voice packets in the network. VoLTE uses also some specific network Quality of service (QoS) optimizations for voice transport to enhance the received quality. VoLTE is a version of the classic telephony service, offered only by network providers. It is also possible to extend VoLTE to offer enriched communication services like Video over LTE

(ViLTE).

The deployment of 4G networks has enabled broadband mobile data usage with significantly better performance (throughput, latency ...) than in the previous generations. 2G networks have a slow speed with a throughput around several dozens of Kbps. 3G networks have a higher speed than 2G networks, with a throughput around several Mbps. In 4G networks, the throughput becomes very high and reaches several dozens of Mbps. This improved performance has enabled mobile VoIP without specific treatment of voice packets. This helped Over The Top (OTT) applications, such as *whatsapp, skype, facebook messenger...*, to evolve in best effort networks and to compete with the classical telephony services.

To develop their service offering, network operators can also propose their own type of OTT services in addition to the VoLTE. They can use software telephony applications or WebRTC soft-phones that operate in best-effort conditions just like other OTT applications. In this case, one cannot assume that QoS guarantees are provided by the network. It must be noted that this type of services is not a complete telephony service, which has to include emergency calls and some legal obligations. However, it is important that softphones developed by network operators have a perceived quality at least as good as competing OTT services.

For both types of services, VoLTE and OTT services, network conditions can heavily impact perceived quality. Mobile terminals must be able to adapt to network variable conditions to compensate for transmission delay variations, packets losses, varying bandwidth, etc. Mobile network operators can use network optimizations through QoS guarantees to enhance their services quality. However, they often let terminal vendors implement their own adaptation algorithms and mechanisms. In this context, our research deals with media adaptation algorithms implemented in terminals for mobile VoIP quality enhancement.

1.2 Main research objectives

VoIP is based on Real time transport Protocol (RTP), which is a generic protocol that can be used for any type of real-time services. This allows the application

to adapt the transport constraints to its needs (for example: packet size, number of frames per packet, use of redundancy). Thus, the application must have the intelligence and flexibility to respond to difficult network conditions. According to the end-to-end principle, the intelligence is located in end points to adapt to the network variable state.

Following these principles, the focus of this work is specifically on media adaptation mechanisms in endpoints to enhance the perceived quality for the two-way communications, based on network conditions. The following adaptation dimensions are addressed in this work.

The first mechanism is about adaptation to delay variation caused by network jitter and clock skew. Network jitter is usually compensated by using a jitter buffer. It can be implemented jointly with time scaling operations to change the duration of active or inactive audio segments. Jitter buffers can be also used to handle clock skew and sound-card jitter. The second mechanism concerns adaptation to packet losses. Many mechanisms have been developed within the encoder and decoder to enhance voice codecs resiliency against packet loss. For example, redundant information or packet loss concealment techniques can be applied to minimize the impact of packet loss on perceived quality. The third mechanism is bit rate adaptation, which can be used to adapt to varying channel conditions

The above dimensions are not necessarily independent. For instance, it is possible to combine jitter buffer operations and packet loss concealment, e.g. to resynchronize decoding when a previously declared lost packet arrives late or take advantage of partial or full redundant information available in different packets in the jitter buffer. Moreover, the jitter buffer may induce some additional packet losses. The adaptive use of redundant information is also a form of bit rate adaptation.

One aspect considered in this work is to adapt jointly the sender and the receiver. Such joint optimization requires feedback which can be provided either in-band in the RTP flow or out-of-band in Real Time Control Protocol (RTCP), in the form of statistics or requests. It is important to optimize the end-to-end (overall sender+network+receiver) system by exploiting feedback.

1.3 Approaches followed in this work

This work was done in a network operator laboratory (Orange Labs), which contributes actively to standardization. After the deployment of VoLTE service, it was found important to define requirements on quality mechanisms embedded in VoLTE terminals and enhance the underlying service performance. This research was motivated by different 3GPP work items to which we actively contributed.

The first work item is called ART_LTE-UED (in 3GPP Release 12) which stands for "acoustic requirements and test methods for IMS-based conversational speech services over LTE-UE delay aspects". The objective of this work item is to define a terminal delay test methods in the VoLTE context and to specify some requirements on terminal delay.

The second work item is called EXT_UED (in 3GPP Release 13) which stands for extension of UE delay test methods. The main objectives of this work item are to:

- add the support of VoWiFi delay testing to the acoustic delay test method defined in the previous work item;
- define some requirements on terminal clock accuracy; 3) extend the test method to characterize the behavior of terminal jitter buffer under realistic LTE radio conditions.

The work reported in Chapter 3 was mainly motivated by these first two work items.

The third work item is a study phase (in 3GPP Release 15) called eVoLP which stands for "enhanced VoLTE Performance", whose objectives are:

- Evaluation of the impact of proprietary terminals implementations such as packet loss concealments and jitter buffer mechanisms;
- Guidelines or requirements to ensure that VoLTE terminals can adapt to the most robust codec modes;
- Specifications of mechanisms to allow terminals to send adaptation requests to the most robust mode.

The work reported in Chapter 5 was mainly motivated by the eVoLP work item.

We followed two different approaches inspired from [10]. The first approach is called black box approach. It consists of evaluating adaptation algorithms without entering inside the algorithm. It is adapted to the evaluation of commercial proprietary terminals, without precise knowledge of the used adaptation algorithm. This includes the use of real test networks with real radio links for data collection and analysis. To be accepted as valid results in the 3GPP, complex tools are used for quality measurement by following only the standardized test methods. This required a lot of time to conduct the measurements according to standard requirements.

The second approach is called glass box approach. It consists of developing and implementing media adaptation algorithms. It is adapted to the case of a softphone where implementation details are available. We make algorithmic improvements and validate them in experiments. Subjective tests are conducted to confirm the effect of algorithmic improvements. We explore here what improvements can be made to the 3GPP Enhanced Voice Services (EVS) codec for VoLTE service.

1.4 Main contributions

The main contributions of this work are as follows:

- We measure some VoLTE quality metrics using a test method described in the 3GPP. We detail and describe a sophisticated method for end-to-end delay measurement in VoLTE context. We focus on VoLTE metrics related to the terminal caused delay. 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.

- We experimentally evaluate WebRTC voice quality over different network conditions including radio coverage, bandwidth limitation and packet loss. We evaluate the impact of the jitter buffer on the different measured quality metrics. We study the impact of varying bit rate on the perceived quality in different network conditions.

- We implement a redundancy mechanism in the EVS codec which we call application layer redundancy. We confirm its impact on received voice quality in lossy channel through objective and subjective evaluation tests.

- We propose a signaling method to request application layer redundancy during a VoLTE call. Our proposition concerns an in band feedback mechanism based on the unused Codec Mode Request (CMR) codes with the EVS codec.

1.5 Dissertation outline

In Chapter 2, we briefly review the main elements of a VoIP system. We also review VoLTE which is the main VoIP service offered by a operators. Then, we present the latest definitions of Quality of experience (QoE) of a VoIP service and the main measurement and evaluation tools.

In Chapter 3, we evaluate metrics to characterize trade-offs between delay and quality of VoLTE mobile phones in various network jitter and loss conditions. We describe the test method usually used for telephony characterization in the 3GPP. We discuss how the underlying test methods allow to evaluate jitter buffer performance implemented in the terminals.

In Chapter 4, we focus on WebRTC. We evaluate the performance of an OTT voice application based on WebRTC in different network conditions (radio coverage, packet loss, bandwidth limitations...) and access (LTE or Ethernet). We evaluate bit rate adaptation impact on quality.

In Chapter 5, we propose an adaptation mechanism based on redundancy for EVS codec, called application-layer redundancy. We evaluate its performance in degraded channel by subjective and objective quality tests. We also propose a signaling method that allows to request application-layer redundancy during a VoLTE call when network condition is degraded.

1.6 List of Publications

1.6.1 Conference articles

- N. Majed, S. Ragot, X. Lagrange and A. Blanc "Delay and quality metrics in Voice over LTE (VoLTE) networks: An end-terminal perspective", International Conference on Computing, Networking and Communications: Communications QoS and System Modeling (ICNC), Silicon Valley, USA 2017.

- N. Majed, S. Ragot, X. Lagrange, A. Blanc, J. Dufour and G. Grao "Experimental evaluation of WebRTC voice quality in LTE coverage tests", 9th International Conference on Quality of Multimedia Experience (Qomex), Erfurt, Germany2017.

- N. Majed, S. Ragot, G. Laetitia, X. Lagrange, A. Blanc "Application-Layer Redundancy for the EVS Codec", 26th European Signal Processing Conference (Eusipco), Rome, Italy 2018.

1.6.2 Contributions to patent applications

- S. Ragot, J. Dufour and N. Majed "Signalisation d'une requête d'adaptation d'une session de communication en voix sur IP", FR 1759203, submitted in Oct 2017.

- S. Ragot, J. Dufour, Minh Tri Vo and N. Majed " Adaptation de débit d'une session de communication en vooix sur IP", submitted may 2018.

1.6.3 Contributions submitted to standardization

This work was motivated as explained before by different 3GPP work items. The results of the different experiments that we conducted were used to contribute to the different 3GPP work items presented in Section 1.3.

- S. Ragot and N. Majed, 3GPP Tdoc S4-160457, S4-160459 "On the influence of DTX on UE LTE delay tests with packet delay and loss profiles" 2016, Source: Orange.

- S. Ragot and N. Majed, 3GPP Tdoc S4-170358, S4-170485 "Extension of UE delay test methods and requirements" 2017, Source: Orange.

- S. Ragot and N. Majed, 3GPP Tdoc S4-170941, S4-171226 "Possible options to signal adaptation requests in VoLTE" 2017, Source: Orange.
- S. Ragot and N. Majed, 3GPP Tdoc S4-180149, "Objective performance results for EVS" 2017, Source: Orange.
- S. Ragot and N. Majed, Tdoc S4-180150, S4-180235 "Subjective test results for EVS with application-layer redundancy" 2018, Source: Orange.

Chapter 2

General Context of Mobile VoIP

Contents

Introduction

VoIP consists of transporting coded voice frames by using real time transport protocol (RTP). RTP is used in mobile telephony as well as in OTT applications. In this chapter, we analyze how media is transported in VoIP services and how feedback is provided. The aim is to identify RTP principles, constraints and limitations for media adaptation. We review VoLTE, which is the main VoIP mobile service offered by network operators. We also review the WebRTC technology, which can be used to develop real-live media applications.

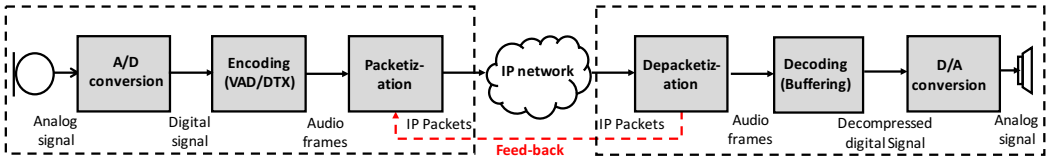


Figure 2.1: VoIP communication system

2.1 Speech packets sending in VoIP

2.1.1 Speech packets generation

Figure 2.1 presents the audio transmission chain in VoIP call. Before being sent in the form of IP packets, the analog audio signal is captured, converted to a digital signal, encoded into a bitstream which is packed in IP packets.

The analog/digital conversion is done in two functional steps:

- The first step is sampling. The continuous-time audio wave is sampled to a sequence of samples according to a given rate to have a discrete-time

Bandwidth	Typical frequency range (Hz)	Typical sampling rate (kHz)
Narrowband (NB)	300-3400	8
Wideband (WB)	50-7000	16
Super-Wideband (SWB)	50-14000	32
Full band (FB)	20-20000	48

Table 2.1: Audio frequency range (ITU-T G.100)

signal. The sampling rate depends on the target band. It should be at least twice as the target band according to the Shannon theorem. We present in table 2.1, the typical audio bandwidth used by various voice codecs and corresponding sampling rate.

- The second step is quantization. The quantization process consists of mapping the magnitude of samples to a discrete subset. This operation is done by a simple mapping between the set of samples magnitude to a smaller set of numbers. The most basic quantizer is Pulse Code Modulation (LPCM), where the amplitude of the analog signal is sampled, and each sample is quantized independently from the other samples, to the nearest value within a range of digital steps. For example, with 16-bit PCM at 8 kHz, the resulting bit stream is at 128 kbit/s.

For telephony service, the amount of data generated after analog/digital conversion is too high. Many efforts have been made to enhance data compression with speech coders to reduce the amount of the transmitted information. In general, speech coders generate compressed audio data in form of bitstream representing a fixed audio frame period at the receiver side. In most cases, the audio frame period called also audio frame length in speech coding is 20 ms. The difference from one voice codec to another is the amount of bits used to code each sample resulting in different possible encoding bit rates.

Multiple speech coders have been proposed, we mention:

- Fixed rate codecs, e.g. ITU-T G.711, G.729, GSM FR, HR, EFR
- Multi-rate codecs, e.g. 3GPP AMR, AMR-WB.

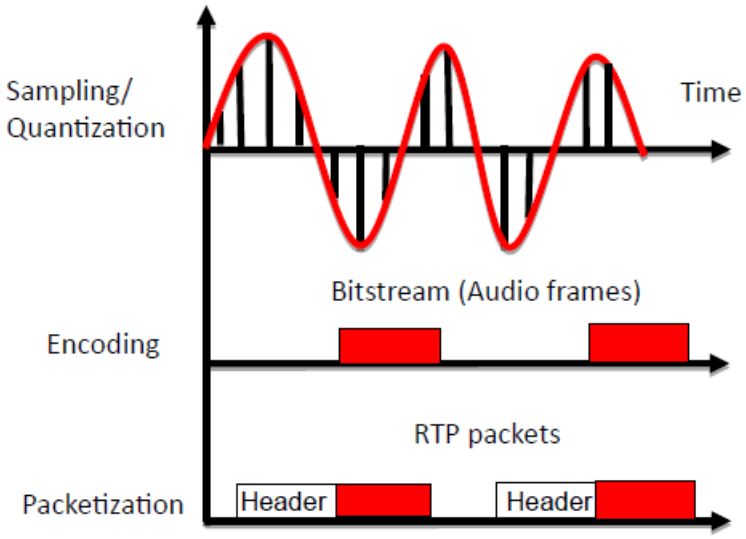


Figure 2.2: Voice packets generation

This category includes scalable (or embedded codecs) such as ITU-T G.729.1 where bit rate can be adapted by truncation of the bitstream.

- Variable bit rate codecs based on different parameters.

  Some variable bit rate codecs are based on the signal classification of the source signal (e.g. voiced, unvoiced frames) targeting a lower average than fixed rate codecs, e.g. EVRC, EVRC-B, EVS 5.9

  Some variable bit rate codecs, e.g. OPUS, have a variable bit rate because they use entropy coding (range coding in OPUS) on coded parameters.

- Multi-mode codecs, e.g. EVS codec.

Multirate/multimode encoding schemes provide flexibility against network variable conditions. Many adaptation algorithms were developed to chose the best encoding scheme according to the network state.

The generated audio frames are encapsulated within RTP/UDP/IP packets. Figure 2.2 presents the different steps to generate voice packets. A packet can contain one or several audio frames. In general, a packet contains a single audio

frame. The packets are sent through the network and received at the other side by a de-packetizer.

2.1.2 Speech codecs used in VoLTE

The standardized speech codecs used for telephony service have been evolving according to the network. In 3G network, 3GPP Adaptive Multi Rate (AMR) and AMR-WB [11, 12], which are Narrowband (NB) and Wideband (WB) speech codecs respectively, have been widely used by telephone operators in their services.

To further enhance voice communication experience and to meet the raising user demand high quality, Fullband (FB) and SWB support have been added to the 3GPP voice codec EVS [13, 14]. Nevertheless, the EVS codec supports all the other voice signal bandwidths, NB and WB. The EVS codec has been standardized by 3GPP in September 2014 to provide new functionalities and improvements for mobile communication which includes:

- Enhanced quality and coding efficiency for NB and WB speech services.
- Enhanced quality by the introduction of SWB and FB speech.
- Enhanced quality for mixed content and music in conversational applications.
- Robustness to packet loss and delay jitter.
- Backward compatibility to the 3GPP AMR-WB codec.

The EVS codec was specified by the 3GPP to be the main speech codec for VoLTE. However, AMR-WB is still widely used in telephony services. Thus, we mainly work with EVS and AMR-WB. We study also some other type of codecs developed mainly for Internet applications. For example, internet Speech Audio Codec (iSAC) [15] and OPUS codec [16].

2.1.3 Discontinuous transmission (DTX)

During a call there are some periods of silence. We do not speak in a continuous way. Some mechanisms have been developed to reduce the amount of voice packets sent during silence periods. This allows to decrease the transmitted

load, increase network capacity, save the terminal battery.

This mechanism of transmission is called DTX. When DTX is activated the encoder sends fewer packets during silence periods than in the active speech periods. A Voice Activity Detection (VAD) mechanism is used within the encoder to detect which parts of the digital signal contain active speech and which parts contain silence. During silence periods detected by the VAD, frames are sent in a larger time interval (typically every 160 ms) which contain only information about background noise. They are called Silence Insertion Descriptor (SID) frames, see Figure 2.3.

For instance, AMR, AMR-WB and EVS codecs support this mechanism. The complete DTX system description and implementation in EVS can be found in [17] and [18].

2.2 RTP packets transport over IP network

2.2.1 Real Time Transport Protocol (RTP)

RTP is a protocol used for media stream transport, including audio and video streams, over IP network. In general, RTP is used for real-time applications. RTP has a companion protocol called RTCP. While RTP transports media stream, RTCP delivers different types of statistics about transmission and quality of service.

RTP runs over User Datagram Protocol (UDP). UDP [19] is a basic packet

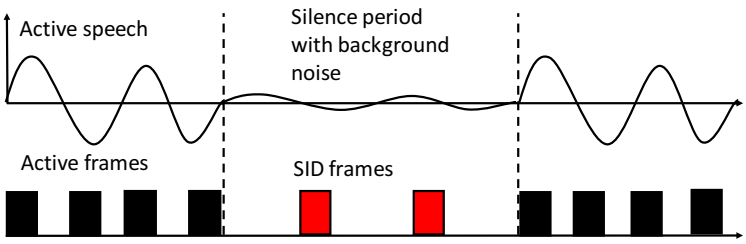


Figure 2.3: Discontinuous transmission illustration.

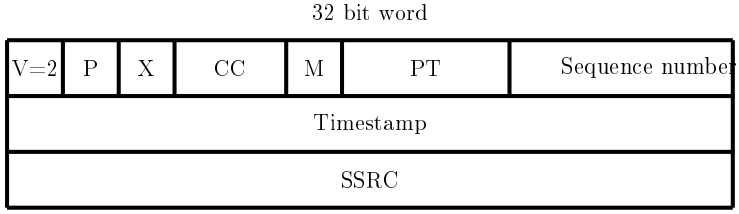


Figure 2.4: RTP header [1].

transport protocol with no retransmission delays which is suitable for applications with real-time constraint. RTP is used with a signaling protocol which establishes connections across the network. A comprehensive review of the signaling plane of VoIP can be found in [20, 2]. A review of media handling (especially voice) in the user plane with RTP can be found in [21, 22].

Each RTP packet consists of a header and media payload. The RTP header, presented in Figure 2.4, includes several fields and the most important ones for media adaptation and proper reproduction of a media stream at the receiver are listed below:

- Timestamp (TS) reflects the sampling instant of the first octet in the RTP data packet. It allows to playout the media at the right timing, i.e. restore the timing at the sender.

- Sequence Number (SN) allows to detect packet losses and handle out-of-order packets by restoring the ordering at the sender.

- Marker bit (M) may be used to indicate the start of a talkspurt and it might be used to reset the jitter buffer at the receiver but in the payload format of some codecs it is just ignored at the receiver.

- The extension bit (X) indicates if the header has an extension with side information.

- Synchronization Source Identifier (SSRC) identifies the synchronization source. This identifier is chosen randomly, with the intent that no two synchronization sources within the same RTP session will have the same SSRC identifier.

RTP [1] is intentionally generic. It provides an incomplete framework for real-time communications. The implementation of RTP is completed with a specific profile and payload format corresponding to the application needs. RTP does not specify any mechanism for QoS [21]. In fact, it does not specify algorithms for media playout and timing regeneration, synchronization between media streams, error concealment and correction, or congestion control. It is up to the application to provide such algorithms using the different information generated by RTP/RTCP. An RTP profile defines only how a class of payloads is carried by RTP. Several RTP profiles have been defined (Audio Video Profile (AVP), Audio Video Profile with minimum Feedback (AVPF), Secure Real Time Protocol (SRTP), Secure AVPF (SAVPF) ...). In this work, we only review AVP and AVPF, given that the "S" prefix stands for secure variants, which is outside media quality issues.

2.2.2 Transport characteristics over IP network

Delay

One important IP network performance factors is transmission delay because of the real time constraint of VoIP services. In our study, we are interested in end-to-end delay, which accounts for terminal delay and network delay. In general, delay includes two parts: deterministic and random delay. Propagation and transmission delays are deterministic if there is no route change because they depend only on the transmission medium nature, length and the link rate. Queuing delay in network routers is random because it depends on the traffic condition.

We take the example of VoIP call between Alice using a fixed terminal and Bob using a VoLTE terminal. Without loss of generality, we consider the link from Alice to Bob. We hereafter make a detailed description of all the components of the delay between Alice's mouth and Bob's ear.

- The first delay τ_1 is the acoustic delay of the voice between Alice's mouth and the microphone of her softphone. This delay is very low and can be neglected.

- The second delay τ_2 is the acoustic processing of the capturing audio signal. This delay depends on terminal performance and can heavily impact the end-to-end delay.

- The third delay τ_3 is due to the codec. The codec generates a voice frame every 20 ms in general. This 20 ms frame is then put in an RTP/UDP/IP packet and sent on the IP fixed network of Alice's provider.
- The third delay τ_4 is due to the routers queuing delay within the IP network. This delay has a great influence on the end-to-end delay and it depends on the state of the network.
- The fifth delay τ_5 is the E-UTRAN delay which includes radio transmission delay and possible retransmissions delay. In fact, the transmission over LTE radio interface is done through blocks of 1 ms which contains a part of an IP packet or a complete one or even several packets. Errors may occur and retransmission is used. Thus, this type of delay depends on the errors that occurs during transmission and can very important in radio cover limit.
- The sixth delay τ_6 is jitter buffer delay. This delay depends on the buffer depth to compensate for delay variation. The jitter buffer delay has an important impact on the end-to-end delay and must be managed in a way to have the best compromise between quality and delay.
- The seventh delay τ_7 is the playout delay including decoder delay and the acoustic processing delay.
- The last delay τ_8 is the acoustic delay of the voice between Alice's mouth and the microphone of her softphone

Network jitter

In VoIP, transmitted packets may take various paths. Network congestion can cause packet loss or packet delay. Therefore, packets can arrive at their destination with different transmission delays. This delay variation is called jitter and must be handled by the receiver.

In real-time applications, continuous packet playout is crucial to have an acceptable quality. In most cases, the speech decoder needs an audio frame every 20 ms so that it can playout speech in a continuous way. A jitter buffer must be implemented within the receiver to compensate for this jitter and provide the decoder with audio frames at a regular time period. In other terms, the jitter buffer receives packets and store them until the decoding time. The jitter buffer depth corresponds to the maximum waiting time for the late packets. The

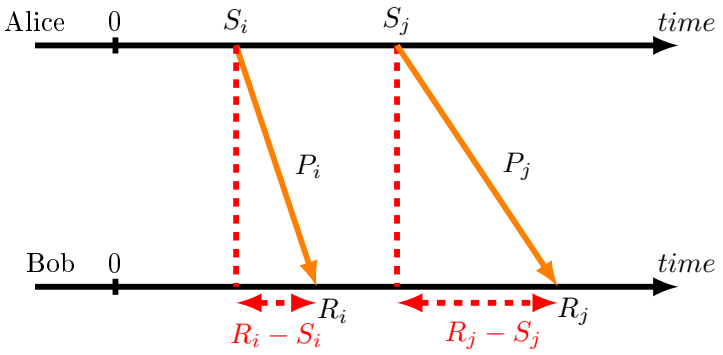


Figure 2.5: VoIP packet transmission jitter presentation

length of a jitter buffer can be fixed or variable according to the implemented Jitter Buffer Management (JBM) algorithm.

Network jitter can be calculated in different ways. We consider a call between Alice and Bob, see Figure 2.5. Two packets P_i and P_j are sent at instant S_i and S_j respectively, and received at instant R_i and R_j respectively. The jitter $D(i,j)$ between packet P_i and P_j can be calculated as the difference of the one way delay.

$$D(i,j) = (R_j - S_j) - (R_i - S_i) = (R_j - R_i) - (S_j - S_i). \tag{2.1}$$

If $j = i+1$, $S_j - S_i$ presents the frame period and $R_j - R_i$ presents the inter-arrival time. Hence, the jitter can be calculated only at th receiver side. The jitter can be smoothed as presented in [1] by a filtering process.

$$J(i) = (1-\alpha)J(i-1) + \alpha|D(i-1,i)|, \tag{2.2}$$

where $\alpha = \frac{1}{16}$.

The jitter reflects the delay variation. It is an important indicator of network condition and an important factor for quality because of the real time constraint of VoIP.

Packet loss

During a VoIP call, packet loss can occur in different cases. In case of network congestion, packets can be lost within routers with high load. Packet loss can also happen because of error transmission in radio coverage limit for example.

Jitter buffer also can be the cause of discarding packets if they are too late. This type of loss is called late loss.

Packet loss impact on speech quality has been deeply studied because of its direct effect on the delivered information. It also is one of the important indicator of network condition.

2.3 Speech packets receiving in VoIP

RTP packets are received in the receiving side by a de-packetizer. The de-packetizer take off the different protocol headers with the different information for playout reconstruction such as send time, arrival time and sequence number. As we have seen in the section 2.2.2, packets may arrive with different transmission delay. Therefore, after being de-packetized, audio frames are stored in a jitter buffer according to their sequence number to compensate for delay variation until decoding.

2.3.1 Clock skew impact on delay

The jitter buffer must compensate for delay variation caused by network state and also from different sender and receiver clocks. The sender produces audio frames with a fixed period controlled by its own audio clock. In the other side, the receiver runs with its own audio clock. In most cases, the two clocks run with different frequencies. We call the difference of clock frequencies clock skew [23, 24]. It can be presented by the linear variation of the one way delay caused by the difference between the sender and the receiver audio clock frequency.

In the following, we study the impact of clock skew on the end-to-end delay. We consider a simple example of sender and receiver as shown in Figure 2.6 running with two different clocks. We consider here that transmission delay is a constant and clock frequencies do not change over time. The delay variability is due only to the possible clock skew. The sender period is ΔTs and the receiver period is ΔTr.

