

ABSTRACT

Title of Dissertation: CROSS-LAYER RESOURCE
 ALLOCATION PROTOCOLS FOR
 MULTIMEDIA CDMA NETWORKS

Andres Kwasinski, Doctor of Philosophy, 2004

Dissertation directed by: Professor Nariman Farvardin
 Department of Electrical and Computer Engineering

The design of mechanisms to efficiently allow many users to maintain simultaneous communications while sharing the same transmission medium is a crucial step during a wireless network design. The resource allocation process needs to meet numerous requirements that are sometimes conflicting, such as high efficiency, network utilization and flexibility and good communication quality. Due to limited resources, wireless cellular networks are normally seen as having some limit on the network capacity, in terms of the maximum number of calls that may be supported. Being able to dynamically extend network operation beyond the set limit at the cost of a smooth and small increase in distortion is a valuable and useful idea because it provides the means to flexibly adjust the network to situations where it is more important to service a call rather than to guarantee the best quality.

In this thesis we study designs for resource allocation in CDMA networks carrying conversational-type calls. The designs are based on a cross-layer approach where the source encoder, the channel encoder and, in some cases, the processing gains are adapted. The primary focus of the study is on optimally multiplexing multimedia sources. Therefore, we study optimal resource allocation to resolve interference-generated congestion for an arbitrary set of real-time variable-rate source encoders in a multimedia CDMA network. Importantly, we show that the problem could be viewed as the one of statistical multiplexing in source-adapted multimedia CDMA. We present analysis and optimal solutions for different system setups. The result is a flexible system that sets an efficient tradeoff between end-to-end distortion and number of users. Because in the presented cross-layer designs channel-induced errors are kept at a subjectively acceptable level, the proposed designs are able to outperform equivalent CDMA systems where capacity is increased in the traditional way, by allowing a reduction in SINR.

An important application and part of this study, is the use of the proposed designs to extend operation of the CDMA network beyond a defined congestion operating point. Also, the general framework for statistical multiplexing in CDMA is used to study some issues in integrated real-time/data networks.

CROSS-LAYER RESOURCE
ALLOCATION PROTOCOLS FOR
MULTIMEDIA CDMA NETWORKS

by

Andres Kwasinski

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2004

Advisory Committee:

Professor Nariman Farvardin, Chairman/Advisor
Professor K. J. Ray Liu
Professor Haralabos Papadopoulos
Professor Min Wu
Professor John J. Benedetto

© Copyright by
Andres Kwasinski
2004

DEDICATION

To

My wife Mariela and daughter Victoria

ACKNOWLEDGEMENTS

This thesis is the culmination of many years of work and, as such, would not have been possible without the help, support and influence from faculty and colleagues at the University of Maryland, as well as friends and family. First and foremost I would like to thank my advisor, Professor Nariman Farvardin. The prospect of working and learning from him was the main reason why I choose the University of Maryland to pursue a doctorate. I greatly appreciate the fact that, despite his very busy schedule, he always took the time to be a true advisor to me. I highly value all what I have learned from him about research and life as a researcher in engineering. Most of all, I would like to thank him for going beyond the duties of an academic advisor and caring and support me as an individual.

I would like to thank Professor K. J. Ray Liu for trusting me to a position as laboratory manager in the Communications and Signal Processing Laboratory. This, and inviting me to be part of his research group meetings, allowed me to interact and cooperate with an active and intellectually motivating research group. I also valued his advice and support.

I would like to thank Professor Min Wu for supporting and encouraging my cooperative work with her research group and Professor Haralabos Papadopoulos for his constant support and encouragement.

I appreciate all of my committee members for serving on my thesis committee

I owe special gratitude to my friends and colleagues from the Communications and Signal Processing Laboratory. I am especially thankful to Mehdi Alasti, Vinay Chande, Zhu Han and Guan-Ming Su for their friendship, advice, help and for working together in different research projects, some of which are included in this thesis. I am especially grateful to Guan-Ming Su for providing the FGS module of the MPEG-4 encoder. After I added error resiliency and concealment functionality, this software became the core tool for the video simulation results reported in this thesis. Also, I am especially indebted to Mehdi Alasti, since the original idea for this thesis spawned from our research discussions.

Also, I would like to thank Professors Daniel Jacoby, Roxana Saint-Nom and Osvaldo Micheloud from the Buenos Aires Institute of Technology for their dedication to my undergraduate education and for their support and encouragement.

Last, but not least, I would like to specially thank my wife, Mariela, daughter, Victoria, and all my family for their support through all these years of hard work.

TABLE OF CONTENTS

List of Tables	viii
List of Figures	ix
1 Introduction	1
1.1 Motivation	1
1.2 Previous Work	6
1.3 Thesis Summary	11
2 Resource Allocation for Fixed Processing Gain DS-CDMA Based on Single State Source Coder	13
2.1 Introduction	13
2.2 System Description	14
2.2.1 Quality Goal and Quality of Service (QoS)	17
2.3 System Analysis	19
2.3.1 Ideal Power Control and Additive White Gaussian Noise Channel Case	19
2.3.2 Design Considering Channel Gain and Transmit Powers	32
2.3.3 Practical Considerations	34
2.4 Theoretical Performance Analysis	36

2.4.1	Dynamic Calls Model	38
2.5	Performance Evaluation	44
2.6	Conclusions	58
3	Resource Allocation Through Statistical Multiplexing of Multimedia Calls in Variable Processing Gain DS-CDMA	61
3.1	Introduction	61
3.2	System Model	62
3.2.1	Model Description	62
3.2.2	Video Telephony Calls as Application Example	65
3.3	CDMA Statistical Multiplexing Resource Allocation and Flow Control .	67
3.3.1	Multiuser Power and Rate Allocation	67
3.3.2	Source Encoder	68
3.3.3	CDMA Statistical Multiplexing, Flow Control and Resource Al- location	69
3.3.4	Flow Control by Transmit Bit Rate Adaptation	75
3.3.5	Flow Control by Transmit Bit Rate and Target SINR Adaptation	78
3.3.6	Rate-Distortion Data Overhead	80
3.4	Analysis for Dynamic Call Traffic and Admission Control	81
3.5	Performance Evaluation	84
3.6	Conclusions	93
4	Real-time and Data Traffic Integration	99
4.1	Introduction	99
4.2	Real-time and Data Integration Through Equivalent Bandwidth Allocation	100
4.3	Data Subsection	101

4.4	Influence of Equivalent Bandwidth Assignment on the Data Subsection	104
4.5	Real-time Traffic Subsection Dependence on Total Assigned Equivalent Bandwidth	113
4.6	Real-time/Data Traffic Integrated Congestion Relief	121
4.7	Conclusions	129
5	Other Related Work	131
5.1	Introduction	131
5.2	Resource Allocation in the Downlink of CDMA by Real-Time Source Encoder Adaptation	132
5.2.1	Resource Allocation Algorithm	136
5.2.2	Performance Bound	140
5.2.3	Performance Evaluation	143
5.3	Resource Allocation in the Downlink of Multicode CDMA for Layered and Embedded Real-Time Video	147
5.3.1	Distortion Management Algorithm	151
5.3.2	Performance Evaluation	157
5.4	Conclusions	163
6	Conclusions and Future Work	164
6.1	Conclusions	164
6.2	Future Work	166
	Bibliography	168

LIST OF TABLES

4.1	Comparison of results from Figures 4.15 and 4.16	129
5.1	Downlink CDMA Resource Allocation Algorithm	139
5.2	Barrier Method for Performance Bound	142
5.3	P_{sum} Relieve Algorithm	153
5.4	Code Assignment to Reduce Distortion	157
5.5	Distortion Reduction by Increasing Power	158

LIST OF FIGURES

1.1	Cellular network reacting to a cell outage. (a) The marked cell goes out of service. (b) The neighboring cells extend their coverage area.	4
2.1	Block diagram of the proposed system	15
2.2	Comparison of allocated power as a function of the number of users with and without using equivalent crosscorrelation.	21
2.3	Measured worst-case speech codec performance and two approximations.	24
2.4	Target SINR, in \log_2 scale, required for the distortion due to channel induced errors to be less than 3% of that of the corresponding source encoding distortion as a function of the source encoding rate.	25
2.5	Markov chain representation of the $M/M/N_L/N_L$ traffic model.	39
2.6	One realization in pseudo congestion state to study the average pseudo congestion length.	41
2.7	Relative increase in the number of users as a function of the normalized average distortion per call (top) and as a function of χ , the source packet reduction relative to the maximum rate (bottom).	47
2.8	Distortion as a function of SNR for each of the six possible operating modes, each defined by the pair (source encoding rate, channel code rate)	50
2.9	Comparison between the proposed scheme (with source rate adaptation) and an equivalent traditional CDMA system (with no adaptation).	52

2.10	Expected normalized distortions and blocking probability as a function of offered load $a = \nu/\mu$	54
2.11	Expected length of time in pseudo congested state as a function of offered load $a = \nu/\mu$	55
2.12	Expected normalized distortion while in pseudo congested state as a function of the expected duration in pseudo congested state.	56
2.13	Erlang capacity for the different operating modes and end-to-end normalized distortion for each operating mode.	57
3.1	Block diagram of the proposed system	63
3.2	Distortion-Rate performance for different video frames.	66
3.3	Total optimal equivalent bandwidth as a function of λ	74
3.4	Markov chain representation of the $M/M/N_L/N_L$ traffic model.	82
3.5	Comparison of three CDMA systems supporting video calls	87
3.6	A frame from the sequence 'Foreman' when the network operates with the scheme with no adaptation and there are 13 ongoing calls in the network.	89
3.7	A frame from the sequence 'Foreman' when the network operates with the scheme with no adaptation and there are 14 ongoing calls in the network.	90
3.8	A frame from the sequence 'Foreman' when the network operates with the scheme with no adaptation and there are 18 ongoing calls in the network.	90
3.9	A frame from the sequence 'Foreman' when the network operates with using algorithm 2 and there are 30 ongoing calls in the network.	91

3.10	A frame from the sequence ‘Foreman’ when the network operates with using algorithm 1 and there are 30 ongoing calls in the network.	91
3.11	A frame from the sequence ‘Foreman’ when the network operates with using algorithm 2 and there are 60 ongoing calls in the network.	92
3.12	Expected distortion (PSNR in dBs) as a function of the offered load. . .	93
3.13	Probability of pseudo congestion as a function of the offered load. . . .	94
3.14	Expected distortion (PSNR in dBs) as a function of probability of pseudo congestion.	95
3.15	Blocking probability as a function of the offered load.	96
4.1	Block diagram of the data section call.	101
4.2	Average data packet delay as a function of average packet arrival rate for a single data call. The total number of data calls was assumed $N_d = 30$.	106
4.3	Optimum processing gain as a function of average packet arrival rate for a single data call. The total number of data calls was assumed $N_d = 30$.	107
4.4	Maximum average packet arrival rate per data call as a function of the total equivalent bandwidth assigned to the data section for total number of data calls $N_d = 10$, $N_d = 30$ and $N_d = 50$	109
4.5	Average data packet delay as a function of the number of data calls. The average packet arrival rate per data call was assumed $\lambda_d = 100$	110
4.6	Optimum processing gain as a function of the number of data calls. The average packet arrival rate per data call was assumed $\lambda_d = 100$	111
4.7	Maximum number of data calls that can be supported as a function of the total equivalent bandwidth assigned to the data section.	112
4.8	Free distance as a function of code rate for two different RCPC codes. .	115
4.9	Averaged distortion-rate performance for two video sequences.	118

4.10	Normalized distortion as a function of the real-time traffic subsection total equivalent bandwidth.	119
4.11	Quality at the receiver (using PSNR in dBs) of a conversational video system as a function of the number of calls in the system, using the real-time traffic subsection total equivalent bandwidth Ω_r as parameter. .	120
4.12	Quality loss (in PSNR loss) versus the data section total equivalent bandwidth.	122
4.13	Normalized distortion of real time calls versus the maximum mean data packet arrival rate.	124
4.14	Normalized distortion of real time calls versus the maximum number of data calls.	125
4.15	Quality at the receiver (using PSNR in dBs) of a conversational video system as a function of the real-time subsection offered load and maximum allowed total equivalent bandwidth.	127
4.16	Data subsection mean packet delay as a function of the real-time subsection offered load and maximum allowed total equivalent bandwidth. .	128
5.1	Proposed system block diagram	133
5.2	Normalized distortion vs. number of calls	144
5.3	Normalized distortion vs. P_{max}	145
5.4	Evaluation of the proposed algorithm as compared to the performance bound.	147
5.5	Block diagram for the proposed desing.	149
5.6	Base layer initialization algorithm.	152
5.7	Resource allocation algorithm for enhancement (FGS) layer.	155

5.8	Power (solid line and scale on the left) and distortion (dashed-dotted line and scale on the right) as a function of the number of assigned codes . .	160
5.9	PSNR results for the sequences corresponding to 4 users.	161
5.10	Performance comparison of the proposed and greedy scheme.	162

Chapter 1

Introduction

1.1 Motivation

The design of mechanisms to efficiently allow many users to maintain simultaneous communications while sharing the same transmission medium is a crucial step during a wireless network design. Then, the problem of efficient distribution of network resources among the calls in service has a direct impact on the network performance. Yet, efficiency is not the only important issue in the design of the network. From the point of view of the network operator, maintaining a high utilization of the network is also important to maintain a cost-effective operation. Even more, since upgrading a network infrastructure may become an expensive venture, a network that has the ability to flexibly adapt to changing operating conditions could add to a cost-effective operation. Nevertheless, the network operator needs are not the only ones that need to be considered in the network design. Customers demand service availability at all times and with good communication quality. This introduces the classic problem in engineering design of conflicting requirements. For example, it is quite possible that a network operating with high utilization may be frequently unable to provide service or service quality may be sub par. Therefore, network need a carefully designed flow control protocol to fairly

distribute the shared resources as needed by each user while maximizing the network utilization and guarantee that communications services are provided with good quality.

Coupled with the flow control protocol, a congestion control protocol is also necessary to act when the network reaches its capacity. In general, these protocols provide functions to many very different layers of the communication stack. In order to provide these functions, network protocols have been divided into a stack of layers, each responsible for providing services to a layer of communication stack. In this approach, each layer is designed independently of the others and only contiguous layers exchange information through clearly defined interfaces. In this way, each layer knows little or nothing at all of the state of the communication link as seen by another layer. A contrasting approach, known as “cross-layer” design, is to design these protocols in such a way that their layers are designed and operate jointly. This is the approach taken in this work and, as we shall see, it results in protocols that are efficient in allocating resources, can flexibly adjust to changing operational conditions and are able to operate at high network utilization while providing continuous service with good quality.

Wireless cellular networks are normally seen as having some limit on the network capacity, in terms of the maximum number of calls that may be supported. This limit may be due to the number of available radio channels, time slots or, in the case of CDMA networks, a maximum interference level. Being able to dynamically extend network operation beyond the set limit at the cost of a smooth and small increase in distortion is a valuable and useful idea because it provides the means to flexibly adjust the network to situations where it is more important to service a call rather than to guarantee the best quality. One example of such situations is a cellular user seeing its call disconnected because there are no resources available when entering moving into a congested cell. Reserving resources at the base station may mitigate this problem but it reduces network

utilization. Instead, the network could deal with this situation by allowing a brief and controlled reduction in quality until the congestion is alleviated. Another situation when this idea can be applied is in the case of a cell that is affected by some disaster (a tornado, for example) and goes out of service (Figure 1.1(a)). When this happens, the surrounding base stations would most likely extend their radio coverage to replace the lost one (Figure 1.1(b)). This action will increase the traffic in these neighboring cells with the calls previously served by the base station that is out of service. This effect may be compounded by customers in the area of the disaster wanting to talk with loved ones. Also, the emergency personnel in the area of the disaster would increase the traffic above normal. In summary it is likely that one or many of the neighboring cells will become congested in a situation where having the best communication quality is not as important as providing service. Other potential applications of this idea include military communications and, in a broad view, dynamically reconfiguration of the network in a controlled way without the need for costly new hardware. Note that this approach is analogous to the operation of a wireline packet network when a router goes out of service. In this case, traffic is routed through other routers. These alternate routers will see an increase in traffic. If necessary, all the sources contributing traffic to a congested network node may reduce their rate, and communication quality, upon notification of the congestion [57].

