UNIVERSITY OF CRETE
COMPUTER SCIENCE DEPARTMENT

A comparative analysis of the perceived quality of VoIP under various wireless network conditions

by

Ilias Tsompanidis

M.Sc. thesis

Heraklion
September 2010

University of Crete

Computer Science Department

A COMPARATIVE ANALYSIS OF THE PERCEIVED QUALITY OF VOIP UNDER VARIOUS WIRELESS NETWORK CONDITIONS

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science in Computer Science

Author:

Ilias Tsompanidis
Computer Science Department, University of Crete, Hellas

Committee Members:

Maria Papadopouli
Assistant Professor
Computer Science Department, University of Crete, Hellas
Committee chair

Evangelos Markatos
Professor
Computer Science Department, University of Crete, Hellas

Panagiotis Tsakalides
Professor
Computer Science Department, University of Crete, Hellas

The thesis of Ilias Tsompanidis is approved:

Panos Trahanias
Professor
Director of Graduate Sudies
Computer Science Department, University of Crete, Hellas

Heraklion, September 2010

"I love deadlines. I like the whooshing sound they make as they fly by."

Douglas Adams

Abstract

A COMPARATIVE ANALYSIS OF THE PERCEIVED QUALITY OF VOIP UNDER VARIOUS WIRELESS NETWORK CONDITIONS

by

Ilias Tsompanidis

Master of Science in Computer Science
University of Crete, Hellas, 2010
Professor Maria Papadopouli, Chair

Wireless local area networks are increasingly being deployed to accommodate the user demand for constant connectivity. Wireless networks though often experience periods of severe impairment (PSIs), characterised by significant packet losses in either or both directions between the wireless Access Points (APs) and wireless hosts, increased TCP-level retransmissions, rate reduction, throughput reduction, increased jitter, and roaming effects. The frequency and intensity of PSI events in modern home and enterprise wireless networks is not well understood. Very few studies analyse the impact of PSI events on the quality of user experience.

This research aims to provide a better understanding of the perceived quality of unidirectional, non-interactive VoIP calls under various wireless network conditions, namely handover and high background traffic. Its main contributions include a novel methodology for performing auditory tests, a performance analysis of VoIP, the statistical analysis and identification of the significant network parameters, and a re-examination of various rules-of-thumb in the context of VoIP over wireless networks. Specifically, in this work, we employed the E-model and performed empirical measurements and subjective auditory tests to analyse the impact of the aforementioned network conditions on the perceived VoIP quality. The reported results show the inability of the E-Model to capture the quality of user experience and the significant impact of various network parameters, such as jitter and packet loss burst inter-arrivals on the perceived quality. Furthermore, this work demonstrates that both the network condition (e.g., roaming, type of background traffic) and the support of Quality of Service (QoS) mechanism exhibit statistically significant differences in terms of their reported opinion score values. The empirical experiments reveal an interesting behaviour of QoS-enabled wireless networks, leading to long packet loss bursts in high-priority flows. Finally, this research proposes a new methodology for performing auditory tests for lengthy VoIP calls.

Περίληψη

ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ ΤΗΣ ΠΟΙΟΤΗΤΑΣ ΤΟΥ VOIP ΚΑΤΩ ΑΠΟ ΔΙΑΦΟΡΕΣ ΣΥΝΘΗΚΕΣ ΑΣΥΡΜΑΤΟΥ ΔΙΚΤΥΟΥ

Συγγραφέας

Ηλίας Τσομπανίδης

Μεταπτυχιακό Δίπλωμα Ειδίκευσης στην Επιστήμη Υπολογιστών
Πανεπιστήμιο Κρήτης, Ελλάδα, 2010
Επόπτης: Καθ. Μαρία Παπαδοπούλη

Ασύρματα δίκτυα τοπικού χαρακτήρα εγκαθιδρύονται με γρήγορους ρυθμούς ώστε να καλύψουν τη ζήτηση των χρηστών για συνεχή συνδεσιμότητα. Τα ασύρματα δίκτυα συχνά αντιμετωπίζουν περιόδους σημαντικών βλαβών (PSIs), που χαρακτηρίζονται από σημαντικές απώλειες πακέτων στη μία ή και στις δύο κατευθύνσεις μεταξύ των σταθμών βάσης (APs) και των ασύρματων συσκευών, αυξημένες επαναμεταδόσεις στο επίπεδο μεταφοράς, μείωση ρυθμού μετάδοσης, μείωση ρυθμοαπόδοσης, αύξηση αστάθειας (jitter) και επιδράσεις περιαγωγής. Η συχνότητα και η ένταση των γεγονότων PSI σε σύγχρονα ασύρματα δίκτυα δεν είναι επαρκώς γνωστές και ελάχιστες έρευνες αναλύουν την επίδραση των γεγονότων PSI στην ποιότητα της εμπειρίας του χρήστη.

Αυτή η μελέτη στοχεύει στην παροχή μιας καλύτερης κατανόησης της ποιότητας μονοκατευθυντικών, μη-αλληλεπιδραστικών κλήσεων VoIP, όπως την αντιλαμβάνεται ο χρήστης, κάτω από διάφορες καταστάσεις στο ασύρματο δίκτυο, όπως περιαγωγή και υψηλό φόρτο περιβάλλουσας κίνησης. Η μελέτη αυτή περιλαμβάνει μια νέα μεθοδολογία για τη διεξαγωγή ακουστικών τεστ, την ανάλυση της απόδοσης του VoIP, τη στατιστική ανάλυση και εύρεση των σημαντικών παραμέτρων δικτύου που επηρεάζουν την ποιότητα υπηρεσίας όπως την αντιλαμβάνεται ο χρήστης. Επίσης επανεξετάζει τους διάφορους εμπειρικούς κανόνες όσον αφορά το VoIP πάνω από ασύρματα δίκτυα. Ειδικότερα, σε αυτή τη μελέτη χρησιμοποιήσαμε το E-Model, πραγματοποιήσαμε πειραματικές μετρήσεις και υποκειμενικά ακουστικά τεστ με σκοπό την ανάλυση της επίδρασης των προαναφερθέντων καταστάσεων δικτύου στην αντιληπτή ποιότητα του VoIP. Τα αποτελέσματα δείχνουν πως το E-Model αδυνατεί να καταγράψει την ποιότητα εμπειρίας του χρήστη καθώς και τη σημαντική επίδραση στην αντιληπτή ποιότητα των διαφόρων παραμέτρων δικτύου, όπως το jitter, και τα διαστήματα μεταξύ ριπών απολεσθέντων πακέτων. Επιπλέον, η εργασία αυτή αναδεικνύει πως ότι τόσο η κατάσταση του ασυρμάτου δικτύου και χρήστη (π.χ., περιαγωγή, τύπος περιβάλλουσας κίνησης) όσο και η υποστήριξη μηχανισμών ποιότητας υπηρεσίας (QoS)

παρουσιάζουν στατιστικά σημαντικές διαφορές ως προς τις καταγεγραμμένες βαθμολογίες άποψης (MOS). Επίσης, οι πειραματικές μετρήσεις αποκαλύπτουν μια ενδιαφέρουσα συμπεριφορά των ασυρμάτων δικτύων με ενεργοποιημένους μηχανισμούς ποιότητας υπηρεσίας, που οδηγούν σε μακρές ριπές απολεσθέντων πακέτων σε ροές υψηλής προτεραιότητας. Επιπροσθέτως, η έρευνα αυτή προτείνει μια νέα μεθοδολογία για την διεξαγωγή ακουστικών τεστ για κλήσεις VoIP μεγάλης διάρκειας.

Acknowledgements

This thesis is the result of a brilliant two-and-a-half-year stop of my life's journey in Crete. The Computer Science Department of the University of Crete and the Telecommunications & Networks Laboratory at the Institute of Computer Science (ICS) of the Foundation for Research and Technology Hellas (FO.R.T.H) have been unimaginably hospitable for my quirks and oddities, gifting me a treasure of knowledge and experiences.

The leading force behind this work was my supervisor, Maria Papadopouli, to which I owe my eternal gratitude for her understanding, help and support. I feel obliged to explicitly thank Giorgos Fortetsanakis, Toni Hirvonen and Yannis Agiomyrgiannakis as well as Magda Chatzaki and Stefanos Papadakis for their invaluable help and advice.

I would also like to thank the -too numerous to mention by name- colleagues and friends at the Telecommunications & Networks Laboratory and the University of Crete. My life has become brighter since I met each and every one of them.

It is important to acknowledge that this work was supported by a graduate fellowship from the ICS-FO.R.T.H.

Finally, I would like to thank my family and Despoina -the lighthouse in my darkest hours- for their love and support.

Contents

Abstract **iv**

Περίληψη **v**

Acknowledgements **vii**

List of Figures **x**

List of Tables **xi**

List of Abbreviations **xii**

1 Introduction **1**
1.1 The emergence of wireless networks and VoIP 1
1.2 Motivation 1
1.3 Objectives 2
1.4 Related work 3
1.5 Contributions 5
1.6 Related publications 6
1.7 Roadmap 7

2 Network conditions and testbed **8**
2.1 Scenarios 8
2.2 Handover testbed 9
2.3 Background traffic testbed 10
2.3.1 Heavy UDP traffic 10
2.3.2 Heavy TCP traffic 11
2.4 Hardware setup 13

3 Measurements and evaluation methodology **14**
3.1 Data collection 14
3.2 Voice recording 15
3.3 Data processing 16
3.4 Network parameters 17
3.4.1 Delay 17
3.4.2 Jitter 17

3.4.3 Packet loss burst size . . . 18
3.4.4 Packet loss burst interarrival . . . 18
3.5 Methodology for the subjective tests . . . 18
3.5.1 Selection of segments . . . 19
3.5.2 Reference samples . . . 24
3.5.3 Auditory tests . . . 24
3.6 E-Model . . . 26

4 Performance analysis 29
4.1 Subjective tests results . . . 29
4.2 Reference samples . . . 29
4.3 Preliminary analysis . . . 31
4.4 Subjective tests and E-Model comparison . . . 33
4.5 Impact of various network parameters on quality . . . 37
4.6 Impact of the statistical properties of the network parameters . . . 39
4.7 Revisiting rules-of-thumb . . . 40
4.8 QoS and packet loss bursts . . . 41

5 Conclusions 44

6 Future work 46

A Auditory test questionnaire 48

Bibliography 50

List of Figures

2.1 Handover scenario. 9
2.2 Handover delay and packet loss. 9
2.3 Heavy UDP traffic scenario. 11
2.4 Heavy UDP delay and packet loss. 11
2.5 Heavy TCP traffic scenario. 12
2.6 Heavy TCP delay and packet loss. 12
2.7 The testbed and the most important components. 13

3.1 Segment constellations. 20
3.2 Segments of QoS-disabled Handover. 21
3.3 Segments of QoS-enabled Handover. 21
3.4 Segments of QoS-disabled Heavy UDP. 22
3.5 Segments of QoS-enabled Heavy UDP. 22
3.6 Segments of QoS-disabled Heavy TCP. 23
3.7 Segments of QoS-enabled Heavy TCP. 23
3.8 Estimated MOS as a function of rating factor R. 28

4.1 Subjective and E-Model grading of the MPLRU samples. 30
4.2 Subjective and E-Model MOS for different network scenarios. 32
4.3 Tukey's HSD with three-way interactions. 34
4.4 Tukey's HSD with criterion effect. 35
4.5 Tukey's HSD with QoS effect. 35
4.6 Tukey's HSD with Scenario effect. 36
4.7 Tukey's HSD with scenario and criterion interaction. 36
4.8 Tukey's HSD with scenario and QoS interaction. 37
4.9 Tukey's HSD with QoS and criterion interaction. 37
4.10 Segments with similar average network parameters jitter CCDF. 39
4.11 Tukey's HSD with group effect. 40
4.12 Packet loss burst length distribution. 42

List of Tables

3.1 MOS scale. 25

4.1 Subjective and E-Model MOS. 30
4.2 Subjective and E-Model grading of the MPLRU samples. 30
4.3 List of scenarios and grading criteria. 31
4.4 Subjective and E-Model MOS segment grading. 33
4.5 ANOVA analysis of Subjective and E-Model MOS per network condition. 34
4.6 ANOVA analysis of impact of network parameters on VoIP quality. 38
4.7 Segments with similar network parameters. 39
4.8 ANOVA analysis of impact of network parameters statistical properties on VoIP quality. 40
4.9 Segments invalidating the rule-of-thumb threshold of 3% packet loss. . . . 41
4.10 Segments invalidating the rule-of-thumb threshold of 10ms delay jitter. . . 41
4.11 Segments with bad quality under the rule-of-thumb thresholds. 41

List of Abbreviations

ANOVA	**AN**alysis **O**f **VA**riance
AP	**A**ccess **P**oint
ARF	**A**uto **R**ate **F**allback
BurstR	**Burst R**atio
CCDF	**C**ounter **C**umulative **D**istribution **F**unction
DCF	**D**istributed **C**oordination **F**unction
EDCA	**E**nhanced **D**istributed **C**hannel **A**ccess
ISP	**I**nternet **S**ervice **P**rovider
MNRU	**M**odulated **N**oise **R**eference **U**nit
MOS	**M**ean **O**pinion **S**core
$\mathbf{MOS_{CQE}}$	**M**ean **O**pinion **S**core of **C**onversational **Q**uality **E**stimation
MPLRU	**M**odulated **P**acket **L**oss **R**eference **U**nit
NGN	**N**ext **G**eneration **N**etwork
NTP	**N**etwork **T**ime **P**rotocol
PESQ	**P**erceptual **E**valuation of **S**peech **Q**uality
Ppl	**P**ercentage of **p**acket **l**oss
PSI	**P**eriod of **S**evere **I**mpairmen?
QoS	**Q**uality **o**f **S**ervice
RSSI	**R**eceived **S**ignal **S**trength **I**ndicator
RTP	**R**eal-Time **T**ransport **P**rotocol
SIP	**S**ession **I**nitiation **P**rotocol
SNR	**S**ignal to **N**oise **R**atio
Tukey's HSD	**Tukey's H**onestly **S**ignificant **D**ifference
VoIP	**V**oice **o**ver **I**nternet **P**rotocol

Chapter 1

Introduction

1.1 The emergence of wireless networks and VoIP

Wireless local area networks are increasingly being deployed to accommodate to the users demand of constant connectivity. More and more users resort to wireless technologies for their every-day activities. In the last years, Voice over IP (VoIP) operators with their mainstream support of mobile devices (such as laptops and smartphones) have started to replace traditional telephony providers and VoIP has become an integral part of the everyday activities. Wireless networks widespread adoption has made them the preferred medium for other uses too. Users enjoy high-quality content over the Internet, greatly increasing the traffic load of the local wireless network. Wireless networks often are misconfigured or the traffic demand is extremely high, resulting in numerous connection problems with significant effect on the user quality of experience.

1.2 Motivation

Wireless networks often experience "periods of severe impairment" (PSIs), characterised by significant packet losses in either or both directions between the wireless Access Points (APs) and wireless hosts, increased TCP-level retransmissions, rate reduction, throughput reduction, increased jitter, and roaming effects. A PSI can last for several seconds to the point that it can be viewed as an outage. The frequency and intensity of PSI events in modern home and enterprise wireless networks is not well understood. Very few studies analyse the impact of PSI events on the quality of user experience.

The throughput, jitter, latency, and packet loss, have been used to quantify network performance and various studies have shown their performance under different network conditions (e.g., handoff, contention, and congestion). Some important observations have been made in the context of wireless networks: (a) handovers result to packet losses (e.g., [1, 2]), (b) queue overflows at APs lead to poor VoIP quality (e.g., [3]), and (c) average delay does not capture well the VoIP quality because of the burstiness of packet losses (e.g., [4]). For various applications, a maximum tolerable end-to-end network delay has been estimated (e.g., about 150ms for VoIP applications [5, 6]). Could such crude statistics accurately denote the quality of experience? They have been extensively used as rules-of-thumb for the performance of VoIP, even though there is evidence that depending on the temporal statistical characteristics of the packet losses and delays during a call, the impact on the user experience varies. It becomes apparent that the rules-of-thumb should be re-examined and the impact of the temporal statistical characteristics of various network parameters on the perceived quality should be studied. However, there are only a few comparative analysis studies of the impact of various network conditions on the perceived quality of experience. Furthermore, this quality is often estimated with the use of various specialised benchmarks/metrics, such as the E-model and PESQ, accurate for a small range of network conditions. There is need to examine the reliability of these benchmarks and study their accuracy in case of PSI events.