In the presented case, the receiver has a lower clock than the sender. Delay will increase with each transmitted packet. We define a parameter ε as follows:

$$\Delta Tr = (1 + \varepsilon)\Delta Ts \qquad \text{with } \varepsilon > 0 \tag{2.3}$$

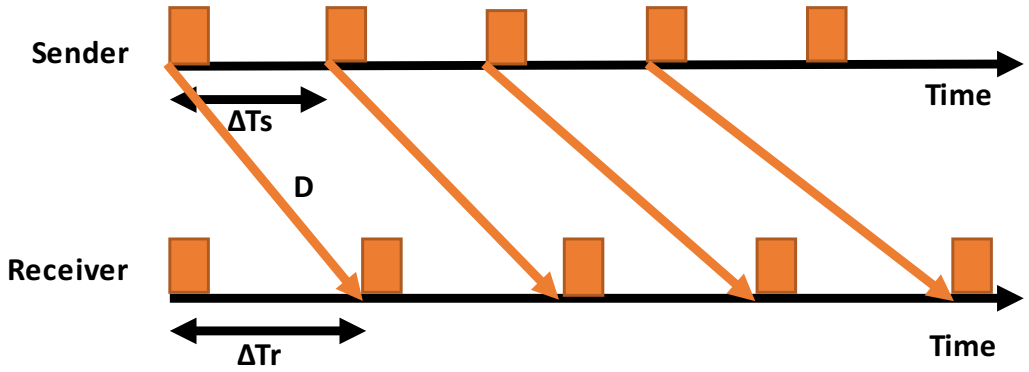


Figure 2.6: Clock skew influence on end-to-end delay

The clock skew μ can be calculated as follows:

$$\mu = \frac{D_{i+1} - D_i}{Ts_{i+1} - Ts_i} \qquad \text{for } i \in \mathbb{N}^*, \tag{2.4}$$

where D_i is the end-to-end delay of packet i which is:

$$D_i = Tr_i - Ts_i \qquad \text{for } i \in \mathbb{N}^*, \tag{2.5}$$

and Ts_i and Tr_i are the sending and receiving time of packet i respectively. Based on equation 2.5 and 2.3, equation 2.4 becomes:

$$\mu = \varepsilon \tag{2.6}$$

Equation 2.6 means that the one way delay increase by ε with each generated packet. Thus, we need N generated packet before having an extra packet caused by clock skew. Thus, we have for a voice frame length equal to ΔTs:

$$N = \frac{\Delta Ts}{\varepsilon} \tag{2.7}$$

We conclude based on equation 2.6 that for a small difference in clock frequency between sender and receiver, the impact on the end-to-end delay can be neglected. However, in cases where the two communicating points operating with clocks with important difference, clock skew can has an impact on the end-to-end delay. The elapsed time to have an extra packet in the jitter buffer can be calculated based on equation 2.7.

2.3.2 Jitter Buffer Management (JBM)

Jitter has a great impact on speech quality in a VoIP call. Jitter buffers are applied to compensate for this delay transmission variability. The coded audio frames must be fed in to the decoder in a regular frequency, see Figure 2.7, to play out voice continuously. JBM design consists in finding an optimal trade-off between quality and end-to-end delay with the constraint that the amount of late losses is kept under an acceptable level.

Figure 2.7: Delay variation compensation by the jitter buffer

Different jitter buffer management strategies have been proposed to handle delay variation [25, 26, 27]. The most simple ones are static jitter buffer with a fixed length. The choice of buffer depth depends on the application type. For those which are more sensitive to packet loss rather than delay, buffer depth should be long enough to wait for late packets. However, real time applications are more delay sensitive and buffer length should be minimized. This strategy is simple but not very efficient because of the instable network condition and delay is almost always variable. A lot of efforts have been made to propose an adaptive jitter buffer strategy that allows an optimal trade-off between quality affected by late losses and end-to-end delay impacted by the buffer depth.

No specific jitter buffer algorithms have been standardized. However, some guidelines and performance requirements have been defined by the 3GPP for terminal vendors, [28]. They can be summarized as follows:

- Limiting the jitter buffering time to provide as low end-to-end delay as possible.
- Limiting the jitter induced concealment operations, i.e. setting limits on the allowed induced losses in the jitter buffer due to late losses, re-bufferings, and buffer overflows.

- Limiting the use of time scaling to adapt the buffering depth in order to avoid introducing time scaling artifacts on the speech media.

In most cases, such as in VoLTE, jitter buffer algorithms are proprietary to terminal vendors. In this dissertation, we use a black box approach to study jitter buffer management in VoLTE. In this case we are only able to propose some requirements to be applied by JBM designers to deliver a minimum acceptable quality.

Jitter buffer adaptation may be combined with other operations to enhance its effectiveness. Time scaling is one of the most mechanisms used within the jitter buffer operations. It consists of modifying the signal by stretching and/or compressing it over the time axis. It gives more flexibility to the jitter buffer to have more time to wait for late packets or inversely if there is to shorten some frames if more frames are available. For instance, time scaling in VoIP and its impact on audio quality has been studied in [29].

For the 3GPP EVS codec, its JBM solution is presented in [30]. It includes jitter estimation, control and buffer adaptation algorithm to manage the inter-arrival jitter of the incoming packet stream. However, in this work EVS JBM will not further studied.

2.4 Feedback mechanisms

In bidirectional communication, keeping the two sides informed about the network state is essential for functional point of view and for service quality optimization. In cases where the sender uses adaptation algorithms to adapt its encoding to the network state, feedback is necessary. Feedback, presented by the red line in Figure 2.1, can contain statistics about network condition such as loss rate, delay and bandwidth. It can also carry adaptation requests. For example, the receiver may request the sender to decrease its encoding rate in case of congestion.

2.4.1 Out-of-band feedback protocol

The main out-of-band feedback protocol is RTCP. It is transported in separate packets other than RTP packets. It has four intended basic functions in RTP AVP [1]: provide quality feedback by sending statistics for flow and congestion control; identify RTCP sources; calculate the number of participants and sender

bit rate; and optionally provide minimal control in a conference. RTCP packets can also be used as a keep-alive mechanism in case of session on hold with no RTP transmission.

Several types of RTCP packets are defined (SR, RR, SDES, BYE, APP) and typically multiple RTCP packets are sent together as a compound RTCP packets. We focus here only on Sender Reports (SR) and Receiver Reports (RR) packets because of their relevance for media adaptation.

SR packets provide quality feedback from active senders. SR packets include data sources, sender statistics (NTP timestamp, RTP timestamp and octet count) and receiver statistics in the form of receiver report block(s). RR packets various receiver statistics: packet loss rate since the previous report sent (fraction lost), long-term packet loss rate (cumulative number of packets lost), smoothed interarrival jitter, time when last SR is received (LSR), delay since last SR (DLSR). Based on sender and receiver reports, QoS parameters at the network level can be computed: delay, jitter and packet loss.

Many constraints are defined in [1] on the RTCP bandwidth (bit rate) and intervals between transmissions of compound RTCP packets. The bandwidth reserved for RTCP packets should be small, around 5% of the session bandwidth is recommended. Transmission interval should not be too small to allow scalability of the number of conference participants, 5 seconds is recommended in [1, Appendix *A*.7]. A randomization of transmission time is specified to avoid unintended synchronization of participants with a maximum waiting time for RTCP reports as defined by a specific scheduling algorithm in [1].

RTCP reports and functions have been extended in [31, 32, 33, 34, 35, 36, 37]. In particular, [31] defines the Extended Report (XR) packet type and signaling methods for RTCP to allow the report of a large number of QoE metrics. It is is used to describe more metrics that are not defined by the RTCP RR and to keep all the participants informed about network performance. The extended RTCP packets includes monitoring information such as: jitter buffer metrics, packet delay variation, delay metrics, burst-gap loss, number of discarded packets. It can also contains Mean Opinion Score (MOS) to evaluate the QoE, [38]. Compared to RTCP RR, the metrics in RTCP XR are closer to perception and actual QoE. RTCP XR defines parametric quality measurement, by combining

different metrics to predict MOS scores, ITU-T P.564. However, the main issue with RTCP XR is that it is not widely supported in terminals. The interval of extended RTCP reports is generally the same as in conventional RTCP, i.e. every 5 s on average, which is sufficient for network supervision but not for real-time adaptation to network conditions.

Real-time applications such as VoIP need more reactive protocols to report network condition changes or to request media adaptation. Application-specific RTCP-APP specified in [28] can be used. However, until the time of writing this dissertation, RTCP-APP is not used for VoLTE. Moreover RTCP-FB messages (TMMBR and TMMBW) are specified only for video. We propose an alternative signaling method in section 5.5 that could be used in VoLTE to request media adaptation in case of network condition change.

2.4.2 In-band feedback mechanisms

Several forms of in-band feedback have been developed and they are specified to a given codec format. We review here the main types of in-band signaling used with different voice codecs and for various purposes.

- Codec Mode Request (CMR): The concept of CMR has been developed for AMR/AMR-WB and then reused for EVS. The payload format of AMR/AMR-WB in [39] includes a fixed field containing a request to codec mode. CMR defines the maximum bit rate that a client wants to receive. It is sent in-band in each RTP packet for AMR and AMR-WB. CMR is useful only if links are bidirectional. The payload format of EVS codec [17, Annex A], includes also an optional CMR field. It can request bit rate, codec mode (EVS Primary or EVS AMR-WB IO), audio bandwidth, or channel-aware mode (CAM). EVS with CMR is standardized for VoLTE in [28].

- Bandwidth estimation: Some codecs estimate the available bandwidth at the receiver and provide the estimation back to the sender. The sender can adjust its bit rate according to the received estimation. The voice codec iSAC includes a field called Bandwidth Estimation Index (BEI) containing an estimation of the available bandwidth as explained.

- Maximum bit rate supported (MBS): The payload of the G.729.1 [40] codec contains a field to tell the other side the maximum bit rate one can receive.

2.4.3 RTP profiles

Two main profiles are defined for RTP. A profile called "RTP Profile for Audio and Video Conferences with Minimal Control" (RTP/AVP) is defined in [41]. This profile lists a set of audio and video codecs and assigns static RTP payload type numbers; payload type values in the range $96-127$ are left to be allocated dynamically. It defines a convention for channel multiplexing in the multichannel audio case. Guidelines for audio are provided on the packetization interval (20 ms recommended), permissible sampling rates, payload structure, and port mapping (RTP port N even and RTCP port $N+1$ recommended). Clarifications are provided, in particular, on the RTCP traffic bandwidth and on the RTP TS. However, one can observe that AVP does not fundamentally fix the limitations of RTP as defined in [1].

An Extended RTP Profile for RTCP-Based Feedback (RTP/AVPF) is defined in [42], as an extension of RTP/AVP. This profile has been developed especially for the video case, to provide more immediate RTCP feedback than RRP/AVP and allow a sender to repair the media stream immediately, by retransmissions Forward Error Correction (FEC). This assumes small round-trip times (RTT) and small groups in the conference. The two main points brought by AVPF are:

- New RTCP messages in the form of RTCP Feedback messages (FB) at the payload, transport and application layer.
- Modified RTCP timing rules to timely report events (e.g. loss or reception of RTP packets) and to use RTCP-FB messages, with the possibility to operate in 3 modes (Regular RTCP mode, Immediate Feedback and Early RTCP mode).

Additional Session Description Protocol (SDP) parameters ("a=rtcp-fb") are also defined to signal AVPF. In general, the use of RTP/AVPF depends on the outcome of the session negotiation. In VoLTE, only RTP/AVP is used for the actual network providers services which limits the possible feedback mechanisms that can be used.

2.5 Overview of VoLTE

In 2008, 3GPP has specified a new radio interface standard to enhance mobile network performance. The new radio interface is called Long-Term Evolution

(LTE) which allows a higher data speed, lower latency and easier to deploy than the previous radio interfaces.

2.5.1 LTE network architecture

The architecture of 4G network presented in Figure 2.8 is based on LTE technology. The network access, named Evolved Universal Terrestrial Radio Access Network (E-UTRAN), is simplified and consists only of evolved Node B (eNB). The eNB is responsible for radio connection between terminals and core network. The core network is called Evolved Packet Core (EPC). EPC includes mainly three entities: Mobility Management Entity (MME), Serving Gateway (SGW) and PDN Gateway (PGW). MME is the network controller that manages terminal access to EPC and user profiles stored in Home Subscriber Service (HSS). SGW and PGW are two gateways that manage data transfer between eNB and EPC and between EPC and IP networks respectively.

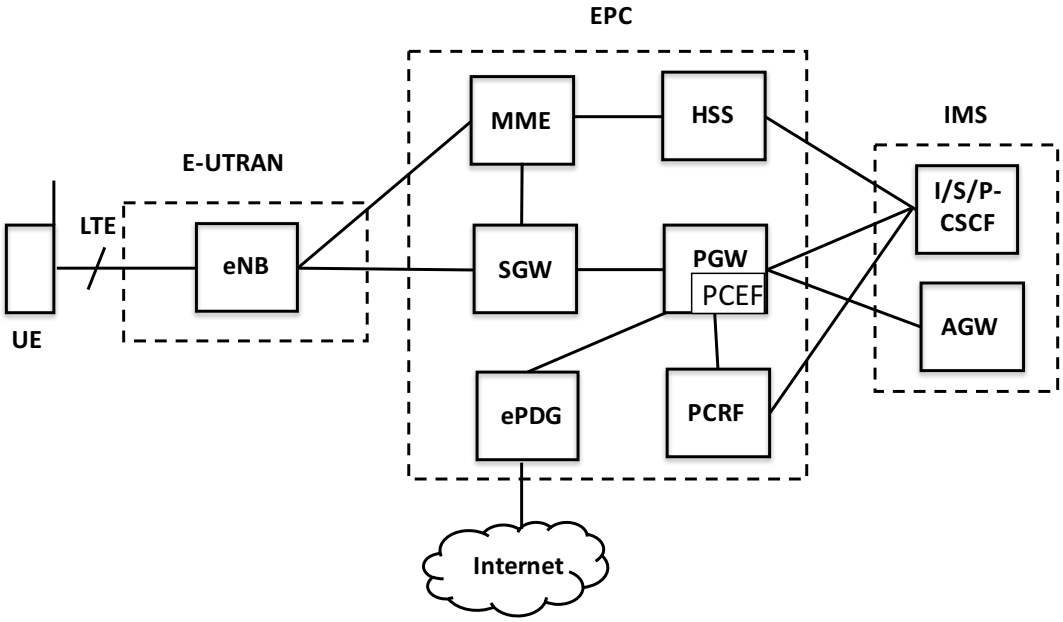


Figure 2.8: 4G network architecture

2.5.2 LTE radio interface characteristics

We present here the main functions and mechanisms in the LTE radio interface that are used for service quality optimization, [43, 2].

- In the MAC layer, a retransmission mechanism called Hybrid Automatic Repeat Request (HARQ) is used to decrease error rate during transmission. HARQ combines two simple methods ARQ and channel coding. The sender sends a block of data and may repeat it in case of NACK with typically 8 ms interval. If the received block has an error, then the receiver buffers the data and requests a re-transmission from the sender. The request is done through ACK/NACK messages. The sender retransmits the same block. When the receiver receives the re-transmitted block, it combines it with buffered block prior to channel decoding and error detection. This helps to enhance the performance of the re-transmissions. HARQ allows a better transmission rate with a low signal to noise ratio than a simple ARQ.

- In the RLC layer, many functions are defined according to the type of service. In section 4.2.1, we will see an example of mechanism used at the RLC layer, which is Acknowledged Mode (AM). It is mainly based on data retransmission in case of packet loss. VoLTE is set to Unacknowledged Mode (UM) to avoid retransmissions delay.

- In the Packet Data Convergence Protocol (PDCP) layer, a compression mechanism called Robust Header Compression (ROHC) is used. It allows to improve the LTE cell capacity by compressing the header of various IP packets. In case of IPv4, the size of IP header goes from 40 bytes in uncompressed version to one or two bytes with ROHC compression.

2.5.3 QoS in EPS

LTE QoS model is defined as part of the policy control and charging architecture in [44]. There is an interface between the application Server (e.g. P-CSCF) and the Policy Charging and Rules Function (PCRF), which manages the QoS policy.

For data transfer that does not have real time constraint, resources with low priority are reserved in form of bearers with non-Garanteed Bit Rate (GBR). The priority of a bearer is defined by its QoS Class Identifier (QCI) value. For example, the default bearer used for signaling is with QCI=5.

For certain delay-sensitive services, such as voice, bearers with high priority are reserved. For VoLTE, dedicated bearers with QCI=1, which is the highest

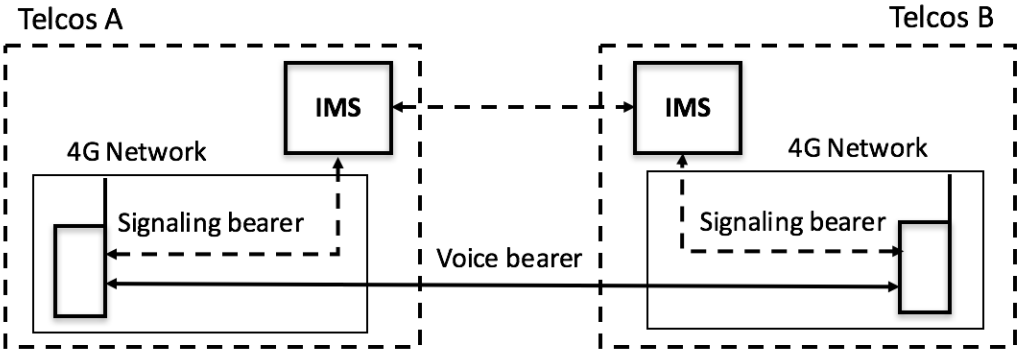


Figure 2.9: VoLTE call [2].

class of service, are reserved. The same dedicated bearer is used in case of multiple concurrent voice sessions (as in call waiting, conference supplementary services, etc...).

2.5.4 VoLTE calls

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 VoIP using RTP for voice transport. It is characterized by a specific network optimizations through IP Multimedia Subsystem (IMS) to provide network QoS [45].

Figure 2.9 shows that the IMS is responsible for Session Initiation (SIP) session negotiation. The complete review of media handling in VxIMS can be found in [28]. An SDP offer is included in a SIP INVITE request which is transmitted to the other side of the communication. If the offered media type is accepted, the other client sends a SIP response message including the SDP answer indicating how to configure the media stream. In IMS there is a constraint that limits the use of a single codec in the SDP answer. Therefore, payload type switching is not possible.

In a VoLTE session, three bearers are used with different QCI degrees 9,

5 and 1. The bearer with QCI 9, called default bearer which is a bearer with low priority, is created whan connecting to the network. For signaling session a bearer with QCI 5 is created as soon as the connection to the IMS APN is established. Once the voice session has been successfully initiated, a dedicated bearer with QCI 1 is created to ensure voice transport.

Most of these mechanisms concerning network optimization are controlled. However, several key elements of VoLTE are left unspecified. They may be proprietary and specific to network equipments (e.g., scheduling in eNodeB) or specific to operator policies (e.g.,radio optimizations). Quality mechanisms that are embedded within the terminals such as de-jitter buffer and potentially media adaptation are proprietary. These aspects can strongly influence performance.

In this dissertation, we study the impact of these unspecified factors on the perceived quality and what type of requirements could be provided by the standard to offer the best quality. In particular, we address the impact of terminal jitter buffer on perceived quality. Furthermore, some media adaptation mechanisms such as redundancy and bit rate adaptation are studied in the VoLTE context.

2.6 Overview of Web Real Time Communication (WebRTC)

The possibilities to make voice calls are no longer limited to telephony services and OTT applications. A technology called Web Real-Time Communication (WebRTC) [46, 47, 48], can also be used to provide conversational services (including telephony) in Web browsers (e.g. *Chrome*, *Firefox*, *Opera*, *Internet Explorer*, *Safari*) [49], non-browser applications (native applications or plugins) or other compatible endpoints. It is specified by the World Wide Web Consortium (W3C) and the Internet Engineering Task Force (IETF). WebRTC is not a service but a media stack with APIs that can be implemented in browsers or in any media capable devices. Media capture, processing (including voice quality enhancement mechanisms such as jitter buffer, noise reduction, echo cancellation), and rendering are integrated in WebRTC.

In particular, W3C has defined several *JavaScript* APIs: *MediaStream* to

handle media (audio/video) capture/rendering, *RTCPeerConnection* which is the most complex API and handles peer network connections, *RTCDataChannel* for data exchange (other than audio/video streams), see [50, 51] for more details. These APIs allow browsers to have real time capabilities. It requires also that the browser can provide the necessary functionalities to exchange audio, video and data through a set of protocols in peer-to-peer mode, including Network Adress Translation (NAT) traversal protocols. The signaling between end points is not fully specified and left to service providers. The only signaling constraint is to rely on JavaScript Session Establishment Protocol (JSEP) [52] which makes use of SDP to exchange media capabilities and other parameters.

WebRTC traffic can be subject to different types of degradation, as it is transported by best effort IP networks. To monitor quality factors including network degradations and media encoding parameters, WebRTC includes a real-time statistics API called *getStat()*. This API allows quality monitoring during audio/video communications. It returns connection statistics: video/audio statistics (bandwidth, input level, packet loss, delay, number of packets and octets sent), connection type (IP address, transport type), network interface informations when it changes from type to another (WiFi to 3G/LTE on a mobile device, or vice-versa).

2.7 VoIP Quality of Experience (QoE)

Understanding QoE meaning allows us to identify the main axes to be improved to enhance the perceived quality. The definition of QoE has been the main objective of many studies, [53, 10, 54]. Its meaning evolves according to existing services. It depends on many factors at different levels including human, system and context level.

2.7.1 QoE definition

QoE has been deeply studied. It presents the major evaluation criterion for any system or service.It is defined in [53, chapter 2] as *the individual stream of perceptions (of feelings, sensory percepts and concepts) that occurs in a particular situation of reference*. QoE may have direct relation ship with feelings and not only with pragmatic concepts. It can be presented as *the judgment of the user based on those feelings and his expectations.*

These definitions are in according with the latest definition of QoE presented in [54]: *QoE is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the users personality and current state.* This definition proposes that QoE is directly related to the user feelings, expectations and personality. This means that to have a proper measurement of QoE, the mentioned features related to user state must be taken into consideration.

To enhance quality, we need in first place to identify QoE factors, which includes human, system and context factors. Human factors can be the demographic and socio-economic background, the physical and mental constitution, or the users emotional state. System factors are more easy to be identified and concern mainly content type, media configuration (encoding, resolution, sampling rate, frame rate, media synchronization...), network performances (bandwidth, delay, jitter, loss, error rate...) and device properties. Context factors are about situational property and user environment.

In practice, non tangible factors are difficult to measure. It is complex to take them into consideration while developing algorithms for QoE management. The more practical way is to identify pragmatic factors or QoS factors that have the major impact on quality. QoS is defined in [55] as the *totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service.*

In general, QoS depends on many factors including quality of the media, cost, usability, security, etc. It can be approached differently depending on the actor in question (user or service provider) and its point of view or role. Mobile VoIP QoS parameters (or technical measures) are defined to characterize several aspects including:

- Data transport (flow rate, loss rate, transmission delay, jitter ...) that can be seen as a dimension of network QoS.
- Access to the service (service availability, call setup time...)
- Mobility management (efficiency of cell change ...)

QoS focuses only on QoE system influencing factors. QoS of a telecommunication service reflects mainly the networks performances from packet loss, delay, jitter and bandwidth. QoS takes only the physical aspects into consideration, while QoE has a larger scope.

QoE covers a larger area of influencing factors and requires a multidisciplinary and multi-methodological approach. However, QoS focuses in pragmatic concepts and relies on analytic and experimental measurement, which is very useful when evaluating a system [54]. This is why we focus on QoS factors that are practical to measure and also have a major impact on quality.

2.7.2 Classification of QoE factors

VoIP quality depends on voice media quality, without considering other dimensions such as pricing, user interface .... Therefore, we focus on factors that are related to voice media.

Media quality factors can be divided into acoustic, codec technology and network factors. Acoustic factors will not be studied in this work. Network factors are mainly loss rate, delay, jitter, bit error rate, congestion and de-sequencing. Network factors have a great impact on VoIP services quality because of it real time constraint. Codec technology factors include encoding (compression format, DTX and VAD) and decoding (JBM, packet-loss concealment (PLC)).

Note that if certain characteristics such as codec can be clearly defined in standards to ensure interoperability and quality, other characteristics such as jitter buffer and acoustic characteristics are generally proprietary and their performance may vary. A solution to ensure minimum quality characteristics is to specify the associated quality requirements and test methods.

2.7.3 QoE measurement techniques

In this work, measuring QoE is a key element in order to evaluate the impact of some factors or the influence of some media adaptation algorithms on voice quality.

Subjective QoE measurement techniques

The most precise way to evaluate QoE is to evaluate user opinion with subjective tests. In this case, real people give their opinion about a service after utilization. A grade is attributed to every test sample. The resulting scores are averaged to give a MOS note as a general measurement of QoE. Table 2.2 presents the MOS scores that we can have according to the ITU-T Rec. P.800 ACR scale. Note that different scales are defined depending on the type of test.

5	Excellent
4	Good
3	Faire
2	Poor
1	Bad

Table 2.2: MOS scores (ITU-T Rec. P.800 ACR scale)

For VoIP services, a review of subjective test methods can be found in the ITU-T handbook of subjective testing practical procedures as well as in [10]. One can distinguish listening quality tests (ITU-T Rec. P.800, ITU-T Rec. P.835, ITU-R Rec. BS.1285, BS.1116, BS.1534) and conversation quality tests (e.g. ITU-T P.805).

The most used methodologies for listening quality tests are Absolute Category Rating (ACR) and Degradation Category Rating (DCR) defined in ITU-T Rec. P.800. ACR tests give the average score of opinions MOS. ACR tests are most commonly used to assess the integral quality of speech. In this type of test, a group of listeners evaluates a series of audio files using five-value scale from Table 2.2, without direct comparison to the original sequence. DCR gives the degradation of the average opinion rating called the Degradation Mean Opinion Score (DMOS). It relies in particular on a degradation scale and is suitable for evaluating good quality speech. The subjects note the level of degradation and discomfort by comparing with the original speech signal.

Subjective tests are normally performed in a well controlled lab conditions as explained in details in [56]. A comprehensive amount of subjective test results are available, e.g. in characterization reports in ITU-T and 3GPP. Subjective

tests are very expensive and complicated to conduct because they require the recruitment of people to do these tests.

Objective QoE measurement techniques

Objective models for measuring speech quality in VoIP communications are reviewed in details in [10, 57].

Intrusive, signal-based quality measurement tools The basic idea of intrusive methods is that a signal is injected into the system under test, and the degraded output is compared by the objective test system to the input signal considered as the reference, see Figure 2.10. Therefore, intrusive assessment techniques require access to both signal at the sending and receiving ends.

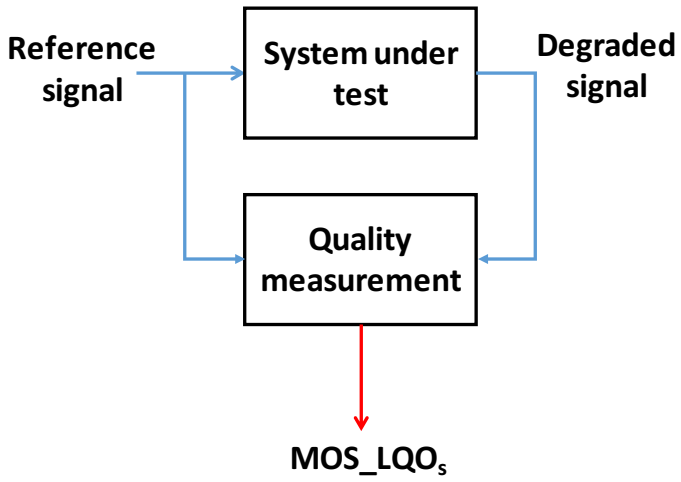


Figure 2.10: Signal-based measurement tool implementation

Several techniques have been developed to measure speech quality objectively: Perceptual Speech Quality Measure (PSQM) (ITU-T Rec. P.861), PESQ (ITU-T Rec. P.862) with its wideband extension WB-PESQ (ITU-T Rec. P.862.2) and Perceptual Objective Listening Quality Assessment (POLQA) (ITU-T Rec. P.863) are intrusive methods which have been trained for ACR speech tests. POLQA uses an advanced psycho-acoustic model for emulating the human perception to transform the sound into an internal neural representation so that

the objectives measures are as close as possible to the subjective quality scores. The final measure is converted to a MOS-LQO score as in table 2.2. POLQA is widely used in this work to measure speech quality.

