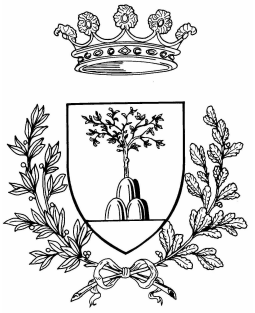
UNIVERSITÀ DEGLI STUDI DI FERRARA



FACOLTÀ DI INGEGNERIA
DOTTORATO DI RICERCA IN SCIENZE DELL'INGEGNERIA
Ciclo XXVI

COORDINATORE Prof. Stefano Trillo

Cross-layer Optimization for Video Delivery over Wireless Networks

Dottorando

Dott. Sergio Cicalò

Tutore

Prof. Velio Tralli

Anni 2010/2013

Acknowledgements

I would like to deeply thank my supervisor, Professor Velio Tralli, for his time, patience, support and especially to convey his passion for high valuable research. His dedication has been inspirational. I'm very grateful for all the opportunities, the ability to travel abroad, the encouragement and the trust he has given to me.

I'm also really glad to thank Dr. Nesrine Changuel and all the brilliant researchers of the Multimedia Team of the Alcatel-Lucent Bell Labs in France, for the opportunity to spend six months in the beautiful city of Paris and, at the same time, make me feel at home.

My sincere thanks to all my friends. First, I'm very thankful to Andrea Peano, Roberto Lapia and Roberto Mattera, who I spend most of the time in these last three years. But I cannot forget to thank the friends that, in different ways, have also helped me go through these years: Marcella, Roberto, Luca, Michele, Alessio, Nicola, Mario, Martina, Paola, Chloé and many others that I cannot mention due to the lack of space.

To my Dad, and to my brothers, Roberto and Mauro, thank you for your support.

Last but not least, I give a special thank to my *Mom*. All my achievements were, are and will be always dedicated to you.

To Mom.
It's impossible to thank you adequately for everything you've done.

Preface

As video streaming is becoming the most popular application of Internet mobile, the design and the optimization of video communications over wireless networks is attracting increasingly attention from both academia and industry. The main challenges are to enhance the quality of service support, and to dynamically adapt the transmitted video streams to the network condition. The cross-layer methods, *i.e.*, the exchange of information among different layers of the system, is one of the key concepts to be exploited to achieve this goals.

In this thesis we propose novel cross-layer optimization frameworks for scalable video coding (SVC) delivery and for HTTP Adaptive Streaming (HAS) over the downlink and the uplink of Long Term Evolution (LTE) wireless networks. They jointly address optimized content-aware rate adaptation and radio resource allocation (RRA) with the aim of maximizing the sum of the achievable rates while minimizing the quality difference among multiple videos.

In order to perform optimized content-aware rate adaptation, we first analyze the video quality metrics that allow to assess the quality of a video sequence and then we provide enhanced low-complexity models to accurately estimate the Rate-Distortion (R-D) relationship of scalable video transmitted over error-free and error-prone channels. For the latter scenario, we design an enhanced Unequal Erasure Protection (UXP) profiler with the objective to provide R-D relationship that keeps the expected distortion almost unchanged at different packet failure rate, with only a rate increase/decrease.

For multi-user SVC delivery over downlink wireless systems, where Orthogonal Frequency Division Multiple Access (OFDMA) is the key Physical (PHY) layer technology and IP/TV is the most representative application, we decompose the optimization problem and we propose the novel iterative local approximation algorithm to derive the optimal solution, by also presenting optimal algorithms to solve the resulting two sub-problems.

For multiple SVC delivery over uplink wireless systems, where Single-Carrier Frequency Division Multiple Access (SC-FDMA) is the key PHY technology and health-related services are one of the most attractive applications, we propose joint

video adaptation and aggregation directly performed at the application layer of the transmitting equipment, which exploits the guaranteed bit-rate (GBR) provided by the low-complexity sub-optimal RRA solutions proposed.

Finally, we propose a quality-fair adaptive streaming solution to deliver fair video quality to HAS clients in a LTE cell by adaptively selecting the prescribed Guaranteed Bit-Rate (GBR) of each user according to the video content in addition to the channel condition.

Extensive numerical evaluations show the significant enhancements of the proposed strategies with respect to other state-of-the-art frameworks. Even though broadband mobile provider are reluctant to include application-aware module in the design of cellular systems, due to management and coordination issues, our research show that significant gains in terms of the Quality of Experience (QoE) of the end-user can be achieved by the proposed content-aware cross-layer strategies.

Most of the contributions presented in this thesis appear in the Author's publications listed at the end of the manuscript.

Contents

Acronyms

3GPP Third Generation Partnership Project

AMC Adaptive Modulation and Coding

ANSI American National Standards Institute

APP Application

AVC Advanced Video Coding

BS Base Station

BL Base Layer

CGS Coarse Grain Scalability

CoMP Coordinated Multi-Point

CSI Channel State Information

CQI Class Quality Information

DASH Dynamic Adaptive Streaming over HTTP

DCT Discrete Cosine Transform

DFT Discrete Fourier Transform

DON Decoding Order Number

EL Enhancement Layer

eNB Enhanced Node-B

EQ Equal Quality

ER Equal Rate

EPP Error Probability Profile

FEP Frame Error Probability

FEC Forward Error Correction

FGS Fine Grain Scalability

FR Full-Reference

GBR Guaranteed Bit-Rate

GOP Group of Pictures

GBS Gradient-Based Scheduling

HAS HTTP Adaptive Streaming

HTTP Hyper Text Transfer Protocol

HVS Human Visual System

ICI Inter-Cell Interference

ICIC Inter-Cell Interference Coordination

IDR Intra-Decoding Refresh

ILA Iterative Local Approximation

IP Internet Protocol

ISI Inter-Symbol Interference

JT Joint-Transmission

LB Load Balance

LTE Long Term Evolution

MAC Medium Access Control

MANE Media Aware Network Element

MBR Maximum Bit-Rate

MGS Medium Grain Scalability

MOS Mean Opinion Score

MP Multimedia Provider

MPD Multimedia Presentation Descriptor

MPEG Moving Picture Experts Group

MSE Mean Square Error

NALU Network Abstraction Layer Unit

NAT Network Address Translation

NR No-Reference

OFDM Orthogonal Frequency Division Multiplexing

OFDMA Orthogonal Frequency Division Multiple Access

PHY Physical

PR Partial-Reference

PAPR Peak-to-Average Power Ratio

PF Proportional Fair

PSNR Peak-to-Signal Noise Ratio

PRB Physical Resource Block

QoS Quality of Service

QoE Quality of Experience

QFAS Quality-Fair Adaptive Streaming

RRA Radio Resource Allocation

RMSE Root Mean Square Error

R-D Rate-Distortion

R-Q Rate-to-Quality

RS Reed Solomon

RTP Real Time Protocol

SC-FDMA Single-Carrier Frequency Division Multiple Access

SSIM Structural SIMilarity

SNIR Signal-to-Noise plus Interference Ratio

SNR Signal-to-Noise Ratio

SVC Scalable Video Coding

TB Transmission Block

TCP Transmission Control Protocol

QP Quantization Parameter

UDP User Datagram Protocol

UE User Equipment

UEP Unequal Error Protection

UXP Unequal Erasure Protection

VCL Video Coding Layer

VQM Video Quality Model

WAN Wireless Access Network

WSAR Weighted Sum of the Average Rates

WSR Weighted Sum of Rates

Chapter 1

Introduction

Today, we are facing an explosion of the video traffic on wireless network due to the proliferation of multimedia-friendly portable devices [1]. In addition, the emergence of high speed networks provides the infrastructure and the possibility for handling a wide set of new applications among which the multimedia contents delivery. Multimedia or more specifically video delivery systems address the problem of streaming multimedia data as a continuous stream. The end-user can start displaying the video data or multimedia data before the entire file has been transmitted.

A high degree of flexibility and adaptivity is required from the video delivery system to meet different levels of quality requirements depending on the different characteristics of end-user devices and access networks. The design and the optimization of video communications over wireless networks is thus attracting a lot of attention from both academia and industry. The main challenges are to enhance the quality of service (QoS) support in terms of packet loss rate, end-to-end delay and minimum guaranteed bit-rate, while providing fairness where needed, and to dynamically adapt the transmitted video streams to the network conditions. One of the key concept to achieve these goals is the cross-layer approach, which allows the exchange of information among different layers of the system.

Traditionally, Real Time Protocol (RTP) is used for video streaming services, since it provides end-to-end delivery for data with real-time characteristics, timing reconstruction, loss detection, security and content identification. RTP also allows for the implementation of source rate adaptation to the different network condition. On-line adaptation of the video sources is enabled by the use of video encoders that support multiple layers which can be sequentially dropped, thereby providing a graceful degradation. One of the most promising tool is the H.264 Advanced Video Coding (AVC) standard with scalable extension, also known as

Scalable Video Coding (SVC) [2]. The main drawback of RTP is that it requires dedicated servers and passes through a port that is often block by firewall and Network Address Translation (NAT). For this reasons most of the video traffic is now transmitted over HTTP protocol, which is NAT transparent, and may exploit the large deployments of cache and content distributed networks (CDN). A new approach referred to as HTTP adaptive streaming (HAS) [3] is becoming popular. HAS is adaptive in the sense that it allows a client to adaptively switch between multiple bit-rates pre-encoded in the server, according to the bandwidth or data rate available between the server and the client. This is a particularly useful feature for a wireless environment since the data rate in mobile systems can vary over time.

The Long Term Evolution (LTE) represents the next generation broadband mobile technology [4]. In comparison to the previous cellular standards, LTE provides improved system capacity and coverage and lower delivery latency. Differently to its predecessors, LTE has selected for the first time OFDMA as a key physical (PHY) layer technology [5]. In the downlink of a multi-user system, OFDMA allows to to allocate a disjoint number of so-called Physical Resource Block (PRB) in the time-frequency grid, in which users experience favorable channel conditions. The better the channel condition are, the higher the rate used in the resource elements. This results in a very flexible access with high spectral efficiency.

The main drawback of OFDMA schemes is that the resulting time-domain waveform exhibits very pronounced envelope fluctuations resulting in a high peak-to-average power ratio (PAPR) which requires highly linear power amplifiers to avoid excessive inter-modulation distortion. This problem is more critical in the up-link transmission where the cost and power consumption of mobile must be kept as lower as possible. To these ends, Single Carrier - Frequency Division Multiple Access (SC-FDMA) [6], has been introduced for the LTE uplink. SC-FDMA provides similar advantages of the OFDMA systems but provide a lower PAPR by introducing a Discrete Fourier Transform (DFT) pre-coding process at the transmitter, which spreads the data power over the entire allocated bandwidth.

1.1 Motivation

The most straightforward approach to deliver video streams to multiple users in bandwidth-limited systems is to divide the available bandwidth equally among all video streams. However, the rate of an encoded video is variable, as the result of the variable temporal and spatial structure of the video frames. Also the

relationship between rate and quality changes within a single video and among different videos [7–10]. For this reason, RTP-based cross-layer video streaming optimization of multiple users in the downlink of wireless systems has been usually addressed in the literature, *e.g.*, in [11][12][13], by formulating a problem where the objective is to adaptively minimize the sum of the average video distortions or, similarly, to maximize the sum of the average objective qualities, *e.g.*, Peak-to-Signal Noise Ratio (PSNR)s, under a particular set of constraints. Such objective usually leads to the provision of the highest quality, *i.e.*, the lowest distortion, to the low-complexity videos, while providing low quality to the more demanding high-complexity videos [14].

The end-user expectation of video streaming is to receive the best feasible quality independently of the particular video complexity. Therefore, quality fairness is an important issue that must be addressed in these applications, and the video models that allow to predict the minimum rate required to achieve a target quality are essential part of the optimization. Moreover, the presence of an optimized source rate adaptation technique at the Application (APP) layer becomes crucial to improve stability, to prevent buffer overflow and to maintain video play-back continuity.

Beside the distortion due to lossy encoding process, the quality of each video can be heavily reduced due to the transmission errors and the consequent loss of part of the video stream. The automatic repeat-request (ARQ) schemes have the main drawback to increase the delay and can not be suitable for many application where the playback time is a stringent constraint. Within the framework of RTP-based SVC video delivery schemes, Forward Error Correction (FEC) has been proposed to recover channel errors and many contributions in the literature have proved its effectiveness [15–17].

The solution for the aforementioned issues in RTP-based SVC video delivery systems requires a Media Aware Network Element (MANE) that is able to extract from the original video sequences a set of scaled streams with a fair assignment of expected end-user quality according to the estimated bandwidth and minimum and maximum rate constraints, even in presence of packet losses.

In the uplink of wireless systems, the transmission of health-related information from an ambulance to a remote hospital is a challenging task, due to the variability and the limitations of the mobile radio link. In particular, the transmission of multiple video streams can improve the efficacy of the tele-consultation service, but requires a large bandwidth to meet the desired quality, not always guaranteed by the mobile network. Moreover, a strict separation into multiple single flows may turn out to be inefficient, especially in case of simultaneous transmission from multiple and heterogeneous co-located sources. We consider two categories

of videos transmitted from the ambulance: (i) ambient videos that allow the hospital staff to visually follow the patient conditions and the activities performed in the ambulance; (ii) diagnostic videos obtained as result of emergency examinations, such as the Focused Assessment with Sonography for Trauma (F.A.S.T.), which is used to rapidly assess the status of heart and abdominal organs of the patient [18]. Due to the different importance of the video flows, a video adaptation module has to manage the inherently different priorities of the video flows generated by the ambulance.

Since HAS-based video delivery is based on a user-centric optimization approach, it suffers from three major problems, *i.e.*, efficiency, stability and fairness. Efficiency and stability issues arise when the clients do not fully exploit the available resources, and perform needless bit-rate switches. The fairness issue mainly arise when users fail to fairly estimate the bandwidth due to periodic request of video chunk, which results in ON-OFF period. In fact, when no limitation on the allocated resources is taken into account, competing players with non-overlapping ON-OFF period may not estimate their fair share of bandwidth correctly. Therefore, also HAS-based systems called for enhanced media-aware optimization strategies, aimed at deriving the minimum and the maximum bit-rate of each user that allows players to fairly estimate the bandwidth and to request quality-fair video streams.

In wireless systems, the throughput experienced by each user depends on how the system exploits the available time and frequency resources. Modern wireless transmission systems make use of suitable Adaptive Modulation and Coding (AMC) scheme to improve the rate of transmission, and/or bit error rates, by exploiting the Channel State Information (CSI) that is present at the transmitter. Especially over fading channels where channel gains vary on time and frequency domains, AMC systems exhibit great performance enhancements compared to systems that do not exploit channel knowledge at the transmitter. In particular, Orthogonal Frequency Division Multiplexing (OFDM)-based systems exploiting AMC schemes have an inherent temporal, frequency and multi-user diversity, which requires suitable adaptive resource allocation and scheduling strategies. Opportunistic schedulers, as for instance, proportional fair (PF) [19] and maximum signal-to-noise ratio (SNR) schedulers, take advantage of the knowledge of the channel state information (CSI) in order to maximize the spectral efficiency. However, with these schedulers, the final share of throughput often results unfair, especially for the cell-edge users which suffer of data-rate limitations due to high path-loss and Inter-Cell Interference (ICI). In real-time streaming the mismatch between the allocated PHY layer rate and the rate required by the delay-constrained application may cause the loss of important parts of the

streams, which significantly degrades the end-user quality of experience (QoE). The provision of acceptable QoE to every user is enabled by the use of a scheduler at the medium access control (MAC) layer which delivers a fair throughput, according to specific utilities and constraints defined by the APP [20].

To summarize, we have to face the following challenges:

- the derivation of accurate and low-complexity models that estimate the minimum bit rate of scalable video stream required to achieve a target quality
- the design of optimized UXP profiler in case of transmission over packet erasure channel, and the derivation of the resulting expected R-D relationship at the end-user
- the study of video quality-fair metrics and the investigation of rate adaptation techniques at the APP layer that allows to extract quality-fair streams, which must also satisfy minimum and maximum rate constraints to ensure the continuity of the video reproduction and to save bandwidth, respectively
- the study and the solutions of the Radio Resource Allocation (RRA) problems of OFDMA and SC-FDMA systems to maximize the sum of the achievable throughput under QoS constraints defined by the applications
- The investigation of enhanced cross-layer strategies, which allows the exchange of information to jointly optimize the APP and the MAC layers

1.2 Objectives

The general objective of this thesis is to develop an optimized analytical cross-layer framework for the delivery of video streams with scalable features to multiple users competing for the same resources. The framework addresses the issues of source rate adaptation, RRA, error protection and the objective is to provide a fair video quality among the video programs.

Therefore, the first aim is to analyze the video quality metrics that allow to assess the quality of a video sequence and to provide enhanced low-complexity models to accurately estimate the R-D relationship of scalable video transmitted over error-free and error-prone channels. For the latter scenario, enhanced UXP scheme have to be investigated. The objective considered here is to provide expected R-D relationship which keeps the expected distortion almost unchanged with only a rate increase/decrease at different packet failure rate. This allows

to model the R-D relationship in error-prone channel with similar function with respect to the case of error-free channels.

According to the scenario, different application have to be investigated. As first and simplest scenario, we consider the multi-user cross-layer SVC delivery problem assuming limited but constant bandwidth and error-prone channels. Then, the objective is to extend such framework to multi-user cross-layer video delivery over single-cell and multi-cell wireless scenario in downlink system where the bandwidth and the user capacity vary on both frequency and time domains. In this frameworks, applications like video on-demand[21], IP-TV[22], sport broadcasting, where an initial transmission delay in the order of seconds can be tolerated by the end-users, as well as real-time streaming [23], are considered.

We then aim at proposing a novel solution for the transmission of multiple videos from an emergency scenario, based on the joint video adaptation and aggregation directly performed at the APP layer of the transmitting equipment. The objective is to deliver the ultrasonography information with sufficiently high quality and the set of ambient videos tuned according to quality fairness criteria. To provide a certain level of QoS, we also investigate enhanced RRA strategies at the MAC layer of SC-FDMA systems.

The last objective considered here is to extend the proposed approach to specific LTE systems and HAS applications.

1.3 Contribution

Here, we briefly summarize the contributions of thesis. A detailed overview of each contribution is provided at the end of the introduction of each chapter.

The main achievement of this thesis is the proposal of novel cross-layer methods for maximizing the aggregate ergodic (average) rate assigned to multiple SVC transmission in the downlink of OFDMA and in the uplink of SC-FDMA systems, while minimizing the distortion or quality difference among the received video sequences.

We first propose continuous low-complexity models to accurately estimate the R-D relationship of SVC and HAS video streams for real-time and near-real-time video transmission, by also designing an optimized UXP strategy.

We then propose method to optimally delivery SVC video streams in the downlink of OFDMA wireless systems. In this case, the optimization problem is "vertically" decomposed into two sub-problems, leading to the rate adaptation at the APP layer and the resource allocation at the MAC layer, and a novel efficient and optimal iterative local approximation (ILA) algorithm is proposed to obtain the

global solution. The ILA algorithm is based on the local approximation of the contour of the ergodic rate region of the OFDMA downlink channel and requires a limited information exchange between the APP and the MAC layers. Moreover, we present and discuss optimal algorithms to solve the two sub-problems, *i.e.*, rate adaptation at the APP layer and RRA at the MAC layer, and finally prove the optimality and convergence of the ILA algorithm. The proposed rate adaptation algorithm can be seen as extension of the special case of the cross-layer optimization SVC delivery problem in shared channel with constant bandwidth and quality-fair constraints.

We also extend the MAC layer algorithm proposed for a single-cell scenario to multi-cell environment. We propose and compare centralized and distributed RRA algorithm aimed at maximizing the sum-rate of a multi-cell clustered system under proportional rate constraints. While the centralized approach allows to optimally solve the Inter-Cell Interference Coordination (ICIC) problem, distributed strategies requires off-line coordinated resource control among the cells in a cluster. In the latter case we propose power planning schemes with preassigned powers. We show that distributed schemes with aggressive reuse manage to approach the capacity of a centralized system when the number of users is large.

For the uplink SC-FDMA wireless network, we propose a novel solution for the transmission of multiple health-related SVC videos, based on the joint video adaptation and aggregation directly performed at the APP layer of the transmitting equipment. In this approach, only a single communication link characterized by given QoS guarantees needs to be managed between the terminal and the receiver, while additional spectrum efficiency is gained from video multiplexing. In our solution the adaptation is designed to optimize quality and fairness by exploiting the information on the available rate assigned by the LTE e-nodeB. The available rate is derived according to the solution of the ergodic sum-rate maximization problem under proportional rate constraints in SC-FDMA systems. For this problem we propose novel sub-optimal algorithmic solution, whose complexity increases only linearly with the number of users and the number of resources and the performance gap to optimal solution is limited to the 10% of the sum-rate.

We finally propose a quality-fair adaptive streaming (QFAS) solution to deliver fair video quality to HAS clients competing for the same resources in an LTE cell. The proposed QFAS solution brings intelligence into the network to adaptively select the prescribed guaranteed bit-rate and maximum bit-rate of each UE according to the contents characteristics in addition to the channel condition.

Extensive numerical evaluations show for each proposed cross-layer solution the significant video quality gain achieved with respect to other state-of-the-art

solutions.

1.4 Overview of the Thesis

This thesis comprises seven chapters and one appendix. The following chapter aims to provide a brief overview of the scalable video approach, in particular of the SVC standard, and of the HAS media preparation. Chapter 3 provides a detailed analysis of video quality assessment metric and of the R-D relationship also in case of error-prone channel of SVC and HAS. In Chapter 4 we first analyze and propose solutions multi-user cross-layer video delivery optimization problem assuming constant bandwidth and error-prone channels. This is the first contribution of the thesis, which allows to understand the benefits of quality-fair adaptive rate-adaptation of multiple SVC videos in a simple scenario. Chapter 5 represents the main contribution of this thesis and aims at extending such approach to the case of multi-user downlink wireless scenario, *i.e.*, a single-cell OFDMA systems (extended to multi-cell scenario in Appendix A), where bandwidth is not constant and depends on how resources are shared among users. The proposed solutions provides a complete novel framework to optimally and jointly perform rate-adaptation at the APP layer and resource allocation at the MAC layer. Chapter 6 focus on the the uplink wireless transmission systems, *i.e.*, SC-FDMA systems, where health-related services are one of the most attractive applications. It aims at proposing a novel solution for the transmission of multiple videos from an emergency scenario, based on the joint video adaptation and aggregation directly performed at the application layer of the transmitting equipment. Also enhanced RRA strategies at the MAC layer are proposed. In chapter 7 we target HAS applications in LTE networks, by considering all the constraints at which such applications must adhere. In All the aforementioned chapters we first introduce the motivation, the objectives, a detailed literature review, as well as the contribution. I finally draw the conclusion of our work in chapter 8.

1.4.1 Notation

Vectors and sets are denoted by bold and calligraphic fonts, respectively. $\boldsymbol{x}^{\mathrm{T}}$ and $\|\boldsymbol{x}\|_{\mathrm{p}}$ indicate transpose and p-norm, respectively, of the vector $\boldsymbol{x}$. Given the vectors $\boldsymbol{x} = [x_1, \ldots, x_N]$, $\boldsymbol{x}' = [x'_1, \ldots, x'_N]$ of N components, we use the following element-wise inequalities:

$$\boldsymbol{x} \succeq \boldsymbol{x}' \Leftrightarrow x_n \geq x'_n, \forall n = 1, \ldots, N$$

$$\boldsymbol{x} \succcurlyeq \boldsymbol{x}' \Leftrightarrow \boldsymbol{x} \succeq \boldsymbol{x}' \wedge \exists m : x_m > x'_m$$

$\mathbb{E}_{\boldsymbol{y}}[\cdot]$ denotes the expectation taken with respect to the random process $\boldsymbol{y}$. We also use notations $[x]^+ = \max(x, 0)$ and $[x]^+_\epsilon = \max(x, \epsilon)$, with ϵ arbitrary close to zero. The operators $\wedge$ indicates "AND".

The most used symbols of this thesis are summarized in Table 1.1 for Readers convenience.

Notation	Description
$\mathcal{K}$, K	Set and total number of users
$\mathcal{S}$, S	Set and total number of subcarriers
$\mathcal{G}$, G	Set and total number of PRBs
$\mathcal{J}$, J	Set and total number of feasible patterns
$\mathcal{R}$	PHY layer rate region
$\mathcal{T}_{\mathcal{R}}$	Tangent space to $\mathcal{R}$
$\mathcal{E}$	Boundary of $\mathcal{R}$
$\mathcal{F}$	Set of rate vectors as in (5.10)
$\mathcal{A}$	Set of feasible allocation policies
$\boldsymbol{\gamma}$	SNR realizations
$\boldsymbol{\psi}, \boldsymbol{p}$	Set of PHY layer allocation variables
$\boldsymbol{R}$	Average PHY layer rate vector
$\boldsymbol{r}$	Instantaneous PHY layer rate vector
$\boldsymbol{\phi}$	Rate direction vector
$\boldsymbol{\mu}$	Weight vector
$\boldsymbol{F}$	Source rate vector
$\boldsymbol{F}^{\max}$	Maximum source rate vector
$\boldsymbol{F}^{\min}$	Minimum source rate vector
γ^{eff}	Effective SNR
d	Discrete distortion
D	Continuous distortion
$D_k^{\min}$	Minimum distortion of video k
$D_k^{\max}$	Maximum distortion of video k
U_k	Utility of user k
α_k, β_k, ξ_k	Parameter of model (3.12)
$P_{\text{e}}^{\text{RTP}}$	RTP error rate
H	Overhead factor
$\Delta(x, y)$	Distortion difference function

Table 1.1: List of of most used symbols

Chapter 2

Scalable Video Encoders

Video streaming is one of the most popular applications of today's Internet. As the Internet is a best effort network, it poses several challenges especially for high quality video streams.

The Advanced Video Coding (H.264/AVC) scalable extension, also called Scalable Video Coding (SVC), provides an attractive solution for the difficulties encountered when a video source is transmitted over RTP/Internet Protocol (IP)-based wireless transmission systems. Such challenges include error prone channels, heterogeneous networks and capacity limitations and fluctuations [2]. SVC allows for QoS adaptation in RTP transmission to variable network conditions or needs or preferences of end-user, as well as video content delivery to a variety of decoding terminals with heterogeneous display resolutions and computational capabilities, by means of a set of scalability features.

While SVC can exploit RTP connection-oriented video transport protocols, which maintain per-session state and use a (proprietary) stateful control protocol to manage the data delivery, more of the video traffic is nowadays transmitted over HTTP. Due to its stateless design, in HTTP-based streaming the video content is segmented in different chunk, and the a client fetches each chunk independently while maintaining the playback session state.

Several proprietary HAS technology has been implemented, *i.e.*, Microsoft Smooth Streaming [24] Apple HTTP Live Streaming [25] and Adobe HTTP adaptive Streaming [26]. However none of them providing a unified standard. The Moving Picture Experts Group (MPEG) has recently finalized a new standard to enable dynamic and adaptive streaming of media over HTTP [27], also known as MPEG-Dynamic Adaptive Streaming over HTTP (DASH). The objective ot the standard is to address the interoperability needs between devices and servers of various providers.

In this chapter we briefly review the SVC standard and the HAS state-of-the-art.

2.1 Scalable Video Coding

SVC is the extension of the H.264/MPEG-4 AVC video compression standard described in the Annex G [2]. SVC standardizes the encoding of a high-quality video bit-stream that also contains one or more subset bit-streams. Within SVC, each sequence is encoded with one base layer (BL) and several enhancement layers (ELs) which can be sequentially dropped by providing a graceful degradation. Each layer is then coded and encapsulated into several Network Abstraction Layer Units (NALUs), which are packets with an integer number of bytes.

Three types of scalabilities, namely spatial, temporal and SNR scalability are supported by the standard, which allows to extract from the encoded video sub-streams of a suitable resolution, frame rate and quality matching various network conditions and terminal capabilities. They corresponds to three key ID values, *i.e.*, *dependency_id*, *temporal_id*, and *quality_id*, which are embedded in the header by means of the high level syntax elements, in order to identify spatial, temporal and quality layers. An optional *priority_id* can be inserted to prioritize each frame in stream [28].

In the next subsection we provide a brief overview of the temporal, spatial and SNR scalability. We refer the interested reader to [2] for a more general overview.

2.1.1 Temporal Scalability

Temporal scalability can be achieved by means of the concept of hierarchical prediction. The pictures of the video sequence are organized in sets of G frames, also called groups of pictures (GOPs). Each picture in one GOP is then identified by a hierarchical temporal index or level $\tau \in \{0, 1, \ldots, T\}$.

The encoding/decoding process starts from the first frame of each GOP with the temporal index $\tau = 0$, which can be intra-coded (I-frame) or inter-coded (P-frame), according to a trade-off between error-resilience and R-D efficiency. The interval (in frames) between two consequent I-frames, also called Intra-Decoding Refresh (IDR) period, is here assumed as multiple of the GOP size G. The remaining frames of the GOP are assumed to be encoded as B-frames using hierarchical prediction, *i.e.*, the encoding of a frame with temporal index τ exploits prediction from frames with temporal index smaller than τ. The remaining frames of one GOP are typically coded as P/B-pictures and predicted according to the hierar-

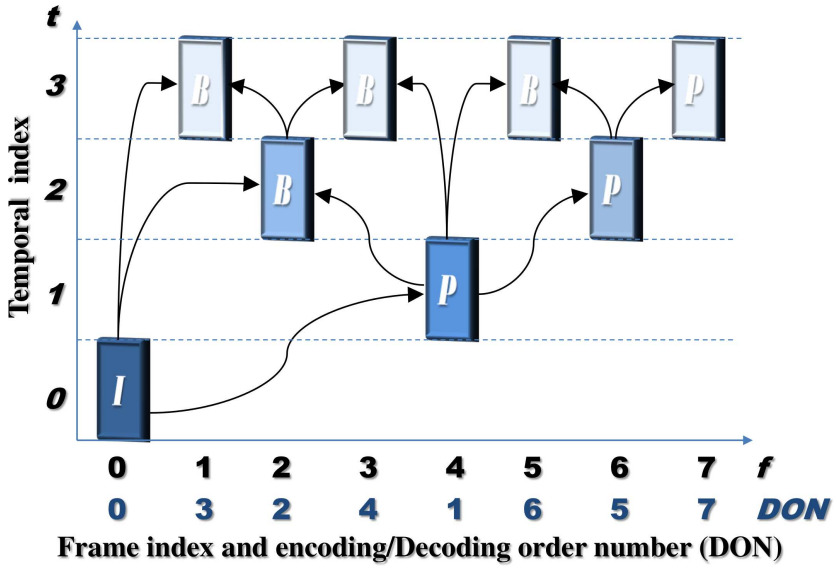


Figure 2.1: Enhancement temporal layer prediction for a GOP of 8 frames.

chical temporal index, thereby allowing to extract a particular frame rate. An implicit encoding/Decoding Order Number (DON) can be set up according to the temporal index and frame number of each frame.

In Figure 2.1 we show an example of the hierarchical prediction structure for a GOP with 8 pictures. The DON is obtained by ordering the pictures according to the temporal index. If more than one frame have the same temporal level, the DON is assigned according to the picture index. In this example the last frame is encoded as P-frame in order to allow a GOP-based decoding.

Temporal scalability is an interesting feature that can be also exploited at the decoder side in case of packet loss. If a picture with temporal index (*temporal_id*) $\tau > 0$ is lost, the decoder is still able to decode and playback the GOP at the τ-th temporal resolution, *e.g.*, by simply replacing the missing picture with the previous one according to a picture copy error-concealment method. Since our work aims to provide quality fairness to the set of served end-users, we assume that the adaptation module extracts the same temporal resolution from each video. Therefore, the temporal scalability is only exploited at the decoder side, when a B/P frame is lost.

2.1.2 Spatial Scalability

Spatial scalability is performed according to a layered coding approach which is used to encode different picture sizes of an input video source. Each layer refers to a target spatial resolution and corresponds to a spatial layer or dependency layer. The lowest spatial resolution, *i.e.*, the spatial base layer, is compatible with H.264/AVC baseline profile and its layer identifier is the lowest one. According to the output frame rate intended for for each spatial layer, it may contains several temporal layers. In particular, the standard specifies a maximum of eight supported dependency layers. To limit the memory requirements and decoding complexity derived from this multi-layer coding approach, the same coding order for all supported spatial layers is used. Specifically, the coding order of each spatial layer is based on an access unit (AU), where an AU is defined as the union of all the representations with different spatial resolutions for a given time instant. In In each spatial layer, the traditional motion-compensated and intra-prediction modes are supported as for non-scalable video coding. Since spatial scalability is not consider in our work, we refer the interested reader to [29] for further details.

2.1.3 SNR Scalability

The SNR scalability allows to increase the quality of the video stream by introducing refinement layers. Two different possibilities are now available in SVC standard and implemented in the reference software [30], namely Coarse Grain Scalability (CGS) and Medium Grain Scalability (MGS). CGS can be achieved by coding quality refinements of a layer using a spatial ratio equal to 1 and inter-layer prediction. However, CGS scalability can only provide a small discrete set of extractable points equal to the number of coded layers. In this thesis we focus on MGS scalability which provides finer granularity with respect to CGS coding by dividing a quality enhancement layer into up to 16 MGS layers.

MGS coding distributes the transform coefficients obtained from a macro-block by dividing them into multiple sets. The R-D relationship and its granularity depends on the number of MGS layers and the coefficient distribution. In [31] the authors analyzed the impact on performance of different numbers of MGS layers with different configurations used to distribute the transform coefficients. We also verified their results, by noting that more than five MGS layers reduce the R-D performance without giving a substantial increase in granularity. This is mainly due to the fragmentation overhead that increases with the number of MGS layers.

While extracting an MGS stream two possibilities are available in the reference software: a flat-quality extraction scheme, and a priority-based extraction

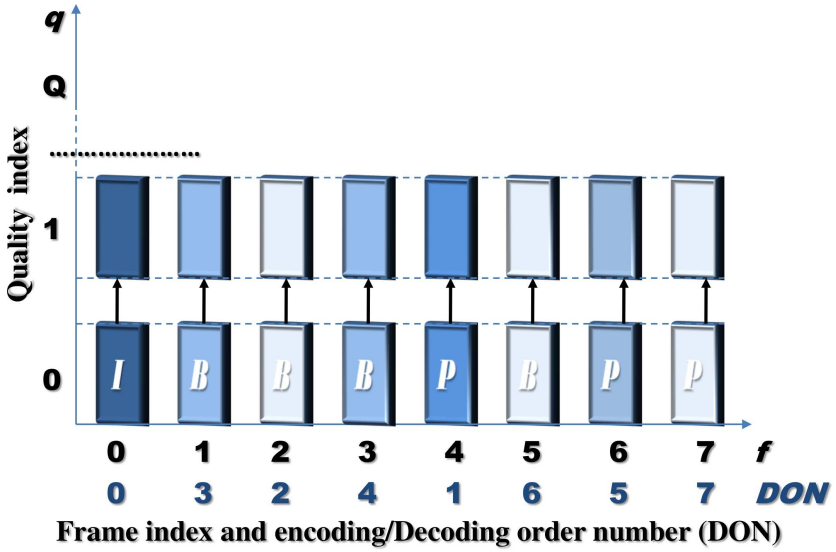


Figure 2.2: Enhancement quality layer prediction for a GOP of 8 frames. The encoding loop is closed at the base layer

scheme. The second scheme requires a post-encoding process, executed by an entity denoted as Priority Level Assigner, that computes a priority level for each NALU. It achieves higher granularity, as well as better R-D-performance [28]. The priority level ranges from 0 to 63, where 63 is intended for the base-layer, and is assigned to each NALU according to quality dependencies and R-D improvement. Nevertheless, in order to exploit the temporal scalability at the decoder side, we re-assign different priority levels to the base-layer frames (those with $q = 0$), according to their temporal indexes, as specified afterwards. This feature is only exploited by the UXP profiler and therefore does not change the 6-bit header of the packet which is necessary to perform the quality-based extraction. The coding efficiency of MGS scalable streams highly depends on the quality layers used for motion compensation. In the basic scheme the quality encoding loop is closed at the base layer as exemplified in Figure 2.2, thus avoiding the drift issue occurring when motion prediction is not synchronized between encoding and decoding process when quality layer are dropped or lost. However, this approach significantly decreases the coding efficiency of enhancement layers.

The R-D performance of the quality layers can be improved by using quality frames for motion compensation and introducing the concept of key-picture, which allows for a trade-off between drifting and coding efficiency as shown in Fig. 2.3. Nevertheless, this tool should be carefully applied in if most or all quality layers

are often discarded by the rate adaptation module.

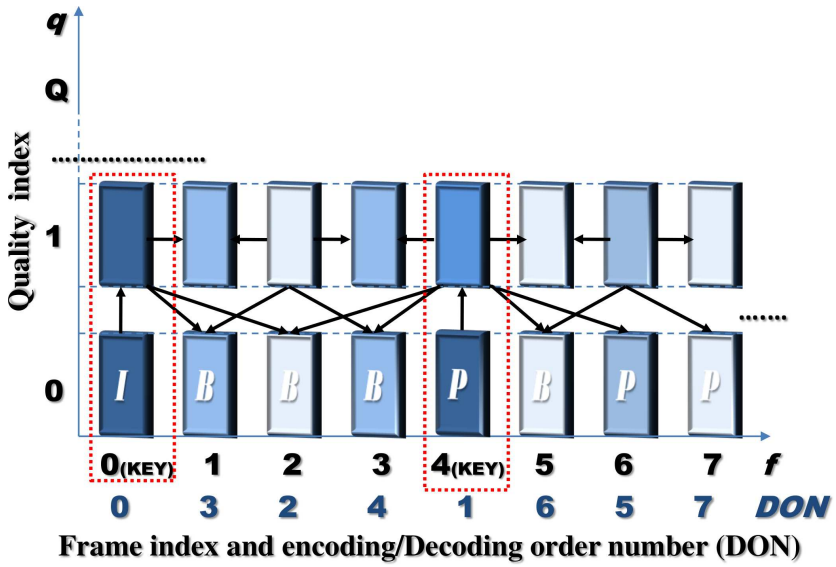


Figure 2.3: Enhancement quality layer prediction for a GOP of 8 frames using key-pictures

In this thesis we focus on MGS with optimized bit-stream extraction (see [32] and [28] for further details).

2.2 HTTP Adaptive Streaming

HTTP adaptive streaming aims to overcome all the issues of RTP streaming as firewalls and NAT traversals, and the requirement of dedicated network infrastructure that cannot be used for other web content.

In HAS approach the video content is encoded at multiple bit-rate, also called profiles, which may consist in different temporal, spatial and SNR resolutions. Even though, HAS can exploit the higher encoding efficiency of H.264/AVC single layer coding compared to SVC, the profiles can be encoded using SVC with benefits resulting in web caching efficiency and saved uplink bandwidth [33].

Each profile is then segmented in several chunks (with duration of 2 to 10 seconds). At the end of the encoding of the profiles or periodically during encoding, the server generates a manifest file, also called Multimedia Presentation Descriptor (MPD) in DASH, in order to provide location and timing information to the client requesting a particular video. An example of HAS approach is provided in Fig. 2.2.

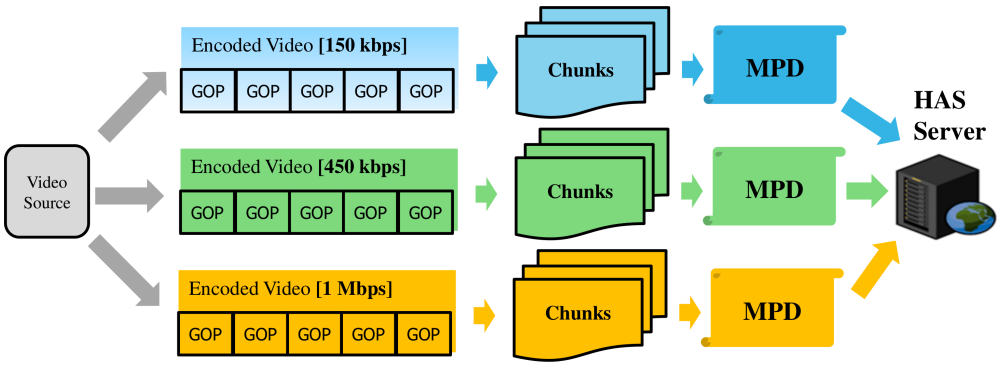


Figure 2.4: HAS media preparation.

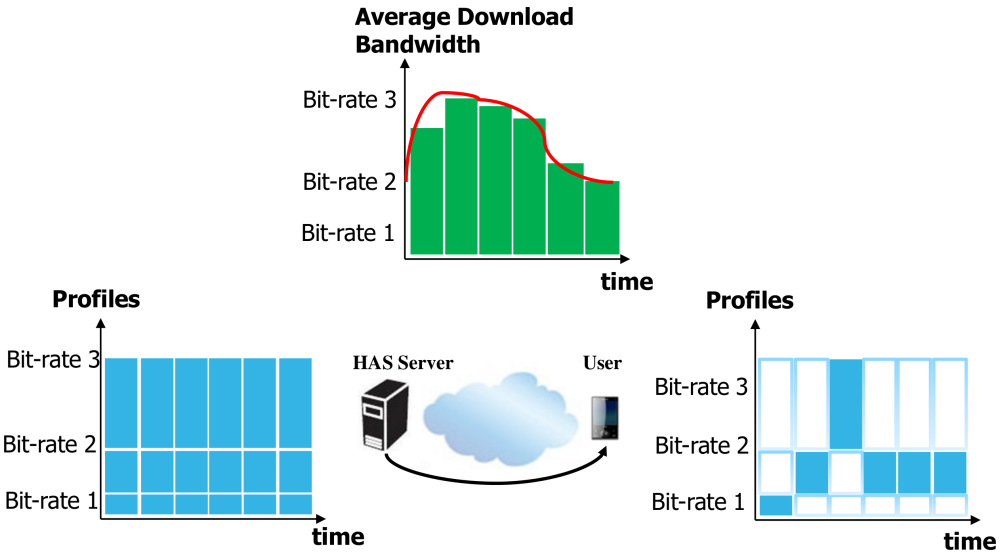


Figure 2.5: HAS-based video stream adaptation.

Generally the MPD file is downloaded using HTTP at the start of the streaming session, but for flexibility, the MPD may also be updated periodically, especially in the case of real-time streaming. After appropriate buffering to allow for network throughput variations, the client continues downloading the subsequent chunks and also monitors the network bandwidth fluctuations. Depending on its measurements, the client decides how to adapt to the available bandwidth by fetching segments of different alternatives (with lower or higher bit-rates) to also maintain an adequate buffer. An example of HAS-based video delivery is depicted in fig. 2.2, for the case of 3 profiles.

Chapter 3

Video Sources: Rate-Distortion Analysis and Models

The rapid growth of video applications into the wireless networks has called for highly media-aware encoder control and enhanced streaming strategies to manage the difficulties of time-varying bandwidth-limited wireless transmission. Capacity restrictions, heterogeneous devices, network capabilities and error prone transmissions are just some of the problems resulting from the characteristics of modern video communication systems, to which scalable video coding (SVC) offers an attractive solution.

Generally, an exhaustive understanding of the quality characteristics of encoded video is the basis for traffic modeling and the development of video transport mechanisms. The most straightforward solution to this problem is to allocate the available bandwidth equally among all video programs. However, due to the different scene content of the programs and the changes of the scene content over time, this approach results in suboptimal R-D performance and perceptual quality differences between the individual sequences.

Models to predict the quality of the encoded video sequence resulting from a certain encoding rate become then a key tool for the video delivery optimization. The computation of the perceptual quality requires in general subjective metrics, which are able to reliably measure the video quality that is perceived by the Human Visual System (HVS). The subjective video quality methods are based on groups of trained/untrained users viewing the video content. The resulting rating are then generally mapped in the so-called Mean Opinion Score (MOS), which is a value increasing with the perceived quality and ranging from 1 to 5.

For real time or near real-time video streaming systems the computation of the relationship between the rate and the quality of the encoded scenes should

be fast enough to deal with the timing constraints of the video stream and of the application. Obviously, subjective quality metric are not suited for these scenarios, but still are crucial for evaluating the performance of objective visual quality metrics. An overview of the the latter metrics used throughout this thesis is provided in section 3.2

R-D or more generally Rate-to-Quality (R-Q) models allow to predict the minimum bit rate required to achieve a target objective distortion or quality, respectively. They can be categorized as full-reference (FR), reduced-reference (RR), and no-reference (NR), depending on whether a reference (FR), partial information about a reference (RR), or no reference (NR) is used in the evaluation of the quality.

NR models are analytical R-D models which predict the rate and distortion of a video sequence prior to the encoding process. They are generally dependent on the probability distribution of Discrete Cosine Transform (DCT) coefficients.

FR models require the decoding of the encoded video sequences and can be further categorized in empirical and semi-analytical models. Empirical models require the computation of all extractable R-D points resulting in a high complexity. Semi-analytical models aim at reducing such complexity by deriving parameterized functions that follow the shape of analytically derived functions, but are evaluated through curve fitting from a subset of the R-D empirical data points. The latter offers an attractive trade-off between computational complexity and accuracy, in case of non-real-time or near-real time video streaming. In this chapter we first analyze and propose semi-analytical models for SVC video with reference to Medium Grain Scalability assuming FR and error-free transmission.

The PR models are derived by introducing new functions dependent only on scalar spatial and temporal parameter of the uncoded/coded video streams, which can be easily extracted during the encoding process. The coefficients of this new functions can be estimated off-line through a prior knowledge of the parameters of a set of sample video sequences, and then used for any future video sequence. We here propose a PR model which aims at estimating the parameter of the previously mentioned semi-analytical model according to two program-dependent indexes. All these models allow to accurately predict the distortion resulting by the lossy encoding process.

However, the quality of each video can be further heavily reduced due to the transmission errors and the consequent loss of part of the video stream. An automatic repeat-request (ARQ) schemes have the main drawback to increase the delay and can not be suitable for many application where the playback time is a stringent constraint. Within the framework of video delivery schemes based on SVC, Forward Error Correction (FEC) has been proposed to recover channel

errors and many contributions in the literature have proved its effectiveness [15, 16, 34]. Due to the different importance and the temporal/quality dependency of the different frames, Unequal Error Protection (UEP) or UXP schemes are generally more effective with respect to schemes based on equal protection. An UXP profiler has the aim to assign a different protection to each frame according to its dependencies and the related R-D improvements, as function of the average estimated packet-loss rate, *e.g.*, the loss rate of RTP packets in RTP transmission.

We then also propose a complete framework to jointly design UXP profiler and derive the resulting expected additional distortion due to error in the channels, as well as the related rate resulting after protection. Our proposal provides to each extractable sub-stream an approximately constant expected distortion for different values of RTP packet failure rate. This means that a change in the packet failure rate only induces a rate increment or decrement. This feature allows to model the expected continuous R-D relationship with the same proposed semi-analytical model for error-free transmission, where only a constant is added for different packet failure rates.

Many R-D models have been proposed in the literature for real time and non-real time video streaming (see for example [34–41] and references therein).

In [35], the authors proposed an accurate semi-analytical square-root model for MGS coding and compared it with linear and semi-linear models. They concluded that the best performance is obtained by changing the model according to a parameter that estimates the temporal complexity, evaluated before encoding the entire sequence. However, a general model for the estimation of the R-D relationship for a large set of video sequences, is necessary to derive analytical solutions for the rate-adaptation problem.

In [37] the authors present a detailed analysis of the R-D relationship in Fine Grain Scalability (FGS) coders and provide an accurate square root R-D model, which requires at least two empirical points. However, as mentioned, FGS has been removed from the SVC standards, due to its complexity.

In [40] the authors proposed a general semi-analytical R-D model for video compression, also verified in [34] for SVC FGS layer, where the relationship between rate and distortion depends on three sequence-dependent parameters which must be estimated through the evaluation of six empirical R-D points. We have verified this model with reference to SNR scalability with MGS and the high accuracy of the results led us to investigate a simplified two-parameters model with lower complexity, where the number of R-D points needed to estimate the parameters is reduced.

An improved real-time R-D model for Medium Grain Scalability (MGS) video coding was proposed in [39]. This model reduces significantly the dependency on

the encoding process. In this model the delay is reduced by extracting the parameters before transformation. Nevertheless it is showed that the model accuracy highly depends on the complexity of the video sequence.