1.3 Objectives

This research aims to provide a better understanding of the perceived quality of unidirectional, non-interactive VoIP calls under various wireless network conditions, namely handover and high background traffic. The research goals of this work are twofold: first, the user perceived quality of experience is studied, performing empirical measurements under various network conditions, namely handover and high background traffic. Such conditions may result to a PSI. In one case the user is carrying a VoIP-enabled mobile device, walking outside the coverage area of an AP, causing a handover. On the other hand, the user initiates a VoIP call, while the wireless medium is saturated by BitTorrent or UDP traffic.

Second, we aim to study the accuracy of objective metrics that estimate the perceived quality of VoIP. One of them, the E-Model is widely used in network planning for support of VoIP applications. We plan to study the behaviour and accuracy of the E-Model under the aforementioned conditions in comparison to a user study. Furthermore, this work intends to perform subjective auditory tests to evaluate the user perceived quality and analyse the impact of various parameters, namely the network condition on the user perceived quality. Finally, we intend to reexamine some rules-of-thumb for the performance of VoIP.

1.4 Related work

While there have been several studies discussing the network statistics under different conditions, most of them focus on the impact of these conditions on the aggregate throughput and capacity. The IEEE802.11 handover has been analyzed and various improvements have been proposed. For example, Forte *et al.* [7] analyzed the various delays involved in the handoff/reassociation process in an experimental testbed and the impact of the handoff on a SIP call. They reduced this overhead by enabling the wireless device to acquire a temporal address. SyncScan [8] reduces the network unavailability during an AP handoff by enabling the client to synchronize the scanning phase with the APs' beacons. Pentikousis *et al.* [9] measured the capacity of a WiMAX testbed in terms of VoIP calls. Li *et al.* [10] proposed a new perceived speech quality driven retransmission mechanism to achieve optimum perceived speech quality for wireless VoIP (in terms of the objective MOS) by switching to the most suitable retransmission schemes (*i.e.*, No Retransmission, Speech Property-Based Retransmission and Full Retransmission) under different communication conditions. Verkaik *et al.* [11] combined uplink TDMA and downlink aggregation mechanisms to develop a system called Softspeak that simultaneously improves VoIP call quality while preserving network capacity for best-effort data transfer over IEEE802.11 networks. Arjona *et al.* [12] studied the performance of VoIP over 3G networks employing a real testbed. They conclude that the end user VoIP experience over HSDPA is still significantly worse than with circuit switched solutions and is not acceptable. Ganguly *et al.* [13] evaluated various packet aggregation, header compression, adaptive routing, and fast handoff techniques. Anjum *et al.* [14] performed an experimental study of the VoIP in WLAN, quantifying the VoIP capacity under light and heavy traffic load, and the practical benefits of implementing backoff control and priority queuing at the AP. Finally, Shin *et al.* [6] performed empirical-based measurements and simulations to estimate the capacity of an IEEE802.11 network in terms of number of VoIP calls and analysed the impact of the preamble size, ARF algorithm, RSSI, packet loss, and scanning. They used as criterion for the quality of calls that the end-to-end delay should not exceed 150ms and the packet loss probability should be 3% or less, whereas Hole *et al.* [15] used the E-Model and simulations to estimate VoIP capacity of IEEE802.11b, depending on delay and MOS constraints.

In the context of Mobisense project, Deutsche Telecom Lab has developed a Next Generation Network (NGN) testbed and implemented a system that enables seamless codec changes to improve the quality during handovers [16]. They performed subjective tests to quantify the degradation in user perceived quality for various types of network changes, namely handovers between various types of networks and changeovers between various codecs [17]. An analysis of the E-model and PESQ quality estimation tools was also performed in the context of the NGN testbed [18] and an enhancement of the E-Model by adding a bandwidth switching impairment factor was proposed [19].

Chen *et al.* [20] analysed the user satisfaction in Skype, employing the call duration as the quality benchmark. Hoene *et al.* [21] evaluated the call quality in adaptive VoIP applications and codecs and showed that high-compression codecs (with relatively low voice quality) may behave better than top-quality codecs under packet losses and limited available bandwidth. Markopoulou *et al.* [22] focused on ISP network problems and showed that ISP networks suffer from PSIs affecting the real-time applications. They also studied the time constants of the “recency effect” presented by the France Telecom R&D [23].

The E-Model [24] and the PESQ [25] have been used as the base of many perceptual quality estimation models presented by the community, even though their accuracy has been questioned. Rix [26] urges that the E-Model should be used only for network planning while Pennock [27] notes that PESQ can be a useful tool in helping identify potential problem areas, but it is not accurate enough to specify speech quality requirements or to use for verification of speech quality performance. Important decisions should still be based on the results of well-known designed subjective studies. Clark [28] described VQmon, a non-intrusive monitoring technique for Voice over IP networks that is computationally efficient and suitable for integrating or embedding into VoIP gateways or IP Phones, by using an extended version of the ITU G.107 E-Model incorporating the effects of time varying packet loss and "recency", while a 4-state Markov model is used to represent the time distribution of packet loss during a VoIP call. Ding *et al.* [29] investigated the effects of packet loss and delay jitter on speech quality in VoIP, proposed a new formula to quantify these effects based on the Internet traffic being self-similar, and incorporated this formula into the E-Model. Sun *et al.* [30] combined the E-Model and the PESQ to present a model covering a variety of codecs, focused on the perceptual optimisation of the playout buffer.

1.5 Contributions

Its main contributions include:

- a novel methodology for performing auditory tests
- an evaluation of the user-perceived quality of VoIP under various network conditions
- a statistical analysis for the identification of the significant network parameters
- a re-examination of various rules-of-thumb in the context of VoIP over wireless networks

Specifically, in this work, we employed the E-model and performed empirical measurements and subjective auditory tests to analyse the impact of the aforementioned network conditions on the perceived VoIP quality. We use various statistical analysis methods, such as ANOVA and Tukey's HSD test to analyse the collected data. We study the impact of the network condition and various network parameters, such as delay, jitter, packet loss percentage, and packet loss bursts, on the perceived quality. We also test the accuracy of the E-Model in regard to the scores obtained by subjects participating in an auditory test.

The reported results show the inability of the E-Model to capture the quality of user experience and the significant impact of various network parameters, such as jitter and packet loss burst inter-arrivals on the perceived quality. Furthermore, we demonstrate that both the network condition, namely, roaming and type of background traffic, and the support of a Quality of Service (QoS) mechanism exhibit statistically significant differences in terms of their reported opinion score values. The empirical experiments reveal an interesting behaviour of QoS-enabled wireless networks, leading to long packet loss bursts in high-priority flows. Finally, this research proposes a new methodology for performing auditory tests for lengthy VoIP calls.

1.6 Related publications

The comparative analysis of the impact various network conditions, codec type, and packet loss concealment mechanisms have on the perceived VoIP quality using the E-Model and PESQ is the main contribution of the paper. **"Analyzing the impact of various wireless network conditions on the perceived quality of VoIP"** [31] published in the *17th IEEE Workshop on Local and Metropolitan Area Networks (LANMAN)* by Ilias Tsompanidis, Georgios Fortetsanakis, Toni Hirvonen, and Maria Papadopouli. This paper received an "honourable mention".

We add on that work and study the performance of VoIP when QoS mechanisms are available on the wireless network in the paper: **"A comparative analysis of the perceived quality of VoIP under various wireless network conditions"** [32] published in the *8th International Conference on Wired/Wireless Internet Communications (WWIC)* by Ilias Tsompanidis, Georgios Fortetsanakis, Toni Hirvonen, and Maria Papadopouli.

The impact of the network parameters on the quality of experience and the enhancement of the E-Model to include them, as well as the introduction of a new subjective auditory test methodology appropriate for calls of long duration are presented in the journal paper (under preparation for submission): **"On the perceived quality of VoIP under various wireless network conditions"** by Ilias Tsompanidis, Toni Hirvonen, Yiannis Agiomyrgiannakis, and Maria Papadopouli.

1.7 Roadmap

Chapter 2 provides a brief description of our testbed, the different network conditions and the traffic scenarios as well as their characteristics. Chapter 3 details the measurement and evaluation methodology. It provides information for the experimental testbed, the configuration of the equipment in addition to the tools and applications used. Then it describes the auditory test methodology, along with a novel model for producing reference samples. It also proposes a method for selecting small-duration segments representative of a lengthy VoIP call recording, appropriate for use in the auditory tests. The E-Model is briefly described at the final part of this Chapter, with emphasis on the network-related attributes and parameters. Chapter 4 discusses the statistical analysis results. It employs ANOVA and Tukey's HSD test to show the significant differences between the E-Model and the subjective scores, the scenarios and network conditions, as well as identify the impact of various network parameters. The analysis of the empirical experiments reveal an interesting behaviour of QoS-enabled wireless networks, leading to long packet loss bursts in high-priority flows. Chapter 5 presents our conclusions and finally, Chapter 6 explores our future work plans.

Chapter 2

Network conditions and testbed

This Chapter presents the testbed, the different network conditions and the traffic scenarios as well as their characteristics.

2.1 Scenarios

Several network conditions that result in PSIs are distinguished and the following scenarios are formed:

- **handover**: no background traffic, user mobility and client handover between wireless APs
- **heavy UDP traffic**: no user mobility, UDP flows saturating the wireless LAN
- **heavy TCP traffic**: no user mobility, TCP flows, generated by a BitTorrent client, saturating the wireless LAN

We setup two testbeds, namely the *handover testbed* in which a user, performing a VoIP call, roams in the premises of FORTH and the *background traffic testbed* in which background traffic that corresponds to the last two scenarios is generated. A recording of a female voice around 1:30 minutes long (source file) is "replayed" under the aforementioned network conditions. In each testbed, we emulate the corresponding conditions (background traffic/user mobility) of each scenario, "replay" the source file, and collect the traces at both endpoints for analysis. Specifically, we analyse the impact of each

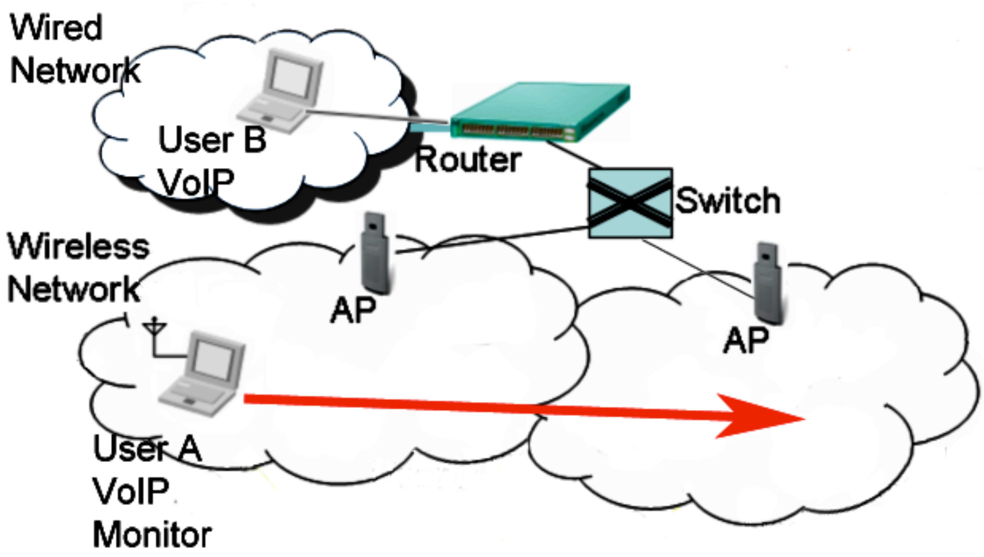
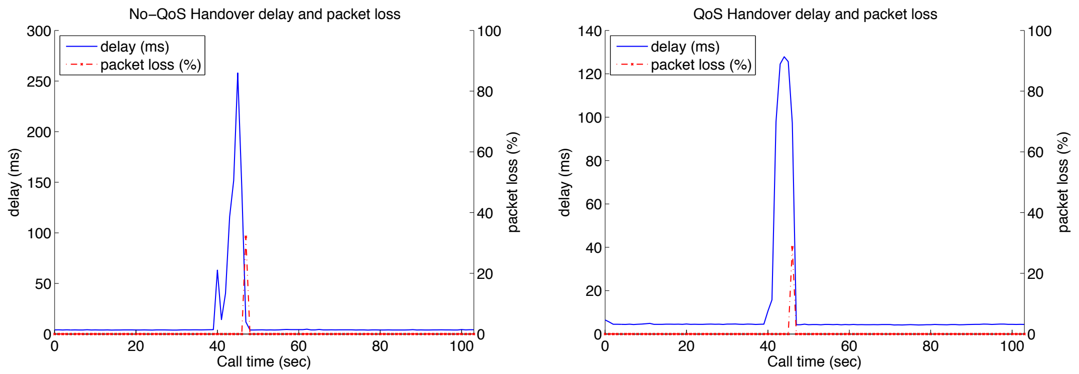


FIGURE 2.1: Handover scenario: User A moves to the coverage area of a different AP while (s)he participates in a VoIP call with user B.



(a) No-QoS Handover delay and packet loss.

(b) QoS Handover delay and packet loss.

FIGURE 2.2: Handover delay and packet loss.

condition on the perceived user experience of the VoIP call. Note that these VoIP calls are essentially unidirectional (streaming-like and non-interactive).

2.2 Handover testbed

The handover testbed includes one VoIP client connected via FastEthernet and one VoIP client connected via IEEE802.11 to the ICS-FORTH infrastructure network. A user holding a wireless laptop (User A) roams in the premises of ICS-FORTH. While moving, the wireless client slowly walks out of range of the AP and a handover is performed.

The ICS-FORTH wireless infrastructure includes several APs broadcasting the same SSID. The handovers are performed between the wireless client and these APs. The wireless client starts scanning for available APs when the received signal-to-noise ratio falls under a specific threshold and decides to associate with a new AP if one is found with Δ dBm better SNR. These values may differ across various IEEE802.11 card manufacturers. If such an AP is found, the re-association process begins, during which the client sends the appropriate messages to the new AP. In the case that no AP is found to satisfy these conditions, the client periodically scans the channels, while the SNR remains under the threshold [33].

As empirical studies have shown (e.g., [2, 7, 8, 34, 35]), handoffs between APs in wireless LANs can consume from one to multiple seconds, as associations and bindings at various layers need to be re-established. Such delays include the acquisition of a new IP address, duplicate address detection, the reestablishment of secure association, discovery of available APs. The overhead of scanning for nearby APs can be of 250ms, far longer than what can be tolerated by VoIP applications. The active scanning in the handoff process of the IEEE802.11 is the primary contributor to the overall handoff latency and can affect the quality of service for many applications.

2.3 Background traffic testbed

The background traffic testbed includes a VoIP client connected via IEEE802.11, a VoIP client connected via FastEthernet, four wireless nodes connected via IEEE802.11 and one node connected via FastEthernet. The four wireless nodes produce the background traffic according to the predefined scenarios. All wireless nodes are connected to a single AP.