Associated with the design of the network protocol (or some of its layers) is the definition of a multiple access technique to allow users to share the same transmission medium while preventing them from interfering with each other [34]. In Frequency-Division Multiple Access (FDMA) users access the transmission medium all the time, each using a different, non-overlapping fraction of the total available frequency bandwidth. In Time-Division Multiple Access (TDMA) sharing of the same transmission

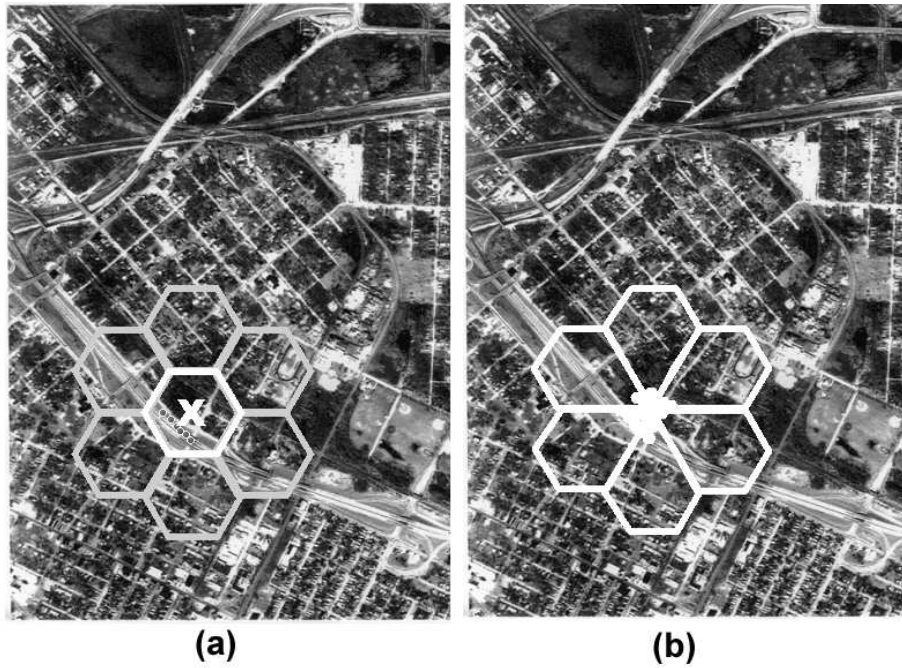


Figure 1.1: Cellular network reacting to a cell outage. (a) The marked cell goes out of service. (b) The neighboring cells extend their coverage area.

medium is achieved by allowing all the users to transmit over the whole available bandwidth on different, non-overlapping time intervals. This work will focus on a third access method, Code-Division Multiple Access (CDMA). Here, all users transmit simultaneously in the same frequency band, separation between them is achieved by making each user transmit using signals that are uncorrelated, or that have low cross-correlation, with the signals sent by the other users. In practice, perfect separation between users is not possible, thus, CDMA is said to be interference limited. Therefore, as much as TDMA and FDMA-based network protocols deal with distributing the available bandwidth between users, a primary goal in protocols intended for a CDMA systems is to control inter-user interference.

In this work we will focus on the design of flow and admission control protocols for conversational communications in CDMA networks. This implies that the proposed schemes will be constrained by strict limits on the delay. As mentioned above, our approach is that of cross-layer designs. We choose the approach because of its efficiency and flexibility. We will see that all the solutions we will present share the common characteristic that they are able to extend operation beyond congestion (defined as the maximum number of calls that can be accepted and still deliver communication service at a promised level) at the cost of a smooth and controlled degradation of quality. The contrasting approach to ours is one where calls are accepted but at the cost of a rapid increase in channel-induced errors and distortion. In our scheme, this does not occur because of the cross-layer adaptable approach and because the designs limit the channel-induced distortion to perceptually acceptable ranges. Importantly, this work lead us to develop an optimal solution to the problem of source-controlled statistical multiplexing in CDMA.

1.2 Previous Work

Since any reduction in interference translates into an increase in user capacity, a significant portion of research in CDMA has been focused on interference control and minimization. This includes techniques such as efficient processing of multipath, multi-user receivers and use of directional antennas [17, 52, 49]. Also, early works introduced the idea of taking advantage of silence periods in between talk spurts to increase the capacity [17, 52, 53]. In [51], Viterbi contended that this idea is one of the two most practical methods (sectored antennas is the other) to increase the number of calls that can be simultaneously serviced. Since it can be considered that voice activity follows a talk spurt-silence model [6], interference can be reduced by avoiding transmission during silence periods. A refinement to this idea, is to consider more than two levels of speech activity. Such is the case in the TIA/EIA IS-95 standard [23], which proposed a multi-state vocoder named QCELP. The QCELP encoder has four different operating states, each of them associated with a different level of speech energy and a different output rate (9600, 4800, 2400 and 1200 bits/sec). The lowest rate is the one associated with a silence period between talk spurts. The multi-rate source coder in IS-95 is used to reduce interference by decreasing the transmitting power of a user when source rate decreases following a reduction in speech energy. The relevant feature of these source-based ideas to reduce interference is that there is no external control on the source encoder, i.e. source coder adaptations were driven by the source, not the network. Being an evolution from IS95, these concepts and ideas have also been extended to the cdma2000 standard [1, 55]. In [44], the authors analyzed coverage and capacity in a cdma2000 network for voice and packet data services. Over the time, new vocoders, with better performance than QCELP, have been introduced with the aim of using them in new cellular standards. Yallapraga and Kripalani studied in [56] the effects that performance

gains in the GSM Advance Multi-Rate (AMR) encoder [12] and the Selectable Mode Vocoder (SMV) [15] have on voice capacity of GSM and CDMA networks. In the case of the GSM system, the multi-rate encoder is used to adapt to different channel conditions by switching between two possible encoding rates. Also, it is proposed in [56] that capacity can be increased by reducing the source encoding rate. In GSM this operation increases capacity by using up less frequency channels, time slots and by reducing interference and allowing a tighter frequency reuse. In the case of CDMA, the SMV codec also reduces interference as done in IS-95 by switching between encoding rates based on the speech characteristics. SMV differs from QCELP in that it has four different operating modes, each associated with different thresholds to switch between encoding rates, and thus it exhibits 4 different average encoding rates. In [56] these 4 different encoding rates are used to estimate their associated change in capacity. The analysis involves only calculation using bit rates as if each CDMA channel would be ideally separated from the others with no inter-user interference. Our work in this thesis goes beyond these works by addressing how a change in source encoding rate for one call affects average distortion and system interference. Our approach is novel in addressing the multiplexing dimension in CDMA using as main control element the source encoder. [7] describes yet another feature of cdma2000, the high rate packet data system IS-856 in this case, where source rate control is used to reduce interference. In this case, a bit named 'Reverse Activity' (RA) bit is set at the base station whenever the interference reaches some value. A change in the RA bit triggers a change in the transmit bit rate. Interestingly, this change is random, following an establish procedure.

Our approach, which was pioneered for a TDMA network in [4], is to add an element of control into the application layer by introducing a source encoder with externally adaptable encoding rate. In the area of adaptable CDMA, a popular research problem

has been the optimal rate and power adaptation to channel conditions. Jafar, et. al. studied in [25] the optimal transmit rate (from a discrete set) and power adaptation subject to an instantaneous BER constraint and channel conditions. Other related work studied rate adaptation for the data portion of traffic subject to the influence of voice calls. In [22], Honig and Kim studied the assignment of power and processing gain, varied by changing the symbol duration, to satisfy a target QoS, which may be bit error rate (BER) or delay. This initial work was extended in [29] to consider a dynamic algorithm to dynamically allocate power and processing gain to voice and data users. The goals of the algorithm are to minimize the total received power within a cell and maximize the average throughput for data users subject to QoS constraints for voice users. With the goal of relieving congestion in CDMA networks, Jacobsmeyer presents in [24] an heuristic solution to reduce interference, where the power and bit rate of data users is reduced while maintaining a constant energy per bit. In [42], Sampath et. al. studied total transmitted power minimization and sum of transmitted rates maximization subject to feasibility conditions on power and rates assignments. Sampath recognized that if the feasibility conditions are not met then either calls must be blocked or the conditions would have to be relaxed somehow. Since most of these works focus on the data portion of the traffic, the goal is to maximize throughput. In these works there is no mechanism to adapt voice calls to network conditions other than power assignment, blocking calls or dropping excessively delayed voice packets.

In the area of cross layer design, more recently, some works have extended the idea of power-controlled networks to include also source and channel coding. In [50], Vishwanath, Jafar and Goldsmith determined the Shannon capacity region (i.e. the set of rates that each user can achieve with arbitrarily small probability of error when assuming the best coding scheme possible and no delay constraints) in the uplink of a mul-

tiuser system. They also studied, as a function of each user channel state, power and rate allocation policies to achieve this region. In [21], the authors present a dynamic programming algorithm to optimally allocate power and source and channel coding parameters in a TDMA network with the goal of achieving delay guarantees for all channels. In [8], a variable rate video coder is used in a design for joint source rate and power allocation that maximizes end-to-end Peak Signal-to-Noise Ratio (PSNR) for a given bit SNR-source coding rate product. This design has the interest of considering a cross layer design involving source coding and power allocation, yet it lack the analysis we will present in this work, assumes homogeneous sources and does not discuss how to distribute each heterogeneous user's contribution to system interference, i.e. the multiplexing dimension is not considered.

Other researchers have focused on the integration of data traffic into the CDMA network. With the purpose of integrating variable bit rate multimedia traffic into the IMT-2000 third generation wireless system [11], Fantacci and Nannicini proposed in [13] a multiple access protocol based on dynamic reservation assignment. The proposed protocol is similar to Packet Reservation Multiple Access (PRMA) [18], [37], in that data packets are transmitted using the bandwidth not being used by real-time sources. It is also similar in that voice terminals release the channel during silence periods and contend for a new one during silence-to-talk spurts transitions in a way similar to slotted ALOHA but over a dedicated control channel and using spreading codes instead of minislots. In essence, the main difference of this work with PRMA is that, instead of allocating time-slots, the protocol allocates spreading codes to transmit in a wideband direct-sequence CDMA system.

Our research can be also considered as an application of multi-resolution coding to control user capacity of a system. Previous work on this area includes a framework that

allocates the minimum power necessary to support a given QoS to each substream [58]. This approach is presented as an alternative to unequal error protection (UEP) since the different levels of channel error protection are provided by a power assignment unique to each substream.

As we shall show, the problem that we will study in this thesis can be seen as the one of source controlled statistical multiplexing in CDMA. In this type of problem it is useful to understand what is the effects on the total system resources, and capacity, a particular setting for a user has. In an asymptotic approach that considers that the number of users and spreading signatures length grow, while keeping their ratio fixed, Tse and Hanly introduced in [48] the concept of effective interference to study the effects of interference on the system capacity considering different multiuser receivers and ideal power control. This study was extended in [60] to the case of imperfect power control.

Finally, there is also a number of previous research work linked to our research on video communication over CDMA. For a multiuser CDMA environment, Zhang *et al.* studied a video bit allocation scheme that considers source and channel coding and power allocation to minimize power consumption while satisfying maximum distortion requirements [61]. The solution first allocates resources so that all calls satisfy the maximum distortion constraint. In [62], the authors present a solution to scalable video distortion minimization subject to total transmit rate and power constraint when communication is carried over a fixed given number of CDMA channels. As we will see, our approach, especially the one presented in chapter 3 is applicable in both [62] and [8] to optimally distribute CDMA channels and system interference among calls, respectively. Also, is applicable in [61] to prevent loss of calls when it is not possible to satisfy the maximum distortion constraint.

Most of the results in this thesis have been presented previously in [31, 30, 32, 20,

33, 46].

1.3 Thesis Summary

Chapter 2 is dedicated to the analysis of a simplified case of the resource allocation problem in a DS-CDMA network through source coder adaptation. In this case all users operate with the same real-time source coders with no change in their encoding rate based on the level of source activity and with a fixed processing gain. This chapter is important in introducing some elements of our study that are common to the rest of this thesis. Such is the case of the description of the system setup and the introduction of the concept of quality goal, as well as the idea of extending operation beyond congestion by accepting a smooth degradation in quality. Chapter 3 extends the study in chapter 2 to the study of the most general problem of optimal adaptation to resolve interference-generated congestion for an arbitrary set of real-time source encoders with arbitrary Signal-to-Interference-plus-Noise ratio (SINR) goal and variable transmit bit rate (variable spreading factor). Also important is the fact that the problem setup is one of a true multimedia system, where our interpretation for this is a system where the distortion-rate performance of each source may change within two consecutive transmit periods and also within two different calls. As important result, we show that our problem, as stated in a multiuser environment subject to power feasibility constraint, can be considered as the optimal source-controlled statistical multiplexing solution in CDMA. In this chapter we also present two optimal solutions (each applicable to different system setup) to the statistical multiplexing problem. Chapter 4 uses the statistical multiplexing concepts and solutions developed in 3 to address the problem of integrating real-time traffic with data traffic. In this chapter we study the sensibility of each section, real-time

and data, to the change in the equivalent bandwidth they are assigned. We use these results to propose and study a scheme for real-time/data integrated congestion relief. While chapters 2 to 4 study problems in the uplink, chapter 5 considers problems in the downlink. In this chapter we consider both constant processing gain and a multicode CDMA system. Finally, chapter 6 discusses the most important results obtained, as well as future extension of this work to other projects.

Chapter 2

Resource Allocation for Fixed Processing Gain

DS-CDMA Based on Single State Source Coder

2.1 Introduction

In this chapter we begin the study of the resource allocation problem in a DS-CDMA network through source coder adaptation. We first consider a simplified case where all users operate with the same real-time source coders, which do not change their encoding rate based on the level of source activity, and where the processing gain is assumed constant. We will start by describing the system setup, followed by establishing the mathematical framework that will form the base for the study in this chapter. Next, we will study the condition necessary for the network to support a given number of active calls. We use this condition as the constraint of the optimization problem that adapts each call operating parameters to allow network operation that avoids congestion. Next, we present the solution to this design problem, as well as an analytical study of the relation between interference, the number of users in a CDMA network and the reconstructed source quality at the output of the source decoder. This study allows us to analyze how these variables can be properly balanced in our design. We will also

consider practical implementation issues followed by the development of useful design and modeling equations. Lastly, we discuss the teletraffic properties of our system. We finish this chapter by summarizing the main conclusions and contributions.

2.2 System Description

Consider a Direct-Sequence Code Division Multiple Access (DS-CDMA) system. Assume, also, that the only calls present are those carrying real-time conversations. Although, in general, each call could be carrying a conversational call of different media (speech, video, etc.), in this chapter we will focus exclusively in calls carrying real-time voice.

Figure refpropresfig shows the main components of the proposed system. The mobile terminals have the same basic components blocks as those found in current commercial DS-CDMA devices: a source encoder that removes unnecessary redundancy from the real-time source, a channel encoder that adds controlled redundancy to the encoded source so as to better protect it against channel errors, a spreader that implements the DS-CDMA functionality, and front-end hardware, such as the modulator and power amplifiers. As shown in Figure refpropresfig, in contrast to current commercial devices, our system introduces a source encoder that has as key property that the output rate can be externally controlled. This can be implemented using either variable rate or embedded encoders. In the former case, the coder generates one bit stream for each of the possible encoding rates. Only one of these will be selected and transmitted based on a control signal that indicates the rate assignment. When compared to embedded encoders, variable rate encoders typically have better coding efficiency. On the other hand, using embedded encoders presents the advantage that only one bit stream is gen-

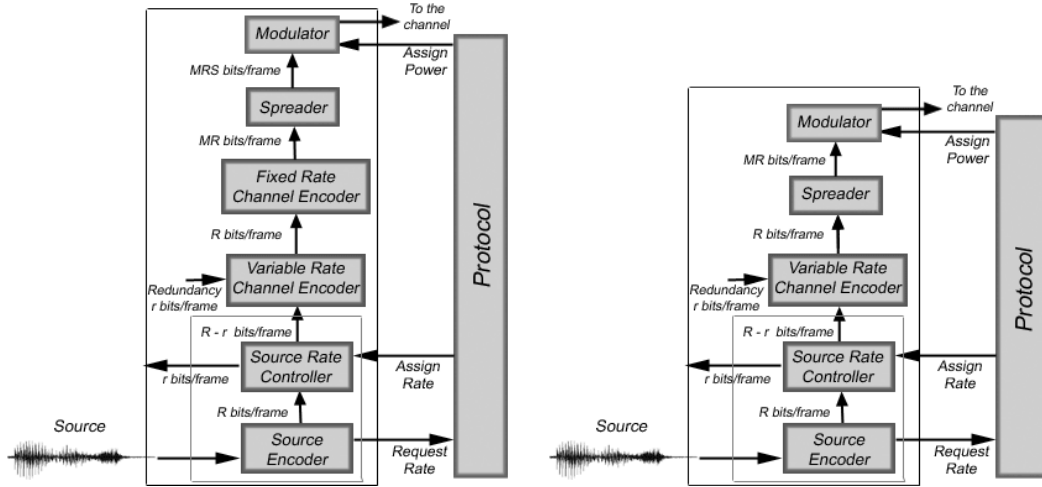


Figure 2.1: Block diagram of the proposed system

erated, making the adaptation to the rate assignment simple by dropping as many bits as necessary from the end of the bit stream. Although the “bit dropping mechanism” is exclusive to the embedded stream, we will loosely use this term to mean a reduction in the source rate regardless of the particular source encoder implementation.