Another model used to predict P.835 scores in the presence of background noise is the ETSI TS 103 281 model for SWB and FB terminals. It can estimate quality in 3 dimensions: S-MOS-LQO (speech quality), N-MOS-LQO (noise intrusiveness), G-MOS-LQO (global quality).

Non-intrusive quality measurement methods Non intrusive models are used usually for network planning but can be integrated within an adaptation algorithm for QoE prediction. Non-intrusive models are also called parameter-based models.

ITU-T has standardized the parametric E-model in [58, 59]. It has around 20 input parameters representing terminal, network and environment quality factors to predict a rating factor. This rating factor R is calculated using all impairment factors contributing to speech quality degradation. E-model considers that all the impairments are additive and can be calculated separately. The score lies in the range of 0 to 100, for NB communications, from bad to high quality. It is calculated as follows:

$$R = R_o - I_s - I_d - I_e + A \tag{2.8}$$

where,

- R_o: basic signal-to-noise ratio, including noise sources such as circuit noise and room noise.
- I_s: combination of all impairments, which occur more or less simultaneously with the voice signal.
- I_d: impairments caused by delay and the equipment impairment factors.
- I_e: impairments caused by low bit rate codecs.
- A: expectation factor that allows for compensation of impairment factors when there are other advantages of access to the user.

To simplify the calculation of the E-model, ITU-T has proposed to reduce the used equation to $R = 93.4 - I_d - I_e$ (ITU-T Rec. G.109), taking into

consideration impairments related to the network transmission and the default values of the other parameters. The transmission rating factor R predicted by the E-model is then converted to a MOS-rating scale.

The E-model has been a key element for evaluating the performance of different network architectures for various telecommunication services. We found in [60] a review for some evaluations of the E-model. Modified versions of E-model have been also presented to be more suitable for VoIP services. The applicability of E-model in the case of VoLTE and the necessity of studying jitter buffer algorithms are discussed because the E-model does not take into consideration conversational delay.

Many other non-intrusive models have been developed such as IQX model [61] and DQX model [62]. IQX model proposes a generic exponential formula for predicting QoE according to a specific QoS variables. DQX model proposes that QoE is based on two different types of variables; increasing and decreasing ones.

Summary

In this chapter, we presented the technical context of this dissertation. We presented the main elements of a VoIP audio chain. We studied the transport protocol used in real time application by presenting the corresponding advantages and limitations. We also studied the characteristics of voice packets transport over IP network and the possible feedback mechanisms. We defined quality of experience (QoE), quality factors and measurement methods.

Chapter 3

VoLTE Speech Quality Metrics Evaluation

Contents

Introduction

This work was mainly motivated by the 3GPP ART_LTE-UED and EXT_UED work items that we presented in Section 1.3. The focus of these two work items was terminal delay testing in VoLTE and jitter buffer characterization. Thus in this chapter, we evaluate metrics specified in 3GPP to characterize the trade-offs between delay and quality of VoLTE mobile phones in various network jitter and loss conditions. We report test results on clock accuracy, terminal delay in Uplink (UL) and Downlink (DL) under error-free conditions, as well as delay and quality in the presence of packet losses and network jitter. We propose some minor enhancements to the existing 3GPP test methods and 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.

3.1 Problem statement and related work

3.1.1 Problem statement

VoLTE presents a major technological turn in the way telephone operators provide voice services. As we discussed in Section 2.5, telephony services which are supported by EPC/LTE are all IP-based. Though VoLTE is deployed with some network optimizations and QoS garantees provided by IMS, several elements are proprietary (e.g., scheduling in eNodeB, de-jitter buffer in mobile phones). One of the objectives of this chapter is to discuss a way to evaluate the elements in terminals that are left unspecified and proprietary and how to derive generic model for network jitter/loss conditions. In particular, we discuss a method to evaluate the de-jitter buffer performance used by mobile phones.

VoLTE QoE is affected by many factors including audio quality, service availability, cost, security. In this work, we limit ourselves to the audio quality dimension, which can be characterized by various metrics, such as MOS [63, 60], mouth-to-ear delay, perceived loudness and frequency spectrum, interruptions (audio gaps). In this work QoE is presented by two main metrics MOS and End-to-End (E2E) delay.

Terminal	Network			
190 ms	E-UTRAN	EPC	Mobile IMS	Transmission
(Source + destination)	80 ms	50 ms	0	10 ms

Table 3.1: VoLTE E2E delay budget [5, Table 3]

All the elements of the VoIP system contribute to the E2E delay at different levels as discussed in Section 2.2.2. Based on [64], E2E delay is mainly composed of terminal and network delays. Terminal delay can be divided into source delay (including encoding, playout buffer, packetization, ...) and destination delay (including de-jitter buffer, decoding ...). De-jitter buffer used in the VoIP receiver must smooth out packet delay variation due to network jitter but also due to clock skew of the sender and the receiver [65]. Network delay includes all the delays induced by the different parts according to the network architecture.

In 4G networks, VoLTE E2E delay is detailed in [5]. Table 3.1 presents the VoLTE E2E delay budget (maximum delay) for both terminal and network elements according to [5]. As we can see terminal delay can greatly affect E2E delay and therefore overall the communication quality. The terminal delay is affected by many factors including de-jitter buffer which is a crucial one. De-jitter buffers are proprietary mechanisms in terminals as we have seen before, and the standards do not propose any methods for on line performance evaluation. In this work we investigate the trade-off between quality and E2E delay with focus on terminal delay.

In particular, this work is based on LTE terminal delay testing specified in 3GPP [7, 6]. 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). The main contribution of this chapter 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. We also propose enhancements to delay/loss packet trace simulations to better represent the actual delay and quality that can be experienced in VoLTE.

3.1.2 Related work

A comprehensive review of quality metrics in VoIP can be found in ITU-T Rec. G.1020 [64] and G.1021 [66], 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 [67]. De-jitter buffer size estimation and optimization with respect to QoE is discussed in [68]. The impact of de-jitter buffer playout delay adjustments has been studied for instance in [69, 29].

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 based on RTP time stamps and not at the acoustic level based on acoustic events at the mouth and ear [25, 70]. Black-box delay measurements at the acoustic level have been reported for instance in [71], 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; audio frames do not have the same state with each call for the same network condition.

Some tests 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 [72, clause 32] under error-free conditions. For 3G, terminal delay tests have been defined in Release 11 of TS 26.131 [7] and TS 26.132 [6]; these tests under error-free conditions have been extended to LTE terminals in Release 12 of these specifications.

3.2 Experimental set-up for VoLTE delay testing

The test set-up used in our experiment follows 3GPP acoustic tests defined in [6]. We used an example of implementation based on test equipments from different vendors, as detailed in our 3GPP contribution [3]. The test set-up presented in figure 3.1, is specific to the case of a mobile phone in handset mode

the headset and hands-free modes are not considered here. Our objective is to measure the E2E delay from the reference client to the used mobile phone through the network emulator as presented in figure 3.1.

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.

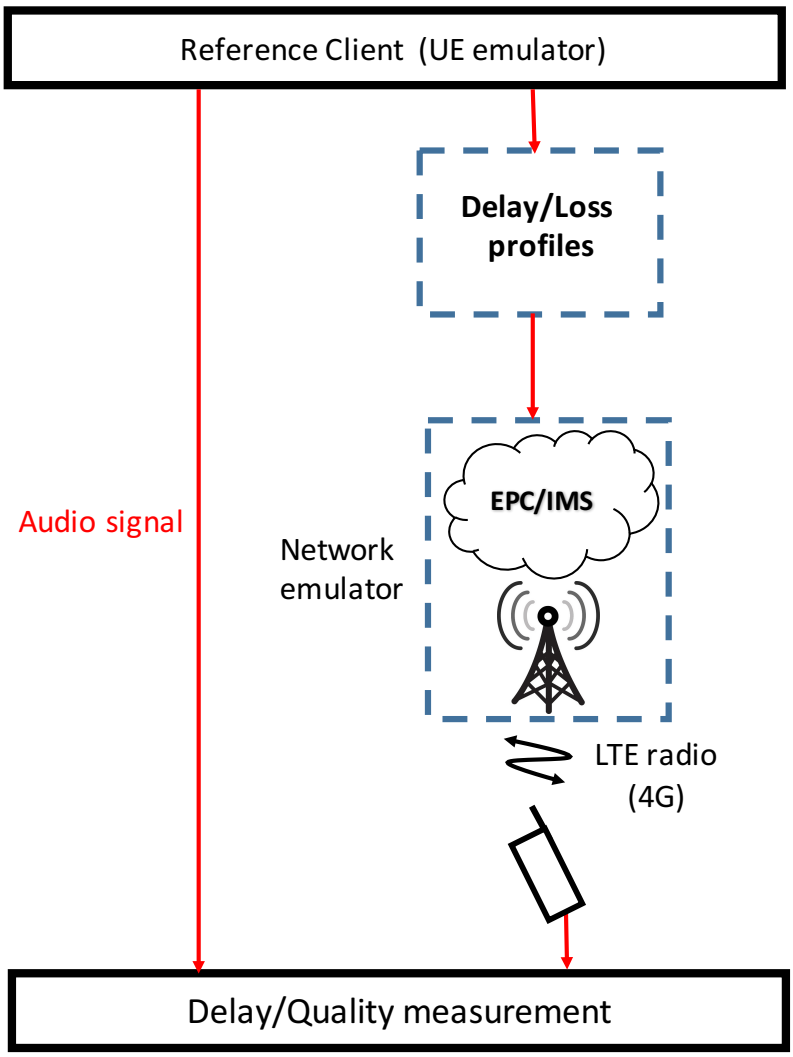


Figure 3.1: Experimental set-up

Unit	Version
ACQUA	3.3.400
MFEVII	1.11.1858
CMW500 base	3.5.80
CMW500 LTE sig.	3.5.30

Table 3.2: Versions of the used equipments

The UE is mounted in handset mode on a manikin as described in ITU-T Rec. P.64 [73] also called head and torso simulator (HATS) with built-in artificial ear which is conform to ITU-T Rec. P.57 [74] (Type 3.3) and artificial mouth which is conform to ITU-T Rec. P.58 [75].

A specific SIM card with proper operator settings is inserted in the mobile phone for testing purposes. A VoLTE call is established with an LTE/EPC network emulator (Rhode & Schwarz CMW500), as illustrated in Figure 3.2 , using its internal IMS server to setup a dedicated bearer with $QCI = 1$ [45]. All tests have been conducted with mobile originated calls with the AMR-WB codec at 12.65 kbit/s with DTX deactivated.

A computer operating an objective measurement system (Head Acoustics ACQUA), is used in compliant with [6] 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), see figure 3.2, which implements AMR-WB codec, 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), shown in Figure 3.2, is used as an audio interface performing A/D and D/A conversion between the manikin and the reference client.

3.3 Clock skew in VoLTE

For delay measurements, it is essential to estimate and compensate for audio clock skew between end points. See Section 2.3.1 for more details about clock

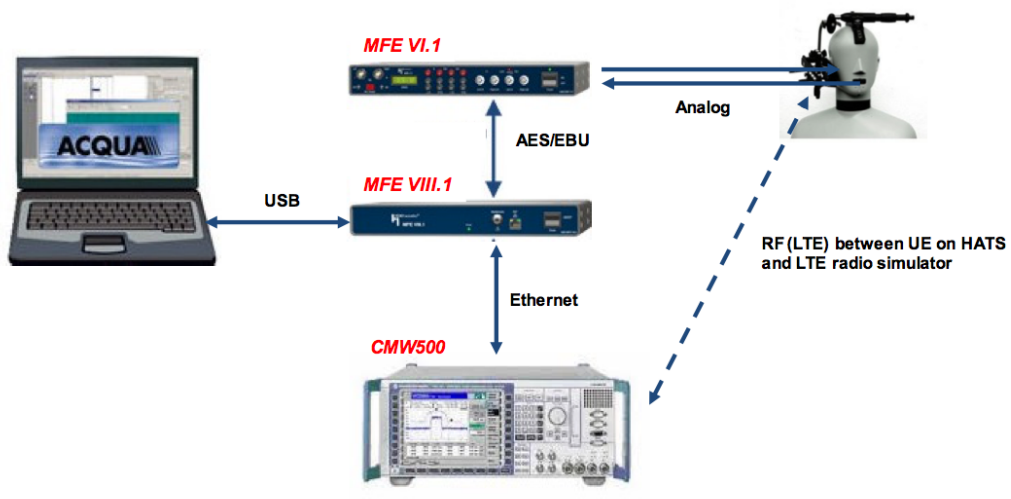


Figure 3.2: Implementation details of the experimental setup, [3]

skew.

In this work, the reference client (MFE VIII.1) has a clock frequency of 48000 Hz; this clock frequency is reset prior to measuring clock skew, and it is then adjusted to compensate the estimated clock skew. All test results presented in this chapter have been performed after synchronizing the clocks of the mobile phone and reference client.

Table 3.3 lists the relative clock skew between several VoLTE mobile phones and the reference client. 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 be adjusted based on the network clock and thus compensate for temperature variation.

Based on the results in Table 3.3, we can target defining a requirement (at least for the handset mode) that the absolute clock accuracy of devices should be less than 10 ppm. As this is in practice a relative clock offset with respect to the reference client, a requirement on the clock accuracy of the reference client would have to be also defined.

For this work, we conclude that clock skew may not have a significant impact on QoE for VoLTE as we also showed in Section 2.3.1 for a small clock skew.

There are potentially some cases where on may consider higher values of clock skew. For example, 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. [6] describes the case of softphone interface, where the audio clock may be derived from a laptop or a mobile device sound card and the telephony service is provided using a software in a non-integrated approach. In this case, the audio clock may not be adjusted with the radio interface clock. Inputs on clock accuracy for this use case would be necessary to check the expected worst-case clock skew.

3.4 Uplink and downlink terminal delay (error-free case)

3.4.1 Measurement objectives and methodology

In this Section, we measure the terminal delay in perfect network condition i.e., no packet loss and nearly no jitter (< 1 ms) from the network emulator. We validate if the used terminals comply with the allocated delay budget that we presented in Table 3.1. We also discuss the possible variabilities of the terminal delay measures in UL and DL directions.

A speech test signal according to ITU-T Rec. P.501 [76] of 32000 samples (at 48 kHz sampling rate) is used as a test signal.

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

Table 3.3: Clock skew in sending and receiving.

3.4.2 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 artificial mouth and the last bit of the corresponding speech frame at the phone antenna [7], as illustrated in figure 3.3. 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 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 emulator).

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:

$$T_S = T_{US} - (T_{TES1} + T_{TES2} + T_{Prop}) = T_{US} - T_{TES} \tag{3.1}$$

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

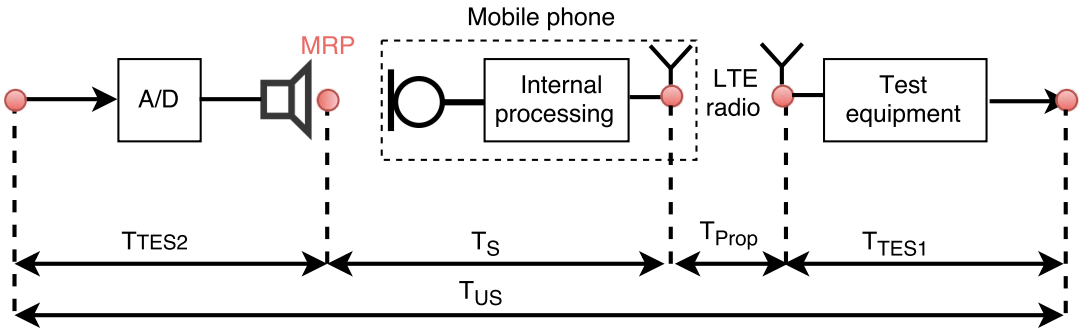


Figure 3.3: Sending delay measurement.

- Network emulator delay: 9.47 ms [77]
- Acoustic front end delay (A/D): 0.4 ms

3.4.3 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 [7], as shown in Figure 3.3. To calculate T_R, we measure the uncompensated delay T_{UR} by cross-correlation analysis between the measured signal at DRP and the original one at test equipment input (reference client).

The calculation of T_R is similar to the sending delay case and we have the following equation:

$$T_R = T_{UR} - T_{TER} \tag{3.2}$$

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

- Reference client delay: 68.5 ms (including resampling and encoding operations)
- Network emulator delay: 8.73 ms [77]
- Acoustic front end delay (A/D): 1.31 ms

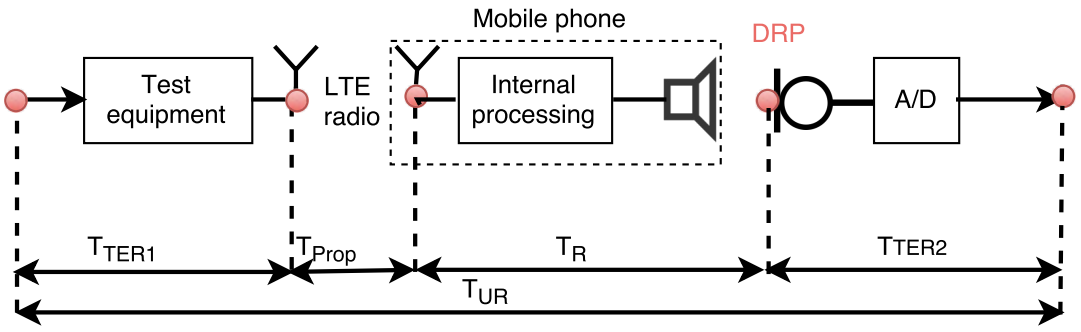


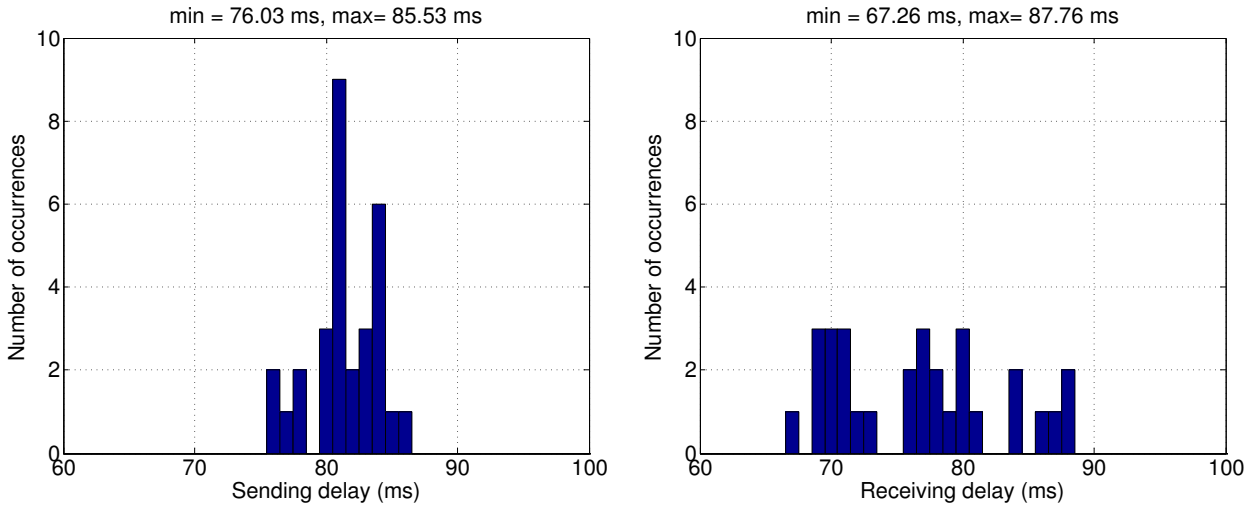
Figure 3.4: Receiving delay measurement.

3.4.4 Discussion: 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 histograms of measured delays are shown in figure 3.5(a) for the UL and figure 3.5(b) for the DL. One can verify that both histograms cover an interval of at most 20 ms which corresponds to the codec frame length. This measurement uncertainty has been attributed in [77] 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.

In [6], 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 Figure 3.5(a) and Figure 3.5(b) respectively 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 in the 3GPP to be a compromise between testing time and accuracy/repeatability.

Note that VoLTE terminal delay targets are specified in [7] as overall (send+receive) delay: $T_S + T_R \leq 190$ ms (mandatory) as we presented also in



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

Figure 3.5: Sending and receiving terminal delay in error free condition.

Table 3.1 and $\leq$ 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 ($\leq$ 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.

3.5 Delay/quality under delay/loss conditions

3.5.1 Measurement objectives and proceedings

In this section, we measure terminal delay and quality in different network conditions by applying different profiles of jitter and packet loss on the transmitted speech frames. The delay tests specified in [6] 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 [45][28, chap. 8], in such non-ideal conditions, both delay and quality are measured.

We used ITU-T Rec. P.863 POLQA for quality measurement. We used a test signal which consists of four 8-second English sentence pairs according to [76, 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.

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

To emulate network impairments, delay/loss degradations at the IP level have been proposed in [78] for delay testing. The profiles are applied at the Ethernet interface between the Reference client and network emulator, see figure 3.1. The 3GPP adopted three end-to-end profiles that have been generated by simulation, with a MATLAB source code given in [6, Annex E].

Note that for WB calls only two profiles are applicable, and the third

Model parameters/statistics	Cond.1	Cond.2	Cond.3
Target BLER (%)	10	10	22
Max. of HARQ retransmission	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

Table 3.4: Jitter/loss profile parameters [6].

profile is intended for (SWB) 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 3.4. The profiles simulate RTP packet impairments between the IP network interfaces of two VoLTE mobile phones, at the antenna reference points shown in Figure 3.3. They model static jitter conditions (i.e., no mobility, no varying cell load) with a simplified handling of a dedicated bearer with $QCI = 1$ and Discontinuous Reception (DRX) [79]. DRX is a mechanism used for UE battery saving in which the UE gets into sleep mode for a certain period of time and wake up for another period of time. The UE receive packets in discontinuous way, only in the waking periods which can be decided by the eNB for UE-Network synchronization.

The 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. Each profile entry indicates the delay of a packet if it is not lost and -1 for lost packets. Therefore voice packets experience the same network degradations in separate calls for a given profile; 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 equation 3.2), except that the minimum delay added by profiles (30 ms for the three profiles) is also subtracted.

3.5.3 Test results and comparison with delay/quality requirements

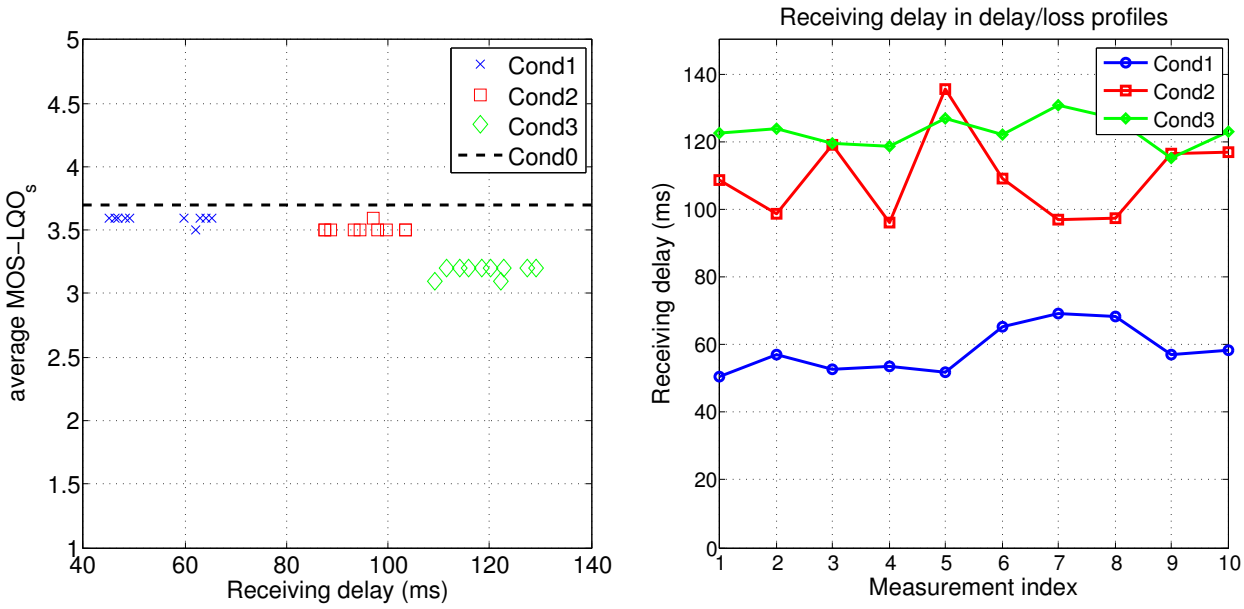
Figures 3.6(a) and 3.6(b) show quality and delay measurement results obtained for the three profiles using the same VoLTE phone (phone A) as in error-free conditions, with ten repeats in separate calls. In [6], the measured receiving delay T_R^{imp} for each condition (profile) is defined as the $95th$ *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 ITU-T Rec.P.863 with POLQA in (SWB) 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-Listening Quality Objective $(LQO)_S$ in (SWB) value. The ten repeats of each condition, one to three, show the same delay variability as in error-frame with an interval of 20 ms, Figure 3.6(b). Similarly, the quality score variability (around 0.1 MOS) is in the expected range for ITU-T Rec.P.863, Figure 3.6(a). In [6], 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. Let us suppose that the minimum jitter buffer length is 40 ms sufficient to contain two audio frames. However, when applying jitter/ profiles it needs to increase its length to 80 ms at least for Cond. 1, see Figure 3.7(a) for Cond. 1 packet delay distribution and more than 100 ms for Cond. 2 & 3, see Figures 3.7.(b) and 3.7(c) for packet delay distribution in Cond. 2 and Con.3 respectively. Yet no official methods are defined to measure the effectiveness of jitter buffer adaptation and to ensure that it fulfills the 3GPP requirements defined in [28]. 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 Figure 3.5(b)); this can be explained by the fact that 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 occurred yet.