2.3.1 Heavy UDP traffic

The *heavy UDP traffic* scenario focuses on the quality of VoIP under congestion caused by a large amount of traffic load generated by a small number of flows, overloading the AP (Figure 2.3). An iperf UDP server runs on the wired server. To produce the background traffic each of the four wireless nodes sends packets of 1500 bytes of UDP traffic to the wired server at a 2Mb/s data rate (a total of 8Mb/s). The AP operates

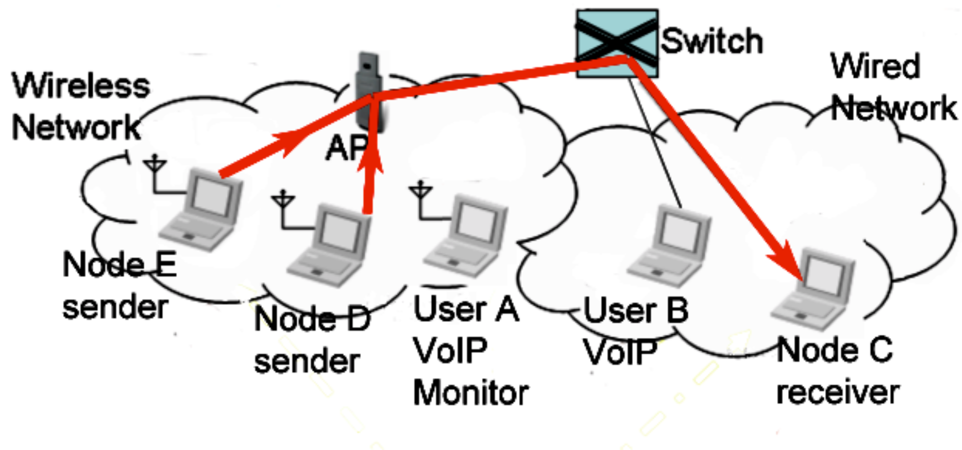
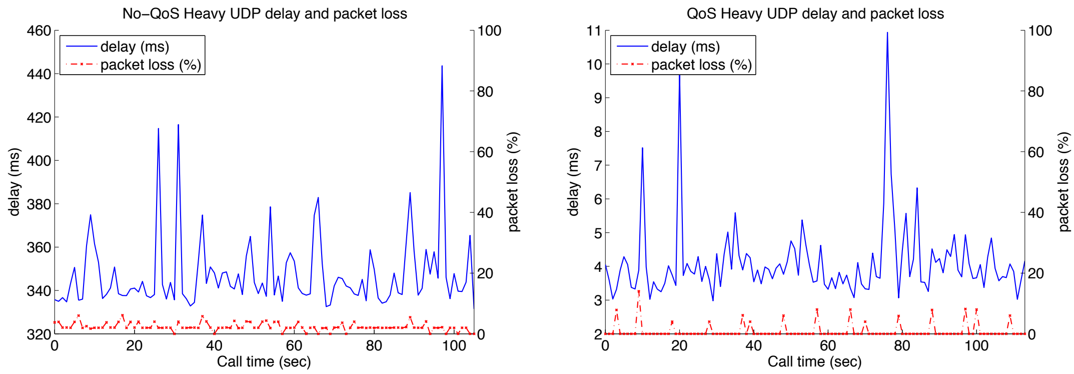


FIGURE 2.3: Heavy UDP traffic scenario: each of the nodes D, E, F and G transmit 2Mb/s UDP traffic towards node C (nodes F and G are not shown).



(a) No-QoS Heavy UDP delay and packet loss.

(b) QoS Heavy UDP delay and packet loss.

FIGURE 2.4: Heavy UDP delay and packet loss.

in IEEE802.11b and the aggregate traffic exceeds the theoretical maximum throughput of an IEEE802.11 network (approximately 6Mb/s [36]). The two VoIP clients initiate a call under these conditions. These scenarios exhibit phenomena of congestion of the wireless channel and continuous contention of the wireless nodes.

2.3.2 Heavy TCP traffic

In the *heavy TCP traffic* scenario, the background traffic is generated by one wireless node running a BitTorrent client, downloading three highly seeded files while the VoIP call takes place (Figure 2.5. The BitTorrent client obtains an IP address from FORTH's DHCP server, and is directly connected to the Internet (no NAT involved). Each file size is more than 5GB. As soon as the downloading speed stabilises at approximately the capacity of the AP, the Sender initiates the VoIP call. The BitTorrent protocol

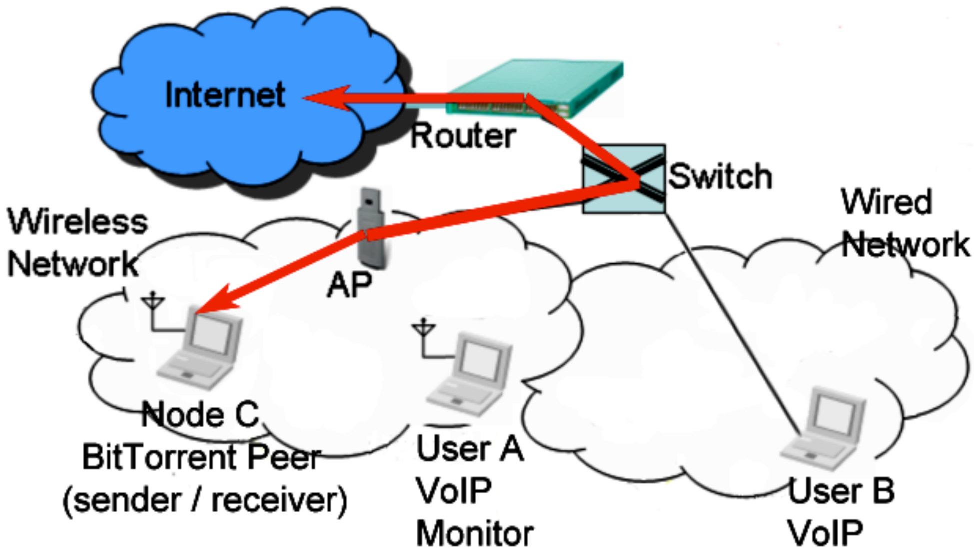
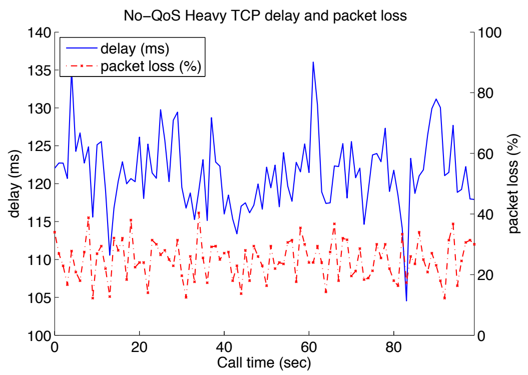
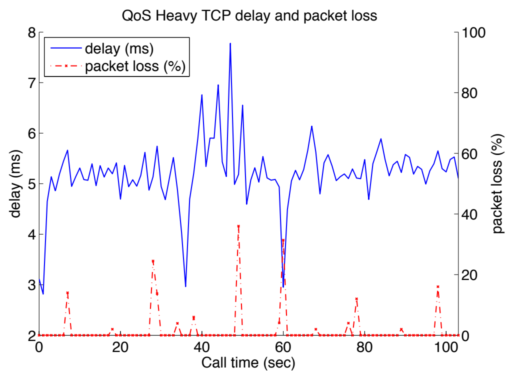


FIGURE 2.5: Heavy TCP traffic scenario. Node C exchanges BitTorrent traffic with Internet peers (both uplink and downlink traffic).



(a) No-QoS Heavy TCP delay and packet loss.



(b) QoS Heavy TCP delay and packet loss.

FIGURE 2.6: Heavy TCP delay and packet loss.

splits the files into small chunks and simultaneously downloads and uploads the shared chunks. In general, the number of generated flows in BitTorrent is high, often causing low-end routers to run out of memory and CPU. As in the previous scenarios, the AP operates in IEEE802.11b mode. The BitTorrent protocol introduces a high number of small TCP flows in both uplink and downlink directions, contending for the medium. This behaviour puts stress on the queue, CPU and memory of APs.

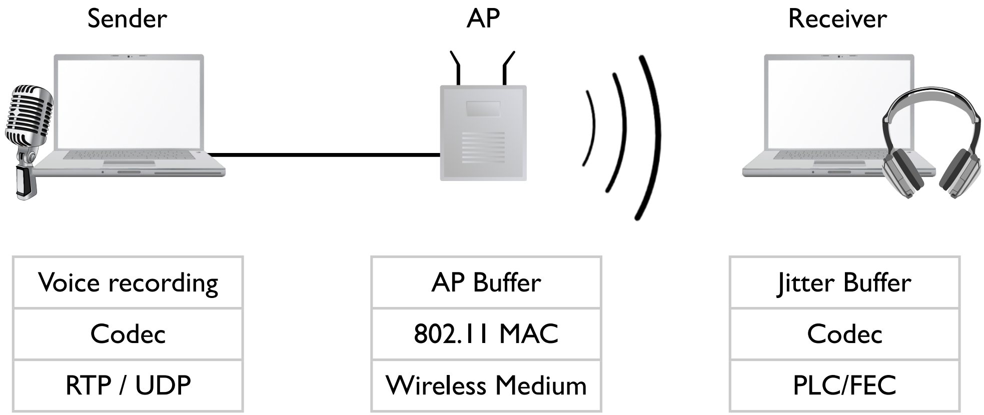


Figure 2.7: The testbed and the most important components.

2.4 Hardware setup

Two laptops are the VoIP endpoints. They both run an H323 client, X-Meeting since the laptops are running OSX, could be Ekiga for Linux or NetMeeting for Windows. One laptop is connected via FastEthernet to the testbed (the Sender) and one is wireless (The Receiver). The Receiver initiates a VoIP call. The Client accepts the call and as soon as the VoIP call is established, starts the playback of the voice recording. Soundflower is used to create a virtual audio interface, which is used as an output device from the VLC media player that is playing the voice recording and as an input device by X-Meeting. This is done in order to "reroute" the playback of the recording into the VoIP call. Ekiga on Linux has the facility to replay an audio file into the call, without the hassle of a virtual audio interface. The VoIP clients use the H323 signalling protocol and the G.711 codec (64kb/s) for the VoIP calls, sending one packet every 20ms. The packet is 200B long, with a 160B RTP payload, a 12B RTP header, a 8B UDP header, and a 20B IP header. H323 was selected over SIP, since H323 allows for easy, direct end-to-end calls with no need for software or hardware infrastructure (*e.g.*, SIP registrar).

Chapter 3

Measurements and evaluation methodology

This Chapter outlines the measurements and evaluation methodology, the monitoring mechanisms that are used to collect traces, the traces collected and the data processing. We present a novel model for packet-loss based reference samples and a method for selecting representative segments from a lengthy VoIP call. The description of the auditory tests methodology follows. Finally, we provide a brief description of the E-Model.

3.1 Data collection

To ensure synchronised traces, right before each call, both endpoints run `ntpdate` to synchronise their clocks. Both endpoints use tcpdump to capture all traffic through the used network interfaces. The network interface of the wired VoIP client captures packets using tcpdump:

```
tcpdump -i en0 -p -w wired_ethernet.pcap
```

On BSD-based systems the network adapters are noted as `enX`. In our case, the Ethernet interface is noted as `en0` and the IEEE802.11 interface as `en1`. The parameter `-i` selects the appropriate interface, `-p` disables promiscuous mode and `-w wired_ethernet.pcap` writes the captured packets to the file `wired_ethernet.pcap`.

The network interface of the wireless VoIP client captures packets in promiscuous mode with IEEE802.11+Radiotap pseudo-header provided by libpcap, using tcpdump with the appropriate settings:

```
tcpdump -i en1 -I -y IEEE802_11_RADIO -w monitor.pcap
```

This argument stream is specific for BSD-based systems. `-I` puts the interface in monitor mode and `-y IEEE802_11_RADIO` uses the IEEE802.11+radiotype data link type, recording the IEEE802.11 MAC header and various physical layer parameters, such as the RSSI value for each packet, the data rate, and the operating channel. For a crude validation of the RSSI values, we used our spectrum analyser and compared the recorded values with its output. The fluctuation of the RSSI values was within 3dBm. A list of known data link types for BSD-based systems can be obtained by:

```
tcpdump -i en1 -I -L
```

Capturing packets in monitor mode leads to deauthentication from the AP in most operating system and wireless adapters. We have been able to verify that we can capture packets in monitor mode while not forcing a deauthentication from the AP by using OSX with a Broadcom WiFi adapter or Linux with an Atheros WiFi adapter and the ath5k driver. On Linux one can configure a virtual interface mirroring the wireless interface in monitor mode:

```
iw dev wlan0 interface add mon0 type monitor
```

In this case tcpdump should be configured to use the mon0 interface. A list of known data link types for Linux systems can be obtained by:

```
tcpdump -i mon0 -L
```

3.2 Voice recording

A recording of a female voice is replayed from the Sender. The recording has a duration around 1:30 minutes during which the female voice rehearses the following passage:

> "In language, infinitely many words can be written with a small set of letters. In arithmetic, infinitely many numbers can be composed from just a few digits with the help of the symbol zero, the principal of position and the concept of base.
>
> Pure systems with base five and six are said to be very rare, but base-20 occurs in English when we use score, as in four-score and seven.
>
> Eventually, no system could keep pace with the decimal or arabic numbering system which has ten numerals, the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 and a decimal point. The numerals take different place values, depending on position. So the number 819.65 can be shown as 8 times 10 to the 2, plus 1 times 10 to the 1, plus 9 times 10 to the 0, plus 6 times 10 to the -1, plus 5 times 10 to the -2.
>
> Monetary systems have evolved to make use of this base-10 notation. France became the first decimal country in Europe in 1799, joined by Belgium, Italy and Switzerland in 1865. Germany's decision followed 8 years later and the Scandinavian states and Russia changed in 1875."

The voice speaks clearly, in a moderate speed, and with a faint British accent.

3.3 Data processing

The initial traces are in binary `pcap` format. Wireshark is used to extract the RTP packet information for each call using the built-in RTP disassembler. The information extracted is saved to a `csv` file, one for every call, that each line contains:

```
<Pcap packet No> <RTP sequence number> <RTP timestamp> <Time>
```

`Tshark` is used to extract the timestamp, RTP sequence number and RSSI for the received RTP packets at the monitor:

```
tshark -r monitor.pcap -Tfields -e frame.time -e rtp.seq \
-e radiotap.dbm_antsignal rtp > rssi.txt
```

The perl script `parse_call_csv.pl` parses the `csv` files from both endpoints, computes the delay of each packet, corresponds each packet to its RSSI and indicates the position of lost packets in the RTP stream. It outputs a plain `txt` that each line contains:

```
<RTP sequence number> <Lost packet (Boolean)> <Delay (ms)> <RSSI (dBm)>
```

The `txt` files are then imported to MATLAB for further statistical analysis. A MATLAB script is used to compute the jitter, packet loss burst interarrival, packet loss burst size and BurstR.

3.4 Network parameters

In this Section we present the network parameters during the data processing phase. The Sender and Receiver are the two VoIP clients of the Figures 2.1, 2.3, 2.5, depending on the scenario.

3.4.1 Delay

The delay is the end-to-end delay of a packet and is defined as:

$$D(x) = T_r(x) - T_s(x) \tag{3.1}$$

where $D(x)$ is the end-to-end delay of packet x, $T_r(x)$ is the time that packet x was captured at the receiver, and $T_s(x)$ is the time that packet x was captured at the sender. The accuracy of the measurements depends on the accuracy of NTP (1ms [37]). The packets from the sender are paired up with the packets from the receiver using the RTP sequence number.

3.4.2 Jitter

In the VoIP context, typically the jitter corresponds to the interarrival jitter and is defined as:

$$j_i(x) = T_r(x) - T_r(x-1) \tag{3.2}$$

where $j_i(x)$ is the interarrival jitter of packet x and $T_r(x)$ is the time that packet x is received. By definition $j_i(1) = 0$. We define the delay jitter to be:

$$j_d(x) = D(x) - D(x-1) \tag{3.3}$$

where $j_d(x)$ is the delay jitter and $D(x)$ is the end-to-end delay of packet x. Since our VoIP application sends a new packet every 20ms:

$$\begin{aligned}
j_i(x) &= T_r(x) - T_r(x-1) \\
j_i(x) &= (T_s(x) + D(x)) - (T_s(x-1) + D(x-1)) \\
j_i(x) &= (T_s(x-1) + 20\text{ms} + D(x)) - (T_s(x-1) + D(x-1)) \\
j_i(x) &= D(x) - D(x-1) + 20\text{ms} \\
j_i(x) &= j_d(x) + 20\text{ms} \qquad (3.4)
\end{aligned}$$

For the rest of this work, *jitter* is defined as *delay jitter*.