In the proposed system, the source encoder output is divided into variable rate source frames. The length of these frames depends on the source encoding rate. Only one source frame is sent per transmission period. While in current commercial DS-CDMA systems the source frames are error protected by a fixed-rate channel encoder (followed by bit repetition if needed), in the proposed scheme source frames are error protected by a variable-rate channel encoder. In this chapter we will consider a system that provides equal error protection for all source bits. Source and channel rate allocation is determined so that any reduction in source encoding rate is matched by an increase in error protection in such a way that the bit rate at the output of the variable-rate channel encoder remains constant. This means that all bits dropped from the source are replaced by error protection bits from the variable-rate channel encoder and that the bitrate at

the output of the variable-rate channel encoder is equal to the product of the maximum source encoder output rate and the maximum channel coding rate. This follows the fixed processing gain assumption and it also assures that delay is kept fixed and small, as required by real-time services. Note that this operation implies that the level of error protection is increased as the source rate is reduced. This approach also simplifies implementation on current system by modifying only those baseband (often software-based) processing unit that normally precedes the spreader and other physical link units. In fact, as shown in Figure refpropresfig (a), one possible implementation may concatenate the variable-rate channel encoder to the already existing fixed-rate one. Not shown in Figure 2.1, is the assumed fact that the source encoder output contains also some bits used to detect frames in error through (assumed ideal) cyclic redundancy check (CRC).

In a general setup the source encoder may be multi-state. This means that the maximum encoding rate will change based on the level of source activity. For the purpose of the current system description we will continue to assume that the source encoder is multi-state. Nevertheless, for the rest of this chapter we will assume that the source is such that the encoder operates always in the same state and we will leave for the next chapter to study the more general problem where each encoder will change state. Since the whole system behavior will follow the state of the encoder we will use the terms "encoder state" and "call state" interchangeably.

A flow control protocol is located in a centralized position (namely the base station) and communicates with all the mobiles in the coverage area. In each transmission period, each mobile terminal sends not only the encoded source data sampled during the previous period but also the rate requirement necessary to transmit the source data sampled during the current period. In effect, transmission of a source frame is delayed by one frame duration with respect to the time when the data was sampled. It is the protocol

job to allocate rate and power to all the users based on the traffic demand estimated from their rate requirements in such a way that average distortion per call is minimized. Rate assignment is communicated to each active user jointly with the power assignment so as to proceed with transmission.

2.2.1 Quality Goal and Quality of Service (QoS)

An important goal for practical implementation of this scheme is to guarantee and provide good communication quality. To quantify quality we will measure the end-to-end distortion, i.e. that composed of the source encoding distortion and the distortion introduced by channel errors.

In general, we have observed during simulations that, for the same distortion measure, channel-induced errors are perceptible more annoying than source encoding distortion. This is because channel-induced errors tend to concentrate the distorting effect on a group of samples as opposed to spread the distortion at lower level to all samples in a frame. This manifests as an artifact that many times is perceptually evident and annoying and that in some cases might affect the understandability of the message. Thus, in this chapter we will have as design goal to keep the relative contribution of channel-induced errors to the end-to-end distortion below a value small enough so that it remains perceptually acceptable. This goal, that we will call the *quality goal*, will determine the design values for target SINR.

We intentionally avoid calling this goal a ‘Quality of Service’ (QoS) goal. This is because, for the type of problems we are studying, the definition of QoS tends to vary for different publications depending on the particular approach. In the present case, the most appropriate definition of QoS is that of the end-to-end quality. We will see that in our design the end-to-end quality will not be kept fixed. In fact, end-to-end

quality will be controllably traded for an increase in the supported calls. Nevertheless, we will also see that our design will be able to effectively maintain the contribution of channel-induced distortion below the threshold when it becomes annoying or it affects the understandability of the message. Meeting this goal will mean that the communication quality remains perceptually good or acceptable, thus the chosen name ‘quality goal’ for this design objective.

In essence, considering a quality goal in terms of the contribution of channel-induced errors is equivalent to the practice in wireless design of meeting a target maximum Frame Error Rate (FER). Nevertheless, our goal not only maintains a closer relation to the end-to-end distortion measure, but it also better considers the source encoder sensitivity to channel errors at different encoding rates. Note that in our scheme the target Signal-to-Interference-plus-Noise ratio (SINR) necessary to meet the quality goal is a function of the source rate. This is because reducing source rate corresponds to increasing the source frame error protection, which, in turn, lowers the SINR needed to achieve the same quality goal. Thus, by reducing some or all calls’ source rate, it is possible to lower transmit power and interference, resolving congestion at the cost of higher source encoding distortion but without increasing channel induced errors. Then, in our scheme the cost of increasing the number of calls beyond congestion is a degradation of the average received source quality. Then, the design problem is to determine what calls should be affected and with what magnitude so as to resolve congestion while meeting the quality goal and minimizing average end-to-end distortion per call.

2.3 System Analysis

2.3.1 Ideal Power Control and Additive White Gaussian Noise Channel Case

Consider the uplink of a chip-sampled Direct-Sequence CDMA system, as in Figure 2.1, with ideal power control and subject to an additive, white, Gaussian noise (AWGN) channel. Assuming that a matched filter is used at the receiver, power assignments and interference from other users are related to the target SINR for each of the N users by,

$$\beta_i \geq \frac{P_i}{\sigma^2 + \sum_{\substack{j=1 \\ j \neq i}}^N P_j \gamma_{ij}^2}, \quad i = 1, 2, \dots, N, \quad (2.1)$$

where P_i is the power assigned to user i , as measured at the receiver, necessary to obtain the target SINR β_i , σ^2 is the channel's noise variance, and γ_{ij} is the crosscorrelation between users i and j unit-energy spreading sequences. For an asynchronous system γ_{ij}^2 models the sum of the squares of the left, right and same-bit correlations. In a synchronous system γ_{ij} models the correlation between pseudo-random spreading sequences or the

If it is possible to find feasible power assignments that satisfies the N inequalities (2.1) with equality, then these assignments minimizes the sum of the transmitted powers, [42]. Therefore, we will consider (2.1) as an equality.

We next assume that our system is able to estimate in each frame an “equivalent crosscorrelation” γ such that equation (2.1) can be equivalently written as [58],

$$\beta_i = \frac{P_i}{\sigma^2 + \gamma^2 \sum_{\substack{j=1 \\ j \neq i}}^N P_j}, \quad i = 1, 2, \dots, N. \quad (2.2)$$

We have found from simulations that this assumption is reasonable. Figure 2.2 shows one of these simulations. In the figure we show the power allocated to a user as the number of calls is increased. The curve labeled “From estimate” shows the result calculated using the equivalent crosscorrelation. The curve “Real” shows the result calculated using the actual crosscorrelations between each call pair. In the simulations we modeled the actual crosscorrelations as a Gaussian random variable with zero mean and variance equal to the inverse of the spreading sequence length (assumed equal to 64), [59]. To estimate the equivalent crosscorrelation we assume that it was possible to obtain a sample of the total interference, with this result the equivalent crosscorrelation was estimated so that $\sum^N P_j \gamma_{ij}^2 = \gamma^2 \sum^N P_j$. We can see from Figure 2.2 that not only the result using equivalent crosscorrelation is close enough to the real one but, more importantly, using equivalent crosscorrelation allows for a good estimation of the point when the power allocation start to grow rapidly. As we shall see in the next pages, this condition determines the congestion point in the CDMA network.

Writing equations (2.2) as a function of the unknowns $P_i, i = 1, 2, \dots, N$, we obtain the linear system,

$$\mathcal{M}\mathbf{P} = \sigma^2\mathbf{1}, \quad (2.3)$$

where $\mathbf{1} = [1, 1, \dots, 1]_{1 \times n}^T$, $\mathbf{P} = [P_1, P_2, \dots, P_n]^T$ and

$$[\mathcal{M}_{ij}] = \begin{cases} \frac{1}{\beta_i} & \text{if } i = j, \\ -\gamma^2 & \text{if } i \neq j. \end{cases} \quad (2.4)$$

Solving this linear system we obtain the power assignment,

$$P_i = \frac{\Psi_i \sigma^2}{\gamma^2 \left(1 - \sum_{j=1}^N \Psi_j\right)}, \quad i = 1, 2, \dots, N, \quad (2.5)$$

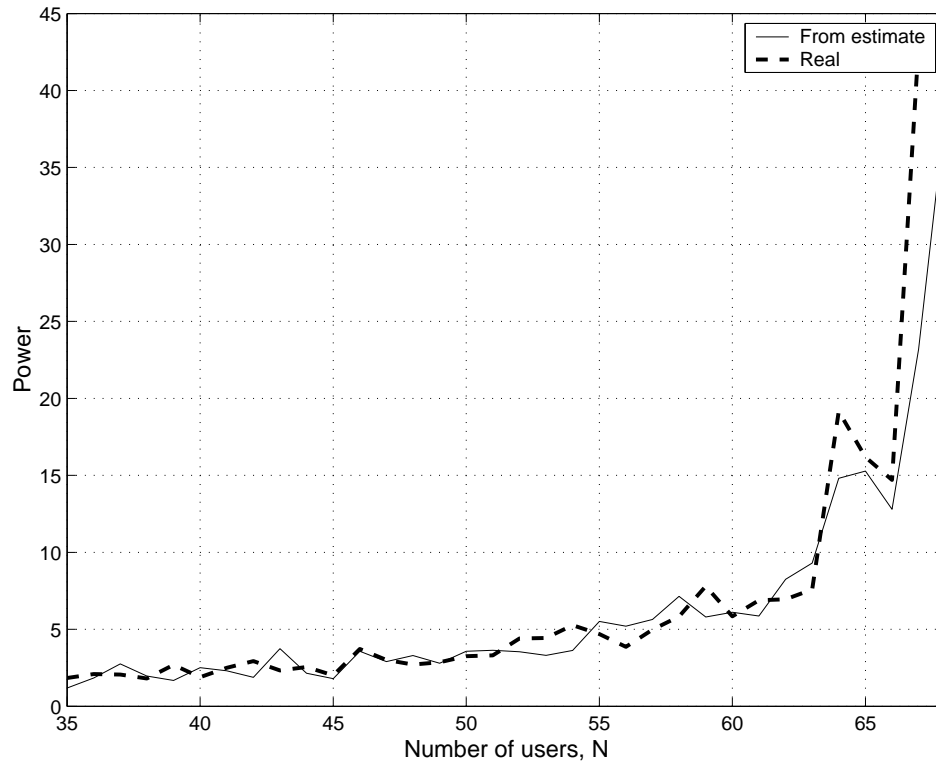


Figure 2.2: Comparison of allocated power as a function of the number of users with and without using equivalent crosscorrelation.

where

$$\Psi_i = \left(1 + \frac{1}{\gamma^2 \beta_i}\right)^{-1}. \quad (2.6)$$

The solution in (2.5) is the power at the receiver necessary to obtain a SINR such that the expected channel induced distortion remains below a preset limit. It is clear that we need $\sum_{j=1}^N \Psi_j \leq 1$ for the power to be positive. This condition determines the maximum number of users that can be accepted into the system. Furthermore, when $\sum_{j=1}^N \Psi_j \approx 1$ the power assignments in (2.5) are likely to be too large to be practically feasible. Thus, the dynamics of the system are such that, as more users are admitted into the system, $\sum_{j=1}^N \Psi_j$ grows up to a point where it exceeds a threshold $1 - \epsilon$, ϵ being a small positive number set during design, that is the congestion point. Therefore, a practical limit on the user capacity will be the one determined by the condition

$$\sum_{i=1}^N \Psi_i \leq 1 - \epsilon. \quad (2.7)$$

We note here that equation (2.5) is equivalent to equation (4) in [43] and the condition (2.7) is equivalent to (6) in the same reference. As is the case for equation (2.7) above, (6) in [43] is also intended to represent limitations on the receiver's dynamic range and system stability ([52]). Both equations differ in that we have used the equivalent crosscorrelation instead of the processing gain. Our intention in doing so is to keep the formulation sufficiently general so as to be able to also consider practical issues that may cause the equivalent crosscorrelation to change over time and to differ from the processing gain.

As explained in Section 2.2.1. The goal of our protocol is to resolve congestion by renegotiating the target SINR (by reducing source rate) so as to bring the system back to the operating point where (2.7) holds. Furthermore, we seek a rate adaptation rule that is optimized in the sense that minimizes the average distortion per call. Let $f_i(x_i)$ be

the distortion-rate (D-R) performance function of the i^{th} user source encoder encoding at rate x_i . Then, the optimization goal can be equivalently written as

$$\min_{x_1, x_2, \dots, x_N} \sum_{i=1}^N f_i(x_i), \quad (2.8)$$

In what follows we will assume that all users use the same single state source encoder, i.e. the encoding rate does not change as a function of the source level of activity. Also, it is reasonable to assume that $f(x)$ is convex and decreasing, as is the case for most well designed source encoders. For simplicity, we will also assume that the rate can be changed continuously without bounds, i.e. $x_i, \forall i = 1, 2, \dots, N$ are real numbers. Obviously, $f(x)$ is minimum when the rate is maximum; $x_i = x_M$, which would be the rate assignment for all users in the absence of congestion. Furthermore, we assume $f(x) = \alpha 2^{-2kx}$. This is a very general form for D-R performance functions that applies for the case of Gaussian sources with squared-error distortion and when the high-rate approximation holds, [16, 10]. In practice, real encoders are very complex; thus, their distortion-rate functions do not strictly follow this rule for all encoding rates, [62]. Nevertheless, we noticed that by carefully choosing α and k , this function is a good representation of an upper bound for the real distortion-rate characteristic and can be safely used as the representation for the worst-case behavior of the encoder. This fact is further illustrated in Figure 2.3. Here we show the measured performance of the GSM Advanced Multirate (AMR) Narrowband speech codec, as well as two approximations. The overall configuration for the measurement was the same as the one explained in detail in Section 2.5. The measured performance represents the worst-case result, i.e. for each encoding rate we only show the result among the many speech sequences with the highest distortion. Of the two approximations shown, we want to focus here only on

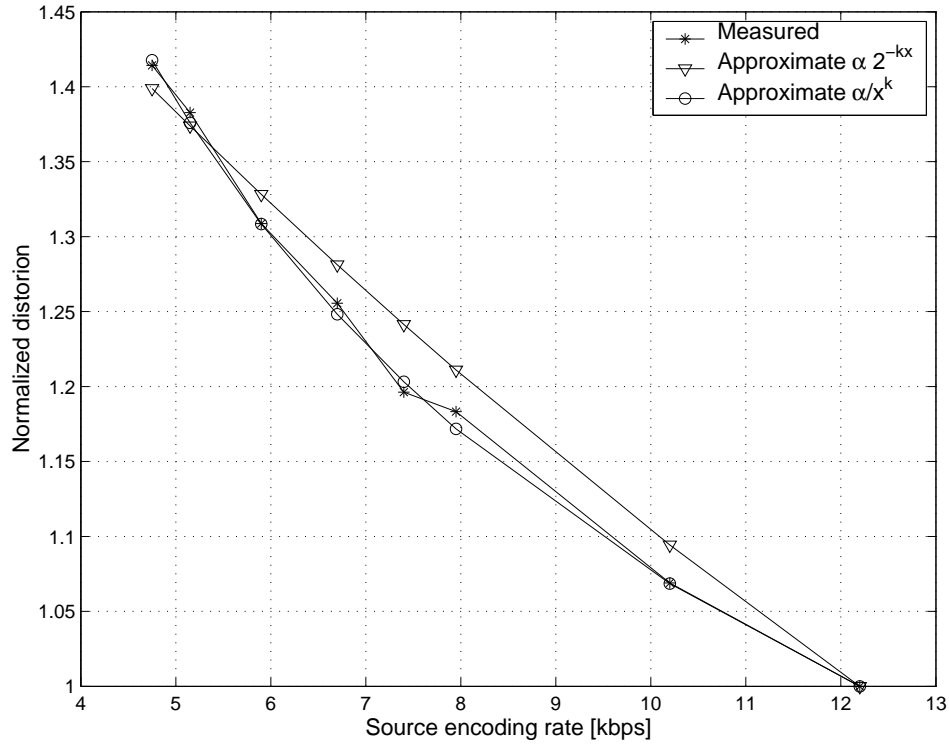


Figure 2.3: Measured worst-case speech codec performance and two approximations.

the one of the form $\alpha 2^{-2kx}$. Therefore, our goal (2.8) can be equivalently written as

$$\min_{x_1, x_2, \dots, x_N} \sum_{i=1}^N \alpha 2^{-2kx_i}. \quad (2.9)$$

Note that we are considering a distortion-rate function that accounts only for source encoding distortion. Nevertheless, will remain applicable to end-to-end distortion if the design meets the quality goal that keeps channel induced errors at a small fraction of the overall distortion.