These test results show that it would be better to repeat delay measurements



(a) Quality (Delay) in jitter/loss conditions

(b) Delay in jitter/loss conditions

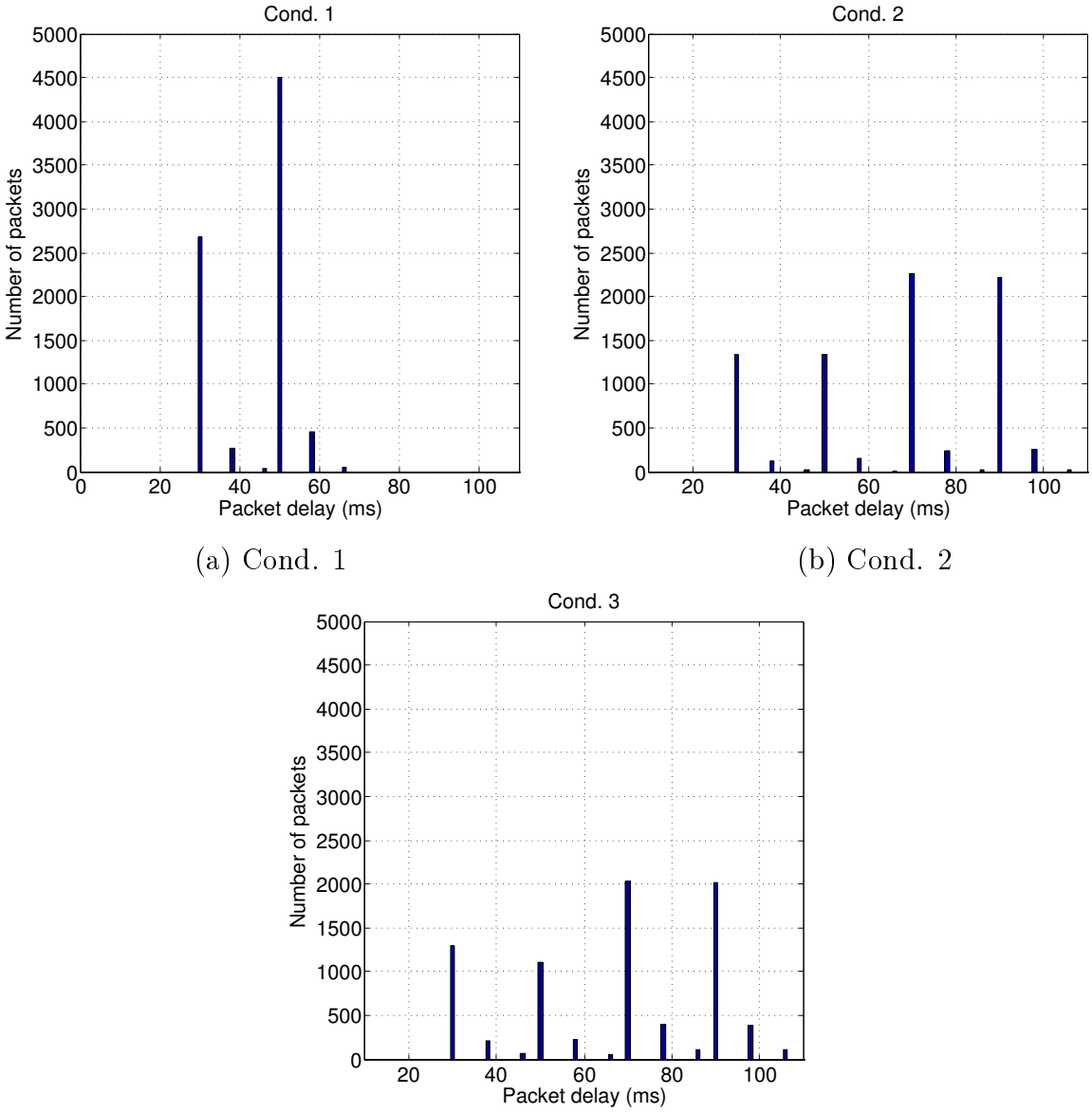
Figure 3.6: Delay and quality in jitter/loss conditions.

in separate calls to capture variability. To avoid repeating this measurement, one could concatenate a CSS signal and a 160 s 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 offset based on the maximum value of T_R previously obtained after several repeats in error-free case.

VoLTE terminal delay targets for impaired conditions are specified in [7] in

	$T_S + T_R^{imp}$ (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$_{Cond0} - 1$

Table 3.5: Delay/quality targets for wideband calls[7] (with an extra condition taken from SWB requirements).



(a) Cond. 1

(b) Cond. 2

(c) Cond. 3

Figure 3.7: Packet delay distribution for the used 3GPP conditions.

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 Cond. 0) as shown in Table 3.5. The tested phone meets all requirements. Figure 3.6(a) illustrates a particular design choice with a certain trade-off between late loss

rate (caused by late packets) and delay.

3.5.4 DTX influence on VoLTE performance

DTX is one of the optimization to enhance the network capacity as we discussed in Section 2.1.3. By the time of conducting this experiment, DTX was disabled in the standard for the VoLTE. We conducted our measurements with DTX deactivated. Then, we tested for each packet/loss profile described earlier the impact of activating the DTX on the measured VoLTE delay and quality. A separate call is established for each profile/DTX mode combination. The same set up configuration is used as before. We only modify DTX mode in the reference client settings for the case where DTX is activated. We use the IP logging in the network emulator to check the codec bit rate and the activation of DTX.

Table 3.6 reports quality and delay at the receiving side (T_R^{imp}) results in case of DTX on and off for the three network profiles. These results were reported in one of our 3GPP contributions [3], where we confirm that the use of DTX does not has an impact on the measured quality neither on the received delay. However, DTX has an indirect impact on service QoE because it optimizes network resources consumption as we explained in Chapter 2.

3.6 Towards de-jitter buffer performance metrics using realistic VoLTE network models

We investigate in this Section, how the 3GPP delay test methodology with delay/loss packet profiles 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,

Condition	DTX OFF		DTX ON		Difference	
	Delay (ms)	MOS-LQO (Avg.)	Delay (ms)	MOS-LQO (Avg.)	Delay (ms)	MOS-LQO (Avg.)
Cond. 1	87.2	3.7	99.3	3.6	−12.1	0.1
Cond. 2	146.3	3.6	120.5	3.6	25.8	0
Cond. 3	150.4	3.3	130.4	3.3	20	0

Table 3.6: DTX impact on delay and quality measurements

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 [64] and G.1021 [66].

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 are generated using a MATLAB simulation as described in [6, 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 modeled. Due to the simplified EPC jitter model, the ratio of out-of-sequence packets is quite high in the used 3GPP profiles with DRX 40 ms,see table 3.4, which is typically not observed in real life. Note that packet delay variations in the three profiles of [6] 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 DTX 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 [6, Annex E] to obtain packet traces that are more representative of real VoLTE networks.

Figure 3.8 (a) shows an example of instantaneous IPDV metrics for a commercial VoLTE network using DRX 40 ms and semi-persistent scheduling (SPS). The instantaneous IPDV is defined as:

$$IPDV(i) = D(i) - D(i-1) \tag{3.3}$$

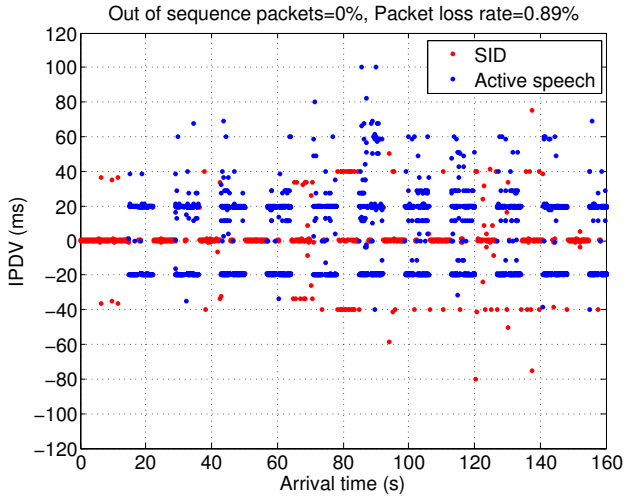
where $D(i)$ denotes the one-way delay of the ith packet.

Measurements were made with one static mobile phone and another mobile phone in a car during a drive test. Figure 3.8 (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 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 ± 20 ms 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.

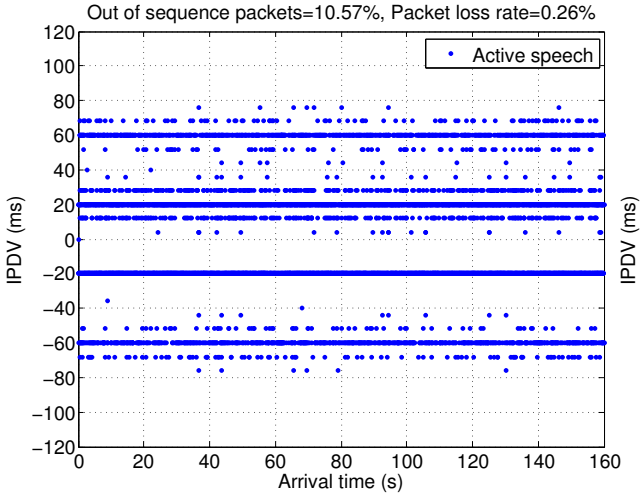
For comparison purposes, IPDV for Cond 2 (DRX 40 ms) from simulated profiles is shown in figure 3.8 (b). DTX is assumed deactivated, and all frames are considered as active speech; this has the advantage that figure 3.8 (b) is independent from any specific speech database. Beside DTX which has no influence on the measured delay as we have seen in the last section, a key difference between Figures 3.8 (a) and 3.8 (b) lies in the amount of out-of-sequence packets (10.575 % for Cond 2) and packets that missed their normal DRX cycle; the target BLER of 10 % in Cond 2 results in quite many HARQ retransmissions.

To better match the example from Figure 3.8 (a), one can modify the MATLAB routine `VoLTEDelayProfile_vPHY` from [6, Annex E], as follows:

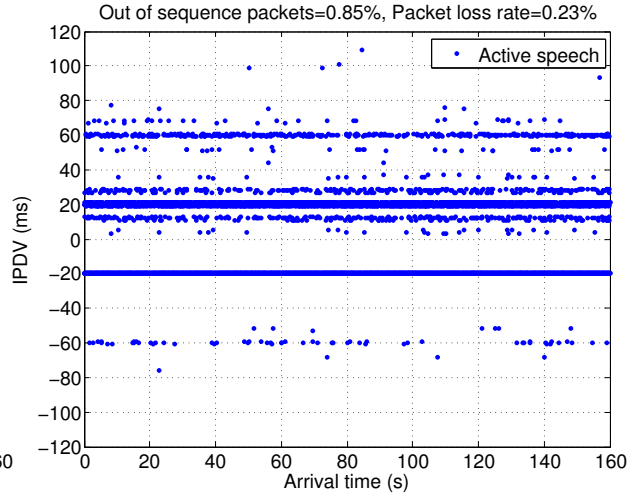
- eNodeB scheduling can be made less periodic, by adding a random jitter of $\{0, 1\}$ ms to scheduling times.
- The uniform distribution for network delay (i.e., delay between two eNB(s) can be replaced by a long-tailed distribution, such as a Weibull mixture model with 2 components, keeping the same minimum network delay as in [6, 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.



(a) Real packet trace (DTX on)



(b) Simulated profile (DTX off)
Condition 2 from [6]



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

Figure 3.8: Comparison of IPDV in real and simulated conditions.

It can be verified from Figure 3.8(c) that, with the modifications listed above, the resulting IPDV better reflects the real example in Figure 3.8(a). The packet loss rate of 0.22% is close to that of the original condition 2, because the HARQ simulation from [6, 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.

Conclusion

In this chapter, we analyzed the existing delay-related metrics for VoLTE terminals. We showed the variation of network conditions can heavily impact VoLTE QoE. We highlighted some shortcomings in the early 3GPP acoustic test methods. In particular the measurement variability was 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 finally proposed improvements to the VoLTE packet delay/loss simulation model used in 3GPP.

Chapter 4

WebRTC Performance Evaluation Over Different Network Conditions

Contents

Introduction

In this chapter, we evaluate the performance of an OTT voice application based on WebRTC over different network conditions. We measure and evaluate WebRTC

speech quality metrics over LTE network. We present experimental results on WebRTC QoE metrics as a function of LTE radio link pathloss level. We complete the study with an evaluation of WebRTC over Ethernet to investigate the impact of bandwidth limitation and packet loss rate on WebRTC speech quality. We evaluate for the different cases the impact of varying bit rate on the measured quality. We use different experimental setups for the two studies including a real LTE platform under controlled lab condition and a simple implementation of WebRTC communication over Ethernet. For speech quality measurement, we use as previous chapter POLQA [80] to assess voice quality in terms of MOS-LQO together with a measurement of mouth-to-ear delay between mobile phones.

4.1 Problem statement and related work

4.1.1 Problem statement

OTT applications including those using WebRTC technology are designed to operate on best effort networks. No QoS guarantees are provided for such services as we have seen in chapter 3 with VoLTE. OTT services are mainly based on RTP over UDP protocol which is a generic transport protocol, see Section 2.2, and does not provide any media adaptation mechanisms. It is up to the application to define its own mechanisms to optimize quality. Bit rate adaptation, redundancy and jitter buffer management are usually implemented.

The general objective of this research is to study the impact of such algorithms on speech quality in the context of OTT services and also in network operators services such as VoLTE. In this chapter, we study the impact of some of these algorithms on speech quality in certain network conditions. The study does not cover all the possibilities of existing network architectures and different network degradations. However, it gives an idea about the effectiveness of certain media adaptation algorithms of enhancing QoE in degraded network conditions. We are interested in a specific OTT based on WebRTC.

EVS has been mainly developed for VoLTE services. Because this work has been conducted in a larger context to compare the performance of VoLTE and WebRTC in LTE radio coverage tests, if possible with the same voice codec, we add the support of EVS to WebRTC to have an idea about EVS quality in OTT/WebRTC context and to be able to conduct tests under different encoding

bit rates. EVS is a voice codec that can operate under multitude number of coding bit rates and modes [13].

We also study the impact of jitter buffer on speech quality and delay, in particular WebRTC jitter buffer [27] is studied in a black box approach.

In this chapter, we deal with several problems related to WebRTC speech quality. For that we conduct several tests to evaluate WebRTC performance in terms of quality and delay. First, we experimentally evaluate WebRTC speech quality over LTE radio link in a way similar to the radio coverage tests but in a controlled lab conditions. We evaluate the impact of varying bit rate when we degrade the LTE radio link by applying pathloss to the radio link. We analyze in a black box approach the jitter buffer impact on the measured speech QoE metrics. Second, we extend the study with another experiment to evaluate the impact of varying encoding bit rate in case of limited bandwidth and lossy channel. We evaluate bit rate impact on speech quality and which feedback on network condition is required to trigger such adaptation.

4.1.2 Related work

In general, a lot of efforts have been devoted to evaluate objectively VoIP quality, as in [81, 82, 83, 84] over different networks. WebRTC performance has been evaluated in many studies. WebRTC congestion control algorithms has been studied in [85, 86]. These algorithms consists of dynamically adapting media bit rate with focus on video according to the available network bandwidth based on several indicators such as packet loss rate and delay variation.

However, the evaluation of WebRTC voice quality has received less attention in the literature [87, 88]. Time scale modification introduced by WebRTC jitter buffer has been evaluated through objective and subjective tests in [88], which shows that objective tools are more sensitive to time scaling modification than real subjects when assessing voice quality.

WebRTC audio and video quality over LTE have been studied using simulations in [87]. To better represent the real usage context of end users and to better evaluate QoE, it is preferable to rely on real system implementations (mobile phones and networks). In this work, we focus on WebRTC speech quality evaluation using real experimental implementations. We report some QoE

metrics representing WebRTC speech quality in different network conditions as detailed in section 4.1.1. Furthermore, we explore the impact of varying voice codec bit rate on perceived quality without proposing any adaptation algorithms.

4.2 LTE coverage impact on WebRTC speech quality

4.2.1 Data transport over LTE

VoLTE is transported over dedicated GBR bearers (QCI 1) with UM data transmission at the LTE RLC layer and specific QoS guarantees. WebRTC media is typically transported with the RLC AM over the Non-GBR default bearer (QCI 6/8/9) [44, Table 6.1.7]. In other words, in AM mode, error correction is applied through ARQ. Figure 4.1 shows an example of packet retransmission when a packet is lost according to ARQ principle in AM mode. It consists of requesting a retransmission of packets with transmission errors (p1 in Figure 4.1) through a feedback message. In case of transmission with errors, a Non-ACKnowledgement (NACK) message is sent from the receiver to indicate a bad reception. In case of good reception, the receiver sends an ACKnowledgement (ACK) message to inform the transmitter that the packet is well received (p2 in Figure 4.1).

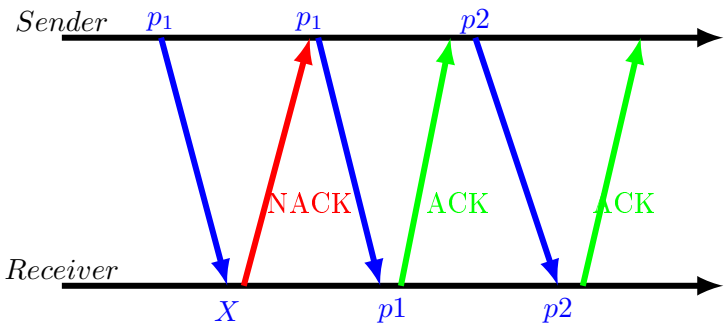


Figure 4.1: ARQ mechanism at RLC layer in AM.

The default bearer used for data traffic over Internet has the following QoS settings: packet delay budget for each LTE radio leg $\leq$ 300 ms, packet loss rate $\leq 10^{-6}$, see [44, Table 6.1.7]. Table 4.1 summarizes the differences and similarities between WebRTC over LTE and VoLTE.

Table 4.1: WebRTC and VoLTE comparison

	VoLTE	WebRTC
Operator	Network operators	OTT
Protocol	IP	IP
Resource reservation	Dedicated bearer with GBR	Default bearer with non-GBR
Prioritization	QCI 1	QCI 6/8/9
RLC error correction	UM	AM
Delay budget	80 ms	300 ms
PLR	$\leq 10^{-2}$	$\leq 10^{-6}$
Radio optimizations	ROHC/HARQ/TTI bundling	None

To make it more tractable, this study is limited to the case of radio coverage tests in reproducible and controlled lab conditions (e.g. one mobile phone per radio cell). Aspects such as network congestion and multi-user environments or performance in live network conditions (drive tests) are not taken into account and out of scope. To be specific, in this section, we measure and analyze speech quality and mouth to ear delay for various LTE coverage conditions and codec bit rates. This section is based on our work in [89].

4.2.2 Experimental setup and measurement methodology

In this section, we evaluate WebRTC speech quality metrics over LTE radio link. The basic idea is to have two softphones based on WebRTC communicating over mobile LTE network as shown in Figure 4.2. The details of the experiments are in the following.

WebRTC application and modifications

We use a native application called AppRTC, which is provided as a demo application in the open-source WebRTC Chromium project [90]. WebRTC defines two voice codecs that are mandatory-to-implement: OPUS and ITU-T G.711. The source code of the WebRTC library from Chromium (also used in AppRTC) has been modified to include the support the voice codec EVS.

Therefore, we apply the following modifications to Chromium's WebRTC library.

- Encryption (DTLS/SRTP) is disabled to allow decoding captured RTP streams.
- EVS encoder and decoder are integrated in WebRTC's audio coding module.
- EVS database is added by replicating the existing integration of the OPUS encoder and decoder; due to this replication, one may consider that the results presented hereafter reflect the audio processing that is also executed with OPUS, even if tests have been conducted with EVS.

Note that the EVS codec is constrained to operate in SWB mode at 9.6 kbit/s or above. The RTP payload format for EVS is implemented. We conduct experiments in the following SDP settings: packetization time `ptime=20` corresponds to speech frame length, header full EVS RTP payload format `hf-only=1` and CMR activated `cmr=1`, hence 2 header bytes are appended to speech data in each 20 ms RTP packet. We only use three bit rates of the EVS codec in SWB mode: 9.6, 13.2, and 24.4 kbit/s.

Testing platform details

We use a real LTE/EPC platform that supports IMS, with an access to the Internet to make sure that WebRTC voice calls could be properly established as shown in Figure 4.2. The test platform used in this work is based on commercial radio equipment found in live networks. Two mobile phones are connected to two different LTE eNodeB's with LTE radio cell in the 2.6 GHz band (20 MHz bandwidth). The two UEs are connected to the same EPC PGW in the same local network. The two IP address have the same subnet mask and the RTP streams remain in the local network.

The UEs under test are radio isolated in separate RF shielded boxes, in which a radio antenna is placed to provide LTE radio access. One of the generated radio signals is degraded and attenuated using a fast fading and variable attenuation generator (SPIRENT VR5) as shown in Figure 4.2. In radio propagation, fast fading presents the effects of the rapid variation of radio channel characteristics in time compared with the duration of data symbol. The EPA (Extended pedestrian A) 3km/h multipath channel model is used as defined in [91, Annex B2]. In our case, only the uplink direction of UE 1 is degraded and UE 2 is in perfect radio condition. UE 1 sends speech to UE 2. Silence is injected in the other direction. This simplification is justified in Section 4.2.3.

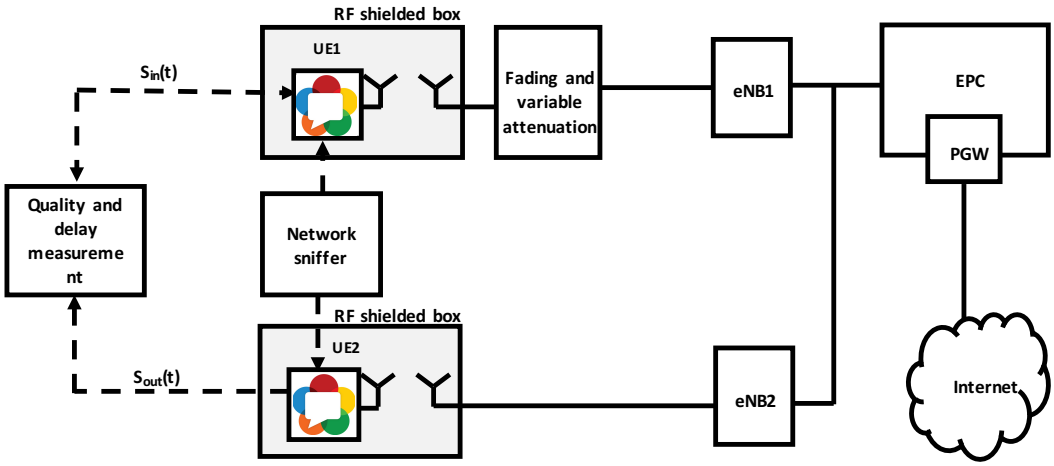


Figure 4.2: WebRTC over LTE experimental setup

No degradation in the downlink direction is done because even in worst cases the available bandwidth is sufficient for speech transport.

The radio signal level is defined in terms of Reference Signal Received Power (RSRP) [92] as the linear average over the power contributions of the resource elements that carry cell-specific reference signals within the considered measurement frequency bandwidth. RSRP measurement (in dBm) is used mainly to rank different candidate cells in accordance with their signal strength. The RSRP level is translated to radio pathloss level expressed in (dB). RSRP level impact on available bandwidth will be discussed later in this section.

Delay and quality measurement proceedings

The UEs are used in headset mode (with an analog audio input/output from the electrical jack interface). A computer operating a voice quality measurement system [93] is used to conduct testing, collect and analyze test data. Voice quality is measured using POLQA (v2.4) which allows measuring quality according to ITU-T Rec. P.863 [80] in terms of MOS scores, see Figure 4.2. The UEs are traced by a network sniffer called (QXDM) [94] to capture the incoming and outgoing data traffic for further analysis. Note that video is deactivated for all

tests to evaluate speech-only communications.

An 8-second test sequence consisting of one 4-second female sentence and one 4-second male sentence in French is used; this sequence is compliant with P.863.1 [95] and it has been selected for drive tests. The 8-second sequences is repeated 20 times, which gives 20 MOS scores. The average of the 20 scores is then computed and is used as quality indicator. In parallel, the end-to-end delay is measured for each 8-second sequence and the average is computed of the 20 repeats in a given network condition. We compute the confidence intervals of 95% for MOS scores as well as for the delay.

4.2.3 Radio signal attenuation impact on available bandwidth

We performed preliminary tests to obtain the average throughput in Mbit/s presenting the available bandwidth as a function of the radio pathloss. Figure 4.3 show UE 1 and UE 2 throughput in UL and DL respectively as a function of the pathloss. In both cases, the measured throughput decreases as function of the radio pathloss. We notice that in bad radio coverage, the DL maintains enough bandwidth (around 2 Mbit/s in 15 dB of pathloss) until the total cut of the communication. The maximum used bit rate of EVS is 24.4 kbit/s. If the different

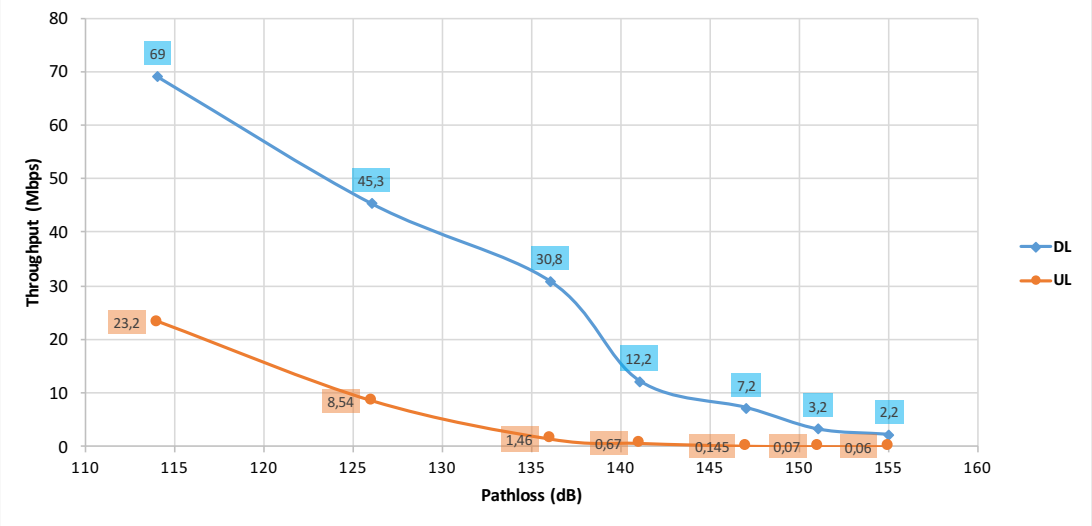


Figure 4.3: UL and DL bandwidth in terms of throughput as function of pathloss

protocol headers are taken into account, the order of magnitude of the total bit rate is around 50 kbit/s. The minimum throughput (2 Mbit/s) is thus largely higher than the total bit rate with EVS. However, in UL direction, bandwidth is rapidly degraded (around 50 kbit/s in 155 dB of pathloss). This is why we present results only in UL direction of UE 1.

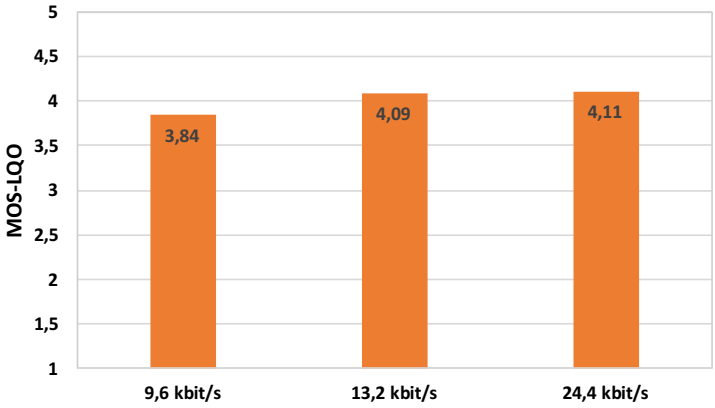
4.2.4 Measurement results

Quality and delay in good LTE radio coverage

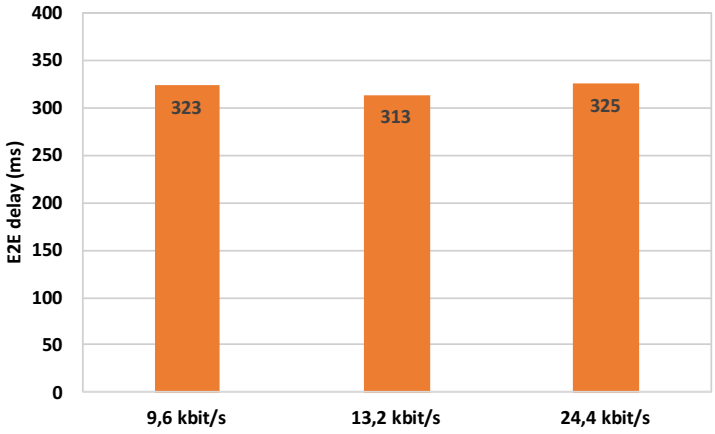
Tests are carried out under good LTE coverage conditions. The RSRP level is around -90 dBm corresponding to a pathloss of 114 dB. The measured quality (MOS-LQO$_s$) increases with bit rates, as shown in figure 4.4 (a). The headset jack used interface and other acoustic characteristics of the UEs (e.g. the frequency response in sending/receiving) as well as the audio processing module in WebRTC can have some impact on POLQA scores. Hence, the reported scores do not reflect the POLQA scores that could be expected from the sole codec contribution (e.g. around 4.75 for EVS at 24.4 kbit/s). For these reasons, POLQA scores depend on the phone models used for testing, and the values in Figure 4.4 (a) should not be taken as absolute values, they should only be understood as indicative examples.