The optimization of video streaming over packet-erasure channel is also highly investigated within the framework of SVC, *e.g.*, [16, 34, 42]. In [34] and in earlier works the authors proposed a complete framework to analyze and model the video streaming system over packet erasure channel, also in presence of play-out deadline. They derived an analytical model to estimate the the R-D in case of base-layer packet losses, while using a semianalytical model for the quality-layers. An UXP profiler, based on the same priority level assigner used in our work, solves a rate-minimizing cost functions. Maani et al. [16] proposed a model to solve the problem of joint bit extraction and channel rate allocation over packet erasure channels, where the level of protection of each enhancement layer is selected according to the expected distortion-to-rate gradient. However, differently from our proposed UXP profile, the resulting R-D relationship significantly depends on packet error probability and may result in a non-convex rate adaptation problem, which is generally much harder to be solved.

Contribution

The contributions of this chapter are summarized as follows

- we evaluate and compare two similar semi-analytical model for the estimation of the R-D relationship for SVC encoded videos transmitted over error-free channel.
- we propose a simple UXP profiler which provides almost similar values of distortion in the low-rate part of the R-D relationship for different values of RTP packet-loss rate; also closed form evaluation of distortion loss is provided. According to the proposed UXP profile, a R-D model considering also error-prone channel is proposed.
- we propose new techniques to further reduce the complexity of semi-analytical models for SVC scalable streams based on the introduction of new functions dependent only on the uncoded video streams. The coefficients of this new functions can be estimated off-line through a prior knowledge of the parameters of a set of sample video sequences, and then used for any future video sequence.
- we extend the proposed SVC R-D models to rate-to-quality models (RQ) for HAS sources.

- we derive, analyze and discuss the accuracy and the complexity of the proposes R-D models, according to extensive numerical evaluations.

3.1 Test Video Sequences

Table 3.1 summarizes the characteristics of the test video sequences that are used throughout this thesis. The first ten video sequences, are well-known video sequences mostly used within the JVT and they are available on-line in [43][44]. They comprises 300 frames, apart from *football* which has 260 frames.

The four sequences numbered from 11 to 14 are extracted from real video programs and comprise 2760 frames each.

The last three video sequences are health-related videos. Sequences 15 and 16 was acquired in realistic on board ambulance scenario thanks to the *"Green Cross Public Assistance Association" of Cesena (Italy)*. They comprise 300 frames each. Finally the ultrasonography sequence was gently provided by the Hospital of Perugia (Italy) and comprises 150 frames. All video sequences have frame rate equal to 30 fps. CIF, QCIF, nHD, VGA resolution (Res. in the Table) corresponds to 352×288 , 704×506, 640×360, 640×480 pixels, respectively.

N.	Sequence	Res.	Spatial Compl.	Temp. Compl.	Description
1	*City*	CIF	Medium	Medium	An urban area with several buildings
2	*Crew*	CIF	Low	Low	A crew walking
3	*Coastguard*	CIF	High	Medium/High	A Small boat in a river
4	*Container*	CIF	Low	Low	A container in the sea
5	*Football*	CIF	Medium	High	A football game
6	*Foreman*	CIF	Medium	Low	A foreman speaking
7	*Harbour*	CIF	Medium	High	Sailing boats slowly moving
8	*Mobile*	CIF	High	Medium	A ball rolling over a desk
9	*News*	CIF	Low	Low	A chinese news
10	*Soccer*	CIF	Medium	Medium	A soccer game
11	*Sport*	4CIF	Very High	Very High	Canoe competition
12	*Interview*	4CIF	Low	Very Low	An interview
13	*Bunny*	4CIF	High	Medium	Extract of *Big Buck Bunny* movie
14	*Home*	4CIF	Medium/High	Low	Extract of *Home* cartoon movie
15	*Ambient 1*	nHD	Medium	Low	Man in a ambulance (far view)
16	*Ambient 2*	nHD	Medium	Medium	Man in a ambulance (close view)
17	*Ultrasound*	VGA	Low	Low	An Ultrasonography video

Table 3.1: Test video sequences: spatial (spat.) and temporal (temp.) complexity (compl.) and general description. The resolution (Res.) CIF, QCIF, nHD, VGA corresponds to 352×288, 704×506, 640×360, 640×480 pixels, respectively.

3.2 Objective Video Quality Metrics

The simplest video quality metric to assess the quality of a reconstructed video sequence are the Mean Square Error (MSE) and PSNR, which were historically adopted in image processing in order to evaluate the performance of the codec of interest. Although simple to implement and to compute, they are generally not considered always reliable. Nevertheless, their use continues to be predominant in the performance evaluation of any video coding system.

Let us define $x[m,n]$ as the original signal at pixel $[m,n]$ of the $i-th$ frame of a video with resolution $M \times N$, and $y[m,n]$ as the associated reconstructed signal. The MSE between the original and reconstructed i-th picture, is evaluated as:

$$MSE[i] = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}(x[m,n]-y[m,n])^2 \tag{3.1}$$

while the average MSE between the original and reconstructed set of pictures $\mathcal{I}$ with cardinality I, composing a video scene is defined as:

$$MSE = \frac{1}{I}\sum_{i\in\mathcal{I}} MSE[i] \tag{3.2}$$

The PSNR of the i-th frame is derived by setting the MSE in relation to the maximum possible value of the luminance (for a typical 8-bit value this is $2^8-1 = 255$) as follows:

$$PSNR[i] = 10\log_{10}\left(\frac{255^2}{MSE[i]}\right) \tag{3.3}$$

The result is a single number in decibels [dB], ranging from 30 to 40 for medium to high quality reconstructed pictures. Two different ways of computing the PSNR of a video scene $\mathcal{I}$, namely PSNR and Average PSNR (APSNR), are possible according on how the average is performed. However, the correct way to calculate average PSNR for a sequence is to calculate average MSE for all frames as in (3.2) and after that to calculate PSNR using ordinary equation for PSNR, *i.e.*,

$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right) \tag{3.4}$$

Nevertheless, sometimes it is needed to take simple average of all the per frame PSNR values, *i.e.*,

$$APSNR = \frac{10}{I}\sum_{i\in\mathcal{I}}\log_{10}\left(\frac{255^2}{MSE[i]}\right) \tag{3.5}$$

Due to their simplicity MSE and PSNR will be mostly used to evaluate the rate-to-quality relationship of the scalable video stream considered throughout this thesis. Nevertheless, due to their poor correlation with subjective quality tests, we will also consider enhanced quality metric, *i.e.*, Structural SIMilarity (SSIM) and the American National Standards Institute (ANSI) Video Quality Model (VQM).

The SSIM index is a method for measuring the similarity between two images proposed by Zhou Wang *et. al.*, [45]. This method differs from the previously described methods, which all are error based, since it uses the structural distortion measurement instead of the error. The idea behind this is that the human vision system is highly specialized in extracting structural information from the viewing field and it is not specialized in extracting the errors. Thus, a measurement on structural distortion should give a better correlation to the subjective impression.

Many different quality assessment methods can be developed from this assumption but Wang proposes a simple but effective index algorithm. The SSIM index of the i-th frame is expressed as

$$SSIM_{\text{index}} = \frac{(2\mu_x\mu_y + a_1)(2\sigma_{x,y} + a_2)}{(\mu_x^2 + \mu_y^2 + a1)(\sigma_x + \sigma_y + a_2)} \tag{3.6}$$

where μ_x, μ_y, σ_x, σ_y, $\sigma_{x,y}$, are the mean of x, the mean of y, the variance of x, the variance of y and the covariance of x and y respectively, while a_1, a_2 are constants. The value of SSIM is between -1 and 1 and gets the best value of 1 if $x[n,m] = y[n,m], \forall n, m$. The quality index is applied to every image using a sliding window with 11×11 circular-symmetric Gaussian weighting function for which the quality index is calculated and the total index of the image is the average of all the quality indexes of the image.

The VQM [46] was developed by the Institute for Telecommunication Science (ITS) to provide an objective measurement for perceived video quality. It measures the perceptual effects of video impairments including blurring, jerky/unnatural motion, global noise, block distortion and color distortion, and combines them into a single metric.

The VQM considers the original and the processed video as input and it is computed according to the following steps:

- *Calibration*: it performs an estimate and a correction of the spatial and temporal shift as well as of the contrast and of brightness offset of the processed video sequence with respect to the original video sequence.

- *Quality Features Extraction*: it extracts a set of quality features that charac-

terizes perceptual changes in the spatial, temporal, and chrominance properties from spatial-temporal sub-regions of video streams.

- *Quality Parameters Computation*: it computes a set of quality parameters that describe perceptual changes in video quality by comparing features extracted from the processed video with those extracted from the original video.

- *VQM Computation*: the VQM value is computed using a linear combination of parameters computed in the previous steps.

The VQM value is a number between 0 and 1 used to judge the visual quality. A low VQM value indicates good perceived quality. Extensive subjective and objective tests were conducted to verify the performance of the VQM. The results show a high Pearson correlation coefficient, around 0.95, between subjective tests and the VQM. For this reason it has been adopted by ANSI as an objective video quality standard.

In the next sections we will analyze the R-D relationship of SVC encoder with respect to SNR scalability with MGS coding in terms of the MSE, as well, as the R-Q relationship of HAS sources in terms of the SSIM. VQM is considered in chapter 5 to further validate the proposed cross-layer framework. All the R-D models are extensively tested for the different video sequences mentioned in the previous section.

3.3 Non-Real-Time Rate-Distortion Models for SVC in Error-Free Channels

In this section we first analyze and propose two semi-analytical models to estimate the R-D relationship at the SVC encoder assuming SNR-scalability.

Let us consider an SNR-scalable video stream resulting from the encoding of a set $\mathcal{I}_k$ of pictures, intended for user k. We define $\mathcal{D}_k^{\text{enc}} = \{d_{1,k}^{\text{enc}}, ..., d_{E_k,k}^{\text{enc}}\}$ as the set of distortion values, one for each extractable sub-stream, whose total number is E_k. The encoder distortion $d_{e,k}^{\text{enc}}$, $e = 1, ..., E_k$, given by the MSE between the original and the reconstructed pictures averaged over $\mathcal{I}_k$ is computed as in eq. (3.2). The R-D theory evaluates the minimum bit-rate F_k required to transmit the k-th stream with a given expected distortion $d_{s,k}$, by defining a function F_k that maps the distortion to the rate, *i.e.*,

$$\begin{aligned} F_k: \quad & \mathcal{D}_k^{\text{enc}} \rightarrow \mathbb{R}^+ \\ & d_{s,k}^{\text{enc}} \rightarrow F_k(d_{s,k}^{\text{enc}}) \end{aligned} \tag{3.7}$$

One of the desirable properties of F_k is the strictly decreasing monotony, *i.e.*

$$F_k(d_{i,k}^{\text{enc}}) > F_k(d_{j,k}^{\text{enc}}), \quad \forall d_{i,k}^{\text{enc}}, d_{j,k}^{\text{enc}} : d_{i,k}^{\text{enc}} < d_{j,k}^{\text{enc}}. \tag{3.8}$$

implying that for any increasing of the information rate corresponds a decrease in distortion. The rate $F_k(d_{i,k}^{\text{enc}})$, evaluated here in bps (bit per second), is generally function of discrete values. Following the approach in [9, 34, 40] the expected R-D relationship is modeled through a parametric function $F_k(D)$ of a continuous variable D.

$$F_k(D) = \frac{\alpha_k}{D + \xi_k} + \beta_k, \ D \in [D_k^{\text{hl}}, D_k^{\text{bl}}], \tag{3.9}$$

where the parameters α_k, ξ_k and β_k, with $\alpha_k, \xi_k > 0$, $\forall k$ depend on the temporal and spatial complexity of the set of pictures $\mathcal{I}_k$ and on on the frame rate. The values of

$$D_k^{\text{bl}} = \max_{s \in D_k^{\text{enc}}} d_{s,k}^{\text{enc}} \tag{3.10}$$

and

$$D_k^{\text{hl}} = \min_{s \in D_k^{\text{enc}}} d_{s,k}^{\text{enc}} \tag{3.11}$$

are the expected distortions of the set of pictures $\mathcal{I}_k$, after decoding the base layer and the highest enhancement layer, respectively.

The drawback of this approach is the need to estimate the three video sequence dependent parameters, α_k, ξ_k and β_k, by using curve-fitting over a subset of the R-D data points. According to extensive simulations, the curve-fitting algorithm requires a minimum of six empirical R-D points and a relevant number of iterations and function evaluations to exhibit high accuracy for most sequences.

In order to reduce the complexity, we have simplified this parametrized model by eliminating one parameter, *i.e.*,

$$F_k(D) = \frac{\alpha_k}{D} + \beta_k \tag{3.12}$$

In this case, four R-D points are generally sufficient to estimate the two sequence-dependent parameters α_k and β_k, with high accuracy; as a result, the number of iterations and function evaluations decreases. Beside the complexity reduction, this model allows a simple derivation of the solution of the fairness-oriented rate adaptation problem, as we will show in the cross-layer optimization

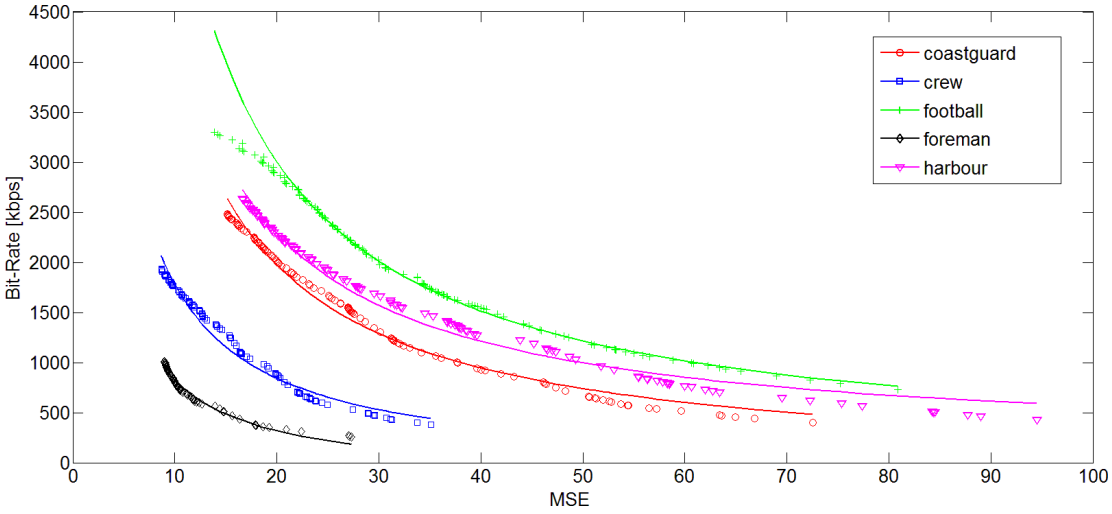


Figure 3.1: R-D Model (straight line), according to eq. (3.12) fitting the empirical R-D relationship for the GOP with the worst $RMSE$ with reference to Table 3.2.

framework proposed in the next chapters. The selection of the empirical points depends on the range where the R-D curve is defined. However, to provide more accuracy the range of interest can be suitably reduced.

Table 3.2 compares the goodness of the two models with respect to the coefficient of determination R^2[47], the RMSE, the average number of iterations (ANoI) and function evaluations (ANoFE) required by non-linear Least Square Trust-Region (LSTR) curve-fitting algorithm to converge. It can be noted how the number of function evaluations, as well as the number of iterations, decreases while a minimum loss occurs in the goodness parameter. In Figure 3.1, we plot the empirical R-D relationship for five test video sequences, as well as their related R-D curves based on model (3.12). All of them are referred to the GOP with the worst RMSE value (the minimum in Table 3.2). We can also appreciate in this figure the achievable granularity of the quality-based extraction method.

3.4 Non-Real-Time Rate-Distortion Models for SVC in Error-Prone Channels

We here extend the proposed semi-analytical R-D models to estimate the expected distortion in case of transmission over error-prone channel.

Let us now consider an SNR-scalable video stream resulting from the encoding

Video	Model	R^2 [min,max]	$RMSE$ [min,max]	ANoI	ANoFE
Coastguard	Model (3.12)	[0.9812 , 0.9921]	[37.895 , 79.992]	30.2	89.6
	Model (3.9)	[0.9956 , 0.9982]	[22.261 , 36.724]	34.7	155.9
Crew	Model (3.12)	[0.9795 , 0.9934]	[23.038 , 89.130]	30.9	94.2
	Model (3.9)	[0.9914 , 0.9972]	[20.019 , 52.489]	35.6	159.9
Football	Model (3.12)	[0.9662 , 0.9891]	[53.403 , 205.572]	29.0	89.5
	Model (3.9)	[0.9839 , 0.9993]	[12.940 , 99.810]	38.0	169.3
Foreman	Model (3.12)	[0.9669 , 0.9955]	[19.710 , 53.371]	25.7	73.2
	Model (3.9)	[0.9914 , 0.9980]	[13.516 , 33.745]	34.1	154.3
Harbour	Model (3.12)	[0.9823 , 0.9929]	[51.860 , 73.344]	37.5	129.8
	Model (3.9)	[0.9952 , 0.9991]	[18.883 , 44.822]	45.3	164.3

Table 3.2: Comparison between the two semi-analytical model in (3.9) and (3.12) with respect to the minimum and maximum $RMSE$, the coefficient of determination R^2, the Average Number of Iterations (ANoI) and the Average Number of Function Evaluation (ANoFE), evaluated for each GOP (GOP size G equal to 8) of five video sequences with CIF resolution and frame rate of 30 fps. The video are encoded with one base layer (QP equal to 38) and two enhancement layers (QP equal to 32 and 26), both with 5 MGS layers and a weights vector equal to [3 2 4 2 5], ($Q = 10$).

of a set $\mathcal{I}_k$ of pictures, intended for user k, which has to be transmitted in an error-prone channel . We define $\mathcal{D}_k = \{d_{1,k}, ..., d_{E_k,k}\}$ as the set of expected distortion values, one for each extractable sub-stream, whose total number is E_k. The distortion $d_{e,k}$, $e = 1, ..., E_k$, given by the Mean Square Error (MSE) between the original and the reconstructed pictures averaged over $\mathcal{I}_k$, is computed as

$$d_{e,k} = d_{e,k}^{\text{enc}} + d_{e,k}^{\text{loss}}, \tag{3.13}$$

where $d_{e,k}^{\text{loss}}$ is the additional distortion due to the packet losses in the error-prone channel, which is function of the frame loss probability as well as on the protection scheme selected as showed next.

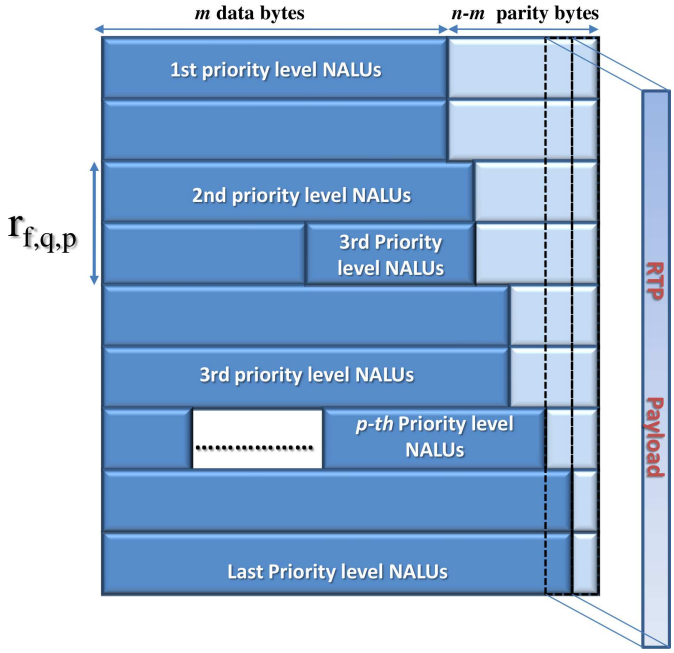


Figure 3.2: Transmission Sub-Block (TSB) structure. Following the priority level, the NALUs of one GOP are placed into one TSB according to a given UXP profile (protection class) from upper left to lower right. The columns of one ore more TSB are then encapsulated into RTP packets

3.4.1 Unequal Erasure Protection for SVC Streams

Due to the different importance and the temporal/quality dependency of the different frames, UXP schemes can generally overcome schemes based on equal protection. In our work, we follow the guidelines presented and discussed in [15] for RTP video transmission over packet-erasure channel, by focusing our attention on a GOP-based transmission. In this approach, each GOP is mapped into one Transmission Sub-Block (TSB) that carries either data and parity bytes, as exemplified in Figure 3.2. Each row of the TSB identifies a RS (n, m) codeword where m is number of data bytes and n is the total bytes of the codeword. If a packet-erasure detection is available at the lower-layers, the RS codes are able to correct up to $n - m$ bytes, equal to the number of parity bytes.

The aim of the UXP profiler is to assign a different protection to each frame according to its dependencies and R-D improvements.

A first step is to order the NALUs according to their protection class. A

priority index greater than 62 is re-assigned to the different temporal base layer frames ($q = 0$), to have lower priority indexes for high temporal indexes. Thus, all the frames are sorted according to the priority level p and sequentially inserted into one TSB, according a given UXP profile $\boldsymbol{M}^* = \{m^*_{f,q,p}\}$, where $m^*_{f,q,p}$ identify the protection class assigned to frame with frame index f, quality index q and priority level p.

Finally, one or more TSB are placed into a transmission block (TB) whose columns become the payload of RTP packets. In this way the RS codewords are interleaved over the different RTP packets. Therefore, RTP packet errors (or erasures) can be assumed as uniformly distributed inside the codewords. In order to reduce the overhead due to the need of padding for compensating the different NALU lengths, the part of the codeword left unused by a given NALU is filled with the data from the subsequent NALU. For simplicity of presentation and without loosing generality, we assume that the size $S_{f,q,p}$ of each NALU is always greater than or equal to the total size n of the RS code:

$$S_{f,q,p} \geq n \tag{3.14}$$

This assumption ensure that each TSB row contains no more than two different frames.

Let us finally note that a Multi Time Aggregation Packet (MTAP) header must be inserted before each priority level NALU in order to deliver the decoding order number (DON) and timing information assignment.

3.4.2 Frame Error Probability and Expected Distortion

Let assume that the RTP packet error rate information $P_{\mathrm{e}}^{\mathrm{RTP}}$, is periodically collected from the lower-layers. According to the proposed UXP scheme a closed formulation of the expected error probability can be derived by using the failure probability of a single (n, m) RS codeword:

$$P(n,m) = \sum_{i=m-n+1}^{n} \binom{n}{i} (P_{\mathrm{e}}^{\mathrm{RTP}})^i (1 - P_{\mathrm{e}}^{\mathrm{RTP}})^{n-i} \tag{3.15}$$

The individual frame error probability now depends on the number of TB rows associated to each frame, *i.e.*,

$$r_{f,q,p} = \left\lceil \frac{S_{f,q,p}}{m^*_{f,q,p}} \right\rceil \tag{3.16}$$

and on whether or not some bytes of the frame are inserted in the row using the protection class of the preceding priority level. Let $z \in \{0, 1\}$ be a boolean variable that indicates whether or not this last event occurs. The frame error probability FEP is then computed as one minus the probability that all codewords of the TB, associated to the frame, can be correctly decoded by the RS decoder:

$$FEP_{f,q} = 1 - \left[\Big(1 - P(m^*_{f,q,p}, n)\Big)^{r_{f,q,p}} \Big(1 - P(m^*_{f,q,p-1}, n)\Big)^{z} \right] \tag{3.17}$$

According to the derived FEP, a closed formula for the expected distortion can be now computed. Let $\Upsilon D_{f,q} = |d_{f,q} - d_{f,q-1}|$ be the quality improvement resulting from the correct decoding of the $f - th$ frame with quality id q, which is computed by the priority level assigner. In order to compute the quality improvement $\Upsilon D_{f,0}$ due to the enhancement (temporal) frames of the base layer we assume an error concealment (EC) method based on the picture copy (PC). Therefore the distortion increment due to the loss of an enhancement picture is computed by considering the difference between the enhancement frame and the copy of the previous one. The expected distortion due to the loss of frames with quality index $q \leq Q$ can be computed as:

$$d_{f,q,loss} = \sum_{r=0}^{q} \Upsilon D_{f,r} \left[FEP_{f,0} u_{f-1} + \sum_{j=1}^{q} FEP_{f,j} \prod_{s}^{j-1} \Big(1 - FEP_{f,s}\Big) \right] \tag{3.18}$$

where u_x is the Heaviside function[1]. The first term of the sum takes into account the distortion due to the loss of a temporal enhancement layer. Since a loss of the I-frame will results in a infinite distortion we assume here that the associated NALUs will receive enough protection to have $FEP_{0,0}$ close to zero.

The second sum, on the other hand, takes into account the cumulative probability that the $j - 1$ quality layers have been successfully received but the j-th quality frame is lost, where $j \leq q$. Finally, the total expected distortion of the entire GOP is the sum of the individual frame loss distortions:

$$d_{s,loss} = \sum_{f=0}^{G-1} d_{f,q,loss} \tag{3.19}$$

Let us note that the number of quality layers of each frame in one GOP can be different after the rate adaptation. Thus, the index s maps the vector whose

[1] $u_x = 0$ if $x < 0$, 1 otherwise, $x \in \mathbb{Z}$

elements are the resulting number of the quality-layer of each frame f: its range is from 0 to GQ. The values of the expected distortion can be finally used, together with the required rate, to reshape the R-D relationship according to the values of the FEP.

3.4.3 Proposed UXP Profiler

The derivation of an optimal UXP profile is hard to achieve. It should be computed according to the solutions of an optimization problem aimed at balancing the trade-off between protection and overhead. This it is a discrete problem since the FEP, as well as the overhead resulting from the RS encoding, strictly depends on the discrete variable m, as shown in Figure 3.3. In order to guarantee a rate distortion relationship strictly decreasing, the FEP of each frame should increase as the quality and the temporal indexes increase. However, due to the granularity of the available values of m, sometimes this condition is not met. This problem could be partially solved by a joint optimization of the encoding process and the UXP profiler. However, this is out of the scope of this work. In our framework the UXP profiler simply drops this cases by slightly compromising the R-D granularity.

We propose a simple strategies by fixing an error probability profile (EPP) $\boldsymbol{\pi}_{f,q,p}$, for each frame f with quality id q and priority level p. Based on this approach, the UXP profile is derived by finding the minimum $m_{f,q,p} \in [\frac{n}{2}+1, n]$ such that

$$FEP_{f,q} \leq \pi_{f,q,p} \tag{3.20}$$

Differently to other solutions in literature, this approach has the main advantage that the expected distortion becomes quasi-independent from the RTP packet failure rate whereas a change of the $P_{e,rtp}$ will only results in a rate increment or decrement. By exploiting the proposed design, the UXP profiler can adaptively adjust the amount of redundancy according to a target value of RTP packet loss rate provided by the BS that serves the destination users. The RTP packet loss rate information can be fixed to a constant conservative value or it can be estimated through error rate measurements.

As a case of study to provide numerical results and illustrate how rate adaptation works when UXP is implemented, we consider here the following choice for the EPP, by differentiating the base and the enhancement layer protections.

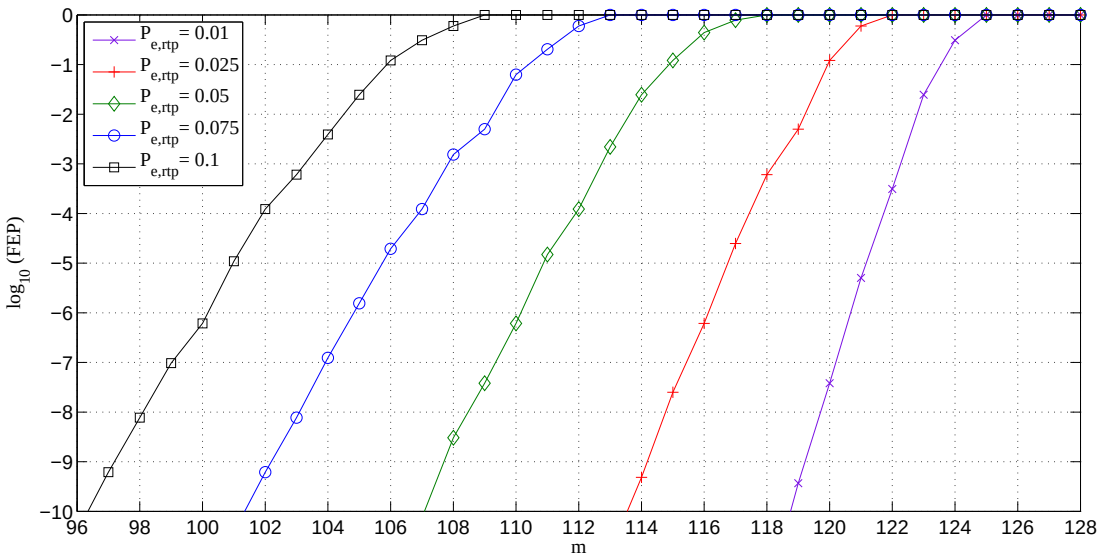


Figure 3.3: Resulting logarithmic FEP for the first I frame of *Football* (byte size equal to 11519) mapped to RS codewords $(128, m)$ at different RTP packet error probability

A Case Study for the Design of EEP

Since the priority level of the quality layers carries both the information of the R-D improvements and the dependency of each frame, the values of the EPP for the quality frames, i.e. $q > 0$, can be derived according to the following formula

$$\pi_{f,q,p} = \begin{cases} \left(\frac{p}{\alpha}\right) 10^{-\frac{p}{\alpha}} & \text{if } p \geq \frac{\alpha}{ln(10)} \\ 1 + \left(\frac{1}{e} - \ln(10)\right)\frac{p}{\alpha} & \text{otherwise} \end{cases} \tag{3.21}$$

where α allows for a trade-off between protection and overhead.

The priority levels for the base layer frames are normally set equal to 63 by the quality processing tool. If the UXP profile used eq. (3.21), it would assign similar protection to the base layer and the first enhancement layers. A smaller frame error rate is ensured for the I-frame, since its loss will produce the drop of all the frames in the GOP. To avoid this we set then $\pi_{0,0,p} = 10^{-6} \quad \forall\alpha$. Moreover, in order to exploit the temporal scalability at the decoder we propose to re-assign to frames of the enhancement temporal layer, with $q = 0$, an higher priority level and to use again the eq. (3.21) to derive the relative EEP values. The choice of the priority level for the enhancement temporal layer depends on the particular

frame rate that must be ensured to each user.

The model (3.12) and (3.9) for the R-D relationship is still applicable in case of frame losses due to the transmission error in the channel. In this case the empirical points of the encoder are replaced by new points taking into account the effects of packet erasures and UXP. These new points are the result of the rate increase due to UXP, *i.e.*,

$$\sum_{f=0}^{G-1} \frac{n - m^*_{f,q,p}}{m^*_{f,q,p}} r_{f,q,p}, \tag{3.22}$$

and the novel expected distortion $d_{s,loss}$ evaluated as in (3.19). In Figure 3.4 we plot the empirical R-D function resulting from the encoder, as the reference curve, and the related R-D functions outcoming from the UXP profiler at different packet error probabilities $P_{\mathrm{e}}^{\mathrm{RTP}} > 0$ for the first GOP of the test-sequence *Footballl*. We can see that the distortion is almost unchanged for the lower points of the curve with respect to the reference case, since high protection is provided to the high priority levels which are the first to be extracted. At larger bit rates the gap with respect to the reference case increases due to insertion of quality frames with lower protection.

Generally a dynamic adaptation of the UXP to different $P_{\mathrm{e}}^{\mathrm{RTP}}$ would require the periodical application of the curve-fitting algorithm to derive the two parameters of the model, thereby increasing the complexity. This problem can be overcome when the UXP profiler adaptively tracks the FEP profile by changing the protection class assigned to the different NALUs. In this way only rate has significant changes while expected distortion practically does not change. While comparing the empirical points resulting from different error probabilities ($P_{\mathrm{e}}^{\mathrm{RTP}} > 0$), we can note in the figure how the proposed UXP profile leads to similar distortion at different $P_{\mathrm{e}}^{\mathrm{RTP}}$ values. Therefore the adaptation module adapts the sequence-dependent parameters by simply adding a constant dependent on the value of $P_{\mathrm{e}}^{\mathrm{RTP}}$. According to extensive simulations the rate shifting is independent of the encoded sequence and can be determined by empirical evaluations.

This feature allows to model the expected R-D relationship through the same parametric function in (3.9) or (3.12), where only β_k changes for different design values of RTP packet-loss rate. This result can also be appreciated in Table 3.3 where the average expected distortion due to different $P_{\mathrm{e}}^{\mathrm{RTP}}$ and the resulting average overhead is evaluated for two video sequences with full quality scalability.

The selection of a small value of α for the EEP results in a small FEP for the quality layers, thereby increasing the overhead. On the other hand, a loss in the expected quality is experienced by doubling α with a consequent rate gain in the

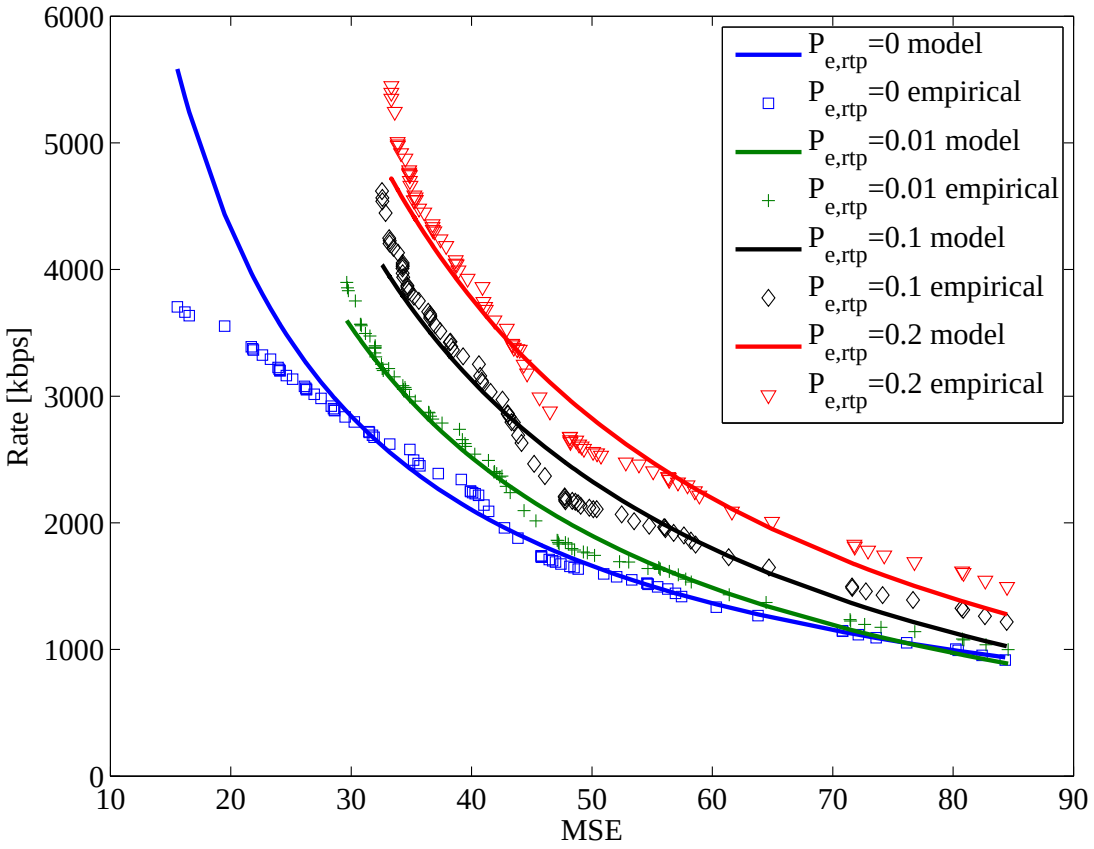


Figure 3.4: R-D Model (straight line), according to eq. (3.12) fitting the empirical R-D relationship for one GOP (size G equal to 8) of the *Football* test-sequence with different error probabilities and α=30. The lower curve refers to the R-D relationship of the encoder.

order of 5%. As mentioned before, the overhead is approximately constant even for video sequences with high spatial and temporal complexity difference, such as *Foreman* and *Harbour*. On the other hand, the loss in the expected quality strictly depends on the range of the distortion values as normally increase with the complexity of the video raises if the same encoding paradigm is used for each sequence.

Video	P_e^{RTP}	$\alpha = 15$		$\alpha = 30$	
		Overhead	$d_{GQ,loss}$ [MSE]	Overhead	$d_{GQ,loss}$ [MSE]
Foreman	0.01	8.4 %	1.82	5.3 %	4.54
	0.05	17.7 %	2.13	13.7 %	5.15
	0.1	28.0 %	2.17	23.1 %	5.31
Harbour	0.01	7.8 %	8.95	5.1 %	19.87
	0.05	17.1 %	9.86	13.3 %	20.32
	0.1	27.6 %	10.13	23.4 %	20.89

Table 3.3: Percentage of the overhead and expected distortion $d_{GQ,loss}$ in term of MSE with respect to the full quality video streams ($Q = 10$ and $G = 8$), for different values of RTP packet error probability and α parameter in the EEP profile

3.5 Real-time Rate-Distortion Models for SVC Streams

The time required to model the R-D curve for a given sequence may drive the decision on the methodology/algorithm to be adopted for the R-D modeling. On the other hand, the performance of the streaming system is directly affected by the accuracy of the R-D model [36]. For real time video streaming systems the computation of the model should be fast enough to deal with the timing constraints of the video stream. Hence, we investigate here techniques to further reduce the complexity of semi-analytical models. This is made possible by introducing new functions dependent only on the uncoded video streams. The coefficients of this new functions can be estimated off-line through a prior knowledge of the parameters of a set of sample video sequences, and then used for any future video sequence. Such new model only uses two parameters, *i.e.*, the Spatial Index (SI) and the Temporal Index (TI), which are calculated taking into account the characteristics of the video sequences through a spatial and a temporal index extracted from the original raw video streams. Moreover, we also use these complexity indexes to calculate BL and EL rates of the given video stream. We consider as a reference R-D model the model in eq. (3.12) introduced for MGS coded video,

As already mentioned, the drawback of this model is the fact that its parameters can only be evaluated by looking for the best fitting of at least 4 R-D points after the encoding process of the video, hence the model is not suited for real time

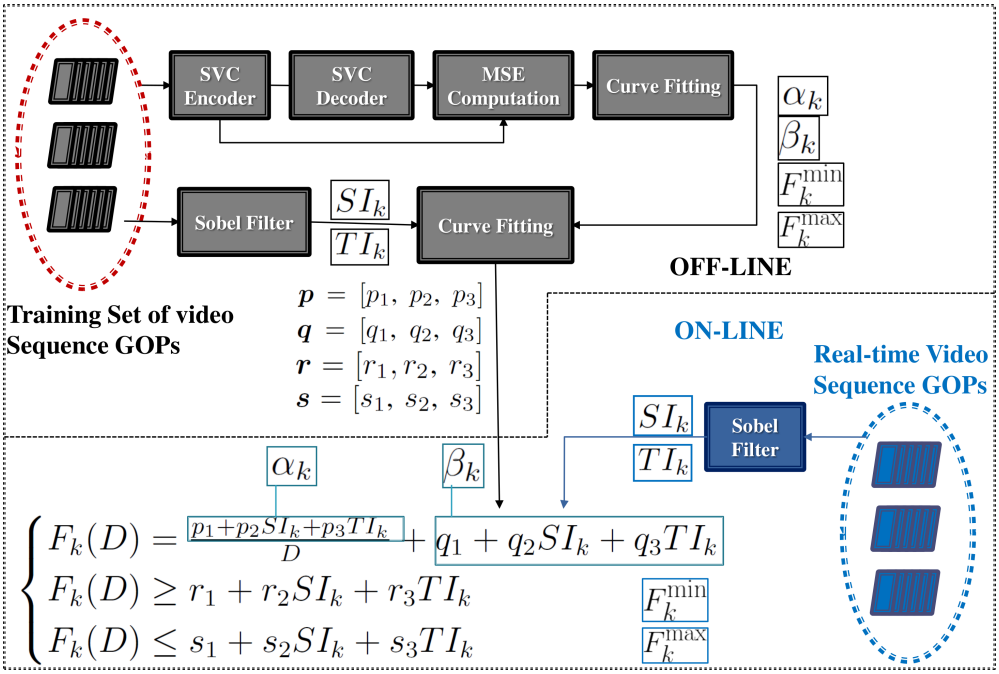


Figure 3.5: Proposed strategy for real-time R-D modeling.

applications. The model proposed here replaces the parameters α_k and β_k with a function of the spatial index SI_k and the temporal index TI_k, as explained in the following:

$$\alpha_k = p_1 + p_2 SI_k + p_3 TI_k \tag{3.23}$$

$$\beta_k = q_1 + q_2 SI_k + q_3 TI_k \tag{3.24}$$

The same approach is used to express the BL (base layer) and EL (enhancement layer) rates:

$$F_k^{\min} = r_1 + r_2 SI_k + r_3 TI_k \tag{3.25}$$

$$F_k^{\max} = s_1 + s_2 SI_k + s_3 TI_k \tag{3.26}$$

The values on the sets $\boldsymbol{p} = [p_1, p_2, p_3]$, $\boldsymbol{q} = [q_1, q_2, q_3]$, $\boldsymbol{r} = [r_1, r_2, r_3]$ and $\boldsymbol{s} = [s_1, s_2, s_3]$ are coefficients that can be calculated by using fitting methods in a sufficiently large set of GOPs from a set of video sequences (training set). As mentioned above, this process is executed off-line only once.

The SI and TI values are evaluated on the luminance component [48] of the video by means of Spatial Information and Temporal Information [49] of the k-th GOP as follows:

$$SI_k = \max_i std_\sigma\{Sobel(x[i](\sigma))\} \tag{3.27}$$

$$TI_k = \max_i std_\sigma\{M[i](\sigma)\} \tag{3.28}$$

where $M[i](\sigma) = x[i](\sigma) - x[i-1](\sigma)$ is the motion difference, $x[i](\sigma)$ is the luminance component and i and σ are the temporal and spatial coordinates, respectively, of the frames used to encode GOP k.

To summarize, the R-D model is obtained by substituting in (3.12) the parameters α_n and β_n from (3.23) and (3.24), and $F_k^{\min}$ and $F^{\max}$ from (3.25) and (3.26), respectively, *i.e.*,

$$\begin{cases} F_k(D) = \frac{p_1+p_2SI_k+p_3TI_k}{D} + q_1 + q_2SI_k + q_3TI_k \\ F_k(D) \geq r_1 + r_2SI_k + r_3TI_k \\ F_k(D) \leq s_1 + s_2SI_k + s_3TI_k \end{cases} \tag{3.29}$$

A diagram block of the proposed strategy is presented in fig. 3.5

The proposed R-D model is verified by considering video sequences generated by the JSVM software [30]. We encoded six video sequences, *i.e.*, *Crew*, *Football*, *Coastguard*, *Soccer*, *City*, and *Mother and Daughter* (*MD*) having different scene complexities, in CIF resolution with a frame rate of 30 fps. We denote this set as the training set. Two ELs are used to obtain SNR scalability where each layer is split into 5 MGS layers with vector distribution of [3 2 4 2 5]. All the videos are coded GOP by GOP with a GOP size of 8 to obtain sequences comprising 26 GOPs. The Quantization Parameter is set to 38, 32 and 26 to obtain the BL and two ELs.

Fig. 3.5 shows α_k, β_k, BL and highest EL models as in (3.23), (3.24), (3.25) and (3.26), respectively, using the spatial and temporal indexes. In the two upper figures the markers are referred to the values of α_k and β_k derived according to model (1) and plotted for each GOP versus the corresponding value of SI_k and TI_k. In the two lower figures the markers are referred to the BL and EL layer rates derived by encoding the sequences with JSVM [30]. It can be observed that the values of the parameters for all the models closely follow a linear behavior. The metrics used to evaluate the goodness of the model in fitting the set of points are reported in the caption. The sets of coefficients, appearing in (3.23), (3.24), (3.25) and (3.26), of the proposed model, are calculated using the linear least square fitting method [50] with Least Absolute Residuals (LAR) [51] for robustness. The resulting values for the training set are the following:

$$\boldsymbol{p} = [-2.4 \times 10^4, 3975, 540.5] \qquad \boldsymbol{q} = [-246.1, 24.1, 3.3]$$
$$\boldsymbol{r} = [41.27, 17.09, 9.12] \qquad \boldsymbol{s} = [-237, 145.6, 34.02]$$

In Fig. 3.7 the different R-D models are shown and compared for two sample GOPs of three video sequences. The accuracy changes GOP by GOP: the upper

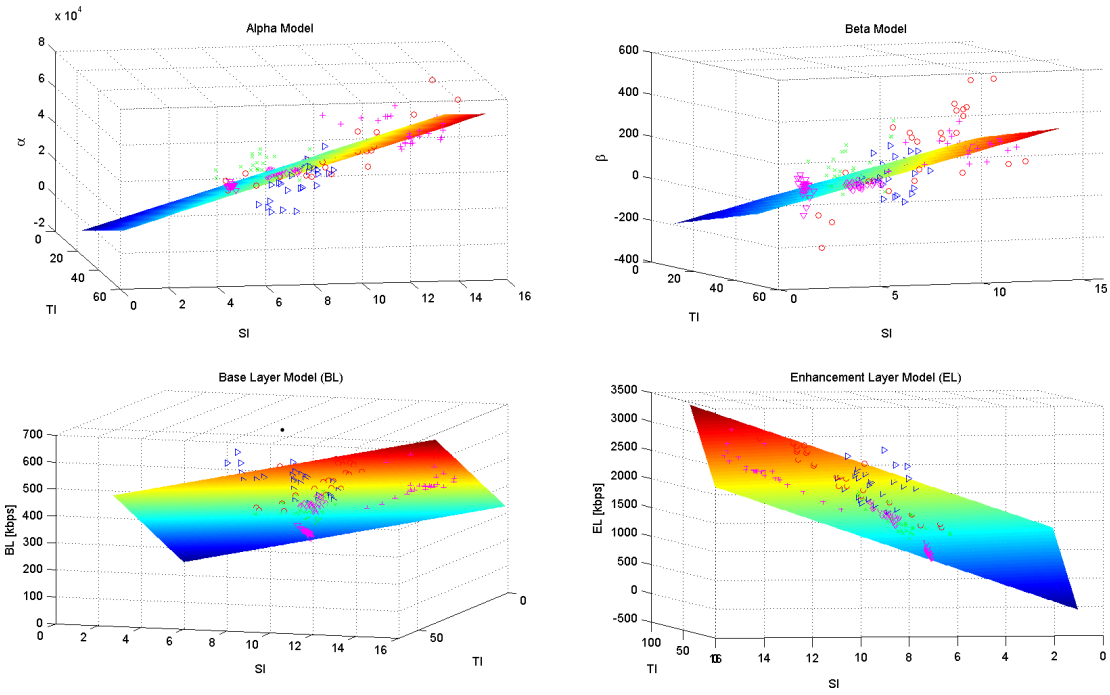
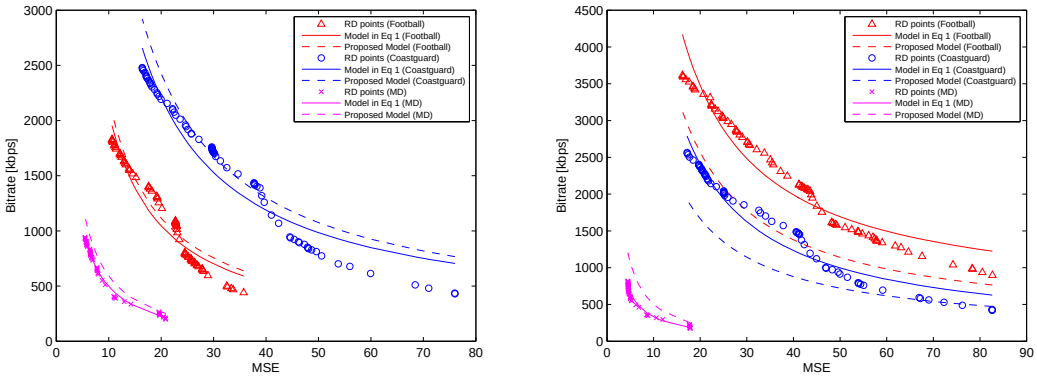


Figure 3.6: Proposed Models for α, β, BL and EL rates. The parameters used for the goodness of the models are the coefficient of determination (R^2) and Root Mean Square Error ($RMSE$). (α) R^2= 0.987 $RMSE$ = 1598, (β) R^2 = 0.973 $RMSE$ = 21.2, (BL) R^2 = 0.979 $RMSE$ = 22.98, (EL) R^2 = 0.985 $RMSE$ = 79.36

figure shows the result for a GOP with good matching between the proposed model and the model in (1), whereas the lower figure shows a result with poor matching. As shown below, the GOPs with less accurate model do not have significant impact on the behavior of rate adaptation strategies in real time multi-video transmission. To evaluate the goodness of BL and EL rate estimation, we compare in Fig. 3 the rates estimated with the model in (3.25) and (3.26), to the original rates obtained from the encoded sequences.