3.4.3 Packet loss burst size

A packet loss burst is defined as a series of consecutive lost packets. The packet loss burst size is the number of the consecutive lost packets.

3.4.4 Packet loss burst interarrival

The packet loss burst interarrival is defined as the number of packets received between two consecutive packet loss bursts.

3.5 Methodology for the subjective tests

This Section outlines the methodology for the subjective tests. Specifically, the main steps of this methodology include the following:

- Select a number of representative segments of the original recording. Such segments ar or relatively short duration (as described in Subsection 3.5.1).
- Normalise the volume of every segment and upsample them to the same sampling frequency.
- Create a number of reference samples of the original recording with various packet losses uniformly distributed (as described in Subsection 3.5.2). The original recording (of 0% packet loss) is also used as a reference.
- Select the control segments that will form the *consistency test* (Subsection 3.5.3).
- Place the reference samples in random order. There is no discernible pattern in the placement of the segments and the users are unaware of the network condition portrayed in each segment.
- Provide a questionnaire with instructions (an english version can be found in Appendix A).
- Decide about the auditory test environment to be used by each subject of the auditory test. Use a laptop with a set of headphones replaying the segments at a fixed volume level.
- Discard the scores of the subjects that fail the consistency test.

The following Subsections describe some aspects of the subjective test methodology in more detail.

3.5.1 Selection of segments

The voice recording has a duration of approximately 90 seconds in order to exceed the time required to capture the full effect of a handover. The same recording is used for consistency for both the heavy UDP and heavy TCP scenarios. However, the length of the recording is unsuitable for subjective MOS tests; subjects listening to long recordings tend to forget prominent characteristics that occur at the beginning of the recording and grade the last 15 seconds or so, with the exception of sudden and dramatic changes of the quality. This creates possible bias in the evaluation process.

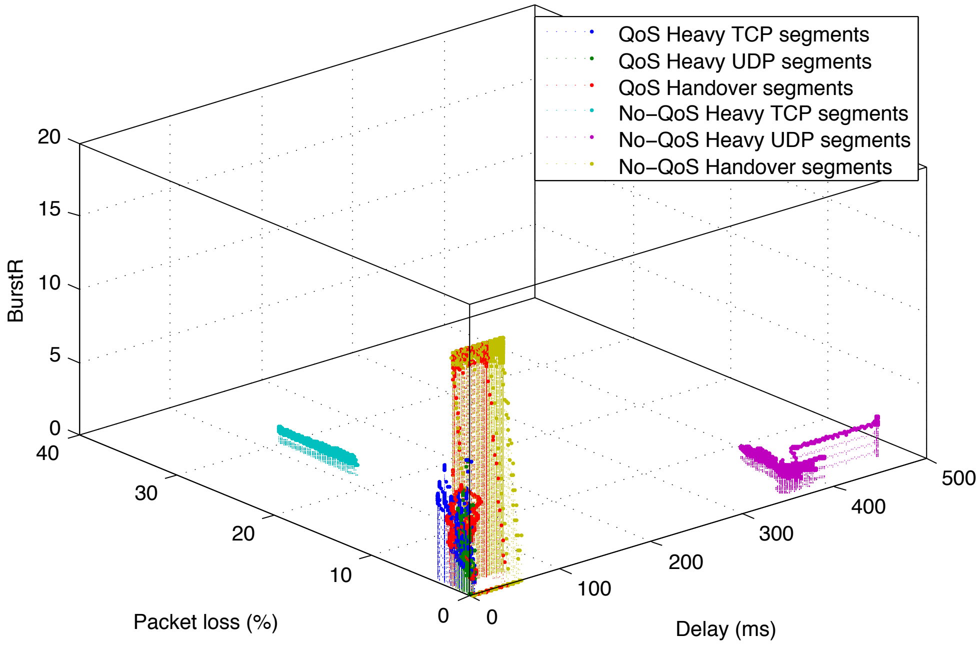


Figure 3.1: Segment constellations.

To prevent such problems, we select 15-second segments representative of the entire call, by computing the Euclidian distance of the entire call and every possible segment characteristic vector. For a VoIP application transmitting one packet every 20ms, a 15-second segment corresponds to 750 packets. The metric that was employed for the selection process was based on the mean delay, packet loss % and BurstR. Therefore, using these parameters, each representative segment shares the same E-Model score with the entire call. Figure 3.1 shows the constellations of all segment characteristic vectors for calls under every scenarios. The axes of the Figure are the delay in milliseconds, the packet loss and the BurstR, and each point corresponds to a characteristic vector. Figures 3.2, 3.3, 3.8, 3.5, 3.6, and 3.7 show the characteristic vectors of an entire call and all segments for every scenario. The axes of the Figures are the delay in milliseconds, the packet loss % and the BurstR. Each data point corresponds to a segment characteristic vector. The red star corresponds to the entire call. Another criterion for the selection is that an audio segment needs to be "coherent". Given that it is not always possible to select segments that correspond to complete sentences, we decided to accept those that contain intact words and the phrase does not end abruptly.

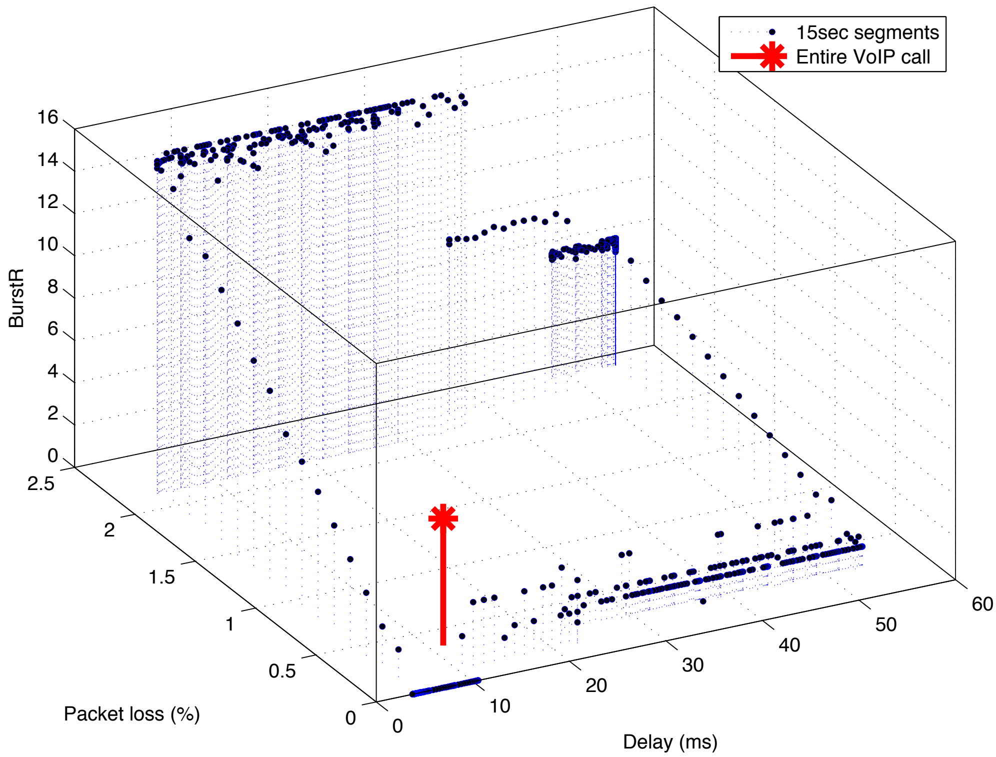


FIGURE 3.2: Segments of QoS-disabled Handover.

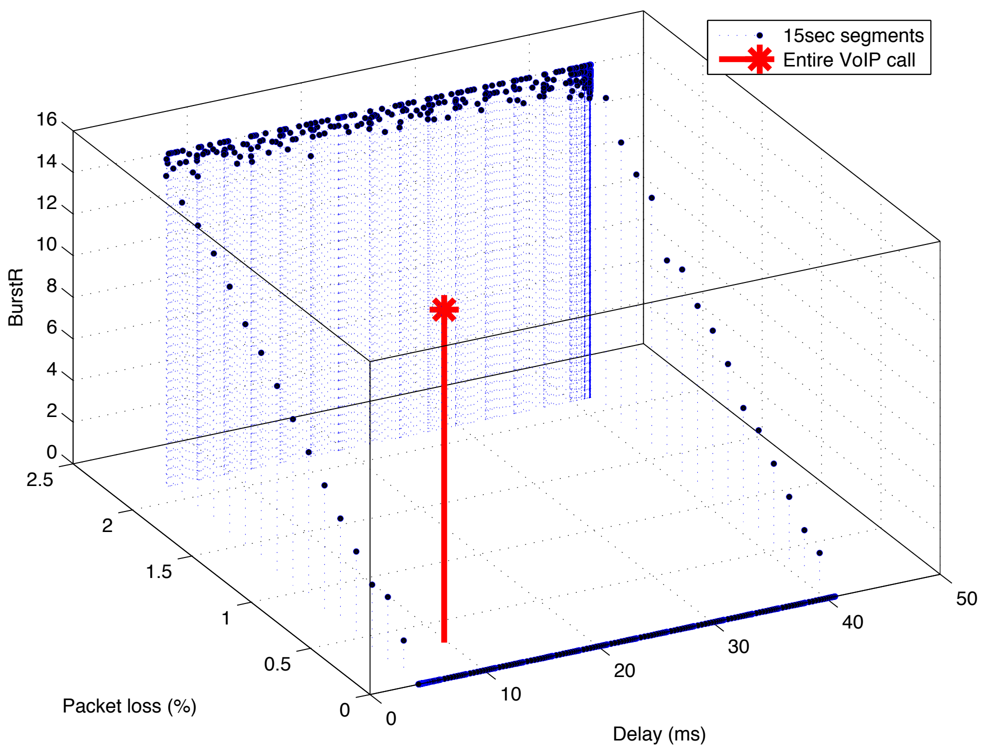


FIGURE 3.3: Segments of QoS-enabled Handover.

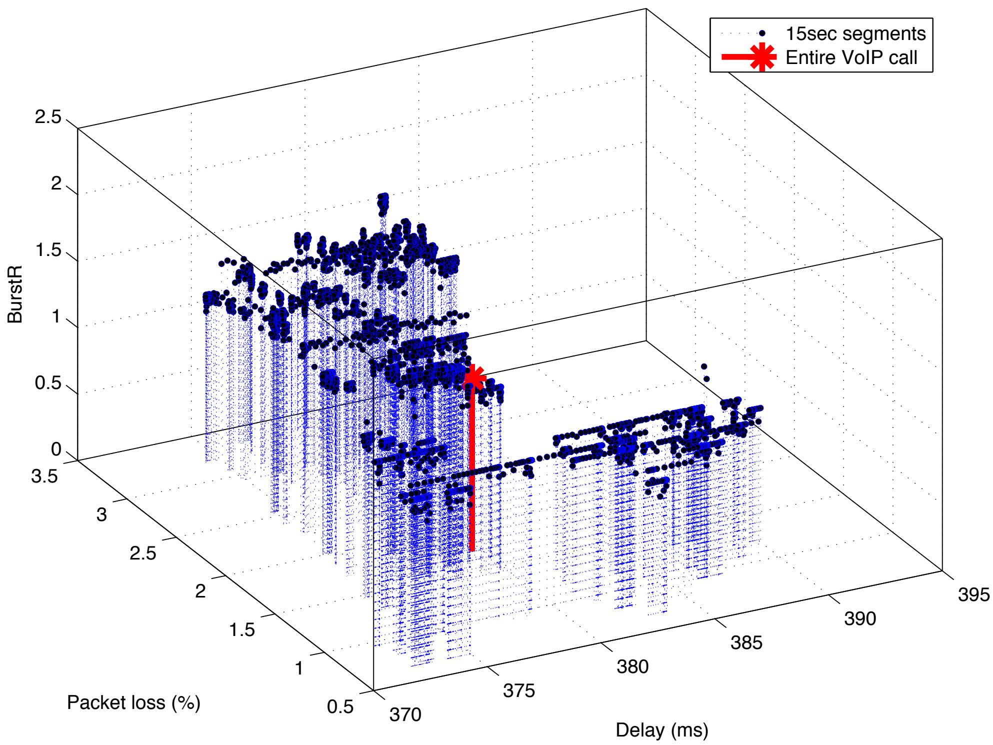


FIGURE 3.4: Segments of QoS-disabled Heavy UDP.

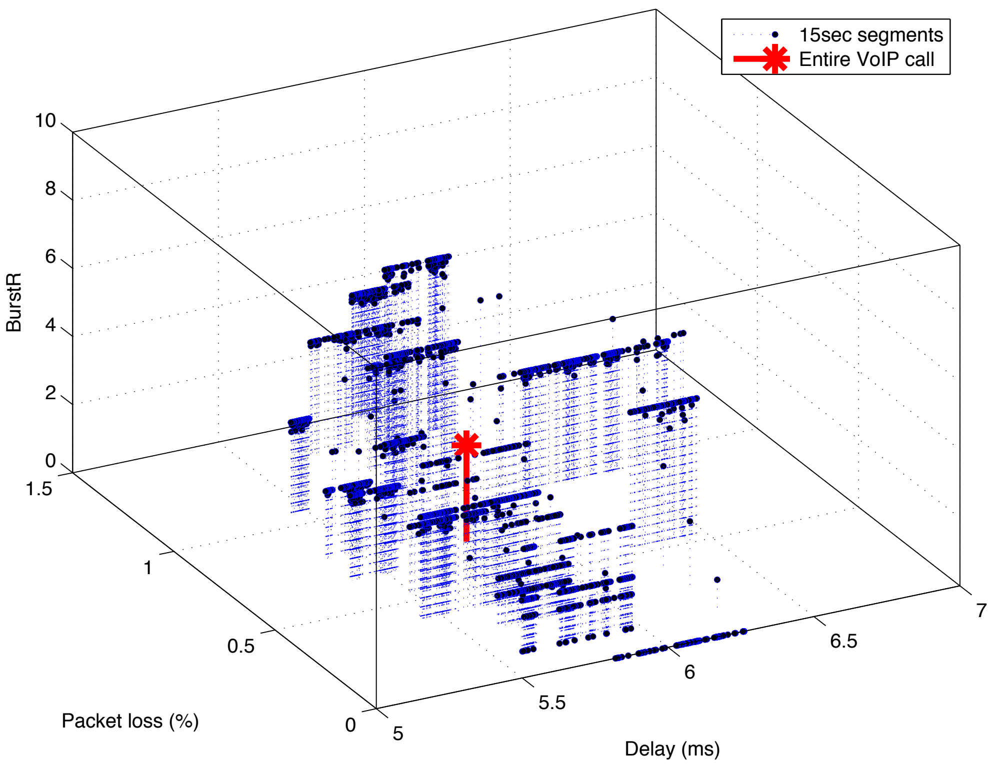


FIGURE 3.5: Segments of QoS-enabled Heavy UDP.