Figure 4.4 (b) gives the corresponding measured mouth-to-ear delay. This delay is close to 300 ms in good coverage conditions, which is higher than the figure of 150 ms recommended by ITU-T G.114 for good user satisfaction [96]. Similar to MOS, it has been noticed that different phone models can yield to significantly different delay values (e.g. up to 200-300 ms extra delay), therefore the specific values reported in Figure 4.4(b) should only be considered as examples that are valid only for specific phones under test.

In good network conditions, the codec bit rate has a slight influence on MOS. The MOS difference between the lowest and highest used bit rate is around 0.3 MOS as shown in Figure 4.4 (a). However, the delay difference is 10 ms, which is negligible compared to the average value 320 ms as shown in Figure 4.4 (b). Thus, in good network conditions, the codec bit rate has a slight impact on delay. It is also important to note that the packet streams do not get out from the same local network, in other scenarios when the two UEs are from different networks and the packet streams cross over other network equipments, mouth-to-ear delay



(a) Average MOS-LQO$_s$



(b) Average E2E delay

Figure 4.4: MOS and E2E delay in low pathloss (114 dB)

can be expected to be higher.

Quality and delay as function of LTE coverage

Tests are conducted under degraded LTE coverage conditions using the fading and variable attenuation emulator to decrease the RSRP level and consequently increase radio pathloss until call drop. The variable attenuation emulator gradually attenuate the signal until having a total communication failure when a severe drop in the channel signal-to-noise ratio happens.

Figure 4.5 shows the evolution of MOS as function of radio pathloss. The MOS value decreases in a similar way for the three tested codec bit rates and

bit rate had a slight influence on coverage. It is expected that lower rates would enable to operate the codec in higher pathloss. This is verified by the trends shown for EVS at 13.2 and 24.4 kbit/s. The MOS curve at EVS 9.6 kbit/s was parallel to the MOS curve at 13.2 kbit/s with no cross-over; this may be explained by the lower intrinsic codec quality at 9.6 kbit/s which did not seem to be compensated in more degraded network conditions, even close to the coverage limit.

With the testbed used in this work there is an uncertainty of about 1 or 2 dB on actual pathloss due to fast fading (resulting in fluctuations in pathloss evaluation). Care should be therefore taken when interpreting the relative coverage of different bit rates. If the coverage limit is for instance set to a MOS threshold of 2.5, the corresponding path loss is be around 142, 143 and 141 dB at 9.6, 13.2 and 24.4 kbit/s, respectively; given the uncertainty on pathloss, coverage may be considered nearly equivalent for all tested bit rates. Note also that the confidence intervals indicate higher variability in higher pathloss and it would have been interesting to use more than 20 repeats of the 8s sentence pair

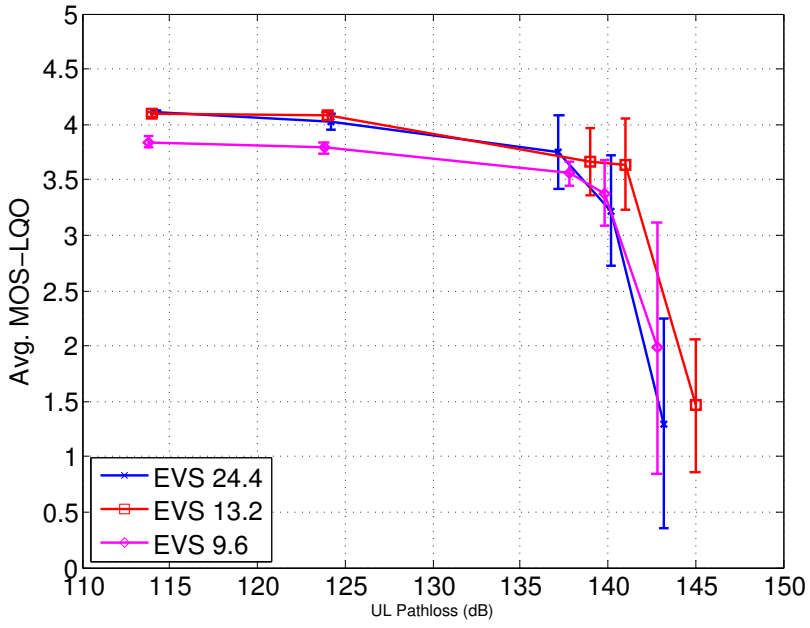


Figure 4.5: MOS scores as a function of pathloss.

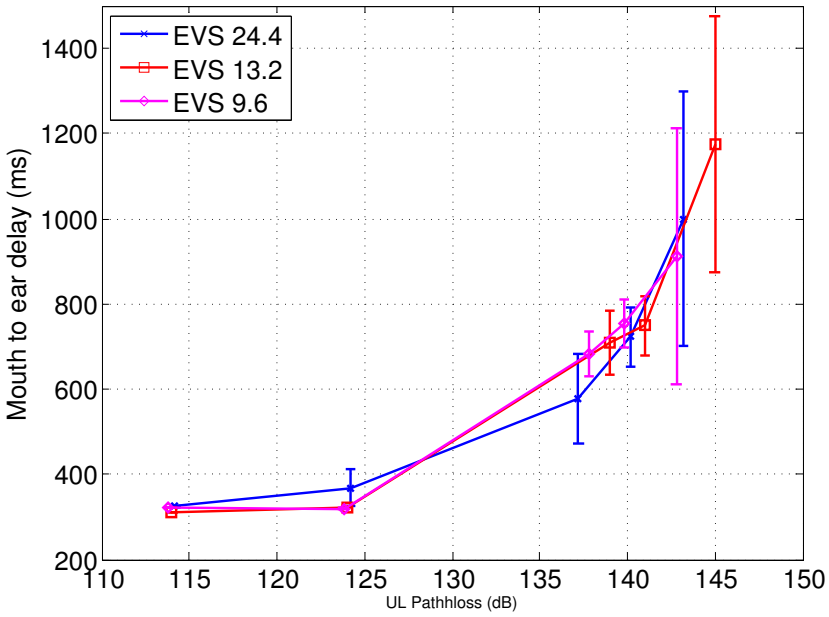


Figure 4.6: Mouth-to-ear (M2E) delay results as a function of pathloss.

to further reduce the size of confidence intervals (at the cost of increased test time). We had access to the test platform only for a short period of time, it was not possible to conduct more tests.

Figure 4.6 shows that the average mouth-to-ear delay of WebRTC voice calls increases sharply with increasing pathloss. The same behavior can be noticed for the three different codec bit rates. This indicates that bit rate has no significant influence on the evolution of mouth-to-ear delay when degrading network condition. As discussed later, one can interpret the delay curves by the fact that data transmission over LTE was configured with AM; in this case, degrading radio coverage translated mainly into increasing end-to-end delay and jitter.

4.2.5 Discussion

Analysis of RTP packet traces

We report the analysis conducted on the incoming and outgoing RTP packet traces captured by the network sniffer at the network interface of the two UEs.

This analysis showed that the packet loss rate is lesser than 0.1% in the worst case (near the coverage limit). This low packet loss rate can be explained by the AM data transfer in the LTE RLC protocol layer where unacknowledged packets are retransmitted noting that HARQ is also used at the physical layer.

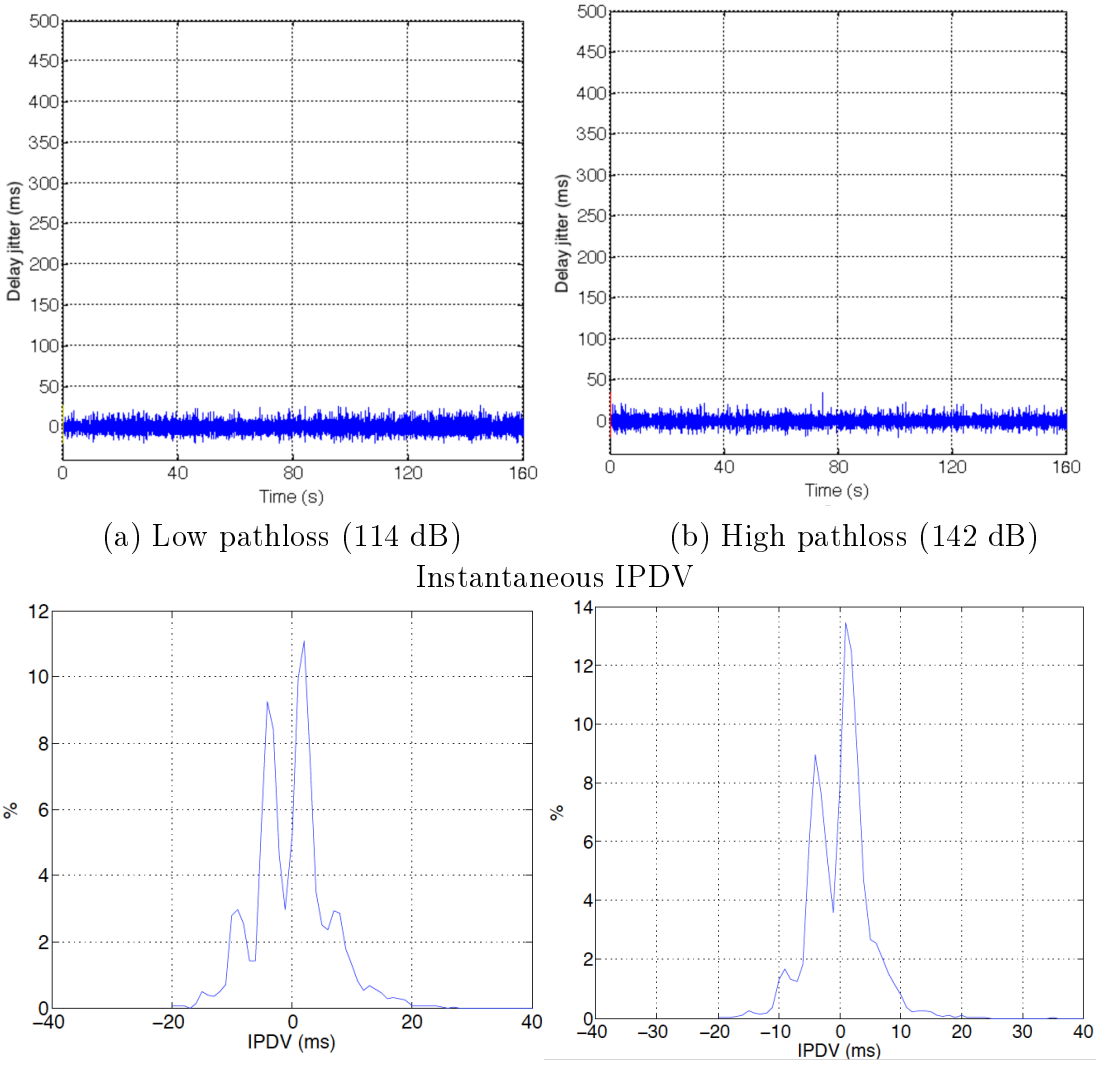
The negligible packet loss rate during transmission cannot fully explain the slight MOS degradation when increasing radio pathloss. MOS mainly decreases due to jitter buffer induced packet losses and delay adjustments when the WebRTC jitter buffer adapts to degrading network conditions. The jitter buffer in Chromium (NetEQ) is designed for a specific compromise between quality and delay [27]. It has a target delay derived from an Inter Arrival Time (IAT) histogram and a specific delay peak detector. WebRTc jitter buffer decisions on received audio can be normal decoding, expand, merge, or accelerate. Based on E2E delay presented in Figure 4.6, one can verify that the jitter buffer adapt its length to wait for late packets or to skip some packets causing quality drop as shown in Figure 4.5. This quality drop cannot be driven from transmission errors thanks to AM data transfer mode as explained before. These losses induced by jitter buffer operations are called late losses. In the following, we look for the main reasons behind jitter buffer length increase and therefore E2E delay increase.

Generally the jitter buffer compensates for packet delay variations (jitter). Therefore, we calculate the IPDV defined by:

$$IPDV(i) = D(i) - D(i-1), \tag{4.1}$$

where $D(i)$ denotes the one-way delay of the i^{th} packet from UE 1 network interface (Network card) to UE 2 network interface. In the absence of jitter (i.e. for a perfectly synchronized transmission), the IPDV is always 0. We present here the IPDV for one EVS bit rate coding (24.4 kbit/s) as an example. The other bit rates show similar behaviors.

Figures 4.7 (a) and 4.7 (b) show the instantaneous jitter in ms in sending direction in good and bad radio coverage, respectively. The jitter in both radio conditions is stable and around +/- 20 ms. Figures 4.7 (c) and 4.7 (d) present the distribution of jitter values presented in Figures 4.7.a and 4.7 (b) respectively. We barely notice any difference between the two graphs. IPDV presents here the delay variation observed at the mobile antenna of the sending phone (in the uplink of UE 1).



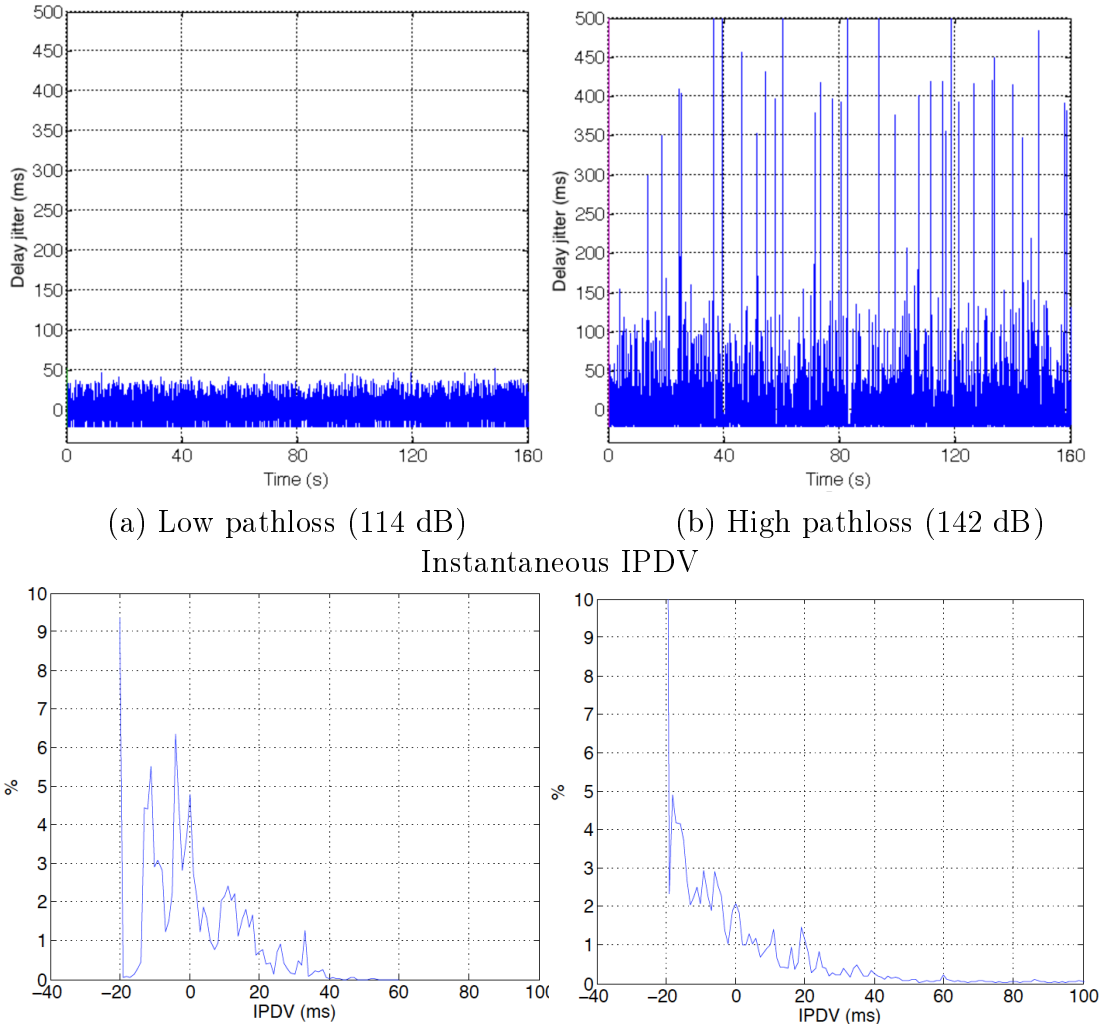
(a) Low pathloss (114 dB) (b) High pathloss (142 dB)
Instantaneous IPDV

(c) Low pathloss (114 dB) (d) High pathloss (142 dB)
IPDV distribution

Figure 4.7: IPDV in low and high pathloss in UL

We conclude that the jitter in sending (at the antenna point of UE 1) does not depend on radio conditions. The IPDV in sending is centered and bounded by ±20 ms. This means that speech frame are transmitted every 20 ms as required by real-time operation, but there are internal timing variations (processing, buffering, operating system scheduling, etc.).

Figures 4.8 (a) and 4.8 (b) show the transmission delay variation in ms observed at the receiving phone antenna (in the downlink of Mobile2) in good and bad radio coverage, respectively. Similar to the sending direction, we present Figures 4.8 (c) and 4.8 (d) for more visual illustration through the IPDV values distribution.



(a) Low pathloss (114 dB) (b) High pathloss (142 dB)

Instantaneous IPDV

(c) Low pathloss (114 dB) (d) High pathloss (142 dB)

IPDV distribution

Figure 4.8: IPDV in low and high pathloss in DL

For low radio pathloss (114 dB), few packets have an IPDV higher than 20 ms. Only 0.2% of packet had an IPDV higher than 40 ms. However, in high radio pathloss (142 dB), jitter is significantly higher. More than 7% of packets have an IPDV higher than 40 ms. Figure 4.8 (d) has a longer tail than Figure 4.8 (c). In high pathloss, IPDV is sometimes higher than 500 ms.

Note that the delay increase in degraded coverage condition cannot be explained only from jitter buffer adaptation; retransmissions (in particular in the LTE RLC layer) can also cause increased one-way transmission delay which impacts mouth-to-ear delay.

Comparison with VoLTE

Coverage tests were conducted with the setup described in section 4.2.2 for VoLTE calls and the AMR-WB codec at 23.85 kbit/s. At that time, we could not conduct VoLTE tests with EVS. Hence, the comparison to WebRTC with EVS was not yet possible at the time of writing. Results showed that coverage is better for VoLTE (with AMR-WB) than for WebRTC (with EVS) with a pathloss gain up around 3 5 dB, thanks to LTE radio optimizations (in particular TTI bundling and robust header compression). Speech quality (MOS) was found to be more stable for VoLTE calls. Moreover, mouth-to-ear delay increase with degraded conditions was smaller for VoLTE. Note that VoLTE is based on UM transmission over LTE, hence quality degradations at cell edge are typically more dominated by packet loss than by jitter effects.

4.3 Bandwidth limitation and packet loss impact on WebRTC speech quality

In the previous section, we evaluated WebRTC over an LTE radio link. In particular, we evaluated the impact of varying voice codec bit rate on the perceived quality when we degrade the radio channel. In this section, we focus on other types of network degradations such as bandwidth limitation and packet loss rate. We experimentally evaluate bandwidth limitation and packet loss influence on WebRTC speech quality and the eventual impact of varying voice codec bit rate in such network conditions. We use a simpler experimental set up than the previous one, including a WebRTC communication over Ethernet.

4.3.1 Experimental setup

The new test setup, figure 4.9 consists of two laptops operating WebRTC Chromium application. The application is operating with EVS. The EVS codec operates in SWB at 9.6, 13.2 et 24.4 kbit/s. Note that the native jitter buffer of the softphone is not the EVS JBM.

Bandwidth and packet loss conditions are emulated at the software level in one of the terminal (a computer running traffic shaping in the firewall as shown in Figure 4.9). The test signal is a French double sentence of 8 s repeated 25 times. We measure objective quality for each double sentence following also ITU-T Rec. P.863 using POLQA.

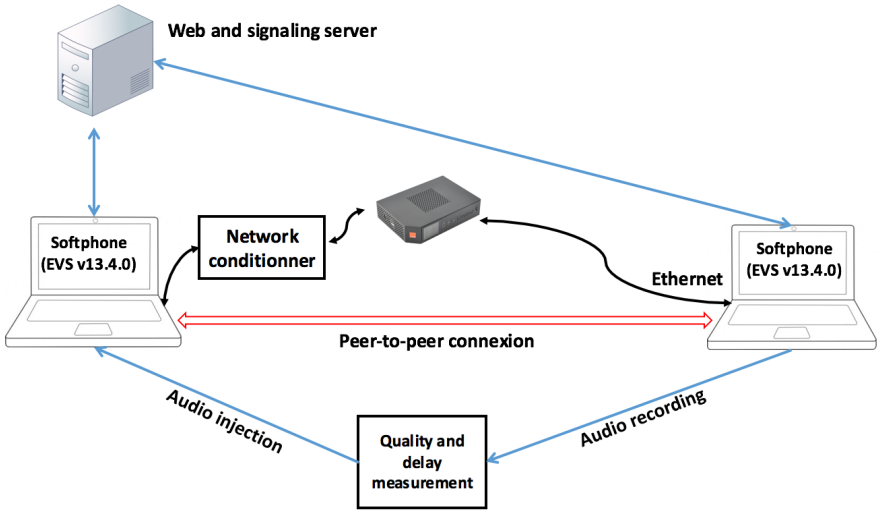


Figure 4.9: WebRTC over Ethernet experimental setup

4.3.2 Bandwidth limitation impact on WebRTC quality

We decrease the available bandwidth using the traffic shaping tool before each call. Figure 4.10 shows the evolution of WebRTC speech quality in terms of average MOS scores for three different bit rates: 9.6, 13.2 and 24.4 kbit/s as function of the bandwidth. The quality decreases slightly once no sufficient bandwidth is available. EVS 24.4 kbit/s drops faster than the other bit rates. Higher bit rates need more resources. Therefore, decreasing bit rate in case of limited bandwidth allows to maintain the communication in lower available bandwidth. We conclude that bandwidth can be used as a bit rate adaptation trigger. Furthermore, available bandwidth can be measured using only information about the received RTP packets such as sending and receiving time, and packet size.

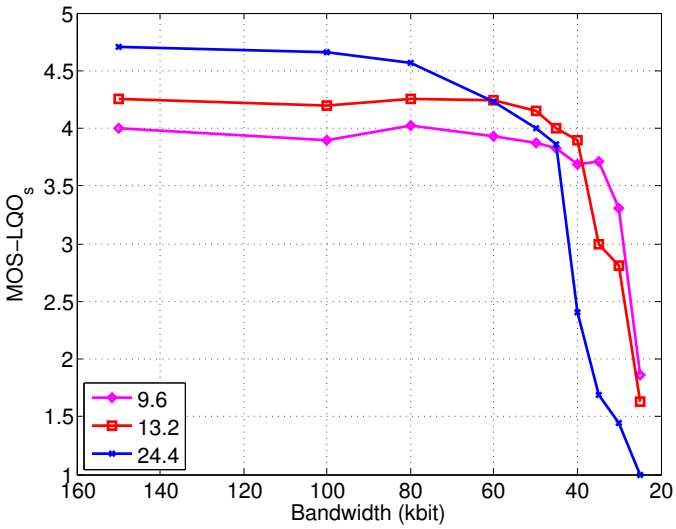


Figure 4.10: MOS as a function of available bandwidth

4.3.3 Packet loss impact on WebRTC quality

Packet losses are also applied using the traffic shaping tool following a random distribution, an i.i.d Bernoulli variable. MOS as function of PLR is presented in figure 4.11 for the three used EVS bit rates. The presented results show that for the three bit rates MOS decreases simultaneously. Varying bit rate has no impact in this case. Packet loss rate was usually used as a trigger for bit rate adaptation indicating for network congestion. However, if packet loss is due to

another problem such as transmission errors, decreasing the bit rate will not have any impact on quality.

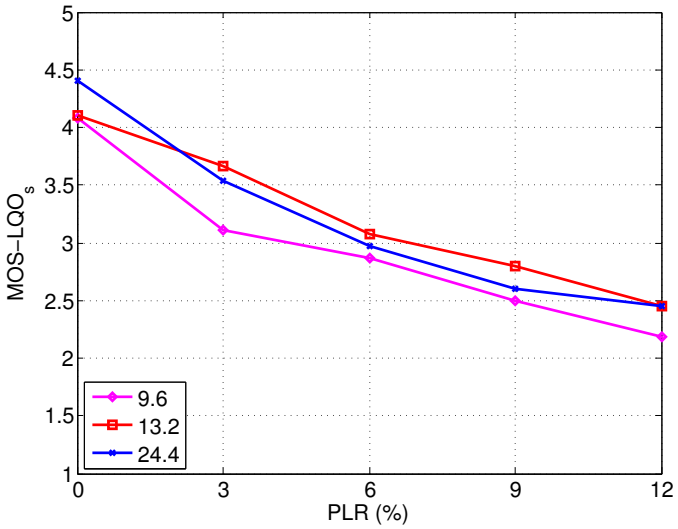


Figure 4.11: MOS as a function of packet loss rate

Summary

The main outcome of this chapter is WebRTC performance evaluation in different network conditions with focus on speech quality. We provided results of WebRTC speech quality as a function of LTE radio signal level, available bandwidth and packet loss rate.

According to the first test results concerning WebRTC over LTE, we conclude that with AM, degraded radio coverage translates into an increased end-to-end delay and jitter and virtually no packet loss. The quality degradation of WebRTC over LTE in radio coverage limit can be explained by WebRTC jitter buffer adaptations where late packets with delay that cannot be compensated by the jitter buffer are discarded. Bit rate variation impact in radio coverage tests needs more investigation to conclude about the benefit from it.

Thanks to the second test, we provided results about WebRTC speech quality as function of available bandwidth and packet loss rate using a real communi-

cation between two WebRTC clients over Ethernet connexion. The outcome of the second test is that bit rate adaptation may enhance speech quality in case of network limited bandwidth. Low bit rates maintain a higher quality in low bandwidth than high bit rates. The impact of packet loss rate is the same for the different bit rates. In this case we need more information about the cause of packet loss. If it is congestion and bandwidth limitation, bit rate adaptation can be a solution. However, in case of packet loss due to transmission errors, other type of adaptations may be used such as redundancy to enhance voice codec resiliency, which we will study in the next chapter.

Chapter 5

Application-Layer Redundancy Impact On Speech Quality

Contents

Introduction

Packet loss is one of the main QoE degradation factors of VoIP services including VoLTE in poor radio condition (e.g., in cell edge). Many mechanisms have been proposed to deal with packet loss and enhance perceived quality by using redundancy and packet loss concealment techniques. In this chapter, we briefly review some of these mechanisms. In particular, we review a special redundancy mechanism defined for EVS to be used in VoLTE services called channel aware mode (CAM). We also present a different type of redundancy mechanism which we call it application-layer redundancy. This type of redundancy has been defined in the literature for other voice codecs. However, it has been never tested for EVS. In this work, we implement application-layer redundancy for the EVS codec. We conduct multiple subjective and objective tests to evaluate the impact of application-layer redundancy on perceived quality and compare its performance to CAM performance. We also discuss some possible RTP/RTCP signaling methods to trigger the use of application-layer redundancy.