We consider not only the video sequences in the training set but also the sequences outside the training set. More emphasis is given to BL rate as it is the minimum rate requirement of each video sequence when transmitted in bandwidth constrained channels. It can be observed from Fig. 3 that our model predicts the BL rate quite accurately for sequences outside the training set, as shown for *Mobile* and *Foreman*. Moreover, it can be seen that the estimation is also good



(a) (b)

Figure 3.7: R-D comparison among model in eq. (3.12), proposed model and actual values for two sample GOPs.

for EL rate.

3.6 Rate-to-Quality Models for HAS Streams

The proposed approach for modeling the R-D relationship of SVC SNR-scalable video stream can be easily extended to consider the case of HAS-based encoded videos.

Similarly, each profile corresponds to an extractable sub-stream, *i.e.*, using the notation introduce in section 3.3 and in section 2.2, $E_k = M$, while the set of pictures $\mathcal{I}_k$ refers to one chunk. We run several simulation by encoding each chunk and each profile with the *x264* encoder [52], *i.e.*, a fast version of the H.264/AVC standard, and we finally extract the average MSE, PSNR and the SSIM. We have verified the model in (3.9) to describe the R-D relationship, which still provide high accuracy. Nevertheless, we are here interested to model the R-Q relationship of HAS sources in terms of SSIM quality metric, which will be used in chapter 7 to quantify the k-th user utility $U_k(R_k)$ of downloading a chunk from video k at a certain rate R_k. We found that the following continuous logarithmic SSIM to rate model in the interval of interest $[A_k, B_k]$, where A_k, B_k are the minimum and maximum available profile rate, have an high correlation with respect to the empirical points:

$$U_k(R_k) = a_1 \log(a_2 R_k + a_3), \quad R_k \in [A_k, B_k] \tag{3.30}$$

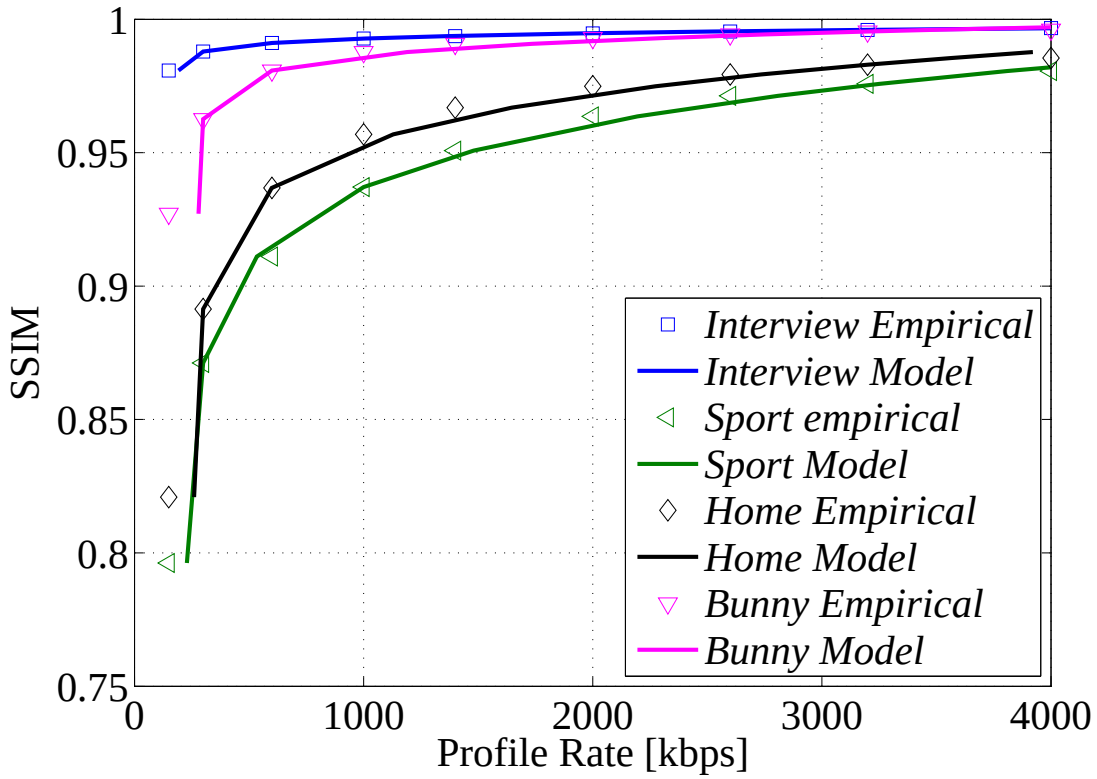


Figure 3.8: SSIM-to-rate model (straight line), according to eq. (3.30) fitting the empirical quality-to-rate relationship for one random chunk (duration equal to 2 sec.). The number of profiles is equal to 9 with rates ranging from 150 kbps to 4 Mbps.

where the parameters a_1, a_2, a_3 are as usual dependent on the spatial and temporal complexity of each chunk and are derived through curve-fitting over the actual discrete empirical points.

The validation results of the model (3.30) have shown almost perfect correlation with a Pearson coefficient always higher than 0.99 for each chunk of the considered video sequences. The parameters values of the SSIM-Rate model can be derived either off-line and on-line and inserted in each MPD as optional information. An example of the resulting empirical vs model relationship of one random chunk of the HAS sources considered in chapter 7 is provided in Fig. 3.8.

Chapter 4

Cross-layer Optimization for SVC Video Delivery in Shared Channel with Constant Bandwidth

In this chapter we analyze a simple scenario that can cover different video applications. The unique assumption is that the multimedia provider is able to perform off-line some computation-expensive processes, such as encoding and quality-computation for each video. In this framework, applications like video on-demand[21], IP-TV[22], sport broadcasting, where an initial transmission delay in the order of seconds can be tolerated by the end-users, as well as real-time streaming [23], are well suited to the low-complexity transmission scheme proposed. Each one of these applications requires a multimedia provider that has to serve several end-users which request different video sources.

Due to the different complexities of the scenes composing a video sequence, the relationships between the rate and the quality can be really different within a set of videos. However, the end-user expectation is to receive the best feasible quality independently of the particular video complexity even in presence of packet losses. If individual video streams are transmitted to different users in a broadcast dedicated channel, an equal rate allocation could lead to unacceptable distortion of high-complexity videos with respect to low-complexity ones. Adaptive transmission strategies have to be investigated to dynamically optimize the overall quality of experience (QoE). Therefore, quality fairness is an important issue that must be addressed. In this light, the adaptation module of the media provider is required to extract from the original video sequences a set of scaled

streams with a fair assignment of expected end-user quality, even in presence of packet losses.

The cross-layer approach considered here assumes that the lower-layers are able to allocated a shared constant bandwidth to a particular set of users, and inform the application layer about channel conditions, in terms of packet losses.

Many contributions exist in the literature that consider fairness-oriented rate adaptation in shared channel with constant bandwidth, but they exploit the Fine Granularity Scalability (FGS) tool, *e.g.*, [53]-[54]. Nevertheless, FGS mode has been removed from SVC, due to its complexity, and these works do not take into account the effects of transmission losses. Cross-layer optimization of video streaming over packet-erasure channel is also highly investigated, within the framework of SVC [34][16][42]. In [34] and in earlier works the authors proposed a complete framework to deliver SVC videos in bandwidth-limited scenario considering packet erasure channel, also in presence of play-out deadline. An UXP profiler, based on the same priority level assigner presented in section 3.4.1, solves a rate-minimizing cost functions. However, the rate adaptation aims at minimizing the distortion of each video without taking into account fairness issues.

We here propose a multi-stream rate adaptation framework with reference to SVC with medium grain scalability (MGS). Rate adaptation is carried out on the temporal and quality domain of the scalable video streams. Nevertheless, the entire framework can be extended to spatially scalable streams.

We first define a general discrete multi-objective problem with the aim to maximize the sum of assigned rates, while minimizing the differences among the expected distortions, under a total bit-rate constraint. A single-objective problem formulation is then derived by applying a continuous relaxation. It is based on the simplified continuous semi-analytical model 3.12 introduced in chapter 3.3, which allows us to derive an optimal and low-complexity procedure to solve the relaxed problem. The Unequal erasure protection (UXP) proposed in section 3.4.3 is also considered to suitably shape the rate-distortion relationship for different values of RTP packet-loss rate. The numerical results show the goodness of our framework in terms of error gap between the relaxed and its related discrete solution, and the significant performance improvement achieved with respect to an equal-rate adaptation scheme.

Contribution

In summary, this chapter collects the following relevant contributions:

- the formulation of a multi-stream rate-adaptation problem which considers minimization of both expected end-user distortion and distortion difference

among users, under bandwidth constraint

- the derivation of a optimal low-complexity algorithm for the solution of the multi-objective problem, based on continuous relaxation

- the derivation, analysis and discussion of simulation results which show the error gap of the low complexity solution and the improvements with respect to equal-rate allocation

4.1 System Architecture

In Figure 4.1 we show the architecture of the video delivery system. Each video sequence is encoded by the SVC encoder to fully support temporal and quality scalability. The resulting streams are encapsulated into Network Abstraction layer Units (NALUs), which are packets of an integer number of bytes, and stored in a media server. The NALUs have different importance according to a certain coding paradigm. To support the features of both Adaptation module and Unequal Erasure Protection (UXP) profiler, the video streams are also processed with the aim of extracting the information on the quality of each stream. After the encoder, the priority level assigner evaluates a priority index for each NALU, by considering the Rate-Distortion (R-D) relationship and the dependency on the other NALUs. Such information is encapsulated in the NALU header and then exploited by both the UXP profiler and the Adaptation module. These two processes are executed off-line.

As proposed in section 3.4.3, The UXP profiler aims at determining for each NALU the level of protection against transmission losses, which is obtained by adding parity bytes according to a specified UXP strategy. This task is executed by taking into account the estimated packet-loss rate of the lower layers which can be supplied at regular intervals. The protection profile is then sent to the Adaptation module which first estimates the expected R-D relationship, then extracts a suitable bit-stream from each video stream to meet fairness and bandwidth constraints. Each outcoming bit-stream is then encoded by the RS encoder. Finally, the resulting codewords are encapsulated in a transmission block and interleaved over RTP packets which are forwarded to the lower layers.

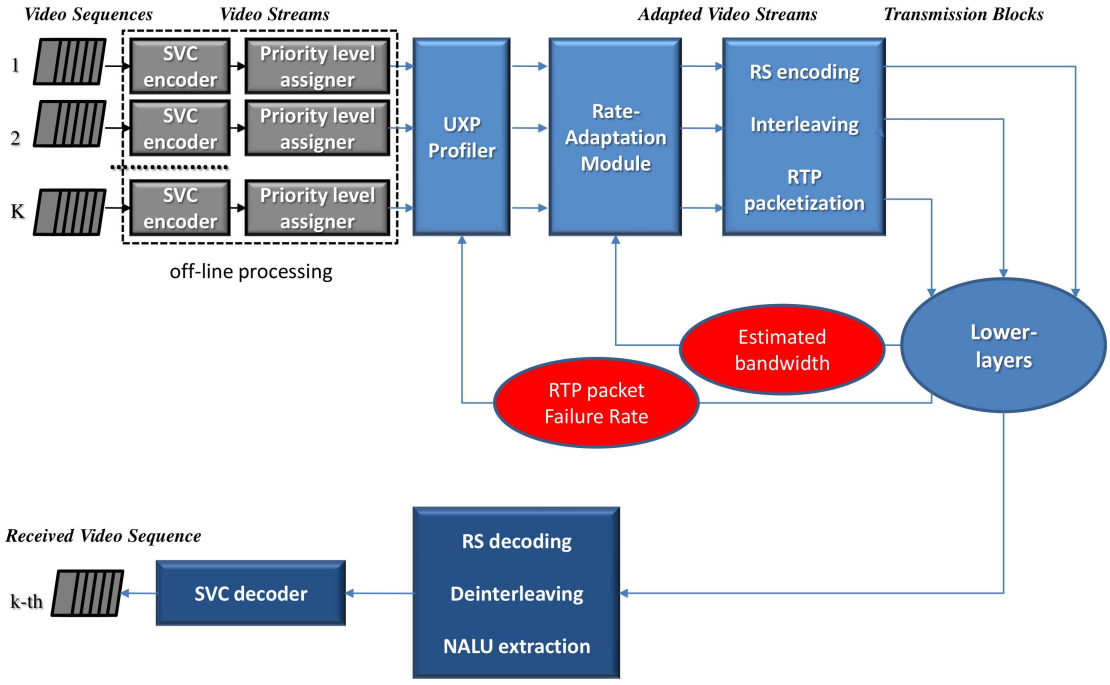


Figure 4.1: System architecture. Each sequence is encoded to fully support temporal and quality scalability and a priority level is assigned to the NALUs. The UXP profiler evaluates the overhead required according to a certain protection policy and RTP packet failure rate, and provides R-D information to the Adaptation module. The Adaptation module extracts sub-streams according to the estimated bandwidth and sends the data bytes to the RS encoder. The resulting codewords are then encapsulated in a transmission block, interleaved in RTP packets and forwarded to the lower layers. The receiver performs the inverse operations (RS decoding and de-interleaving) in order to extract the NALUs which are sent to the SVC decoder.

4.2 Problem Formulation for Multi-Stream Rate Adaptation

We first propose a general problem formulation, which can be suitable for different video coding schemes. At the end we restrict our attention to the proposed system architecture.

Let K be the number of streams involved in the optimization, indexed by the set $\mathcal{K} = \{1, \ldots, K\}$ and E_k the number of the available encoding schemes char-

acterized by different SNR resolution. Note that the cardinality E_k of the set $\mathcal{D}_k$ is generally not the same for each video source, and depends on the particular coding/extraction scheme applied. We recall the definition of set of expected distortion values for the k-th stream: $\mathcal{D}_k = \{d_{1,k}, ..., d_{E_k,k}\}$, $k \in \mathcal{K}$ where $d_{s,k}$ is evaluated according to eq. 3.13. The values in the set $\mathcal{D}_k$ take into account the distortion due to the lossy encoding techniques $d_{s,k}^{\text{enc}}$, and the expected distortion $d_{s,k}^{\text{enc}}$ due to the packet loss in the error-prone channel. The rate adaptation algorithm must choose at each time slot and according to the optimization strategy, the best vector $\boldsymbol{d} = [d_1, ..., d_K] \in \boldsymbol{\mathcal{D}} = \mathcal{D}_1 \times \ldots \times \mathcal{D}_K$. $\boldsymbol{\mathcal{D}}$ contains all the possible combinations of the elements of $\mathcal{D}_k$ and has cardinality $E = \prod_{k=1}^{K} E_k$. Optimization strategies for video rate adaptation has in general the aim to assign to each video the distortion that minimize the sum of the distortion, or equivalently that maximize the sum of the achievable PSNRs, under a total bit-rate constraints R_c [55]. However, the solution of such problem can usually lead to large distortion variations among different streams, due to the different complexity of video sources. As already mentioned, quality fairness is an important issue that must be addressed when multiple videos from different sources are transmitted in a shared channel.

The general objective of our proposed framework is to minimize the differences among the distortions assigned to each video stream while maximizing the sum of the rates until a maximum bit-rate is met. We then formulate the general problem as a multi-objective problem:

$$\min_{\boldsymbol{d}\in\boldsymbol{\mathcal{D}}} \sum_{i\in\mathcal{K}} \sum_{j\in\mathcal{K}, j<i} \Delta(d_i, d_j) \tag{4.1a}$$

$$\max_{\boldsymbol{d}\in\boldsymbol{\mathcal{D}}} \|\boldsymbol{F}\|_1 \tag{4.1b}$$

$$s.t.\ \|\boldsymbol{F}\|_1 \leq R_{\text{c}} \tag{4.1c}$$

where $\boldsymbol{F} = [F_1(d_1), \ldots, F_K(d_K)]$ is the vector of rates necessary to achieve the distortion $\boldsymbol{d}$ In case of video delivery over error-prone channel. As shown in section 3.4.3, the rate F_k depends on the rate of the encoder and on the overhead for error control required to obtain the expected distortion $d_{s,k}$.

The distortion-fairness metric in the objective (4.1a) is defined as:

$$\Delta(d_i, d_j) = \begin{cases} 0 & \text{if } d_i = D_i^{\min} \wedge d_j < d_i \\ 0 & \text{if } d_j = D_j^{\min} \wedge d_i < d_j \\ 0 & \text{if } d_i = D_i^{\max} \wedge d_j > d_i \\ 0 & \text{if } d_j = D_j^{\max} \wedge d_i > d_j \\ |d_i - d_j| & \text{otherwise.} \end{cases} \tag{4.2}$$

and $D_i^{\min}$ and $D_i^{\max}$ are the minimum and maximum distortion in the set $\mathcal{D}$. Such expression can be explained by the following considerations.

Ideal fairness among the distortion values assigned to the multiple video streams would require $d_i = d_j, \quad \forall i \neq j$. This is hard to be achieved due to (i) the discretization of the R-D relationship and (ii) the presence of a minimum and a maximum distortion values for each source, which are related to the encoding scheme and to the complexity of each video and can be very different. The definition of the fairness metric $\Delta(d_i, d_j)$ takes this fact into account, by introducing the effects of the minimum and maximum distortion constraints. In fact if d_i (or d_j) takes the maximum or minimum values and the difference $|d_i - d_j|$ can not be further decreased by moving some rate from video with small d to video with large d, then the fairness metric is set to 0.

It is worth noting that, by assuming a strict decreasing relationship between the rate and the distortion, this problem admits a feasible solution only if at least the minimum rates of all the of the video streams, *i.e.*, $\boldsymbol{F}^{\min} = [F_1^{\min}, \ldots, F_K^{\min}]$, with $F_k^{\min} = F_k(D_k^{\max})$ are supported by the transmission bandwidth R_{c}, *i.e.*,

$$\|\boldsymbol{F}^{\min}\|_1 \leq R_{\mathrm{c}} \tag{4.3}$$

otherwise a certain number of videos are not admitted in the transmission in order to keep this constraint satisfied.

The solution of the problem in (4.1) requires an exhaustive search in the space $\boldsymbol{\mathcal{D}}$ of all possible vectors. If E becomes large the required complexity can be not suitable for real-time adaptation. On the other hand if E is small, *i.e.*, there are few video sources as well as few related R-D points, the problem solution can lead to a waste of the available bandwidth and to large distortion differences among multiple videos. In the next section, we then propose a continuous relaxation of the problem, which implying a reasonable number of extractable sub-stream.

4.2.1 Continuous Relaxation

Considering all the discussions in the previous sections, we apply to the optimization problem a continuous relaxation based on the model (3.12). Therefore, we assume that the discrete variable d_k becomes continuous (with notation D_k), but limited by the minimum and maximum distortion values, *i.e.*,

$$D_k \in [D_k^{\min}, D_k^{\max}]. \tag{4.4}$$

With reference to the SNR scalability, the points $(D_k^{\max}, F_k^{\min})$ and $(D_k^{\min}, F_k^{\min})$ refer to the base layer and the highest enhancement layer streams, respectively.

It is worth noting that a trivial solution can be derived if the sum of the full quality encoded stream rates is less than or equal to the available bandwidth, that corresponds to transmitting the entire encoded streams without adaptation. Thus, we analyze the non-trivial case where the following constraint holds:

$$\|\boldsymbol{F}^{\max}\|_1 > R_c \tag{4.5}$$

According to the continuous relaxation (4.4) and the assumptions (4.3) and (4.5), a feasible solution is obtained when the constraint on the overall channel bandwidth is active with equality. A single-objective problem where the second objective, *i.e.*, (4.1b) in the problem formulation, is eliminated and replaced by an equality constraints can be then formulated. Nevertheless, as a result of the relaxation of the problem, the two constraints on the maximum and minimum available rates of each stream must be added. They imply that each video sequence has to obtain at least the base layer and not more than the maximum available bit-rate must be allocated to each video source to save bandwidth.

Thus, the relaxed problem can be formulated as

$$\min_{\boldsymbol{D}} \sum_i \sum_{j<i} \Delta(D_i, D_j) \tag{4.6a}$$

$$s.t.\ \|\boldsymbol{F}\|_1 = R_{\mathrm{c}} \tag{4.6b}$$

$$\boldsymbol{F} \succeq \boldsymbol{F}^{\min} \tag{4.6c}$$

$$\boldsymbol{F} \preceq \boldsymbol{F}^{\max} \tag{4.6d}$$

Note that, with a slight abuse of notation, the model $F_k(D_k)$ replaces the actual R-D relationship $F_k(d_k)$. In the next subsection we will derive an optimal procedure to solve this relaxed problem using methods that are computationally efficient and without the use of heuristics or brute-force search.

4.3 Adaptation Algorithms

A solution to the relaxed problem (4.6) can be derived by using sub-optimal procedures as the golden search algorithm proposed in [53] for a piecewise linear model. Nevertheless, the continuous formulation of model (3.12) allows us to derive a low-complexity optimal procedure, by noting that the solutions to the problem without the constraints (4.6c) and (4.6d) can be easily derived as follows:

$$D^* = D_k^* = \frac{\sum_{k\in\mathcal{K}} \alpha_k}{R_c - \sum_{k\in\mathcal{K}} \beta_k}, \quad \forall k. \tag{4.7}$$

Since those constraints imply that a minimum (maximum) or a maximum (minimum) rate (distortion) has to be allocated to each video stream, these solutions can be improved successively through a simple iterative procedure.

Let $\boldsymbol{x} = [x_1, \ldots, x_K]$, $\boldsymbol{y} = [y_1, \ldots, y_K]$, with $x_k, y_k \in \{0, 1\}$, $k \in \mathcal{K}$, be binary vectors that indicate whether (1) or not (0) the two constraints are active for the video stream k and these variables will be updated during the procedure. We can then define:

$$A(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k \in \mathcal{K}} x_k y_k \alpha_k \tag{4.8}$$

$$B(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k \in \mathcal{K}} x_k y_k \beta_k \tag{4.9}$$

$$\Omega(\boldsymbol{x}, \boldsymbol{y}) = R_{\mathrm{c}} - \left[\sum_{k \in \mathcal{K}} (1 - x_k) F_k^{\max} + \sum_{k \in \mathcal{K}} (1 - y_k) F_k^{\min} \right] \tag{4.10}$$

where $\Omega(\boldsymbol{x}, \boldsymbol{y})$ is the available rate for the videos which have not active constraints. The iterative procedure works as showed in Algorithm 1.

The algorithm requires in the worst case, a maximum of $K(K-1)/2$ iterations which happens in the unpractical case $\boldsymbol{F}^{\min} \simeq \boldsymbol{F}^{\max}$. At the first iteration, due to the initialization, D_k^* is computed as in (4.7). Then at each iteration the algorithm checks if the related rate solutions violate one of the constraints (4.6c), (4.6d). If it happens for one video, the algorithm assigns the relative minimum or maximum rate to this particular video and re-evaluates the distortion for the other video streams.

The optimality of the solutions (4.11) and (4.12) can be easily proved, by noting that the sum of the difference functions in (4.6a) is always kept to zero, *i.e.*, $\sum_i \sum_{j<i} \Delta(D_i^*, D_j^*) = 0$ and the sum of the rates is always equal to the available bandwidth. A rigorous proof is provided in section 5.5, lemma 2 for an extended version of the algorithm.

Algorithm 1 Pseudo code to solve problem (4.6)

if $\|\boldsymbol{F}^{\min}\|_1 > R_{\mathrm{c}}$ **then**
  report infeasibility
else if $\|\boldsymbol{F}^{\max}\|_1 \leq R_{\mathrm{c}}$ **then**
  report infeasibility and set $\tilde{F}_k = F_k^{\max}$, $\forall k \in \mathcal{K}$
else
  $y_k = 1$, $\forall\ k \in \mathcal{K}$;
  repeat
    $cond_{\mathrm{HL}} =$ false;
    $x_k = 1$, $\forall\ k \in \mathcal{K}$;
    repeat
      $cond_{\mathrm{BL}} =$ false;
      $\tilde{D} = \frac{A(\boldsymbol{x},\boldsymbol{y})}{\Omega(\boldsymbol{x},\boldsymbol{y}) - B(\boldsymbol{x},\boldsymbol{y})}$;
      for all $k \in \mathcal{K} : x_k y_k = 1$ **do**
        $\tilde{F}_k = \frac{\alpha_k}{\tilde{D}} + \beta_k$;
        if $\tilde{F}_k < F_k^{\min}$ **then**
          $\tilde{F}_k = F_k^{\min}$; $x_k = 0$; $cond_{\mathrm{BL}} =$ true;
        end if
      end for
    until $cond_{\mathrm{BL}}$ is false
    for all $k \in \mathcal{K} : x_k y_k = 1$ **do**
      if $\tilde{F}_k > F_k^{\max}$ **then**
        $\tilde{F}_k = F_k^{\max}$; $y_k = 0$; $cond_{\mathrm{HL}} =$ true;
      end if
    end for
  until $cond_{\mathrm{HL}}$ is false
end if

The final relaxed solutions, given $\boldsymbol{x}, \boldsymbol{y}$, are then given by:

$$F_k^* = \begin{cases} \frac{\alpha_k}{D_k^*} + \beta_k & \text{if } x_k y_k = 1 \\ F_k^{\min} & \text{if } x_k = 0 \\ F_k^{\max} & \text{if } y_k = 0 \end{cases} \tag{4.11}$$

with

$$D_k^* = \begin{cases} \frac{A(\boldsymbol{x},\boldsymbol{y})}{\Omega(\boldsymbol{x},\boldsymbol{y}) - B(\boldsymbol{x},\boldsymbol{y})} & \text{if } x_k y_k = 1 \\ D_k^{\max} & \text{if } x_k = 0 \\ D_k^{\min} & \text{if } y_k = 0 \end{cases} \tag{4.12}$$

From a mathematical perspective the optimal discrete solution $\boldsymbol{d}^*$, starting from the relaxed one $\boldsymbol{D}^*$, should be derived by applying optimization techniques, *e.g.*, branch & bound search. Nevertheless, such techniques will increase the

complexity. To keep the complexity low, it is common practice to extract the higher discrete bit-rate under the optimal relaxed solution, by paying a minimum waste of bandwidth due to the granularity of the empirical R-D relationship. When the packet loss is taken into account, *i.e.*, when the probability of losing RTP packets is such that $P_{\mathrm{e}}^{\mathrm{RTP}} > 0$, the solutions (4.11) are referred to the rate values which include the overhead. In order to perform the desired bit-stream extraction the information overhead is fed into the the adaptation module, thereby allowing the evaluation of the related encoder rate solution, whose distortion is denoted as D^*_{enc}.

4.4 Numerical Results

In this section we evaluate the performance of the proposed rate adaptation framework by using the JSVM reference software [30] and a C++ ad-hoc simulator. We encode five video sequences with different scene complexity, *i.e.*, *Coastguard*, *Crew*, *Football*, *Foreman*, *Harbour* in CIF resolution with a frame-rate of 30 fps (see Table 3.1 for further details). Each sequence is coded GOP-by-GOP and we analyze the performance with two different GOP sizes, *i.e.*, $G = 8$ and $G = 16$. In both cases the coding structure is based on the maximum coding efficiency that allows to decode GOPs independently, *i.e.*, IDR-period is equal to the GOP size, and to insert the maximum number of temporal resolutions, *i.e.*, $T = \log_2(G)$. Thus, in the former case we suppose an $IBBBPBPP$ encoding structure as depicted in Fig. 2.1, with 4 temporal layers, while in the latter case the encoding sequence is $IBBBBBBBPBBBPBPP$ with 5 temporal layers. The SNR-scalability is obtained through 2 enhancement layers, each one split in 5 MGS layers with vector distribution [3 2 4 2 5] resulting in a maximum of $Q = 10$ quality layers. The Quantization Parameter (QP) of the base and enhancement layers are equally spaced and set to 38, 32 and 26, respectively. The post-processing priority level assignment is then applied, as described in section 2.1.3, which provides the priority level information as well as the distortion increment of each layer.

We compare the solution of the proposed algorithm (OPT) with an equal-rate (ER) scheme where no quality-based adaptation is performed, *i.e.*, the same portion of the available bandwidth is assigned to each video.

To have a fair comparison we apply to ER scheme the constraints (4.6c) and (4.6d) in order to guarantee the resource to the base-layer of each video and to fulfill the available bandwidth. Therefore, after sorting the streams in two vectors, one into decreasing order with respect to base-layer bit-rate and the other into increasing order with respect to highest layer bit-rate, respectively, we iteratively

check if the bit-rate $R_k = R_c/K$ required by each ordered stream violates one of those constraints. If it happens, we assign the corresponding bit-rate and equally re-distribute the remaining bandwidth to the other streams.

The fairness is computed according to three different metrics: the average MSE difference

$$\delta_{\text{av}} = \frac{1}{L} \sum_i \sum_{j<i} |D_i^* - D_j^*| \quad , \tag{4.13}$$

where the average is computed over $L = K(K-1)/2$ possible MSE difference terms, the modified average modified difference

$$\Delta_{\text{av}} = \frac{1}{L} \sum_i \sum_{j<i} \Delta(D_i^*, D_j^*) \tag{4.14}$$

and the most used MSE variance for each GOP.

We first analyze the performance of the adaptation algorithm by assuming error-free channel, *i.e.*, $P_{\text{e}}^{\text{RTP}} = 0$, and GOP size equal to 8.

4.4.1 Error-free channel

In Table 4.1, we show the improvements of our proposed scheme with respect to ER when the available bandwidth is fixed to R_c= 3000 kbps. The average modified MSE difference is significantly reduced and equivalently the variance is decreased up to ten times. Let us note that Δ_{av} also gives us the information on the error generated when the discrete solution replaces the continuous solution in the relaxed problem, (where Δ_{av} is zero). This error includes two contributions: the estimation error of the model and the integrality gap. As expected, the average error is not small due to mainly the granularity of the low-rate points. Moreover, in this particular case of bandwidth, the MSE difference (variance) is still quite high, due to the minimum rate constraints. Our algorithm, while providing fairness, is able to improve the performance of the most demanding videos, by allocating more bits to sequences with more complex scenes. This is more clear in Figure 4.2 where we plot the rate assigned by our adaptation algorithm to each video sequence GOP-by-GOP. More bit-rate is assigned to *Coastguard*, *Football* and *Harbour* video sequences, allowing them to achieve more quality. In Figure 4.3, the MSE variance averaged over 30 GOPs is evaluated for different bandwidths. In the bandwidth interval considered, the assumptions (4.3) and (4.5) hold for each GOP. When the bandwidth is very low both schemes show high MSE variance, because the optimization range is limited by the minimum rate constraints. When the bandwidth increases, our procedure improves the fairness leading the variance close to 0. A slight variance increase occurs at large

bandwidths when the maximum rate constraints limit the achievable distortion. On the other hand, the ER scheme generally increases the MSE variance until the base-layer constraints are active for most of the streams. This behavior can be partially improved by controlling the base-layer bit-rate [56] of each video according to its complexity, as performed for instance in [53].

GOP index	Δ_{av}		δ_{av}		Variance	
	ER	OPT	ER	OPT	ER	OPT
1	34.54	2.04	35.07	23.40	719.4	216.1
2	35.08	2.36	35.09	25.38	715.0	262.2
3	34.27	1.45	34.79	23.56	772.4	217.6
4	33.13	0.29	37.63	19.50	780.5	227.0
5	29.62	0.26	35.16	21.95	652.0	258.2
6	33.67	0.55	37.99	23.36	774.9	281.8
7	26.88	0.31	31.78	17.63	551.3	170.8
8	30.07	1.28	34.76	25.58	636.0	241.6
9	25.57	0.38	31.18	15.58	493.3	139.8
10	29.46	1.14	40.94	17.75	902.9	164.2
11	38.84	0.20	38.84	18.34	810.6	185.8
12	34.68	0.25	34.68	14.43	666.7	111.6
13	39.09	0.43	39.09	20.33	811.4	223.3
14	32.80	0.19	38.25	16.92	741.0	172.6
15	36.21	0.05	36.21	15.17	680.5	85.4
Av.	32.92	0.74	36.09	19.92	713.9	197.2

Table 4.1: Average modified MSE difference Δ_{av}, average MSE difference δ_{av} and MSE variance in each GOP interval. Comparison between the proposed algorithm (OPT) and equal-rate (ER) assignment with bandwidth equal to 3000 kbps.

4.4.2 Packet-erasure channel

In this subsection we assess the performance in the case of transmission over packet-erasure channel, by evaluating only the proposed algorithm with two different GOP sizes. The number of bytes per RS codeword is set equal to $n = 128$ (as a shortened version of the code with natural length 255) by allowing the insertion of more than one GOP into a TB and then filling the payload of each RTP packet with a reasonable number of bytes. In order to limit the overhead to about 20% for the worst case considered, *i.e.*, $P_{\text{e}}^{\text{RTP}} = 0.1$, the parameter α of the proposed UXP scheme is set equal to 30 (see Table 3.3). According to

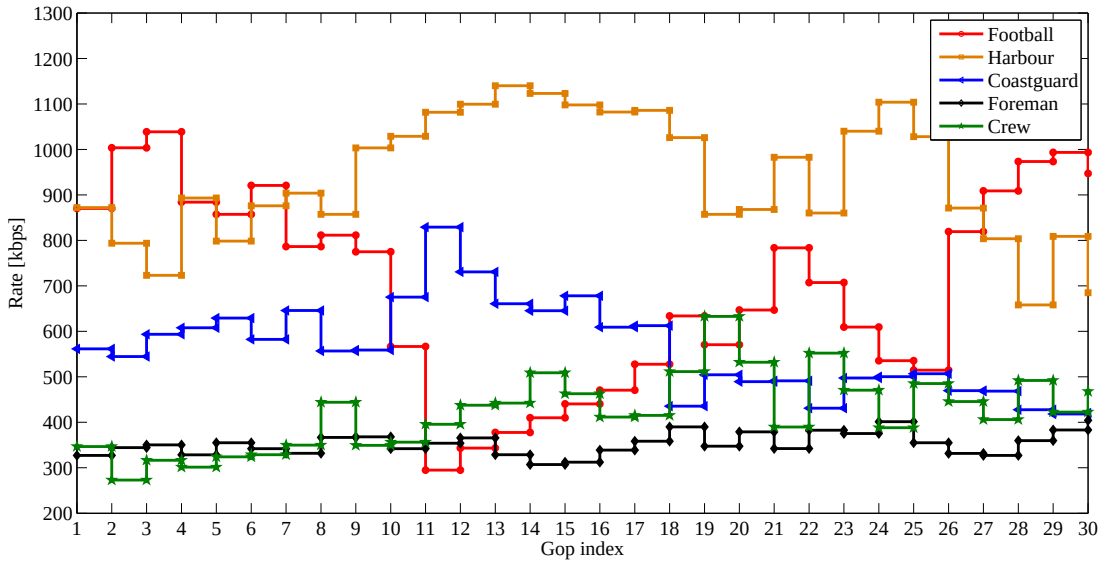


Figure 4.2: Rate assigned GOP-by-GOP by our adaptation algorithm (GOP size G equal to 8), when bandwidth is equal to 3000 kbps.

extensive simulations we define the range of the EPP values for the enhancement temporal layers between 10^{-6}, which is intended to the I-frame, and $10^{-(6-T)}$. We also consider a value of bandwidth sufficiently high, *i.e.*, R_c= 7000 kbps, to allow the insertion of the higher quality layers which have less protection.

Table 4.2 shows the average distortion resulting at different $P_{\mathrm{e}}^{\mathrm{RTP}}$ for the different video sequences. The average is obtained by looping the first 240 frames of each sequences for 1000 times. Here, $D_{\mathrm{av}}^{*\mathrm{rec}}$ is the average received MSE; D_{av}^{*} is the average expected distortion which is the discrete solution of the adaptation algorithm, and $D_{\mathrm{av}}^{*\mathrm{enc}}$ is its related encoding distortion. We can note that the expected distortions as well as the received distortions at the same RTP packet failure rate $P_{\mathrm{e}}^{\mathrm{RTP}}$ are approximately equal, showing the goodness of the framework even in presence of packet erasures. The distortion values decrease for most of the video sequences, while the packet error rate increases, due to the effect of bandwidth constraint. At large values of $P_{\mathrm{e}}^{\mathrm{RTP}}$ the outcoming overhead from the UXP profiler increases and the Adaptation module reacts by reshaping the rate of each sequence, thereby increasing the distortion to provide fairness. This behavior is less marked in the case of GOP size equal to 8 for the *Foreman* sequence whose distortion does not change significantly, since it receives in most cases only the base-layer with the highest protection. The slight increase of distortion with

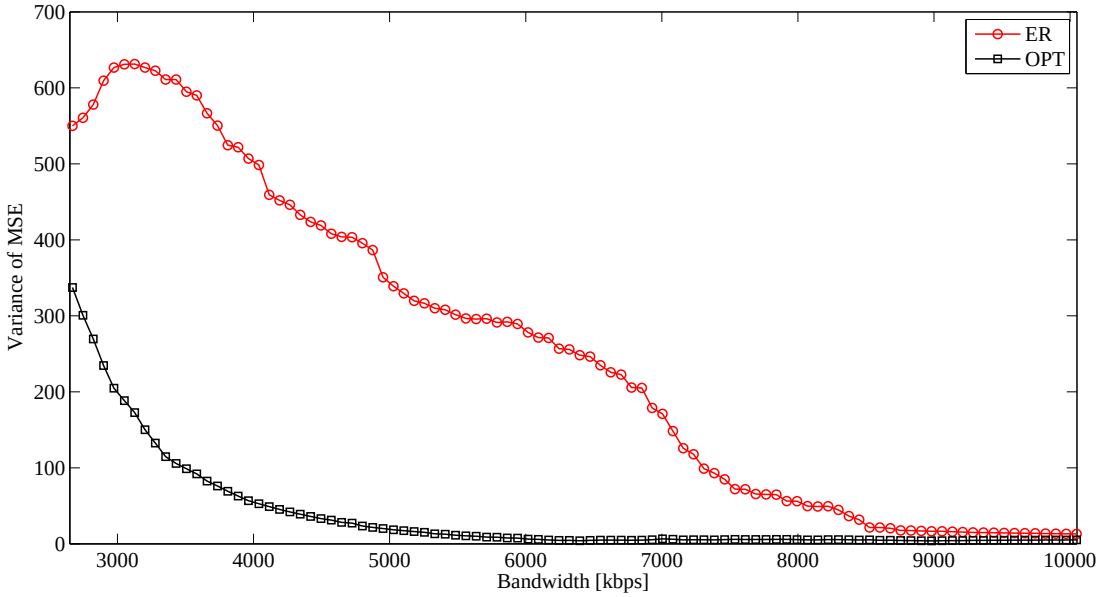


Figure 4.3: Variance of the MSE averaged over 30 GOPs, with different bandwidth values. Comparison between the proposed algorithm (OPT) and equal-rate (ER) assignment.

respect to the encoding MSE is due to the loss of certain enhancement temporal layers.

As expected, an higher GOP size decreases the distortion thanks to the higher coding efficiency, which allows to improve the R-D performance of the base layer. Nevertheless, such gain is reduced with respect to the case of error-free channel, since more quality layers with low protection are transmitted. This behavior can be improved with a more careful design of the EPP aimed at balancing overhead and degree of protection according to the available bandwidth.

Video	P_e^{RTP}	$G = 8$			$G = 16$		
		D_{av}^{*rec}	D_{av}^{*}	D_{av}^{*enc}	D_{av}^{*rec}	D_{av}^{*}	D_{av}^{*enc}
Coastguard	0.01	33.9	37.4	29.6	27.3	29.4	19.8
	0.05	37.5	40.1	33.6	31.2	32.0	22.4
	0.1	40.8	42.3	37.8	36.1	37.7	27.0
Crew	0.01	36.5	36.6	36.2	28.4	28.4	28.2
	0.05	39.3	39.4	39.1	32.4	32.5	32.3
	0.1	41.4	41.5	41.3	36.6	37.0	36.0
Football	0.01	35.2	35.6	34.0	27.9	28.4	26.4
	0.05	38.4	38.9	37.1	30.8	31.6	29.2
	0.1	41.8	41.8	40.5	35.9	37.3	34.3
Foreman	0.01	35.7	35.6	34.2	28.1	28.7	27.9
	0.05	35.9	36.0	35.4	30.4	30.8	30.1
	0.1	36.2	37.1	36.1	33.8	34.9	33.2
Harbour	0.01	35.3	38.8	23.7	29.8	30.3	18.2
	0.05	40.6	42.2	26.5	32.0	32.3	20.3
	0.1	42.8	44.2	31.0	34.4	37.8	22.9

Table 4.2: Average received distortion, D_{av}^{*rec}, expected distortion, D_{av}^{*}, and encoding distortion, D_{av}^{*enc}, in term of the MSE for different video sequences, GOP size G, and packet-erasure rate values P_e^{RTP}, resulting from the proposed rate-adaptation algorithm. Available bandwidth is R_c =7000 kbps.

Chapter 5

Cross-layer Optimization for SVC Video Delivery in Downlink OFDMA Channels

In beyond-3G and 4G wireless system orthogonal frequency division multiple access (OFDMA) has been selected as a key physical (PHY) layer technology to support a very flexible access with high spectral efficiency. In OFDMA wireless systems, the channel capacity of each user depends on how the channel is shared by the multiple users and on the fading correlation properties, which are not static in both time and frequency domains. In order to exploit the available temporal, frequency and multi-user diversity, and to provide a given level of QoS, suitable adaptive resource allocation and scheduling strategies have to be implemented. Opportunistic schedulers, as for instance, Proportional Fair (PF) [19] and maximum signal-to-noise ratio (SNR) schedulers, take advantage of the knowledge of the channel state information (CSI) in order to maximize the spectral efficiency. However, with these schedulers, the final share of throughput often results unfair, especially for the cell-edge users which suffer of data-rate limitations due to high path-loss and inter-cell interference.

In real-time streaming the mismatch between the allocated PHY layer rate and the rate required by the delay-constrained application may cause the loss of important parts of the streams, which significantly degrades the end-user quality of experience (QoE). The provision of acceptable QoE to every user is enabled by the use of a scheduler at the medium access control (MAC) layer which delivers a fair throughput, according to specific utilities and constraints defined by the application [20]. Moreover, the presence of an optimized source rate adaptation technique at the application (APP) layer becomes crucial to improve stability, to

prevent buffer overflow and to maintain video play-back continuity. As already mentioned, source rate adaptation is enabled by the use of video encoders, *e.g.*, SVC, that support multiple layers which can be sequentially dropped, thereby providing a graceful degradation.

In this chapter we extend the framework proposed in Chapter 4, where we have assumed constant bandwidth, to the more general case of OFDMA wireless scenario where the user capacity and the total bandwidth vary on the time and strictly depends on how the resources are allocated to each user. As in chapter 4, we target the delivery of quality-fair SVC video streams.

In the literature, several researchers proposed a cross-layer approach for the optimization of multi-user wireless communications systems.

The Authors in [57] proposed a cross-layer approach for the delivery of one scalable video in a TDMA-based wireless local-area network under a predefined service time constraint. It is based on an unequal error protection scheme which jointly selects the different rates for each scalable video layer and the amount of enhancement layers permitted in order to maximize the PSNR of the delivered video. They showed that such intelligent link adaptation scheme significantly improves the end-video quality with respect to conventional layer drop solutions.

In [58] the framework has been extended to also consider traffic control for a multi-user scalable video delivery. The optimization framework specifies for each video the PHY layer rate of each layer and the amount of the packets that should be dropped from each video.

Both frameworks assume quasi-static fading channel in the time scale of one group of pictures where the rate can be predicted with enough accuracy. However, these assumptions can not be applied to realistic OFDMA wireless systems where the channel capacity depends on how the channel is shared by the multiple users and fading is not static in both time and frequency domains. Moreover, temporal fairness constraints simplify the resource allocation in TDMA-based scenarios, but they are not able to capture the frequency and multi-user diversity of the OFDMA systems. In this paper we specifically address optimal resource allocation for multiple users in OFDMA scenario where fading is variable in both time and frequency domains.

In [12] the Authors presented a cross-layer method to solve the problem of multiuser SVC streaming over ODFMA networks. The framework is based on a gradient scheduling algorithm where user-priority weights are derived heuristically according to video contents, deadline requirements, and previous transmission results. However, differently from our work, optimized source adaptation is not addressed, leading to the loss of important parts of the streams, in case of scarce resources.

The work in [13] addressed the maximization of the weighted sum of the average PSNRs achieved by a set of users sharing a wireless channels, but without addressing fairness and OFDMA systems, as in our framework. As already mentioned, the solution of such problem can usually lead to large quality variations among different streams.

The Authors in [14] proposed a fairness-oriented coo-petition strategy for multi-user multimedia radio resource allocation (RRA) under the assumption of a general PHY layer setup with convex rate region. The problem is solved by using the layering as optimization decomposition (LOD) method, which enables a simple implementation in a layered transmission system. It is shown that it improves the number of satisfied users by providing a video quality proportionally fair to the user channel condition, but requires a careful adaptive selection of the minimum PSNR thresholds for each user according to system throughput, which is left in future works. As in [14], we propose a decomposition method for the optimization problem resulting in algorithmic solutions that handle parameter and constraints of a single layer, but differently from [14], our framework provides video quality fairness and does not depend on specific thresholds selection.

To the best of our knowledge only the work in [7] addressed the issue of transmitting quality-fair SVC streams by jointly optimizing APP and MAC layers in OFDMA downlink. The fairness problem is handled by minimizing the maximal end-to-end distortion among all users at each transmission time interval (TTI), under rate constraints. Due to the NP-hard nature of the problem, the Authors proposed a suboptimal algorithmic solution. However, the TTI-based optimization does not allow to fully capture the time diversity of the channel and requires extensive exchange of information between MAC and APP layers.

In our work we show that an ergodic-based optimization problem can be optimally solved resulting in a limited scalar information exchange among the involved layers. In fact, in practical applications, the definition of utilities and constraints should be function of the rate averaged over a certain time period [59], *e.g.*, an interval related to the structure of the encoded video streams. When the objective of the optimization is to maximize of the sum of concave utility functions of the ergodic rates, the optimal solution for the downlink of an OFDMA system can be derived through dual decomposition, which results in MAC layer scheduling algorithms with decoupled subcarrier and power allocations. Similar frameworks were proposed in [60] and [61], which proved that quasi-optimal solutions have linear complexity with respect to the number of both subchannels and users. The main drawback of such solutions is that the MAC layer has to directly manipulate the utility functions of the APP layer, thus limiting the applicability to layered transmission systems where only limited scalar information can be exchanged between

APP and MAC layers.

We here propose a cross-layer method for maximizing the aggregate ergodic (average) rate assigned to multiple SVC transmission in an OFDMA wireless network, while minimizing the distortion difference among the received video sequences. The optimization problem is "vertically" decomposed into two sub-problems, leading to rate adaptation at the APP layer and resource allocation at the MAC layer, and a novel efficient iterative local approximation (ILA) algorithm is proposed to obtain the global solution. The ILA algorithm is based on the local approximation of the contour of the ergodic rate region of the OFDMA downlink channel and requires a limited information exchange between the APP and the MAC layers. Moreover, we present and discuss the algorithms to solve the two sub-problems and prove the optimality and convergence of the ILA algorithm.

It should be pointed out that a similar approach has been developed in [62] to solve the maximization of a general concave utility function. In such approach, the APP layer derives iterative solutions on the space tangent to the rate region. But differently from our approach, a gradient-based update of the utility function is proposed, hence requiring a careful selection of the related step-size to ensure convergence. In our work, since the utility is replaced by a one-dimensional manifold representing the fairness constraints, such issue is overcome. Moreover, the Authors in [62] proposed to project the APP solutions on the contour of the rate region, orthogonally to the tangent space. Differently, our approach projects the APP solution by using a parametric line representing a proportionality constraint.