The condition (2.7) for system stability and power amplifiers dynamic range in (2.7) imposes a constraint on the design problem. This condition implicitly depends on the target SINR, which is design to meet the quality goal and is a function of the source

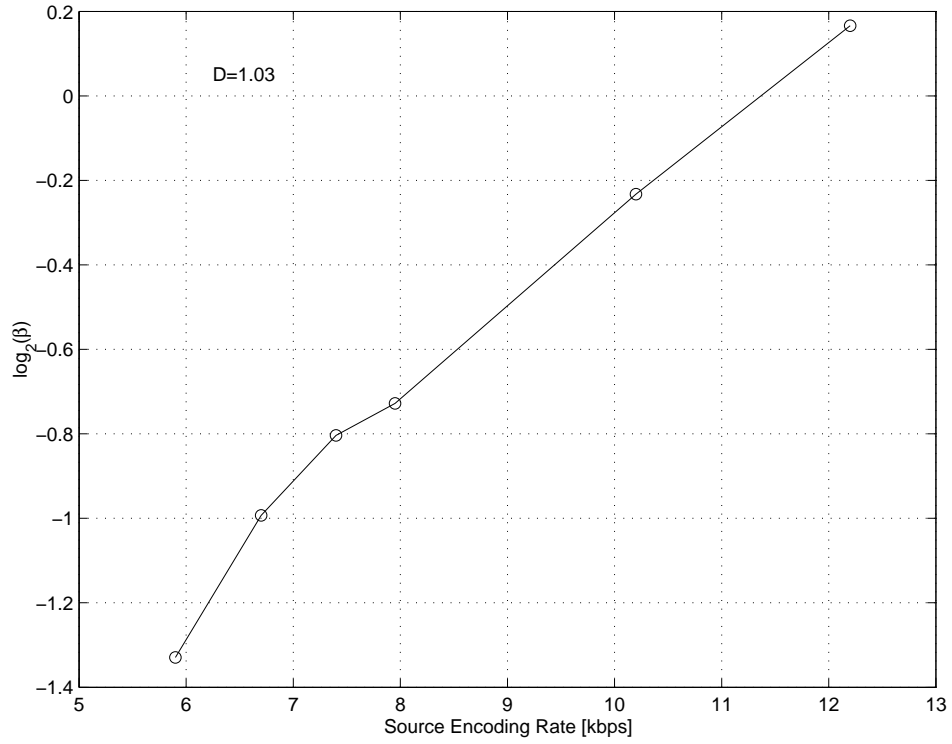


Figure 2.4: Target SINR, in \log_2 scale, required for the distortion due to channel induced errors to be less than 3% of that of the corresponding source encoding distortion as a function of the source encoding rate.

rate. With the goal of finding an analytical expression that approximates this function, we performed a series of simulations to evaluate what is the target SINR required to meet the quality goal. The setup was common to all simulations and is described in Section 2.5. Figure 2.4 is one of these simulations. It shows, as a function of the source encoding rate, the target SINR required for the distortion due to channel induced errors to be less than 3% of that of the corresponding source encoding distortion. We concluded from the simulations that a reasonable approximation for the relation between target SINR and source encoding rate is

$$\beta_i = 2^{Ax_i+B}. \quad (2.10)$$

The parameters A and B , which depend on the error control coding scheme, if we fixed the source encoder, are obtained from the simulations. Note that this is a convex function that increases with the source rate. Using this expression in (2.7) we get

$$\sum_{i=1}^N \frac{1}{2^B \gamma^2 + 2^{-Ax_i}} = \frac{1 - \epsilon}{2^B \gamma^2}, \quad (2.11)$$

which leads into the following proposition:

Proposition 1 *The problem of optimal rate adaptation that resolves congestion is*

$$\min_{x_1, \dots, x_N} \sum_{i=1}^N \alpha 2^{-2kx_i} \quad \text{subject to} \quad \sum_{i=1}^N \frac{1}{2^B \gamma^2 + 2^{-Ax_i}} = 2^{-B} \gamma^{-2} (1 - \epsilon). \quad (2.12)$$

The solution to this problem is

$$x_i = x^* = \frac{1}{A} \log_2 \left[\frac{2^{-B}}{\gamma^2} \left(\frac{N}{1 - \epsilon} - 1 \right)^{-1} \right], \quad \forall i, \quad (2.13)$$

which corresponds to a target SINR assignment to all calls equal to

$$\beta^* = \frac{1}{\gamma^2} \left(\frac{N}{1 - \epsilon} - 1 \right)^{-1} \quad (2.14)$$

Proof: Let

$$\Upsilon = \frac{1 - \epsilon}{2^B \gamma^2} \quad (2.15)$$

and

$$y_i = \frac{1}{2^B \gamma^2 + 2^{-Ax_i}}. \quad (2.16)$$

Then, since the distortion-rate function can be written as

$$f(y_i) = \alpha (y_i^{-1} - 2^B \gamma^2)^{2k/A}, \quad (2.17)$$

the problem in Proposition 1 becomes

$$\min_{y_1, y_2, \dots, y_N} \sum_{i=1}^N \alpha (y_i^{-1} - 2^B \gamma^2)^{2k/A} \quad \text{subject to} \quad \sum_{i=1}^N y_i = \Upsilon. \quad (2.18)$$

It is easy to show that in the presence of congestion, average distortion is minimized when the constraint (2.7) is active [45]. Note that k is positive if the D-R performance would be nonincreasing with the rate. Also, A is positive for the error protection to increase with decreasing channel coding rate. Thus, $f(y_i)$ is always a convex function in practice. Consider now the constraint $\sum_{i=1}^N y_i = \Upsilon$ with $N = 2$, i.e. $y_1 + y_2 = \Upsilon$. Because $f(y_i)$ is convex, we have in (2.8), as a function of y_i ,

$$f(y_1) + f(y_2) \geq 2f\left(\frac{y_1 + y_2}{2}\right) = 2f\left(\frac{\Upsilon}{2}\right). \quad (2.19)$$

Then, $\min_{y_1, y_2} [f(y_1) + f(y_2)] = 2f(\Upsilon/2)$, with the optimal assignment $y_1 = y_2 = \Upsilon/2$, or, in the general case when there are N terms in the sum,

$$y_i = \frac{\Upsilon}{N} = \frac{1 - \epsilon}{N 2^B \gamma^2},$$

Furthermore, since x_i and y_i are related by the one-to-one expression $x_i = (1/A) \log_2(y_i^{-1} - 2^B \gamma^2)^{-1}$, (2.9) is minimized by choosing all source rates equal to

$$x_i = x^* = \frac{1}{A} \log_2 \left[\frac{2^{-B}}{\gamma^2} \left(\frac{N}{1 - \epsilon} - 1 \right)^{-1} \right]. \quad (2.20)$$

Also, this implies that the optimal target SINR assignment β^* is the one where all users are assigned the same target SINR, $\beta_i = \beta^*$ and, from (2.5), the same received power. Then, from (2.7), we have

$$\frac{N}{1 + \frac{1}{\gamma^2\beta}} = 1 - \epsilon, \quad (2.21)$$

and

$$\beta^* = \frac{1}{\gamma^2} \left(\frac{N}{1 - \epsilon} - 1 \right)^{-1}. \quad (2.22)$$

□

Let N_M be the maximum number of users that can be supported without congestion, i.e. those that can be supported when the source encoding rate is x_M and target SINR is $\beta_M = 2^{Ax_M+B}$.

Proposition 2 β^* and N_M can be approximated as

$$\beta^* \approx \frac{1 - \epsilon}{\gamma^2 N} \quad (2.23)$$

$$N_M \approx \frac{1 - \epsilon}{\beta_M \gamma^2}. \quad (2.24)$$

Proof: Considering β^* from Proposition 1, its approximation follow from N being generally one or two orders of magnitude larger than 1 and $1 - \epsilon$ being close to 1. Proof for N_M follows from the fact that all users operate with maximum target SINR when source encoding rate is maximum. □

Let D_N be the average distortion per call when there are N ongoing calls and D_N^* be the minimum attainable average distortion when the rate follows the optimal assignment in Proposition 1 (i.e. when assignment is such that condition (2.7) holds). Also, let $\delta = \alpha 2^{-2kx_M}$ be the absolute minimum distortion. Then,

Proposition 3 *The number of supported calls is related to the minimum attainable average normalized distortion per call through the expression*

$$\frac{N}{N_M} \approx \left(\frac{D_N^*}{\delta} \right)^{A/(2k)}. \quad (2.25)$$

Also, the number of users in the system is related to the rate reduction per call through the expression

$$\frac{N}{N_M} \approx 2^{Ax_M\chi}, \quad (2.26)$$

where $\chi = 1 - x/x_M$ is the source packet reduction relative to the maximum rate.

Proof: From Proposition 1, all source encoders are assigned the same rate. Then $D_N = f(x) = \delta 2^{2k(x_M - x)}$ and

$$D_N^* = f(x^*) = \delta \left[2^{(Ax_M + B)} \gamma^2 \left(\frac{N}{1 - \epsilon} - 1 \right) \right]^{(2k/A)}. \quad (2.27)$$

Next, as in Proposition 2, we can approximate

$$D_N^* \approx \delta 2^{\frac{2k}{A}(Ax_M + B)} \left(\frac{N\gamma^2}{1 - \epsilon} \right)^{(2k/A)}. \quad (2.28)$$

Since $\beta_M = 2^{Ax_M + B}$, we have

$$D^* \approx \delta \left(\frac{\beta_M \gamma^2 N}{1 - \epsilon} \right)^{2k/A}, \quad (2.29)$$

and using the approximation for N_M in Proposition 2 we get

$$D_N^*/\delta \approx (N/N_M)^{(2k/A)}. \quad (2.30)$$

To prove the second part of the proposition consider that, using both approximations in Proposition 2, we have $\frac{N}{N_M} \approx \frac{\beta_M}{\beta^*} = \frac{2^{(Ax_M + B)}}{2^{(Ax^* + B)}} = 2^{Ax_M\chi}$. \square

As discussed above, our distortion-rate assumption follows a well-known function in rate distortion theory that, for the case of complex encoders, can be used as a good upper

bound for the worst-case encoder performance. In the process of the same simulations we carried out to test this function, we found that a function that represents a tighter bound is $f(x) = \alpha x^{-k}$, with $k > 0$. This can be seen in the example codec performance shown in Figure 2.3. Consequently, we study next how results are affected when using this approximation.

Proposition 4 *The solution to the problem of optimal rate adaptation to resolve congestion when $f(x) = \alpha x^{-k}$, $k > 0$, is*

$$x_i = x^* = \frac{1}{A} \ln \left[\frac{e^{-B}}{\gamma^2} \left(\frac{N}{1-\epsilon} - 1 \right)^{-1} \right], \quad \forall i, \quad (2.31)$$

which corresponds to a target SINR assignment $\beta^* = \frac{1}{\gamma^2} \left(\frac{N}{1-\epsilon} - 1 \right)^{-1}$ to all calls. Furthermore, with this allocation the average distortion per call is

$$D_N^* = \delta \left[\frac{Ax_M}{\ln \left(\frac{e^{-B}}{\gamma^2} \left(\frac{N}{1-\epsilon} - 1 \right)^{-1} \right)} \right]^k. \quad (2.32)$$

Also, the number of supported calls is related to the minimum attainable average normalized distortion per call through the expression,

$$\frac{N}{N_M} \approx e^{Ax_M(1-(D/\delta)^{-1/k})}, \quad (2.33)$$

$$(2.34)$$

and to the rate reduction per call through

$$\frac{N}{N_M} \approx e^{Ax_M \chi}. \quad (2.35)$$

Proof: For simplicity, we use a different base for the exponent of the target SINR-source encoding rate function (2.10); $\beta_i = e^{Ax_i+B}$. When using these functions, equation (2.18)

becomes

$$\min_{y_1, y_2, \dots, y_N} \sum_{i=1}^N \frac{\alpha A^k}{\left[\ln (y_i^{-1} - e^B \gamma^2)^{-1} \right]^k} \text{ subject to } \sum_{i=1}^N y_i = \Upsilon. \quad (2.36)$$

with $y_i^{-1} = \gamma^2 e^B + e^{-Ax}$ and

$$\Upsilon = \frac{1 - \epsilon}{e^B \gamma^2}. \quad (2.37)$$

The proof follows as in Propositions 1 and 3 where we only need to show that the function

$$f(y) = \left[\ln (y^{-1} - e^B \gamma^2)^{-1} \right]^{-k} \quad (2.38)$$

is convex. This is equivalent to show that $d^2 f / dy^2 > 0$ for all values of y in the domain of f .

Let $\zeta = e^B \gamma^2$, then, it is straightforward to show, after algebraic operations, that

$$\frac{d^2 f}{dy^2} = \frac{k y^{-3}}{(y^{-1} - \zeta) \left[\ln (y^{-1} - \zeta)^{-1} \right]^{(k+1)}} \left(2 + \frac{\frac{k+1}{\ln(y^{-1} - \zeta)^{-1}} - 1}{y^{-1} (y^{-1} - \zeta)} \right) \quad (2.39)$$

Since $y = (\gamma^2 e^B + e^{-Ax})^{-1}$, $y_i^{-1} - \zeta = e^{-Ax} > 0$ and $y > 0$, using (2.39) the condition $d^2 f / dy^2 > 0$ becomes,

$$2 + (\zeta + e^{-Ax}) e^{Ax} \left(\frac{k+1}{Ax} - 1 \right) > 0 \quad (2.40)$$

Thus,

$$\frac{k+1}{Ax} > \frac{-2}{(\gamma^2 e^{Ax+B} + 1)} + 1 \quad (2.41)$$

Since the left hand side member is always positive, f is convex if the right hand side member is less than or equal to zero. For this to occur we need that $\gamma^2 e^{Ax+B} + 1 < 2$, or $e^{Ax+B} < 1/\gamma^2$. Since e^{Ax+B} is the target SINR and γ^2 is a number in the order of 10^{-2} ,

f would be convex if the target SINR is less than 20dB approximately, which is a very likely condition. In the unlikely case that

$$\frac{-2}{(\gamma^2 e^{Ax+B} + 1)} + 1 > 0,$$

the convexity of f depends on the design parameters k (from the source encoder) and A (from the channel encoder). \square

2.3.2 Design Considering Channel Gain and Transmit Powers

The equations developed so far follow a simplified approach that focus on studying the core dynamics of the system. However, they don't consider channel conditions and constraints on each user's transmit power. Specifically, users undergoing large channel attenuation may be unable to transmit at a power level that meets the optimal target SINR (2.22). Although this condition is not desirable, it is practically unfeasible to avoid this situation due to limits on the mobiles' transmit power, battery life, etc. These users will not only be forced to transmit at a lower SINR, thus experiencing higher distortion, they will also generate less interference to others, thus making it possible for the rest of the users to increase their transmit power and reduce end-to-end distortion.

Consider equation (2.1) modified so as to consider that users transmit with power T_i over a channel with power gain h_i^2 :

$$\beta_i = \frac{T_i h_i^2}{\sigma^2 + \gamma^2 \sum_{\substack{j=1 \\ j \neq i}}^N h_j^2 T_j}, \quad i = 1, 2, \dots, N \quad (2.42)$$

Following the same procedure as before, the transmit power assignment, T_i , to user i necessary to obtain the target SINR is,

$$T_i = \frac{\Psi_i \sigma^2}{\gamma^2 h_i^2 (1 - \sum_{j=1}^N \Psi_j)}, \quad i = 1, 2, \dots, N \quad (2.43)$$

Note that we have the same system feasibility condition as in (2.7). In addition to this constraint, there is also a limit on each user's maximum transmit power T_M , i.e. $0 \leq T_i \leq T_M$. From (2.43), and assuming that the target SINR for all users have been set so that (2.7) holds true, this limit becomes

$$0 \leq \frac{\sigma^2}{h_i^2} \left(\frac{1}{\beta_i} + \gamma^2 \right)^{-1} \leq T_M \epsilon, \quad (2.44)$$

or, writing this condition in terms of the SINR,

$$\beta_i \leq \frac{h_i^2 T_M \epsilon}{\sigma^2 - \gamma^2 h_i^2 T_M \epsilon} \triangleq \hat{\beta}_i. \quad (2.45)$$

Let \mathcal{S} be the set of users' indices such that $\hat{\beta}_i \leq \beta^*$ and assume it has cardinality N' . This set represents the users that are unable to transmit at a power level high enough to reach the optimal target SINR β^* . Modifying equation (2.7) to separate the sum in two terms, one for those users in the set \mathcal{S} and one for those that are not,

$$\sum_{\substack{i=1 \\ i \notin \mathcal{S}}}^N \left(1 + \frac{1}{\gamma^2 \beta_i} \right)^{-1} + \sum_{\substack{i=1 \\ i \in \mathcal{S}}}^N \left(1 + \frac{1}{\gamma^2 \beta_i} \right)^{-1} = 1 - \epsilon. \quad (2.46)$$

Since each user's distortion is minimized by setting a target SINR as large as possible (i.e. larger source encoding rate), the target SINR of users in the set \mathcal{S} is set by taking equation (2.45) with equality, i.e. $\beta_i = \hat{\beta}_i$. Since, for users in \mathcal{S} , $\hat{\beta}_i \leq \beta^*$, the reduction in target SINR reduces also the system interference and, thus, allows an increase in the target SINR for the users not in the set \mathcal{S} (to reduce distortion). The new (higher) target SINR for users not in \mathcal{S} is calculated by modifying equation (2.46) as

$$\sum_{\substack{i=1 \\ i \notin \mathcal{S}}}^N \left(1 + \frac{1}{\gamma^2 \beta_i} \right)^{-1} = \Omega = 1 - \epsilon - \sum_{\substack{i=1 \\ i \in \mathcal{S}}}^N \left(1 + \frac{1}{\gamma^2 \hat{\beta}_i} \right)^{-1}. \quad (2.47)$$

This equation is of the same form as (2.7) with the sum in the leftmost member having $N'' = N - N'$ terms. Therefore, we can apply again all the conclusions developed from

(2.7). Based on this, the new target SINR for users not in the set \mathcal{S} can be calculated from (2.22) and (2.47) as $\beta_i^{(2)} = \min[\hat{\beta}_i, \beta^{*(2)}]$, where the superscript denotes the second pass through the algorithm that allocates target SINR, power and rate, and

$$\beta^{*(2)} = \frac{1}{\gamma^2 \left(\frac{\hat{N}}{\Omega} - 1 \right)}. \quad (2.48)$$

Since $\beta^* < \beta^{*(2)}$, there might be users for which $\hat{\beta}_i < \beta^{*(2)}$, i.e. they may be unable (based on channel conditions) to set the transmit power necessary to achieve $\beta^{*(2)}$. Therefore, the above procedure is repeated until all SINR have been allocated.