This work was motivated by eVoLP study within the 3GPP, see Chapter 1 for more details. We recall here the main objectives of eVoLP: 1) investigate guidelines or requirements to ensure that VoLTE clients adapt to the most robust codec modes, study performance; results for different conditions and adaptation procedures; 2) study how terminals can indicate at setup their ability to send adaptation triggers to robust modes; 3) evaluate the impact of proprietary client implementations of PLC and JBM. In this chapter, we focus on the first two objectives. We quantify the performance of the EVS codec with redundancy in degraded channel with different packet loss rates; we also briefly review media signaling methods (based on RTP or RTCP) that can be used to trigger the use of application-layer redundancy.

5.1 Related work

Many approaches have been proposed to address and adapt to packet losses in speech and audio coding, and we classify them as sender/encoder vs. receiver/decoder-based methods. A method frequently used in networks to correct errors is an end-to-end retransmission (ARQ mechanism as we have seen in Chapter 4 for data retransmission over LTE). However, this cannot be used for VoLTE because latency and system constraints.

Encoder-based methods typically consist in adding redundancy [97] or limiting the use of memory (prediction) [98][16, Sec. 2.1.6][99]. Two redundant coding approaches are reviewed in [97]: Multiple Description Coding (MDC) and FEC. MDC consists in encoding the signal in complementary descriptions that are sent separately. If some descriptions are lost, a coarser reconstruction is obtained. FEC is based on channel-coding principles. Some FEC variants apply error-correcting codes such as Reed-Solomon codes at the bitstream level [21, Chap. 9]. In this work we refer to 100% application-layer redundancy as an FEC variant where the bitstream of a given speech frame is fully repeated in a subsequent packet [100]. Application-layer redundancy is defined in more details in [28, Sec. 9.2][45] and in section 5.2.1. The FEC principle can also be applied to the coded parameters (signal class, energy) to minimize the bit rate penalty of redundancy [100]. For instance, the ITU-T G.729.1 codec sends some frame coded parameters with few bits to guide packet loss concealment [101]. Partial redundancy coding is also used in the LBRR (Low-Bit-Rate Redundancy) of the OPUS codec [16, Sec. 2.1.7] or the Channel-Aware Mode (CAM) of the EVS codec [102] see section 5.2.2 for more details on CAM. It can be noted that some sender-based adaptation strategies to packet losses may rely on rate/congestion control mechanisms, for instance by reducing codec bit rate or even packet rate to deal with insufficient throughput.

Decoder-based methods are mainly based on PLC techniques to fill and recover from missing frames [103]. Examples are given in [104, 105, 106]. It can be noted that in VoIP PLC may be integrated with JBM [27], and concealment and recovery may be implemented by JBM expand and merge operations at the reconstructed signal level.

5.2 Redundancy mechanisms for the EVS codec

We review here two approaches to use redundancy for the EVS codec: channel aware mode (CAM) and 100% application-layer redundancy.

5.2.1 Application-layer redundancy

Application-layer redundancy can be used with any codec [28, sec. 9.2]. In normal operation, when redundancy is not used, each RTP packet transports

a single codec frame. When application-layer redundancy is used, a packet transports the bitstream of the current frame (N) as well as the bitstream of one or several past (redundant) frames ($N - k$), where $k > 0$.

We focus in this chapter on 100% redundancy on single frames, where only one redundant frame is added. However, in principle it is possible to use more redundancy (e.g. 200% with two redundant frames per packet) and several frames per packet (e.g. two frames per packet).

The RTP payload format of codecs such as 3GPP AMR, AMR-WB [39] or EVS [17, Annex A] includes a 'max-red' media type parameter, which restricts the maximal time interval (offset) for redundancy. It is noted in [28, sec. 9.2] that this type of redundancy may not be an appropriate solution in scenarios with packet losses due to limited throughput or congestion.

As shown in figure 5.1, in the 100% redundancy case, an offset $k = 1$ implies that an RTP packet includes the RTP header followed by an optional table of content (ToC) and frames N and $N - 1$; when $k = 2$ a dummy frame $N - 1$ (referred to as NO_DATA) has to be inserted between frames N and $N - 2$, typically this insertion is done implicitly in the ToC part [39]. The bit rate overhead depends on the redundancy level (e.g. 100%) and the ToC length.

In principle end-to-end delay may be increased if redundancy is used; in practice for small offset values (e.g $k = 1$ or 2) and assuming a sufficient jitter buffer depth, the redundant frame may often be available if it is received as a future packet before decoding and playing out the current frame.

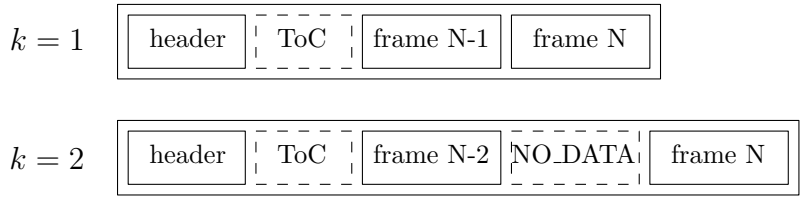


Figure 5.1: Example structure of an RTP packet with 100% application-layer redundancy for offset $k = 1$ or 2: RTP header followed by the RTP payload including an optional table of content (ToC) and dummy (NO_DATA) frame.

Performance results for 100% application-layer redundancy with 3GPP AMR are reported in [100, 107].

5.2.2 Channel-Aware Mode (CAM) of EVS

The EVS-CAM [102], [17, section 5] is a partial redundancy mode supported at a single bit rate (13.2 kbit/s) in WB and in SWB. The redundant data is defined in the form of partially coded frame $N-k$ embedded within the bit stream of the current frame N. The offset is restricted to $k = 2, 3, 5$ or 7. The high offsets ($k = 5$ or 7) may be used to deal with long bursts of losses. However, they imply significantly higher receiver delay.

A key feature of EVS-CAM is that it keeps a fixed bit rate (13.2 kbit/s) at the application level, so the activation of CAM is transparent to the network. When CAM is activated, the partial copy of the past frame $N-k$ is coded using about 3 kbit/s, therefore the remaining bit rate budget to code the current frame N is reduced to about 10 kbit/s. The performance of CAM has been reported in [108] and [109]. In clean channel conditions (i.e. no packet loss) the intrinsic quality of CAM is close to the 9.6 kbit/s mode of EVS. However, CAM is significantly better than the regular EVS modes at 9.6 or 13.2 kbit/s for packet loss rate greater than 3% - see also results reported in section 5.4.

The EVS RTP payload format [17, Annex A] defines a media type parameter 'evs-ch-aw' to control the use of EVS-CAM. The signaling methods to trigger EVS-CAM relies on RTP CMR or RTCP-APP [28].

5.3 Experimental setup and EVS codec modifications

We test four EVS bit rates (9.6, 13.2, 16.4 and 24.4 kbit/s) with or without application-layer redundancy, together with EVS-CAM at 13.2 kbit/s. For all tests, we use EVS with DTX activated.

5.3.1 Modifications to the EVS codec

We describe here how the source code of the EVS codec is modified to simulate application-layer redundancy. The EVS bitstream is compliant with the serial

bitstream format in ITU-T G.192 [110, App. I.2]. This format is convenient to simulate transmission of frames over a synchronized noisy channel between the encoder and decoder.

In this work, we only modify the G.192 bitstream formatting in the EVS encoder part; the actual EVS encoding algorithm is not changed. We add a buffer of encoded frames (bitstreams) outside the main EVS encoding loop to produce an extended G.192 bitstream.

At the decoder side, a receiving buffer is added as a pre-processing step to bitstream decoding, with an extra decoder delay to allow detecting if the (lost) current frame is available as a redundant frame in a future packet at a given offset.

5.3.2 Generation of processed audio samples

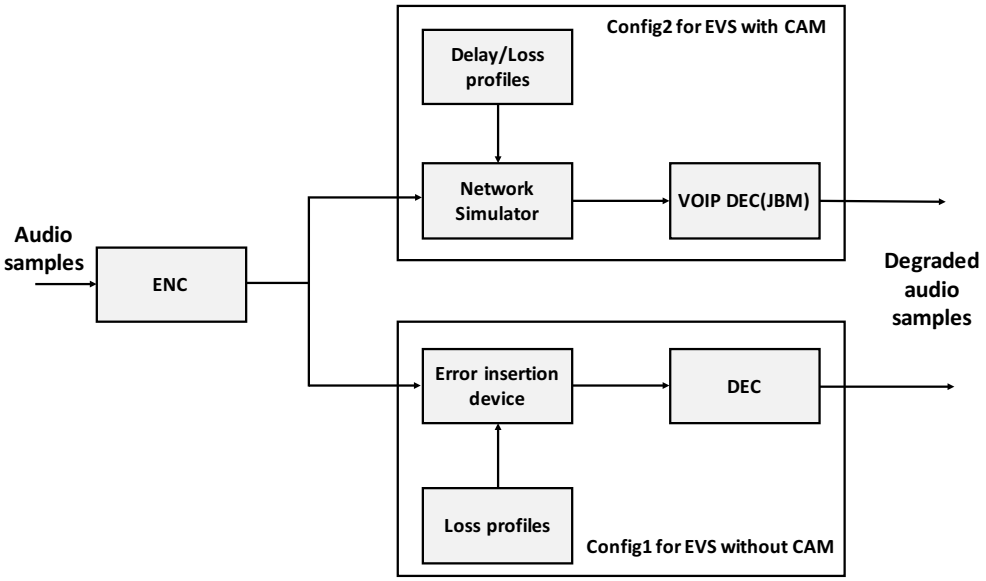


Figure 5.2: Audio samples processing setup

To generate the audio samples required for subjective and objective tests, we reuse EVS qualification scripts [111] to automate the encoding and decoding

tasks. The original clean speech samples in French (16-bit linear PCM) sampled at 48 kHz are high-pass filtered, down-sampled to 32 kHz, normalized to -26 dBov and encoded with EVS. When application-layer redundancy is used, we refer to the $2 \times X$ bit-rate where X is the regular EVS mode and we use the modified EVS codec described in section 5.3.1. We apply a redundancy offset $k = 2$.

We use the ITU-T 'gen-patt' tool to generate loss profiles. Each channel profile contained one entry per packet indicating whether the packet is received or not. These loss profiles are applied to the bitstream by the ITU-T 'eid-xor' tool which stands for error insertion device as shown in figure 5.2 in the first configuration.

For EVS-CAM conditions, we use a network simulator developed in 3GPP [112] to simulate VoIP transmission using delay/loss profiles as shown in Figure 5.2 in the second configuration. We use a CAM offset of 2 with default. The EVS decoder operated in VoIP mode in conjunction with the EVS JBM algorithm. The loss-only profiles generated by 'gen-patt' are converted to delay/loss profiles with a fixed delay and identical packet loss distribution.

5.3.3 Packet loss models

We use three different loss profiles. Each loss profile is characterized with a different loss distribution. This allows us to evaluate the performance of application layer redundancy in different types of lossy channels. The channel is presented by three known loss models, which are Bernoulli, Gilbert and Bell-Core loss models. We present here the theoretical distribution of losses of each used model.

Bernoulli Channel Model

The first model consists of random packet loss profile following an i.i.d Bernoulli variables. We define p the probability of having a packet loss. Thus, the probability of receiving a packet (without error) is $1 - p$. Random loss profiles may contains some losses in form of bursts. A loss burst is a number of successive lost packets. The distribution of losses can be presented by the probability of having a loss burst with length n_L:

$$\mathbb{P}\left(n_L = k | n_L > 0\right) = p^{k-1}(1-p), \qquad \text{for } k \geq 1 \tag{5.1}$$

The loss distribution depends on the Bernoulli probability distribution. Therefore the distribution of losses is equivalent for all used packet loss rate for the same Bernoulli probability distribution. Figure 5.5(a) presents the theoretical probabilities of loss bursts at 6% of packet loss rate.

Gilbert Channel Model

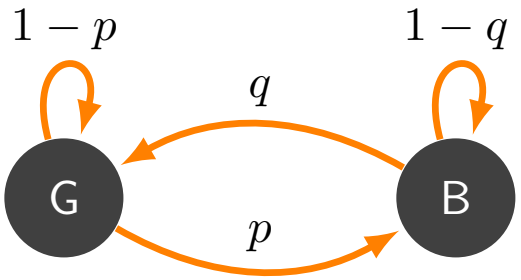


Figure 5.3: Gilbert channel model [4]

The second model is the Gilbert channel model [113]. It consists of 2-state discrete time Markov chain, see Figure 5.3, one state without errors (state G for good) and the other one with errors (state B for bad); we define the following transition probabilities shown in Figure 5.3 :

- $\mathbb{P}(G \to B) = p$
- $\mathbb{P}(B \to G) = q$
- $\mathbb{P}(G \to G) = 1 - p$
- $\mathbb{P}(B \to B) = 1 - q$

We assume there is a relative stability of the chain. In other words, the probability to stay in the same state is higher than or equal to the probability to leave the state. Hence, $p \leq \frac{1}{2}$, $q \leq \frac{1}{2}$. We define a parameter γ as follows:

$$\gamma = 1 - p - q, \tag{5.2}$$

where $\gamma \in [0, 1]$. The probability to have an error in state G and state B are P_G and P_B, respectively. The steady state equation is:

$$\Pi_G = (1 - p)\Pi_G + q\Pi_B. \tag{5.3}$$

Knowing that $\Pi_G + \Pi_B = 1$ and based on equation 5.3:

$$\Pi_G = \frac{q}{q+p}, \tag{5.4}$$

and

$$\Pi_B = \frac{p}{q+p}. \tag{5.5}$$

The average loss probability is:

$$P_L = \Pi_G P_{G,L} + \Pi_B P_{B,L}, \tag{5.6}$$

where P_L is the average loss rate probability, $P_{G,L}$ and $P_{B,L}$ are loss probabilities in state G and B, respectively. Our work is based on ITU-T G.192-STL [4] where $P_{G,L}$ and $P_{B,L}$ are fixed to 0 and 0.5. Thus, Equation 5.6 becomes:

$$P_L = \frac{p}{2(p+q)}. \tag{5.7}$$

Based on equation 5.2, 5.7 becomes:

$$P_L = \frac{p}{2(1-\gamma)}. \tag{5.8}$$

A loss of successive $n_L = k$ packets means that the Markov chain is in state B and there is packet losses. Thus, the Markov chain stays in state B k-1 times with a loss each time. Finally, the burst is ended by receiving a packet without errors. Thus, we have for $k \geq 1$:

$$\begin{aligned}
\mathbb{P}\left(n_L = k / n_L > 0\right) &= \frac{\mathbb{P}\left(n_L = k \cap n_L > 0\right)}{\mathbb{P}(n_L > 0)} && (5.9)\\
&= \frac{\mathbb{P}\left(n_L = k\right)}{\mathbb{P}(n_L > 0)} && (5.10)\\
&= \frac{\Pi_B P_{B,L}((1-q)P_{B,L})^{(k-1)}[q + (1-q)P_{B,L}]}{\Pi_B P_{B,L}} && (5.11)\\
&= [1 - (\gamma + p)(1 - P_{B,L})](\gamma + p)^{k-1} P_{B,L}^{k-1}. && (5.12)
\end{aligned}$$

We have $P_{B,L} = 0.5$. Thus, equation 5.12 becomes:

$$\mathbb{P}\left(n_L = k / n_L > 0\right) = 0.5^{k-1}[1 - \frac{1}{2}(\gamma + p)](\gamma + p)^{k-1}, \qquad \text{for } k \geq 1. \tag{5.13}$$

Equation 5.13 shows that the probability of having a loss burst of a given length depends only on transition probabilities of the markov chain. These probabilities

depend on the parameter γ as we have seen in Equation 5.2. Therefore, the distribution of bursts depends on γ. Case $\gamma \approx 0$ corresponds to random losses. Case $\gamma = 1 - \varepsilon$ where ε>0, is the most extreme case of bursty channel.

Our objective is to compare the performance of application layer redundancy for different error models but for the same average packet loss. Hence, P_L is an input parameter and we tune p to get the objective P_L. We fix $\gamma = 0.5$ to have more bursts than in the random case which results in $p = P_L$ and $q = \frac{1}{2} - P_L$. Figure 5.5(b) shows the burst distribution using the Gilbert model at 6% of packet loss rate and $\gamma = 0.5$. We can see that the bursts of losses are more frequent and higher than in the random model.

Bell-core channel model

The third model is Bell-Core model [114], which is presented by an N-state discrete time Markov chain as show in Figure 5.4. Each state j presents a burst of losses of length j. State 0 corresponds to a received frame without errors. A transition from state j to state $j + 1$ indicates that packet $j + 1$ is also lost as the previous j packets with a transition probability p_j. A transition from state j to state 0 indicates that a burst of packet loss with length j is followed by a received packet with a transition probability $1 - p_j$.

The probability of receiving a packet without error is Π_0 which is the steady state probability of state 0. Thus, the loss probability P_L is:

$$P_L = 1 - \Pi_0. \tag{5.14}$$

The balance equations of the markov chain are:

$$\Pi_j = p_j \Pi_{j\text{-}1} \qquad \text{for } 1 \leq j \leq N - 1, \tag{5.15}$$

which results in

$$\Pi_j = \prod_{k=0}^{j-1} p_k \Pi_0 \qquad \text{for } 1 \leq j \leq N - 1. \tag{5.16}$$

In the steady state, we have:

$$\sum_{j=0}^{N-1} \Pi_j = 1. \tag{5.17}$$

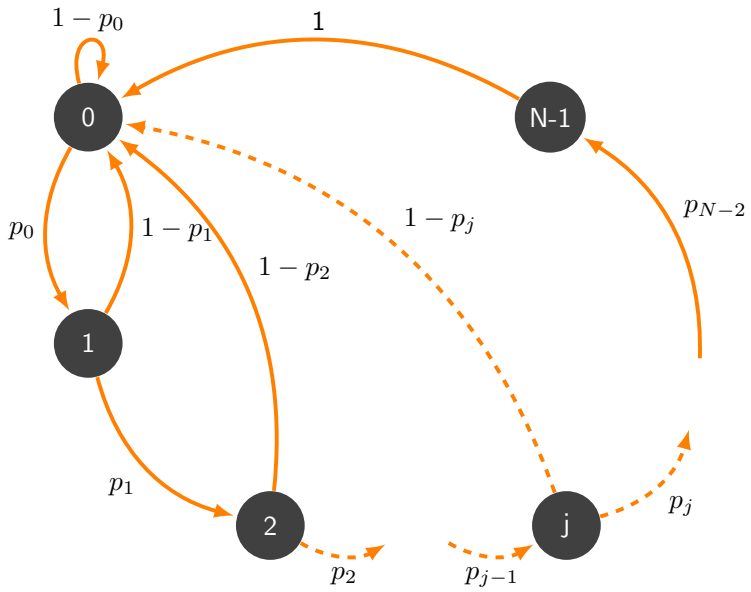


Figure 5.4: N-state Markov chain (Bell-Core transitions graph)

From equation 5.16 and 5.17, we have:

$$\Pi_0 = \frac{1}{1 + \sum_{j=0}^{N-2}(\prod_{k=0}^{j} p_k)} \tag{5.18}$$

By fixing $\alpha = 1 + \sum_{j=1}^{N-2}(\prod_{k=1}^{j} p_k)$, which is independent of p_0, we have:

$$\Pi_0 = \frac{1}{1 + \alpha p_0}. \tag{5.19}$$

To calculate the distribution of bursts in the Bell-core model, we calculate the probability of burst losses of length $n_L = k$ where $k \geq 1$:

$$\mathbb{P}\left(n_L = k | n_L > 0\right) = \frac{\mathbb{P}\left(n_L = k\right)}{\mathbb{P}(n_L > 0)}. \tag{5.20}$$

A loss of successive $n_L = k$ packets means that the Markov chain enters successively state 0 to k and then goes back to state 0. Therefore, we have:

$$\mathbb{P}\left(n_L = k\right) = \frac{\Pi_k(1 - p_k)}{\Pi_0}. \tag{5.21}$$

From Equations 5.14 and 5.16, we have:

p_0	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
variable	0.85	0.825	0.8	0.775	0.75	0.725	0.7	0.6	0.45

Table 5.1: Used Bellcore transition probabilities for the simulator(N=11)

$$\mathbb{P}\left(n_L = k\right) = (1 - p_k) \prod_{i=0}^{k-1} p_i. \tag{5.22}$$

Based on Equation 5.22, we can write Equation 5.20 as follows:

$$\mathbb{P}\left(n_L = k | n_L > 0\right) = (1 - p_k) \prod_{i=0}^{k-1} p_i = (1 - p_k) \prod_{i=1}^{k-1} p_i. \tag{5.23}$$

$\mathbb{P}\left(n_L = k | n_L > 0\right)$ is a function of the transition probabilities p_j for $1 \leq j \leq N-1$ and not of p_0.

For this work, we use the Bell-Core model as proposed in ITU-T Rec. G.191 STL-2009 Manual [4], with a maximum loss burst length of 10 packets, $N = 11$. We use the presented transition probabilities p_j for $1 \leq j \leq N-1$ in Table 5.1. To vary the target packet loss rate P_L, only p_0 is changed. The equation of p_0 can be derived from equation 5.14 and 5.19 as follows:

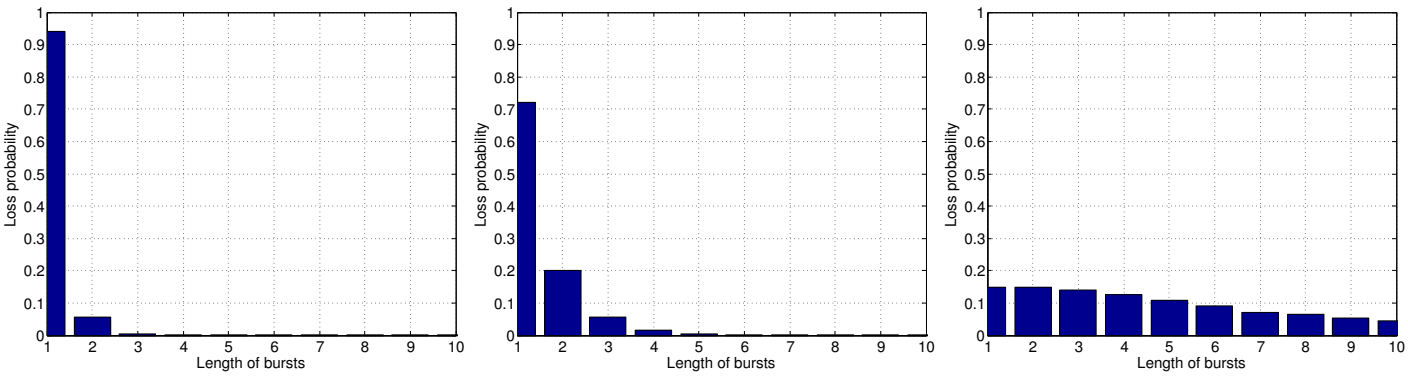
$$p_0 = \frac{P_L}{\alpha(1 - P_L)}. \tag{5.24}$$

The distribution of bursts in the used Bell-Core model are presented in Figure 5.5(c) for packet loss rate of 6% after generating the loss profiles. The Bell-Core model model presents more losses in form of bursts than the Gilbert model, see Figure 5.5(c) in comparison with Figure 5.5(b). Bell-core model is an extreme model where the losses are mainly in bursts.

5.3.4 Subjective test (P.800 ACR)

We use the P.800 ACR methodology [56] to allow comparisons with P.863 predictions. In ACR tests, groups of listeners evaluate series of processed audio files using a five-category scale. The experimenter allocates the following categories to scores: Excellent=5, Good=4, Fair=3, Poor=2, Bad=1. We recruited 48 naive listeners for the subjective test.

Table 5.3.4 describes the settings used for the subjective test. This resulted in 1152 processed sequences with 24 blocks for 6 panels (of 8 listeners), 4 blocks per panel. Each block contained 48 conditions and 4 talkers equally. The number of votes per condition is $4 \times 48 = 192$. The overall listening/scoring duration is around 42 minutes for each subject ($192 \times 13s$). The list of conditions and randomizations can be found in [115]. The statistical analysis was based on independent-group t tests.



(a) Random model (b) Gilbert model (c)Bell-core model

Figure 5.5: Loss bursts probabilities for the three used models at 6% of loss rate

Table 5.2: Subjective test plan settings.

Voice codec	EVS [116] with DTX on
Rating scale	ACR [56]
Listening	73 dB SPL, naive listeners, 6 panels of 8 listeners
Loss profiles	static with no jitter, only random losses are applied 0%, 3%, 6%, 9%, and 12 %
Coding mode	9.6, 13.2, 16.4, and 24.4 kbit/s 2×9.6, 2×13.2, 2×16.4, 2×24.4 kbit/s with offset=2 13.2 kbit/s CAM with offset=2
Calibration	P.50 MNRU: 10, 16, 22, 28, 34, and 40 dB
Talkers	4 (two males and two females)
Samples	6 clean speech samples (8s double sentences) per talker

5.3.5 Objective Test (P.863)

The objective quality evaluation used ITU-T Rec. P.863 [80] using POLQA (v2.4) in SWB mode. MOS-LQO$_s$ scores are computed by providing the audio sequences (SWB) and the degraded ones to the measurement tool POLQA. For the random channel, the audio sequences used in the objective test were the same as in the subjective test.

5.4 Tests Results

5.4.1 Results for Random Channel (no channel memory)

Figure 5.6 shows a bar chart with the average subjective scores, including 95% confidence intervals (in the order of ±0.1 MOS). Figures 5.7 (a) and 5.7 (b) present subjective and objective test results, as a function of packet loss rate. Note that Figure 5.7 (a) is just an alternative representation of the same scores as in Figure 5.6, and confidence intervals are not shown in Figure 5.7 (a) to improve readability.