In this chapter we also address some issues arising in practical implementations, by designing a suboptimal solution based on the outcome of a single step of the ILA algorithm and on the use of stochastic algorithms for resource allocation. Our numerical evaluations show (i) the fast convergence of the ILA algorithm, (ii) the resulting low gaps in terms of efficiency and fairness between optimal and suboptimal proposed strategies, and (iii) the significant video quality improvements with respect to other state-of-the-art solutions.

The remainder of this chapter is organized as follows. Section 5.1 introduces the system architecture, The PHY layer model is presented in 5.2. In Section 5.3 the optimization problem and its "vertical" decomposition are formulated and discussed, whereas the ILA algorithm is proposed in Section 5.4. The solutions of the APP and MAC sub-problems are provided in Section 5.5 and 5.6, respectively, whereas in Section 5.7 optimal and suboptimal solutions suited for realistic implementation are discussed. The performance of the proposed schemes is finally evaluated in Section 5.8.

Contribution

In summary, this chapter collects the following relevant contributions:

- we formulate the cross-layer optimization problem for maximizing the aggregate ergodic (average) rate assigned to multiple SVC transmission in an OFDMA wireless network, while minimizing the distortion difference among the received video sequences. We prove that the global optimal solution is unique.
- we decompose the cross-layer optimization problem into two sub-problems that handle parameter and constraints of a single layer. They results in rate adaptation at the APP layer and resource allocation at the MAC layer,
- we propose a novel efficient Iterative Local Approximation (ILA) algorithm to obtain the global solution and we rigorously prove its convergence and optimality.
- We propose optimal algorithm for the solution of the APP layer sub-problem which has linear complexity in the number of users for practical scenario
- We re-trace the optimal algorithmic solution for the resource allocation sub-problem at the MAC layer, by also analyzing its impact to the computational complexity of the ILA algorithm.
- we propose and design suboptimal solution for practical implementations based on the use of stochastic algorithms for resource allocation.
- in order to further reduce the complexity, we propose the 1-step ILA algorithm, which is based on the outcome of a single step of the ILA algorithm. Due to its sub-optimality, we also investigate methods to adaptively compensate its residual error.
- we finally provide extensive numerical evaluation by comparing optimal and suboptimal proposed solution with respect to other state-of-the-art frameworks.

5.1 System Architecture

In Figure 5.1 we show the architecture of a video delivery system, where the three key elements taken into account in this work are outlined, *i.e.*, the multimedia

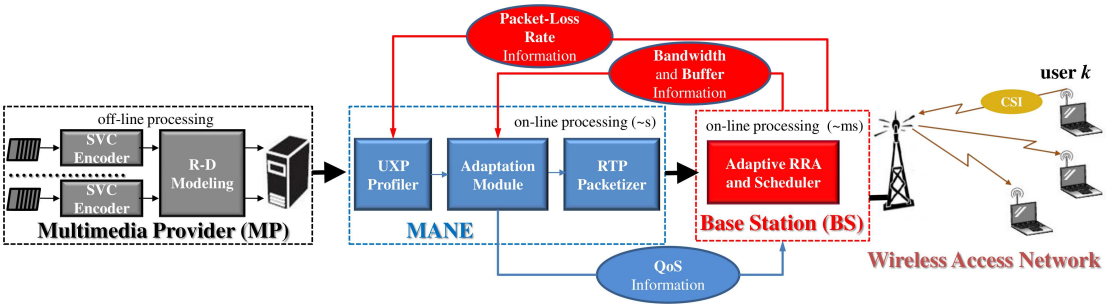


Figure 5.1: System architecture

provider (MP), the media aware network element (MANE) and the OFDMA-based wireless access network (WAN) which includes a base station (BS) that serves K users indexed by the set $\mathcal{K} = \{1, \ldots, K\}$. Each mobile user in the WAN requests a video sequence and the MP encodes the requested video to fully support temporal and quality scalability. The video streams are further processed to extract a priority index for each frame [28] and the R-D information for each layer. The R-D modeling block collects the R-D information and evaluates the set of parameters describing the R-D relationship, according to the parametric model introduced in the section 3.4. Specifically, we consider the three parameter model in eq. (3.9). Priority indexes and R-D parameters are then sent as side information to the MANE.

The UXP (unequal erasure protection) profiler assigns a different protection to each frame according to its dependencies and the related R-D improvements. This task is executed by taking into account the estimated average packet-loss rate at the lower layers of the systems. According to the scheme described in section 3.4.1, the profiler also computes the rate and the expected distortion after error protection, which will be used to update the parameters of the expected R-D characteristics. The resulting information is then sent to the adaptation module which extracts a suitable bit-stream from each encoded video stream, according to the outcome of the adaptation algorithm. The parameters of the R-D relationships available at the MANE, as well as bandwidth and buffer information provided by the BS, are the input of the adaptation algorithm.

The packetization process is carried out according to the guidelines presented in section 3.4.1. Each GOP of the adapted video stream is mapped into one transmission block (TB) that carries both data and UXP parity bytes. After interleaving and packetization, the TB is then re-organized into a sequence of RTP packets which are finally forwarded to the MAC/PHY layers through a

suitable protocol stack, *e.g.*, with UDP/IP/link layers. It is worth noting that, due to the GOP interleaving over RTP packets, the receiver is not able to decode any frame until the entire TB is received.

The MANE procedures are executed at regular time intervals (in the order of seconds), here named as *application frame intervals*. During each *application frame interval* the MANE and the BS of the WAN exchange limited information related to QoS constraints, PHY layer bandwidth constraints, buffer status and packet-loss rate, according to a cross-layer paradigm. We assume here that this information exchange introduces a negligible delay, thanks to the high-speed connection in the fixed network. Radio resource allocation (RRA) and scheduling at the BS are based on adaptive algorithms, which aim to maximize the spectral efficiency of the OFDMA network, using QoS constraints provided by the MANE and CSI information from the PHY layer.

5.2 Physical Layer Model for the Downlink of the Wireless Access Network

In this chapter we consider a single-cell time-slotted OFDMA system where the BS and users are equipped with one antenna. Methods and algorithms developed here are also extended to multi-cell scenario in Appendix A and they can be easily extended to multi-antenna configurations [63].

The total available bandwidth B is divided into S orthogonal subcarriers indexed by the set $\mathcal{S} = \{1, \ldots, S\}$, with subcarrier spacing $\Delta B = B/S$. The channel gain $h_{k,s}[n]$ between the BS and user k, on subcarrier s and time slot n, is modeled as a complex Gaussian random process (Rayleigh fading), in general correlated across subcarriers and time slots. We define the normalized SNR of user k, on subcarrier s and time slot n, as

$$\gamma_{k,s}[n] = \frac{|h_{k,s}[n]|^2}{\sigma^2} \tag{5.1}$$

where σ^2 is the noise power.

The RRA at the BS aims to allocate the available resources, *i.e.*, subcarriers and power, at each time slot, to the users according to a predefined allocation strategy. We first assume that subcarriers can be shared by multiple users over non-overlapping fractions of the total time slot duration t_{slot}. We denote with $\psi_{k,s}[n] \in [0, 1]$ and $p_{k,s}[n] > 0$ the fraction of time slot and the power, respectively, allocated to user k, on subcarrier s and time slot n. By using a suitable adaptive

modulation and coding (AMC) scheme, the rate achieved by user k on subcarrier s can be evaluated with the following model:

$$r_{k,s}(\psi_{k,s}[n], p_{k,s}[n]) = \Delta B \psi_{k,s}[n] C\left(\frac{\gamma_{k,s}[n] p_{k,s}[n]}{\psi_{k,s}[n]}\right) \tag{5.2}$$

if $\psi_{k,s}[n] > 0$ and $r_{k,s}(\psi_{k,s}[n], p_{k,s}[n]) = 0$ otherwise, where $C(x) = a_1 \log_2(1 + x/a_2)$ and a_1, a_2 are two parameters, namely the rate adjustment and the SNR-gap, respectively, depending on the specific AMC scheme adopted [64]. To summarize, given the set $\boldsymbol{\gamma} = \{\gamma_{k,s}, k \in \mathcal{K}, s \in \mathcal{S}\}$ of the SK realizations of the SNR random process, the RRA algorithm at the BS determines the set of allocation variables $\boldsymbol{\psi} = \{\psi_{k,s}, k \in \mathcal{K}, s \in \mathcal{S}\}$ and $\boldsymbol{p} = \{p_{k,s}, k \in \mathcal{K}, s \in \mathcal{S}\}$ functions of the SNR realizations $\boldsymbol{\gamma}$, *i.e.*, $\boldsymbol{p}(\boldsymbol{\gamma}), \boldsymbol{\psi}(\boldsymbol{\gamma})$. Although we use for the sake of clarity a simplified notation, it should be noted that $\boldsymbol{\gamma}$, $\boldsymbol{p}$ and $\boldsymbol{\psi}$ are sets of random processes along time dimension n.

Finally, we assume that the *application frame interval* t_{I} is sufficiently large to support ergodic approximation for the average rate provided to users. Specifically, we assume that the rate assigned to user k averaged over the discrete time window $W_{\mathrm{I}} = \left\lfloor \frac{t_{\mathrm{I}}}{t_{\mathrm{slot}}} \right\rfloor \gg 1$ can be approximated by its expected value with respect to the random process $\boldsymbol{\gamma}$, *i.e.*, the ergodic rate:

$$\begin{aligned} R_k(\boldsymbol{\psi}, \boldsymbol{p}) = & \frac{1}{W_{\mathrm{I}}} \sum_{n=1}^{W_{\mathrm{I}}} \left[\sum_{s \in \mathcal{S}} r_{k,s}(\psi_{k,s}[n], p_{k,s}[n]) \right] \cong \\ & \mathbb{E}_{\boldsymbol{\gamma}} \left[\sum_{s \in \mathcal{S}} r_{k,s}(\psi_{k,s}(\boldsymbol{\gamma}), p_{k,s}(\boldsymbol{\gamma})) \right]. \end{aligned} \tag{5.3}$$

According to the proposed source rate-distortion model in chapter 4, The average PHY rate can be mapped to the average rate required by the source with the relationship

$$R_k(\boldsymbol{\psi}, \boldsymbol{p}) = H F_k(D) \tag{5.4}$$

where $H \geq 1$ is a constant that takes into account the overhead introduced by the different layers of the network architecture. Therefore, the continuous distortion D_k of the set of pictures delivered to user k has an implicit dependence on the allocation variables $\boldsymbol{\psi}, \boldsymbol{p}$, *i.e.*,

$$D_k = F_k^{-1}\big(R_k(\boldsymbol{\psi}, \boldsymbol{p})/H\big). \tag{5.5}$$

5.3 The Optimization Problem

Similarly to the framework proposed in chapter 4, our objective is to provide a fair video quality by maximizing the overall video quality while minimizing the quality difference among the different videos, under the minimum and maximum rate constraints.

Let us denote with $\mathcal{A}$ the set of feasible allocation policies $\boldsymbol{\psi}(\boldsymbol{\gamma}), \boldsymbol{p}(\boldsymbol{\gamma})$, *i.e.*,

$$\mathcal{A} = \left\{(\boldsymbol{\psi}, \boldsymbol{p}) : \psi_{k,s}(\boldsymbol{\gamma}) \geq 0, p_{k,s}(\boldsymbol{\gamma}) \geq 0, \sum_{k \in \mathcal{K}} \psi_{k,s}(\boldsymbol{\gamma}) \leq 1\right\} \tag{5.6}$$

and with $\mathcal{P}$ the set of the feasible allocation policies $(\boldsymbol{\psi}, \boldsymbol{p}) \in \mathcal{A}$ which also satisfy an average sum-power constraints, *i.e.*,

$$\sum_{k \in \mathcal{K}} \sum_{s \in \mathcal{S}} \mathbb{E}_{\boldsymbol{\gamma}}[p_{k,s}(\boldsymbol{\gamma})] \leq P \tag{5.7}$$

where P is the *average* power budget of the OFDMA transmitter. The achievable ergodic rate region is then given by

$$\mathcal{R} = \bigcup_{(\boldsymbol{\psi}, \boldsymbol{p}) \in \mathcal{P}} \left\{\boldsymbol{\varrho} : \boldsymbol{\varrho} \preceq \boldsymbol{R}(\boldsymbol{\psi}, \boldsymbol{p})\right\} \tag{5.8}$$

where $\boldsymbol{R}(\boldsymbol{\psi}, \boldsymbol{p}) = [R_1(\boldsymbol{\psi}, \boldsymbol{p}), \ldots, R_K(\boldsymbol{\psi}, \boldsymbol{p})]^{\mathrm{T}}$ is the ergodic rate vector and $\boldsymbol{\varrho} = [\varrho_1, \ldots, \varrho_K]^{\mathrm{T}}$. When subcarrier sharing is considered, as here, the rate region results in a convex set of the rate vectors [61].

The optimization problem can be then described by the following constrained sum-rate maximization:

$$\max_{(\boldsymbol{\psi}, \boldsymbol{p}) \in \mathcal{A}} \|\boldsymbol{R}(\boldsymbol{\psi}, \boldsymbol{p})\|_1 \tag{5.9a}$$

$$s.t.\ \Delta(D_i, D_j) = 0 \quad \forall i, j \in \mathcal{K}, i \neq j \tag{5.9b}$$

$$H\boldsymbol{F}^{\min} \preceq \boldsymbol{R}(\boldsymbol{\psi}, \boldsymbol{p}) \preceq H\boldsymbol{F}^{\max} \tag{5.9c}$$

$$\boldsymbol{R}(\boldsymbol{\psi}, \boldsymbol{p}) \in \mathcal{R} \tag{5.9d}$$

where the fairness constraints in (5.9b) are translated into rate constraints through $D_k = F_k^{-1}\big(R_k(\boldsymbol{\psi}, \boldsymbol{p})/H\big)$, $\forall k \in \mathcal{K}$ and $\boldsymbol{F}^{\min} = [F_1^{\min}, \ldots, F_K^{\min}]^{\mathrm{T}}$, with $F_k^{\min} = F_k(D_k^{\mathrm{bl}})$, and $\boldsymbol{F}^{\max} = [F_1^{\max}, \ldots, F_K^{\max}]^{\mathrm{T}}$, with $F_k^{\max} = F_k(D_k^{\mathrm{hl}})$, are the minimum and the maximum rates, respectively, of the SNR scalable video streams in the given *application frame interval*. The relationship between the rate and the distortion is here modeled according to eq. 3.9. The definition of the the distortion-fairness metric in the constraint (5.9b) can be found in eq. (4.2).

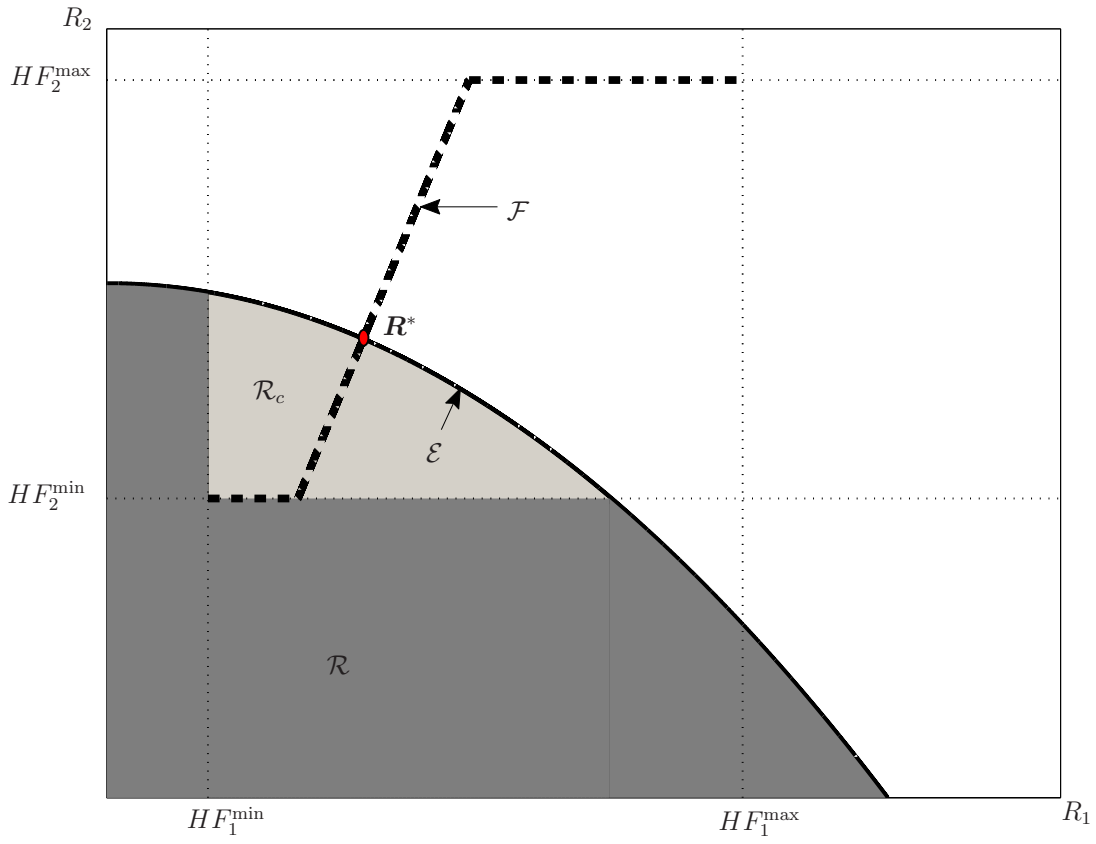


Figure 5.2: An example of two-user optimization problem as in (5.9). $\boldsymbol{R}^*$ is the optimal solution given by the intersection between the boundary of the rate region $\mathcal{E}$ and the piece-wise (bold dashed) curve $\mathcal{F}$ related to the constraint $\Delta(D_1, D_2) = 0$. The problem is feasible because $H\boldsymbol{F}^{\min}$ belongs to $\mathcal{R}$.

According to the constraints (5.9c) and (5.9d), any feasible solution of the problem should belong to $\mathcal{R}_c = \{\boldsymbol{R} \in \mathcal{R} : H\boldsymbol{F}^{\min} \preceq \boldsymbol{R} \preceq H\boldsymbol{F}^{\max}\}$, if it is a non-empty set. This happens if and only if the rate vector $H\boldsymbol{F}^{\min}$ belongs to the interior of $\mathcal{R}$, *i.e.*, if transmission at minimum rate for all videos is supported by the PHY layer. It is also worth noting that a trivial solution to the problem can be derived when all the full quality encoded streams are supported by the rate region, *i.e.*, if $H\boldsymbol{F}^{\max} \in \mathcal{R}$, that corresponds to transmitting all the encoded streams without any adaptation. In Fig. 5.2 we draw an example of the optimization problem for a two-user case. The feasible solutions also lie on the piece-wise curve (dashed

line in the figure):

$$\mathcal{F} = \left\{ \boldsymbol{\varrho} : \Delta\big(F_i^{-1}(\varrho_i/H), F_j^{-1}(\varrho_j/H)\big) = 0, \forall i,j \in \mathcal{K} \right\} \tag{5.10}$$

representing the constraints (5.9b).

We have the following property:

Property 1. *The set $\mathcal{F}$ in eq. (5.10) describes a one-dimensional monotonically increasing manifold with boundary in the $\mathbb{R}^K$ space, i.e., the coordinates in $\mathbb{R}^{K-1}$ are expressed explicitly as a function of one coordinate:*

$$\mathbb{R} \xrightarrow{\mathcal{F}} \mathbb{R}^{K-1}, \tag{5.11}$$

and for any given $\tilde{\boldsymbol{\varrho}}, \tilde{\boldsymbol{\varrho}}' \in \mathcal{F}$, if $\tilde{\varrho}_i > \tilde{\varrho}'_i$ then

$$[\tilde{\varrho}_1, \ldots, \tilde{\varrho}_{i-1}, \tilde{\varrho}_{i+1}, \ldots, \tilde{\varrho}_K] \succeq [\tilde{\varrho}'_1, \ldots, \tilde{\varrho}'_{i-1}, \tilde{\varrho}'_{i+1}, \ldots, \tilde{\varrho}'_K] \tag{5.12}$$

Proof. According to the definition of $\mathcal{F}$ in (5.10), any $\boldsymbol{\varrho} \in \mathcal{F}$ is constrained by $K(K-1)/2$ equations. By fixing one component $\tilde{\varrho}_i \in [HF_i^{\min}, HF_i^{\max}]$, the constraint equations can be reduced to $K-1$ equations, *i.e.*,

$$\frac{\alpha_j}{\varrho_j/H - \beta_j} - \xi_j = \begin{cases} D_j^{\mathrm{bl}}, & \text{if} \quad D_j^{\mathrm{bl}} \le \tilde{D}_i \\ D_j^{\mathrm{hl}}, & \text{if} \quad D_j^{\mathrm{hl}} \ge \tilde{D}_i \\ \tilde{D}_i & \text{otherwise} \end{cases} \qquad \forall j \in \mathcal{K} \setminus \{i\} \tag{5.13}$$

where $\tilde{D}_i = F_i^{-1}(\tilde{\varrho}_i/H)$, which readily proves the one-dimensionality of the manifold with boundary $\mathcal{F}$. The monotonically increasing property is straightforward from the last equation of (5.13), by considering that the inverse of the R-D function given in eq. 3.9, *i.e.*, $\frac{\alpha_k}{F_k - \beta_k} - \xi_k$, is a monotonically strictly decreasing function of F_k. □

According to property 1, since the objective (5.9a) is concave [65], increasing and uniformly bounded $\forall\, \boldsymbol{R} \in \mathcal{R}$ [61], if we assume $H\boldsymbol{F}^{\max} \notin \mathcal{R}$ and $H\boldsymbol{F}^{\min} \in \mathcal{R}$ the optimal solution $\boldsymbol{R}^*$ is clearly attained at the boundary of the rate region $\mathcal{R}$, identified by the Pareto-efficient set:

$$\mathcal{E} = \{\boldsymbol{R} \in \mathcal{R} : \nexists\, \boldsymbol{\varrho} \in \mathcal{R} \text{ s.t. } \boldsymbol{\varrho} \succcurlyeq \boldsymbol{R}\}. \tag{5.14}$$

and is given by the intersection of the piece-wise curve with the rate region boundary $\mathcal{E}$. The optimal solution $\boldsymbol{R}^*$ is unique as proved in lemma 1.

We finally remark that the optimization provides a continuous rate solution, whereas the scalable encoding works with a discrete set of rates. A discrete

rate solution could be evaluated, starting from the continuous one, by further applying proper optimization techniques, *e.g.*, branch & bound search. To keep the complexity low, it is common practice to convert the continuous rate into the nearest discrete rate value smaller than the continuous one, at the expense of a minimum waste of bandwidth.

The evaluation of the optimal solution of the problem in (5.9) would generally require a controller that manages both APP and MAC layers variables and constraints, which is not suitable for realistic network implementations. A desirable solution is the possibility to have single-layer entities that exchange a limited information in a cross-layer fashion, as indicated in Fig. 5.1. This motivates us to decompose problem (5.9) into two sub-problems, each one handling parameters and optimization constraints which are characteristics of a single layer, *i.e.*, in our case, the APP or the MAC layer. In the next subsection we will describe this vertical problem decomposition.

5.3.1 Problem Decomposition

If we first assume that the APP layer has a perfect knowledge of the boundary $\mathcal{E}$ of the rate region $\mathcal{R}$, the problem (5.9) can be simplified into a multi-dimensional constraint-satisfaction problem that aims to find $\boldsymbol{F}$ such that

$$\begin{cases} H\boldsymbol{F} \in \mathcal{E} \cap \mathcal{R}_c \\ \Delta(D_i, D_j) = 0, \quad \forall i,j \in \mathcal{K} \end{cases} \tag{5.15}$$

This is a rate adaptation problem that can be handled by the APP layer. Note that it does not include any objective since the objective of maximizing the source rates is achieved on the boundary $\mathcal{E}$ due to the convexity of the R-D functions.

On the other hand, if we assume that the information about the line where the optimal rate vector lies, which is identified by the parametric equation $\boldsymbol{R}^* = \boldsymbol{\phi}\rho$, is available, the problem (5.9) can be simplified into a problem that can be handled by the MAC layer. The rate direction vector $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_K]^{\mathrm{T}} \succeq \mathbf{0}$ defines the direction of the line and $\rho \in \mathbb{R}^+$ is the parameter. This line departs from $\boldsymbol{R} = \mathbf{0}$ and intersects the boundary $\mathcal{E}$ in $\boldsymbol{R} = \boldsymbol{R}^*$. By assuming $\|\boldsymbol{\phi}\|_1 = 1$, we also obtain $\|\boldsymbol{R}^*\|_1 = \|\boldsymbol{\phi}\|_1\rho = \rho$, *i.e.*, the parameter is the sum-rate. By exploiting this information, the second problem becomes a constrained sum-rate maximization where the objective is to find the optimal allocation policy $(\boldsymbol{\psi}, \boldsymbol{p})$ that maximizes

the sum-rate under the aforementioned proportionality constraints:

$$\max_{(\boldsymbol{\psi},\boldsymbol{p})\in\mathcal{A}} \rho \tag{5.16a}$$

$$s.t.\ \boldsymbol{R}(\boldsymbol{\psi},\boldsymbol{p}) \in \mathcal{R} \tag{5.16b}$$

$$\boldsymbol{R}(\boldsymbol{\psi},\boldsymbol{p}) \succeq \boldsymbol{\phi}\rho \tag{5.16c}$$

This is a simple resource allocation problem. Optimal and efficient solutions of it are well known in literature [66] and do not require the a-priori knowledge of the rate region $\mathcal{R}$. Only the information on the vector $\boldsymbol{\phi}$ is needed and this could be provided by the APP layer. In fact, once the solution $\boldsymbol{F}^*$ of the first problem in (5.15) is known, vector $\boldsymbol{\phi}$ can be easily evaluated as

$$\boldsymbol{\phi} = \frac{\boldsymbol{F}^*}{\|\boldsymbol{F}^*\|_1} \tag{5.17}$$

The main challenge is still on setting up and solving problem (5.15). In fact, the boundary $\mathcal{E}$ of the rate region for the OFDMA scenario cannot be explicitly derived in a fading environment, even when a perfect channel distribution information (CDI) is available at the BS side. To overcome this challenge we propose an efficient iterative method based on the local approximation of the boundary $\mathcal{E}$, which simplifies problem (5.15).

5.4 Iterative Local Approximation (ILA) Algorithm

The starting point for developing the algorithm is the following proposition [61]:

Proposition 1. *Each point on the boundary $\mathcal{E}$ of the rate region $\mathcal{R}$ is the result of the maximization of a weighted sum of average rates (WSAR), i.e.,*

$$\max_{(\boldsymbol{\psi},\boldsymbol{p})\in\mathcal{A}} \boldsymbol{\mu}^T\boldsymbol{R}(\boldsymbol{\psi},\boldsymbol{p}) \tag{5.18a}$$

$$s.t.\ \boldsymbol{R}(\boldsymbol{\psi},\boldsymbol{p}) \in \mathcal{R} \tag{5.18b}$$

for a given $\boldsymbol{\mu} = [\mu_1,\ldots,\mu_K]^T \succcurlyeq \mathbf{0}$.

The WSAR maximization problem is a well-investigated problem and low-complexity procedures can be derived to obtain almost-sure optimal solutions for OFDMA wireless systems [61,67]. In this problem the vector $\boldsymbol{\mu}$ is usually selected

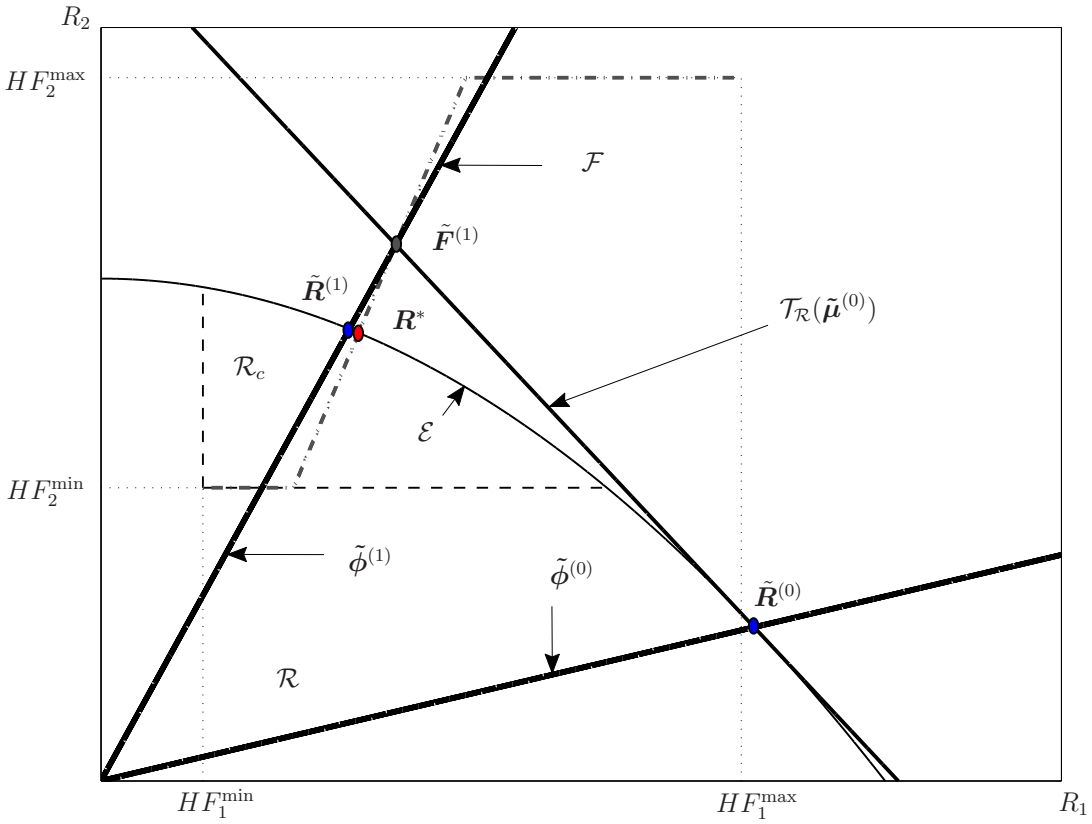


Figure 5.3: An example of the first step of the ILA algorithm for a system with two users.

to enforce some notions of fairness, efficiency, *etc.*, commonly embedded inside utility functions [62].

However, it is shown in [66] that even the solution of problem (5.16) can be obtained through a WSAR maximization problem, where the weights $\boldsymbol{\mu}$ are derived in the dual domain.

Interestingly, the null space of the weight vector $\boldsymbol{\mu}$ also identifies the tangent space to the boundary $\mathcal{E}$ of the rate region at the point where the optimal solution of the WSAR problem is located [62]. The key idea proposed here is to exploit the tangent space as a local approximation of $\mathcal{E}$ to build an iterative procedure between APP and MAC layers that converges to the optimal solution of the problem. To this end, let us denote with $\tilde{\boldsymbol{R}}$ the optimal rate solution of the WSAR problem with weights $\tilde{\boldsymbol{\mu}}$. The tangent space of $\mathcal{R}$ at the point $\tilde{\boldsymbol{R}}$ is then defined

by the following set:

$$\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{\mu}}) = \{\boldsymbol{\varrho} : \tilde{\boldsymbol{\mu}}^{\mathrm{T}}(\boldsymbol{\varrho} - \tilde{\boldsymbol{R}}) = 0\}. \tag{5.19}$$

The cross-layer procedure, named iterative local approximation (ILA) algorithm, can be presented as follows.

Given an initial values of the vector $\boldsymbol{\phi}$, *i.e.*, $\tilde{\boldsymbol{\phi}}^{(0)}$, the MAC layer solves the problem in (5.16). The resulting optimal rate solution $\tilde{\boldsymbol{R}}^{(0)}$ and the weights $\tilde{\boldsymbol{\mu}}^{(0)}$, which identify the tangent space $\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{\mu}}^{(0)})$, are forwarded to the APP layer. The APP layer exploits this information to derive the optimal distortion-fair solution $\tilde{\boldsymbol{F}}^{(1)}$ such that $H\tilde{\boldsymbol{F}}^{(1)}$ is on the tangent space, *i.e.*, $H\tilde{\boldsymbol{F}}^{(1)} \in \mathcal{F} \cap \mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{\mu}}^{(0)})$. hence outside the achievable rate region due to the convexity of $\mathcal{R}$.

The solution for vector $\boldsymbol{\phi}$, *i.e.*, $\tilde{\boldsymbol{\phi}}^{(1)}$ (see eq. (5.17)), is then forwarded to the MAC layer, which projects the solution on the boundary of $\mathcal{R}$ by solving the problem (5.16) to get $\tilde{\boldsymbol{R}}^{(1)}$, and the related weights $\tilde{\boldsymbol{\mu}}^{(1)}$. These steps are iterated until convergence, according to a closed loop strategy. The procedure can be stopped when the error between APP and MAC solutions, which is $\delta^{(i)} = \|H\tilde{\boldsymbol{F}}^{(i)} - \tilde{\boldsymbol{R}}^{(i)}\|_1$, is sufficiently small. An example of the first step of the ILA algorithm for two users is depicted in Fig. 5.3, whereas the details are reported in Algorithm 2 below. The optimality and convergence is stated in the following

Algorithm 2 ILA algorithm

1: $i = 0$; set $\tilde{\boldsymbol{\phi}}^{(0)}$ and error bound ϵ
2: Solve problem (5.16) to get $\tilde{\boldsymbol{\mu}}^{(0)}$, $\tilde{\boldsymbol{R}}^{(0)}$
3: **repeat**
4: $\quad i = i + 1$
5: $\quad$ Find $\tilde{\boldsymbol{F}}^{(i)} : H\tilde{\boldsymbol{F}}^{(i)} \in \mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{\mu}}^{(i-1)}) \cap \mathcal{F}$
6: $\quad \tilde{\boldsymbol{\phi}}^{(i)} = \frac{\tilde{\boldsymbol{F}}^{(i)}}{\|\tilde{\boldsymbol{F}}^{(i)}\|_1}$
7: $\quad$ Solve problem (5.16) to get $\tilde{\boldsymbol{\mu}}^{(i)}$, $\tilde{\boldsymbol{R}}^{(i)}$
8: **until** $\delta^{(i)} < \epsilon$

lemma:

Lemma 1. *ILA algorithm converges to the unique optimal rate solution* $\boldsymbol{R}^* \in \mathcal{E} \cap \mathcal{F}$ *of problem (5.9) under the assumptions* $H\boldsymbol{F}^{max} \notin \mathcal{R}$ *and* $H\boldsymbol{F}^{min} \in \mathcal{R}$, *i.e.*

$$\lim_{i \to \infty} \tilde{F}^{(i)} = \boldsymbol{R}^*/H. \tag{5.20}$$

Proof. We first prove that the optimal solution is unique.

Since the optimal solution $\boldsymbol{R}^*$ is given by the intersection of the boundary $\mathcal{E}$ of the convex rate region, which is a $(K-1)$-dimensional manifold with boundary,

with the monotonically increasing one-dimensional manifold with boundary $\mathcal{F}$, then $\boldsymbol{R}^*$ is a unique point in $\mathbb{R}^K$. In fact, if we had two intersections belonging to $\mathcal{F}$, *i.e.*, $\boldsymbol{R}'$ and $\boldsymbol{R}''$, we would obtain $\boldsymbol{R}' \preceq \boldsymbol{R}''$ if at least one k exists such that $R'_k < R''_k$. But $\boldsymbol{R}''$ should also be below the tangent space touching $\boldsymbol{R}'$, *i.e.*, $\tilde{\boldsymbol{\mu}}^{\mathrm{T}}(\boldsymbol{R}' - \boldsymbol{R}'') \leq 0$, thus contradicting $\boldsymbol{R}' \preceq \boldsymbol{R}''$ if $\tilde{\boldsymbol{\mu}} \succ \mathbf{0}$. Note that we can never have $\tilde{\mu}_k = 0$, as in this case the solution of the WSAR problem would lead to $R_k = 0$ which is the coordinate of a point that can not be touched by any line with $\tilde{\phi}_k = \tilde{F}_k / \|\tilde{\boldsymbol{F}}\|_1 > 0$.

We now prove that $\tilde{\boldsymbol{F}}^{(i)}$ is a monotonically decreasing sequence. According to the problem formulation in (5.31), we have $H\tilde{\boldsymbol{F}}^{(i)} \in \mathcal{F}$, $\forall i$, and

$$\tilde{\boldsymbol{\mu}}^{(i)\mathrm{T}} H \tilde{\boldsymbol{F}}^{(i+1)} = \tilde{\boldsymbol{\mu}}^{(i)\mathrm{T}} \tilde{\boldsymbol{R}}^{(i)} \tag{5.21}$$

if the manifold $\mathcal{F}$ intersects the tangent space $\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{\mu}}^{(i)})$. Since $\tilde{\boldsymbol{R}}^{(i)}$ is the projection of $H\tilde{\boldsymbol{F}}^{(i)}$ on $\mathcal{E}$ through the line defined by the proportionality contraints

$$\tilde{\boldsymbol{R}}^{(i)} = \tilde{\boldsymbol{\phi}}^{(i)} \rho = \frac{\tilde{\boldsymbol{F}}^{(i)}}{\|\tilde{\boldsymbol{F}}^{(i)}\|_1} \rho, \tag{5.22}$$

we have

$$H\tilde{\boldsymbol{F}}^{(i)} \succcurlyeq \tilde{\boldsymbol{R}}^{(i)}, \forall i \tag{5.23}$$

By combining it with eq. (5.21) and by observing that $\tilde{\mu}_k^{(i)} \neq 0$, $\forall k \in \mathcal{K}$, we find

$$\tilde{\boldsymbol{\mu}}^{(i)\mathrm{T}} H \tilde{\boldsymbol{F}}^{(i+1)} < \tilde{\boldsymbol{\mu}}^{(i)\mathrm{T}} H \tilde{\boldsymbol{F}}^{(i)} \tag{5.24}$$

Therefore, from (5.12) and (5.24) it follows that $\boldsymbol{F}^{(i)}$ is a component-wise monotonic sequence, *i.e.*,

$$\tilde{\boldsymbol{F}}^{(i+1)} \preccurlyeq \tilde{\boldsymbol{F}}^{(i)}, \forall i. \tag{5.25}$$

If there were no intersection between $\mathcal{F}$ and $\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{\mu}}^{(i)})$, we would have $\tilde{\boldsymbol{\mu}}^{(i)\mathrm{T}} H \boldsymbol{F}^{\max} < \tilde{\boldsymbol{\mu}}^{(i)\mathrm{T}} \tilde{\boldsymbol{R}}^{(i)}$, thus implying $\tilde{\boldsymbol{\mu}}^{(i)\mathrm{T}} H \boldsymbol{F}^{\max} < \tilde{\boldsymbol{\mu}}^{(i)\mathrm{T}} H \tilde{\boldsymbol{F}}^{(i)}$. But this can not happen, because $\boldsymbol{F}^{(i)} \preceq \boldsymbol{F}^{\max}$.

We now prove that the sequence $\tilde{\boldsymbol{F}}^{(i)}$ converges to the limiting fixed point $\boldsymbol{R}^*/H$. A sufficient condition is given by

$$\|\tilde{\boldsymbol{F}}^{(i+1)} - \tilde{\boldsymbol{F}}^{(i)}\|_2 \xrightarrow{i\to\infty} 0 \Rightarrow \tilde{\boldsymbol{F}}^{(i)} - \boldsymbol{R}^*/H \xrightarrow{i\to\infty} \mathbf{0} \tag{5.26}$$

By exploiting (5.21), we have:

$$H\tilde{\boldsymbol{F}}^{(i)} - \tilde{\boldsymbol{R}}^{(i)} = \tilde{\boldsymbol{\phi}}^{(i)}\delta^{(i)} \tag{5.27a}$$

$$\tilde{\boldsymbol{\mu}}^{(i)\mathrm{T}}(H\tilde{\boldsymbol{F}}^{(i)} - \tilde{\boldsymbol{R}}^{(i)}) = \tilde{\boldsymbol{\mu}}^{(i)\mathrm{T}}\tilde{\boldsymbol{\phi}}^{(i)}\delta^{(i)} \tag{5.27b}$$

$$H\tilde{\boldsymbol{\mu}}^{(i)\mathrm{T}}(\tilde{\boldsymbol{F}}^{(i)} - \tilde{\boldsymbol{F}}^{(i+1)}) = \delta^{(i)} \tag{5.27c}$$

$$H\|\tilde{\boldsymbol{F}}^{(i)} - \tilde{\boldsymbol{F}}^{(i+1)}\|_2 \geq \frac{\delta^{(i)}}{\|\tilde{\boldsymbol{\mu}}^{(i)}\|_2} \tag{5.27d}$$

where (5.27c) holds because $\tilde{\boldsymbol{\mu}}^{(i)T}\tilde{\boldsymbol{\phi}}^{(i)} = 1$ and (5.27d) follows from the Cauchy-Schwarz inequality. If $\|\tilde{\boldsymbol{\mu}}^{(i)}\|_2$ is bounded and $\|\tilde{\boldsymbol{F}}^{(i+1)} - \tilde{\boldsymbol{F}}^{(i)}\|_2 \xrightarrow{i\to\infty} 0$, eq. (5.27d) implies that $\delta^{(i)} \xrightarrow{i\to\infty} 0$, proving the lemma. To show that $\|\tilde{\boldsymbol{\mu}}^{(i)}\|_2$ is bounded, we first consider that

$$\boldsymbol{\phi}^{(i)} \succeq \boldsymbol{F}^{\min}/\|\boldsymbol{F}^{\max}\|_1 \succeq \mathbf{1}\min_k(F_k^{\min})/\|\boldsymbol{F}^{\max}\|_1. \tag{5.28}$$

Then, from $\tilde{\boldsymbol{\mu}}^{(i)T}\tilde{\boldsymbol{\phi}}^{(i)} = 1$ we obtain

$$1 \geq \tilde{\boldsymbol{\mu}}^{(i)T}\mathbf{1}\min_k(F_k^{\min})/\|\boldsymbol{F}^{\max}\|_1 = \|\tilde{\boldsymbol{\mu}}^{(i)}\|_1\min_k(F_k^{\min})/\|\boldsymbol{F}^{\max}\|_1 \tag{5.29}$$

which implies that $\|\tilde{\boldsymbol{\mu}}^{(i)}\|_1$ is bounded $\forall i$:

$$\|\tilde{\boldsymbol{\mu}}^{(i)}\|_1 \leq \frac{\|\boldsymbol{F}^{\max}\|_1}{\min_k(F_k^{\min})} \tag{5.30}$$

Since $\|\tilde{\boldsymbol{\mu}}^{(i)}\|_2 \leq \|\tilde{\boldsymbol{\mu}}^{(i)}\|_1$, $\|\tilde{\boldsymbol{\mu}}^{(i)}\|_2$ is also bounded. □

Let us finally underlining that, for our proposed projection, the error at step i, given by $\delta^{(i)} = \|\delta^{(i)}\tilde{\boldsymbol{\phi}}^{(i)}\|_1$, is proportionally distributed across individual rates according to $\tilde{\boldsymbol{\phi}}^{(i)}$, *i.e.*, the error for the k-th user rates is given by $\delta_k^{(i)} = \tilde{\phi}_k\delta^{(i)}$. This property enforces the fairness also for the intermediate solutions of the ILA algorithm when they are used as suboptimal solutions for practical applications (see Section 5.7).

5.5 Application Layer Algorithm: Rate Adaptation

By exploiting the local approximation of $\mathcal{E}$ given by the tangent space $\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{\mu}})$ at $\tilde{\boldsymbol{R}}$, the problem (5.15) at the APP layer will be simplified into the following

constraint-satisfaction problem:

$$\begin{cases} \tilde{\boldsymbol{\mu}}^{\mathrm{T}} H \boldsymbol{F} - \Lambda = 0 \\ \boldsymbol{F}^{\min} \preceq \boldsymbol{F} \preceq \boldsymbol{F}^{\max} \\ \Delta(D_i, D_j) = 0 \quad \forall i, j \in \mathcal{K} \end{cases} \tag{5.31}$$

where $\Lambda = \tilde{\boldsymbol{\mu}}^{\mathrm{T}} \tilde{\boldsymbol{R}}$ is the value of the WSAR resulting from the solutions of (5.18). The problem admits a feasible solution under two conditions, $\tilde{\boldsymbol{\mu}}^{\mathrm{T}} H \boldsymbol{F}^{\max} \geq \Lambda$ and $\tilde{\boldsymbol{\mu}}^{\mathrm{T}} H \boldsymbol{F}^{\min} \leq \Lambda$, which relax the two feasibility conditions of the main problem, *i.e.*, $H\boldsymbol{F}^{\max} \notin \mathcal{R}$ and $H\boldsymbol{F}^{\min} \in \mathcal{R}$, respectively. In the ILA algorithm, if the first condition is violated at the iteration i, the vector $\tilde{\boldsymbol{F}}^{(i)} = \boldsymbol{F}^{\max}$ can be used to replace the APP layer solution. Conversely, if the second condition is violated at the iteration i, the APP layer can terminate the ILA algorithm, since no adaptation is feasible.

In Chapter 4 we derived an optimal low-complexity procedure based on the simplified SVC model with two parameters to solve problem (5.31), in the special case where all the weights are equal to 1 and sum-rate is a fixed value. The procedure can be extended to the more general case considered here. Let $x_k, y_k \in \{0, 1\}$, $k \in \mathcal{K}$, with $(x_k, y_k) \neq (0, 0)$, be binary variables that indicate whether (1) or not (0) the two constraints $F_k \geq F_k^{\min}$ and $F_k \leq F_k^{\max}$, respectively, are satisfied. We then define the function

$$\mathit{\Gamma}(\boldsymbol{x}, \boldsymbol{y}, D) = \sum_{k \in \mathcal{K}} x_k y_k \tilde{\mu}_k \left(\frac{\alpha_k}{D + \xi_k} + \beta_k \right) - \mathit{\Lambda}(\boldsymbol{x}, \boldsymbol{y}) \tag{5.32}$$

where

$$\mathit{\Lambda}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\Lambda}{H} - \sum_{k \in \mathcal{K}} \tilde{\mu}_k \big[(1 - x_k) F_k^{\min} + (1 - y_k) F_k^{\max} \big] \tag{5.33}$$

generalizes the similar function defined in equation 4.10. By applying the procedure in Algorithm 1 we obtain the pseudo-code of the algorithm summarized as follows.

Algorithm 3 Pseudo code to solve problem (5.31)

if $\tilde{\boldsymbol{\mu}}^{\mathrm{T}} H \boldsymbol{F}^{\min} > \Lambda$ **then**
  report infeasibility
else if $\tilde{\boldsymbol{\mu}}^{\mathrm{T}} H \boldsymbol{F}^{\max} \leq \Lambda$ **then**
  report infeasibility and set $\tilde{F}_k = F_k^{\max}$, $\forall k \in \mathcal{K}$
else
  $y_k = 1$, $\forall\ k \in \mathcal{K}$;
  repeat
    $cond_{\mathrm{HL}}$ = false;
    $x_k = 1$, $\forall\ k \in \mathcal{K}$;
    repeat
      $cond_{\mathrm{BL}}$ = false;
      Compute $\tilde{D} : \Gamma(\boldsymbol{x}, \boldsymbol{y}, \tilde{D}) = 0$;
      for all $k \in \mathcal{K} : x_k y_k = 1$ **do**
        $\tilde{F}_k = \frac{\alpha_k}{\tilde{D} + \xi_k} + \beta_k$;
        if $\tilde{F}_k < F_k^{\min}$ **then**
          $\tilde{F}_k = F_k^{\min}$; $x_k = 0$; $cond_{\mathrm{BL}}$ = true;
        end if
      end for
    until $cond_{\mathrm{BL}}$ is false
    for all $k \in \mathcal{K} : x_k y_k = 1$ **do**
      if $\tilde{F}_k > F_k^{\max}$ **then**
        $\tilde{F}_k = F_k^{\max}$; $y_k = 0$; $cond_{\mathrm{HL}}$ = true;
      end if
    end for
  until $cond_{\mathrm{HL}}$ is false
end if

As in previous case, the algorithm requires in the worst case a maximum of $K(K-1)/2$ iterations, which happens in the unpractical case when $\boldsymbol{F}^{\min} \simeq \boldsymbol{F}^{\max}$. According to extensive simulations in practical scenarios, K iterations are enough to terminate the procedure. The optimality of the algorithm is stated in the following lemma:

Lemma 2. *Algorithm 3 converges to the unique optimal rate solution* $\boldsymbol{F}^*$ *of problem (5.31) under the assumptions* $\tilde{\boldsymbol{\mu}}^T H \boldsymbol{F}^{min} \leq \Lambda$ *and* $\tilde{\boldsymbol{\mu}}^T H \boldsymbol{F}^{max} \geq \Lambda$.