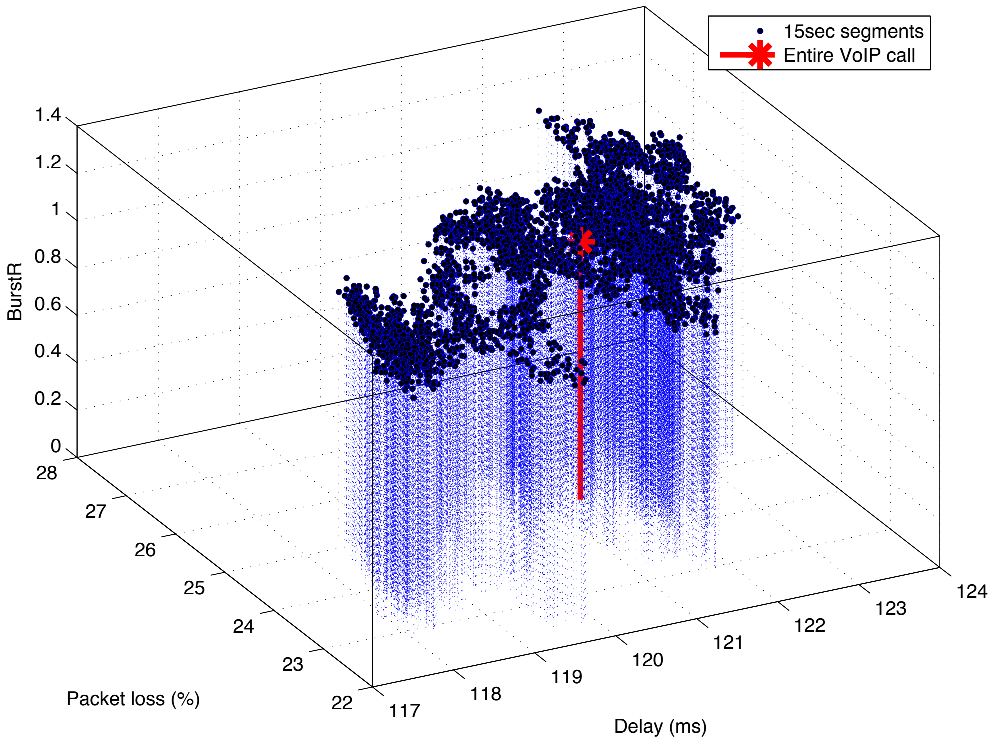


FIGURE 3.6: Segments of QoS-disabled Heavy TCP.

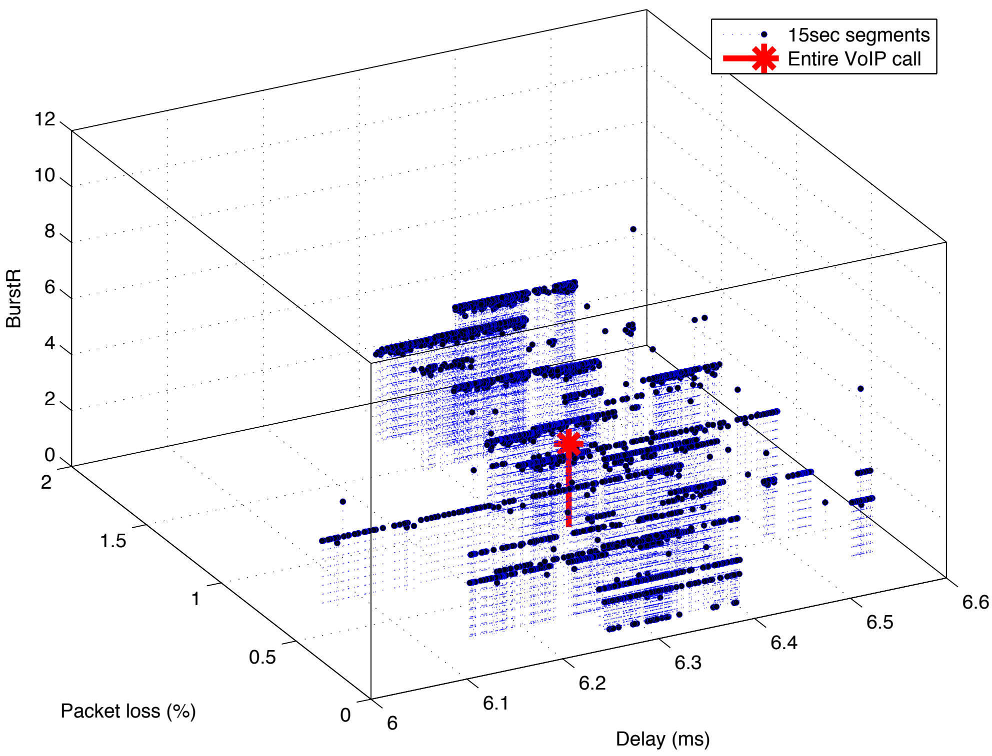


FIGURE 3.7: Segments of QoS-enabled Heavy TCP.

3.5.2 Reference samples

The Modulated Noise Reference Unit (MNRU [38]) is a noise model used widely for reference conditions for telephony applications. Modulated noise is introduced to the signal and the signal-to-noise level is referenced to equivalent mean opinion score (MOS). We argue that MNRU is not applicable in user studies of packetised speech under packet loss. MNRU introduces noise, similar to quantisation error or equipment impairment, whereas the system at focus has gaps in audio and possibly error compensation leading to loss of voice or robotised voice. The use of MNRU produced reference samples in the subjective auditory test would introduce a new type of audio degradation, irrelevant to the types prominent in the selected segments. This can confuse the subjects and lead them to grade with bias.

We propose the Modulated Packet Loss Reference Unit (MPLRU), a novel reference sample model for use in subjective auditory tests of VoIP calls. MPLRU is in the same spirit with MNRU, albeit with modulation of the packet losses instead of the signal-to-noise ratio. Packet loss following a *uniform distribution* for various packet loss rates is introduced to the same samples that are to be graded by the subjects. Although, in a production network, packet losses are often not uniformly distributed, uniform distribution of packet losses in the context of audio reference samples provide a qualitatively constant degradation. The subjects also grade the MPLRU samples. The interpolated curve produced can be used as a reference for the grading behaviour of the specific subject group. We generate MPLRU samples with 0, 3, 6, 10, 15, 20, 25, and 30% packet loss with the same voice content.

3.5.3 Auditory tests

We follow a strict auditory test procedure according to ITU recommendations ([39], [40] and [41]), adjusted to use the MPLRU reference samples. The selected segments from Subsection 3.5.1 and the reference samples from Section 3.5.2 are converted to 16000Hz sampling rate using SoX [42], and their volume is normalised using the speech voltmeter software from ITU [43]. This is to ensure that the volume level and the available voice spectrum during playback are the same for all segments. On each segment a leading 2-second silence is added, to ensure that the subjects will wait for at least 2 seconds before

Score	**Quality**
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

Table 3.1: MOS scale.

they listen to the next segment. This way, the subjects' hearing has the opportunity to "reset" and grade the next segment without bias.

The test is split into two parts, namely the training phase and the grading phase. The subjects are asked to listen to, but not grade, the original file and the MPLRU samples, in order to become acquainted with the possible level of sound impairments. This is to ensure that the subjects know beforehand how bad the quality can be and grade the segments in a consistent manner.

For the second part, the subjects are asked to listen to all 18 selected segments and all 8 MPLRU samples in random order. 20% of the segments (*i.e.*, five segments) as well as the 0% packet loss MPLRU sample are repeated as a consistency test and these segments are called control segments. The consistency is tested by examining the grades of the control segments. If a subject grades a control segment with 2 or more points difference from the score of the original segment, or grades 4 or more control segments differently than their original counterparts, the subject is considered to have failed the consistency test. The number of the segments is limited by the desirable test duration.

The subjects are forced to listen to each segment at least two times and grade it in accordance with the widely used MOS scale (Table 3.1). Once a segment is graded, the subjects are neither allowed to listen to it again, nor change their score. The total number of segments including the MPLRU and the control samples is 32.

The duration of both parts sum up to approximately 20 minutes. In order to avoid fatigue and boredom, the subjects are asked to take a small break after the first ten minutes.

The auditory test took place in ICS-FORTH premises, in a small room to minimise echo and reverberations. The equipment used for the test was the same for all subjects. The subjects were either graduate students or researchers in ICS-FORTH. Each subject was given the greek version of the questionnaire in Appendix A. Each subject read the instructions on the questionnaire carefully and was given the opportunity to ask any question they might have. All subjects were asked to listen to the same 32 segments.

3.6 E-Model

The E-model is a psycho-acoustic computational model that takes into consideration transmission parameters to estimate the quality of the user's experience during a VoIP call. Various factors as voice loudness, background noise, equipment impairment, packetisation distortion, voice codec's robustness under packet losses, end-to-end delay and round trip time are some of the parameters producing a *R-factor*, a rating that estimates the voice quality.

The R is produced by adding the perceived impairments of different nature into one value while the MOS is the average of scores of each call from 1 to 5, given by individuals in a controlled environment.

$$R = R_o - I_s - I_d - I_{e-eff} + A \tag{3.5}$$

The R_o expresses the basic signal-to-noise ratio the user receives. It includes the loudness of the voice, the noise introduced by the circuit and by background sources. I_s represents voice specific impairments, such as: too loud speech level (non-optimum OLR), non-optimum sidetone (STMR) and quantization noise (qdu). The term I_d represents the impairments introduced by delay and echo effects. I_{e-eff} is the equipment impairment factor, which represents impairments caused by low bit-rate codecs and the effect due to packet losses. A is an "advantage factor" that depends on the information regarding how users have for trading voice quality over convenience. All factors are extensively analysed in ITU-T's G.107 recommendation (E-model), but we will briefly describe the ones affected by different network conditions.

The delay impairment factor I_d, representing all impairments due to delay is composed of:

$$I_d = I_{dte} + I_{dle} + I_{dd} \tag{3.6}$$

where I_{dte} accounts for Talker Echo and I_{dle} for Listener Echo. The I_{dd} represents the impairment when absolute delay T_a is too long. This can happen even with perfect echo cancellation. The I_{dd} is defined as follows:

For $T_a \leq 100$ ms,

$$I_{dd} = 0 \tag{3.7}$$

while for $T_a > 100$ ms

$$I_{dd} = 25\left\{\left(1 + X^6\right)^{\frac{1}{6}} - 3\left(1 + \left[\frac{X}{3}\right]^6\right)^{\frac{1}{6}} + 2\right\}, \tag{3.8}$$

where

$$X = \frac{log\left(\frac{T_a}{100}\right)}{log2} \tag{3.9}$$

The equipment imparement factor I_{e-eff} represents the impairment introduced by faulty or lossy equipment (I_e). It accounts for the Packet-loss Probability (Ppl), the Packet-loss Robustness (Bpl) which is codec-specific and the Burst Ratio ($BurstR$). I_{e-eff} is computed using this formula:

$$I_{e-eff} = I_e + (95 - I_e) \cdot \frac{Ppl}{\frac{Ppl}{BurstR} + Bpl} \tag{3.10}$$

$BurstR$ is an index of the packet loss burstiness; it indicates whether packet losses occur closely together.

$$BurstR = \frac{\text{Average length of observed bursts in an arrival sequence}}{\text{Average length of bursts expected for the network under “random” loss}} \tag{3.11}$$

An estimated mean opinion score (MOS_{CQE}) for the conversational situation in the scale 1 - 5 can be obtained from the R-factor using the equations:

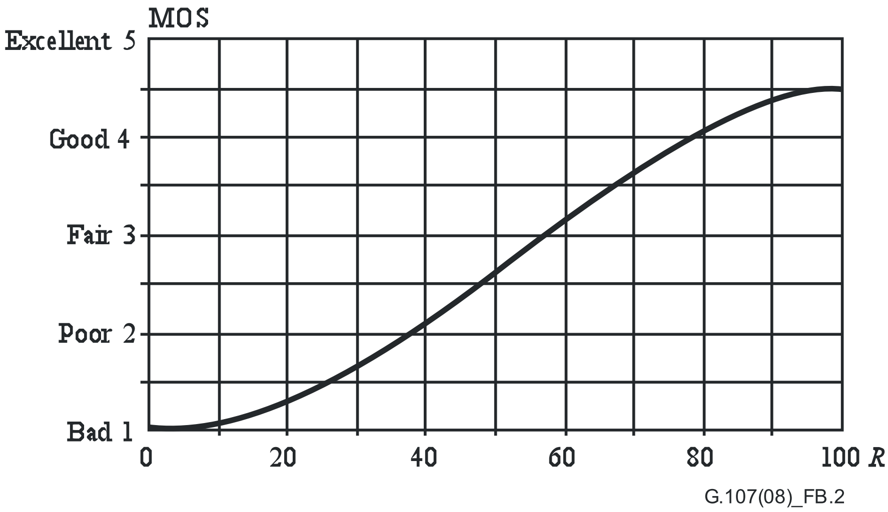


FIGURE 3.8: Estimated MOS as a function of rating factor R.

$$\begin{aligned}
\text{For } R < 0: \quad \text{MOS}_{\text{CQE}} &= 1 \\
\text{For } 0 < R < 1000: \quad \text{MOS}_{\text{CQE}} &= 1 + 0.035R + R(R-60)(100-R)7 \cdot 10^{-6} \\
\text{For } R > 100: \quad \text{MOS}_{\text{CQE}} &= 4.5
\end{aligned} \tag{3.12}$$

Chapter 4

Performance analysis

This Chapter presents the results of the statistical analysis. We identify the impact of various network parameters and traffic conditions on the perceived quality and the shortcomings of the E-Model. We also revisit and evaluate rules-of thumb for VoIP network support. Finally this Chapter presents an interesting behaviour of the packet loss burst sizes on QoS-enabled wireless networks.

4.1 Subjective tests results

35 subjects participated in the auditory tests and evaluated the selected segments according to the proposed methodology discussed in Chapter 3. Specifically, there were nine subjects that failed the consistency test. Six subjects have regraded at least one control segment with 2 or more points difference than their original score, while the other three have regraded four of the control segments with one point difference than their original score. The scores of these nine subjects were discarded. The average of the opinion scores for each segment across all subjects was computed, producing the Mean Opinion Score (MOS), as shown in Table 4.1.

4.2 Reference samples

The users evaluated the MPLRU samples in addition to the selected audio segments (Table 4.2 and Figure 4.1). The E-Model and subjective MOS of each MPLRU sample

Segment	Subjective MOS	E-Model MOS
No-QoS heavy TCP seg. 1	1.42	2.15
No-QoS heavy TCP seg. 2	1.12	2.23
No-QoS heavy TCP seg. 3	1.08	1.91
No-QoS handover seg. 1	3.58	4.39
No-QoS handover seg. 2	3.04	4.36
No-QoS handover seg. 3	2.73	4.38
No-QoS heavy UDP seg. 1	3.85	2.58
No-QoS heavy UDP seg. 2	3.27	2.91
No-QoS heavy UDP seg. 3	3.77	2.88
QoS heavy TCP seg. 1	3.85	3.93
QoS heavy TCP seg. 2	4.42	4.31
QoS heavy TCP seg. 3	4.38	4.27
QoS handover seg. 1	2.73	4.09
QoS handover seg. 2	3.96	4.25
QoS handover seg. 3	3.35	3.90
QoS heavy UDP seg. 1	4.58	4.33
QoS heavy UDP seg. 2	4.04	4.25
QoS heavy UDP seg. 3	3.54	4.35

TABLE 4.1: Subjective and E-Model MOS.

Packet loss %	Subjective MOS	E-Model MOS
0	4.42	4.41
1	4.35	4.33
3	4.12	4.13
6	2.77	3.81
10	2.50	3.41
15	2.23	2.97
20	1.54	2.63
30	1.15	2.13

TABLE 4.2: Subjective and E-Model grading of the MPLRU samples.