Finally, note that the target SINR renegotiation is such that all users are assigned the same SINR (to reduce average distortion) except those in bad channel conditions which are assigned their highest possible SINR (to minimize individual distortion).

2.3.3 Practical Considerations

So far we have considered that the source encoding rate could be any real number between zero and x_M . In practice, it is to expect that the source encoding rate could only take a discrete number of values. To reconcile this reality there are two possible approaches.

The first approach is to consider that the achievable target SINRs also take a discrete number of values, each of them associated with one specific source encoding rate through $\beta_i = 2^{Ax_i+B}$ or $\beta_i = e^{Ax_i+B}$. Once the final β^* has been found, all users that had not been assigned a target SINR $\hat{\beta} < \beta^*$ are assigned as target SINR (or equivalently, source encoding rate) the one from the discrete set that is immediately larger than β^* . Then, one user at a time is assigned the target SINR from the discrete set that is immediately smaller than β^* . The order of user assignment is from the users in worst channel condition to those in the best one. This order is because, by reducing the trans-

mit power to those users transmitting at larger values, the general effect is to roughly equalize transmit power and battery life among users. After each assignment, condition (2.7) is tested. If the test results in compliance with the condition the assignment process stops and power assignment is calculated using (2.43). The overall result is that both the optimal target SINR and source encoding rate are achieved as an average among all calls, with users' source encoding rate differing in at most one discrete step. The only drawback of this solution is that is slightly sub optimal because it narrowly departs from the goal of assigning all users the same encoding rate. In what follows we will denote this approach as *type 1 rate adapted*.

The second approach is to consider that each source encoding rate is associated with a range of possible values for target SINR where overall distortion is minimum. To see this consider that, for a given rate, as the SINR decreases the distortion increases up to a point where a lower encoding rate with better channel protection would have lower overall distortion (although it would always have larger source encoding distortion than the other larger rate). Then, with this approach, all users that had not been assigned a target SINR $\hat{\beta} < \beta^*$ are assigned the same encoding rate: the one such that β^* is in its range of SINR values. Power assignment is calculated using (2.43). The small drawback of this approach is that it might be the case that some assignments corresponds to a larger contribution of channel induced errors than the quality goal. In what follows we will denote this approach as *type 2 rate adapted*.

Both approaches presents some differences but we will see in Section 2.5 that both have the same performance. Nevertheless, type 2 implementation may be less preferable over some restricted values of target SINR because in these ranges the subjective quality may be worse than the one for type 1 implementation due to channel induced distortion exceeding the quality goal.

2.4 Theoretical Performance Analysis

In this section we will develop a theoretical model to analyze the performance of our system. To this end we take a snapshot of the system at some random time and analyze the performance based on the observations of the system state and the random processes involved at that particular time.

The average distortion per call depends on two phenomena. One is the distortion of users who can set their transmit power at a level such that meets the optimal target SINR assignment. The other is the distortion of users who are undergoing deep fades and have to transmit at an SINR lower than the optimal one. The percentage of user in either of the two scenarios will depend on different design parameters as well as the environment the system is operating in. In order to obtain an statistical model for the behavior of our system we will assume that the iterative procedure to allocate SINR (as described in section 2.3.2) takes place in one stage; i.e. all users that cannot achieve the optimal SINR are assigned the target value as in equation (2.45), the rest of users are assigned the SINR as in (2.22). The results obtained with this assumption should approximate the optimal ones if very few users are assigned non-optimal SINRs in passes through the algorithm after the first one, and if there are not significant differences between the optimal target SINR in the first pass and those in following passes through the algorithm. These assumptions are likely to occur in practice since in well design systems the expected situation is that some few users would be in such a bad channel that they would be unable to reach any target SINR and most of the rest would be in such a good channel state that they could achieve any operational target SINR. Therefore the expected distortion \bar{D} is

$$\bar{D}_N = E[D] = D_N^* [1 - P_{\beta_{dB}^*}] + \int_{-\infty}^{\beta_{dB}^*} D(y) f_{\beta_{dB}}(y) dy \quad (2.49)$$

where $D(\beta_{dB})$ is the distortion of a call undergoing an SINR of β_{dB} , β_{dB} is the SINR assignment (2.45) and β_{dB}^* is the optimal SINR assignment, both measured in dBs, and $f_{\beta_{dB}}(y)$ is the probability density function (pdf) of β_{dB} . In effect, β_{dB} is the SINR assigned to users that cannot achieve the optimal target value. The probability of this event is, from (2.45),

$$P_{\beta_{dB}^*} = P[\beta_{dB} < \beta_{dB}^*] = P[\beta < \beta^*] = P \left[h^2 < \frac{\sigma^2(1-\epsilon)}{T_M \epsilon \gamma^2 N} \right]$$

If we assume a channel fading model following a normalized ($E[h_i^2] = 1$) Rayleigh distribution with users' gains being mutually independent, h^2 follows an exponential distribution with parameter $\lambda = 1$ and

$$P_{\beta_{dB}^*} = 1 - e^{-\lambda \frac{1-\epsilon}{\Theta \gamma^2 N}}, \quad (2.50)$$

where $\Theta = T_M \epsilon / \sigma^2$. This equation (2.50) is a valuable design tool to compute ϵ , given a target value $P_{\beta_{dB}^*}$ and the maximum transmit power T_M . Equation (2.50) allows to control the percentage of users that cannot reach the target SINR in relation to practical design parameters such as T_M and ϵ .

Next, from (2.45), we have

$$\beta_{dB} = -10 \log \left(\frac{1}{h^2 \Theta} - \gamma^2 \right), \quad (2.51)$$

which is a function of the random variable h^2 . Assuming that $\beta_{dB} \gg 10 \log \gamma^2$, which is reasonable for practical values of β_{dB} and γ^2 , it can be shown that

$$f_{\beta_{dB}}(y) = \lambda \frac{\ln 10}{10} 10^{(y-\Theta_{dB})/10} e^{-\lambda 10^{(y-\Theta_{dB})/10}}, \quad (2.52)$$

where $\Theta_{dB} = 10 \log \Theta$.

In order to be able to obtain a close form solution for the integral in (2.49) we used Taylor approximation to obtain,

$$f_{\beta_{dB}}(y) \approx \lambda \frac{\ln 10}{10} 10^{(y-\Theta_{dB})/10} \left(t_0 + t_1 y + \frac{t_2}{2} y^2 \right), \quad (2.53)$$

where

$$\begin{aligned}
t_0 &= e^{-10^{-\lambda\Theta_{dB}/10}} \\
t_1 &= -\lambda \frac{\ln 10}{10} 10^{-\hat{\Theta}/10} e^{-\lambda 10^{-\Theta_{dB}/10}} \\
t_2 &= \lambda \left(\frac{\ln 10}{10} \right)^2 10^{-\Theta_{dB}/10} e^{-\lambda 10^{-\Theta_{dB}/10}} (\lambda 10^{-\Theta_{dB}/10} - 1).
\end{aligned}$$

Also, we represented the distortion-SINR (in dB) function by piecewise linear approximation, i.e. $D(\beta_{dB}) \approx a_i \beta_{dB} + b_i$ in the interval $[\beta_{dB_i}, \hat{\beta}_{dB_{i+1}}]$, for $i = 0, 1, 2, \dots, m-1$ and $\beta_{dB_0} = -\infty, \beta_{dB_m} = \beta_{dB}^*$. Since it is to expect that $D(\beta_{dB})$ is an smooth function that saturates at both large and small values of β_{dB} , the approximation could be done reasonable well for a small number of intervals m . Therefore, (2.49) becomes

$$\bar{D}_N = \lambda \sum_{i=1}^{m-1} \int_{\beta_{dB_i}}^{\beta_{dB_{i+1}}} \frac{\ln 10}{10} 10^{\frac{y-\Theta_{dB}}{10}} (a_i y + b_i) \left(t_0 + t_1 y + t_2 \frac{y^2}{2} \right) dy + D_N^* e^{-\lambda \frac{1-\epsilon}{\Theta_{dB}^2 N}}, \quad (2.54)$$

2.4.1 Dynamic Calls Model

The equations developed so far depend on the optimal SINR assignment β^* obtained using equation (2.22) converted to dBs, which in turn depends on the number of users in the system N . Then, the expected distortion is a function of N , as emphasized in equation (2.54). The number of users in the system is itself a random variable that depends on the traffic within the cell under consideration. Assume that calls enter the cell at a rate ν following a Poisson arrival process (i.e. exponential interarrival time) and that the random calls duration follow an exponential distribution with mean $1/\mu$. We also assume that the system implements a call admission control policy that limits the maximum number of users to a maximum $N_L > N_M$. The value of N_L is set by a QoS goal $\bar{D}(N_L) = D_M$, D_M being the maximum tolerable distortion limit. Users

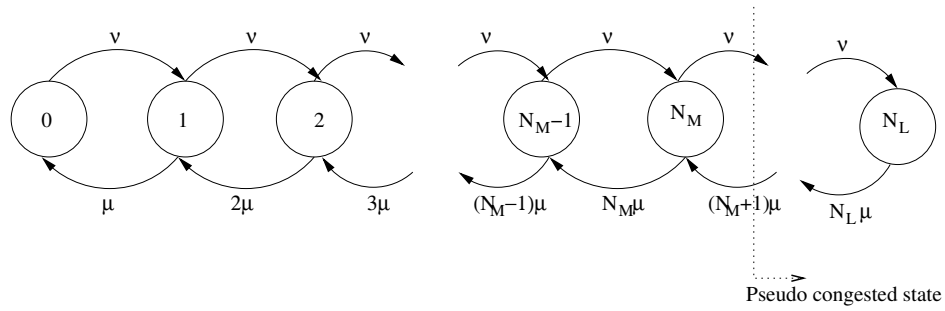


Figure 2.5: Markov chain representation of the $M/M/N_L/N_L$ traffic model.

that arrive when there are already N_L ongoing calls in the system are denied service and considered lost. Therefore, the traffic model follows an $M/M/N_L/N_L$ queuing model. To facilitate its study, the queue can be represented in the form of a transition diagram as the one shown in Figure 2.5.

From the theory of $M/M/N_L/N_L$ queues the steady-state probability that at any time there are N users in the system is, [3]

$$q_n = P(n = N) = \frac{a^N/N!}{\sum_{i=0}^{N_L} a^i/i!} \quad (2.55)$$

where $a = \nu/\mu$ is the offered load. Therefore, over a sufficiently large period of time the average distortion per call will be

$$E[\bar{D}_n] = \frac{\sum_{n=0}^{N_L} \bar{D}_n a^n/n!}{\sum_{i=0}^{N_L} a^i/i!} \quad (2.56)$$

where we have assumed that processes are ergodic and that the channel gain process is independent of the number of users in the system. Note that, as stated, $E[\bar{D}_n]$ is a function of the offered load a . In addition, when evaluating and designing the proposed

system it is important to also consider the average length of time that users will be operating at an encoding rate lower than the maximum. Consider the following definitions:

Definition 1 We call pseudo congestion state the operational state when users are operating at a source rate lower than the maximum. This corresponds to the situation when $N_M + 1 \leq N \leq N_L$. We refer to the average time in this state as the average pseudo congestion duration, T_{pc}

Definition 2 We call the operational state when new incoming calls need to be blocked as congestion state. This corresponds to the situation when $N = N_L$ and a new call arrives. The probability of this event is queuing theory's blocking probability.

The following theorem states the two most important operating magnitudes related to the pseudocongestion state:

Theorem 1 The average pseudo congestion duration is equal to

$$T_{pc} = \frac{\sum_{n=N_M+1}^{N_L} (n\mu)^{-1} q_n}{\sum_{n=N_M+1}^{N_L} [1 - \nu(n\mu)^{-1}] q_n}, \quad (2.57)$$

and the average distortion while in the pseudo-congested state is

$$E \left[\bar{D}_n \middle| N_M < n \leq N_L \right] = \frac{\sum_{i=N_M+1}^{N_L} \bar{D}_i q_i}{\sum_{i=N_M+1}^{N_L} q_i}. \quad (2.58)$$

Proof: The proof for the average pseudo congestion length, T_{pc} is an adapted version, to the $M/G/N_L/N_L$ queue, of the approach followed in [38] to find the average busy time of an $M/G/1$. The present case differentiates from the referenced in that we want

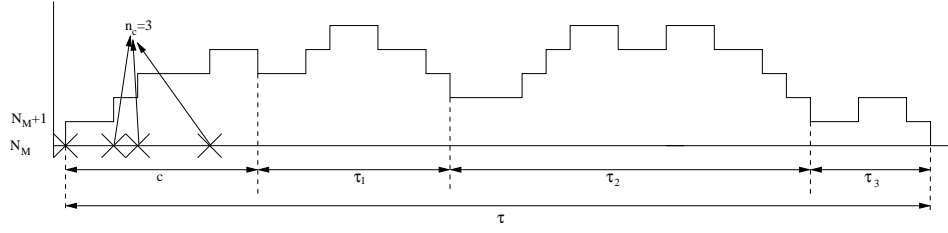


Figure 2.6: One realization in pseudo congestion state to study the average pseudo congestion length.

to find the expression for the average time the system spends in a group of contiguous states (those between $N_M + 1$ and N_L). Also, in the present case average service times change for each state. Specifically, the event of interest starts when there are N_M calls in the system and a new call arrives. The call is serviced as soon as it arrives. As shown in figure 2.6, we assume that the call spends a time c in the system before departing. During this time n_c new calls arrive into the system. As in [38], we define a busy period τ_i as a random variable equal to the time interval from $t = t_i$ when there are $N(t_i)$ calls in the system to $t = t_i + \tau_i$ when $N(t_i + \tau_i) = N(t_i) - 1$ for the first time. A useful property associated with busy periods is that the variations of the number of calls in the system, $N(t)$, during the time interval do not depend on the number of calls at the beginning of the interval. Hence, as shown in figure 2.6, the period of interest τ , which is itself a busy period, can be considered as the sum of the first arrived call service time and as many independent and identically distributed busy periods as calls arrived while servicing the first arrived call. Hence

$$\tau = c + \tau_1 + \tau_2 + \dots + \tau_{n_c}. \quad (2.59)$$

Note that $E[n_c] = \nu E[c]$, where $E[c]$ is the average time to service a call (or equivalently, the average time for any call to finish) when the system is operating in the pseudo

congestion mode. Hence

$$E[c] = \frac{\sum_{n=N_M+1}^{N_L} (n\mu)^{-1} q_n}{\sum_{n=N_M+1}^{N_L} q_n}. \quad (2.60)$$

Also, assuming independence between n_c and all busy periods τ_i the expectation of the random sum $\sum_{i=1}^{n_c} \tau_i$ is

$$E \left[\sum_{i=1}^{n_c} \tau_i \right] = E[n_c] E[\tau_i] = \nu E[c] E[\tau]. \quad (2.61)$$

Therefore, taking expectation of (2.59) and using (2.61) we get $E[\tau] = E[c] + \nu E[c] E[\tau]$.