Proof. For a given pair of vectors $\boldsymbol{x} = [x_1, \ldots, x_K]^{\mathrm{T}}$ and $\boldsymbol{y} = [y_1, \ldots, y_K]^{\mathrm{T}}$, the value of distortion D^* to be assigned to all videos that have $x_k = y_k = 1$, can be obtained from the equation

$$\Gamma(\boldsymbol{x}, \boldsymbol{y}, D) = 0 \tag{5.34}$$

by using a numerical method. With other words, this value of D^* is the unique solution of

$$\begin{cases} \tilde{\boldsymbol{\mu}}^{\mathrm{T}} H \boldsymbol{F} - \Lambda = 0 & \\ F_k = F_k(D) & \text{if} \quad x_k = y_k = 1 \\ F_k = F_k^{\min} & \text{if} \quad x_k = 0 \\ F_k = F_k^{\max} & \text{if} \quad y_k = 0 \end{cases} \tag{5.35}$$

By denoting with $\boldsymbol{F}^*$ the resulting source rate vector, if $F_k^{\min} \leq F_k^* \leq F_k^{\max}$, $\forall k \in \mathcal{K}$, then $\boldsymbol{F}^*$ is also the unique solution of problem (5.31). Hence, the solution of problem (5.31) can be obtained through a search in the space of all $(\boldsymbol{x}, \boldsymbol{y})$ to find the optimal pair $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ that gives $\boldsymbol{F}^{\min} \preceq \boldsymbol{F}^* \preceq \boldsymbol{F}^{\max}$. Algorithm 2 performs this search by reducing the complexity from exponential to quadratic. To show the optimality of the algorithm, let us consider the two following propositions.

Proposition 1: Let D' be the solution of eq. (5.34) for a given $(\boldsymbol{x}, \boldsymbol{y})$ having $x_m = y_m = 1$. If x_m is changed from 1 to 0 and at the same time $F_m(D') < F_m^{\min}$, the new solution D'' is such that $D'' > D'$ and, consequently, $F_k'' < F_k', \forall k : x_k y_k = 1$.

This can be proved by evaluating from (5.34) the difference

$$D'' - D' = \tilde{\mu}_m (F_m^{\min} - F_m(D')) \left[\sum_{k \neq m} \frac{x_k y_k \tilde{\mu}_k \alpha_k}{(D' + \xi_k)(D'' + \xi_k)} \right]^{-1}$$

which is always greater than 0.

Proposition 2: Let $\boldsymbol{F}'$ be the solution of the problem

$$\begin{cases} \tilde{\boldsymbol{\mu}}^{\mathrm{T}} H \boldsymbol{F} - \Lambda = 0 & \\ \Delta(D_i, D_j) = 0 & \forall i, j \in \mathcal{K} \\ F_k \geq F_k^{\min} & \text{if} \quad y_k = 1 \\ F_k = F_k^{\max} & \text{if} \quad y_k = 0 \end{cases} \tag{5.36}$$

for a given $\boldsymbol{y}$ having $y_m = 1$. If $F_m' > F_m^{\max}$ and y_m is changed from 1 to 0, the new solution $\boldsymbol{F}''$ is such that $F_k'' > F_k', \forall k : x_k y_k = 1$. The solution of this problem can be obtained through a search in the space of all $\boldsymbol{x}$, after having fixed $\boldsymbol{y}$, to find the optimal vector $\boldsymbol{x}'$ that satisfies the last two contraints of (5.36).
The proof is straightforward.

The two propositions can be used to show that Algorithm 2 is able to find the solution of the original problem (5.31). In the algorithm, the inner procedure from line 8 to line 19, for a given vector $\boldsymbol{y}$, finds the values of F_k, $\forall k \in \mathcal{K}$, that solve problem (5.36). It starts from $\boldsymbol{x} = \mathbf{1}$, evaluates D, evaluates F_k for all users

and finally sets $F_k = F_k^{\min}$ and $x_k = 0$ for those users resulting with $F_k < F_k^{\min}$. Then, it repeats the same steps until no update is done on vector $\boldsymbol{x}$. Note that, according to Proposition 1, the updated values of F_k can only decrease. The outer procedure from line 6 to line 25 finds the values of $\boldsymbol{F}$ that solve problem (5.31). It starts from $\boldsymbol{y} = \mathbf{1}$, evaluates F_k for all users by solving the problem (5.36) in the inner loop, and finally sets $F_k = F_k^{\max}$ and $y_k = 0$ for those users resulting with $F_k > F_k^{\max}$. Then, it repeats the same steps until no update is done on vector $\boldsymbol{y}$. Note that, according to Proposition 2, the updated values of F_k can only increase. □

5.6 MAC Layer Algorithm: Resource Allocation

By exploiting the value of vector $\tilde{\boldsymbol{\phi}}$, provided by the APP layer at each ILA algorithm iteration, the MAC layer is now able to find the solution of problem (5.16). The solution is derived through dual decomposition, as in [66], which also provides the dual geometric multipliers, *i.e.*, the weight vector $\boldsymbol{\mu}$ of the WSAR. We retrace here the main results of [66] according to the time-sharing assumption of the subcarriers.

Let $L(\boldsymbol{\psi}, \boldsymbol{p}, \lambda, \boldsymbol{\mu})$ be the Lagrangian function, where λ is the dual variable related to the average power constraint implicitly considered in (5.16b), and $\boldsymbol{\mu}$ is the dual vector related to the proportionality constraint (5.16c). The dual problem becomes

$$\begin{aligned} &\min_{\lambda, \boldsymbol{\mu}} \;\; g(\lambda, \boldsymbol{\mu}) \\ &s.t. \; \lambda > 0, \; \boldsymbol{\mu} \succcurlyeq \mathbf{0}, \; (1 - \boldsymbol{\mu}^{\mathrm{T}} \tilde{\boldsymbol{\phi}}) = 0 \end{aligned} \tag{5.37}$$

where the third constraint holds to avoid sum-rate diverging to infinity or being zero and $g(\lambda, \boldsymbol{\mu}) = \max_{\boldsymbol{\psi}, \boldsymbol{p}} L(\boldsymbol{\psi}, \boldsymbol{p}, \lambda, \boldsymbol{\mu})$ is the dual objective. In order to derive $g(\lambda, \boldsymbol{\mu})$, given λ and $\boldsymbol{\mu}$, the expression of the Lagrangian function can be suitably manipulated leading to:

$$g(\lambda, \boldsymbol{\mu}) = \lambda P + S\mathbb{E}\left[\max_{\boldsymbol{\psi}, \boldsymbol{p}} M(\psi_{k,s}, p_{k,s})\right] \tag{5.38}$$

where $M(\psi_{k,s}, p_{k,s}) = \mu_k r_{k,s}(\psi_{k,s}, p_{k,s}) - \lambda p_{k,s}$. The unique solution for the dual objective is obtained when each couple, subcarrier s and time slot n, is assigned to a unique user $u_s[n]$, *i.e.*,

$$u_s^*[n] = \arg\max_{k \in \mathcal{K}} M(\psi_{k,s}^*[n], p_{k,s}^*[n]) \tag{5.39}$$

where the optimal sharing factors and powers, are [61]:

$$\psi^*_{k,s}[n] = 1;\ p^*_{k,s}[n] = \left[\frac{a_1 \Delta B \mu_k}{\lambda \ln 2} - \frac{a_2}{\gamma_{k,s}[n]}\right]^+. \tag{5.40}$$

which means exclusive subcarrier assignment, *i.e.*, $\psi^*_{k,s}[n] = p^*_{k,s}[n] = 0,\ \forall k \neq u^*_s[n]$.

Due to the convex definition of $\mathcal{R}$ and the concavity of the constraints, the strong duality holds, *i.e.*, the optimal dual solution is equal to the primal one, resulting in a zero duality-gap. Since the dual problem is in general not tractable analytically, an iterative sub-gradient method as in [66] can be used to solve it. It is important to remark that the scheduling algorithm described by (5.38) and (5.39) works on the time scale of slot intervals (order of milliseconds), whereas the dual problem solution needs to be updated every *application frame intervals* (order of seconds). Such solution is the input for the ILA algorithm, and therefore should be computed N times, where N is the number of iterations required by the ILA algorithm to converge.

5.7 Practical Issues for the Implementation

The cross-layer framework, based on the ILA approach presented in section 5.4, requires the iteration of the two algorithms presented in section 5.5 and 5.6, *i.e.*, those that solve problems (5.31) and (5.16), respectively. The main practical challenges come from the solution of (5.37), which generally requires perfect knowledge of CDI to compute the expectation of rate and power. Although some methods to estimate the CDI and compute the expectation are known in literature (see [68] and [67]), they are computationally expensive. A viable alternative for the solution of (5.37) is to implement an adaptive stochastic sub-gradient algorithm as in [60][67], where the update equations are evaluated along time, once for each time slot, and the average power and rate in the subgradients are computed through a stochastic approximation. The implementation details of the ILA with stochastic algorithm at MAC layer are discussed in the next section. It is shown that parallel processing is required to execute the ILA algorithm. To drastically reduce the complexity with a small performance penalty, the 1-step ILA algorithm is proposed.

5.7.1 ILA with Stochastic Algorithm at MAC Layer

In the adaptive implementation, the solution of (5.37) is evaluated through an adaptive stochastic sub-gradient algorithm, after user and power allocations are

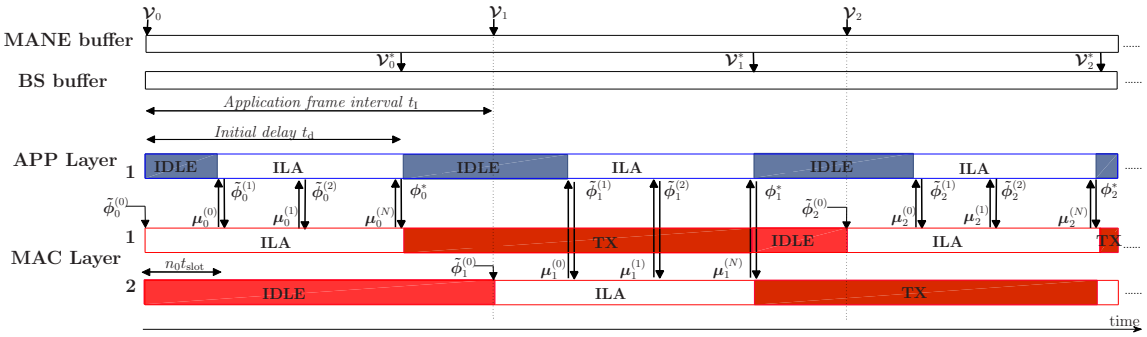


Figure 5.4: Evolutions of the ILA algorithm iterations according to the stochastic approximation framework.

obtained as in (5.39) and (5.40). The dual variables are updated at each time slot n according to the following equations:

$$\begin{cases} \lambda[n+1] = [\lambda[n] - \delta_\lambda g_\lambda[n]]_\epsilon^+ \\ \boldsymbol{\mu}[n+1] = [\boldsymbol{\mu}[n] - \delta_{\boldsymbol{\mu}} \boldsymbol{g}_{\boldsymbol{\mu}}[n]]^+ \end{cases} \tag{5.41}$$

where

$$g_\lambda[n] = P - \sum_{s \in \mathcal{S}} p_{k,s}^*[n], \tag{5.42}$$

$$\boldsymbol{g}_{\boldsymbol{\mu}}[n] = \boldsymbol{r}^*[n] - \tilde{\boldsymbol{\phi}} \|\boldsymbol{r}^*[n]\|_1 \tag{5.43}$$

with $\boldsymbol{r}^*[n] = [r_1^*, \ldots, r_K^*]$, and

$$r_k^* = \sum_{s \in \mathcal{S}} r_{k,s}^*(\psi_{k,s}[n], p_{k,s}[n]) \tag{5.44}$$

Finally, δ_λ, $\delta_{\boldsymbol{\mu}}$ are step-sizes suitably selected to ensure convergence [66].

The time required for the MAC layer algorithms to converge may be significant and with great impact on the ILA algorithm. However, the proposed framework can still work by introducing parallel processing and by assuming perfect synchronization between APP and MAC algorithms. The final cost will be the presence of a latency t_d in the transmission of video frames, which is the time required to execute N steps of the ILA algorithm.

The implementation details of the ILA algorithm based on the stochastic algorithm at the MAC layer are illustrated in Fig. 5.4 and explained as follows.

Let us assume that the first set of encoded videos $\boldsymbol{\mathcal{V}}_0$ at the *application frame* 0 is ready for transmission. the encoded pictures set $\mathcal{I}_k$ of each video. Given any

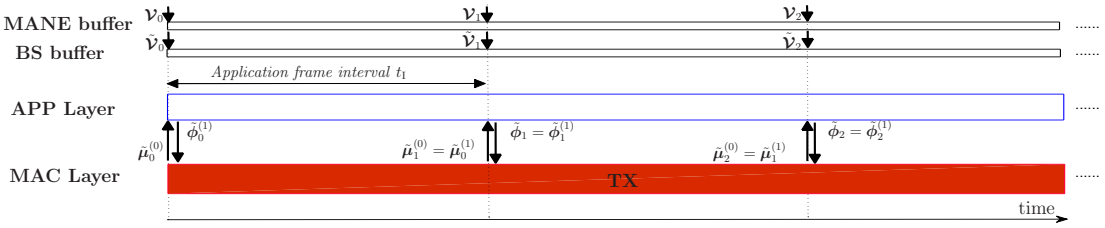


Figure 5.5: The 1-step ILA algorithm.

initial $\tilde{\boldsymbol{\phi}}$, *i.e.*, $\tilde{\boldsymbol{\phi}}_0^{(0)}$, the MAC layer performs the iterations along n_0 time slots to find the dual variables $\tilde{\boldsymbol{\mu}}_0^{(0)}$, and the related rate vector $\tilde{\boldsymbol{R}}_0^{(0)}$, which are forwarded to the APP layer. The APP layer derives the vector $\tilde{\boldsymbol{\phi}}_0^{(1)}$ by solving problem (5.31), and sends it to the MAC layer, which again finds the related optimal solutions $\tilde{\boldsymbol{\mu}}_0^{(1)}$ along n_1 time slots.

After N iterations, required by the ILA algorithm to converge, the APP layer extracts the sub-streams of the videos in $\boldsymbol{\mathcal{V}}_0$ that meet the optimal rate solution $\boldsymbol{F}_0^*$ and forward the resulting TBs, *i.e.*, the set $\boldsymbol{\mathcal{V}}_0^*$, to the lower layers. of the related $\boldsymbol{\mathcal{I}_0}$ sub-streams. Simultaneously to the transmission of $\boldsymbol{\mathcal{V}}_0^*$, MAC and APP layers restart the same procedure for the next set $\boldsymbol{\mathcal{V}}_1$ of videos, in order to derive the optimal transmission parameters $\tilde{\boldsymbol{\mu}}_1^{(N)}$ and $\boldsymbol{F}_1^*$. The transmission of the a-th set $\boldsymbol{\mathcal{V}}_a^*$ of adapted sub-streams can starts after $at_\mathrm{I} + t_\mathrm{d}$ seconds, where t_d is the time allowed for ILA algorithm to converge.

The number of parallel processes depends on the relationship between t_d and t_I. By assuming $t_\mathrm{d} = ct_\mathrm{I}$, $c \in \mathbb{R}^+$, it is intuitively provable from Fig. 5.4 that APP and MAC layers require $\lceil c \rceil$ and $\lceil c \rceil + 1$ parallel processes, respectively. Therefore, as the delay t_d increases, also the complexity increases. The most interesting approach to achieve a good trade off between complexity and performance is the 1-step-based ILA algorithm, which does not require parallel computation.

5.7.2 1-step ILA Algorithm

The 1-step ILA algorithm is built on the assumption that the R-D relationship of the encoded sets of pictures does not significantly change over two consecutive *application frames*. This allows to eliminate the parallel processing illustrated in the previous section, to finally obtain the operations described in Fig. 5.5 where only one step of the ILA algorithm is executed. As shown in the figure, the APP layer solution $\tilde{\boldsymbol{\phi}}$ for the current application frame with index a, denoted with $\tilde{\boldsymbol{\phi}}_a$,

is derived instantaneously, because the one-step ILA outcomes are obtained from the results of MAC algorithm in the $(a-1)$-th *application frame*, *i.e.*, by using the initial vectors[1] $\tilde{\boldsymbol{\mu}}_a^{(0)} = \tilde{\boldsymbol{\mu}}_{a-1}^{(1)}$ and $\tilde{\boldsymbol{R}}_a^{(0)} = \tilde{\boldsymbol{R}}_{a-1}^{(1)}$. Therefore, the rate adaptation process is based on the local approximation of the rate region resulting from the preceding *application frame*, and the set of encoded sub-streams ($\tilde{\boldsymbol{\mathcal{V}}}_a$ in the figure) can be sent to the BS buffer at the beginning of the current *application frame* without delay. This approach greatly reduces the algorithm complexity and the number of cross-layer iterations, and makes transmission latency negligible.

However, the residual rate error $\delta^{(1)}$ of the ILA algorithm at the first iteration is not negligible. Hence, the transmission of the adapted sub-stream may require more than one *application frame interval*. Such error is expected to be small as long as the video complexities and the channel conditions present slight variations between two consecutive *application frames*. When such conditions do not hold buffer underflow may arise with uncomfortable pauses during the video reproduction. We will propose next a method to compensate the rate error $\delta^{(1)}$.

5.7.3 Residual Error Compensation in the 1-step ILA Algorithm

As usually done in practical applications, we introduce a minimum initial play-out delay t_0, before starting to reproduce the video, in order to allow the transmission of the data still in the queue at the beginning of each *application frame interval*. Let B_k be the total amount of data still in the buffer of the BS for user k. If more than one TB has not been transmitted, the queue contains several blocks of bits $b_{k,g}$, which have different time-to-deadlines $t_{k,g}^{\mathrm{dl}}$, updated to include the play-out delay t_0, where $g = 1, 2, \ldots$, is the TB index. We then define

$$Q_k = \max_g \frac{\sum_{i=1}^{g} b_{k,i}}{t_{k,g}^{\mathrm{dl}}} \tag{5.45}$$

as the minimum rate required to ensure the transmission of all TBs, before their time-to-deadline. When $t_0 > 0$, the APP layer will be able to compensate the effects of rate mismatch if it updates the R-D functions $F_k(D)$ of the next *application frame* in order to satisfy the following two not-exclusive constraints:

1. F_k must support the transmission of the incoming encoded set of pictures also considering the residual data in the buffer, *i.e.*,

$$F_k(D) \geq \frac{\alpha_k}{D + \xi_k} + \beta_k + B_k / t_{\mathrm{I}} \tag{5.46}$$

[1]Here and in Fig. 5.5, the subscript in the vectors refers to the *application frame* index.

2. (ii) $F_k^{\min}$ must be greater than or equal to the minimum rate required to transmit the residual data in the buffer before its time-to-deadline, *i.e.*,

$$F_k^{\min} \geq Q_k. \tag{5.47}$$

Such constraints are always satisfied by simply re-evaluating: the parameter β_k of the R-D functions as $\beta'_k = \beta_k + B_k/t_{\mathrm{I}}$ to obtain a new function $F'_k(D)$; the maximum rate as $F_k^{\max} = F'_k(D_k^{\mathrm{hl}})$; the minimum rate as $F_k^{\min} = \max\{Q_k, F'_k(D_k^{\mathrm{bl}})\}$, $\forall k \in \mathcal{K}$. The compensation method proposed here can be easily implemented without changing the proposed algorithms.

5.8 Numerical Evaluations

We consider an OFDMA access network with frequency spacing $\Delta B = 15$ kHz, time slot duration $t_{\mathrm{slot}} = 10$ ms and a maximum average power budget $P = 1$ W. The number of available subcarriers is set to $S = 64$ if not specified otherwise. A total of $K = 10$ users are uniformly distributed in a cell with radius equal to 300 m, with resulting average SNR ranging from 7 to 28 dB. The adaptive modulation and coding system is characterized by a rate adjustment $a_1 = 0.905$ and an SNR gap $a_2 = 1.34$ [64]. The users request different video sequences with different spatial and temporal complexities, *i.e.*, *City*, *Crew*, *Coastguard*, *Container*, *Football*, *Foreman*, *Harbour*, *Mobile*, *News* and *Soccer* in CIF resolution with a frame-rate of 30 fps (see Table 3.1). Each sequence is looped 10 times and encoded with the JSVM reference software [30] with one base layer and two enhancement layers, and quantization parameters 40, 34 and 28, respectively. Each enhancement layer is split into five MGS layers with vector distribution [3 2 4 2 5] and the post-processing priority level assignment is then applied. GOP size and IDR period are set to 8 and to 32 frames, respectively. The three parameters of model (3.9) are evaluated for each IDR period, resulting in an *application frame* window $W_{\mathrm{I}} = 106$. We set the overhead factor $H = 1$.

Individual video qualities are evaluated according to the PSNR computed through the MSE D_k^{mean} averaged over all the transmitted frames as in eq. 3.4:

$$\mathrm{PSNR}_k = 10 \log_{10} \left(\frac{255^2}{D_k^{\mathrm{mean}}} \right) \tag{5.48}$$

while the overall performance is evaluated with the global PSNR [11] as

$$\mathrm{GPSNR} = 10 \log_{10} \left(\frac{255^2 K}{\sum_{k \in \mathcal{K}} D_k^{\mathrm{mean}}} \right) \tag{5.49}$$

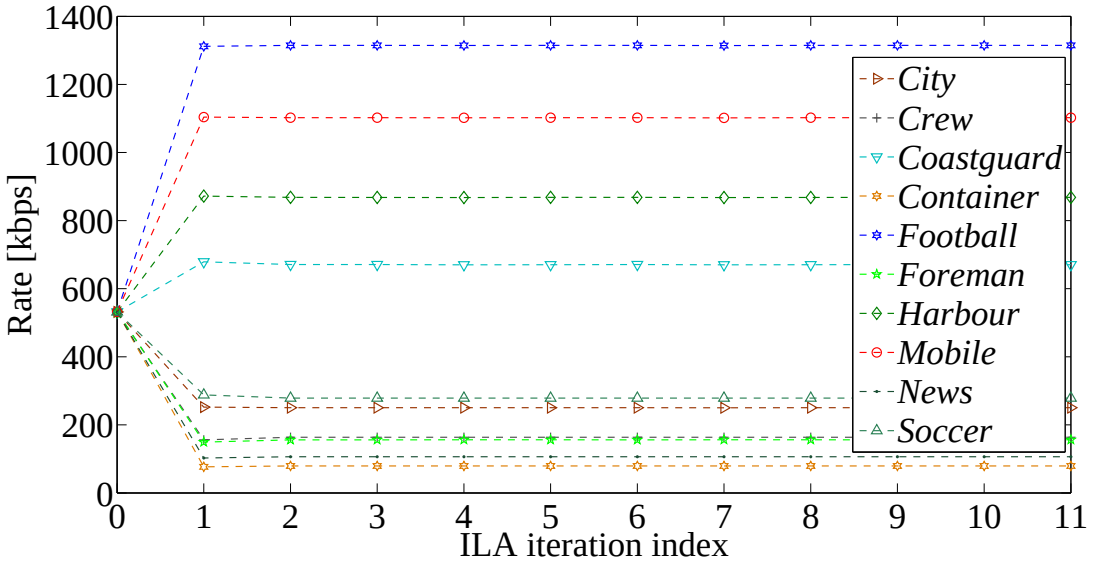


Figure 5.6: PHY layer rates at each step of the ILA algorithm for a randomly selected *application frame interval*. iterations provide solutions close to the optimal ones.

Moreover, to better assess the improvements of our strategy in terms of end-user perceived quality, we also provide results in terms of the standardized ANSI VQM, assuming full reference calibration (see [46]). The VQM value is a number between 0 and 1 used to judge the visual quality, which shows high correlation with subjective quality test. A low VQM value indicates good perceived quality.

We compare the performance of the ILA and the 1-step ILA algorithms, denoted with EQ-ILA and EQ-1STEP, respectively, aiming at equalizing the distortion, with the two following strategies:

- equal rate strategy, denoted with ER-1STEP, which aims to provide fairness only in terms of assigned video rate, while satisfying the minimum and maximum rate constraints. It is unaware of the individual R-D relationship of each video and is built by replacing the constraints in (5.9b) with $\Delta_F(F_i, F_j) = 0$ where the function Δ_F is defined in a way similar to (4.2). It can be implemented through a 1-step of the ILA algorithm by replacing line 14 of algorithm 2 with

$$\tilde{F}_k = \frac{\Lambda(\boldsymbol{x}, \boldsymbol{y})}{\sum_{k \in \mathcal{K}} x_k y_k \tilde{\mu}_k} \tag{5.50}$$

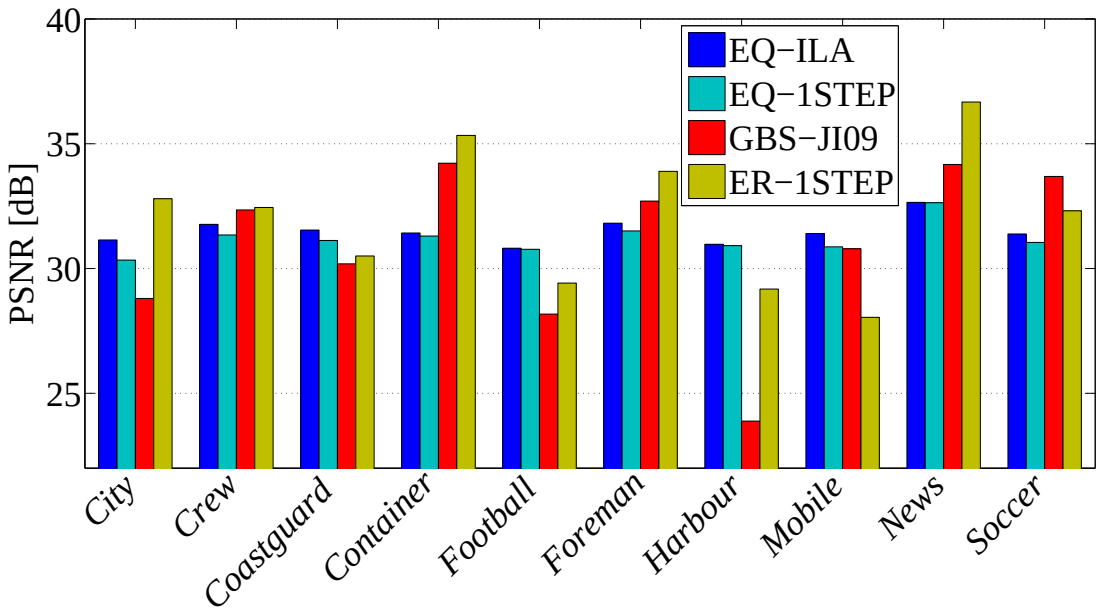


Figure 5.7: PSNR of each video resulting from distortion-fair ILA (EQ-ILA) algorithm, 1-step ILA (EQ-1STEP) algorithm, equal-rate algorithm (ER-1STEP) and the strategy proposed in [12] (GBS-JI09). The initial playout deadline is set to 200 ms.

as in [10], which is an extension of the approach used for numerical comparison in section 4.4.

- cross-layer gradient-based scheduling strategy proposed by Ji *et al.* [12], denoted with GBS-JI09, which maximizes an instantaneous weighted sum-rate. In this framework, the APP layer collects and sorts the frames of each GOP into several sub-flows according to their dependencies, and the MAC layer, at each scheduling interval and for each sub-flow, updates the weights according to: distortion reduction achieved through successful delivery of the sub-flow, time-to-deadline and sub-flow length. Let us note that the framework exploits the temporal and quality scalability of the SVC streams only when the packets violate their play-out deadline, *i.e.*, no prior rate adaptation is performed. However, in order to have a fair comparison, we assume that a preliminary rate adaptation for each video sequence is carried out to not exceed a maximum average PSNR of 35 dB as in [12].

For performance evaluation we consider two different scenarios. In the first scenario, the transmissions are assumed error-free and only affected by fast fading

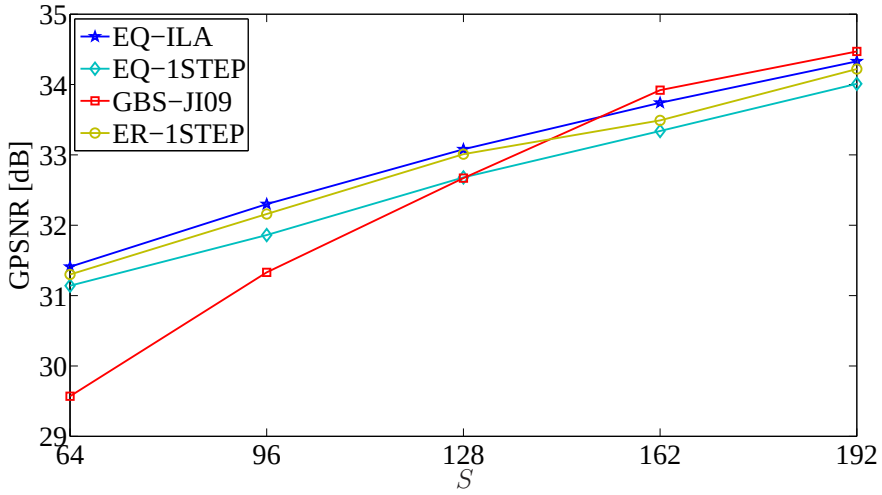
which is obtained from a multi-path channel model with delay spread of 2.3 µs and Doppler bandwidth of 6 Hz. In the second scenario, the transmissions are affected by RTP packet losses, with loss-rate of 10% as in [11][10], and by additional mobility effects modeled through a log-normal shadowing process (std. deviation: 6 dB) with an exponential auto-correlation (correlation time: 20s).

5.8.1 Static Scenario with Error-free Transmission

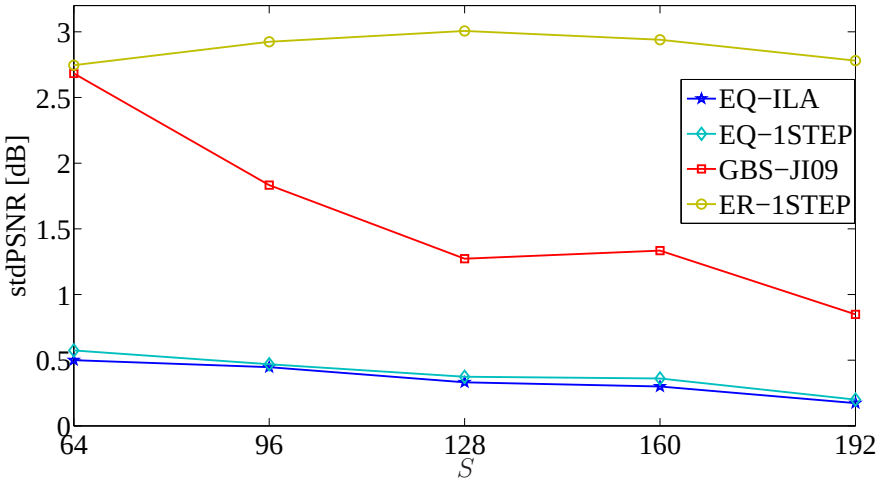
In Fig. 5.6 the PHY layer rate of each user resulting at each iteration of the EQ-ILA algorithm is plotted for a randomly selected *application frame interval* (similar results are obtained for the other *application frame intervals*), by using the initialization $\phi_k^{(0)} = 1/K$, $\forall k \in \mathcal{K}$. As expected, the ILA algorithm quickly converges, achieving the optimal solutions, *i.e.*, $\delta \approx 0$, in no more than $10 - 11$ iterations in all the investigated cases. However, two iterations are always enough to approach the optimal solution with a relative error $(HF_k - R_k)/R_k < 10^{-4}$ in all the investigated cases. This result further justifies the use of the suboptimal 1-step ILA algorithm.

Fig. 5.7 compares the individual video qualities obtained with play-out deadline $t_0 = 200$ms for the different scheduling and adaptation strategies. We first note that the gap between best and worst PSNR resulting from the optimal (EQ-ILA) and the sub-optimal (EQ-1STEP) is relatively small. It ranges from 0.1 to 0.8 dB. Both the approaches provide approximately uniform quality to each video, with the exception of *News* whose minimum rate constraints are active for most of the time. Moreover, the proposed EQ-1STEP compared to the GBS-JI09 strategy is able to improve the video quality of the users requesting the most demanding videos, *i.e.*, *Coastguard*, *Football*, *Harbour* and *Mobile*, and experiencing relatively bad channel conditions, *e.g.*, *City* and *Coastguard*, in the scenario simulated here, while exhibiting similar complexity[2]. For these videos EQ-1STEP achieves a gain ranging between 1.5 to 7 dB. It is also interesting to note that the ER-1STEP strategy generally outperforms the gradient-based scheduling, thanks to the adaptation process, even based on rate information only. In the GBS-JI09 strategy, the loss of base-layer frames, due to several deadline violations, highly reduces the received PSNR. It should also be noted that the PSNR values of GBS-JI09 are computed by excluding the set of pictures with missing I-frames, thus overestimating the actual quality.

[2]EQ-1STEP requires stream adaptation at each *application frame interval* (in the order of seconds), but GBS-JI09 requires information exchange at each time slot (in the order of tenth ms) for the evaluation of the weights.



(a)



(b)

Figure 5.8: Overall performance in terms of global PSNR (GPSNR) (a) and standard deviation of the PSNRs (b) of distortion-fair ILA (EQ-ILA) algorithm, 1-step ILA (EQ-1STEP) algorithm, equal-rate algorithm (ER-1STEP) and the strategy proposed in [12] (GBS-JI09), for different numbers of available subcarriers S.

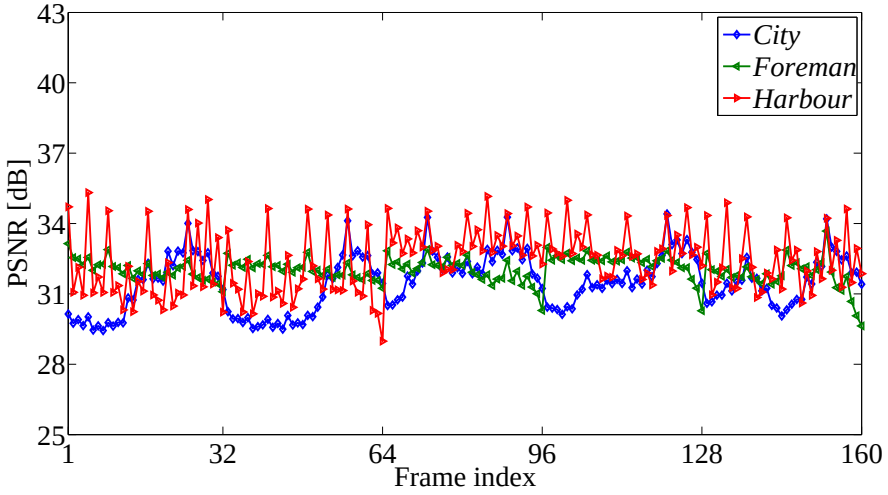
Such behaviour is more clear in Table 5.1 where the ANSI VQM values are reported for $S = 128$. We can first note the significant improvement of the adaptation-based strategies, *i.e.*, EQ-1STEP and ER-1STEP, over GBS-JI09, since the perceived video quality is more sensitive to the temporal impairments

ANSI VQM	EQ-1STEP	ER-1STEP	GBS-JI09
City	0.128	0.068	0.711
Crew	0.133	0.110	0.518
Coastguard	0.072	0.090	0.548
Container	0.157	0.082	0.636
Football	0.121	0.135	0.837
Foreman	0.129	0.070	0.767
Harbour	0.045	0.063	0.830
Mobile	0.027	0.046	0.255
News	0.109	0.033	0.723
Soccer	0.114	0.102	0.904
Average	0.103	0.085	0.672
Std. Dev.	0.041	0.039	0.192

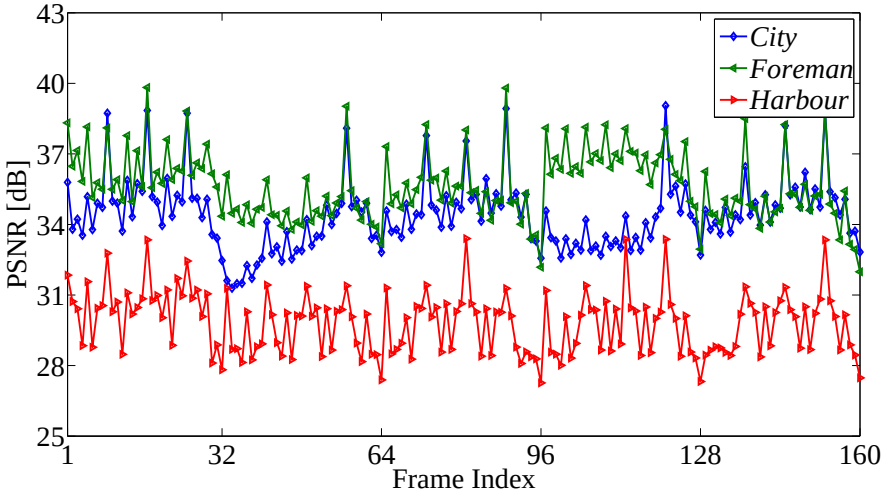
Table 5.1: Individual and global (average and standard deviation) values of the ANSI VQM for different strategies

of the videos. We also note the comparable performance of ER-1STEP and EQ-1STEP. In fact, in the absence of frame losses, small encoding rate variations do not significantly impact the perceived quality. However, when the VQM values are smaller than 0.1-0.2, the VQM metric becomes less useful for comparing the different strategies, because the correlation between VQM values and subjective scores decreases [46].

The trade-off between efficiency and fairness is investigated in Fig. 5.8(a) and 5.8(b), where the global PSNR and the standard deviation of the user PSNRs (stdPSNR) are plotted for different numbers of allocable subcarriers. We first note how the fairness achieved by the proposed strategies improves as the number of available sub-carriers increases, due to the de-activation of the minimum rate constraints for the low-complexity videos. which limits the minimum allocable distortions. Even though the GBS-JI09 strategy is aimed at maximizing the efficiency, the loss of the base layer frames due to the limited resources (up to 64 and 96 available sub-carriers) and to the lack of adaptation between APP and MAC layer does not allow GBS-JI09 to outperform the fairness-oriented strategies, unless the number of available resources becomes quite large. We finally plot in Fig. 5.9(a) and 5.9(b) the end-user quality, frame-by-frame, resulting from EQ-1STEP and ER-1STEP, respectively, for a subset of three videos. The figures further assess the benefits of the proposed strategy in terms of PSNR fairness with respect to rate-oriented strategies.



(a)



(b)

Figure 5.9: End-user frame-by-frame PSNR resulting from distortion-fair 1-step ILA (EQ-1STEP) algorithm (a) and equal-rate algorithm (ER-1STEP), with S=128.

5.8.2 Scenario with User Mobility and Error-prone Transmission

In this scenario, the presence of mobility makes the channel non-stationary in a time scale larger than the *application frame interval*. Therefore, we first inves-

Strategy	GPSNR [dB]	min-max PSNR [dB]	stdPSNR [dB]
EQ-ILA	31.9	31.4-32.6	0.29
EQ-1STEP	31.3	30.8-32.6	0.42
ER-1STEP	31.8	28.5-37.4	2.75

Table 5.2: GPSNR and fairness (minimum and maximum PSNR, PSNR standard deviation) for different strategies.

tigate this issue in error-free channel by comparing the behaviour of both the proposed 1-step ILA algorithm and the GBS-JI09 strategy in presence of mobility with the behaviour of the same algorithms in static scenario. The GPSNR achieved with 64 subcarriers decreases from 31.3 dB to 31 dB for EQ-1STEP and from 29.7 dB to 27.7 dB for GBS-JI09, which indicates the robustness of the 1-step ILA to moderate non-stationarity.

We move now to error-prone transmission by assuming a initial play-out-deadline $t_0 = 500$ms and a number of available subcarriers $S = 128$, which allows the system to support at least the base layer rate (including the UXP parity bytes) of each video for all the simulation time. As benchmark we consider an ideal ILA algorithm virtually running without shadowing variations in the channel. We do not consider the GBS-JI09 strategy, since it can not support the UXP. Table 5.2 compares the performance in terms of global PSNR and fairness. We can note the small gap of the EQ-1STEP strategy with respect to the benchmark, even in case of error-prone channel and the significant fairness improvement compared to ER-1STEP.

Chapter 6

Cross-layer Optimization for Health-com Services Delivery in Uplink SC-FDMA Channels

In the last decade, e-health has become one of the most promising applications of emerging information and communication technologies (ICT) [69]. The fundamental concept of the e-health is to provide innovative healthcare services supported by electronic/digital processes and data remote transmission. In particular, telemedicine services can highly benefit from the recent advances offered by mobile communication systems [70], which are nowadays potentially able to support a wide range of ubiquitous healthcare applications, such as tele-diagnosis [71], real-time monitoring of vital parameters [72], remote treatment of patients and even tele-surgery. In the next future, this kind of mobile-health (m-health) services are expected to spread rapidly, increasing the efficacy and efficiency of the healthcare offer with decreasing costs.

M-health services can also play an important role in the management of emergency situations, such as those involving one or more ambulances rushed to the scene of an accident. In this case, the presence of a 3G/4G radio access network can be exploited to establish a communication link between the emergency area and a remote hospital, enabling real-time and interactive tele-consultation services through the exchange of audio, video and other medical information [73].

In this context, the transmission of health-related information from an ambulance to a remote hospital is a challenging task, due to the variability and the limitations of the mobile radio link. In particular, the transmission of multiple video streams can improve the efficacy of the tele-consultation service, but requires a large bandwidth to meet the desired quality, not always guaranteed by

the mobile network. Therefore, the possibility to adapt the video encoding to the current transmission conditions becomes particularly important in the context of emergency m-health. In these cases, in fact, just a limited buffering can be adopted, since a low delivery latency is imposed by the need of real-time interaction with remote specialists, while high quality videos are always fundamental to provide an effective medical support. As already mentioned, SVC conjugates good compression efficiency with high flexibility in rate adaptation. For these reasons several solutions have been recently proposed for e-health applications based on SVC [74][75].

In LTE, healthcare related video traffic can be prioritized over less critical traffic. This can be done through enhanced RRA techniques, allowing to support a certain level of QoS. However, differently from the uplink where OFDMA is used and optimized QoS aware RRA strategies have already been proposed, Single Carrier - Frequency Division Multiple Access (SC-FDMA) [6], also known as DFT-spread FDMA, has been selected as key-technology for the LTE uplink. The main reason is that in OFDMA schemes the resulting time-domain waveform exhibits very pronounced envelope fluctuations resulting in a high peak-to-average power ratio (PAPR). Signals with a high PAPR require highly linear power amplifiers to avoid excessive inter-modulation distortion. To achieve this linearity, the amplifiers have to operate with a large back-off from their peak power, resulting in an increasing cost and power consumption. Such problem is clearly more critical in the up-link transmission where the cost and power consumption of mobile must be kept as lower as possible. SC-FDMA keeps similar advantages of the OFDMA systems but provide a lower PAPR by introducing a DFT pre-coding process at the transmitter (see Fig. 6.1), which spreads the data power over the entire allocated bandwidth. Such advantages is paid in an increased Inter-Symbol Interference (ISI) at the receiver, which requires adaptive frequency domain equalization to cancel this interference. These trade-offs are then well balanced in a cellular system since an increasing cost of complex signal processing (frequency domain equalization) at the base station is acceptable if followed with a reduction of the burden of linear amplification in portable terminals.

There are two types of SC-FDMA: localized-FDMA (L-FDMA) in which the sub-channels assigned to a user are adjacent to each other, and interleaved-FDMA (I-FDMA) in which users are assigned with sub-channels distributed over the entire frequency band. Only L-FDMA is taken into account by LTE. A detailed overview of the SC-FDMA scheme can be found in [76] and [6].

After the introduction of the SC-FDMA for the uplink, by LTE in 2004, the related RRA problem has gather increasingly interest. SC-FDMA, and more specifically L-FDMA, consider only contiguous PRBs allocation, making the problem a

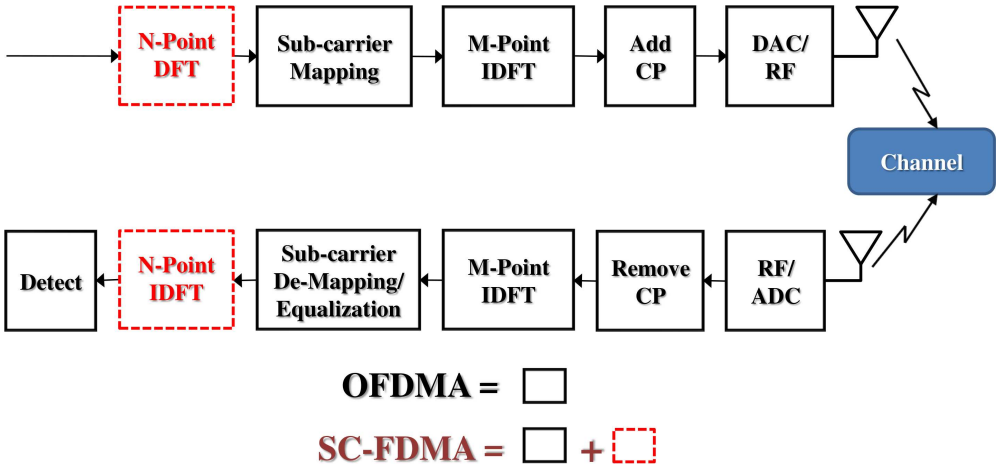


Figure 6.1: Transmitter and receiver structure of SC-FDMA and OFDMA modulation. SC-FDMA introduces a N-point DFT ($N > M$) pre-coding to spread the data power over the entire allocated bandwidth.

strict NP-hard problem [77], where the techniques traditionally used to optimally solve OFDMA RRA problem (*e.g.*, dual decomposition as proposed in section 5.6) cannot be directly applied. For this reason, the first scheduler proposals, *e.g.*, [6] [78][77], were based on greedy approaches, following the idea of maximizing the per-user marginal utility, *i.e.*, multiple contiguous PRBs are assigned to users which have the maximum increase of benefits. The first proposal published by Lim *et. al.*[78] consider and equal-bit-equal-power (EBEP) allocation for each sub-channel in order to maximize a proportional fairness utility function. Instead of solving the optimization problem, they provide a sub-carrier allocation scheme which improves the marginal utility using both L-FDMA and I-FDMA schemes. In both cases the heuristic scheduler aims to find the subchannel which has the highest marginal utility for one user. It is defined as the difference between the utility obtained when the s-th subchannel is allocated to user k and the utility of user k in the absence of a subchannel allocation. The heuristic is derived to ensure contiguous or interleaved subchannel allocation, according to L-FDMA and I-FDMA, respectively. Specifically, when a user is allocated to a set of subchannels, the algorithm deletes such set from the set of available subchannels, without checking if other users can exhibit a larger marginal utility gain. The complexity results linear in the number of users and subchannels. The authors show the significant gain, in terms of aggregate throughput of the localized scheme compared to interleaved one. They also extended their work to take into account the outdated CSI as in practical high-mobility scenario [79].

The authors in [80] provide three different proposal, namely the First Maximum Expansion (FME), Recursive Maximum Expansion (RME), Minimum Area Difference (MAD), based on a proportional fairness metric. FME and RME algorithm are extensions to the one proposed in [78]. Both algorithms assign the subchannels starting from that with the highest metric of the $S \times K$ matrix representing the channel gain. Then it expands the allocation for a selected user on both the right and the left sides of the subchannel with the largest metric. However the subchannels are not dropped from each allocation as in [78], but the algorithms also verify if an other user can provide a largest marginal utility improvement. The main difference between FME and RME is the action taken when reaching a subchannel allocation whose maximum metric is found for two user. Finally MAD algorithms aims to provide the minimum difference between the cumulative utility of different users and the the maximum utility value for any given subchannels.

A comparison study of the performance of these scheduler proposals was provided in [81] with respect to the Search-Tree Based Packet Scheduling (STBPS) algorithms proposed in [82]. They also proposed a modified version of the FME algorithm. The performance comparison are carried out according to throughput, spectral efficiency and fairness for different number of users in the systems. They shows how all the schemes are able to exploit multi-user diversity by increasing throughput and spectral efficiency as the number of users increases. However, as expected, all algorithms lacks for QoS support due to the absence of GBR constraints in the problem statements. STBPS algorithm exhibits better performance in terms of fairness data-rate with respect to all scheme but pays in a loss of sum-throughput. On the other hand, MAD provides the highest efficiency, both in terms of sum of throughput and spectral efficiency, also providing more fairness with respect to FME and RME which exhibit the worst overall performance.