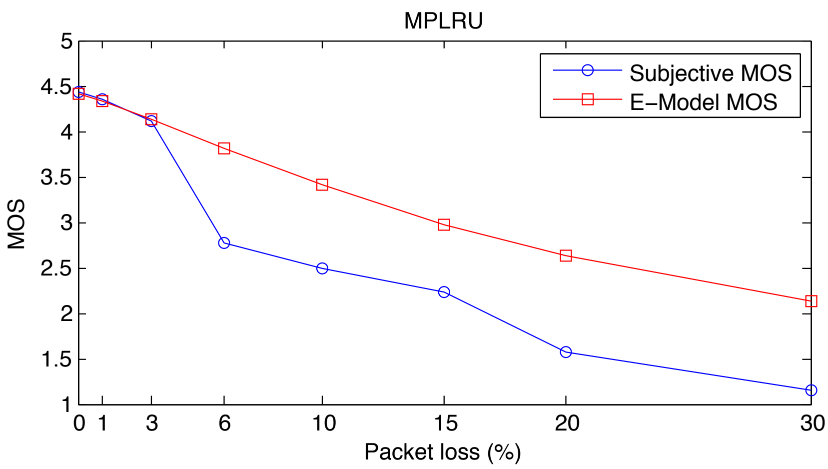


FIGURE 4.1: Subjective and E-Model grading of the MPLRU samples.

Scenarios	Grading criteria
Heavy TCP traffic	Subjective tests
Heavy UDP traffic	E-Model
Handover	

Table 4.3: List of scenarios and grading criteria.

are plotted against its packet loss rate. The users scores and the E-Model scores for packet loss up to 3% agree perfectly, supporting our E-Model parameter choice. For packet loss higher than 3% the users graded the samples with significantly lower scores than the E-Model. We believe that this outcome is due to the fact that the PLC mechanism of the codec is worse than the one that the E-Model assumes. The first significant drop of the subjective scores is at 6% packet loss and the second at 20%, whereas the E-Model scores decrease in an almost linear fashion.

4.3 Preliminary analysis

We computed the E-Model estimated MOS for each segment and paired it with the corresponding subjective MOS (Table 4.1). Table 4.3 lists the Scenarios and the grading criteria. Figure 4.2 reports the average MOS for each scenario. The E-Model overestimates the user perceived quality compared to the subjective tests. Furthermore, the QoS mechanisms improve the user perceived quality in all the cases except for the handover scenario. We discuss these results in more detail in the following paragraphs.

Within a handover (Figure 2.1 and 2.2), the client initiates an active AP discovery during which packets are queued up. The AP tries to retransmit an unacknowledged packet, until it reaches a given retransmission threshold. If this limit is reached, the AP discards the packet. Packet losses occur during the scanning phase and the handover. The QoS mechanisms of the IEEE802.11e cannot ensure the successful delivery of packets under these conditions, and thus, it cannot improve the user perceived quality. This explains the lack of improvement in the case of the handover scenario as shown in Figure 4.2.

In the VoIP under heavy TCP traffic scenario (e.g., Figure 2.5 and 2.6), the calls suffer from relatively high packet losses and delays, leading to extremely low perceived quality. The nature of the BitTorrent protocol can explain this behaviour: a BitTorrent client initiates many flows, with small payload sizes. Each flow tries to expand its TCP window,

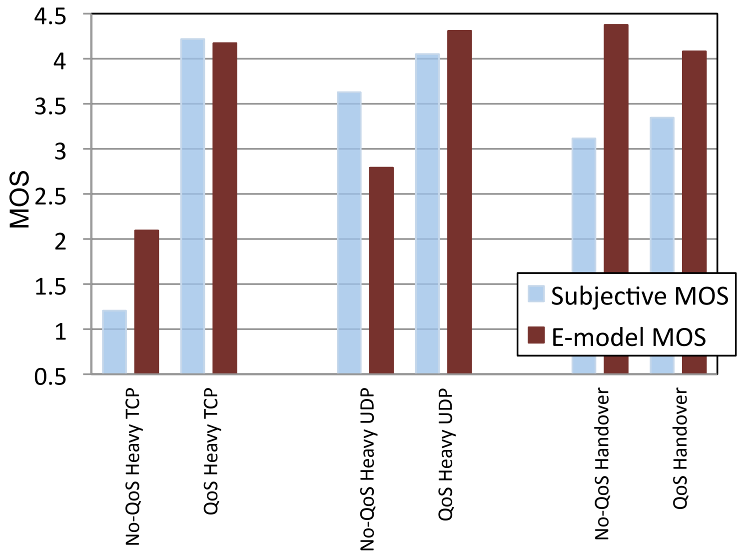


FIGURE 4.2: Subjective and E-Model MOS for different network scenarios.

up to the point that packet losses occur, triggering the TCP congestion control which will drop the throughput of that flow. Other flows active at that time will also manifest this behaviour. Since the number of flows at any given time is large, this phenomenon is repeated frequently, causing severe performance degradation (e.g., packet drops at the AP). In some calls, the large number of flows initiated by the BitTorrent client saturates the wireless LAN.

In the VoIP under heavy UDP traffic scenario (e.g., Figure 2.3 and 2.4), the MOS deteriorates due to the high packet delays. In this scenario, the very large delays are due to the presence of heavy background traffic resulting in an arrival rate higher than the 'service' rate at the AP (also observed in other studies, e.g., [44]). Indeed, a saturated network with full buffers will increase the mean delay values, trying to deliver all packets and occasionally dropping packets from the queue when a timeout occurs. The E-model reports a mediocre quality for these VoIP calls while the subjects in the auditory tests report a reasonably high opinion score value. This is due to the "unidirectional" (and non-interactive) nature of these VoIP tests. In general, this scenario highlights the need for a prioritisation scheme for different traffic classes, such as IEEE802.11e (also indicated in other studies, e.g., [3]).

As expected, the QoS mechanisms improve the user experience of VoIP calls under all background traffic scenarios. Specifically, the QoS mechanisms improve the performance

	MOS < 3.5	**MOS ≥ 3.5**
E-Model MOS ≥ 3.5	22.22%	44.44%
E-Model MOS < 3.5	22.22%	11.11%

TABLE 4.4: Subjective and E-Model MOS segment grading.

of VoIP calls under heavy UDP traffic. Especially, in the case of the E-model, their benefits are noticeable (as shown in Figure 4.2), while in heavy TCP traffic, the improvement is even more prominent, exceeding 100%.

4.4 Subjective tests and E-Model comparison

In order to compare the MOS scores of the subjective tests with E-Model, we first performed a simple preliminary test. We consider that a MOS greater or equal than 3.5 indicates "good quality", while MOS lower than 3.5 "bad quality" (Table 4.4). First, we check whether the E-Model is in agreement with the evaluation of the subjective tests. For that, we add the percentage of the segments that are graded as good or bad by both the subjective tests and the E-Model. The two grading methodologies are in agreement for the 66.66% of the segments (as shown in Table 4.4). This simple preliminary test indicates a significant difference between the subjective tests and E-Model that "calls" for further research.

The next step would be to perform a Student T-test to examine if the grading criterion is statistically important, *i.e.*, whether the E-Model MOS and the subjective MOS have statistically significant differences. It would also be interesting to examine whether the existence of the QoS mechanism and the Scenario type have a statistically significant impact on MOS. Doing multiple two-sample t-tests would result in a largely increased chance of committing a *type I* error [45]. The Analysis of variance (ANOVA) method allows for multiple tests without increasing the *type I* error.

ANOVA computes the sum of the square deviation for each component of a group (**Sump Sq.**), the group's degrees of freedom (**d.f.**), the mean square deviation (**Mean Sq.**), the F ratio (**F**), and performs an F-Test, calculating the probability that the F ratio is greater than the critical F (**Prob $>$ F**). If this probability is significantly low, the components of the group have statistically differences between them.

Source	**Sum Sq.**	**d.f.**	**Mean Sq.**	**F**	**Prob > F**
Scenario	4.9959	2	2.498	24.77	0
QoS	12.1419	1	12.1419	120.4	0
Criterion	1.277	1	1.277	12.66	0.0016
Scenario*QoS	10.1238	2	5.0619	50.19	0
Scenario*Criterion	2.4938	2	1.2469	12.36	0.0002
QoS*Criterion	0.0332	1	0.0332	0.33	0.5716
Scenario*QoS*Criterion	1.7271	2	0.8635	8.56	0.0016
Error	2.4204	24	0.1008		
Total	35.213	35			

TABLE 4.5: ANOVA analysis of Subjective and E-Model MOS per network condition.

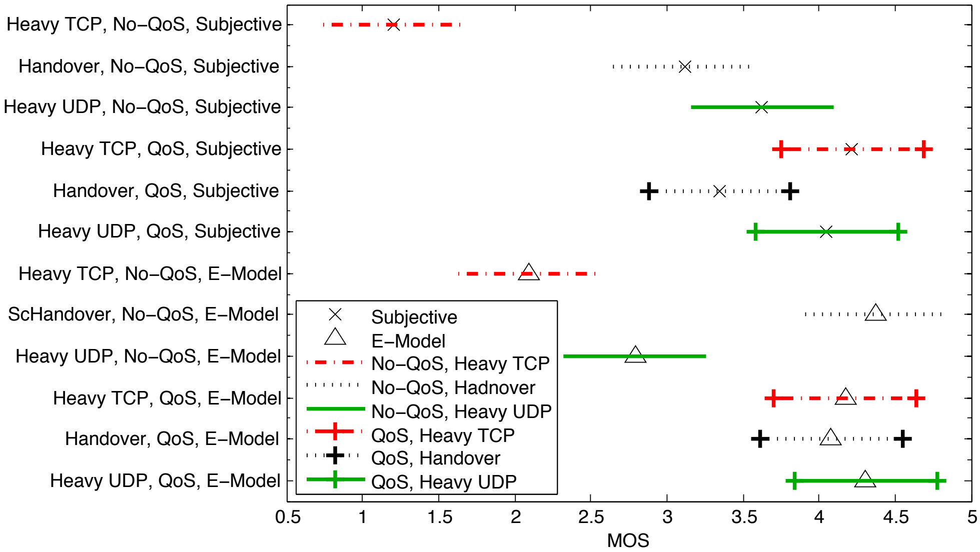


FIGURE 4.3: Tukey's HSD with three-way interactions.

We employed ANOVA to statistically analyse the differences of the subjective tests and E-Model MOS taking into consideration the QoS status and the Scenario (Table 4.5). ANOVA with three-way interactions indicates that the Scenario, QoS status, Criterion as well as the interaction of Scenario and QoS, Scenario and Criterion, and Scenario, QoS and Criterion are statistically significant and that the interaction of QoS and Criterion is not significant. We analysed these results further with Tukey's HSD test, for all possible dimensions.

The three-way interaction Tukey's HSD test (Figure 4.3) shows that the E-model tends to overestimate the perceived quality, except for the heavy UDP with no QoS scenario. There are significant statistical differences between the Heavy-TCP with no QoS enabled segments for both evaluation methods respectively. The subjective MOS values of the handover with no QoS enabled scenario are significantly different from their E-Model

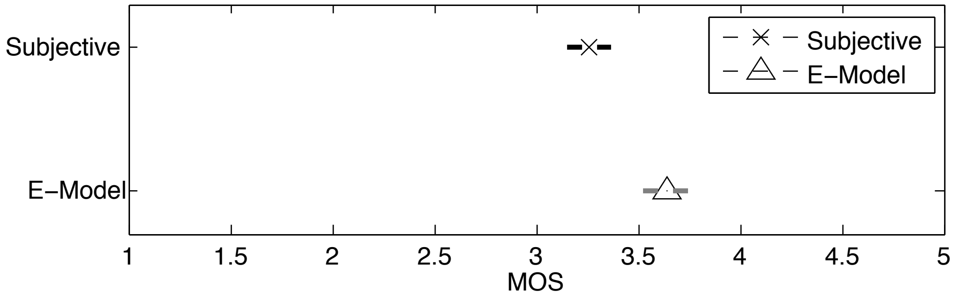


FIGURE 4.4: Tukey's HSD with criterion effect.

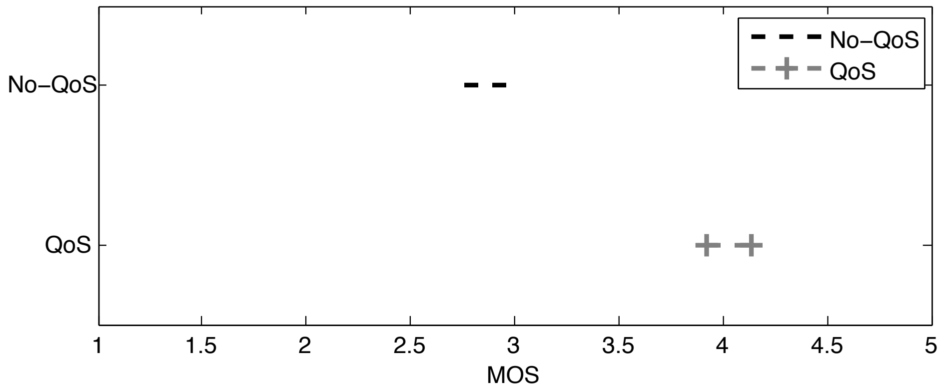


FIGURE 4.5: Tukey's HSD with QoS effect.

counterparts. This discrepancy is due to the bursts of the packet losses and their relative placement in the segment. Due to the active scanning of the client during a handover, the packet loss bursts occur closely together. This issue motivated us to examine more network related parameters as discussed in Section 4.5. Finally, QoS mechanisms seem to be ineffective for the handover scenario, whereas the background traffic scenarios benefit from QoS. We further analyse the ANOVA results with two-way and one-way interaction Tukey's HSD tests.

There are statistically significant differences between the subjective MOS and the E-Model estimated MOS (Figure 4.4). The E-Model seems to significantly overestimate the quality compared to the evaluation of the subjects. There also are statistically significant differences between the combined MOS values of the QoS-enabled and QoS-disabled testbeds (Figure 4.5). Enabling QoS makes a significant difference on the quality of experience for VoIP calls. Finally the heavy TCP scenario differs significantly from the handover and heavy UDP scenarios (Figure 4.6), since the high packet loss rates of the heavy TCP scenario took their toll on the perceived voice quality.

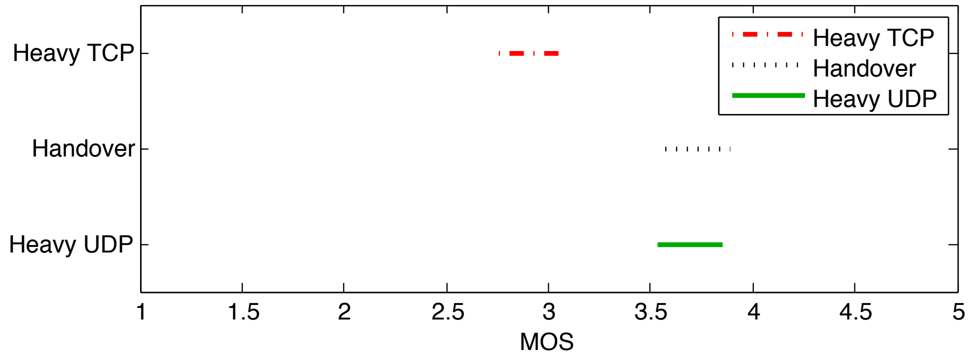


FIGURE 4.6: Tukey's HSD with Scenario effect.

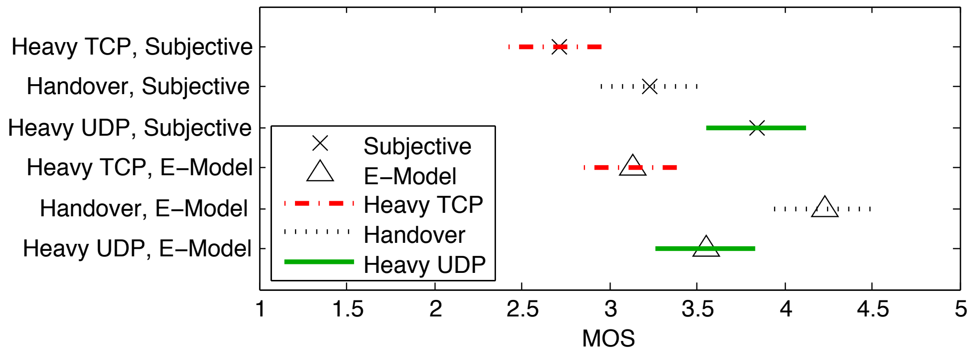


FIGURE 4.7: Tukey's HSD with scenario and criterion interaction.