Subjective and objective results show similar trends, however P.863 predic-

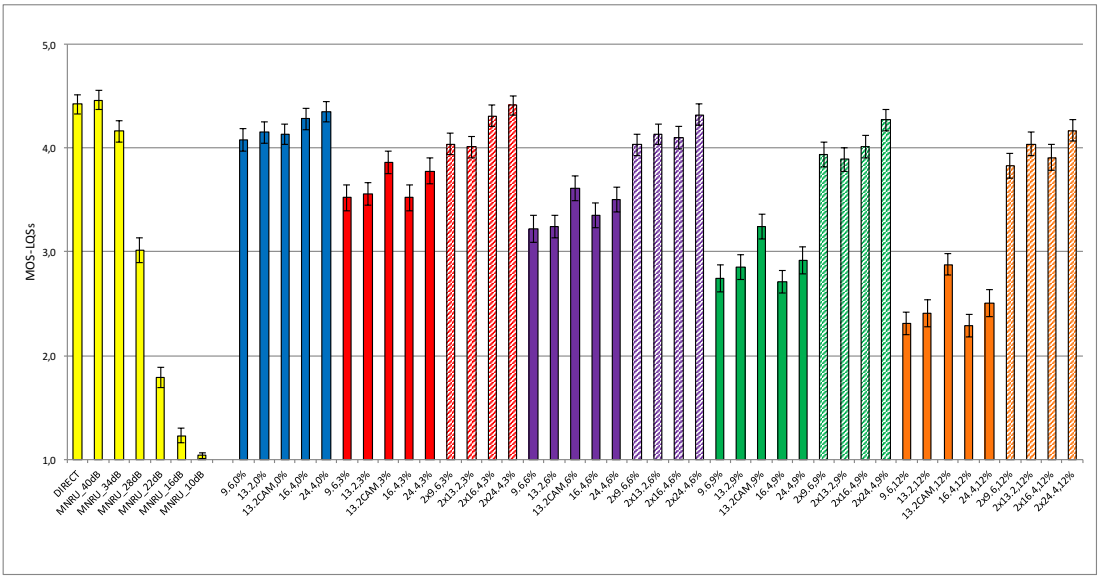
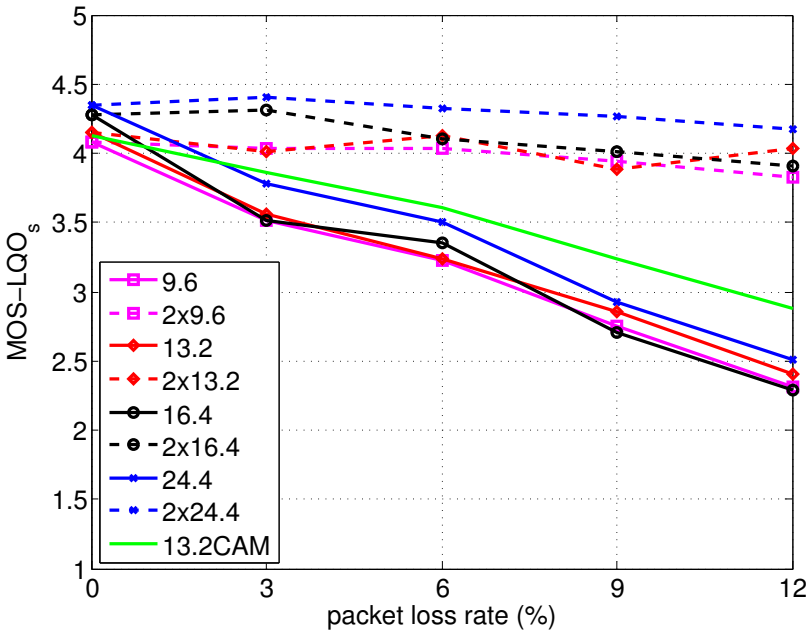
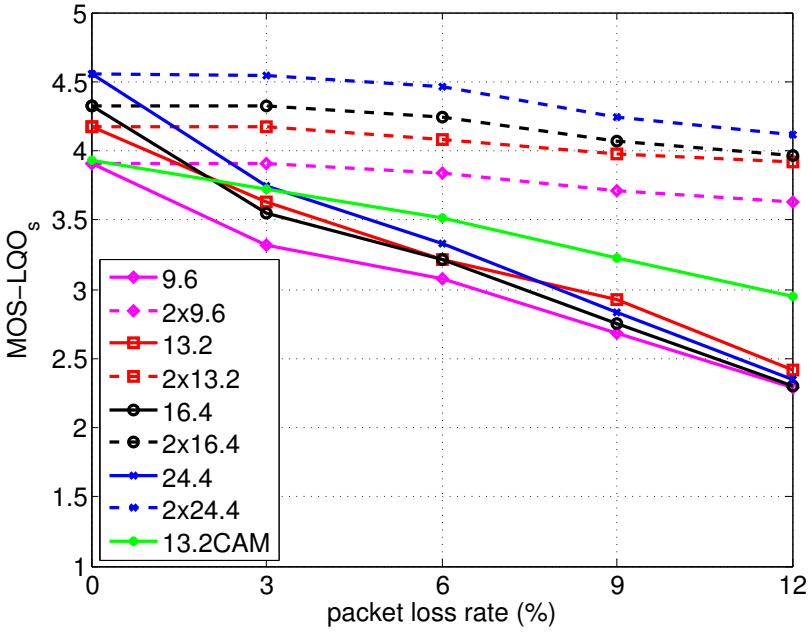


Figure 5.6: Subjective test results (random losses).



(a) Subjective results ($MOS - LQS_s$)



(b) Objective results ($MOS - LQO_s$)

Figure 5.7: EVS codec performance (random losses), as a function of packet loss rate and operation mode (bit-rate, redundancy).

tions emphasized the intrinsic quality difference between EVS bit rates; on can observe that subjective scores are more compressed, and this may be explained by that fact that subjects may have focused their assessment on artifacts related to packet losses more than on the impact of codec rate. Assuming VoLTE bearers configured for EVS up to 24.4 kbit/s, application layer redundancy at 2×9.6 kbit/s gives the best performance for packet loss rate $\geq 3\%$ for all compatible EVS operation modes. The MOS score stays close to 4 even at a high packet loss rate (12%); this could be predicted, given that application-layer redundancy in the random channel case reduces packet loss rate from p to p^2; the MOS score at 12 % packet loss rate with 2×9.6 kbit/s is theoretical the same as the MOS score at 1.44% packet loss rate for 9.6 kbit/s.

Results also confirm that EVS-CAM at 13.2 kbit/s is significantly better than EVS from 9.6 to 16.4 kbit/s for packet loss rate greater than or equal to 3%. However, its performance is significantly worse than 2×9.6 kbit/s for packet loss rate greater than or equal to 3%. Note that application-layer redundancy used an extra decoder delay of 40 ms (due to the offset $k = 2$). As discussed in section 5.2.2 there may be no impact on receiver delay when using EVS-CAM with offset 2 (assuming a minimal jitter buffer depth of 2 extra frames).

5.4.2 Results for Gilbert channel model

Objective test results for the Gilbert model are provided in Figure 5.8. Compared to Figure 5.7 (b), we can see that in a bursty channel the performance decreases faster at a given packet loss rate and the use of redundancy is less efficient. Indeed, due to longer burst of losses, the redundant frames can only be exploited to compensate for losses at the end of a burst.

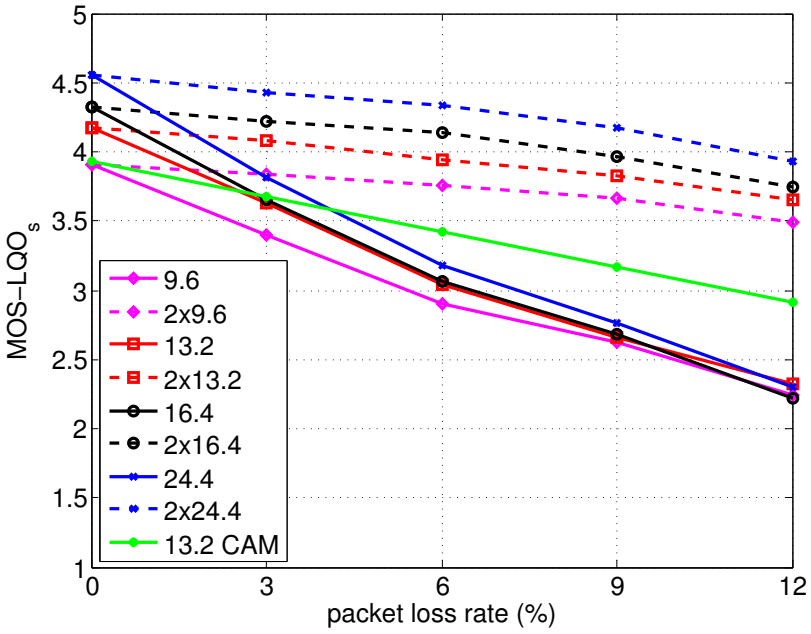


Figure 5.8: Objective results ($MOS - LQO_s$) with Gilbert channel model, as a function of packet loss rate and operation mode (bit-rate, redundancy).

5.4.3 Results for Bell-Core Channel Model

Objective test results for the Bell-Core model are provided in Figure 5.9. Compared to Figure 5.8, we can see that the performances decreases even faster than within Gilbert model at a given packet loss rate. In fact, the percentage and the length of burst in Bell-Core channel model are higher than the in Gilbert channel model as shown in Figure 5.5 (c) in comparison with Figure 5.5 (d) which explains the limited impact of application-layer redundancy with the last channel model.

The more and the longer the bursts are, the lesser the impact of application-layer redundancy is.

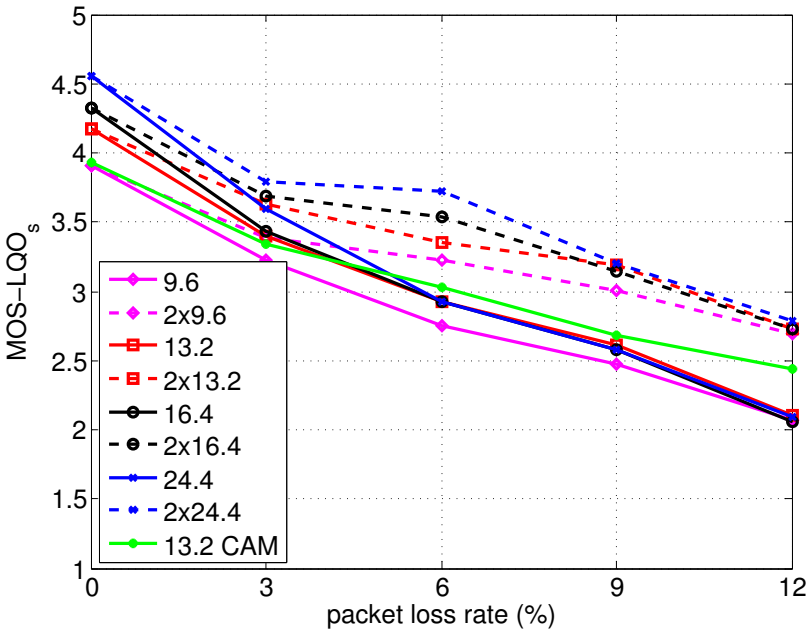


Figure 5.9: Objective results ($MOS - LQO_s$) with Bell-Core channel model, as a function of packet loss rate and operation mode (bit-rate, redundancy).

5.5 Proposal of a method to request for application-layer redundancy

In Section 5.4, we showed that application layer redundancy enhances the codec resiliency in case of lossy channel. In some cases, where packet loss is lower than 3%, application-layer redundancy is useless. In some cases, it can be just a waste of bandwidth. Therefore, it must be activated according to this threshold.

In a bidirectional communication, the receiver is the one who can detect packet loss and network channel degradation. The sender can decide to use application layer redundancy based on the receiver feedback. However, a simpler way is that the receiver decides or not that the sender activates application layer redundancy

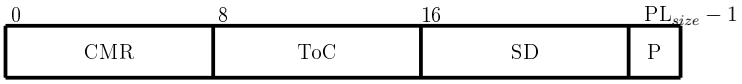


Figure 5.10: EVS RTP payload in header-full format

to enhance codec resiliency. Only the decision must be sent back from the receiver to the sender (from decoder to encoder). In general, this feedback or adaptation request can be sent through several methods. The media handling of VoLTE specified in [28] defines two methods to send adaptation requests for speech: RTP CMR and application-specific RTCP (RTCP-APP).

We recall here the principal of RTP-CMR and RTCP-APP that we have seen in Section 2.4. RTP CMR consists of sending adaptation requests in-band within the codec payload. It is defined for AMR and AMR-WB in [39] and for EVS in [17, Annex A]. Several CMR codes are left reserved or unused and there are no CMR codes specified to request application-layer redundancy. RTCP-APP is a form of out-of-band feedback [1] used for speech adaptation in [28, clause 10.2.1] to more general signal adaptation requests than CMR, in particular it can be used to request the activation of application-layer redundancy.

VoLTE deployments are based on RTP-AVP profile, presented in chapter 2.4.3. Within this minimum profile, the use of RTCP is very limited to SR and RR reports which contain only information about network condition to maintain the connection. Therefore, sender based adaptation is very constrained by the limited and irregular feedback which has only few basic quality indicators. RTCP-APP can be a solution. However, we recall also that the use of the RTP-AVPF profile is forbidden for speech in the VoLTE profile specified in GSMA IR.92 [117]. Therefore, RTCP-APP cannot be used in VoLTE as it requires RTP-AVPF profile.

There are two possible options that may be used in VoLTE to request the use of application-layer redundancy. The first option consists in using the RTP CMR codes that are currently left 'reserved' or 'not used', if this is negotiated at call setup with an appropriate SDP parameter. The CMR can be added at the end of RTP header as shown in Figure 5.11.

We present here a use case for EVS codec which RTP payload format is

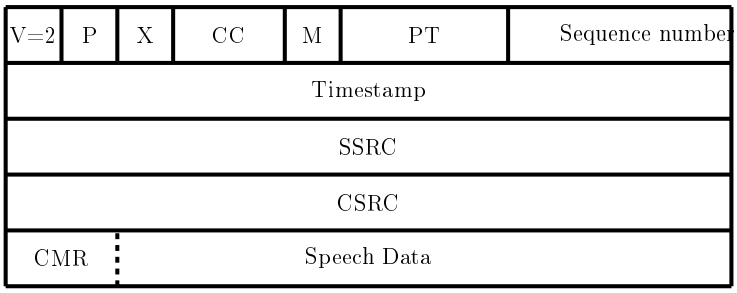


Figure 5.11: RTP packet with CMR included for application layer redundancy requests

presented in [17, Annex A]. An example of EVS packet payload is presented in Figure 5.10 with four main fields, CMR(1 byte), Table of Contents (ToC)(1 byte), Speech Data (SD) and a padding field P. The CMR field contains the request. The ToC field indicates the type of coded frame (EVS Primary, EVS AMR-WB IO, or SID). We can have multiple ToC fields if the packet contains multiple frames. One ToC for each frame in the packet. The field P contains padding as described in RFC 3550. The CMR field is composed of three parts as shown in Figure 5.12.

- H (1 bit): always set to 1.
- T (3 bits): indicates EVS coding modes (EVS Primary in NB, WB, SWB, and FB mode , or EVS AMR-WB IO).
- D (4 bits): contains the CMR requests.

The possible values of CMR are defined in [17, Annexe A]. The main idea of this option is to use the reserved values of CMR codes presented in [17, Annex A] with $T = 111$ other than the one with $D = 1111$. We propose to use CMR codes in Table 5.3 for EVS. In VoLTE, we are constrained to a maximum bit

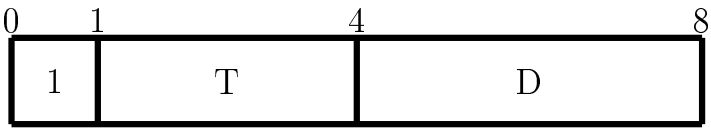


Figure 5.12: CMR

rate of 24.4 kbit/s. Therefore, application layer redundancy is limited to the lowest possible bit rates of EVS, 9.6 kbit/s or lower. The same methodology can be applied to AMR/AMR-WB codec to define CMR codes to request application-layer redundancy with the different AMR/AMR-WB bit rates.

The second option consists in using RTCP-APP requests with AVP. We proposed the first option in the 3GPP as a new method to send application layer redundancy in VoLTE. The idea has been accepted within the feasibility study eVoLP. However, option two which consists of using RTCP-APP is also under consideration in the standard. The choice between the two methods has not yet been decided within the 3GPP by the time of writing of this report.

Conclusion

In this chapter, we presented performance results for the 3GPP EVS codec with or without applicaton-layer redundancy. We evaluated speech quality with objective and subjective tests as function of packet loss rate and operation mode. We

Code CMR	**EVS request**
111 0000	RED 2 × 7.2-NB
111 0001	RED 2 × 8-NB
111 0010	RED 2 × 9.6-NB
111 0011	RED 2 × 13.2-NB
111 0100	RED 2 × 7.2-WB
111 0101	RED 2 × 8-WB
111 0110	RED 2 × 9.6-WB
111 0111	RED 2 × 13.2-WB
111 1000	RED 2 × 7.2-SWB
111 1001	RED 2 × 8-SWB
111 1010	RED 2 × 9.6-SWB
111 1011	RED 2 × 13.2-SWB
111 1100	RED 2 × 6.6
111 1101	RED 2 × 8.85
111 1110	RED 2 × 12.65

Table 5.3: CMR codes for EVS application-layer redundancy requests

demonstrated that application-layer redundancy (2×9.6 kbit/s) has better quality than the 13.2 kbit/s CAM mode in adverse conditions (packet loss rate $\geq 3\%$). Future work will investigate performance delay/loss profiles that reflect measured VoLTE conditions. Moreover, the actual relationship between payload bit rate, radio pathloss and packet loss will be further studied to be able to map test results as a function of radio pathloss.

Chapter 6

Conclusion

Contents

6.1 Main contributions

The main objective of this dissertation was to study, develop and evaluate media adaptation algorithms that can enhance the quality of mobile VoIP services.

First of all, we identified the approaches used to conduct this research. We reviewed the different factors, metrics and adaptation algorithms of mobile VoIP quality. Then, we conducted several experiments according to 3GPP terminal acoustic test methods and operator radio coverage tests to evaluate VoLTE and WebRTC speech quality in different network condition. Finally, we proposed two different adaptation mechanisms that can be implemented in any type of mobile VoIP including VoLTE.

We recall here the main outcomes of this work.

Enhanced method for VoLTE delay measurement (See chapter 3)

We analyzed a test method described in 3GPP to measure VoLTE QoE metrics (quality and delay) in different network delay/loss profiles. We evaluated in particular the impact of network condition on terminal delay. We proposed enhancements to the underlying test method. We concluded also that this methodology can be extended to evaluate terminal jitter buffer performance. This allows to have some requirement on the proprietary jitter buffer algorithms.

WebRTC over LTE coverage tests evaluation (See chapter 4)

We conducted several experiments to evaluate WebRTC voice quality in LTE radio coverage tests. We added the support of the EVS codec to WebRTC to test WebRTC with multiple bit rates of EVS. The different tests allowed us to conclude that the jitter buffer has a significant impact on enhancing the perceived quality in bad LTE radio conditions. The end-to-end delay increases considerably in bad LTE radio conditions due to multiple retransmissions. In acknowledge mode, data is retransmitted in case of loss. Bit rate variation may not enhance the voice quality in bad radio condition but it can enhance the radio coverage.

Evaluation of application-layer redundancy for the EVS codec (See chapter 5)

We evaluated the benefit of application-layer redundancy for the EVS codec. We conducted subjective and objective tests to evaluate its impact on the quality. We used different profiles of packet losses to evaluate the real effectiveness of application layer redundancy to compensate for loss packets. We confirmed that application layer redundancy can be very benefit in case of random losses. However, its impact decreases if bursts of losses increase during transmission. The used offset of the application layer redundancy must be adapted according to the type of the losses. The offset should increases in bursty channel and offset of two is sufficient in case of random losses.

Proposal of an in-band signaling for adaptation requests (See chapter 5)

In most cases, adaptation mechanisms use some extended versions of RTCP as a feedback mechanism to send adaptation requests. The use of RTCP extensions requires RTP AVPF. In VoLTE, RTP AVPF is not allowed. Thus, we proposed to use in-band feedback through CMRs to request application layer redundancy in VoLTE.

6.2 Perspectives

Media adaptation algorithms in real-time applications is an old subject. However, the work on such mechanisms for VoLTE and VoWifi terminals is still an important topic. In particular, media adaptation standardization is on going for

enhanced VoLTE end-to-end performance. Some terminal vendors implement bit rate adaptation mechanisms in their terminals with no requirements for such mechanisms.

In chapter 3, we evaluated VoLTE quality metrics including delay and quality metrics. In some cases, the performance of LTE network may suffer some limitations in indoor environments. Some network operators have deployed VoWifi to replace the VoLTE in indoor. Thus, future work can focus on VoWifi to assess the underlying service quality. The Wifi network is different from LTE and yet the service quality must be equivalent. The studied test method in chapter 3 has been extended to support VoWifi testing in the 3GPP EXT_UED work item. By the time writing this dissertation no results have been presented in the 3GPP about VoWifi quality.

In chapter 5, we studied application-layer redundancy for the EVS codec. We showed through subjectives and objectives tests that application-layer redundancy can enhance speech quality in case of packet loss. However, we used only static loss profiles without introducing jitter. Future work, can focus on evaluating application-layer redundancy using delay/loss profiles to evaluate how jitter buffer can manage the redundant packets in case of network jitter.

Congestion control has been widely studied, but has not been sufficiently considered for VoIP. With the development of future immersion audio codecs with high bit rates, congestion control may become important for audio services. In this context, we realized some work on a bit rate adaptation based on bandwidth estimation and CMR requests. We used an existing bandwidth estimator that we enhanced by using bursts of packets with redundancy to test the capacity of the channel temporary. Future work can concentrate on defining the frequency of sending the redundancy bursts to test the channel and validate the overall mechanism.

In a VoIP communication, voice packets can be transported over different types of networks including LTE, Wifi and Ethernet. It is possible to optimize media adaptation decisions by exploiting information from the physical layer. Some cross-layer mechanisms have been proposed such as CQIC presented in [118]. CQIC is a cross-layer congestion control for cellular networks that uses network estimates to provide optimal packet sending. This mechanism keeps

track of the variable capacity of mobile network based on some indicators extracted from the physical layer. The base idea of this mechanism can be used in real-time communication services.

Machine learning has been used for QoE prediction and management for various services. For instance, a method based on machine learning for VoIP QoE prediction is presented in [119]. Media adaptation algorithms for VoIP can be developed based on machine learning.

Finally, the quality of the telephony service remains an important criteria for choosing a mobile network operator. Thus, it is crucial for network operators to offer a telephony service with the best possible quality.

Appendix A

Résumé

A.1 Contexte

Le concept de téléphonie est très ancien. La première communication transmettant un discours entre deux personnes date de 1873, grâce à l'invention attribuée à Alexander Graham Bell. Le service de téléphonie n'était offert que sur des réseaux filaires jusqu'à la commercialisation du premier réseau mobile par AT&T en 1949. Après cela, le réseau mobile a évolué d'une manière continue. Le réseau mobile transmet aujourd'hui une grande partie de l'ensemble des communications. En France, l'Autorité de régulation des communications électroniques et des postes (ARCEP) déclare que le nombre d'abonnés au service de téléphonie fixe est d'environ 39 millions en 2017, sans augmentation importante depuis plusieurs années, [8]. Cependant, le nombre d'abonnés à la téléphonie mobile est d'environ 90 millions. Ce nombre augmente depuis le déploiement du réseau de quatrième génération (4G) avec un volume global d'appels mobiles supérieur à 40 milliards de minutes en 2017 [8]. Aux États-Unis, l'augmentation du nombre d'utilisateurs de téléphones mobiles est encore plus rapide qu'en France. D'une part, 90% des foyers disposent d'un service téléphonique fixe en 2004 contre moins de 50% en 2017 [9]. D'autre part, le nombre de maisons disposant uniquement de connexions sans fil augmente constamment pour atteindre les 50% en 2017 [9]. Ces statistiques montrent qu'une partie importante des communications est transmise sur des réseaux mobiles. Il est donc important que les opérateurs fournissent un service de téléphonie mobile avec la meilleure qualité possible.

Les réseau 4G est un réseau tout IP. La voix est transmise par commutation de paquets contrairement aux réseaux de générations précédentes. Dans les réseaux 2G et 3G, la commutation de circuits est utilisée pour le transport de la voix. Dans les réseaux à commutation de circuits, des ressources didiées (circuits) sont réservées à la transmission de la voix. Dans les réseaux à commutation de paquets, les données et la voix partagent les mêmes ressources réseaux. Le traffic

des données et de la voix est transporté dans des paquets IP.

La commutation de paquets offre un partage de bande passante plus efficace que la commutation de circuits. Cela permet de réduire les coûts de déploiement du réseau en unifiant l'infrastructure pour la voix et les données. Cependant, dans les réseaux IP, les paquets transmis peuvent avoir plusieurs problèmes lors de la transmission comme le retard, la gigue, la perte de paquets, etc. Ces imperfections dépendent des conditions du réseau. Les performances du réseau mobile sont variables en raison de multiples facteurs. Le canal de propagation radioélectrique est sujet aux fluctuations dues aux réflexions ou diffractions multiples causées par les obstacles mobiles et fixes.

La téléphonie est un service temps réel extrêmement sensible aux retards, à la gigue et à la perte de paquets. Dans les communications temps réel, les paquets doivent être reçus à une fréquence régulière pour reproduire la voix sans interruption. Pendant les conversations, des retards importants peuvent limiter l'interaction entre les deux interlocuteurs. L'UIT-T a défini des seuils pour le retard de bout en bout afin de considérer que le service a une qualité acceptable. Peu de pertes peuvent être tolérées lors d'une communication. En effet, un taux de perte de paquets élevé dégrade considérablement la qualité de la communication.

La commutation de paquets est utilisée dans VoLTE. VoLTE est une forme de VoIP avec un traitement spécifique des paquets vocaux sur le réseau. VoLTE utilise également certaines optimisations réseau QoS spécifiques pour le transport de la voix afin d'améliorer la qualité reçue. VoLTE est une version du service de téléphonie classique, proposée uniquement par les fournisseurs de réseau. Il est également possible d'étendre VoLTE pour offrir des services de communication enrichis tels que ViLTE.

Le déploiement des réseaux 4G a permis l'utilisation de données mobiles en large bande avec des performances nettement meilleures (débit, latence ...) que dans les précédentes générations. Les réseaux 2G ont une vitesse lente avec un débit d'environ plusieurs dizaines de Kbits/s. Les réseaux 3G ont une vitesse supérieure à celle des réseaux 2G, avec un débit de plusieurs Mbps. Dans les réseaux 4G, le débit devient très élevé et atteint plusieurs dizaines de Mbps. Cette amélioration des performances a permis au VoIP mobile sans traitement

spécifique des paquets vocaux. Cela a aidé les applications OTT, telles que *whatsapp, skype, facebook messenger ...*, à évoluer et à concurrencer les services de téléphonie classiques.

Pour développer leur offre de services, les opérateurs de réseau peuvent également proposer leur propre type de services OTT en plus de la VoLTE. Ils peuvent utiliser des applications de téléphonie logicielles ou des téléphones logiciels WebRTC qui fonctionnent dans des conditions optimales, tout comme les autres applications OTT. Dans ce cas, on ne peut pas supposer que les garanties de qualité de service sont fournies par le réseau. Il convient de noter que ce type de service n'est pas un service de téléphonie complet, qui doit inclure les appels d'urgence et certaines obligations légales. Cependant, il est important que les softphones développés par les opérateurs de réseau aient une qualité perçue au moins aussi bonne que les services OTT concurrents.

A.2 Objectifs

La VoIP est basée sur RTP, un protocole générique pouvant être utilisé pour tout type de services en temps réel. Cela permet à l'application d'adapter les contraintes de transport à ses besoins (par exemple: taille de paquet, nombre de trames par paquet, utilisation de la redondance). Ainsi, l'application doit disposer de l'intelligence et de la souplesse nécessaires pour répondre aux conditions de réseau difficiles. Selon le principe de bout en bout, l'intelligence est localisée dans des points d'extrémité pour s'adapter à l'état de variable réseau.

Conformément à ces principes, nos travaux portent en particulier sur les mécanismes d'adaptation des supports dans les terminaux pour améliorer la qualité perçue des communications bidirectionnelles, en fonction des conditions du réseau. Les dimensions d'adaptation suivantes sont abordées dans ce travail.

Le premier mécanisme concerne l'adaptation à la variation de retard provoquée par la gigue du réseau et l'horloge oblique. La gigue réseau est généralement compensée à l'aide d'un tampon de gigue. Il peut être mis en œuvre conjointement avec des opérations de mise à l'échelle temporelle pour modifier la durée des segments audio actifs ou inactifs. Les tampons de gigue peuvent également être utilisés pour gérer l'inclinaison de l'horloge et la gigue

de la carte son. Le deuxième mécanisme concerne l'adaptation aux pertes de paquets. De nombreux mécanismes ont été développés au sein du codeur et du décodeur pour améliorer la résilience des codecs vocaux contre la perte de paquets. Par exemple, des techniques de masquage d'informations redondantes ou de perte de paquets peuvent être appliquées pour minimiser l'impact de la perte de paquets sur la qualité perçue. Le troisième mécanisme est l'adaptation de débit, qui peut être utilisée pour s'adapter à différentes conditions de canal.