An extended comparison study was provided by the same authors in [83], where mixed traffic is used as prescribed by 3GPP for practical performance evaluation. In their study an heuristic localized gradient algorithm (HLGA) as proposed in [84] was also included. HLGA is designed to include Hybrid -Automatic Request (H-ARQ) schemes as contemplated by LTE, where a subset of RBs are reserved for H-ARQ process for previous unsuccessful transmissions. Performance evaluations were carried out in terms of per-user throughput, packet loss and fairness. However, none of the mentioned framework provides optimal solutions to understand the overall benefits of each proposal in terms of performance and complexity.

Recently, Wong *et al.* [85] provided a novel reformulation of the RRA problem into a pure binary-integer program (BIP), namely a set partitioning problem, which allows to compute the optimal allocation through known methods of the op-

eration research, *e.g.*, linear programming (LP) relaxation, thus avoiding exhaustive enumeration. Ahmad and Assaad [86] took advantage of such reformulation to transform the BIP problem into a canonical dual problem in the continuous space. They analytically proved that under certain conditions, the solution of the canonical dual problem is identical to the solution of the primal problem and finally showed that such solution is close to the optimal one. The algorithmic complexity of the proposals in [85]-[86] is still unsuited for practical applications and, although both frameworks are able to capture frequency and multi-user diversity of OFDM-based systems, the TTI-based maximization does not allow to fully capture the time-diversity of the channel. Moreover, the consideration of user-specific rate requirements is not addressed, which may be a drawback in QoS-aware applications [87][88].

In order to ensure the required level of QoS to the m-health application considered in this chapter, we will first address the problem of ergodic sum-rate maximization under proportional rate constraints for the uplink of SC-FDMA systems. To the Author knowledge, this problem has not yet been investigated in the literature. We then propose a novel cross-layer adaptation strategy for multiple SVC videos delivered over a single LTE channel, which dynamically adjusts the overall transmitted throughput to meet the actual available bandwidth, while being able to provide high quality to diagnostic video sequences and lower (but fair) quality to less critical ambient videos

Contribution

The novel contribution of this chapter is two-fold:

- We analyze the ergodic sum-rate maximization problem under proportional rate constraints in SC-FDMA systems and propose a novel sub-optimal algorithmic solution, whose complexity increases only linearly with the number of users and the number of resources. Numerical results show that the performance gap to optimal solution is limited to the 10% of the sum-rate.

- We propose a novel solution for the transmission of multiple videos from an emergency scenario, based on the joint video adaptation and aggregation directly performed at the application layer of the transmitting equipment. In fact, a strict separation into multiple single flows may turn out to be inefficient, especially in case of simultaneous transmission from multiple and heterogeneous co-located sources. we consider two categories of videos transmitted from the ambulance: (i) ambient videos that allow the hospital staff to visually follow the patient conditions and the activities performed

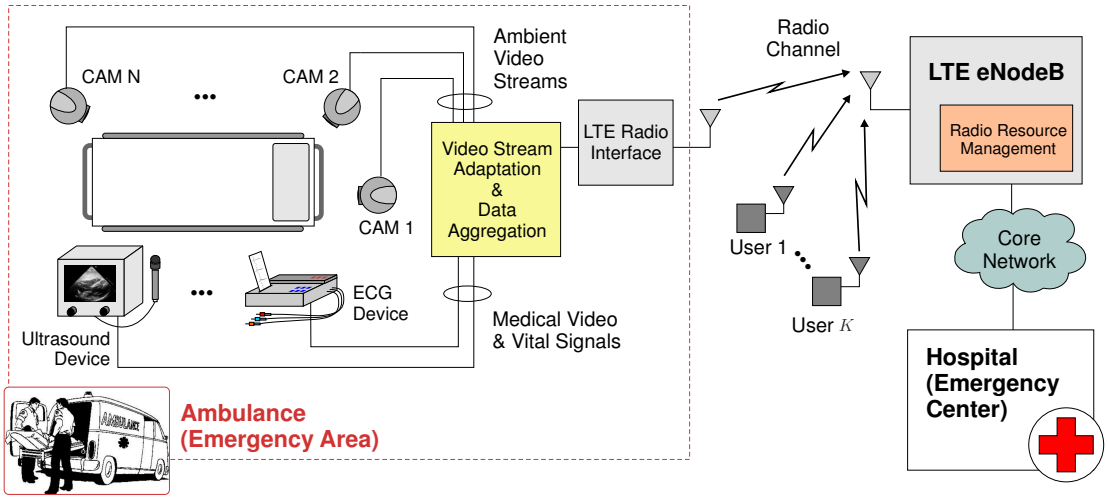


Figure 6.2: The proposed m-health architecture for emergency scenarios.

in the ambulance; (ii) diagnostic videos obtained as result of emergency examinations, such as the Focused Assessment with Sonography for Trauma (F.A.S.T.), which is used to rapidly assess the status of heart and abdominal organs of the patient [18]. From the LTE e-NodeB perspective, only a single communication link characterized by given QoS guarantees needs to be managed between the ambulance and the remote hospital, while additional spectrum efficiency is gained from video multiplexing. In our solution the adaptation is designed to optimize quality and fairness by exploiting the information on the available rate assigned by the LTE e-nodeB. It is shown that the proposed strategy permits to achieve a good end-to-end quality even in the presence of rate limitations and fluctuations due to the wireless channel and intense traffic within the LTE cell.

6.1 System Architecture

The m-health scenario addressed in this chapter is depicted in Fig. 6.2. An ambulance equipped with multiple cameras and medical devices reaches an emergency area where one or multiple injured persons need immediate medical assistance. Multiple information is sent through the available LTE radio access network to the emergency management center at the hospital, where specialized medical staff can follow the first-aid operations, coordinate the intervention and acquire the health-

state information necessary to prearrange the treatment at the hospital. The doctors at the hospital interact in real-time with the ambulance staff, receiving both ambient and ultrasonography videos, as well as other important information on the vital parameters of the patients, such as ECG data, respiratory and cardiac frequency, blood pressure and oxygen concentration. In this work, the real-time video streams consist of one diagnostic F.A.S.T. sequence and N ambient videos acquired by a set of cameras installed on the ambulance. The multimedia flows are processed in real-time and multiplexed with the other medical information by the video adaptation and data aggregation unit before transmission over the LTE radio channel.

From the LTE network perspective, the ambulance (also indicated in the following as m-health user) competes for radio resources with other K users within the cell, indexed by the set $\mathcal{K}$, subdivided into K_1 GBR users and K_2 best-effort users, indexed by the sets $\mathcal{K}_1$ and $\mathcal{K}_2$, respectively. The e-NodeB tries to guarantee the transmission rates $\bar{R}_0$ to the m-health user[1] and $\bar{R}_k$ to the k-th GBR user, with $k \in \mathcal{K}_1$, while the throughput to best-effort users is provided fairly, according to the resources left after allocating all GBR users.

The video adaptation unit performs two fundamental tasks:

- It manages the inherently different priorities of the data flows generated by the m-health user. In particular, it optimally adapts the SVC-encoded streams, in order to deliver the ultrasonography information with sufficiently high quality and the set of ambient videos tuned according to quality fairness criteria.

- It produces an aggregated throughput adapted to the radio channel and cell traffic conditions, according to the amount of resources assigned by the e-NodeB to the m-health user.

6.2 Physical Layer Model for the Uplink of the Wireless Access Network

We consider a single-cell time-slotted SC-FDMA system where multiple users and base station are equipped with a single antenna. The total available bandwidth B is divided into S orthogonal subcarriers, which are grouped in G subchannels of 12 adjacent subcarriers indexed by the set $\mathcal{G} = \{1, \ldots, G\}$. Each group has

[1]Note that here, as well as in the rest of this chapter, we indicate the m-health GBR user with the subscript 0.

bandwidth ΔB (180 KHz in LTE). The piece of frame composed of one group and one slot (0.5ms) is defined as the Physical Resource Block (PRB), which is the elementary resource unit for RRA. Let $h_{k,s}[n]$ be the channel gain between the BS and the user k, on subcarrier s and time slot n. It is modeled as a complex Gaussian ergodic random process (Rayleigh fading), generally correlated across subcarriers and time slots. We define the normalized SNR of user k on subcarrier s and time slot n, as

$$\gamma_{k,s}[n] = \frac{|h_{k,s}[n]|^2}{\sigma^2} \tag{6.1}$$

where σ^2 is the noise power. As already mentioned, in SC-FDMA the available resources, *i.e.*, PRBs, are assigned in a contiguous manner. By assuming a MMSE receiver, the effective SNR experienced by user k over a "pattern" j of G_j adjacent PRBs is given by:

$$\gamma_{k,j}^{\text{eff}}[n] = \left[\left(\frac{1}{12G_j} \sum_{s \in \mathcal{G}_j} \frac{p_{k,s}\gamma_{k,s}[n]}{1 + p_{k,s}\gamma_{k,s}[n]} \right)^{-1} - 1 \right]^{-1} \tag{6.2}$$

where $p_{k,s}$ is the power allocated to user k on subcarrier s and $\mathcal{G}_j$ is the set of subcarriers belonging to the PRBs of pattern j. As in [85][86], we consider constant power allocation, *i.e.*, $p_{k,s} = \frac{P_k}{12G_j}$ where P_k is the per-user power budget. Moreover, power control schemes as in LTE [89] can be easily considered as well. By using a suitable AMC scheme, the rate achieved by user k on slot n when pattern j is allocated, can be evaluated with the following model:

$$r_k(\gamma_{k,j}^{\text{eff}})[n] = G_j \Delta B \log_2\big(1 + \gamma_{k,j}^{\text{eff}}[n]\big). \tag{6.3}$$

As in [85], we define $\boldsymbol{A}$ as the matrix of all feasible allocation patterns in a system with G PRBs. The matrix has G rows and $J = (G^2 + G)/2 + 1$ columns. Each column represents a pattern and each element indicates whether (1) or not (0) the PRB is allocated. It can be easily built by including for each integer $l \in \mathcal{G}$ one column with index $j_l = 1 + \sum_{t=0}^{l-1}(G - t)$ composed of l ones and $G - l$ zeros, and all the shifted version of it with indexes $j_l + m$, $m = 1, ..., G - l - 1$. An example of the matrix $\boldsymbol{A}$ for a set-up with G=5 PRB is given by:

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

This construction facilitates the reduction of the search space when the information on the maximum number of PRBs to be allocated to each user is known. In fact, the patterns with $\tilde{G} \geq 1$ allocated PRB have indexes in the set

$$\mathcal{J}_{\tilde{G}} = \left\{ j : 2 + \sum_{t=0}^{G_j-2} (G-t) \leq j \leq 1 + \sum_{t=0}^{G_j-1} (G-t) \right\} \tag{6.4}$$

which has cardinality $|\mathcal{J}_{\tilde{G}}| = G - \tilde{G} + 1$.

Given the matrix $\boldsymbol{A}$, the RRA algorithm computes for each slot n the $K \times J$ allocation matrix $\boldsymbol{I}[n]$ which has the generic form $\boldsymbol{I} = [\boldsymbol{i}_1, \ldots, \boldsymbol{i}_K]^{\mathrm{T}}$, where the row $\boldsymbol{i}_k = [i_{k,1}, \ldots, i_{k,J}]$ indicates which pattern out of J is allocated to user k, *i.e.*, $i_{k,j} = 1$ if pattern j is allocated to user k and 0 otherwise. A feasible allocation matrix satisfies the constraints:

$$\|\boldsymbol{I}\boldsymbol{a}_g^{\mathrm{T}}\|_1 = 1, \quad \forall g \in \mathcal{G},$$

$$\|\boldsymbol{i}_k\|_1 = 1, \quad \forall k \in \mathcal{K},$$

where $\boldsymbol{a}_g$ is the g-th row of matrix $\boldsymbol{A}$. The set of all feasible allocation matrices is here denoted as $\mathcal{A}$.

6.3 MAC Layer: Radio Resource Allocation

In this section, we analyze the framework of ergodic sum-rate maximization for continuous (capacity based) rates under proportional rate constraints, as done in [66] for downlink OFDM systems. Let us denote with $\boldsymbol{R} = [R_1, \ldots, R_K]^{\mathrm{T}}$, the vector of the ergodic achievable rates where

$$R_k = \mathbb{E}_{\boldsymbol{\gamma}} \left[\sum_{j=1}^{J} i_{k,j}[n] r_k(\gamma_{k,j}^{\mathrm{eff}})[n] \right]. \tag{6.5}$$

The MAC scheduler within the e-NodeB allocates the rates to the users in the cell according to the solution of the following optimization problem,

$$\max \|\boldsymbol{R}\|_1 \tag{6.6a}$$

$$R_k \geq \bar{R}_k, \qquad \forall k \in \mathcal{K}_1 \tag{6.6b}$$

$$R_k \geq \theta_k \left(\sum_{l \in \mathcal{K}} R_l - \sum_{l \in \mathcal{K}_1} \bar{R}_l \right), \ \forall k \in \mathcal{K}_2 \tag{6.6c}$$

$$\boldsymbol{I}[n] \in \mathcal{A} \tag{6.6d}$$

where $\theta_k \geq 0$, $k \in \mathcal{K}_2$, $\sum_{l\in\mathcal{K}_2} \theta_l = 1$, define the required average share of throughput for the best-effort users.

The constraint (5.16c) accounts for the target rate requirement holding for the m-health user and the other GBR users within the cell. On the contrary, the rate allocated to non GBR users must satisfy the inequality (6.6c), which requires that the residual rate after serving all the GBR users (see the term in parenthesis at the right hand side) is assigned to the best effort users based on the ϕ_k. Problem (6.6) admits feasible solutions if and only if the required GBR $\bar{R}_k$ are supported by the rate region, and thus it requires a careful selection of $\bar{R}_k$ according to traffic load and channel conditions. To overcome such issue we translate the problem in (6.6) to an ergodic sum-rate maximization with proportional rate constraints, *i.e.*,

$$\max \|\boldsymbol{R}\|_1 \tag{6.7a}$$
$$s.t.\ \boldsymbol{R} \succeq \boldsymbol{\phi}\|\boldsymbol{R}\|_1 \tag{6.7b}$$
$$\boldsymbol{I}[n] \in \mathcal{A} \tag{6.7c}$$

where now the vector $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_K]^{\mathrm{T}} \succeq \mathbf{0}$, defines the required average share of throughput among all users which must satisfy $\|\boldsymbol{\phi}\|_1 = 1$. In this problem, the values of $\phi_k, \forall k \in \mathcal{K}_1$ can be statistically configured, or adaptively updated to follow the prescribed GBR rate $\bar{R}_k, \forall k \in \mathcal{K}_1$.

Problem (6.7) is a non-linear combinatorial problem which is difficult to be solved directly. As in [66], we decompose the problem by considering a Lagrangian relaxation of the rate constraint. It is important to remark that since the problem is not defined on a convex set and the objective is not differentiable, this is not a convex optimization problem, thus the resulting duality gap may not be zero. Nevertheless, the following Lagrangian relaxation will help to find an excellent suboptimal solution.

6.3.1 Lagrangian Relaxation

Let $L(\boldsymbol{\mu})$ be the Lagrangian function, associated to the problem (6.7a)-(6.7b), and $\boldsymbol{\mu}$ be the dual vector related to the constraint (6.7b). The dual problem becomes

$$\min_{\boldsymbol{\mu}} \Theta(\boldsymbol{\mu}) \tag{6.8a}$$
$$s.t. \quad \boldsymbol{\mu} \succcurlyeq \mathbf{0} \tag{6.8b}$$
$$\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\phi} = 1 \tag{6.8c}$$

where constraint (6.8c) holds to avoid sum-rate diverging to infinity or being zero. Given the dual vector $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_K]$, the dual objective

$$\Theta(\boldsymbol{\mu}) = \max_{\boldsymbol{I}[n]\in\mathcal{A}} \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{R} \tag{6.9}$$

is evaluated by using the allocation matrix $\boldsymbol{I}[n]$ that maximize, slot by slot, the instantaneous weighted sum-rate (WSR) under the SC-FDMA constraints, *i.e.*,

$$\Theta(\boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{\gamma}}\left[\max_{\boldsymbol{I}[n]\in\mathcal{A}} \sum_{k\in\mathcal{K}} \mu_k \sum_{j=1}^{J} i_{k,j}[n] r_{k,j}[n]\right] \tag{6.10}$$

where $r_{k,j}[n]$ is used to denote $r_k(\gamma_{k,j}^{\mathrm{eff}})[n]$. The solution $\boldsymbol{I}[n]$ of the WSR problem can be evaluated as the optimal solution of a set partitioning problem as in [85], or as a quasi-optimally solution through a canonical dual relaxation of the SC-FDMA constraints in the continuous space as in [86].

Since the dual problem is in general not tractable analytically, an iterative sub-gradient method as in [66] can be used to solve it. However, in realistic applications, the adaptive implementation is suggested, where the dual variables are updated at each time slot as:

$$\boldsymbol{\mu}[n+1] = \boldsymbol{\mu}[n] - \delta_{\boldsymbol{\mu}}(\boldsymbol{r}^*[n] - \phi\|\boldsymbol{r}^*[n]\|_1) \tag{6.11}$$

where $\boldsymbol{r}^*$=$[r_1^*, \ldots, r_K^*]$, with

$$r_k^*[n] = \sum_{j=1}^{J} i_{k,j}^*[n] r_{k,j}[n] \tag{6.12}$$

and $\delta_{\boldsymbol{\mu}}$ is a step-size suitably selected to ensure convergence [61].

By varying the proportionality constraints and, consequently, the dual geometric multiplier $\boldsymbol{\mu}^*$, the convex hull of ergodic rate region can be parameterized. Note that, since the original problem is not convex, the optimum may not lie on the convex hull and the dual problem may lead to a suboptimal solution which is the point that lies on the convex hull and is the closest to the optimum. The algorithmic complexity of this solution is mainly due to the search for the optimal patterns to be used in $\boldsymbol{I}^*[n]$. Such complexity can be highly reduced when an estimate of the average number of PRBs required to achieve the prescribed portion of rate is available as shown in the next section.

6.3.2 Estimation of the Average Amount of Allocated Resources

We derive here a linear estimate of the average number of PRBs allocated to each user, *i.e.*, $\bar{\boldsymbol{G}}^* = [\bar{G}_1^*, \ldots, \bar{G}_K^*]$, with $\bar{G}_k^* = \mathbb{E}[G_k^*]$, when the optimal rate $\boldsymbol{R}^*$ is achieved. Since $\boldsymbol{R}^*$ depends on $\boldsymbol{\phi}$, we would like to obtain $\bar{\boldsymbol{G}}^*$ as function of $\boldsymbol{\phi}$. The first step of the derivation is to build a simple local approximation of the convex hull of the rate region around $\boldsymbol{R}^*$, which can be expressed in parametric form as function of the vector $\tilde{\boldsymbol{G}} = [\tilde{G}_1, \ldots, \tilde{G}_K] \succeq \mathbf{0}$, denoting the average number of allocated PRBs. We approximate the convex hull of the rate region with the $(K-1)$-dimensional hyperplane

$$\mathcal{R}^{\text{v}} = \{\tilde{\boldsymbol{R}} : \tilde{R}_k = a_k \tilde{G}_k,\ \forall k, \|\tilde{\boldsymbol{G}}\|_1 = G\}, \tag{6.13}$$

where the constraint $\|\tilde{\boldsymbol{G}}\|_1 = G$, implying that all the resources are allocated at each slot, is usually reasonable when K is sufficiently large. As an example, in the two-user case $\mathcal{R}^{\text{v}}$ is a parametric line with parameters $\tilde{G}_1$ and $\tilde{G}_2 = G - \tilde{G}_1$, leading to $\tilde{R}_1 = a_1\tilde{G}_1$, $\tilde{R}_2 = a_2(G - \tilde{G}_1)$. To make $\mathcal{R}^{\text{v}}$ a local approximation, the two following constraints must hold:

1. the optimal rate vector $\boldsymbol{R}^*$ must belong to $\mathcal{R}^{\text{v}}$
2. $\mathcal{R}^{\text{v}}$ is tangent to the rate region at $\boldsymbol{R}^*$.

Condition 1. implies that $R_k^* = a_k G_k^*$, $\forall k \in \mathcal{K}$, which follows $a_k = R_k^*/G_k^*$. Condition 2. implies that $\mathcal{R}^{\text{v}}$ is orthogonal to $\boldsymbol{\mu}^*$, since, from convex optimization theory, it is known that the dual geometric multiplier $\boldsymbol{\mu}^*$ associated to the solution of the dual problem defines the normal vector to the convex hull of the rate region at the optimal rate vector $\boldsymbol{R}^*$ [90]. This means that

$$\mu_k^* = c\left(\frac{d\tilde{R}_k}{d\tilde{G}_k}\right)^{-1} = \frac{c}{a_k} = c\frac{G_k^*}{R_k^*}, \quad \forall k \in \mathcal{K} \tag{6.14}$$

where c is a real constant which can be derived by combining (6.14) with constraint (6.8c), as follows

$$c = \left(\sum_{k\in\mathcal{K}} \frac{\phi_k \bar{G}_k^*}{R_k^*}\right)^{-1}. \tag{6.15}$$

By considering $R_k^* = \phi_k \|\boldsymbol{R}^*\|_1$ and $\|\bar{\boldsymbol{G}}_k^*\|_1 = G$, we obtain:

$$\mu_k^* \approx \frac{R_k^*}{\bar{G}_k^*}\left(\sum_{k\in\mathcal{K}} \frac{\phi_k \bar{G}_k^*}{R_k^*}\right)^{-1} = \left(\frac{\bar{G}_k^* \|\boldsymbol{R}^*\|_1}{\phi_k \|\boldsymbol{R}^*\|_1 \|\bar{\boldsymbol{G}}_k^*\|_1}\right) = \frac{\bar{G}_k^*}{G\phi_k} \tag{6.16}$$

Thus

$$\bar{G}_k^* \approx G\phi_k\mu_k^*, \quad \forall k \in \mathcal{K}. \tag{6.17}$$

The proposed estimate of the average number of allocated PRBs depends on the dual variables μ_k. By exploiting the adaptive implementation of $\boldsymbol{\mu}$ as in (6.11), eq. (6.17) can be used to adaptively estimate the average number of allocated PRBs, *i.e.*, $\bar{G}_k[n] = \mu_k[n]\phi_k G$.

6.3.3 Proposed RRA Algorithm

The knowledge of an estimate of the average number of allocated PRBs G_k^* is here exploited to reduce the search space of optimal patterns in the WSR maximization problem in (6.10). In fact, we consider for allocation only the patterns having up to $\bar{G}_k[n]$ PRBs, leading to a sub-optimal solution of problem (6.7).

The pseudo-code of the proposed algorithm 4, which is is composed of two stages, is listed in Algorithm 1. In the first stage (lines 6-18) users are sequentially scheduled according to their best value of $\mu_k r_{k,j}$, *i.e.*, with $r_{k,j}$ obtained among the not yet allocated patterns of $\tilde{G}_k$ PRBs, where $\tilde{G}_k$ is not larger than the estimate in (6.17). The initial value of $\tilde{G}_k$ is initialized to $\bar{G}_k[n] = \lceil \mu_k[n]\phi_k G \rceil$, where the Ceil operator projects the estimate on the discrete space $\mathcal{G}$ and ensures that $\sum_k \bar{G}_k[n] \geq G$.

After each pattern allocation, $\tilde{G}_k$ is reduced or preserved in order to consider only set of patterns that have no PRBs in commons with the ones already allocated. This is performed by evaluating the maximum number of still allocable adjacent PRBs (lines 15-16), thereby ensuring contiguous PRB allocation. Note that the operations in lines 15 and 16 can be performed through a look-up table, which can be easily built off-line. Since each search is performed in the pattern set $\mathcal{J}_{\tilde{G}_k}$ with cardinality $G - \tilde{G}_k + 1$, the worst-case complexity of this stage which corresponds to the case $\lceil \mu_k[n]\phi_k G \rceil = 1/G$, is $O(GK)$. In case the final values $\tilde{G}_k$ are such that $Q = \sum_k \tilde{G}_k < G$, there may still be a maximum of $G - Q$ unallocated PRBs. An increase of the objective, *i.e.*, the sum-rate, can be still obtained by allocating these PRBs. The second stage (lines 21-28) tries to allocate each not yet allocated PRB to users having already allocated neighbor PRBs, according to the maximum increase in marginal weighted rate. The complexity of this second stage is $O(KQ)$, where $Q \ll G$. Therefore, the overall algorithm complexity increases only linearly with the number of users and the number of resources.

Algorithm 4 Sub-optimal algorithm to solve problem (6.7)

Input $\boldsymbol{\phi}$, $\boldsymbol{\gamma}$
Initialize $\boldsymbol{\mu}[0]$
for all n **do**
  $\mu_k[n] \leftarrow \mu_k[n]/(\boldsymbol{\mu}^{\mathrm{T}}[n]\boldsymbol{\phi})$, $\forall\ k \in \mathcal{K}$;
  $\mathcal{J}' = \mathcal{J}$, $\mathcal{G}' = \mathcal{G}$, $\mathcal{K}' = \mathcal{K}$, $G^{\max} = G$;
  repeat
    for all $k \in \mathcal{K}'$ **do**
      $\tilde{G}_k \leftarrow \min(\lceil \mu_k[n]\phi_k G \rceil, G^{\max})$, $\forall\ k \in \mathcal{K}$;
      $\tilde{j}_k \leftarrow \underset{j \in \mathcal{J}_{\tilde{G}_k}}{\operatorname{argmax}}\ r_k(\gamma^{\mathrm{eff}}_{k,j}[n])$;
    end for
    $k^* \leftarrow \underset{k \in \mathcal{K}'}{\operatorname{argmax}}\ \mu_k r_k(\gamma^{\mathrm{eff}}_{k,\tilde{j}}[n])$;
    $j^*_{k^*} = \tilde{j}_{k^*}$;
    $\mathcal{K}' \leftarrow \mathcal{K}' \setminus \{k^*\}$;
    $\mathcal{G}' \leftarrow \mathcal{G}' \setminus \mathcal{G}_{j^*_{k^*}}$;
    $\mathcal{J}' \leftarrow \mathcal{J}' \setminus \{j : \mathcal{G}_{j^*_{k^*}} \cap \mathcal{G}_j \neq \{0\}\}$
    $G^{\max} \leftarrow \underset{g}{\operatorname{argmax}}\{g : \mathcal{J}' \cap \mathcal{J}_g \neq \{0\}\}$
    $g^{\min}_{k^*} \leftarrow \min_{\mathcal{G}_{j^*_{k^*}}} g$, $g^{\max}_{k^*} \leftarrow \max_{\mathcal{G}_{j^*_{k^*}}} g$
  until $\mathcal{G}' \neq \{0\} \vee \mathcal{K}' \neq \{0\}$
  if $\mathcal{G}' \neq \{0\}$ **then**
    define $\delta_j := G - G_j$
    for all $g \in \mathcal{G}'$ **do**
      $k^{\mathrm{t}} = \underset{g^{\min}_k > g}{\operatorname{argmin}}\ g^{\min}_k$
      $k^{\mathrm{b}} = \underset{g^{\min}_k < g}{\operatorname{argmax}}\ g^{\max}_k$
      define $u_k := 1$ if $k = k^{\mathrm{b}}$, 0 otherwise;
      $\Delta r_k = \mu_k[n](r_{k,j^*_k+\delta_{j^*_k}+u_k}[n] - r_{k,j^*_k}[n]), k = k^{\mathrm{t}}, k^{\mathrm{b}}$
      $k^* = \operatorname{argmax}(\Delta r_{k^{\mathrm{t}}}, \Delta r_{k^{\mathrm{b}}})$
      $j^*_{k^*} \leftarrow j^*_{k^*} + \delta_{j^*_k} + u_{k^*}$
    end for
  end if
  compute $\tilde{\boldsymbol{r}}[n]$ and update $\boldsymbol{\mu}[n+1]$ according to (6.3) and (6.11)
end for

6.3.4 Comparative Study

In this section we first evaluate and compare the performance of the proposed algorithm in realistic scenario through simulation. We consider an uplink SC-FDMA system with subcarrier spacing equal to 15 kHz and time slot duration equal to 10 ms. We investigate the performance of the proposed solution by varying the number of users, which are uniformly distributed in a cell with radius equal to 500 m, and the values of total bandwidth, *i.e.*, $B = 1.75, 3, 5$ MHz resulting in $G = 6, 15, 25$ allocable PRBs, respectively. The user power budget is set to 23 dBm. Each user is affected by fast fading which is obtained from a multi-path channel model with delay spread of 2.3 µs and Doppler bandwidth of 6 Hz. In the evaluation of the assigned rates an SNR gap of 3dB is taken into account.

We first validate the estimate in (eq. (6.17)) by running Monte Carlo simulation with different number of users and $B = 3$ MHz. For simplicity, the weights $\boldsymbol{\mu}$ are fixed and the values of $\boldsymbol{\phi}$ are then evaluated. The built-in *bintprog* matlab procedure is used to solve the WSR problem (6.10) at each slot [85]. Fig. 6.3 reports two examples of the results obtained when the weights μ_k of each user are selected as: (i) inversely proportional to its long term channel capacity (Fig. 6.3(a)), and (ii) randomly with uniform distribution (Fig. 6.3(b)). We can observe that in the first case (Fig. 6.3(a)), the estimation is highly accurate due to the fact that the resulting share of rate among users is quite balanced. In the second case the final share of throughput is generally unfair, with few users having large $\bar{G}_k$. Nevertheless, the estimate is still quite effective.

We next compare the performance of the proposed linear resources estimate (LRE) algorithm with (i) the optimal solution (OPT) obtained by using the built-in Matlab function *bintprog* [85], which performs a branch & bound search procedure and a LP relaxation at each iteration to solve (6.10), and (ii) a quasi-optimal solution obtained through the canonical dual method (CDM) proposed in [86], to solve (6.10). The results presented here refers to $\phi_k = 1/K, \forall k \in \mathcal{K}$. Table 6.1 shows for each algorithm the complexity order and the typical number of iterations required to converge in the test-case of 10 users and 25 PRBs. The *bintprog* procedure achieves an optimal solution within $I_{\text{mjr}} \approx 2$ major iterations of branch & bound search at which corresponds $I_{\text{mnr}} \approx 40$ LP relaxations with complexity $O((G+K)^2)$. The sub-gradient based iterative approach used in the CDM strategy requires the computation of KJ, K and G variables in the vectors, $\boldsymbol{\rho}, \boldsymbol{\lambda}$ and $\boldsymbol{\epsilon}$, each one requiring $I_{\boldsymbol{\rho}}, I_{\boldsymbol{\lambda}}$ and $I_{\boldsymbol{\epsilon}}$ iterations, respectively. The number of iterations highly depends on the choice of the initial variables. In our numerical evaluations the overall complexity of the CDM approach is generally comparable to that of

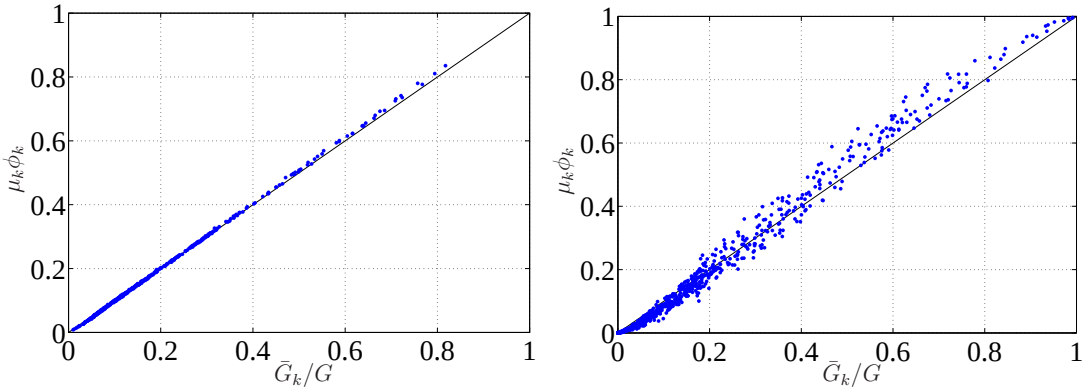


Figure 6.3: Scatter plots showing the empirical correlation between $\bar{G}_k/G$ and $\mu_k\phi_k$. μ_k is selected as the inverse of the achievable long term user capacity (a), and randomly with uniform distribution (b), considering G=15 PRBs. Pearson correlation coefficient is 0.9999 (a), and 0.9945 (b).

Strategy	Complexity	Typical value (K=10, G=25)
OPT	$O(I_{\text{mjr}}I_{\text{mnr}}(K+G)^2)$	$\sim 10^4$
CDM	$O(I_{\boldsymbol{\rho}}K[(G^2+G)/2+1]+I_{\boldsymbol{\epsilon}}G+I_{\boldsymbol{\lambda}}K)$	$\sim 10^5$
LRE	$O(GK)$	200

Table 6.1: Complexity Comparison.

the optimal procedure, even when the initial values are chosen as the optimal values computed at the previous slot. The linear complexity of our proposed algorithm greatly reduces the number of required iterations. Fig. 6.4 provides a performance comparison in terms of the average sum-rate, by varying the number of available PRBs with $K = 10$ users (left histogram) and by varying the number of users with $G = 15$ PRBs (right histogram). The fairness is evaluated through the Jain index [91], which is always greater than 0.99 in all investigated cases. The CDM algorithm approaches the performance of the optimal solution with a loss always less than 1%. The performance loss of the proposed LRE algorithm ranges between 9% and 11%, thus offering an attractive trade-off between computational complexity and average throughput in practical QoS aware applications.

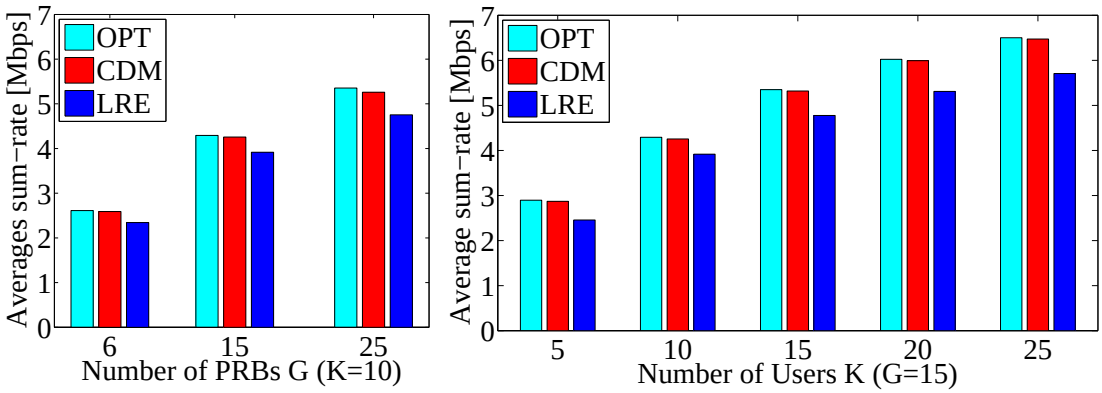


Figure 6.4: Average sum-rate resulting from the proposed LRE algorithm, CDM-based algorithm and the optimal (OPT) solution, by varying the number of available PRBs with K=10 users (left histogram) and by varying the number of users with G=15 PRBs (right histogram).

6.4 APP Layer: Video Coding and Adaptation

For the m-health application addressed in this chapter, we adopt the GOP encoding format IBPBPBPP and exploit temporal and SNR scalability with a fixed spatial resolution. More precisely, for the high quality ultrasound video temporal scalability is allowed with two available temporal decimations, whereas only SNR scalability is allowed for ambient videos. We here focus on MGS scalability using one enhancement layers and optimized bit-stream extraction (see section 2.1.3 for further details). RD Rate adaptation techniques dynamically adapt the amount of transmitted information to the available channel bandwidth by taking into account the content of the videos and its impact on the end user quality. Rate-distortion (R-D) models enable to predict the minimum bit rate required (in bit/s, or bps) to achieve a target distortion, defined in terms of Mean Square Error (MSE). In this chapter we consider the semi-analytical R-D model proposed in eq. (3.9).

We assume that the ambulance equipment negotiates with the LTE access network a guaranteed bit rate $\bar{R}_0$ to support the emergency m-health services. Such value $\bar{R}_0$ is derived by defining the value of ϕ_0 in problem (6.7) and eventually updated every W seconds in case of critical cell-load conditions or bad channel conditions for the ambulance. The m-health user exploits this guaranteed bandwidth to deliver the best video quality according to priority and fairness constraints defined for the different videos. This is obtained through a dynamic rate adaptation

strategy, consisting in maximizing the overall video quality while minimizing a weighted quality difference among the different videos under minimum and maximum rate constraints. This strategy extends the one proposed in chapter 4 by addressing the problem for both empirical and semi-analytical R-D model with two-parameters and then extended in chapter 5 with three-parameters models, but without considering weighted quality difference. As shown in section 4.2.1, when the parametric R-D model is sufficiently accurate, it can be used to relax the multi-objective optimization problem leading to a much simpler constraint satisfaction problem. Here, we follow this approach by defining an adaptation strategy which derives the transmission rates as the solution of the following set of equations and constraints:

$$\sum_{v\in\mathcal{V}} HF_v = \bar{R}_0, \tag{6.18a}$$

$$\Delta(D_i, D_j; w_i, w_j) = 0 \qquad \forall i, j \in \mathcal{V}, i > j \tag{6.18b}$$

$$F_v^{\min} < F_v < F_v^{\max} \qquad \forall v \in \mathcal{V} \tag{6.18c}$$

where $\mathcal{V}$ is the set of videos handled by the ambulance for e-health emergency services, H is the estimated overhead introduced at the different layers of the network architecture, w_v, $v \in \mathcal{V}$ are the weights used to account for the different priorities mentioned in the introduction and $\Delta(D_i, D_j; w_i, w_j)$ is the extended distortion-fairness metric for each pair of videos, defined as

$$\Delta(D_i, D_j; w_i, w_j) = \begin{cases} 0 & \text{if } d_i = D_i^{\min} \wedge d_j < d_i \\ 0 & \text{if } d_j = D_j^{\min} \wedge d_i < d_j \\ 0 & \text{if } d_i = D_i^{\max} \wedge d_j > d_i \\ 0 & \text{if } d_j = D_j^{\max} \wedge d_i > d_j \\ |w_i D_i - w_j D_j| & \text{otherwise.} \end{cases} \tag{6.19}$$

The algorithm 3 proposed in section 5.5 to solve the similar problem 5.31 can be suitable extended to solve the problem considered here. Recall that $x_v, y_v \in \{0, 1\}$, $v \in \mathcal{V}$, are binary variables that indicate whether (1) or not (0) the two constraints $F_v > F_v^{\min}$ and $F_v < F_v^{\max}$ are active for the video v. We then can then extend the function defined in eq.5.32 and 5.33 as

$$\Gamma(\boldsymbol{x}, \boldsymbol{y}, D) = \sum_{v\in\mathcal{V}} x_v y_v \left(\frac{\alpha_v}{w_v D + \xi_v} + \beta_v \right) - \Lambda(\boldsymbol{x}, \boldsymbol{y}) \tag{6.20}$$

where

$$\Lambda(\boldsymbol{x}, \boldsymbol{y}) = \frac{\bar{R}_0}{H} - \sum_{v \in \mathcal{V}} \left[(1 - x_v) F_v^{\min} - (1 - y_v) F_v^{\max} \right] \tag{6.21}$$

Algorithm 3 in page 51 can be then used to optimally solve problem (6.18).

6.5 Numerical Results

In our simulations we consider an LTE-like access network with subcarrier spacing $\Delta B = 15$ kHz and system bandwidth set to 5 MHz, resulting in G=25 allocable PRBs. A total of K =20 users with a maximum per-user power budget of 23 dBm are uniformly distributed in a cell, resulting in an average SNR ranging from 5 to 28 dB. More specifically the ambulance user experiences an average SNR of 13 dB and receives from the LTE access network a GBR with values negotiated in the range from 2.5 to 7 Mbps, which are obtained by varying the related value $\phi_0 \in [0.2, 1]$ in problem (6.7). The other users are assumed best-effort, *i.e.*, with

$$\phi_k = \frac{1 - \phi_0}{K - 1}, \quad \forall k \in \mathcal{K}_2 \tag{6.22}$$

The radio channel for all users is modeled according to the ITU extended vehicular A model, with a Doppler frequency of 70Hz.

The ambulance sends N=2 raw ambient videos consisting of 300 frames with resolution 640x360 and one raw ultrasound sequence of 150 frames with resolution 640x480. Each video is acquired with a frame-rate of 30 fps and is looped until a sequence of 50 seconds is obtained. The video sequences are encoded with the JSVM reference software [30] with one base layer and two enhancement layers. Each enhancement layer is split into five MGS layers with vector distribution [3 2 4 2 5]. The GOP size and the IDR period are set to 8 and to 32 frames, respectively. After encoding, the resulting quality in terms of the average PSNR, ranges from 29 to 35 dB for the ambient videos and from 32 to 40 dB for the ultrasound video. The three parameters of model (3.9) are evaluated for each IDR period, resulting in adaptation interval of about 1 sec. The video distortion weights w_v of the ambient videos are set to 1, whereas the weight of the ultrasound video is set to 2. Finally, video playout deadline at the receiver is set to 200 ms and the overhead factor is set to $H = 1$.

Figure 6.5 shows the PSNR at the receiver averaged over each adaptation interval for the three video sequences when a GBR $\bar{R}_0 = 4.5$ Mbps is provided to the ambulance. We can note how the resulting qualities closely follow the selected

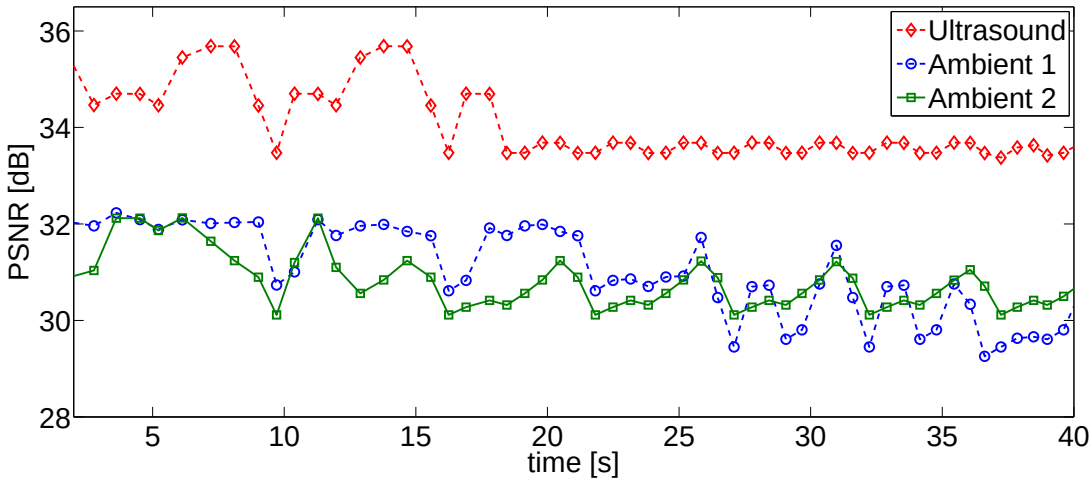


Figure 6.5: Average PSNR received at the hospital when a GBR $\bar{R}_0 = 4.5$ Mbps is guaranteed to the Ambulance.

video quality priorities providing a PSNR difference of approximately 3 dB between ultrasound and ambient videos. Fluctuations on the PSNR are mainly due to the different spatial and temporal complexity of the scenes composing the video sequences. Such behaviour can also be appreciated in Figure 6.6 where the average PSNR of each video sequence is plotted against the different GBRs granted to the ambulance. By looking also at Table 6.2, where the average rates provided to ambulance and to the other best-effort users are reported for different $\bar{R}_0$ settings, we can note that our strategy adjusts the quality of each video in a proportional way by reshaping the source rate, thus keeping a reasonable throughput also for the best-effort users. On the other hand, an high GBR requirement, *i.e.*, 7 Mbps, allows to transmit the highest enhancement layer for most of the time, but a large part of the physical resources is drained by the ambulance, thereby starving the best-effort users.

Finally, in Figures 6.7 and 6.8 we report a few examples of received frames, as they result from the joint adaptation process. Frames in Fig. 6.7 are extracted from to the ultrasound video sequence, while frames Fig. 6.8 from the ambient video 1. Comparing the received frames with the original ones, we note that the proposed adaptation strategy is capable to achieve good visual results even when the overall available throughput for the m-health user is low ($\bar{R}_0 = 3.3$ Mbps). At

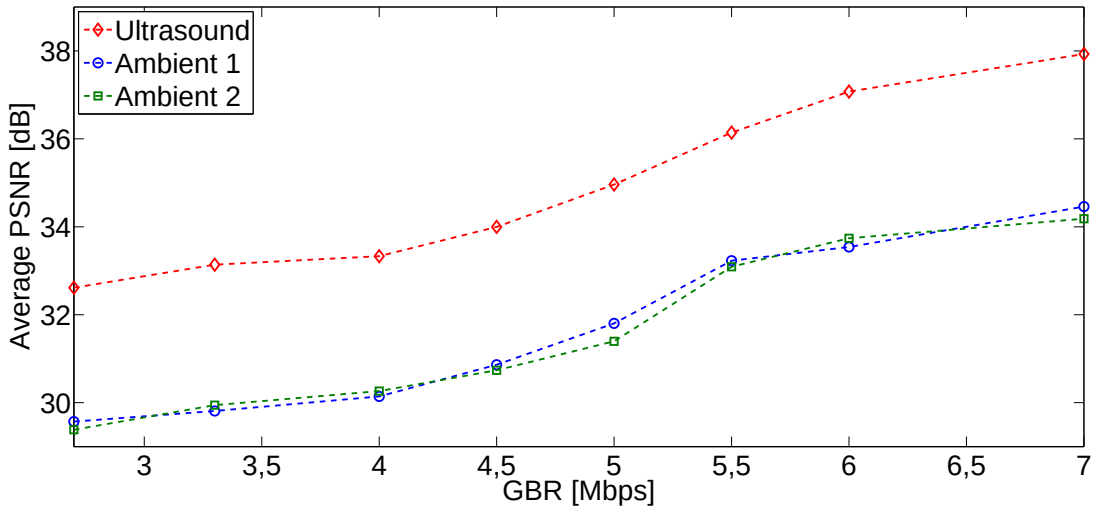
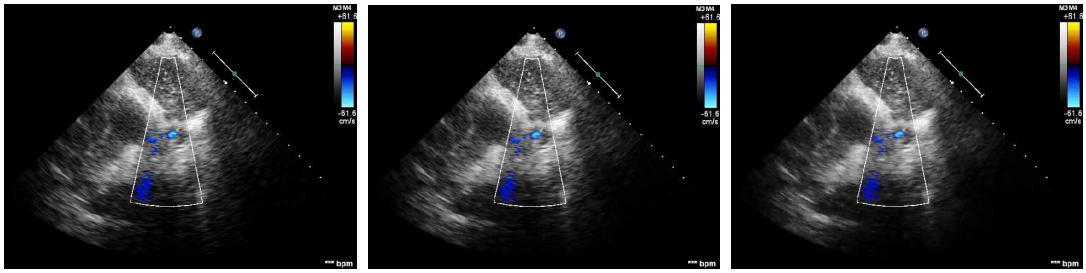


Figure 6.6: Average PSNR of the video sequences received at the hospital for different values of GBR granted to the ambulance.

$\bar{R}_0$ [Mbps] (guaranteed)	2.5	3	4	4.5	5	6	7
R_0 [Mbps] (achieved)	2.7	3.3	4.2	4.5	5	6	6.8
R_k, $k \in \mathcal{K}_2$ [Mbps]	1.35	1.21	0.93	0.81	0.67	0.27	0.08

Table 6.2: Average rate R_0 provided to m-healt user and average rate R_k, $k \in \mathcal{K}_2$ provided to the best-effort users, for different values of GBR $\bar{R}_0$ settings.

the same time, when the amount of available resources is higher ($\bar{R}_0 = 7$ Mbps), the optimization strategy adaptively increases the final quality of all the videos. We conclude by observing that in all cases, the diagnostic ultrasound sequence is transmitted with higher quality with respect to ambient videos, according to the weights used in (6.18), enabling effective tele-diagnosis services.