We analyse the interaction of the scenario and the criterion on Tukey's HSD test (Figure 4.7), and find that the subjective handover MOS and E-Model handover MOS are significantly different. The E-model does not take into consideration the packet loss bursts occurring closely together (small packet loss burst interarrival), thus overestimating the quality of the handover scenario. While not being statistically significant, the E-model "slightly" overestimates the heavy TCP MOS and underestimates the heavy UDP MOS. The difference between the subjective MOS and the E-Model MOS of the heavy UDP scenario is due to the streaming-like nature of the VoIP calls, practically "masking" the effects of consistently long, one-way delays.

The examination of the QoS and scenario also reveals a several interesting issues (Figure 4.8). First, the no-QoS heavy TCP scenario is statistically different from every other one. Second, the no-QoS heavy UDP scenario is significantly different from the QoS-enabled background traffic scenarios. Moreover, the two handover scenarios perform similarly and do not differ significantly from the QoS-enabled background traffic scenarios. The packets during a handover are not lost due to improper MAC-layer priority assignment, but due to loss of signal while scanning for APs in different channels. Therefore, the

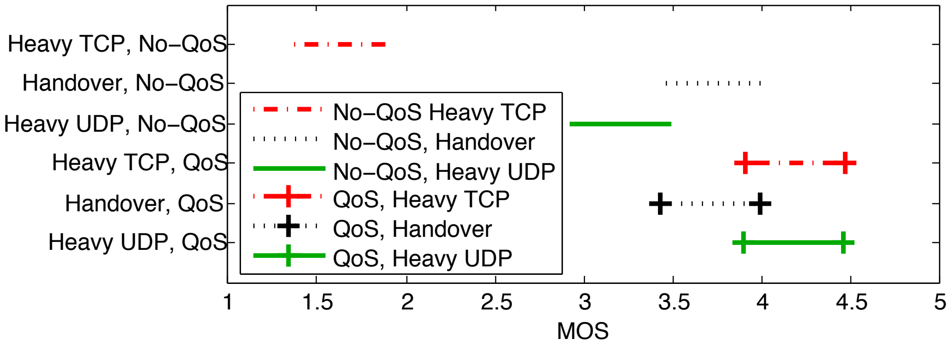


FIGURE 4.8: Tukey's HSD with scenario and QoS interaction.

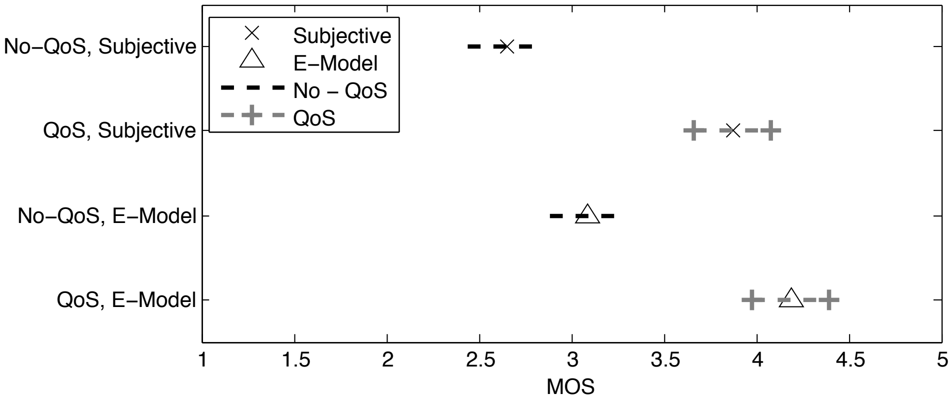


FIGURE 4.9: Tukey's HSD with QoS and criterion interaction.

support of IEEE802.11e mechanisms cannot improve this performance. Finally, the QoS heavy TCP and QoS heavy UDP scenarios have similar statistical behaviour. The IEEE802.11e assures low packet loss and delay for the voice flows, leading to MOS higher than 4. The impact of QoS on the quality of experience is once again prominent.

The E-model MOS has significant statistical differences from the subjective MOS for the QoS-disabled scenario (Figure 4.9). Both criteria have significant differences from their QoS-enabled counterparts.

4.5 Impact of various network parameters on quality

We are interested in evaluating the impact of various network parameters, namely, the packet loss, the BurstR, the delay, the jitter, the packet loss burst size and the packet loss burst interarrival. To statistically analyse the impact of the network parameters on the users opinion scores, we employed ANOVA (Table 4.6). It should be noted that

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob > F
avg delay	1.137	1	1.1368	2.47	0.1165
packet loss	23.039	1	23.0387	50.11	0
avg jitter	1.272	1	1.2718	2.77	0.0969
BurstR	0.691	1	0.6906	1.5	0.221
avg burst ia	3.552	1	3.5518	7.73	0.0057
avg burst size	0.725	1	0.7249	1.58	0.2099
var burst size	0.179	1	0.1791	0.39	0.5328
var delay	9.505	1	9.5046	20.67	0
var jitter	2.89	1	2.89	6.29	0.0125
var burst ia	0.725	1	0.7251	1.58	0.2098
Error	210.101	457	0.4597		
Total	698.197	467			

Table 4.6: ANOVA analysis of impact of network parameters on VoIP quality.

ANOVA was run on continuous variables, thus the degrees of freedom for all parameters is 1. The MATLAB's implementation of ANOVA indicates with 0 the probabilities less than 0.0001. The ANOVA analysis shows that the following parameters have significant impact on the quality:

- Packet loss
- Average packet loss burst interarrival
- Delay variance
- Jitter variance

The average delay is not included in the significant parameters list, as our VoIP calls are more streaming-like and not two-way conversations. However, the delay variance is significant, as fluctuation of the delay can fill the playout buffer of the application and lead to lost packets. It is no surprise that packet loss is an important parameter, since lost packets are responsible for loss of voice during a VoIP call. The jitter is the per packet fluctuation of delay. The average jitter is not significant, but the variance of jitter is. The BurstR is an index of the length of packet loss bursts, relative to the theoretical burst length for random packet losses for the same packet loss probability. The BurstR and the packet loss burst size seem to have lower impact than the average packet loss burst interarrival. The packet loss burst length interarrival is an index of how closely together packet loss bursts occur, which although the E-Model ignores, it can have a significant impact on user perceived quality.

Segment	Delay	Packet loss	jitter	MOS	Group
QoS heavy UDP seg. 1	4.11	0.9%	2.76	4.5	group1
QoS heavy UDP seg. 2	5.20	1.7%	3.85	4.03	group1
QoS heavy UDP seg. 3	3.75	0.7%	2.50	3.53	group1
QoS heavy TCP seg. 2	7.20	1%	1.63	4.42	group2
QoS heavy TCP seg. 3	6.31	1.5%	1.70	4.38	group2
No- QoS handover seg. 1	4.99	0.2%	0.94	3.57	group3
No- QoS handover seg. 2	4.23	0.5%	0.59	3.03	group3
No- QoS handover seg. 3	4.67	0.4%	0.69	2.73	group3

TABLE 4.7: Segments with similar network parameters.

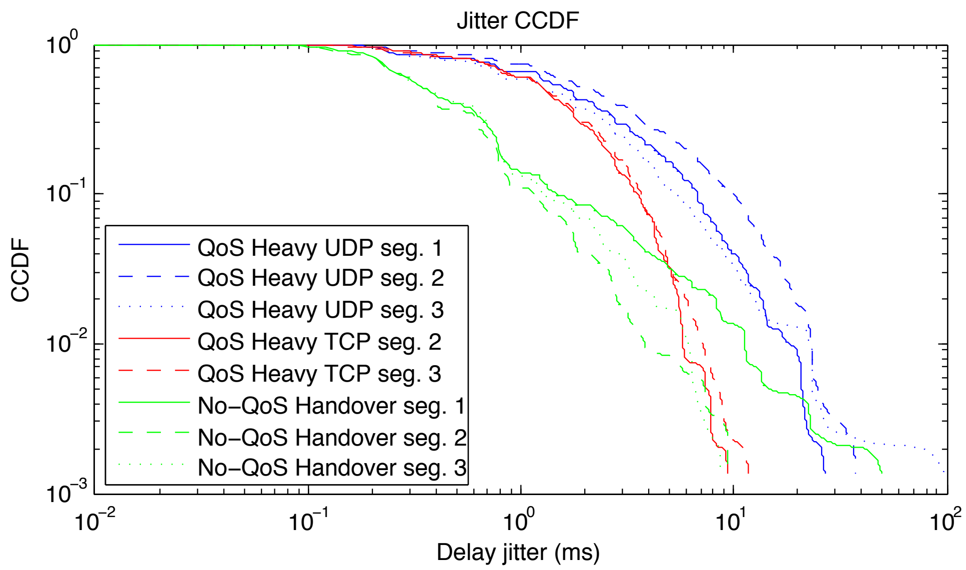


FIGURE 4.10: Segments with similar average network parameters jitter CCDF.

4.6 Impact of the statistical properties of the network parameters

The E-model includes BurstR, a statistical packet loss burst index to take into consideration the packet loss burst length, but in terms of the delay, it considers only the average delay. Are average-based values of packet loss and delay sufficient to characterise the MOS or finer-level statistics are required? To answer this question, we considered eight segments with similar average values for packet loss, delay, and jitter, while the distribution of jitter is as shown in Table 4.7 and Figure 4.10. More precisely, all selected segments have up to 1.7% packet loss, an average delay of 7.5ms or less, and an average jitter of 4ms or less. We sorted each segment to a group depending on the stochastic

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob > F
group	2.32372	2	1.16186	6.4	0.0418
Error	0.90705	5	0.18141		
Total	3.23077	7			

TABLE 4.8: ANOVA analysis of impact of network parameters statistical properties on VoIP quality.

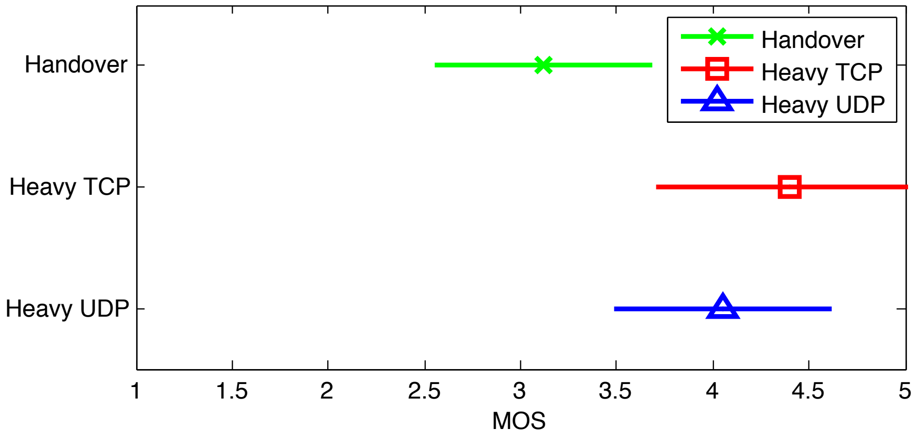


FIGURE 4.11: Tukey's HSD with group effect.

order of its jitter CCDF. The groups created are the heavy UDP group (group1), the heavy TCP group (group2) and the handover group (group3).

To test whether the groups are different in terms of their MOS we performed a null-hypothesis test using ANOVA (Table 4.8) and Tukey's HSD test (Figure 4.11). The test revealed that the handover group is significantly different from the other two, indicating that the statistical properties of the networks parameters are important and affect the perceived quality. A more extensive analysis with more data is required for more solid conclusions.

4.7 Revisiting rules-of-thumb

Typically network planning for VoIP is based on various well-known rules-of-thumb. For example, in order for a VoIP call to have good quality the packet loss must be less than 3% packet loss, the delay less than 150ms, and the interarrival jitter less than 30ms ([46], [47], and [48] . We re-examine these thresholds by indicating cases where the thresholds are either too strict or too relaxed. Since our study focused on one-way VoIP calls we

Segment	packet loss(%)	Subj. MOS	E-Model
No-QoS Heavy UDP seg. 1	4.4	3.84	2.57
QoS Heavy TCP seg. 1	4.1	3.84	3.93
QoS Handover seg. 3	4.4	3.34	3.90

Table 4.9: Segments invalidating the rule-of-thumb threshold of 3% packet loss.

Segment	jitter (ms)	Subj. MOS	E-Model
No-QoS Heavy UDP seg. 1	11.68	3.84	2.57
No-QoS Heavy UDP seg. 2	11.4	3.26	2.90
No-QoS Heavy UDP seg. 3	10.77	3.76	2.88

Table 4.10: Segments invalidating the rule-of-thumb threshold of 10ms delay jitter.

Segment	packet loss(%)	jitter(ms)	Subj. MOS	E-Model
No-QoS Handover seg. 2	0.53	0.59	3.03	4.36
No-QoS Handover seg. 3	0.4	0.68	2.73	4.37
QoS Heavy UDP seg. 3	0.66	2.5	3.53	4.35

Table 4.11: Segments with bad quality under the rule-of-thumb thresholds.

decided to ignore the delay-based rule-of-thumb. The rule-of-thumb thresholds relevant to our testbed are the "3% packet loss", and the "30ms interarrival jitter" (which equals to 10ms delay jitter). In Section 4.4, we considered that a MOS greater or equal than 3.5 indicates "good quality" while MOS lower than 3.5 "bad quality". We found counter-examples with metrics above the rule-of-thumb thresholds with good quality (Table 4.9 and 4.10) and one case with metrics below the thresholds with bad quality (Table 4.11).

4.8 QoS and packet loss bursts

We have identified an interesting behaviour of QoS-enabled wireless networks, resulting to long packet loss bursts in high-priority flows (as shown in Figure 4.12).

The handover scenario is relatively unaffected from QoS. One can see that the greatest discrepancy is for bursts of 1 to 2 packets long. This is easily explained by the behaviour of EDCA used for QoS. For packets of higher priority, EDCA has a faster retransmission rate than DCF. During a handover, a client can be out of the AP's range or scanning for other APs in other channels. On either case, the packets sent from the AP are lost. Since EDCA retransmits packets faster, it is will also be quicker to reach the retransmission

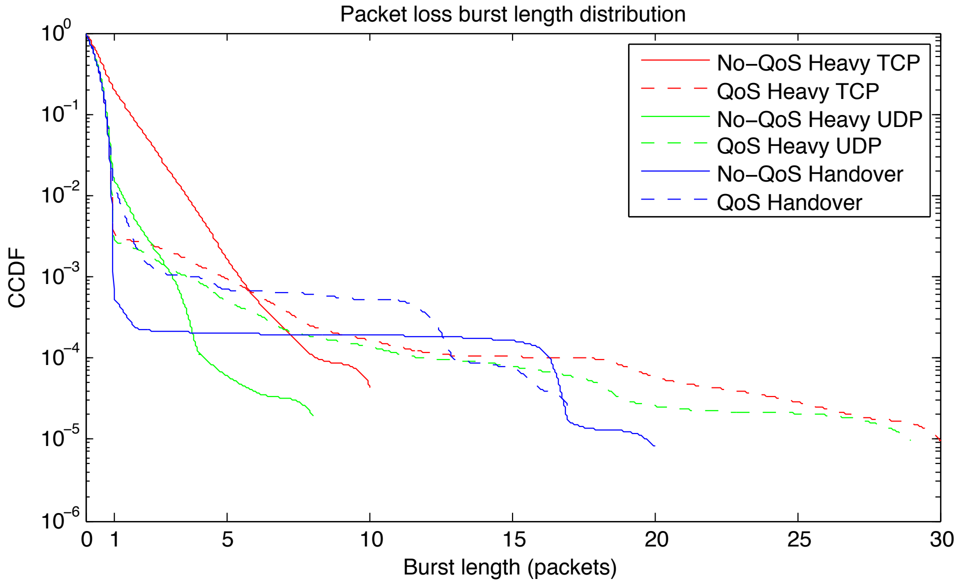


FIGURE 4.12: Packet loss burst length distribution.

threshold for a packet. This behaviour makes the loss of 1 to 2 consecutive packets more probable. The rest of the distribution does not present any significant differences.