Les dimensions ci-dessus ne sont pas nécessairement indépendantes. Par exemple, il est possible de combiner des opérations de tampon de gigue et de dissimulation de perte de paquets, par ex. resynchroniser le décodage lorsqu'un paquet perdu précédemment déclaré arrive en retard ou tirer parti des informations redondantes partielles ou complètes disponibles dans différents paquets du tampon de gigue. De plus, le tampon de gigue peut induire des pertes de paquets supplémentaires. L'utilisation adaptative d'informations redondantes est également une forme d'adaptation du débit.

Un aspect considéré dans ce travail est d'adapter conjointement l'émetteur et le destinataire. Une telle optimisation conjointe nécessite des informations en retour pouvant être fournies soit dans la bande du flux RTP, soit hors de la bande dans RTCP, sous forme de statistiques ou de requêtes. Il est important d'optimiser le système de bout en bout (émetteur global + réseau + récepteur) en exploitant le retour d'information.

A.3 Approches

Ce travail a été effectué dans un laboratoire d'un opérateur (Orange Labs), qui contribue activement à la normalisation. Après le déploiement de la VoLTE, il a été jugé important de définir des exigences relatives aux mécanismes de qualité intégrés dans les terminaux VoLTE pour améliorer les performances du service. Ce travail de recherche était motivé par différentes études au sein du 3GPP auxquelles nous avons activement contribué.

Le premier work item s'appelle ART_LTE-UED (dans Release 12 du 3GPP), qui correspond à "Acoustic requirements and test methods for IMS-based conversational speech services over LTE-UE delay aspects". L'objectif de ce work item est de définir des méthodes de test de délai terminal dans le contexte

VoLTE et de spécifier certaines exigences relatives au délai terminal.

Le deuxième work item s'appelle EXT_UED (dans Release 13 de 3GPP), qui correspond à l'extension des méthodes de test de délai terminal. Les principaux objectifs de ce work item sont les suivants:

- Ajout de la prise en charge du test de délai VoWiFi à la méthode de test de délai acoustique définie dans ART_LTE-UED;
- Définition de certaines exigences sur la précision de l'horloge du terminal;
- Extension de la méthode de test pour caractériser le comportement du buffer de gigue terminal dans des conditions radio LTE réalistes.

Le travail rapporté dans le chapitre A.4 était principalement motivé par ces deux premiers work items.

La troisième work item est en phase d'étude (dans la Release 15 du 3GPP) appelé eVoLP, qui signifie «Enhanced VoLTE Performance», dont les objectifs sont:

- Evaluation de l'impact des implémentations de terminaux propriétaires telles que la dissimulation de perte de paquets et les mécanismes de tampon de gigue;
- Des directives ou des exigences pour que les terminaux VoLTE puissent s'adapter aux modes de codec les plus robustes;
- Spécification des mécanismes permettant aux terminaux d'envoyer des requetes d'adaptation au mode le plus robuste.

Le travail décrit dans section A.6 était principalement motivé par le work item eVoLP.

Nous avons suivi deux approches différentes inspirées de [10]. La première approche est appelée approche en boîte noire. Elle consiste à évaluer des algorithmes d'adaptation sans entrer à l'intérieur de l'algorithme. Elle est adaptée à l'évaluation des terminaux propriétaires, sans connaissance précise de l'algorithme d'adaptation utilisé. Cela inclut l'utilisation des réseaux de test réels avec des liaisons radio réelles pour la collecte et l'analyse des données. Pour être acceptés comme résultats valables dans le 3GPP, des outils complexes

sont utilisés pour mesurer la qualité en suivant uniquement les méthodes de test standardisées. Cela a pris beaucoup de temps pour effectuer les mesures conformément aux exigences des standards.

La deuxième approche est appelée approche en boîte de verre. Elle consiste à développer et à implémenter des algorithmes d'adaptation de média aux conditions réseaux. Elle est utilisée dans le cas des softphones en open sources ou les détails d'implémentation sont disponibles. Nous apportons des améliorations algorithmiques et les validons par des expériences. Des tests subjectifs sont effectués pour confirmer l'effet des améliorations algorithmiques. Nous explorons ici les améliorations pouvant être apportées au codec EVS pour le service VoLTE.

A.4 Evaluation et mesure des métriques de qualité de la VoLTE

Cette partie du travail était principalement motivée par les deux work items du 3GPP appelés ART_LTE-UED et EXT_UED. Ainsi, nous avons évalué les métriques spécifiées dans la méthodes de test du 3GPP afin de caractériser le compromis entre le délai et la qualité de la VoLTE dans différentes conditions de gigue et de perte dans le réseau. Nous avons rapporté les résultats de test sur la précision de l'horloge des terminaux utilisés, le délai terminal dans le lien montant et descendant dans des conditions sans erreur, ainsi que le délai et la qualité en présence de pertes de paquets et de gigue réseau. Nous avons montré que la variation des conditions du réseau peut avoir un impact important sur la qualité de service VoLTE. Nous avons proposé quelques améliorations aux méthodes de test 3GPP existantes. Nous avons mis en évidence certaines lacunes dans la méthode de test acoustique 3GPP. En particulier, la variabilité de la mesure n'a pas été pleinement prise en compte avec un nombre limité d'essais et les dégradations du réseau avec trois profils simulés ne capturent pas le comportement du buffer de gigue dans des conditions réelles de VoLTE. Nous avons discuté une manière dont la méthodologie peut être étendue pour évaluer les performances du buffer de gigue en utilisant une approche boîte noire, et comment modéliser les caractéristiques de retard / perte de paquets VoLTE dans d'une manière réaliste. Nous avons enfin proposé des améliorations au modèle de simulation de perte / retard de paquets VoLTE utilisé dans le 3GPP.

A.5 Evaluation des performances de WebRTC dans différents conditions réseaux

Dans cette partie, nous avons évalué les performances d'une application OTT basée sur WebRTC dans différentes conditions de réseau. Nous avons mesuré et évalué la qualité WebRTC sur le réseau LTE. Nous avons présenté des résultats expérimentaux sur les métriques de qualité WebRTC en fonction du niveau de perte du lien radio LTE. Nous avons complété l'étude par une évaluation de WebRTC sur Ethernet afin d'étudier l'impact de la limitation de la bande passante et du taux de perte de paquets sur la qualité de WebRTC. Nous avons évalué pour des différents cas l'impact de la variation du débit sur la qualité. Nous avons utilisé différentes configurations expérimentales pour les deux études, notamment une véritable plate-forme LTE dans des conditions contrôlées et une mise en œuvre simple de la communication WebRTC sur Ethernet. Nous avons pu montrer que WebRTC s'adapte aux conditions réseaux en adaptant la longueur de son buffer de gigue pour attendre les paquets transmis. En fait, dans un bearer LTE autre que celui de la VoLTE, un mécanisme de retransmission est mis en place au niveau de la couche RLC. Cela permet d'éviter la perte des paquets par contre ça engendre une augmentation du retards de bout en bout. Les OTTs bénéficient énormément de ce genre de mécanisme, par contre un compromis entre retard et qualité doit etre mis en place.

A.6 Evaluation de l'impact de la redondance au niveau applicatif sur la qualité perçue

La perte de paquets est l'un des principaux facteurs de dégradation de la QoE des services VoIP, y compris VoLTE dans des conditions radio difficiles. De nombreux mécanismes ont été proposés pour traiter les pertes de paquets et améliorer la qualité perçue en utilisant des techniques de redondance. Dans cette partie, nous avons étudié un mécanisme de redondance spécial défini pour EVS à utiliser dans les services VoLTE, appelé "Channel Aware Mode" (CAM). Nous avons présenté un autre type de redondance, que nous avons appelé "Application-Layer Redundancy". Ce type de redondance a été défini dans la littérature pour d'autres codecs. Cependant, il n'a jamais été testé pour EVS. Dans ce travail, nous avons implémenté "Application-Layer Redundancy" pour le codec EVS. Nous avons effectué de nombreux tests subjectifs et objectifs pour évaluer l'impact de ce

type de redondance sur la qualité perçue et comparer ses performances à celles de la CAM. Nous avons montré que "Application-Layer Redundancy" permet de garder une qualité optimale en cas de perte de paquets, Par contre cela est fait en augmentant le débit codec utilisé. Nous avons aussi discuté des méthodes de signalisation RTP / RTCP possibles pour déclencher l'utilisation de la redondance au niveau applicatif au cours d'un appel VoLTE.

A.7 Conclusion

Avec le déploiement du service VoLTE et la forte concurrence que subissent les opérateurs de la part des OTTs, la qualité VoIP est de plus en plus nécessaire et importante pour survivre. Etant donné que les OTT ont beaucoup plus d'expérience dans le domaine de la VoIP, les opérateurs doivent mettre en place une stratégie non seulement pour optimiser les performances de leurs réseaux mais aussi pour contrôler les performances des terminaux VoLTE. En fait, en VoIP, le terminal a un role primordial dans la qualité perçue. Dans ce travail de recherche, on a étudié des méthodes de test pour bien évaluer la qualité des services VoIP et les performances des terminaux utilisés et leurs contributions dans la qualité mesurée. On a aussi exploré des mécaismes d'adaptation permettant un couplage optimal entre média et conditions réseaux.

Bibliography

[1] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," IETF RFC 3550, July 2003.

[2] Andre Perez, *La voix sur LTE: reseau et architecture IMS*, Lavoisier, 2013.

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

[4] ITU-T Rec. G.192 STL manual, "ITU-T Software Tool Library Manual," 2009.

[5] ITU-T Rec. G.1028, "End-to-end quality of service for voice over 4G mobile networks," April 2016.

[6] 3GPP TS 26.132, "Speech and video telephony terminal acoustic test specification," .

[7] 3GPP TS 26.131, "Terminal acoustic characteristics for telephony; Requirements," .

[8] Regulatory Authority for Electronic Communications and Posts(ARCEP), "ELECTRONIC COMMUNICATIONS MARKET OBSERVATORY," `www.arcep.fr/index.php?id=13652#c96025`, 2017.

[9] "Landline phones in the united states," `https://www.statista.com/chart/2072/landline-phones-in-the-united-states/`.

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

[11] 3GPP TS 26.071, "Mandatory speech codec speech processing functions; AMR speech Codec; General description," .

[12] 3GPP TS 26.190, "Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Transcoding functions (Release 14) ," .

[13] 3GPP TS 26.441, "Codec for Enhanced Voice Services (EVS); General Overview," 2017.

[14] Stefan Bruhn, Harald Pobloth, M. Schnell, B. Grill, Jon Gibbs, L. Miao, Kari Järvinen, Lasse Laaksonen, N. Harada, N. Naka, Stéphane Ragot, Stéphane Proust, T. Sanda, Imre Varga, C. Greer, Milan Jelinek, M. Xie, and P. Usai, "Standardization of the new 3GPP EVS codec," in *Proc. ICASSP*, 2015.

[15] P. Jones and al., "RTP Payload Format for the iSAC Codec," `http://tools.ietf.org/html/draft-ietf-avt-rtp-isac-04`, 2013.

[16] JM. Valin, K. Vos, and T. Terriberry, "Definition of the Opus Audio Codec," IETF RFC 6716, Sept. 2012.

[17] 3GPP TS 26.445, "3GPP TS 26.445, Codec for Enhanced Voice Services (EVS);Detailed Algorithmic Description," .

[18] 3GPP TS 26.450, "Codec for Enhanced Voice Services (EVS); Discontinuous Transmission (DTX) (Release 14," .

[19] J. Postel, "User Datagram Protocol," IETF RFC 768, 1980.

[20] O.Hersent, D.Gurle, and J.P. Petit, *Voice over IP: Codecs, H323, SIP, MGCP deploiement and dimensionnement*, Dunod, 2004.

[21] C. Perkins, *RTP: Audio and Video for the Internet*, Addison-Wesley, 2003.

[22] J. Benesty, M.M Sondhi, and Y. Huang, *Handbook of Speech Processing*, Springer, 2008.

[23] S.B. Moon, P. Skelly, and D. Towsley, "Estimation and removal of clock skew from network delay measurements," in *Proc. INFOCOM*, 1999.

[24] O. Hodson, C. Perkins, and W.W. Hardman, "Skew detection and compensation for internet audio applications," in *Proc. ICME*, 2000.

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

[26] R. Ramjee and J. Kurose and D. Towsley and H. Schulzrinne, "Adaptive playout mechanisms for packetized audio applications in wide-area networks," in *Proc. INFOCOM*, Jun. 1994.

[27] US Patent 7733893, "Method and receiver for determining a jitter buffer level," 2008.

[28] 3GPP TS 26.114, "IP Multimedia Subsystem (IMS); Multimedia telephony; Media handling and interaction," .

[29] 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," in *Proc. Acta Acustica United with Acustica*, 2015.

[30] 3GPP TS 26.448, "3GPP TS 26.448 codec for enhanced voice services (EVS); jitter buffer management," .

[31] T. Friedman, R. Caceres, and A. Clark, "RTP Control Protocol Extended Reports (RTCP XR)," IETF RFC 3611, Nov. 2003.

[32] Q. Wu, G. Hunt, and P. Arden, "Guidelines for Use of the RTP Monitoring Framework," IETF RFC 6792, Nov. 2012.

[33] A. Clark and Q. Wu, "RTP Control Protocol (RTCP) Extended Report (XR) Block for Packet Delay Variation Metric Reporting," IETF RFC 6798, Nov. 2012.

[34] A. Clark, K. Gross, and Q. Wu, "RTP Control Protocol (RTCP) Extended Report (XR) Block for Delay Metric Reporting," IETF RFC 6843, Jan. 2013.

[35] A. Clark, S. Zhang, J. Zhao, and Q. Wu, "RTP Control Protocol (RTCP) Extended Report (XR) Block for Burst/Gap Loss Metric Reporting," IETF RFC 6958, May 2013.

[36] A. Clark, G. Zorn, and Q. Wu, "RTP Control Protocol (RTCP) Extended Report (XR) Block for Discard Count Metric Reporting," IETF RFC 7002, Sept. 2013.

[37] A. Clark, V. Singh, and Q. Wu, "RTP Control Protocol (RTCP) Extended Report (XR) Block for De-Jitter Buffer Metric Reporting," IETF RFC 7005, Sept. 2013.

[38] M. Masuda, K. Yamamoto, S. Majima, T. Hayashi, and K. Kawashima, "Performance Evaluation of VoIP QoE Monitoring Using RTCP XR," in *Management Enabling the Future Internet for Changing Business and New Computing Services*, vol. 5787, pp. 435 439. Springer, 2009.

[39] J. Sjoberg, M. Westerlund, A. Lakaniemi, and Q. Xie, "RTP Payload Format and File Storage Format for the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) Audio Codecs," IETF RFC 4867, Apr. 2007.

[40] A. Sollaud, "RTP Payload Format for the G.729.1 Audio Codec," IETF RFC 4749, Oct. 2006.

[41] H. Schulzrinne and S. Casner, "RTP Profile for Audio and Video Conferences with Minimal Control," IETF RFC 3551, July 2003.

[42] J. Ott, S. Wenger, N. Sato, C. Burmeister, and J. Rey, "Extended RTP Profile for Real-time Transport Control Protocol (RTCP)-Based Feedback (RTP/AVPF)," IETF RFC 4585, July 2006.

[43] Xavier Lagrange, "Principes de fonctionnement de l'interface radio lte," Technique de l'ingénieur, Te 7374, 2013.

[44] 3GPP TS 23.203, " 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Policy and charging control architecture (Release 15) ," 2017.

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

[46] "WebRTC organization," `https://webrtc.org/`.

[47] "WebRTC 1.0: Real-time Communication Between Browsers," `https://www.w3.org/TR/webrtc/`.

[48] C. Perkins and al., " Web Real-Time Communication (WebRTC): Media Transport and Use of RTP," draft-ietf-rtcweb-rtp-usage-26, 2016.

[49] "Browser support scorecard," `http://iswebrtcreadyyet.com/`.

[50] S. Loreto and S.P. Romano, "Real-Time Communications in the Web, Issues, Achivements and ongoing Standardization Efforts," *IEEE Internet Computing*, 2012.

[51] C. Jennings and T. Hardie and M. Westerlund, "Real-Time Communications for the Web," *IEEE communications Magazine*, 2013.

[52] "Javascript Session Establishment Protocol," `https://tools.ietf.org/html/draft-ietf-rtcweb-jsep-18`.

[53] Sebastian Möller and Alexander Raake, *Quality of Experience: Advanced Concepts, Applications and Methods*, Springer, 2014.

[54] Kjell et al. Brunnström, "Qualinet White Paper on Definitions of Quality of Experience," Mar. 2013, Qualinet White Paper on Definitions of Quality of Experience Output from the fifth Qualinet meeting, Novi Sad, March 12, 2013.

[55] ITU-T Rec. E800, "Quality of telecommunication services: concepts, models, objectives and dependability planning. Terms and definitions related to the quality of telecommunication services," Sep 2008.

[56] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," 2006.

[57] Sibiri TIEMOUNOU, *Developpement d une methode de diagnostic technique des degradations de qualite vocale percue des communication telephoniques a partir d une analyse du signal de parole*, Ph.D. thesis, Universite Rennes1, 2014.

[58] ITU-T Rec. G.107, "The E-Model, a computational model for use in transmission planning," 2003.

[59] ITU-T Rec. G.107.1, "The E-Model, a computational model for use in transmission planning," 2015.

[60] R. Sanchez-Iborra, M.D. Cano, and J. Garcia-Haro, "Revisiting VoIP QoE assessment methods: are they suitable for VoLTE?," *Network Protocols and Algorithms*, 2016.

[61] Markus Fiedler, Tobias Hossfeld, and Phuoc Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, 2010.

[62] Christos Tsiaras and Burkhard Stiller, "A deterministic QoE formalization of user satisfaction demands (DQX)," in *Proc. LCN*, 2014.

[63] 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*, 2012.

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

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

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

[67] 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.

[68] 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.

[69] 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.

[70] R.G. Cole and J.H. Rosenbluth, "Voice over IP Performance Monitoring," in *Proc. SIGCOMM*, 2001, pp. 9 24.

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

[72] 3GPP TS 51.010-1, "Mobile Station (MS) conformance specification; Part 1: Conformance specification," .

[73] ITU-T Rec. P.64, "Determination of sensitivity/frequency characteristics of local telephone systems," Nov. 2007.

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

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

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

[77] 3GPP Tdoc S4-140079, "Method for determining one way delays of LTE radio network simulators," Source: HEAD acoustics.

[78] 3GPP Tdoc S4-AHQ077, "Delay profiles for ART-LTE-UED," Source: Qualcomm.

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

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

[81] ETSI TR 126 935, "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Packet Switched (PS) conversational multimedia applications; Performance characterization of default codecs," 2010.

[82] A. Bacioccola, C. Cicconetti, and G. Stea, "User-level Performance Evaluation of VoIP Using Ns-2," in *International Conference on Performance Evaluation Methodologies and Tools*, 2007.

[83] S. Jadhav, H. Zhang, and Z. Huang, "Performance Evaluation of Quality of VoIP in WiMAX and UMTS," in *Proc. International Conference on Parallel and Distributed Computing, Applications and Technologies*, Oct 2011.

[84] M. Pradhan and L. Sun, *Performance Analysis of Voice Call using Skype*, Plymouth University, 2012.

[85] V. Singh, A. A. Lozano, and J. Ott, "Performance Analysis of Receive-Side Real-Time Congestion Control for WebRTC," in *Proc. International Packet Video Workshop*, 2013.

[86] L.D. Cicco, G. Carlucci, and S. Mascolo, "Congestion Control For WebRTC: Standardization Status And Open Issues," *IEEE Communications Standards Magazine*, 2017.

[87] G. Carullo, M. Tambasco, M. D. Mauro, and M. Longo, "A performance evaluation of WebRTC over LTE," in *Proc. Annual Conference on Wireless On-demand Network Systems and Services (WONS)*, Jan 2016.

[88] M. Al-Ahmadi, Y. Cinar, H. Melvin, and P. Pocta, "Investigating the Extent and Impact of Time-Scaling in WebRTC Voice Traffic Under Light, Moderate and Heavily Congested Wi-Fi APs," in *Proc. ISCA/DEGA Workshop on Perceptual Quality of System (PQS)*, Aug. 2016.

[89] N. Majed, S. Ragot, X. Lagrange, and A. Blanc, "Experimental evaluation of WebRTC voice quality in LTE coverage tests," in *Proc. QoMEX*, 2017.

[90] "Chromium project," `https://www.chromium.org/`.

[91] 3GPP TS 36.101, "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception (Release 15)," .

[92] ETSI TS 136 214, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer; Measurements," .

[93] "MultiDSLA," `http://www.j3tel.fr/solutions/malden/`.

[94] "QXDM," `https://www.qualcomm.com/documents/qxdm`.

[95] ITU-T Rec. P.863.1, "Application guide for Recommendation ITU-T P.863," Sept. 2014.

[96] ITU-T Rec. G.114, "One-way transmission time," May 2003.

[97] J. Skoglund and al., "Voice over IP: Speech Transmission over Packet Networks," in *Handbook of Speech Processing (eds. J. Benesty and M.H. Sondhi and Y.A. Huang)*, pp. 307 330.

[98] S. Andersen, A. Duric, H. Astrom, R. Hagen, W. Kleijn, and J. Linden, "Internet Low Bit Rate Codec (iLBC)," IETF RFC 3951, Dec. 2004.

[99] Václav Eksler and Milan Jelinek, "Transmission mode coding for source controlled CELP codecs," in *Proc. ICASSP*, 2008.

[100] I. Johansson, T. Frankkila, and P. Synnergren, "Bandwidth efficient AMR operation for VoIP," in *Proc. Speech Coding Workshop*, 2002.

[101] S. Ragot and .al, "ITU-T G.729.1: AN 8-32 Kbit/S Scalable Coder Interoperable with G.729 for Wideband Telephony and Voice Over IP," *Proc. ICASSP*, 2007.

[102] V. Atti, D.J. Sinder, S. Subasingha, V. Rajendran, D. Dewasurendra, V. Chebiyyam, I. Varga, V. Krishnan, B. Schubert, J. Lecomte, X. Zhang, and L. Miao, "Improved error resilience for volte and voip with 3gpp evs channel aware coding," in *Proc. ICASSP*, April 2015.

[103] C. Perkins, O. Hodson, and V.Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, 1998.

[104] R.V. Cox, D. Malah, and D. Kapilow, "Improving upon toll quality speech for VOIP," in *Proc. 38th Asilomar*, 2004.

[105] B. Kövesi and S. Ragot, "A low complexity packet loss concealment algorithm for ITU-T G.722," in *Proc. ICASSP*, 2007.

[106] J. Lecomte and al., "Packet-loss concealment technology advances in EVS," in *Proc. ICASSP*, 2015.

[107] 3GPP Tdoc S4-060448, "Results from Subjective Listening Test with Redundancy," Aug 2006, Source: Ericsson.

[108] 3GPP TR 26.952, "Codec for Enhanced Voice Services (EVS); Performance Characterization (Release 14)," .

[109] Anssi Ramo, Antti Kurittu, and Henri Toukomaa, "Evs channel aware mode robustness to frame erasures," *Proc. Interspeech*, 2016.

[110] ITU-T Rec. G.192, "A common digital parallel interface for speech standardisation activities," March 1996.

[111] 3GPP Tdoc S4-130155, "EVS Permanent Document EVS-7a: Processing functions for qualification phase, v1.3," Jan.-Feb. 2013, Source: Editor (Fraunhofer IIS).

[112] 3GPP Tdoc AHEVS-197, "Updated network simulator for JBM," Sept. 2012, Source: Fraunhofer IIS.

[113] E.N. Gilbert, "Capacity of a Burst-Noise Channel," *Bell System Technical Journal*, , no. 39, pp. 1253 -1265, 1960.

[114] V.K. Varma, "Testing Speech Coders for Usage in Wireless Communications Systems," in *Proc. Speech Coding for Telecommunications*, 1993.

[115] 3GPP Tdoc S4-180150, "Subjective test results for EVS with application-layer redundancy," February 2018, Source: Orange.

[116] 3GPP TS 26.442, "Codec for Enhanced Voice Services (EVS); ANSI C code (fixed-point) (Release 14)," .

[117] GSMA IR.92, "IMS Profile for Voice and SMS," Version 11.0, 15 June 2017.

[118] Feng Lu, Hao Du, Ankur Jain, Geoffrey M. Voelker, Alex C. Snoeren, and Andreas Terzis, "Cqic: Revisiting cross-layer congestion control f or cellular networks," in *The 16th International Workshop on Mobile Computing Systems and Applications (HotMobile)*, 2015.

[119] P. Charonyktakis, M. Plakia, I. Tsamardinos, and M. Papadopouli, "On user-centric modular qoe prediction for voip based on machine-learning algorithms," *IEEE Transactions on Mobile Computing*, June 2016.

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Titre : Mesure et amélioration de la qualité d'expérience des services Voix sur IP mobiles.

Mots clés : Réseaux mobiles, VoIP, VoLTE, WebRTC, QoE, Adaptation

Résumé : Les réseaux mobiles 4G basés sur la norme LTE (Long Term Evolution), sont des réseaux tout IP. Les différents problèmes de transport IP comme le retard, la gigue et la perte des paquets peuvent fortement dégrader la qualité des communications temps réel telles que la téléphonie. Les opérateurs ont mis en œuvre des mécanismes d'optimisation du transport de la voix dans le réseau afin d'améliorer la qualité perçue. Cependant, les algorithmes propriétaires de gestion de la qualité dans les terminaux ne sont pas spécifiés dans les standards. Dans ce contexte, nous nous intéressons aux mécanismes d'adaptation de média, intégrés dans les terminaux afin d'améliorer la qualité d'expérience (QoE). En particulier, nous évaluons de manière expérimentale des métriques QoE de la voix sur LTE (VoLTE) en utilisant une méthode de test standardisée. Nous proposons d'améliorer la méthode de test et discutons la manière dont cette méthode peut être étendue pour évaluer les performances du buffer de gigue. Nous évaluons également de manière expérimentale la qualité de WebRTC dans différentes conditions radios en utilisant un réseau réel. Nous évaluons l'impact du buffer de gigue et de la variation du débit sur la qualité mesurée. Pour améliorer la robustesse des codecs contre la perte de paquets, nous proposons d'utiliser une redondance simple au niveau applicatif. Nous implémentons cette redondance pour le codec EVS (Enhanced Voice Service) et nous évaluons ses performances. Enfin, nous proposons un protocole de signalisation qui permet d'envoyer des requêtes de redondance au cours d'une communication afin d'activer ou désactiver celle-ci dynamiquement.

Title : Measuring and Improving the Quality of Experience of Mobile Voice Over IP

Keywords : Mobile networks, VoIP, VoLTE, WebRTC, QoE, Adaptation

Abstract: Fourth-generation mobile networks, based on the Long Term Evolution (LTE) standard, are all-—IP networks. Thus, mobile telephony providers are facing new types of quality degradations related to the voice packet transport over IP network such as delay, jitter and packet loss. These factors can heavily degrade voice communications quality. The real-time constraint of such services makes them highly sensitive to delay and loss. Network providers have implemented several network optimizations for voice transport to enhance perceived quality. However, the proprietary quality management algorithms implemented in terminals are left unspecified in the standards. In this context, we are interested in media adaptation mechanisms integrated in terminals to enhance the overall Quality of Experience (QoE). In particular, we experimentally evaluate Voice over LTE (VoLTE) QoE metrics such as delay and Mean Opinion Score (MOS) sing a standardized test method. We propose some enhancements to the actual test method and discuss how this method can be extended to evaluate de-jitter buffer performance. We also experimentally evaluate WebRTC voice quality in different radio conditions using a real LTE test network. We evaluate the impact of jitter buffer and bit rate variations on the measured quality. To enhance voice codec robustness against packet loss, we propose a simple application-layer redundancy. We implemented it for the Enhanced Voice Service (EVS) codec and evaluate it. Finally, we propose a signaling protocol that allows sending redundancy requests during a call to dynamically activate or deactivate the redundancy mechanism.