(a) (b) (c)

Figure 6.7: Examples of original and received frames for the echo sequence with different GBR settings for the m-health user: original frame (a), $\bar{R}_0 = 7$Mbps. PSNR= 39.46dB (b) and $\bar{R}_0 = 3.3$Mbps. PSNR= 34.70dB (c).

(a) (b) (c)

Figure 6.8: Examples of original and received frames for on sequence with different GBR settings for the m-health user: original frame (a),$\bar{R}_0 = 7$ Mbps. PSNR= 35.93 dB.. (b) and $\bar{R}_0 = 3.3$Mbps. PSNR= 29.58dB (c).

Chapter 7

Cross-layer Optimization for HTTP Adaptive Streaming in LTE Networks

In the previous chapters, we have proposed a complete framework to fully optimize the video delivery over IP/UDP/RTP protocol stacks in both downlink and uplink, *i.e.*, OFDMA and SC-FDMA, systems of the next generation wireless systems. We have assumed the possibility to handle both the MAC layer and and the APP layer *i.e.*, by performing enhanced RRA and source rate adaptation/control, respectively. The transmissions based on RTP have the main drawback that requires dedicated servers and pass through a port that is often blocked by firewall and NAT. For this reasons most of the video traffic is now transmitted over HTTP protocol, which is NAT transparent, and may exploit the large deployments of cache and content distributed networks (CDN).

A new approach referred to as HTTP adaptive streaming (HAS) [3] is becoming popular. HAS is adaptive in the sense that it allows a client to adaptively switch between multiple bit-rates, depending on the bandwidth or data rate available between the server and the client. This is a particularly useful feature for a wireless environment since the data rate of the wireless link can vary over time. Based on TCP, one of the objectives of HAS is keeping the fairness among multiple homogeneous/heterogeneous connections in the network. In fact, fair share of network resources among multiple heterogeneous connections is one of key issues especially for the commercial use of the Internet [92].

On the MAC layer side, optimized RRA was performed by assuming the complete and perfect knowledge of the CSI of each user. Nevertheless, in order to limit the large feedback required by such information, LTE allows only a limited

feedback, that is the sub-band with the highest CSI and its index. In such case, RRA scheduler is often based on a proportional fair rule [93, 94] with GBR and MBR contraints, which basically allows only to optimize the GBR values according to the type of services and traffic load. In this chapter, we then extended the multi-user cross-layer proposed solution presented in chapter 5 to cope with HAS and LTE RRA constraints.

In literature, multi-user HAS video delivery optimization has attracted increasingly attention in the last few years. In [95], an overview of the recently standardized quality metrics for HAS and an end-to-end evaluation study are presented. They concluded that network-level and radio-level adaptation is required for enhancing service capacity and user perceived quality. Recently, Authors in [93] propose an efficient method to optimally and adaptively set up the GBR of each video flow in a LTE network with heterogeneous traffic. The approach is intended to achieve a level of fairness among the video flows while preventing starvation of other data flows. However, the definition of the utilities is not content-aware and may not lead to the best possible quality fairness among the video flows. To the best of our knowledge only [96] investigated an optimized content-aware multi-user HAS video delivery framework in LTE networks. Similarly to here, a media-aware network element (MANE) is in charge of selecting the streaming rate required by each client in order to maximize the aggregate video utilities under resource constraint. However, differently from here, video quality fairness is not considered and the peculiarities of pull-based delivery strategy of HAS technology are not taken into account.

We here propose a quality-fair adaptive streaming (QFAS) solution to deliver fair video quality to HAS clients competing for the same resources in an LTE cell. Using a similar mechanism as [93], the proposed QFAS solution brings intelligence into the network to adaptively select the prescribed GBR of each UE according to the contents characteristics in addition to the channel condition. Such GBR values are derived by solving an optimization problem aimed at maximizing the aggregate video utility under minimum and maximum rate constraints, available resource, and quality-fair constraint across multiple video clients. Numerical evaluations resulting from extensive and detailed *ns2* simulations show that QFAS solution provides significant improvement to the quality received by the end-users demanding more complex video, even when they are experiencing bad channel condition, with a tolerable degradation of the other low-complexity videos. The quality fairness is thus well improved among heterogeneous clients compared to best effort and AGBR approaches.

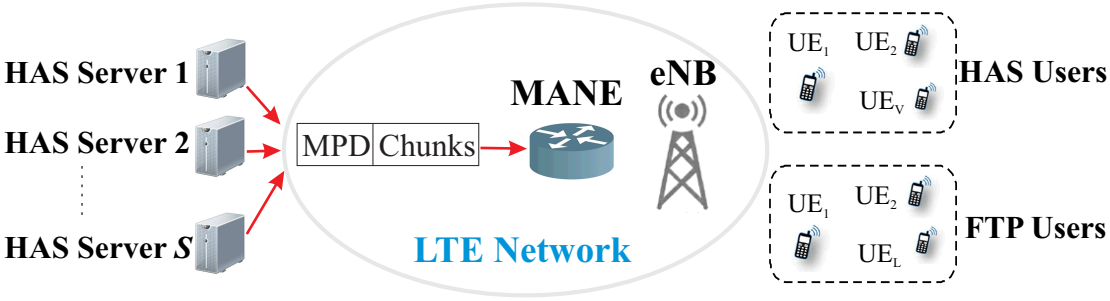


Figure 7.1: System Architecture.

7.1 System Model and Assumptions

As depicted in Fig. 7.1, we consider an LTE wireless access network serving a total of K UEs, subdivided in V HAS users indexed by the set $\mathcal{V} = \{1, \ldots, V\}$, and L data users. One or more HAS servers encode the video sequences at multiple bit-rates and, after segmentation, generate a manifest file, also named media presentation descriptor (MPD). We assume that each HAS server extracts synthetic quality information from each segment (also called chunk in the following) and inserts them in the MPD. A MANE, located close to the e-NodeB (eNB), is able to intercept and process the MPD requested by each HAS client in order to get rate and quality information. The eNB allocates the available resources according to a general proportional fair scheduler with minimum rate, *i.e.*, GBR, constraints, which are dynamically updated by the MANE.

In the following we omit the index of the client and we details the adaptation process for a single client-server link in an ideal scenario as illustrated in Fig. 7.2. Let $\mathcal{R}$ be the set of M available rate profiles $r_m, m = 1, \ldots, M$, listed in the MPD and assume that the client is able to follow the GBR provided by the eNB. This means that once a chunk, *i.e.*, chunk $(n-1)$ in Fig. 7.2, is received, the rate decision algorithm (RDA) at the client completes the measurement of the chunk download rate $\hat{R}[n-1]$ and requests chunk n with a profile rate

$$r^*[n] = \max_{r_m \leq \hat{R}[n-1]} r_m, \tag{7.1}$$

according to a pull-based approach.

When the MANE intercepts the request, it collects the channel state information (CSI) from the eNB and updates the GBR value $R[n]$ that the scheduler at the eNB will use to send chunk n. In case of ideal rate measurement, which requires that the channel state information do not vary significantly between time

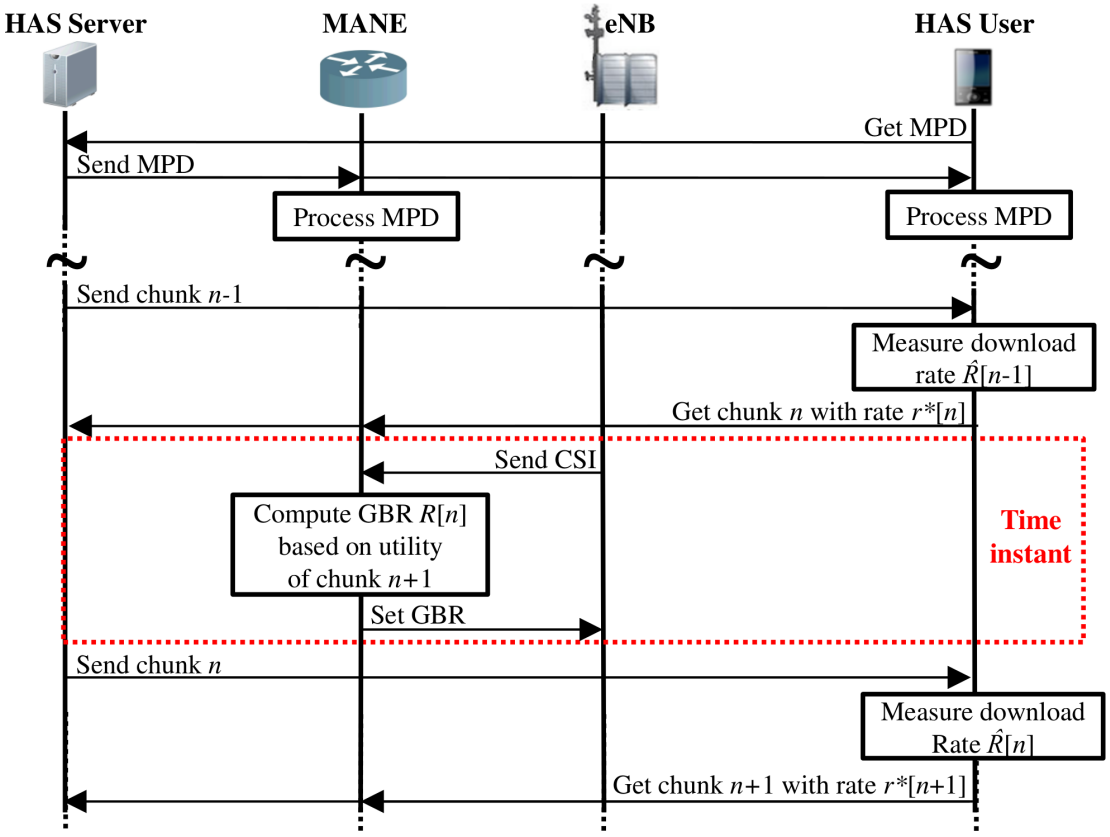


Figure 7.2: Proposed approach for a single HAS video delivery optimization.

instant n and time instant $(n+1)$, we would have $R[n] = \hat{R}[n]$, *i.e.*, the GBR value $R[n]$ is then used for the rate request of chunk $(n+1)$. Thus, to avoid mismatch due to video characteristic changing over time, the GBR value $R[n]$ is computed based on the video utility of the chunk $(n+1)$. With this approach, the MANE only acts as pre-scheduler that dynamically selects the feasible guaranteed minimum rate, transparently to the actual physical LTE scheduler. More details are provided in the next Section.

7.2 Optimization Problem and Solutions

The objective of our quality-based approach is to derive the rate which allows to maximize the overall video quality under quality fairness constraint and according to users channel condition. Let n be the chunk index requested by user k, we define U_k as the *utility* of requesting chunk $(n+1)$ in terms of video quality metric. The

following parametric rate-utility model is used to describe the evolution of the utility U_k as a function of the rate R_k:

$$U_k = f\left(\mathbf{a}_k, R_k\right), \tag{7.2}$$

where $\mathbf{a}_k \in \mathcal{A} \subset \mathbb{R}^{N_a}$ is a time-varying and program-dependent parameter vector. For all values of $\mathbf{a}$ belonging to the set of admissible parameter values $\mathcal{A}$, $f\left(\mathbf{a}_k, R\right)$ is assumed to be a continuous, invertible and strictly increasing function of R. The model (7.2) may represent the variation of the SNR, the PSNR, the SSIM (see section 3.6), or any other strictly increasing quality metric as function of the encoding rate [97].

Following the approach in [93,96], we consider a simplified air interface model where the maximum achievable rate for each UE is estimated according to its average channel condition. Let γ_k be the average signal-to-noise plus interference ratio (SNIR) experienced by UE k. As in [93], we define $w_k = [\log_2(1+\gamma_k)]^{-1}$ as the inverse of the estimated average rate per unit bandwidth. The optimization problem is then stated as follows:

$$\max \sum_{k \in \mathcal{V}} f\left(\mathbf{a}_k, R_k\right) \tag{7.3a}$$

$$s.t.\ A_k \leq R_k \leq B_k, \forall k \in \mathcal{V} \tag{7.3b}$$

$$\sum_{k \in \mathcal{V}} w_k R_k \leq \Pi \tag{7.3c}$$

$$\Delta(U_i, U_j) = 0 \quad \forall i, j \in \mathcal{V}, i \neq j \tag{7.3d}$$

where A_k, B_k are the minimum and maximum rates from the MPD of the video requested by UE k. The value of Π defines the amount of resources dedicated to the HAS UEs, which can be statically configured or dynamically computed at each time transmission interval (TTI) based on number of users and scaling factors [93].

The utility-fairness metric in the constraint (7.3d) is defined as:

$$\Delta(U_i, U_j) = \begin{cases} 0 & \text{if } U_i = f(\mathbf{a}_i, A_i) \wedge U_j < U_i \\ 0 & \text{if } U_j = f(\mathbf{a}_j, A_j) \wedge U_i < U_j \\ 0 & \text{if } U_i = f(\mathbf{a}_i, B_i) \wedge U_j > U_i \\ 0 & \text{if } U_j = f(\mathbf{a}_j, B_j) \wedge U_i > U_j \\ |U_i - U_j| & \text{otherwise.} \end{cases} \tag{7.4}$$

The metric, introduced in eq. (4.2) in terms of video distortion extends the simple fairness metric $|U_i - U_j|$ to the case where R_i, R_j are constrained to their minimum

and maximum values. The motivation behind such formulation are similar to the one mentioned for the video distortion and can be explained as follows: Ideal fairness among the utility values assigned to the multiple HAS users would require $U_i = U_j, \forall i \neq j$, if the utilities were not constrained to their maximum or minimum rate values. This may not be guaranteed due to the constraints on the minimum and the maximum rates which are different for each flow. In fact, in presence of rate constraints, if a video achieves its maximum utility, it is reasonable to use the available resources to increase the utilities of other videos. On the other hand, in a case of scarce resources, if decreasing the rate of the i-th video is not possible since its minimum utility value has been already reached, it is necessary to decrease the rate of the other videos, at the price of decreasing the related utility.

The optimization problem in (7.3) admits a feasible solution under the condition $\sum_{k\in\mathcal{V}} w_k A_k \leq \Pi$. By considering the trivial condition $\sum_{k\in\mathcal{V}} w_k B_k \geq \Pi$, we have already show in chapter 4 that the problem (7.3) collapses in a constraint-satisfaction problem where the objective is achieved by fulfilling constraint (7.3c) with an equality constraint. Optimal solution can be derived by relaxing constraint (7.3b) with two boolean variables and applying a procedure with quadratic complexity in the worst case. More specifically, we can re-formulate the function presented in eq. 5.32 , based on the utility-to-rate model as:

$$\mathit{\Gamma}(\boldsymbol{x}, \boldsymbol{y}, U) = \sum_{k\in\mathcal{V}} x_k y_k w_k f^{-1}(\mathbf{a}_k, U) - \mathit{\Pi}(\boldsymbol{x}, \boldsymbol{y}) \tag{7.5}$$

where

$$\mathit{\Pi}(\boldsymbol{x}, \boldsymbol{y}) = \Pi - \sum_{k\in\mathcal{V}} w_k \big[(1-x_k)A_k + (1-y_k)B_k\big], \tag{7.6}$$

and f^{-1} is the inverse function of f. Since $f(\boldsymbol{a}, R)$ is a continuous and strictly increasing function of R, $f^{-1}(\boldsymbol{a}, U)$ is continuous and strictly increasing function of U.

The algorithm is then the same as the one proposed in Algorithm 3, where $\Pi = \Lambda/H$ and $\mu_k = w_k$, $A_k = F_k^{\min}$, $B_k = F_k^{\max}$, $\forall k$ and the solutions are $\tilde{R}_k = \tilde{F}_k$. The outcoming solutions $\tilde{R}_k$ of such algorithm are send to the eNB as GBR contraints and will be set equal to the Maximum Bit-Rate (MBR) constraints.

7.3 Numerical Results

We consider 4 video sequences extracted from real time programs, *i.e.*, *Interview*, *Sport*, *Bunny* and *Home*, in 4CIF format at 30 fps (see Table 3.1. The sequences

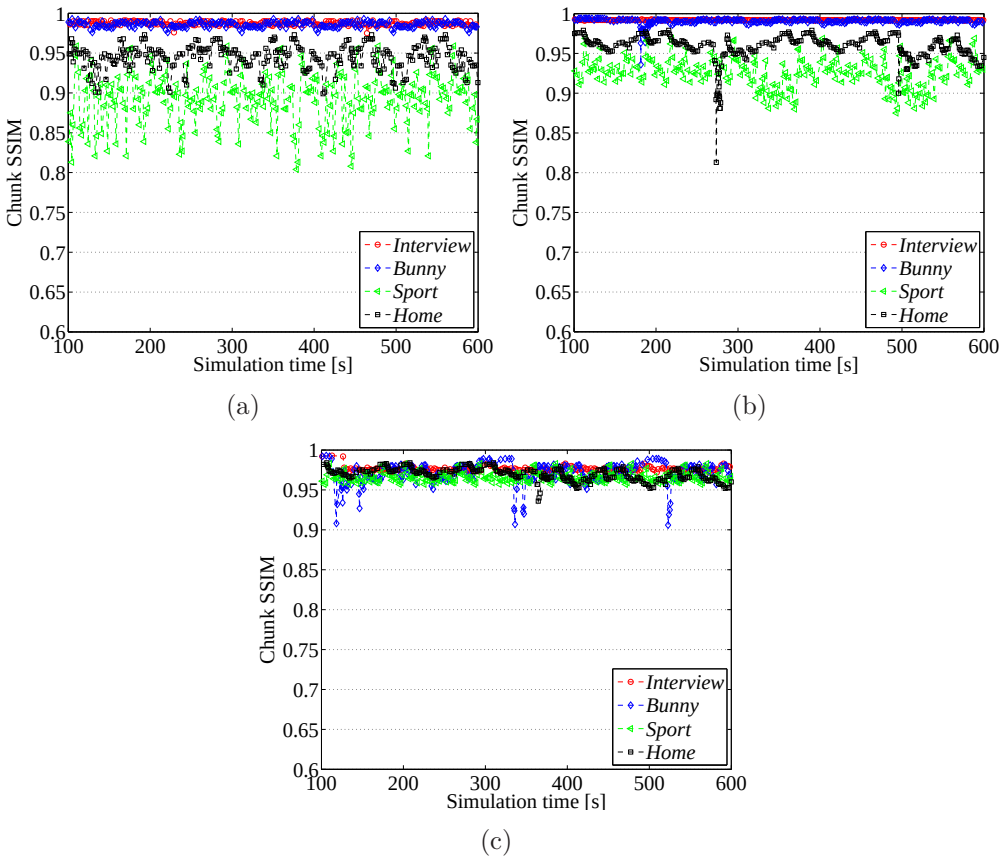


Figure 7.3: Chunk-by-chunk SSIM at the client for scenario (A) resulting from BE (a), AGBR (b) and the proposed QFAS (c).

are looped 10 times and encoded through DASH Encoder [98] with 10 profiles with rate ranging from 150 *kbps* to 5 *Mbps*. Chunk duration is set to 2 seconds.

We have considered both PSNR and SSIM metric [45] to assess the video quality. Due to the lack of space, we here provide results only in terms of SSIM. Specifically, to model the dependency between the utility (here SSIM) and the rate, we consider a logarithmic SSIM to rate continuous utility in the interval of interest $[A_k, B_k]$ proposed in section 3.6.

Simulations are carried out on the *ns2*-platform which includes HAS servers and clients, LTE radio interface and radio resource management as well as the different protocol layers (TCP/IP, PDCP and RLC). Specifically, we consider a single cell where a total of 20 UEs (L=16 FTP UEs and $V = 4$ HAS UEs) are uniformly distributed in one cell with a radius of 1 km.

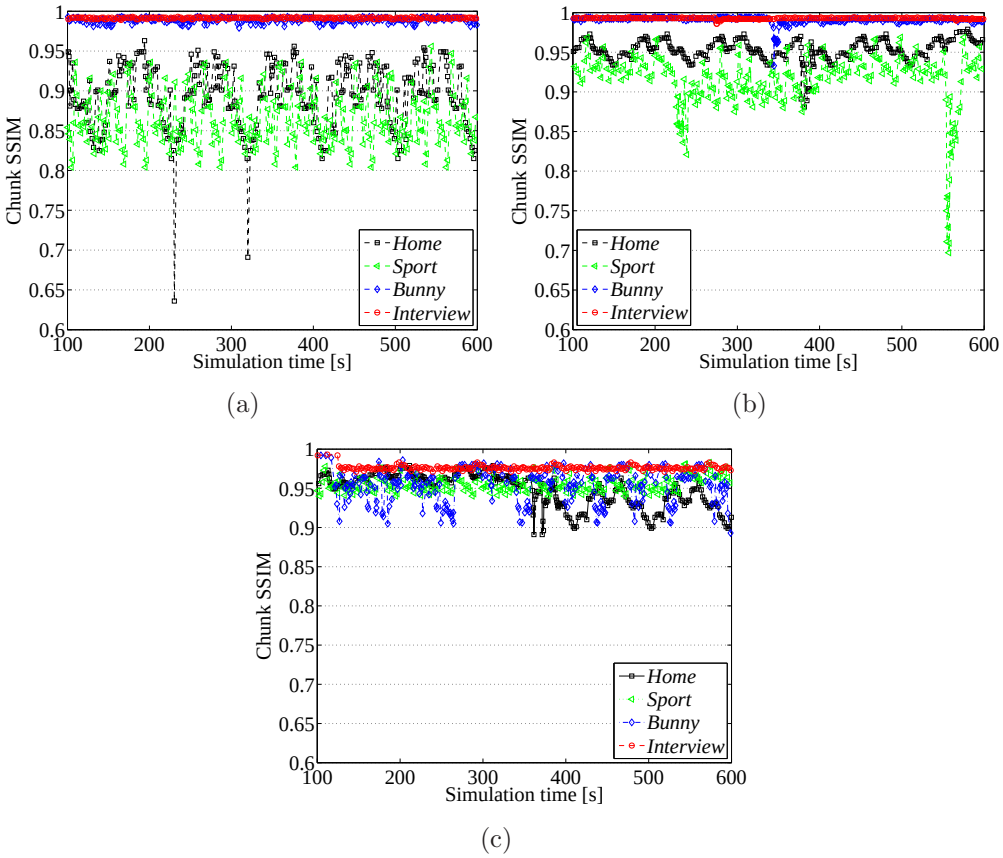


Figure 7.4: Chunk-by-chunk SSIM at the client for scenario (B) resulting from BE (a), AGBR (b) and the proposed QFAS (c).

The radio channel is modeled according to the ITU extended pedestrian A model [99] and users are also affected by log-normal shadowing (std. deviation: 8 dB) with an exponential auto-correlation (correlation distance: 100 m). Each HAS client requests one of the video sequences mentioned above. The amount of available resources Π dedicated to HAS UEs is derived according to the on-line implementation proposed in [93]. A maximum receiver buffer of 40 s is considered allowing to absorb the possible mismatch between the rate at which the chunk is encoded and the rate actually resulting after transmission.

In order to better assess the goodness of the proposed framework, we investigate two different scenarios; in the first scenario (A) UEs requesting high complex videos, *i.e.*, *Home* and *Sport*, are in good channel conditions while in the second (B) such UEs experience bad channel conditions.

Scen.	Average Rate [Mbps]				Average SNIR [dB]			
	Traffic	BE	AGBR	QFAS	*Bunny*	*Home*	*Interv.*	*Sport*
A	FTP	1.19	0.93	0.94	8.1	16.8	10.2	19.6
	HAS	0.76	1.11	1.26				
B	FTP	1.13	0.94	0.97	16.8	8.1	19.6	10.2
	HAS	0.73	1.11	0.95				

Table 7.1: Average SNIRs in dB of HAS clients and resulting average rates for FTP and HAS users.

We compare the proposed strategy with the two following approaches: (i) best effort (BE), where all UEs are non-GBR (QCI equal to 9) [100]; (ii) AGBR approach [93] where the GBR values are updated every 2 seconds for each HAS UEs (QCI equal to 4). Table 7.1 reports the average SNIRs experienced by each client as well as the average MAC rate provided by the three approaches in the two scenarios. Fig. 7.3 shows the received chunk-by-chunk SSIM at the client in the first scenario (A) for each strategy, while Table 7.2 reports the overall average and the standard deviation of the SSIM. We can note how AGBR approach allows to increase the quality of the high complex video with respect to BE approach by increasing the average rate provided to HAS clients. However, both approaches (BE

Scen.	SSIM	BE	AGBR	QFAS
A	Average	0.946	0.966	0.971
	Std. Dev.	0.042	0.031	0.005
B	Average	0.934	0.960	0.959
	Std. Dev.	0.063	0.045	0.012

Table 7.2: Overall average and standard deviation of the SSIM at the clients for two different scenarios resulting from BE, AGBR and the proposed QFAS.

and AGBR) experience less than ideal quality fairness with a standard deviation of the SSIM at the client higher than 0.031. Moreover, high quality fluctuations are experienced by users requesting high-complexity video, although they are in good channel condition. Our proposed QFAS strategy allows to significantly increase the quality of the high-complexity video up to 0.058 in average SSIM and 0.011 in overall SSIM, while keeping reasonable high quality to the low-complexity ones. However some quality drops in QFAS are still experienced by the *Bunny* client due to the gap between the limited number of available rate profiles and the continuous utility.

The benefits of our QFAS approach are also significant in the scenario B, as showed in Fig. 7.4. Due to the unfavorable channel condition, BE causes intolerable average and instantaneous quality degradation. AGBR improves video quality over BE. However, compared to QFAS, AGBR still results in lesser quality to *Sport*, exhibits higher fluctuations, and provides unfair quality in contrast to *Interview*. QFAS better distributes the available resources, as confirmed by the resulting average rates in Table 7.1, by providing a chunk SSIM equal or higher than 0.9 to all video programs which ensure a good quality level for all users. As reported in Table 7.2, similar overall SSIM are provided by both approaches but the fairness in terms of standard deviation is highly improved. We also verified that both AGBR and QFAS approaches maintains similar buffer stability at the client. However, some drop in quality experienced by *Home* client in AGBR are due to the RDA at the client, which is selecting the minimum rate profile to prevent buffer underflow.

Chapter 8

Conclusions

As video streaming has become the most popular application of Internet mobile, the requirements of enhanced video QoE of the end-user have called for content-aware optimized video delivery wireless systems. The main challenges still resides on a better Quality of Service (QoS) support, and on a dynamic adaptation of the transmitted video streams to meet the network condition. Without an efficient and optimized exchange of information among the different layers of the transmission systems, such goals are hard to be achieved.

In this thesis we have proposed novel cross-layer optimization frameworks for SVC delivery and for HAS application over the downlink and the uplink of Long Term Evolution (LTE) wireless networks. They jointly addressed optimized content-aware rate adaptation and radio resource allocation (RRA) with the aim of maximizing the sum of the achievable rates while minimizing the quality difference among multiple videos.

In order to perform optimized content-aware rate adaptation, we have first analyzed the video quality metrics that allow to assess the quality of a video sequence and then we have provided enhanced low-complexity models to accurately estimate the R-D relationship of scalable video transmitted over error-free and error-prone channels. For the latter scenario, an enhanced UXP profiler has been designed with the objective to provide R-D relationship that keeps the expected distortion almost unchanged with only a rate increase/decrease at different packet failure rate.

For the multi-user SVC delivery over downlink wireless systems, where OFDMA is the key PHY layer technology and IP/TV is one of the most representative application, we have first formulated the optimization problem and discussed its feasibility, showing that the optimal solution is unique and lays on the boundary of the convex rate region. Then, the problem has been "vertically" decomposed

into two sub-problems, each one characterized by parameters and optimization constraints confined within a single layer. The novel ILA algorithm has been proposed to achieve the global solution and its convergence and optimality have been rigorously proved. Also efficient methods to solve the two sub-problems has been presented by proving their optimality. Finally, in order to reduce the overall complexity and the latency of the proposed algorithm, a suboptimal low-complexity strategy based on the first-step of the ILA algorithm has been designed. From the implementation perspective, the proposed cross-layer strategy only requires that the base station is able to allocate resources according to a weighted sum-rate maximization, where the weights can be dynamically updated to track the existing rate constraints. In the timescale of the *application frame interval* the MAC layer sends the weights to the APP layer while the APP layer sends constraint parameters to MAC layer. Our numerical evaluations have shown that the ILA algorithm converges in few iterations and the suboptimal one-step version achieves almost the same performance of the ILA algorithm. Moreover, it is shown that the 1-step ILA algorithm is able to obtain significant overall and individual video quality gains, up to 1.5 and 7 dB in average PSNR, respectively, compared to other state-of-the-art frameworks exhibiting similar complexity.

For multiple SVC delivery over uplink wireless systems, where SC-FDMA is the key PHY layer technology and healt-care services are the most attractive and challenging application, we have proposed joint video adaptation and aggregation directly performed at the application layer of the transmitting equipment, which exploits the guaranteed bit-rate (GBR) provided by the e-NodeB. The proposed approach is able to manage the inherently different priorities of the data flows generated by the m-health user. In particular, it optimally adapts the SVC-encoded streams, in order to deliver the ultrasonography information with sufficiently high quality and the set of ambient videos tuned according to quality fairness criteria

Due to the NP-hardness of the RRA resource allocation problem and the requirements of QoS-aware scheduling strategies, we have also analyzed the ergodic sum-rate maximization problem under proportional rate constraints in SC-FDMA systems and we have proposed a novel sub-optimal algorithmic solution, whose complexity increases only linearly with the number of users and the number of resources. Numerical results have shown that the performance gap to optimal solution is limited to the 10% of the sum-rate.

Finally, we have proposed a quality-fair adaptive streaming solution to deliver fair video quality to HAS clients in a LTE cell by adaptively selecting the prescribed (GBR) of each user according to the video content in addition to the channel condition. By adding intelligence in the network, *i.e.*, through the use of a MANE, the proposed approach is able to control the rate provided to each

HAS user in order to obtain fair video quality among multiple HAS clients. This is achieved even when HAS users are requesting programs with significant differences in video complexity and are experiencing different channel conditions. Numerical results have shown that, compared to other state-of-the-art approaches, the proposed QFAS solution provides significant improvement in the overall quality delivered to user demanding complex video with a tolerable degradation of the other low-complex videos. However, some quality fluctuations dependent on the RDA at the client, are still present. Future works will consider the possibility of optimizing the rate request, *e.g.*, by overwriting the chunk request at the MANE, according to the solution of a problem aimed at provisioning fair video quality and buffer stability.

Even though broadband mobile provider are reluctant to include application-aware module in the design of cellular systems, due to management and coordination issues, our research have shown that tremendous gains in terms of the QoE of the end-user can be achieved by the proposed cross-layer strategies.

Appendix A

MAC Layer Algorithm: Extension to Multi-cell Scenario

RRA in the downlink of OFDMA systems has been studied extensively for the single-cell case [101]-[102]. However, to address realistic scenarios, a multi-cell environment has to be considered [103, 104]. In this Appendix we extend the downlink RRA algorithm for OFDMA system for a single cell scenario proposed in section 5.6 as part of the ILA algorithm, to multi-cell environment. In this case, LTE specifications suggest aggressive frequency reuse and distributed low-complexity implementations [105]. Nevertheless, if the frequency resource is fully reused in every cell of the network and no inter-Enhanced Node-B (eNB) cooperation/coordination is supported, the cell throughput will be reduced in the attempt of serving the users at the cell edge, due to inter-cell interference (ICI). Radio resource management with ICI coordination is a key issue under investigation by LTE research community.

ICI coordination can be achieved through the implementation of different degrees of network coordination and complexity. High complexity MIMO network approach [106] requires the availability of user data to be transmitted at all the eNB's, as well as collaborative processing based on dirty-paper coding and suitable precoding. A first step toward complexity reduction is to keep network coordination to only perform RRA. In a centralized RRA, a control unit collects all the channel state information (CSI) of every user in the system and allocates the available Physical Resources Block (PRB) of each eNB trying to maximize the capacity according to fairness and power constraints. Without an efficient and fast infrastructure, centralized scheduling is an hard task due to the stringent time required to exchange the inter-cell scheduling information and the large feedback required by the User Equipments (UEs) to send all the CSI. Some strategies

to reduce this complexity are under study for LTE-advanced [107] where joint transmission coordinated multi-point (JT-COMP) is proposed [108].

However, distributed RRA at each eNB is suggested for low-complexity multi-cell radio design. In distributed RRA each eNB allocates resources to its users only, and UEs feed back a partial CSI. ICI can be partially avoided by means of an off-line coordinated resource control among the cells in a cluster. Examples are given by Fractional Frequency Reuse (FFR) and Soft Frequency Reuse(SFR) [109][110], power planning techniques [111][112], load balancing (LB) [113], or by partial coordination as in [114] where only cell-edge users send CSI of interfering links. The lack of full coordination in RRA simplifies the implementation at the expense of some capacity degradation with respect to centralized RRA. Few contributions until now, *e.g.*, [114],[115], are addressing RRA in multicell OFDMA systems.

In this Appendix we first introduce a centralized RRA algorithm aimed at maximizing the sum-rate of a multi-cell clustered system under proportional rate constraints contraints, extending the strategy proposed in 5.6. The algorithm is obtained from an ergodic optimization framework presented in [102]. A stochastic approximation is applied to derive on-line implementation. After, we reformulate the algorithm to consider the constraints of a distributed RRA based on power planning schemes with pre-assigned powers. The distributed RRA algorithm preserves intra-cell fairness, but requires a LB algorithm to ensure inter-cell fairness. All the algorithms support PRB-based allocation typical of LTE systems.

By comparing centralized and distributed schemes, this work shows that distributed schemes with aggressive reuse manage to approach the capacity of a centralized system when the number of users is large. However, a fractional reuse between 2/3 and 4/5 helps to reduce the gap. This work is organized as follows. The framework proposed here can be then used, when multiple video has to be transmitted from a set of eNB connected to the MANE.

A.1 System Model

We consider a a cluster of Q cells with a total of K UEs or users and a multi-carrier transmission system with S available subcarriers divided in G groups $\mathcal{G}_g$, $g = 1, \ldots, G$, of $N = 12$ adjacent subcarriers. The piece of frame composed of N allocable subcarrier and one slot is defined as PRB, which is the elementary resource unit for RRA. Each cell is served by one eNB. In this paper, only the basic configuration where eNBs and UEs are equipped with one antenna is investigated. However, methods and algorithms developed here can also be extended to multi-

antenna configurations [102].

We use the discrete variable or index $u_{g,q} \in \mathcal{K}_0 = \{0, 1, .., K\}$ to indicate the user (*i.e.*, 0 means no user) that is scheduled to use cell q on PRB g. Note that only one user or none can be scheduled for each PRB and each cell. The whole set of these variables is the matrix $\boldsymbol{U} \in \mathcal{K}_0^{G \times Q}$, whereas the whole set of powers assigned for transmission is the matrix $\boldsymbol{P} \in \mathbb{R}^{+,S \times Q} \cup \{\boldsymbol{0}\}$. It is implicitly assumed that if $u_{g,q} = 0$ then $p_{s,q} = 0$, $\forall s \in \mathcal{G}_g$ that also means that $\boldsymbol{P}$ has an implicit dependence on $\boldsymbol{U}$ and viceversa as shown afterwards.

Network coordination is only exploited to perform RRA, *i.e.*, no co coming from other eNB are considered as inter-cell interference. We define the Signal-to-Interference plus Noise Ratio (SINR) of user k at frequency s, when the signal is coming from cell q, as

$$\gamma_{k,s,q}(\mathbf{p}_s) = \frac{p_{s,q} c_{k,s,q}}{1 + \sum_{m=1, m \neq q}^{Q} p_{s,m} c_{k,s,m}} \tag{A.1}$$

where $\mathbf{p}_s = [p_{s,1}, \ldots, p_{s,Q}]$ and $c_{k,s,q} = |h_{k,s,q}|^2 / \sigma_n^2$, being σ_n^2 the noise power and $|h_{k,s,q}|^2$ the channel power gain between eNB q and UE k on subcarrier s. The channel gain that includes the contribution of path-loss, shadowing and fast fading.

A.2 Centralized RRA

We consider here a centralized architecture where a control unit collects all the CSI of every user in the system and allocates the resource units of the cluster trying to maximize the capacity according to fairness and power constraints.

Following the approach in [101], we consider the framework of ergodic sum-rate maximization for continuous (capacity based) rates extended to the multi-cell case. The problem can be formulated, similarly to problem (5.16), as

$$\begin{aligned} & \max_{\boldsymbol{U},\boldsymbol{P}} \|\boldsymbol{R}(\boldsymbol{U},\boldsymbol{P})\|_1 \\ s.t.\ & P_q(\boldsymbol{U},\boldsymbol{P}) \leq \bar{P}_q, \\ & \boldsymbol{R}(\boldsymbol{U},\boldsymbol{P}) \succeq \boldsymbol{\phi} \|\boldsymbol{R}(\boldsymbol{U},\boldsymbol{P})\|_1, \end{aligned} \tag{A.2}$$

where $\boldsymbol{R}(\boldsymbol{U},\boldsymbol{P}) = [R_1(\boldsymbol{U},\boldsymbol{P}), \ldots, R_K(\boldsymbol{U},\boldsymbol{P})]$ and

$$R_k(\boldsymbol{U},\boldsymbol{P}) = \sum_{g \in \mathcal{G}} \sum_{m \in \mathcal{G}_g} \sum_{q=1}^{Q} \mathbb{E}[\delta_k^{u_{g,q}} C(\gamma_{k,m,q}(\boldsymbol{p}_m))] \tag{A.3}$$

is the average rate per unit bandwidth provided to user k and

$$\mathcal{P}_q(\boldsymbol{U}, \boldsymbol{P}) = \sum_{m=1}^{M} \mathbb{E}[p_{m,q}] \tag{A.4}$$

is the total average power spent by cell q to serve the allocated users. Finally, δ_k^u is the Kronecker's delta[1] and $C(x) = \log_2(1+x)$.

The first constraint refers to the total average power used by q-th eNB, which must be less than or equal to a maximum amount $\bar{P}_q$. Let us note that this constraint allows instantaneous power levels to exceed the average power when necessary. The second constraint determines the share of throughput finally achieved by each user. Therefore $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_K]^T$ defines the required QoS by each user and must satisfy the condition $\sum_{k=1}^{K} \phi_k = 1$.

A.2.1 Solutions for the Allocation Problem

Here, we follow the approach presented in [102] for a single-cell multi-antenna system, which is based on a dual optimization framework that utilizes the Lagrangian function $L(\mathbf{U}, \mathbf{P}, \boldsymbol{\lambda}, \boldsymbol{\mu})$, where the dual variables $\boldsymbol{\lambda} = [\lambda_1, ..., \lambda_Q]^T, \boldsymbol{\mu} = [\mu_1, \ldots, \mu_K]^T$ relax the cost function. The dual problem becomes

$$\begin{aligned} &\min_{\boldsymbol{\lambda}, \boldsymbol{\mu}} \; g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\ &s.t. \; \boldsymbol{\lambda} > 0, \; \boldsymbol{\mu} \geq 0, \; (1 - \boldsymbol{\mu}^T \boldsymbol{\phi}) = 0 \end{aligned} \tag{A.5}$$

where the third constraint on $\boldsymbol{\mu}$ holds if sum-rate is not diverging to infinity and $g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \max_{\mathbf{U}, \mathbf{P}} L(\mathbf{U}, \mathbf{P}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ is the dual objective.

Although the system utility in (A.2) is non-concave, for ergodic optimization it is proved that duality gap is zero if the cumulative density function (CDF) of channel gains is continuous, which happens in classical Rayleigh and Ricean scenarios. However, it is not possible to guarantee the dual problem is differentiable. Hence, an iterative sub-gradient method which updates the $Q + K$ solutions $\boldsymbol{\lambda}, \boldsymbol{\mu}$ of the dual problem (A.5) at each iteration can be applied. Nevertheless, in the practical applications, the adaptive implementation is suggested, where the iterations are performed along time and the evaluation of the average power and rate in the subgradients can be done through a stochastic approximation, as outlined in [101, 102].

[1] $\delta_k^u = 1$ if $u = k$, 0 otherwise

In order to derive the dual objective $g(\boldsymbol{\lambda}, \boldsymbol{\mu})$ given $\boldsymbol{\lambda}, \boldsymbol{\mu}$ the expression of the Lagrangian function can be suitably manipulated as in [102], leading to:

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \sum_{q=1}^{Q} \lambda_q \bar{P}_q + S\mathbb{E}\left[\max_{\boldsymbol{u}_g, \boldsymbol{p}_m, m \in \mathcal{G}_g} M(\boldsymbol{u}_g, \boldsymbol{P})\right] \tag{A.6}$$

where $\mathbf{u}_g = [u_{g,1}, \ldots, u_{g,Q}]$ and

$$M(\boldsymbol{u}_g, \boldsymbol{P}) = \sum_{q=1, u_{g,q} \neq 0}^{Q} \sum_{m \in \mathcal{G}_g} \left[\mu_{u_{g,q}} C(\gamma_{u_{g,q},m,q}(\boldsymbol{p}_m)) - \lambda_q p_{m,q}\right] \tag{A.7}$$

The optimal solutions for the evaluation of the dual objective, given $\boldsymbol{\lambda}, \boldsymbol{\mu}$, denoted as $\boldsymbol{U}^* = [\boldsymbol{u}_1^*, ..., \boldsymbol{u}_G^*]^T$, $\boldsymbol{P}^* = [\boldsymbol{p}_1^*, ..., \boldsymbol{p}_M^*]^T$, becomes:

$$\boldsymbol{u}_g^* = \arg\max_{\boldsymbol{u}_g} M^*(\boldsymbol{u}_g) \tag{A.8}$$

with

$$M^*(\boldsymbol{u}_g) = \max_{\boldsymbol{p}_s, s \in \mathcal{G}_g} M(\boldsymbol{u}_g, \boldsymbol{P}) \tag{A.9}$$

and $\boldsymbol{p}_s^*$ is the argument that finally leads to $M^*(\boldsymbol{u}_g)$.

It should be noted that user allocation in (A.8) represents a discrete optimization problem which requires in general an exhaustive search in the space of all possible vectors $\boldsymbol{u}_g$. This huge search space can eventually be reduced by using suboptimal heuristic algorithms as in [102]. However, for each element of this search space the power allocation solution, *i.e.*, (A.9), has to be computed. This non-convex problem can be solved by using successive convex approximation methods as in [116]. A suboptimal solution is obtained by simplifying power allocation with the water-filling solution evaluated by assuming constant uniform power for the interfering cells, *i.e.*,

$$\tilde{p}_{s,q} = \left[\frac{\mu_{u_{g,q}}}{\lambda_q \ln_2} - \frac{V}{\gamma_{u_{g,q},s,q}(\boldsymbol{V}_s)}\right]^+ \tag{A.10}$$

where the components of $\boldsymbol{V}_s$ are $v_{s,q} = V\delta_0^{u_{g,q}}$. The power V is a parameter which estimates the power of interfering cells in each subcarrier.

Even though the power allocation algorithm based on (A.10), as well as the update of variables $\boldsymbol{\lambda}$ in the adaptive implementation, can be distributed on each base station, user allocation algorithm as in (A.8) requires a centralized controller which determines, for all sub-carriers, the vectors $\boldsymbol{u_g}$ and sends them to eNB through signaling. Therefore, each eNB has to forward the received CSIs of each

UE to the centralized controller and to receive back the allocation informations before transmitting, resulting in a high black-haul signaling. Besides, the signaling interface in realistic systems, *i.e.*, LTE X2 interface, has not a negligible latency, resulting in an additional delay on the CSIs report, that should be taken into account, especially in very fast fading environments. Thus, in a realistic network, centralized resource allocation is practically difficult to be realized. In the next section we provide a distributed solution and analyze techniques to mitigate the inter-cell interference.

A.3 Distributed RRA

In distributed RRA each eNB allocates resources to its users only without any knowledge of the allocation process at the other eNBs, meaning without knowledge of the actual ICI. Only partial ICI information is available at each eNB, which essentially consists of a "power mask" $\overline{\mathbf{V}}_s = [\overline{V}_{s,1}, \ldots, \overline{V}_{s,Q}]$, $s = 1, \ldots, S$, used to limit the power allocated by each eNB q on each subcarrier s. The effects of ICI can be further mitigated by means of an off-line coordinated resource control among the cells in a cluster, which is the subject of next Section.

To formulate the distributed RRA problem, let us first define $\mathcal{K}^{(q)}$, with cardinality $K^{(q)}$, as the set of users served by the q-th eNB, where $\mathcal{K}^{(1)} \bigcap \ldots \bigcap \mathcal{K}^{(Q)} = \{0\}$ and $\bigcup_{q=1}^{Q} \mathcal{K}^{(q)} = \mathcal{K}$. In the distributed setting, only the lower-bound of the SINR which depends on the power mask $\overline{\mathbf{V}}_s$, $s = 1, \ldots, S$ is known at the eNB, *i.e.*, $\gamma_{k,s,q}(\mathbf{p}_s) \geq p_{s,q}\eta_{k,s,q}$ where

$$\eta_{k,s,q} = \frac{c_{k,s,q}}{1 + \sum_{m=1, m\neq q}^{Q} \overline{V}_{s,m} c_{k,s,m}} \tag{A.11}$$

is the normalized SNIR lower-bound. The UE can measure $\eta_{k,s,q}$ and send it through a feedback channel to its eNB only. The average rate per unit bandwidth that can be provided to user k in cell q, without incurring in outages, is given by

$$\tilde{R}_k(\boldsymbol{U}, \boldsymbol{P}) = \sum_{g \in \mathcal{G}} \sum_{s \in \mathcal{G}_g} \mathbb{E}[\delta_k^{u_{g,q}} C(p_{s,q}\eta_{k,s,q})],\ k \in \mathcal{K}^{(q)} \tag{A.12}$$

We can then write the distributed problem as a maximization problem for each

cell $q = 1, \ldots, Q$:

$$\begin{aligned} &\max_{\boldsymbol{U},\boldsymbol{P}} \|\tilde{\boldsymbol{R}}(\boldsymbol{U},\boldsymbol{P})\|_1 \\ &s.t.\ \ P_q(\boldsymbol{U},\boldsymbol{P}) \leq \bar{P}_q, \\ &\qquad \tilde{R}_k(\boldsymbol{U},\boldsymbol{P}) \geq \phi_k \sum_{m\in\mathcal{K}^q} \tilde{R}_m(\boldsymbol{U},\boldsymbol{P}), \quad \forall k \in \mathcal{K}^q \\ &\qquad p_{s,q} \leq \overline{V}_{s,q}, \quad \forall m \end{aligned} \tag{A.13}$$

where now there is a restriction on the set of allocation variables $\boldsymbol{U}$, *i.e.*, $u_{g,q} \in \mathcal{K}^q$, and ϕ_k must satisfy the constraint

$$\sum_{k\in\mathcal{K}^q} \phi_k = 1, \quad \forall q \tag{A.14}$$

In this way each cell tries to maximize its sum-rate, taking care of intra-cell fairness only.

A solution for the RRA problem can derived for each cell q by following the same approach of Sec.A.2 through dual optimization and adaptive algorithms. The set of dual variables to be updated is still the same, but now the constraints on $\boldsymbol{\mu}$ change as $(\boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{\phi})^{(q)} = \sum_{k\in\mathcal{K}^q} \mu_k\phi_k = 1$. The allocation problem, due to cell decoupling, is now simpler than before, because it avoids the non-convex multi-cell power allocation, and requires, for each cell q and PRB g, the maximization of the following metric

$$M^{(q)}(u_{g,q}, \mathbf{P}) = \sum_{s\in\mathcal{G}_g} \left[\mu_{u_{g,q}} C(p_{s,q}\eta_{u_{g,q},s,q}) - \lambda_q p_{s,q}\right] \tag{A.15}$$

leading to the optimal solution [61]:

$$p^*_{s,q} = \min\left\{\overline{V}_{s,q}, \left[\frac{\mu_{u_{g,q}}}{\lambda_q \ln 2} - \frac{1}{\eta_{u_{s,q},s,q}}\right]^+\right\} \tag{A.16}$$

$$u^*_{g,q} = \arg\max_{u_{g,q}} \mathcal{M}^{(q)}(u_{g,q}, \mathbf{P}^*) \tag{A.17}$$

The main drawback of this solution is the fact that the fairness of rate allocation is confined within each cell, whereas global fairness depends on load and channel conditions on each cell. The fairness issue can be partially solved with an off-line algorithm that balances the eNBs load, as contemplated by LTE [105]. Although LB is not within the scope of our paper, we will evaluate the performance of distributed RRA when a simple LB algorithm, is running, which is reported next for the sake of completeness.