Thus,

$$T_{pc} = E[\tau] = \frac{E[c]}{1 - \nu E[c]}.$$

With $E[c]$ calculated as in (2.60). After doing some algebraic operations we conclude that

$$T_{pc} = \frac{\sum_{n=N_M+1}^{N_L} (n\mu)^{-1} q_n}{\sum_{n=N_M+1}^{N_L} [1 - \nu(n\mu)^{-1}] q_n}, \quad (2.62)$$

□

So far we have considered a $M/M/N_L/N_L$ queuing model. As discussed, this model represents a system where users are admitted up to a maximum number of ongoing calls (N_M or N_L). This represents a strict limit on the number of users set based on the average distortion. On one side this model better represents the problem of call admission control from the network operator's viewpoint, i.e. calls requesting service when there are a certain maximum number of calls already being served are rejected

in order to maintain the quality for the existing calls. On the other side, in CDMA it is possible to accept more calls beyond the set limit. This is why researchers have studied the Erlang capacity of CDMA systems following a $M/M/\infty$ model in terms of the numbers of calls that can be supported such that the outage probability remains below some threshold, where outage occurs when the system exceeds some operational parameter (typically related to interference in CDMA) [52], [43]. Therefore, it is of interest to also study the impact of our approach on this other capacity measure.

In [52] the outage condition was defined in terms of the relation between the amount of interference density and background noise level. Noticing again the equivalence between equations (2.5) and (2.7) with (4) and (6) in Ref. [43] it is straightforward to show that the outage condition derived from the feasibility constraint (2.7) can be written as

$$\sum_{i=1}^N \omega_i > (1 - \epsilon) \left(1 + \frac{1}{\gamma^2 \beta} \right) \quad (2.63)$$

where ω_i is a binary random variables that is equal to one when user i is in a talk spurt. Let $\rho = P(\omega_i = 1)$. Typically $\rho = 0.4$. If the outage probability is defined as

$$P_{out} = P \left[\sum_{i=1}^N \omega_i > (1 - \epsilon) \left(1 + \frac{1}{\gamma^2 \beta} \right) \right], \quad (2.64)$$

the Erlang capacity is the maximum offered load such that the outage probability is kept below some target, typically 1% or 2%. Let the random variable $Z = \sum_{i=1}^N \omega_i$. Defining

$$K_0 = (1 - \epsilon) \left(1 + \frac{1}{\gamma^2 \beta} \right), \quad (2.65)$$

and using the characteristic function it can be shown that, [52],

$$P_{out} = e^{-\rho\nu/\mu} \sum_{[K_0]}^{\infty} \frac{(\rho\nu/\mu)^k}{k!} \quad (2.66)$$

Since K_0 increases as the target SINR β is reduced we can see that the outage probability is reduced as source encoding rate is also reduced. Equivalently, for the same outage

probability, as the source encoding rate is reduced it is possible to support larger offered loads.

2.5 Performance Evaluation

In this chapter we have so far studied how to renegotiate the target SINR in a CDMA system by reducing the source encoding rate so as to extend operation beyond the nominal congestion point while maintaining a quality goal (that of keeping channel induced distortion at a perceptually acceptable level). This reduction in source encoding rate increases the end-to-end distortion in a smooth and controllable way by setting as ‘quality goal’ a limit to the channel induced distortion. This section is concerned with the evaluation of the techniques studied in this chapter and the overall performance of the proposed system. Most of this evaluation will be based on simulations. Since the focus of the study is on real-time sources, we will consider that the system is carrying conversational voice traffic. As input for the simulation tests, we used eighteen speech sequences from the NIST speech corpus [47]. These sequences were chosen to represent different male and female speakers. We encoded these sequences using the GSM AMR (Advance Multi-Rate) Narrowband Speech Encoder [12]. This encoder operates with 20 ms frames, 5 ms look-ahead and includes an error concealment mode, which was used in simulations. Of the eight possible encoding rates: 12.2, 10.2, 7.95, 7.4, 6.7, 5.9, 5.15 and 4.75 kbps, we used only the six highest ones.

To measure the end-to-end distortion of the speech sequences we choose a perceptually weighted log-spectral distortion measure calculated by numerical approximation of the function

$$SD(\hat{A}(f), A(f)) = \sqrt{\int |W_B(f)|^2 \left| 10 \log \frac{|\hat{A}(f)|^2}{|A(f)|^2} \right|^2 df}, \quad (2.67)$$

where $A(f)$ and $\hat{A}(f)$ are the FFT-approximated spectra of the original and the reconstructed speech frames, respectively. $W_B(f)$ is a subjective sensitivity weighting function defined by [9],

$$W_B(f) = \frac{1}{25 + 75(1 + 1.4(f/1000)^2)^{0.69}}. \quad (2.68)$$

This distortion is measured on a frame-by-frame basis and then averaged over all frames. Log-spectral distortions are frequently used to objectively measure speech distortion. They see their most common application in the evaluation of speech vocoders. In the present case, we choose this measure not only because of its good mathematical properties but also because of its good correspondence to subjective perception [41]. Contrasting to the application of this distortion measure to the evaluation of vocoders, we included in the measure computation outliers frames. This is to better capture the effects of channel errors. For design and to report our results, we used a normalized distortion measure, computed as the ratio of the spectral distortions with that of the speech sequence encoded at the highest rate (12.2kbps) in the absence of channel noise, δ . Incidentally, we noted that both approximations to the distortion-rate function (using $f(x) = \alpha 2^{-2kx}$ and $f(x) = \alpha x^{-k}$) were applicable, as discussed in section 2.3.

Traditionally, it is possible to increase the number of users in a CDMA network by accepting an increase in interference. Therefore, as more users are admitted, the decrease in the SINR reduces the source quality at the receiver due to the increase in channel-induced errors. Our scheme presents an alternative to this approach. In our scheme, as more users are admitted into the system the source distortion also grows but, in this case, due to the source rate reduction. Channel-induced distortion does not change since its contribution to the end-to-end distortion is kept constant. This idea of increasing the number of users by source encoding rate adaptation is highlighted in our results in equations (2.25) and (2.26) for a distortion rate function of the form

$f(x) = \alpha 2^{-2kx}$ and in equations (2.33) and (2.35) for the more tight distortion-rate approximation of the form $f(x) = \alpha x^{-k}$. Figures 2.7 (a) and (b) show different realizations of equations (2.33) and (2.35) respectively, for different values of the parameter A. The parameters used in this figure were obtained from simulations. Specifically, the different values of parameter A are representative of those values obtained from different configurations of the channel encoder subunit, with and without concatenation of fixed and variable rate convolutional encoder. The analysis of these results shows these equations substantiate the claim that our system is able to increase capacity at the cost of a controlled smooth degradation of reconstructed source quality. For example it is possible to increase 70% the number of users for a 20% increase in average normalized distortion.

We have already emphasized the fact that in our system the ability to adapt the source encoding rate allows for a change in the target SINR so that the system feasibility condition (2.7) holds. In a traditional CDMA system the lack of control over the source encoding rate allows for two options when increasing the number of users. One option is to change the target SINR so that (2.7) still holds. In this case the received quality would eventually be dominated by channel-induced distortion because a traditional system is not able to change the source encoding rate and change the target SINR without increasing the proportion of channel errors to the end-to-end distortion. We will denote this option as *type 1 non-rate adapted*. The second option is the one where a call admission algorithm is responsible for preventing that the number of users reaches a value such that (2.7) does not hold any more. This system has no functionality to change the target SINR. If more users enter the system the first effect would be the one following the rapid approach of the denominator in the power assignment (2.43) to zero. As a consequence, the power assignment to meet the target SINR significantly increases, as does

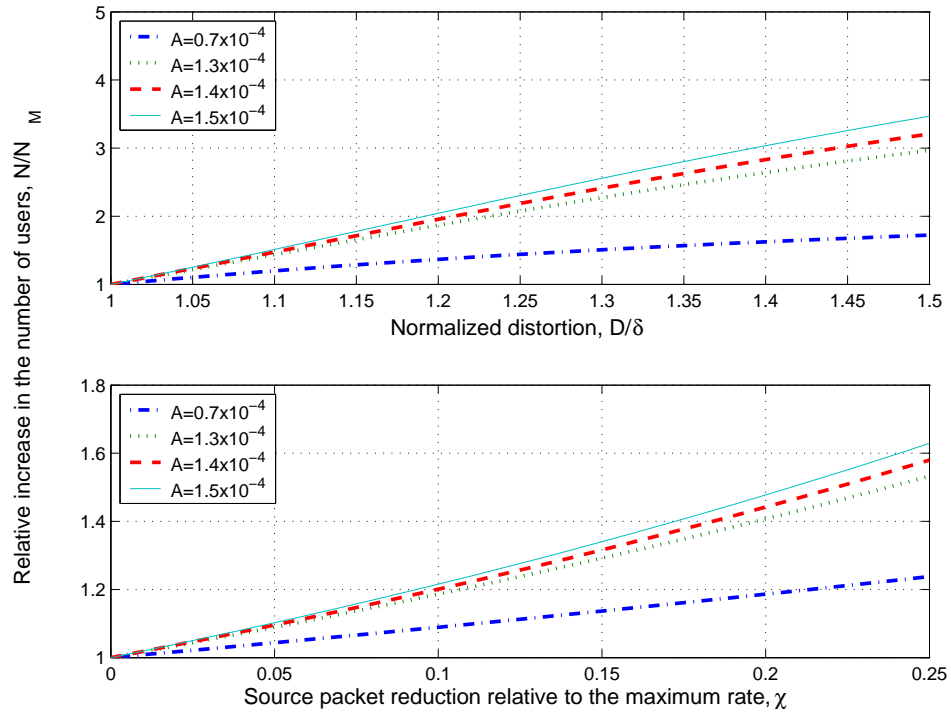


Figure 2.7: Relative increase in the number of users as a function of the normalized average distortion per call (top) and as a function of χ , the source packet reduction relative to the maximum rate (bottom).

the number of users that become limited by the maximum power constrain. Therefore, as the number of users increases beyond the congestion point, those of them that see their SINR reduced also significantly increases. When eventually the number of users is such that the power assignments become a negative value, the system would have gone beyond the controlled operation range and would assign the maximum transmit power to all users; i.e. an assignment where each user is left by itself. Even when assigning maximum transmit power to all users, all of them will suffer from increasing degradation of SINR. We will denote this option as *type 2 non-rate adapted*.

We evaluated the performance of our scheme by comparing it with an equivalent CDMA system. This system is one that shares the same operational blocks and configuration as our system in Figure 2.1 with the difference that no adaptation is possible, i.e. the system operates always at the maximum source encoding rate and corresponding larger channel encoding rate. Operation beyond the congestion point is carried out following either type 1 or type 2 non-rate adapted options as described above. For all the systems under consideration we assumed BPSK modulation. For channel error protection we choose a memory 4, puncturing period 8, mother code rate $1/4$ (variable rate in our system) Rate-Compatible Punctured Convolutional (RCPC) code [19] decoded with a soft Viterbi decoder. The constant frame size that is input into the spreader was chosen to be equal to 500 bits.

Based on this configuration our source rate-adapted system can change between six possible operating modes, one for each possible source encoder rate. Describing the modes by pairs (source encoding rate, channel code rate) the six possible modes are (12.2 kbps, $1/2$), (10.2 kbps, $8/19$), (7.95 kbps, $1/3$), (7.4 kbps, $4/13$), (6.7 kbps, $2/7$) and (5.9 kbps, $1/4$). Figure 2.8 shows the distortion as a function of an AWGN channel SNR for each of the six possible operating modes. We set the quality goal so

that the end-to-end distortion was 3% more than the one for the same source encoding rate with no channel errors. By approximating the interference to an AWGN process, we found from simulations the target SINR for each mode so that the quality goal was met. The equivalent traditional CDMA system is the one that always operates at the maximum source encoding rate, i.e. it only uses the mode (12.2 kbps, 1/2). The target SINR for this mode was set with the same criteria detailed above. Assuming $\epsilon = 0.05$ and $\gamma^2 = 0.01$, we set the maximum number of users that can be supported without congestion to $\hat{N}_M = 85$. From this operating point we increased the number of users by adapting the source rate (changing mode and target SINR) in our system and by reducing the SINR in the equivalent CDMA system following the two above options. We modeled the channel gain as having a normalized ($E[h_i^2] = 1$) Rayleigh distribution. The limit on transmit power, T_M , was set so that at the congestion point no more than approximately 6% of users were unable to achieve the target SINR due to bad channel conditions. We also assumed that the base station could perfectly estimate the users' channel gains. For the cases we did simulations, we used Monte Carlo method setting as stopping criteria a convergence in the relative error below 1%.

Figure 2.9 shows the simulations results. Our first observation is that there is no statistical difference (less than 1 %) in the performance of the two types of rate-adapted approaches described in Section 2.3.3. In addition, we can see that the proposed system significantly outperforms the traditional approach for any admissible additional distortion. Specifically, when compared to the type 1 non-adapted system, our system can accept 30% and 55% more users for 15% and 25% extra distortion, respectively and when compared to the type 2 non-adapted system, the proposed system can accept 32% and 76% more users for 15% and 25% extra distortion, respectively. These distortion values were of interest because they correspond to acceptable quality for telephone commu-

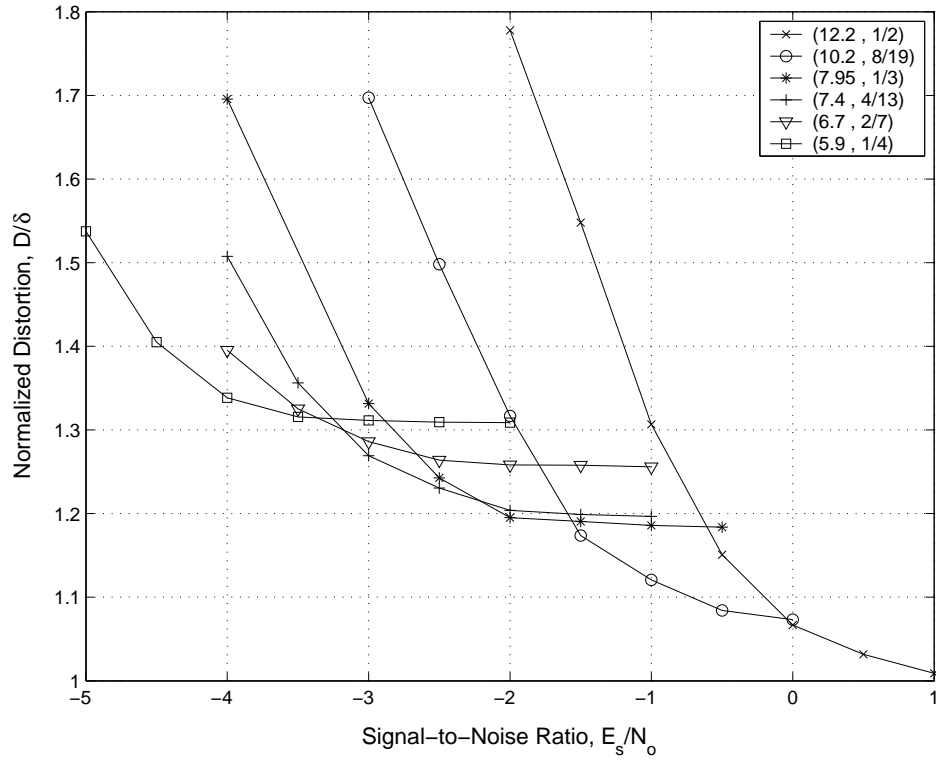


Figure 2.8: Distortion as a function of SNR for each of the six possible operating modes, each defined by the pair (source encoding rate, channel code rate)

nication when considering subjective perception and results variability due to channel errors. Even larger gains are achievable if higher levels of distortion can be accepted. It is also interesting to note that, as predicted, the type 2 traditional CDMA system undergoes a steep increase in distortion shortly after the congestion point. In practical terms, this observation justifies the choice for the congestion point since it follows the natural guideline of accepting the largest possible number of users while avoiding the region of steep increase in distortion. Even more, notice that this concept is also implied in the definition of outage condition (2.63). In addition to this simulation results, Figure 2.9 includes results from the analysis in sections 2.3 and 2.4. The curve labeled “Approximate” corresponds to equation (2.33) using parameters measured from the simulation. We can see that this result is a very good one as an initial design prediction of the system behavior that only requires knowledge of a few simple parameters from the source and channel encoder units. The difference between this result and simulations are readily justified by the fact that equation (2.33) considers all channel fadings equal to one and that the distortion-rate function is a tight bound on the worst case performance of the source encoder. In fact, examining this result, and in view of the simplified model it represents, we can consider that the approximations for the distortion-rate curve and the SINR-source encoding rate function are sufficiently representative of the system behavior. Figure 2.9 also includes, labeled as “Theoretical”, the analytical prediction of the system behavior using equation (2.54). Because the relative error between this result and the one from simulations never exceed 4% (and in many cases it is much less than this value), we can say that equation (2.54) is a good representation of the system behavior. Finally note that for N_M users the distortion is 1.07 instead of 1.03. The difference is due to the users that are in deep fades and cannot achieve the target SINR. This difference could be controlled at design time using equations (2.50) and (2.54).