In the case of the heavy TCP scenario though, the effects of QoS are significant. With QoS disabled, the packet losses are high (20%) and have a maximum packet loss burst length of 10 packets. On the QoS-enabled testbed though, packet losses fall to less than 0.4%. This is a significant drop with tremendous positive impact on the user perceived quality. Despite the decrease of the packet loss rate, the maximum packet loss burst length increases noticeably to 30 packets, albeit with a probability less than 0.01% for such long bursts to occur.

The behaviour of the heavy UDP scenario is similar to the heavy TCP scenario, except for the lower packet loss rate to begin with. With QoS disabled, the packet losses rate is 2% and have a maximum packet loss burst length of 8 packets. The QoS-enabled heavy UDP scenario is quite similar with the QoS-enabled heavy TCP scenario. The packet loss rate is less than 0.4% and the maximum packet loss burst length is 29 packets, with a probability of occurrence less than 0.01%.

As it was expected the QoS provided by IEEE802.11e lowers the delay to 10ms or less and packet losses. However, in some cases as in high contention or congestion scenarios, it raises the number of retransmissions and the length of packet loss bursts. One of the reasons that the community was not aware of this phenomenon was the low likelihood of the occurrence of such long packet bursts.

Chapter 5

Conclusions

This work offers an evaluation of the perceived quality of unidirectional, non-interactive VoIP calls under various wireless network conditions, namely handover and high background traffic. Its main contributions include a novel methodology for performing auditory tests, an evaluation of the user-perceived quality of VoIP, a statistical analysis for the identification of the significant network parameters, and a re-examination of various rules-of-thumb in the context of VoIP over wireless networks, by performing empirical measurements on a wireless network.

We performed a statistical analysis using ANOVA and Tukey's HSD tests. In the context of VoIP calls the analysis shows that the impact of network conditions, QoS, and their interplay on the perceived quality of experience varies significantly. The analysis highlights the benefits of the QoS mechanisms under these network conditions. VoIP calls under heavy background traffic benefit the most from the existence of QoS mechanisms, raising the MOS to 4 (good quality). However, the perceived quality under a handover does not improve with the support of QoS mechanisms; in fact, the MOS remains stationary at 3 (Fair). As shown in this analysis, not all network conditions impact the quality of VoIP applications in the same manner. Specifically, the ANOVA analysis indicates that the packet loss percentage, the average packet loss burst interarrival, the delay variance and the jitter variance have a significant effect on the perceived user quality. Understanding which network parameters cause severe impairment in the VoIP application and which cross-layer measurements can be used to predict such impairment is important in the design of adaptation mechanisms.

Our analysis reveals the inability of the E-model to capture the perceived user experience under specific network conditions, especially during a handover. A comparative evaluation of the quality of VoIP calls using E-Model and subjective tests demonstrates the need for more accurate metrics, tailored to the specific requirements of the application at hand. In our case, the E-Model fails to capture the effect transient phenomena have on the perceived quality of a VoIP call. More precisely, we indicate that the perceived quality is affected by the jitter and its variance, the distribution and the interarrival of the packet loss bursts, metrics that are ignored by the E-Model.

We revisited situations in which common “rule-of-thumb” metrics for packet loss and jitter cannot reflect the user-perceived quality. Our experiments reveal test cases that both thresholds of “3% packet loss” and “10ms delay jitter” fail to estimate the perceived quality correctly.

We presented a novel methodology for performing subjective auditory tests tailored for lengthy VoIP calls. In order to capture the full effects of a handover in a real testbed on the quality of a VoIP call, the call must have a duration of more than 60 seconds. However, this duration is unsuitable for subjective tests. We described a methodology that selects the representative 15-second segments of the entire call based on the delay and packet loss characteristics of the packets of the call. It then employs these segments in the subjective tests. A novel reference sample model was introduced, applying to the packet-loss based audio impairments caused by the wireless network conditions.

Finally, we have identified an interesting behaviour of QoS-enabled wireless networks, resulting to long packet loss bursts in high-priority flows. The QoS provided by IEEE802.11e lowers the delay to 10ms or less and packet losses, improving the user perceived quality significantly. However, in some cases as in high contention or congestion scenarios, it raises the number of retransmissions and the length of packet loss bursts.

Chapter 6

Future work

This work can be extended in the following directions: First, we will further research the network parameters that have significant impact on the perceived quality. We have identified the packet loss, the variance of the delay jitter as well as the packet loss burst interarrivals to affect the user perceived quality. We plan to perform an extended study of various network parameters with more experiments, to better understand their effect and more accurately identify the important ones.

Second, enhancements can be made to the E-Model to take into consideration the significant network parameters, such as the distribution of packet loss bursts. In that way E-Model can provide a more accurate estimation of the user perceived quality under various network conditions. Also, it will be interesting to investigate the use of the E-Model to estimate only the perceptual effect of the delay, in order to facilitate for streaming, real-time communications and will revisit the use of PESQ to provide an estimate of the qualitative degradation.

We intend to repeat this analysis with more codecs in a variety of real testbeds. Our previous work highlighted the benefits of the packet loss concealment of the AMR 12.2kb/s. We will build on that work and employ popular and state of the art codecs, such as Silk, iLBC and Speex. This research was based on experiments performed in a real testbed in ICS-FORTH. We plan to extend the analysis on even more realistic environments, such as a hotel, an airport or a home network, employing different devices, such as smartphones and bluetooth headsets.

It would be interesting to statistically analyse the ability of various cross-layer measurements, such as MAC retransmissions and delays, type of control packets, and SNR values, to predict when the MOS will reach a certain threshold and the current network condition in order to trigger an adaptation process. For example, a continuous increase of the average delay in the presence of UDP background traffic will likely result in low MOS. A smoother degradation in the MOS is more likely in the presence of a few TCP flows as background traffic.

Appendix A

Auditory test questionnaire

Please, listen to ALL audio samples.

You should first listen to the audio sample named REFERENCE, and the TRAINING 1 - TRAINING 8 afterwards. During this phase you will familiarise with the possible audio quality degradation. You will not grade these samples.

Afterwards, you will listen to the rest of the audio samples and grade their quality, in accordance with the following scale:

Score	**Quality**
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

You must listen to the audio samples **at least two times**. You can listen to an audio segment again if needed. Please make sure that you listen to the audio samples **in the given order and not listen to an audio sample you have already graded. When you have listened to the samples A1 - A14, you should take a 2-minute break.**

Grades:

A1		B1		B10	
A2		B2		B11	
A3		B3		B12	
A4		B4		B13	
A5		B5		B14	
A6		B6		B15	
A7		B7		B16	
A8		B8		B17	
A9		B9		B18	
A10					
A11					
A12					
A13					
A14					

Thank you for you participation!

Bibliography

[1] Sangheon Pack, Jaeyoung Choi, Taekyoung Kwon, and Yanghee Choi. Fast handoff support in IEEE802.11 wireless networks. *IEEE Communications Surveys and Tutorials*, 9(1):2–12, 2007.

[2] Haitao Wu, Kun Tan, Yongguang Zhang, and Qian Zhang. Proactive scan: Fast handoff with smart triggers for 802.11 wireless LAN. In *IEEE INFOCOM*, Anchorage, Alaska, May 2007.

[3] Choi Sunghyun, Prado Javier, Shankar Sai, and Mangold N. Stefan. IEEE802.11e contention-based channel access (EDCF) performance evaluation. In *IEEE International Conference on Communications*, Anchorage, Alaska, May 2003.

[4] Alan Clark. Extensions to the E-Model to incorporate the effects of time varying packet loss and recency. T1A1.1/2001-037, April 2001.

[5] ITU. ITU-T recommendation G.113: Transmission impairments due to speech processing, 2007.

[6] Sangho Shin and Henning Schulzrinne. Experimental measurement of the capacity for VoIP traffic in IEEE802.11 WLANs. In *IEEE INFOCOM*, Anchorage, AK, USA, May 2007.

[7] Andrea Forte, Sangho Shin, and Henning Schulzrinne. Improving layer-3 handoff delay in IEEE802.11 wireless networks. In *ICST WICON*, Boston, Massachusetts, August 2006.

[8] I Ramani and S Savage. Syncscan: practical fast handoff for 802.11 infrastructure networks. In *IEEE INFOCOM*, volume 1, pages 675 – 684, Miami, FL, USA, March 2005.

[9] Kostas Pentikousis, Esa Piri, Jarno Pinola, Frerk Fitzek, Tuomas Nissilä, and Ilkka Harjula. Empirical evaluation of VoIP aggregation over a fixed WiMAX testbed. In *ICST TRIDENTCOM*, Austria, March 2008.

[10] Z. Li, L. Sun, Z. Qiao, and E. Ifeachor. Perceived speech quality driven retransmission mechanism for wireless VoIP. In *IEE 3G*, London, UK, June 2003.

[11] Patrick Verkaik, Yuvraj Agarwal, Rajesh Gupta, and Alex C. Snoeren. Softspeak: making VoIP play well in existing 802.11 deployments. In *USENIX NSDI*, Boston, MA, April 2009.

[12] A. Arjona, C. Westphal, A. Ylä-Jääski, and M. Kristensson. Towards high quality VoIP in 3G networks: An empirical study. In *IEEE AICT*, Athens, Greece, June 2008.

[13] Samrat Ganguly, Vishnu Navda, Kyungtae Kim, Anand Kashyap, Dragos Niculescu, Rauf Izmailov, Sangjin Hong, and Samir R. Das. Performance optimizations for deploying VoIP services in mesh networks. *IEEE Journal of Selected Areas of Communications*, 24(11):2147–2158, 2006.

[14] Farooq Anjum, Moncef Elaoud, David Famolari, Abhrajit Ghosh, and Ravichander Vaidyanathan. Voice performance in WLAN networks - an experimental study. In *IEEE GLOBECOM*, San Francisco, December 2003.

[15] David P. Hole and Fouad A. Tobagi. Capacity of an IEEE 802.11b wireless lan supporting VoIP. In *IEEE ICC*, Paris, France, June 2004.

[16] Pablo Vidales, Niklas Kirschnick, Blazej Lewcio, Frank Steuer, Marcel Wältermann, and Sebastian Möller. Mobisense testbed: Merging user perception and network performance. In *ICST TRIDENTCOM*, Innsbruck, Austria, March 2008.

[17] Sebastian Möller, Marcel Wältermann, Blazej Lewcio, Niklas Kirschnick, and Pablo Vidales. Speech quality while roaming in next generation networks. In *IEEE International Conference on Communications*, Dresden, Germany, June 2009.

[18] Blazej Lewcio, Marcel Wältermann, Pablo Vidales, Alexander Raake, and Sebastian Möller. Performance of instrumental speech quality measures for next generation wireless networks. In *IEEE International Conference on Acoustics*, Rotterdam, March 2009.

[19] Blazej Lewcio, Marcel Wältermann, Sebastian Möller, and Pablo Vidales. E-model supported switching between narrowband and wideband speech quality. In *First International Workshop on Quality of Multimedia Experience (QoMex)*, San Diego, July 2009.

[20] Kuan-Ta Chen, Chun-Ying Huang, Polly Huang, and Chin-Laung Lei. Quantifying skype user satisfaction. In *ACM SIGCOMM*, Pisa, Italy, August 2006.

[21] Christian Hoene, Holger Karl, and Adam Wolisz. A perceptual quality model intended for adaptive voip applications. *International Journal of Communication Systems*, 19(3):299–316, 2006.

[22] A Markopoulou, F Tobagi, and M Karam. Assessment of VoIP quality over internet backbones. In *IEEE INFOCOM*, volume 1, pages 150 – 159, New York, NY, USA, June 2002.

[23] France Telecom R&D. Continuous assessment of time-varying subjective vocal quality and its relationship with overall subjective quality. ITU Study Group 12, Contribution COM 12-94-E, 1999.

[24] ITU. ITU-T recommendation G.107: The E-model, a computational model for use in transmission planning, 2005.

[25] ITU. ITU-T recommendation P.862: Perceptual evaluation of speech quality PESQ: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, 2001.

[26] Antony W. Rix. Perceptual speech quality assessment - a review. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, Montreal, Canada, May 2004.

[27] Scott Pennock. Accuracy of the perceptual evaluation of speech quality (PESQ) algorithm. In *MESAQIN*, Prague, Czech Republic, January 2002.

[28] Alan Clark. Modelling the effects of burst packet loss and recency on subjective voice quality. In *IP Telephony Workshop*, New York, March 2001.

[29] Lijing Ding and Rafik A. Gourban. Speech quality prediction in voip using the extended E-Model. In *IEEE GLOBECOM*, San Francisco, December 2003.

[30] Lingfen Sun and Emmanuel Ifeachor. New models for perceived voice quality prediction and their applications in playout buffer optimization for voip networks. In *IEEE ICC*, Paris, France, June 2004.

[31] Ilias Tsompanidis, Georgios Fortetsanakis, Toni Hirvonen, and Maria Papadopouli. Analyzing the impact of various wireless network conditions on the perceived quality of VoIP. In *IEEE LANMAN*, New Jersey, USA, May 2010.

[32] Ilias Tsompanidis, Georgios Fortetsanakis, Toni Hirvonen, and Maria Papadopouli. Analysis of the perceived voip quality under various wireless network conditions. In *Eighth International Conference on Wired/Wireless Internet Communications (WWIC10)*, Luleå, Sweden, June 2010.

[33] Ilias Tsompanidis, Georgios Fortetsanakis, and Maria Papadopouli. Measuring the perceived VoIP quality under various wireless network conditions. Technical Report 401, ICS-FORTH, Heraklion, Crete, Greece, January 2010.

[34] Hector Velayos and Gunnar Karlsson. Techniques to reduce IEEE802.11b MAC layer handover time. In *IEEE International conference on communications*, Paris, June 2004.

[35] Sangho Shin, Andrea Forte, Anshuman Singh Rawat, and Henning Schulzrinne. Reducing mac layer handoff latency in ieee 802.11 wireless lans. In *ACM MobiWac*, Philadelphia, PA, September 2004.

[36] Jangeun Jun, P Peddabachagari, and M Sichitiu. Theoretical maximum throughput of IEEE802.11 and its applications. In *IEEE NCA*, pages 249 – 256, Cambridge, MA, USA, April 2003.

[37] David L. Mills. Network time protocol (version 3) specification, implementation and analysis. Technical Report 90-6-1, University of Delaware, June 1990.

[38] ITU. ITU-T recommendation P.810: Modulated noise reference unit (mnru), 1996.

[39] ITU. ITU-T recommendation P.800: Methods for subjective determination of transmission quality, 1996.

[40] ITU. ITU-T recommendation P.830: Subjective performance assessment of telephone-band and wideband digital codecs, 1996.

[41] ITU. ITU-T recommendation P.833: Methodology for derivation of equipment impairment factors from subjective listening-only tests, 1996.

[42] SoX. SoX - Sound eXchange, 2010. URL `http://sox.sourceforge.net/`.

[43] ITU. ITU-T recommendation G.191: Software tools for speech and audio coding standardization, 2007.

[44] Shao-Cheng Wang and Ahmed Helmy. Performance limits and analysis of contention-based IEEE802.11 MAC. In *IEEE LCN*, Tampa, Florida, U.S.A., November 2006.

[45] Barbara G. Tabachnick and Linda S. Fidell. *Using Multivariate Statistics (5th Edition)*. Allyn & Bacon, 5 edition, March 2006.

[46] K. Mase and Y. Toyama. End-to-end measurement based admission control for voip networks. In *IEEE ICC*, New York, USA, April 2002.

[47] Durga Shankar Dash, Arjan Durresi B, and Raj Jain B. Routing of voip traffic in multi-layered satellite network. In *Performance and control of next-generation communications networks*, Orlando, USA, September 2003.

[48] ITU. ITU-T recommendation G.114: One-way transmission time, 2003.