A.3.1 A Greedy Load Balancing Algorithm

The algorithm considers for each cell q of the cluster the following load metric:

$$L_q = \sum_{k \in \mathcal{K}^q} l_{k,q} \tag{A.18}$$

where

$$l_{k,q} = \frac{\phi_k K^{(q)}}{C(\overline{SIR}_{k,q})} \tag{A.19}$$

and $\overline{SIR}_{k,q}$ is the signal-to-interference ratio of user k with respect to cell q evaluated by taking only into account distance-based attenuation. The load metric $l_{k,q}$ estimates the amounts of resource units needed by user k, if served by cell q, to achieve the required portion ϕ_k of the sum-rate. The pseudo-code of the algorithm is listed in Algorithm.

After an initial assignment of each users k to cell q (line 2-6) having on the minimum estimated resources $l_{k,q}$, the LB algorithm aims to minimize the difference between the maximum and minimum load of the eNB, *i.e.*, $\Delta L = L_{q^{\max}} - L_{q^{\min}}$. This is iteratively done by moving the user belonging to the cell $q^{\max}$ with the highest load to the cell $q^{\min}$ with the minimum load, and also having the minimum *positive* difference $\Delta l_k = l_{k,q^{\min}} - l_{k,q^{\max}}$ between the estimated amount of resource $l_{k,q}$, which would drain by the cell $q^{\min}$ and $q^{\max}$. It easy to see that if $\Delta L[i]$ is the difference between the maximum and minimum load at iteration i then $\Delta L[i+1]$ after the new assignment of user u is such that

$$\begin{aligned}
\Delta L[i+1] =& (L_{q^{\max}}[i] - l_{u,q^{\max}}) - (L_{q^{\min}}[i] + l_{u,q^{\min}}) && \text{(A.20a)}\\
=& \Delta L[i] - l_{u,q^{\min}} - l_{u,q^{\max}} && \text{(A.20b)}\\
\leq& \Delta L[i] && \text{(A.20c)}
\end{aligned}$$

where the inequality holds given the condition in line 19.

Algorithm 5 Pseudo code of the Greedy Load Balance Algorithm

1: **Initialize** tolerance ϵ
2: $\mathcal{K}^{(q)} = \{0\}$, $\forall q$
3: **for all** $k \in \mathcal{K}$ **do**
4: $\quad q^* \leftarrow \underset{q}{\text{argmin}}\ l_{k,q}$
5: $\quad \mathcal{K}^{(q^*)} \leftarrow \mathcal{K}^{(q^*)} \cup k$
6: **end for**
7: $q^{\min} \leftarrow \underset{q}{\text{argmin}}\ L_q$;
8: $q^{\max} = \underset{q}{\text{argmax}}\ L_q$;
9: $\Delta L \leftarrow L_{q^{\max}} - L_{q^{\min}}$
10: $\Delta l_k \leftarrow l_{k,q^{\min}} - l_{k,q^{\max}}$, $\forall k \in \mathcal{K}^{q^{\max}}$
11: $\mathcal{K}^{\text{cand}} = \{0\}$
12: **for all** $k \in \mathcal{K}^{q^{\max}}$ **do**
13: $\quad$ **if** $\Delta l_k > 0$ **then**
14: $\quad\quad \mathcal{K}^{\text{cand}} \leftarrow \mathcal{K}^{\text{cand}} \cup k$
15: $\quad$ **end if**
16: **end for**
17: **if** $\mathcal{K}^{\text{cand}} \neq \{0\}$ **then**
18: $\quad u = \underset{k \in \mathcal{K}^{\text{cand}}}{\text{argmin}} \Delta l_k$
19: $\quad$ **if** $\Delta l_u > \epsilon \vee \Delta L \geq l_{k,q^{\min}} + l_{k,q^{\max}}$ **then**
20: $\quad\quad \mathcal{K}^{(q^{\max})} \leftarrow \mathcal{K}^{(q^{\max})} \setminus \{u\}$
21: $\quad\quad \mathcal{K}^{(q^{\min})} \leftarrow \mathcal{K}^{(q^{\min})} \cup \{u\}$
22: $\quad\quad$ **Go** to line 9
23: $\quad$ **end if**
24: **end if**

A.3.2 PRB-based Power/Rate Allocation

In order to decrease feedback complexity, in a realistic LTE scenario also rate and power are allocated per PRB, as channel state feedback is reduced to no more than one value per PRB. In this downlink case, the Exponential Effective SNIR Mapping (EESM)[117] is used to evaluate the SNIR of each PRB, by taking into account that power mask is constant inside each PRB, *i.e.*, $\overline{\mathbf{V}}_s = \overline{\mathbf{V}}_g$, $\forall s \in \mathcal{G}_g$, with $\overline{\mathbf{V}}_g = [\overline{V}_{g,1}, \ldots, \overline{V}_{g,Q}]$.

The EESM for user k, PRB g, cell q is defined as

$$\gamma_{k,g,q}^{(\text{eff})} = -\beta \log \left(\frac{1}{N} \sum_{s \in \mathcal{G}_g} e^{-\frac{(\overline{V}_{g,q}\eta_{k,s,q} - \gamma_{k,g,q}^{(\min)})}{\beta}} \right) + \gamma_{k,g,q}^{(\min)} \tag{A.21}$$

where

$$\gamma_{k,g,q}^{(\min)} = \overline{V}_{g,q} \min_{s \in \mathcal{G}_g} \eta_{k,s,q} \tag{A.22}$$

and the parameter β is used to tune the approximation. Here, with respect to [117], the term $\gamma_{k,g,q}^{(\min)}$ is introduced as an offset that keeps the range of exponential function limited. The effective instantaneous rate for user k using PRB g in cell q, when a power $p_{g,q}$ is allocated for transmission, becomes

$$r_{k,g,q}^{(\text{eff})} = N \cdot C \left(p_{g,q} \frac{\gamma_{k,g,q}^{(\text{eff})}}{\overline{V}_{g,q}} \right) \tag{A.23}$$

Parameter β in (A.21) has to be designed to keep the probability $P(\tilde{r}_{k,g,q} \leq r_{k,g,q}^{(\text{eff})})$, where $\tilde{r}_{k,g,q} = \sum_{m \in \mathcal{G}_g} C(p_{g,q}\eta_{k,m,q})$, below a given threshold to prevent outage events.

In distributed RRA using EESM feedback for each PRB, the equations (A.12), (A.15) and (A.16) change by replacing $\eta_{k,s,q}$ with $\gamma_{k,g,q}^{(\text{eff})}/\overline{V}_{g,q}$ and by setting $p_{s,q} = p_{g,q}$, $\overline{V}_{s,q} = \overline{V}_{g,q}$, $\forall s \in \mathcal{G}_g$.

A.4 Power Planning for ICI Coordination

Generally, in distributed RRA each eNB allocates resources without complete knowledge of ICI. However, when the maximum value of the transmitted power is predefined and known to all the eNB, partial information on ICI is available, which enables some kinds of ICI coordination. Since the loss of performance is usually due to the unlucky UEs close to the cell border which drain radio resources in the attempt to obtain the same capacity of the lucky UEs closer to the eNB, ICI coordination techniques can be used to reduce interference in resource units assigned to unlucky users.

In this work we denote in general as "power planning" techniques those technique aimed at determining the set of values $\overline{\mathbf{V}}_g$, $g = 1, \ldots, G$ that allow the best distribution of ICI across resource units as function of system and load conditions. Within this framework, the simplest techniques that consider an off-line static design of maximum power values are FFR and SFR. In this paper we consider FFR techniques to be combined with distributed RRA. Improvements might come from more sophisticated power planning techniques which optimize all the values of $\overline{\mathbf{V}}_g$ with a limited set of constraints, but this calls for further investigation.

When FFR is applied, $\overline{\mathbf{V}}_g$ is one of the elements of a finite set $\mathcal{V}$; we assume that if the vector $\mathbf{V}$ is an element of $\mathcal{V}$, then all the vectors obtained as cyclic shifts

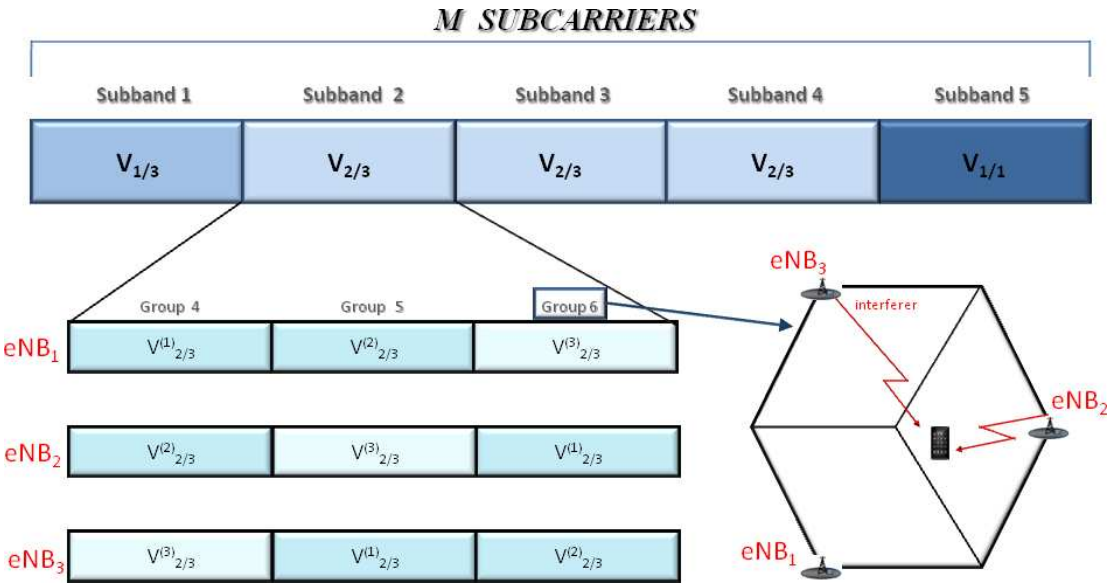


Figure A.1: An example of FFR based power planning for a cluster of $Q = 3$ cells with overall reuse factor 2/3

Sub-band	1	2	3	4	5	Overall Reuse Factor
Reuse	1/3	1/3	1/3	1/3	1/3	**1/3**
Reuse	1/3	1/3	2/3	2/3	2/3	**8/15**
Reuse	1/3	2/3	2/3	2/3	1	**2/3**
Reuse	2/3	2/3	2/3	1	1	**4/5**
Reuse	1	1	1	1	1	**1/1**

Table A.1: Frequency partitioning for different FFR schemes with $Q = 3$

of $\mathbf{V}$ are elements of $\mathcal{V}$. All the available PRBs are partitioned in G/Q sub-bands of Q PRBs. Inside each sub-band the Q shifts of the same power vector $\mathbf{V} \in \mathcal{V}$ are assigned to the Q PRBs. If G/Q is not integer we may apply the partitioning and assignment of power vectors to a pool of GQ PRBs over Q slots. If one or more of the elements of power vector $\mathbf{V} \in \mathcal{V}$ are zero, we say that the resource unit or PRB is working with power vector $\mathbf{V}$ has a reuse factor smaller than 1. This means that for one cell the PRB has a reduced ICI, because one or more of the other cells are not allowed to transmit in the PRB.

As an example for $Q = 3$ we consider the following elements of $\mathcal{V}$:

- $\boldsymbol{V}_{\frac{1}{3}} = [V_{\frac{1}{3}}^{(0)}, V_{\frac{1}{3}}^{(1)}, V_{\frac{1}{3}}^{(2)}] = [\infty, 0, 0]$
- $\boldsymbol{V}_{\frac{2}{3}} = [V_{\frac{2}{3}}^{(0)}, V_{\frac{2}{3}}^{(1)}, V_{\frac{2}{3}}^{(2)}] = [\frac{3}{2}V, \frac{3}{2}V, 0]$
- $\boldsymbol{V}_{\frac{1}{1}} = [V_{\frac{1}{1}}^{(0)}, V_{\frac{1}{1}}^{(1)}, V_{\frac{1}{1}}^{(2)}] = [V, V, V]$

and we also assume that $\bar{P}_q = \bar{P}, \forall q$ and $V = \xi \bar{P}/G$. The parameter ξ allows a simple off-line optimization of power levels. Note that in the power vector $\boldsymbol{V}_{\frac{1}{3}}$ the non zero elements is infinity, because there is no need to limit the power when all the other cells are not allowed to use the PRB. The assignment of these vectors to sub-bands and PRBs is illustrated in Fig.A.1 with reference to a system with overall reuse factor 2/3. The just introduced concept allow us to define other FFR schemes. We summarize in table A.1 the most relevant used to obtain numerical results.

A.5 Numerical Results

The performance evaluation is carried out through Monte-Carlo simulations according to models and assumptions summarized in Tab. A.2, also following the guidelines for LTE in [105][118]. In the evaluation of assigned rates an SNR gap of 3dB is taken into account. We consider a downlink scenario with a cluster of $Q = 3$ cells, where the centralized and distributed RRA techniques described in the paper are evaluated and compared. We denote with S1 the distributed system where the power and the user rates are evaluated per subcarriers (sect. A.3), whereas S2 indicates the distributed system with per-PRB power and rate allocation based on the EESM metrics (subsect. A.3.2). The fairness in rate allocation is evaluated through the well-known Jain Index [91]

$$J = \frac{(\sum_{k=1}^{K} x_k)^2}{K \sum_{k=1}^{K} x_k^2} \tag{A.24}$$

where x_k is modified to take account of inter-class fairness, *i.e.*, $x_k = \frac{R_k}{\phi_k}$. However, without loosing generality, we presents here the results for one class, *i.e.*, $\phi_k = \frac{1}{K}, \forall k$ in centralized RRA and $\phi_k = \frac{1}{K^{(q)}}, \forall k \in \mathcal{K}^q$ in distributed RRA.

Fig. A.2 provides a comparison of the different RRA techniques proposed in the paper by showing the capacity loss of distributed RRA with various FFR schemes with respect to centralized RRA for different number of UEs in the cluster. The LB algorithm is implemented for the distributed RRA. All the RRA schemes

System model	
User distribution	Uniform, in average 5, 10, 15, 20 per sector
Cell layout	Single Hexagonal Cluster, with 3 sectors
Inter-eNB distance	520 m
Data generation	Full buffer
Channel model	
Path Loss	40 + 15.2log(d), d = distance in meter
Doppler Bandwidth	6Hz
Shadowing model	Log-normal with 6dB standard deviation
Fast Fading	3GPP Pedestrian model
Delay spread	2.3 μs
PHY model	
System Bandwidth	3 MHz
Subcarrier spacing	15 KHz
Number of allocable subcarriers	180
Number of carrier per PRB	15
Frame duration	10 ms
Slot duration	0.5 ms
DL slots per frame	8 (TDD - Configuration 1 [105])
OFDM symbols per slot	7
CSI update	5 ms
Transmission Time Interval (TTI)	1 ms
Average Maximum eNB Power	1 W
Noise Power Density	$2 \cdot 10^{-20}$ W/Hz

Table A.2: Simulation model

allocate rates with a Jain's index ranging from 0.97 with K=15 user to nearly 1. We note that the capacity loss decreases when the number of users increases, emphasizing that multiuser diversity significantly help distributed RRA. We also note that FFR schemes with reuse factor of 4/5 reduce the capacity gap with respect to full-reuse RRA, and a loss of some percent units is due to to per-PRB allocation. Optimal reuse factor moves towards 1 as the number of users get large. In all the investigated cases the capacity loss is between 5% and 30%. In the centralized RRA investigated here power allocation is evaluated with suboptimal solution (A.10). However, it has been checked that optimal power allocation provides a limited capacity gain around 7% for $K = 30$ users, at the expense of more complexity.

Table A.3 collects results illustrating the impact of LB techniques to ensure

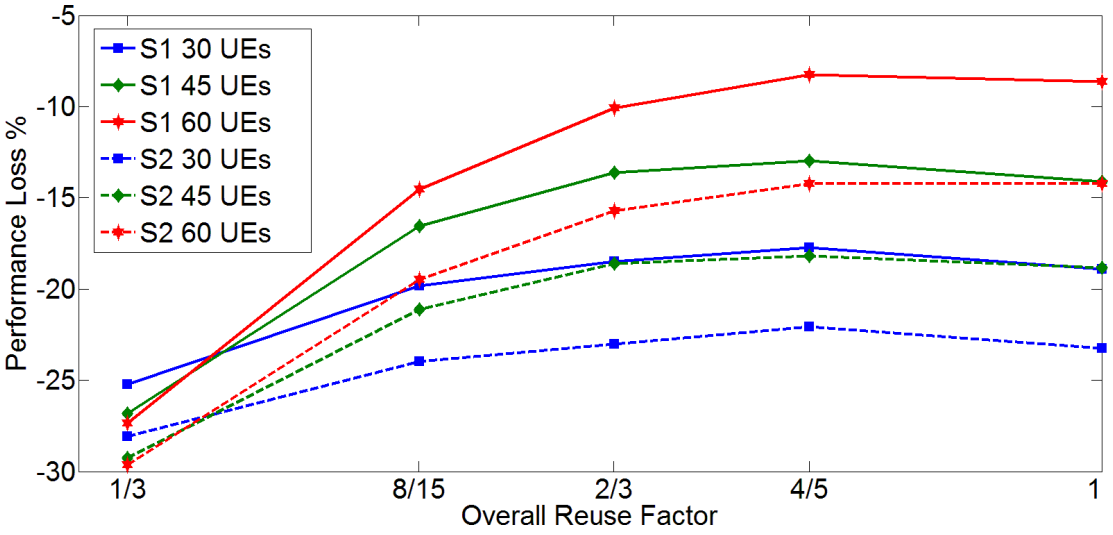


Figure A.2: Capacity loss of distributed RRA with respect to centralized RRA

fairness in distributed RRA. The table shows the sum-rate and Jain's index of the whole system, with and without LB, as well as the load metric in each cell. Without LB a larger sum-rate is achieved at the expense of fairness among users. It is also interesting to note that in case of spatially uniform distribution of user, LB looses relevance when the number of users get large.

Next table, Tab. A.4 shows the trade-off between allocated sum-rate and outage rate as function of parameter β in the EESM metric. According to these results we set $\beta = 410$ for all simulations in order to keep $P(\tilde{r}_{k,g,q} \leq r_{k,g,q}^{(\text{eff})})$ below 1%.

Finally, in Fig. A.3 we investigate the sensitivity of achieved sum-rate to the choice of parameter ξ, which selects the values in the power vectors used in distributed RRA with $K = 30$ users. It is shown that the maximum capacity is obtained with $\xi = 1$, which is the value used in all simulations. This result does not change for different values of K.

K	q	**without LB**			**with LB**		
		L_q	J	R[kbps]	L_q	J	R[kbps]
15	1	1.371	0.845	5402	2.756	0.971	4969
	2	1.773			3.054		
	3	4.188			2.441		
30	1	6.178	0.941	6532	5.106	0.990	6621
	2	4.341			4.341		
	3	3.046			4.344		
45	1	3.131	0.944	8103	7.628	0.998	7398
	2	8.498			6.933		
	3	7.226			7.226		
60	1	5.811	0.968	9636	8.965	0.999	9501
	2	7.068			8.601		
	3	12.582			9.122		

Table A.3: Load metric, Jain's index and sum-rate in each cell ($Q = 3$) of the cluster for different values of K and full-reuse distributed RRA in a single simulation scenario

β	**350**	**395**	**400**	**405**	**410**	**420**	**450**
Sum-rate [Kbps]	6402	6433	6440	6442	6444	6449	6466
Outage Prob. %	0.00	0.34	0.51	0.72	0.96	1.56	3.96

Table A.4: Sum-rate and outage probability $P(\tilde{r}_{k,g,q} \leq r_{k,g,q}^{(\text{eff})})$ vs EESM parameter β. Distributed RRA with full-reuse and $K = 30$

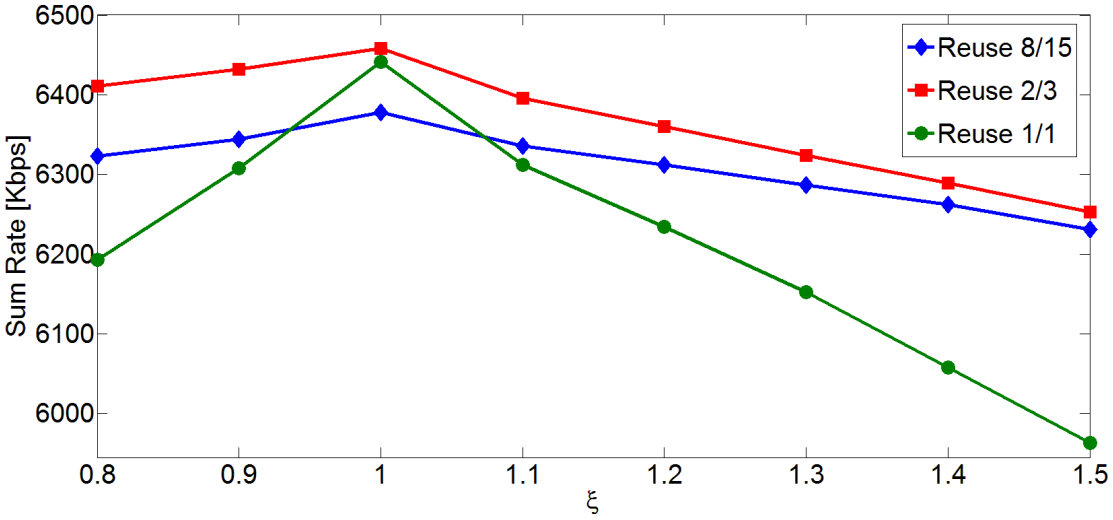


Figure A.3: Sum-rate vs parameter ξ in S2 systems with $K = 30$ UEs

Bibliography

[1] "Cisco visual networking index: Forecast and methodology, 2008-2013," Cisco, Tech. Rep., 2009.

[2] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, 2007.

[3] T. Stockhammer, "Dynamic adaptive streaming over HTTP design principles and standards," *in Proc. ACM conference on Multimedia systems*, pp. 133–144, 2011.

[4] D. Astely, E. Dahlman, A. Furuskar, Y. Jading, M. Lindstrom, and S. Parkvall, "LTE: the evolution of mobile broadband," *IEEE Communications Magazine*, vol. 47, no. 4, pp. 44–51, April 2009.

[5] C. Ciochina and H. Sari, "A review of OFDMA and single-carrier FDMA," in *Wireless Conference (EW), 2010 European*, April 2010, pp. 706–710.

[6] H. G. Myung, J. Lim, and D. J. Goodman, "Single carrier FDMA for uplink wireless transmission," *IEEE Vehicular Technology Magazine*, vol. 1, no. 3, pp. 30–38, 2006.

[7] G.-M. Su, Z. Han, M. Wu, and K. J. R. Liu, "A scalable multiuser framework for video over OFDM networks: Fairness and efficiency," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 10, pp. 1217–1231, 2006.

[8] N. Changuel, B. Sayadi, and M. Kieffer, "Control of distributed servers for quality-fair delivery of multiple video streams," in *Proc. of the 20th ACM Int. Conf. on Multimedia*, 2012, pp. 269–278.

[9] S. Cicalò, A. Haseeb, and V. Tralli, "Multi-stream rate adaptation using scalable video coding with medium grain scalability," in *Mobile Multimedia Communications, Lecture Notes of the Institute for Computer Sciences,*

Social Informatics and Telecommunications Engineering. Springer Berlin Heidelberg, 2012, vol. 79, pp. 152–167.

[10] ——, "Fairness-oriented multi-stream rate adaptation using scalable video coding," *Elseveir Signal Processing: Image Communication*, vol. 27, no. 8, pp. 800 – 813, 2012.

[11] H. Mansour, V. Krishnamurthy, and P. Nasiopoulos, "Channel aware multiuser scalable video streaming over lossy under-provisioned channels: Modeling and analysis," *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1366–1381, 2008.

[12] X. Ji, J. Huang, M. Chiang, G. Lafruit, and F. Catthoor, "Scheduling and resource allocation for SVC streaming over OFDM downlink systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 10, pp. 1549–1555, 2009.

[13] H. Zhang, Y. Zheng, M. Khojastepour, and S. Rangarajan, "Cross-layer optimization for streaming scalable video over fading wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 344–353, 2010.

[14] Z. Guan, D. Yuan, and H. Zhang, "Optimal and fair resource allocation for multiuser wireless multimedia transmissions," *EURASIP J. Wirel. Commun.*, vol. 2009, no. 1, 2009.

[15] T. Schierl, H. Schwarz, D. Marpe, and T. Wiegand, "Wireless broadcasting using the scalable extension of H. 264/AVC," in *Proc. IEEE Int. Conf. Multimedia and Expo ICME 2005*, 2005, pp. 884–887.

[16] E. Maani and A. K. Katsaggelos, "Unequal error protection for robust streaming of scalable video over packet lossy networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 3, pp. 407–416, 2010.

[17] S. Cicalò and V. Tralli, "Cross-layer algorithms for distortion-fair scalable video delivery over OFDMA wireless systems," in *Globecom Workshops (GC Wkshps), 2012 IEEE*, 2012, pp. 1287–1292.

[18] Y. Chu and A. Ganz, "A mobile teletrauma system using 3G networks," *IEEE Transactions on Information Technology in Biomedicine*, vol. 8, no. 4, pp. 456–462, Dec 2004.

[19] V. Vukadinovic and G. Karlsson, "Video streaming performance under proportional fair scheduling," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 399–408, 2010.

[20] M. Van Der Schaar and S. N. Sai, "Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms," *IEEE Wireless Commun. Mag.*, vol. 12, no. 4, pp. 50–58, 2005.

[21] Y. Sanchez, T. Schierl, C. Hellge, T. Wiegand, D. Hong, D. De Vleeschauwer, W. Van Leekwijck, and Y. Lelouedec, "Improved caching for HTTP-based video on demand using scalable video coding," in *Consumer Communications and Networking Conference (CCNC), 2011 IEEE*, Jan 2011, pp. 595–599.

[22] T. Wiegand, L. Noblet, and F. Rovati, "Scalable video coding for IPTV services," *IEEE Transactions on Broadcasting*, vol. 55, no. 2, pp. 527–538, June 2009.

[23] T. Schierl, C. Hellge, S. Mirta, K. Gruneberg, and T. Wiegand, "Using H.264/AVC-based scalable video coding (SVC) for real time streaming in wireless IP networks," in *IEEE International Symposium on Circuits and Systems, 2007. ISCAS 2007*, May 2007, pp. 3455–3458.

[24] Microsoft smooth streaming. [Online]. Available: http://www.iis.net/downloads/microsoft/smooth-streaming

[25] Apple HTTP live streaming. [Online]. Available: http://tools.ietf.org/html/draft-pantos-http-live-streaming-07

[26] Adobe HTTP adaptive streaming. [Online]. Available: http://www.adobe.com/products/hds-dynamic-streaming.html

[27] I. Sodagar, "The MPEG-DASH standard for multimedia streaming over the internet," *IEEE MultiMedia*, vol. 18, no. 4, pp. 62–67, April 2011.

[28] I. Amonou, N. Cammas, S. Kervadec, and S. Pateux, "Optimized rate-distortion extraction with quality layers in the scalable extension of H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1186–1193, 2007.

[29] C. Segall and G. Sullivan, "Spatial scalability within the H.264/AVC scalable video coding extension," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1121–1135, Sept 2007.

[30] *JSVM 9.19.11 Reference Software Febrary 2011.*

[31] B. Gorkemli, Y. Sadi, and A. Tekalp, "Effects of MGS fragmentation, slice mode and extraction strategies on the performance of SVC with medium-grained scalability," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, Sept 2010, pp. 4201–4204.

[32] P. Seeling and M. Reisslein, "Video transport evaluation with H.264 video traces," *IEEE Commun. Surv. Tutor.*, vol. 14, no. 4, pp. 1142 –1165, quarter 2012.

[33] "Efficient HTTP-based streaming using scalable video coding," *Signal Processing: Image Communication*, vol. 27, no. 4, pp. 329 – 342, 2012.

[34] H. Mansour, V. Krishnamurthy, and P. Nasiopoulos, "Channel aware multiuser scalable video streaming over lossy under-provisioned channels: Modeling and analysis," *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1366–1381, Nov 2008.

[35] M. Cesari, L. Favalli, and M. Folli, "Quality modeling for the medium grain scalability option of h.264/svc," in *Proceedings of the 5th International ICST Mobile Multimedia Communications Conference*, ser. Mobimedia '09, 2009, pp. 9:1–9:6.

[36] H. Cheng-Hsin and M. Hefeeda, "On the accuracy and complexity of rate-distortion models for fine grained scalable video sequences," *ACM Trans. on Multimedia Computing, Communications and Applications*, 2006.

[37] M. Dai, D. Loguinov, and H. Radha, "Rate-distortion analysis and quality control in scalable internet streaming," *IEEE Trans. on Multimedia*, vol. 8 issue 6, pp. 1135–1146, 2006.

[38] K. Do-Kyoung, S. Mei-Yin, and K. C. C. Jay, "Rate control for H.264 video with enhanced rate and distortion models," *IEEE Trans. Circuits Syst. Video Tech.*, vol. 17, no.5, pp. 517–529, 2007.

[39] H. Mansour, V. Krishnamurthy, and P. Nasiopoulos, "Rate and distortion modeling of medium grain scalable video coding," in *Proc. 15th IEEE Int. Conf. Image Processing ICIP 2008*, 2008, pp. 2564–2567.

[40] K. Stuhlmuller, N. Farber, M. Link, and B. Girod, "Analysis of video transmission over lossy channels," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 6, pp. 1012–1032, 2000.

[41] H. Seferoglu, O. Gurbuz, and Y. Altunbasak, "Rate-distortion based real-time wireless video streaming," *Elsevier Signal Processing: Image Communication*, vol. 22 Issue 6, pp. 529–542, 2007.

[42] D. Munaretto, D. Jurca, and J. Widmer, "A fast rate-adaptation algorithm for robust wireless scalable streaming applications," in *Wireless and Mobile Computing, Networking and Communications, 2009. WIMOB 2009. IEEE International Conference on*, Oct 2009, pp. 246–251.

[43] Xiph.org video test media [derf's collection]. [Online]. Available: https://www.media.xiph.org/video/derf/

[44] Yuv video sequences [trace website]. [Online]. Available: http://trace.eas.asu.edu/yuv

[45] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[46] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[47] D.-K. Kwon, M.-Y. Shen, and C.-C. J. Kuo, "Rate control for H.264 video with enhanced rate and distortion models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 5, pp. 517–529, 2007.

[48] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H.264/AVC video database for the evaluation of quality metrics," in *Proc. IEEE Int Acoustics Speech and Signal Processing (ICASSP) Conf*, Mar. 2010.

[49] ITU-T, "Subjective video quality assessment methods for multimedia application, recommendation," *ITU-T*, p. 910, Sept. 1999.

[50] P. Chaffe-Stengel and D. N. Stengel, *Working With Sample Data: Exploration and Inference*. Business Expert Press, Aug. 2011.

[51] Y. Dodge and J. Jureckov, *Adaptive Regression*, BPOD, Ed. Springer, 2000.

[52] x264: a free software library and application. [Online]. Available: http://www.videolan.org/developers/x264.html

[53] Y. Wang, L.-P. Chau, and K.-H. Yap, "Joint rate allocation for multiprogram video coding using FGS," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 829–837, 2010.

[54] M. Jacobs, J. Barbarien, S. Tondeur, R. Van De Walle, T. Paridaens, and P. Schelkens, "Statistical multiplexing using SVC," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, 2008*, March 2008, pp. 1–6.

[55] T. C. Thang, J.-G. Kim, J. W. Kang, and J.-J. Yoo, "SVC adaptation: Standard tools and supporting methods," *Signal Processing: Image Communication*, vol. 24, no. 3, pp. 214–228, 2009.

[56] Y. Liu, Z. G. Li, and Y. C. Soh, "Rate control of H.264/AVC scalable extension," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 1, pp. 116–121, Jan 2008.

[57] Y. P. Fallah, H. Mansour, S. Khan, P. Nasiopoulos, and H. M. Alnuweiri, "A link adaptation scheme for efficient transmission of H.264 scalable video over multirate WLANs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 7, pp. 875–887, 2008.

[58] H. Mansour, Y. P. Fallah, P. Nasiopoulos, and V. Krishnamurthy, "Dynamic resource allocation for MGS H.264/AVC video transmission over link-adaptive networks," *IEEE Trans. Multimedia*, vol. 11, no. 8, pp. 1478–1491, 2009.

[59] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks-part II: algorithm development," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 625–634, 2005.

[60] X. Wang and N. Gao, "Stochastic resource allocation over fading multiple access and broadcast channels," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2382–2391, 2010.

[61] X. Wang and G. B. Giannakis, "Resource allocation for wireless multiuser OFDM networks," *IEEE Trans. on Information Theory*, vol. 57, no. 7, pp. 4359–4372, 2011.

[62] J. Brehmer and W. Utschick, "A decomposition of the downlink utility maximization problem," in *Proc. 41st Annual Conf. Inform. Sciences and Syst. CISS '07*, 2007, pp. 437–441.

[63] P. Henarejos, A. I. Perez-Neira, V. Tralli, and M. A. Lagunas, "Low-complexity resource allocation with rate balancing for the MISO-OFDMA broadcast channel," *Signal Processing Elsevier*, vol. 92, no. 12, pp. 2975 – 2989, 2012.

[64] M. Mazzotti, S. Moretti, and M. Chiani, "Multiuser resource allocation with adaptive modulation and LDPC coding for heterogeneous traffic in OFDMA downlink," *IEEE Trans. Commun.*, vol. 60, no. 10, pp. 2915 –2925, october 2012.

[65] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Tran. Wireless Commun.*, vol. 4, no. 6, pp. 2726–2737, 2005.

[66] I. C. Wong and B. L. Evans, "Adaptive downlink OFDMA resource allocation," in *Proc. 42nd Asilomar Conf. Signals, Systems and Computers*, 2008, pp. 2203–2207.

[67] ——, "Optimal downlink OFDMA resource allocation with linear complexity to maximize ergodic rates," *IEEE Trans. on Wireless Commun.*, vol. 7, no. 3, pp. 962–971, 2008.

[68] S. M. Ross, *Simulation, 4th Edition.* Elseveir Academic Press, 2006.

[69] G. Eysenbach, "What is E-Health?,," *J Med Internet Res*, vol. 3, no. 2, p. e20, Jun 2001. [Online]. Available: http://www.jmir.org/2001/2/e20/

[70] K. Perakis, *Third Generation (3G) Cellular Networks in Telemedicine: Technological Overview, Applications, and Limitations.* IGI Global, 2009., 2014.

[71] M. Martini, R. S. H. Istepanian, M. Mazzotti, and N. Philip, "Robust multilayer control for enhanced wireless telemedical video streaming," *IEEE Transactions on Mobile Computing*, vol. 9, no. 1, pp. 5–16, Jan 2010.

[72] R. Paradiso, A. Alonso, D. Cianflone, A. Milsis, T. Vavouras, and C. Malliopoulos, "Remote health monitoring with wearable non-invasive mobile system: The healthwear project," in *30th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society, EMBS 2008.*, Aug 2008, pp. 1699–1702.

[73] J. Gallego, A. Hernandez-Solana, M. Canales, J. Lafuente, A. Valdovinos, and J. Fernandez-Navajas, "Performance analysis of multiplexed medical data transmission for mobile emergency care over the umts channel," *IEEE Transactions on Information Technology in Biomedicine*, vol. 9, no. 1, pp. 13–22, March 2005.

[74] C. Doukas and I. Maglogiannis, "Adaptive transmission of medical image and video using scalable coding and context-aware wireless medical networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2008, no. 1, p. 428397, 2008. [Online]. Available: http://jwcn.eurasipjournals.com/content/2008/1/428397

[75] S. Cicalò, M. Mazzotti, S. Moretti, V. Tralli, and M. Chiani, "Cross-layer optimization for m-health SVC multiple video transmission over LTE uplink," in *e-Health Networking, Applications Services (Healthcom), 2013 IEEE 15th International Conference on*, Oct 2013, pp. 212–217.

[76] M. Rumney, "3GPP LTE: Introducing single-carrier FDMA," *Agilent measurement journal*, vol. 4, pp. 18–27, 2008.

[77] H. Yang, F. Ren, C. Lin, and J. Zhang, "Frequency-domain packet scheduling for 3GPP LTE uplink," in *in Proc. of IEEE INFOCOM, 2010*, 2010, pp. 1–9.

[78] J. Lim, H. G. Myung, K. Oh, and D. J. Goodman, "Proportional fair scheduling of uplink single-carrier FDMA systems," in *Proc. IEEE 17th Int Personal, Indoor and Mobile Radio Comm. Symp.*, 2006, pp. 1–6.

[79] H. G. Myung, K. Oh, J. Lim, and D. J. Goodman, "Channel-dependent scheduling of an uplink SC-FDMA system with imperfect channel information," in *Proc. IEEE Wireless Communications and Networking Conf. WCNC 2008*, 2008, pp. 1860–1864.

[80] L. Ruiz de Temino, G. Berardinelli, S. Frattasi, and P. Mogensen, "Channel-aware scheduling algorithms for SC-FDMA in LTE uplink," in *Proc. IEEE 19th Int. Symp. Personal, Indoor and Mobile Radio Communications PIMRC 2008*, 2008, pp. 1–6.

[81] K. Elgazzar, M. Salah, A.-E. M. Taha, and H. Hassanein, "Comparing uplink schedulers for LTE," in *Proceedings of the 6th International Wireless Communications and Mobile Computing Conference*, ser. IWCMC '10, 2010, pp. 189–193.

[82] F. D. Calabrese, P. H. Michaelsen, C. Rosa, M. Anas, C. U. Castellanos, D. L. Villa, K. I. Pedersen, and P. E. Mogensen, "Search-tree based uplink channel aware packet scheduling for UTRAN LTE," in *Proc. IEEE Vehicular Technology Conf. VTC Spring 2008*, 2008, pp. 1949–1953.

[83] M. Salah, N. A. Ali, A.-E. Taha, and H. Hassanein, "Evaluating uplink schedulers in LTE in mixed traffic environments," in *Proc. IEEE Int Communications (ICC) Conf*, 2011, pp. 1–5.

[84] M. Al-Rawi, R. Jantti, J. Torsner, and M. Sagfors, "Opportunistic uplink scheduling for 3G LTE systems," in *Proc. 4th Int. Conf. Innovations in Information Technology IIT '07*, 2007, pp. 705–709.

[85] I. C. Wong, O. Oteri, and W. Mccoy, "Optimal resource allocation in uplink SC-FDMA systems," *IEEE Transactions on Wireless Communications*, vol. 8, no. 5, pp. 2161–2165, 2009.

[86] A. Ahmad and M. Assaad, "Polynomial-complexity optimal resource allocation framework for uplink SC-FDMA systems," in *Proc. IEEE Global Telecommunications Conf. (GLOBECOM 2011)*, 2011, pp. 1–5.

[87] S. Cicalò and V. Tralli, "Distortion-fair cross-layer resource allocation for scalable video transmission in OFDMA wireless networks," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 848–863, April 2014.

[88] ——, "Adaptive resource allocation with proportional rate constraints for uplink SC-FDMA systems," *Submitted to IEEE Communications Letters*, 2014.

[89] "E-UTRA: Physical layer procedures (release 10)," 3GPP TS 36.213, version 10.1.0, Tech. Rep., 2011.

[90] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2009.

[91] R. Jain, D.-M. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," *ACM Transactions on Computer Systems*, 1984.

[92] G. Hasegawa and M. Murata, "Survey on fairness issues in TCP congestion control mechanisms," *IEICE TRANSACTIONS on Communications*, vol. E84-B, no. 8, pp. 1461–1472, 2001.

[93] D. De Vleeschauwer, H. Viswanathan, A. Beck, S. Benno, G. Li, and R. Miller, "Optimization of HTTP adaptive streaming over mobile cellular networks," in *INFOCOM, 2013 Proceedings IEEE*, 2013, pp. 898–997.

[94] S. Cicalò, N. Changuel, R. Miller, B. Sayadi, and V. Tralli, "Quality-fair adaptive streaming over LTE networks," in *in Proc.of IEEE 39th International conference on Acoustic, Speech and Signal Processing*, May 2014, pp. 1–5.

[95] O. Oyman and S. Singh, "Quality of experience for HTTP adaptive streaming services," *IEEE Communications Magazine*, vol. 50, no. 4, pp. 20–27, 2012.

[96] A. E. Essaili, D. Schroeder, D. Staehle, M. Shehada, W. Kellerer, and E. Steinbach, "Quality-of-experience driven adaptive HTTP media delivery," in *IEEE Int. Conf. on Commun. (ICC 2013)*, Budapest, Hungary, Jun 2013.

[97] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427 – 1441, June 2010.

[98] S. Lederer, C. Müller, and C. Timmerer, "Dynamic adaptive streaming over HTTP dataset," in *Proceedings of the 3rd Multimedia Systems Conference*, ser. MMSys '12. New York, NY, USA: ACM, 2012, pp. 89–94. [Online]. Available: http://doi.acm.org/10.1145/2155555.2155570

[99] G. T. 36.521-1, "User equipment (UE) conformance specification, radio transmission and reception. part 1: Conformance testing," version 11.0.1 Release 11, Tech. Rep., 2013.

[100] G. T. 23.203, "Policy and charging control architecture," version 10.7.0 Release 10, Tech. Rep., 2012.

[101] I. Wong and B. Evans, *Resource Allocation in Multiuser Multicarrier Wireless Systems*. Springer, 2008.

[102] P. Henarejos, A. I. Perez-Neira, V. Tralli, and M. Angel Lagunas, "Low-complexity resource allocation with rate balancing for the miso-ofdma broadcast channel," *Signal Process.*, vol. 92, no. 12, pp. 2975–2989, dec. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.sigpro.2012.05.031

[103] S. Cicalò, V. Tralli, and A. I. Perez-Neira, "On the performance of distributed power allocation and scheduling in multi-cell OFDMA systems," in *In proc. of NEWCOM++ / COST 2100 Joint Workshop on Wireless Communications, 1 - 2 March, 2011, Paris, France*, 2011, pp. 1–6.

[104] ——, "Centralized vs distributed resource allocation in multi-cell OFDMA systems," in *Proc. IEEE 73rd Vehicular Technology Conf. (VTC Spring)*, 2011, pp. 1–6.

[105] G. T. 36.300, "Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); overall description; stage 2," version 10.8.0 Release 10, Tech. Rep., 2012.

[106] F. Boccardi and H. Huang, "Limited downlink network coordination in cellular networks," in *Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on*, Sept 2007, pp. 1–5.

[107] G. T. 36.814, "Further advancements for E-UTRA physical layer aspects (release 9)," version 1.5.2 Release 9, Tech. Rep., 2009.

[108] S. Brueck, L. Zhao, J. Giese, and M. Amin, "Centralized scheduling for joint transmission coordinated multi-point in LTE-Advanced," in *Smart Antennas (WSA), 2010 International ITG Workshop on*, Feb 2010, pp. 177–184.

[109] N. Himayat, S. Talwar, A. Rao, and R. Soni, "Interference management for 4G cellular standards [WIMAX/LTE update]," *IEEE Communications Magazine*, vol. 48, no. 8, pp. 86–92, August 2010.

[110] J. Xiang, Y. Luo and C. Hartmann, "Inter-cell interference mitigation through flexible resource reuse in OFDMA based communication networks," in *Proc. 13th European Wireless Conf., Paris, France*, Apr 2007.

[111] V. Corvino, D. Gesbert, and R. Verdone, "A novel distributed interference mitigation technique using power planning," in *Wireless Communications and Networking Conference, 2009. WCNC 2009. IEEE*, April 2009, pp. 1–6.

[112] V. Tralli, R. Veronesi, and M. Zorzi, "Power-shaped advanced resource assignment (PSARA) for fixed broadband wireless access systems," *Wireless Communications, IEEE Transactions on*, vol. 3, no. 6, pp. 2207–2220, Nov 2004.

[113] K. Son, S. Chong, and G. Veciana, "Dynamic association for load balancing and interference avoidance in multi-cell networks," *Wireless Communications, IEEE Transactions on*, vol. 8, no. 7, pp. 3566–3576, July 2009.

[114] H. Huh, H. Papadopoulos, and G. Caire, "Multiuser MISO transmitter optimization for intercell interference mitigation," *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 4272–4285, Aug 2010.

[115] M. Rahman and H. Yanikomeroglu, "Enhancing cell-edge performance: a downlink dynamic interference avoidance scheme with inter-cell coordination," *IEEE Transactions on Wireless Communications*, vol. 9, no. 4, pp. 1414–1425, April 2010.

[116] J. Papandriopoulos and J. Evans, "SCALE: a low-complexity distributed protocol for spectrum balancing in multiuser DSL networks," *IEEE Transactions on Information Theory*, vol. 55, no. 8, pp. 3711–3724, Aug 2009.

[117] "OFDM EESM simulation results for system-level performance evaluations, and text proposal for section a. 4.5 of tr 25.892," Tech. Rep.

[118] G. T. 36.104, "Evolved universal terrestrial radio access (E-UTRA); base station radio transmission and reception," version 9.5.0 Release 10, Tech. Rep., 2010.

Author's Publications List

[**J1**] Cicalò S., Haseeb A., Tralli V., "Fairness-oriented Multi-stream Rate Adaptation using Scalable Video Coding", *Elsevier Signal Processing: Image Communication, Volume 27, Issue 8, September 2012, Pages 800-813,*
doi: 10.1016/j.image.2012.01.005.

[**J2**] Cicalò, S.; Tralli, V., "Distortion-Fair Cross-Layer Resource Allocation for Scalable Video Transmission in OFDMA Wireless Networks," *IEEE Transactions on Multimedia, vol.16, no.3, pp. 848-863, April 2014*
doi: 10.1109/TMM.2014.2300442

[**J3**] Cicalò, S.; Tralli, V., "Adaptive Resource Allocation with Proportional Rate Constraints for Uplink SC-FDMA Systems," *Submitted to IEEE Communication Letters.*

[**C1**] Cicalò, S., Tralli, V., Perez-Neira, A.I., "On the performance of Distributed Power Allocation and Scheduling in Multi-Cell OFDMA Systems, *In proc. of NEWCOM++ / COST 2100 Joint Workshop on Wireless Communications, 1 - 2 March, 2011, Paris, France*,
url: http://hdl.handle.net/2117/14790

[**C2**] Cicalò, S., Tralli, V., Perez-Neira, A.I., "Centralized vs Distributed Resource Allocation in Multi-Cell OFDMA Systems," *In proc. of 2011 IEEE 73rd Vehicular Technology Conference (VTC Spring), pp. 1-6, 15-18 May 2011*,
doi: 10.1109/VETECS.2011.5956553

[**C3**] Cicalò S., Haseeb A., Tralli V., "Multi-stream Rate Adaptation Using Scalable Video Coding with Medium Grain Scalability", *Mobile Multimedia Communications: Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 79, pp. 152-167, Springer Berlin Heidelberg, 2012* ,
doi: 10.1109/VETECS.2011.595655310.1007/978-3-642-30419-4_14

[**C4**] Haseeb, A.; Martini, M.G.; Cicalo, S.; Tralli, V., "Rate and distortion modeling for real-time MGS coding and adaptation," *In proc. IEEE of Wireless Advanced (WiAd) Conference, 2012 pp.85,89, 25-27 June 2012*
doi: 10.1109/WiAd.2012.6296574

[**C5**] Cicalò, S., Mazzotti, M., Moretti, S., Tralli, V., Chiani, M., "Cross-layer optimization for m-health SVC multiple video transmission over LTE uplink," *In Proc. of 2013 IEEE 15th International Conference on e-Health Networking, Applications & Services (Healthcom), pp.212,217, 9-12 Oct. 2013*,
doi: 10.1109/HealthCom.2013.6720669

[**C6**] Cicalò, S.; Tralli, V., "Cross-layer algorithms for distortion-fair scalable video delivery over OFDMA wireless systems," *In proc. of 2012 IEEE Globecom Workshops (GC Wkshps), pp.1287,1292, 3-7 Dec. 2012*
doi: 10.1109/GLOCOMW.2012.6477767

[**C7**] Cicalò, S., Changuel N., Miller R., Sayadi B. and Tralli V., "Quality-Fair Adaptive Streaming over LTE networks", *To Appear in Proc.of IEEE 39th International conference on Acoustic, Speech and Signal Processing, 4-9 May, 2014 Florence, Italy.*