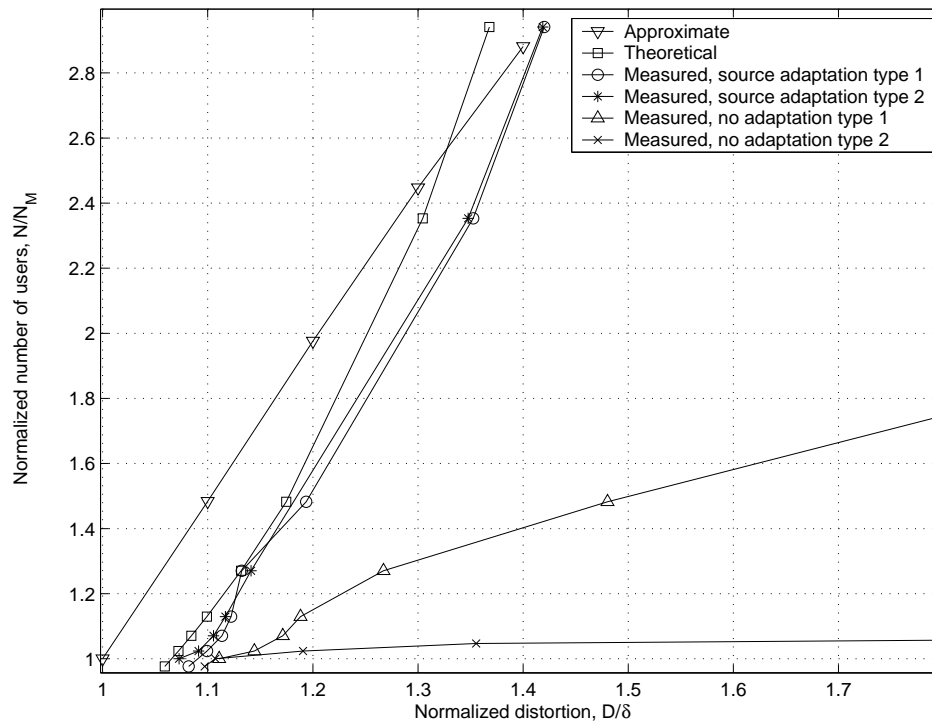


Figure 2.9: Comparison between the proposed scheme (with source rate adaptation) and an equivalent traditional CDMA system (with no adaptation).

When considering dynamic change over time in the number of ongoing calls and assuming a $M/M/N_L/N_L$ model, i.e. acceptance of calls up to a maximum number, the performance of the rate adaptable system could be assessed by considering three elements: expected distortion, average duration in pseudo congestion state and call blocking probability (probability that an incoming call cannot be provided with service). Using the same system setup as for the previous simulation, we show in Figure 2.10(a) the analytical results for both the expected normalized distortion and expected normalized distortion while in pseudo congested state as a function of offered load $a = \nu/\mu$. In the model we assumed also an average call duration of three minutes and $N_L = 240$ (maximum average normalized distortion equal to 1.4 approximately, which we found corresponds to the maximum acceptable distortion). As expected, as the offered load increases, both expectations converge to the same value because the probability of the system being in the pseudo congestion state increases with the offered load to a value of 1. In Figure 2.10(b) we compare the blocking probability of our system with a non adaptable system, i.e. one that is either congested or uncongested with no pseudocongestion operation or increase in distortion due to source rate adaptation. For these systems, calls are blocked when the number of users in the system is N_M . As our system can extend operation up to N_L calls by reducing source encoding rate, it can support much higher offered loads for the same level of blocking probability. The increase in offered load depends on the maximum acceptable expected normalized distortion. For example, the increase in offered load is 27% if $N_L = 106$ (acceptable maximum expected normalized distortion is 1.1 or 10% additional distortion, which is almost imperceptible), or 61% if $N_L = 132$ (maximum distortion equal to 1.2, which is slightly perceptible), or 205% if $N_L = 140$.

Figure 2.11 shows the average pseudocongestion duration. The figure does not show

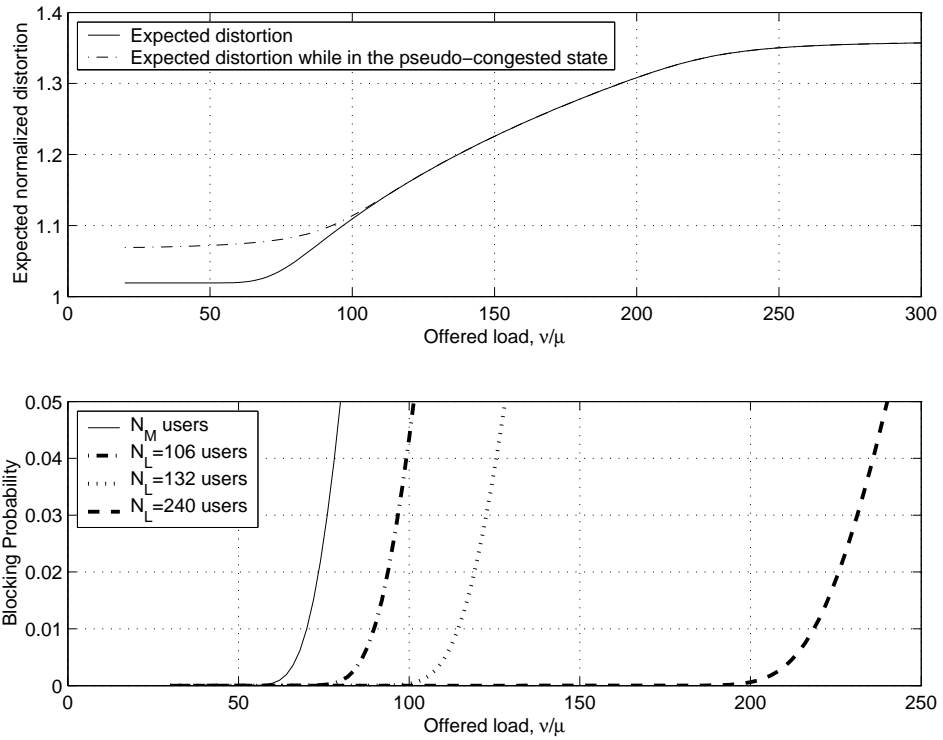


Figure 2.10: Expected normalized distortions and blocking probability as a function of offered load $a = \nu/\mu$.

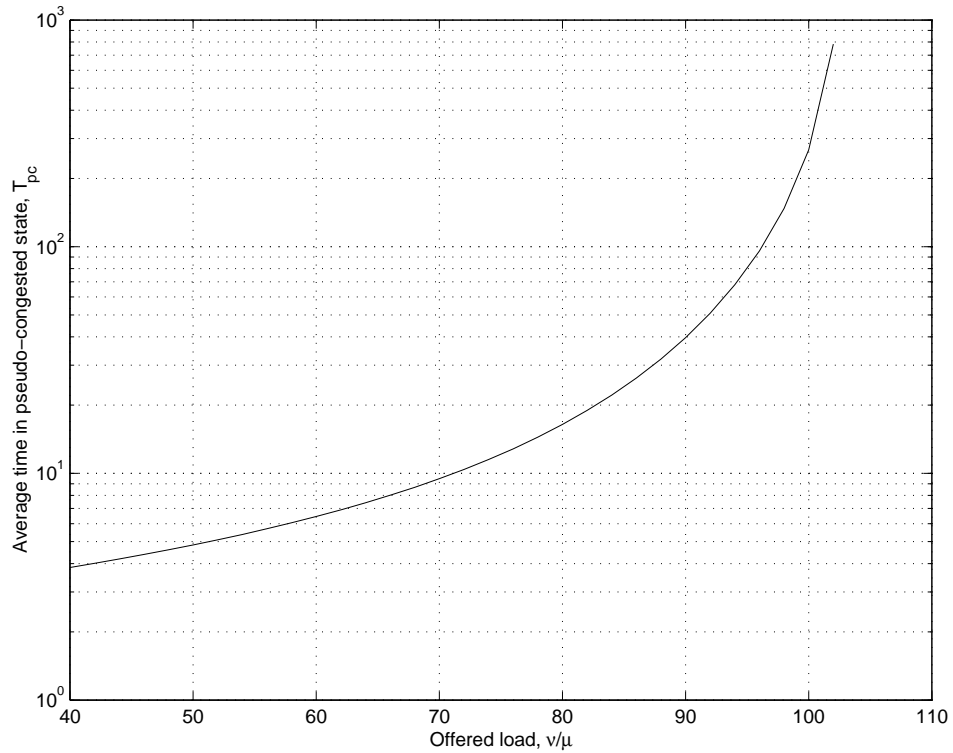


Figure 2.11: Expected length of time in pseudo congested state as a function of offered load $a = \nu/\mu$.

a range of offered load values as large as the others because the duration in the pseudocongestion state for large offered loads become “infinite”, meaning that the system is permanently in the pseudo congestion state. This result is useful in highlighting the fact that although the extension of operation beyond a congestion point comes at the cost of increasing distortion (albeit being always controllable and often small), this increase is only transient in many cases. As a complement to this figure, Figure 2.12 shows the expected normalized distortion while in pseudo congested state as a function of the expected duration of time the system would spend each time it transitions into the pseudocongestion state.

Figure 2.13 shows the Erlang capacity for each operating mode assuming a $M/M/\infty$

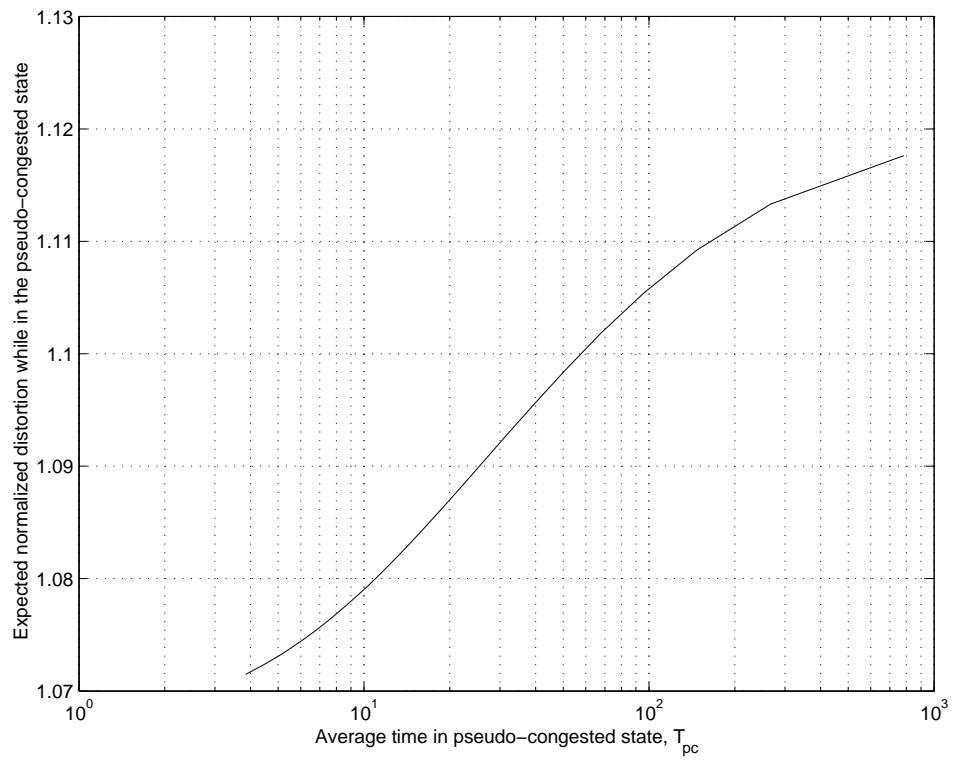


Figure 2.12: Expected normalized distortion while in pseudo congested state as a function of the expected duration in pseudo congested state.

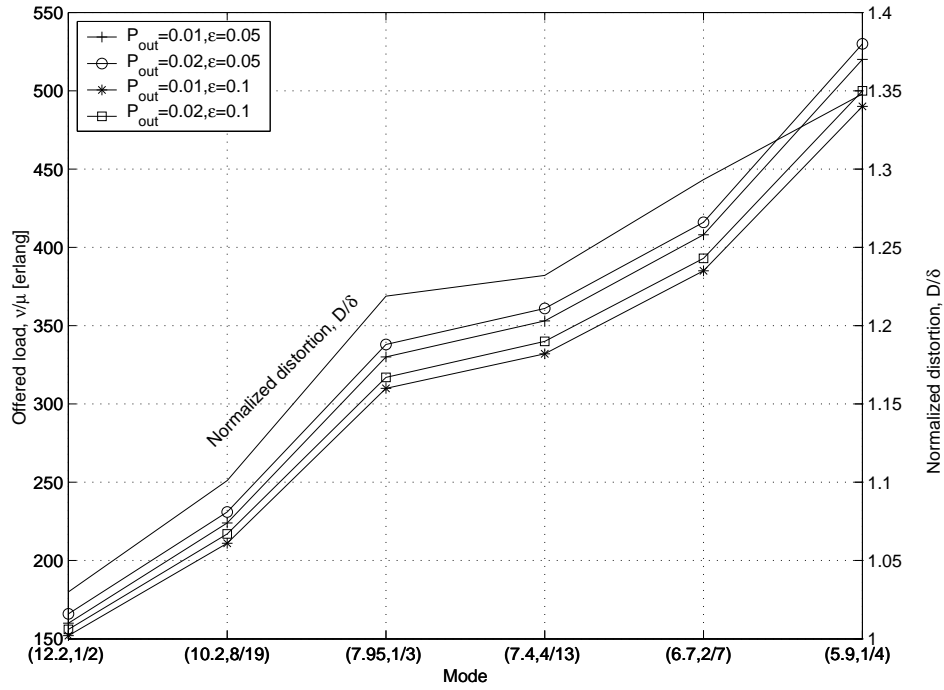


Figure 2.13: Erlang capacity for the different operating modes and end-to-end normalized distortion for each operating mode.

model. Of course, a non adapted system coincides with the (12.2 Kbps, 1/2) mode. To obtain the curves we used equation (2.66), changing K_0 with each mode's target SINR, to find what was the maximum offered load so that the outage probability was 1% and 2%. We assumed $\rho = 0.4$. Figure 2.13 also includes, with scale on the right, the normalized distortion corresponding to each operating mode. It can be seen that, depending on the acceptable increase in distortion, the erlang capacity could be notably increased with values ranging from approximately 40% increase in Erlang capacity for a distortion of 1.1 to approximately 220% increase in Erlang capacity for a distortion of 1.35. Roughly speaking, results considering this model where comparable to the ones obtained with the $M/M/N_L/N_L$ model.

Finally, note that our system has the extra advantage that the increase in distur-

tion is smooth, controllable and predictable. This is because, as the dominant process is the reduction in source encoding rate, performance follows the D-R curve. By setting the target SINR according to the quality goal, channel induced distortion is kept small. In contrast, in a traditional CDMA approach, the increase in distortion is a consequence of the increase in channel-induced errors. Here, the system behavior is much less predictable and distortion is subjectively more annoying, because performance is dominated by the random errors in the channel.

2.6 Conclusions

In this chapter we have presented a design for real-time communication in a CDMA network that is able to dynamically extend system operation beyond a congestion point at the cost of smooth, controllable quality degradation. The basis of the idea lies in optimally renegotiating the target SINRs with the goal of accommodating the real-time traffic demands while minimizing average distortion. Target SINRs are adapted by changing the source-encoding rate while the contribution of channel-induced distortion to overall quality is kept below a fixed small threshold. This changes the phenomena by which distortion in CDMA increases with the number of calls from the one dominated by the growth in channel-induced errors to one that follows the source encoder distortion-rate performance. From this viewpoint, a contribution in this chapter is the departure from the established idea that quality settings of a real-time call are set at least for the duration of the call. Our contributed idea of extending operation beyond congestion at the cost of a smooth degradation of quality has very important application in many situations where servicing a call is more important than guaranteeing an strict quality setting.

Another contribution in this chapter is the mathematical model and analysis of a sim-

ple model where all source encoders are the same and operate in the same state, ideal power control is assumed, processing gain is fixed and the channel is AWGN. The simplicity of this approach allowed isolating the main elements of our design and focused on the core elements affecting performance. The study was presented for two assumed distortion-rate functions; one that is directly related with popular such functions in rate-distortion theory, and another that was observed to be a better fit for complex practical encoders. As part of this study, we were able to develop useful design equations that permit predicting a fairly accurate behavior of our system. An important conclusion from this approach is that our system is able to significantly increase capacity at the cost of a moderate controlled smooth degradation of reconstructed source quality.

Next we showed how this study could be extended for situations where transmit powers and channel gains are taken into considerations. For this case, SINR adaptation is made in such a fashion that all users are assigned the same SINR (to reduce average distortion) except those in bad channel conditions that are assigned their highest possible SINR (to minimize individual distortion). Based on this model, and assuming a channel dominated by Rayleigh fading, we were able to develop an expression that predicts the performance of our system within a small relative error. In addition, we also discussed practical considerations related to the practical design implementation.

The final part of our study is a traffic analysis where we developed expressions for the performance of the system considering the offered load. As part of this study we developed an expression for the average duration in the pseudo congestion state assuming that the system behaves as an $M/M/N/N$ queue. We also developed expressions to calculate the Erlang capacity following the model used in [52] ($M/M/\infty$ queue).

Based on the analytical study in this chapter, we presented new results that show that it is possible to increase the number of serviced calls by a smooth reduction in

quality. Numerical results from this analysis show that it is possible to increase 70% the number of users for a 20% increase in average normalized distortion or that it is possible to support at least a 100% increase in offered load for also a 20% additional expected normalized distortion for both call admission models considered. Finally we compared through simulations the proposed scheme with an equivalent system that cannot perform adaptation. The results show that the proposed system can accept 30% and 55% or 32% and 76% more users for 15% and 25% extra distortion, respectively depending on the type of non adapted system.

We finally highlighted the fact that the rate adapted system is also preferable from the subjective quality viewpoint due to the different process that dominate the increase in distortion: a smooth and predictable increase following the source encoder distortion-rate performance as opposed to a random process that increases distortion through channel induced error.

Chapter 3

Resource Allocation Through Statistical Multiplexing of Multimedia Calls in Variable Processing Gain DS-CDMA

3.1 Introduction

In this chapter we will extend the study of resource allocation in CDMA systems carrying real-time traffic. We now consider the most general problem of optimal adaptation to resolve interference-generated congestion for an arbitrary set of source encoders with arbitrary SINR goal and variable transmit bit rate (variable spreading factor). As important result, we show that our problem, as stated in a multiuser environment subject to a system stability and power amplifier dynamic range constraint, is analogous to the problem of efficient bit budget allocation to an arbitrary set of quantizers [45] and, more importantly, can be further considered as the optimal source-controlled statistical multiplexing solution in CDMA. The overall result of the proposed solution is a flexible system that inherently establishes an efficient tradeoff between end-to-end distortion and number of conversational calls. Also important is the fact that the problem setup is one of a true multimedia system, where our interpretation for this is a system where the

distortion-rate performance of each source may change within two consecutive transmit periods and also within two different calls.

In the previous chapter we used speech as real-time source. In this chapter we change the type of real-time source we use as application example and in simulations and we consider layered, embedded conversational video. This choice maintains our assumption of an externally controllable source encoder and adds extra new challenges due to video high and widely variable demand for network resources.

As we have seen in the previous chapter, our solution is a centralized scheme that requires for the true multimedia setup addressed in this chapter the transmission of information about each source encoder distortion-rate performance. Clearly, this approach might in principle add a notable overhead to the communication and become bandwidth-inefficient. In this chapter we solve this potential problem by presenting a scheme that compressed the distortion-rate information, i.e. reduced the overhead to acceptable values, and does not affect performance.

Finally, we study the teletraffic characteristics of our system and its relation with end-to-end distortion, traffic load and resource demand. Here we also design a simple call admission control rule. We finish this chapter by summarizing the main conclusions and contributions.

3.2 System Model

3.2.1 Model Description

Consider the uplink of a single cell, chip-sampled Direct-Sequence CDMA system with bandwidth W . Assume that there are N users in the system, each carrying on an independent conversational call. Figure 3.1 shows the block diagram of the main compo-

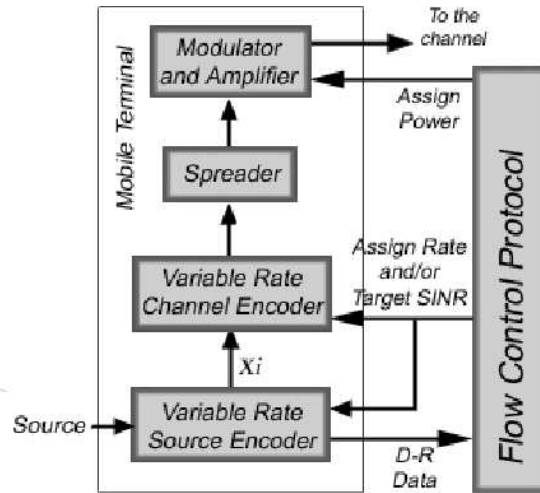


Figure 3.1: Block diagram of the proposed system

nents of the proposed system. This system is similar to the one presented in 2. A block of samples from a real-time source is encoded into a *source frame* using an encoder with the key property that it is possible to externally control the encoding rate. This means that, for each user i , it is possible to choose the source encoding rate x_i . We will assume here that the possible choices for source encoding rate belong to a finite set. Each user could be operating with a different encoder and although users will not change encoder during a call, its distortion-rate (D-R) performance is allowed to change from frame to frame based on the changing source statistics. A variable rate channel encoder provides channel error protection for the source frame. In this chapter we are not going to impose a restriction of equally protecting all sources-encoded bits. As will become clear later in this chapter, the proposed design could be applied to certain Unequal Error Protection (UEP) schemes. The channel encoder output, with a transmit bit rate equal to r_i , is fed into the spreader for transmission. In contrast to the system studied in chapter 2, the system in Figure 3.1 has a Variable Spreading Factor (VSF) spreader that can adapt the call's processing gain.

Note that this system design allows for each user to dynamically switch between different combinations of a source and a channel coding rate. Each such combination will form an *operating mode*. A flow control protocol located at the base station allocates to each mobile in the coverage area an operating mode and power, which are jointly communicated to each mobile terminal so as to proceed with transmission. We will see that its function is analogous to the one of a statistical multiplexer. In this chapter we will focus on the design and analysis of this protocol. The design optimization criteria will be that the allocation needs to minimize mean end-to-end distortion subject to traffic demand. As was the case in chapter 2, we will require the allocation to satisfy a *quality goal*. The quality goal represents a condition that ensures that the communication will not be noticeably impaired by channel-introduced errors. In chapter 2 the quality goal was specified as a limit on the proportion of channel-induced distortion to the overall distortion. Alternatively, the quality goal could also be specified using the Frame Error Rate (FER) or, as will be the case in this chapter, the Bit Error Rate (BER).

Because we consider that each call distortion-rate performance may change from one transmission period to the next, each call needs to send information about this performance to the flow control protocol. The flow control protocol performs optimal statistical multiplexing using the estimates of traffic demands from each call D-R performance information. Therefore, during each transmission period, each mobile sends not only the encoded source data sampled during the previous period but also information about the source encoder D-R performance corresponding to the source data sampled during the current period. In effect, transmission of a source frame is delayed by exactly one frame duration with respect to the time when data was sampled. We will discuss later in this work how the transmission overhead associated with the D-R information could be kept small.

3.2.2 Video Telephony Calls as Application Example

Consider the application of the proposed system in Figure 3.1 to provide a conversational video communication service. Assume that all mobile terminals use an MPEG4 FGS (Fine Granularity Scalability) coder [36] that is error-protected with Rate compatible punctured convolutional (RCPC) codes [14]. Transmission parameters are chosen so that the end-to-end quality is good for conversational communication, which means that source encoding distortion should be kept as small as possible and channel-induced errors should not introduce annoying effects. This case matches well our model since the MPEG4 FGS encoder generates a two-layer (base and enhancement) source coded bit stream. The enhancement layer bit stream is embedded, which easily allows controlling encoding bit rate. Furthermore, the D-R performance of this encoder (in fact, all video codecs) may change for each call and frame because it depends on various characteristics such as frame texture and type of temporal prediction (I or P) used and the amount of motion in the video sequence. This is illustrated in Figure 3.2, which shows the D-R performance of several representatives frames from two QCIF, 30 frames per seconds, video sequences: Foreman (with high motion) and Akiyo (with low motion). Both sequence were encoded with 29 P frames between each I frame. The figure shows results with and without channel errors. Channel errors were introduced at a BER equal to 5×10^{-6} and 10^{-5} for the base and enhancement layers respectively. These values were chosen following the *quality goal* criterion: after exhaustive simulations using the MPEG4 FGS encoder with error resilience and concealment we noticed that these BER values corresponds to the limit where the end-to-end subjective quality was good (comparable to ‘toll’ quality in telephone communications) and channel errors did not introduced annoying artifacts or impaired understandability of the source. Notice, then, that in this case the quality goal is specified in terms of BER. Also note that the dif-

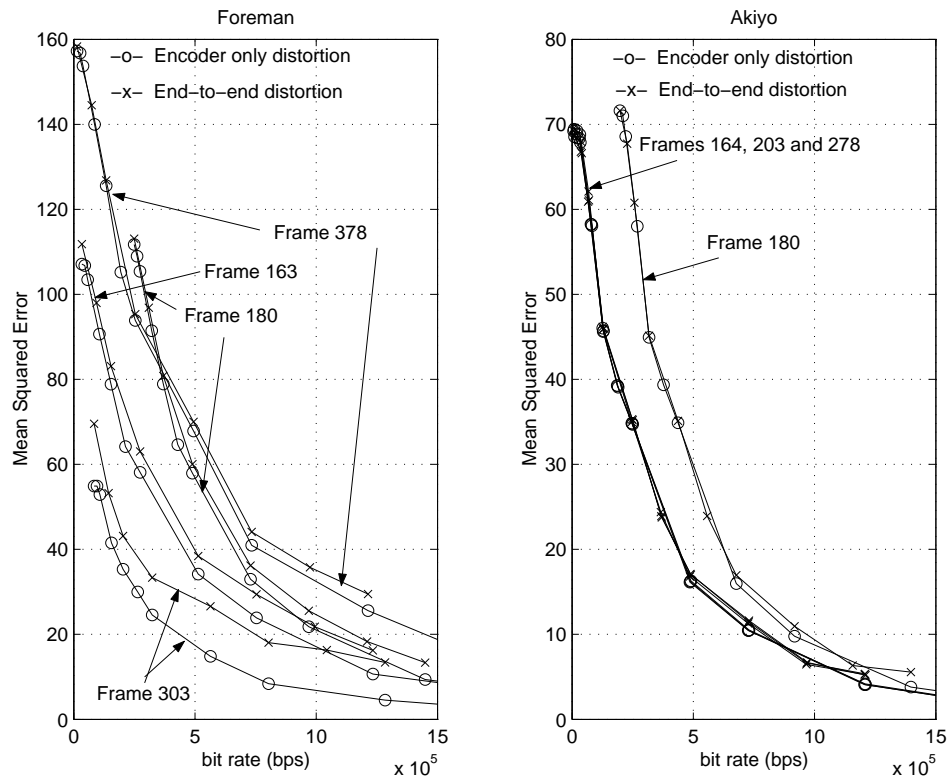


Figure 3.2: Distortion-Rate performance for different video frames.

ferentiated target BER specification for the base and the enhancement layer constitute in effect a simple but effective Unequal Error Protection scheme. Importantly, note in Figure 3.2 that the D-R performance changes from frame to frame, being more stable for sequences with low motion. Also, in most of the cases the contribution of channel errors to the end-to-end distortion is negligible. Nevertheless, there are cases, as in Foreman's frame 303 (part of a camera panning section), where this is not true. In these cases, channel-induced distortion is approximately the same for all encoding rates.

Overall, we can see that the application example just describes matches well the system setup described in the previous section. Although we will keep the study in the rest of this chapter general, in the sense of addressing real-time multimedia sources, we are

occasionally used this well matched application as a reference in our study. Nevertheless, it is worth noticing that this is not the only application example that matches well the system setup. In fact, the setup is general enough that it is applicable to many other real-time communication examples. Finally, we note that, in this case, at 30 frames per second the single-frame delay introduced in our system is completely acceptable for conversational video.

3.3 CDMA Statistical Multiplexing Resource Allocation and Flow Control

This section discusses the problem of allocating operating modes and power to all users based on traffic load and subject to the conditions that average source distortion per call is minimized, a maximum distortion is not exceeded and BER requirements are satisfied.

3.3.1 Multiuser Power and Rate Allocation

Assume the system setup in Section 3.2 with ideal power control, an additive, white, Gaussian noise (AWGN) channel and that a matched filter at the receiver. Then power assignment and interference from other users are related to the target SINR required by each call as [29],

$$\beta_i \geq \frac{(W/r_i) P_i}{\sigma^2 + \sum_{j \neq i} P_j}, \quad i = 1, 2, \dots, N, \quad (3.1)$$

where P_i is the power assigned to user i , as measured at the receiver, necessary to obtain the target SINR β_i and σ^2 is the background noise variance, which accounts for intercell interference [29]. The target SINRs are set based on the quality goal, i.e., following the discussion in Section 3.2, the target SINRs are set so as to not exceed the BER threshold

that limits channel errors to an acceptable magnitude. W/r_i is the processing gain. If it is possible to find feasible power assignments that satisfies the N inequalities (3.1) with equality, then these assignments minimizes the sum of the transmitted powers, [42]. Taking (3.1) as an equality and following the same procedure as in chapter 2, Equations (2.3)-(2.5), and [42] the power assignment is

$$P_i = \frac{\Psi_i \sigma^2}{1 - \sum_{j=1}^N \Psi_j}, \quad i = 1, 2, \dots, N. \quad (3.2)$$

where, again,

$$\Psi_i = \left(1 + \frac{W}{r_i \beta_i}\right)^{-1}. \quad (3.3)$$

As in 2, from (3.2) we can derive the following condition that limits the number of simultaneous serviced calls

$$\sum_{i=1}^N \Psi_i \leq 1 - \epsilon, \quad (3.4)$$

where ϵ is a small positive number set during design. Recall from chapter 2 that (3.4) represents limitations on the power amplifier dynamic range and system stability.

3.3.2 Source Encoder

We want the same optimization criterion for the flow control protocol adaptation rule as the one used in chapter 2, i.e. the adaptation needs to minimizes the average distortion per call. Let $f_i(x_i)$ be the distortion-rate (D-R) performance function of the i^{th} user source encoder at rate x_i . Then, the optimization goal can be equivalently written as

$$\min_{x_1, x_2, \dots, x_N} \sum_{i=1}^N f_i(x_i), \quad (3.5)$$

Typically, $f_i(x_i)$ would be a decreasing function. Since it is possible that goal (3.5) might be achieved with some users undergoing excessive distortion, we will limit each

user's source rate to some minimum value. On the other extreme, $f_i(x_i)$ is minimum when the rate is maximum; i.e. $x_i = x_{M_i}$. Also, as in chapter 2, we assume $f_i(x_i) = \alpha_i 2^{-k_i x_i}$. Since now we allow the processing gain to change, the channel coding rate is not automatically specified once the source encoding rate is known. Yet, the channel coding rate is specified once both the source encoding rate and the transmit bit rate (equivalently the processing) are known. Therefore, the optimization problem can be written as,

$$\min_{x_1, r_1, x_2, r_2, \dots, x_N, r_N} \sum_{i=1}^N f_i(x_i) \quad \text{subject to} \quad \sum_{i=1}^N \Psi_i(\beta_i, r_i) \leq 1 - \epsilon, \quad (3.6)$$

In addition, as discussed in Section 3.2, channel-induced distortion needs to be considered. When the design follows a quality goal that aims at preventing annoying channel impairments, in most of the cases this distortion can be neglected. Nevertheless, this distortion could be numerically significant in some cases of video sources, as is the case in our application example. Because in these cases the channel-induced distortion is independent of the source encoding rate, i.e. it is approximately constant for all encoding rates with a magnitude that is made acceptable by system design, we can ignore it from our formulation without loss of optimality. Note that meeting the quality goal is implied in the constraint of (3.6) since Ψ_i depends on the target SINR.

3.3.3 CDMA Statistical Multiplexing, Flow Control and Resource Allocation

Next, we state the optimization problem in a form that highlights its close relation to a family of problems in source encoding research.

Proposition 5 *Let b_i be user i 's transmit rate and target SINR allocation pair (r_i, β_i) . Then, the problem (3.6) of optimal operating mode and power allocation to minimize*

average end-to-end distortion in the uplink of multiuser, single cell CDMA system with an arbitrary set of source encoders can be stated as

$$\min_{b_1, b_2, \dots, b_N} \sum_{i=1}^N D_i(b_i), \quad \text{s.t.} \quad \sum_{i=1}^N \Psi_i(b_i) \leq 1 - \epsilon, \quad (3.7)$$

where D_i is a distortion function.

Proof: The proof shows the equivalence between source and channel rate (operating modes) and the pair $b_i = (r_i, \beta_i)$ and the functional relation between b_i and D_i . From its definition, clearly Ψ_i is a function of only r_i and β_i . Let D_i be user's i source encoder distortion, which depends only on user i 's source encoder rate x_i when the target SINR is set to prevent annoying channel impairment (as discussed in the previous section). Consider that in our setup each call's transmit bit rate is divided between source coding and channel error protection bit rates. Then, given any two of user i 's source encoding, channel coding and transmit bit rate, the third is automatically determined. This is because the natural allocation guideline is to maximize the transmit bit rate utilization by maximizing source encoding rate or minimizing channel coding rate (maximizing error protection). Also, if β_i is given, user i 's channel coding rate is automatically determined as the one that provides enough error protection to achieve the quality goal. Therefore, if user's i target SINR is changed, then the channel rate will need to be changed so as to maintain the quality goal and x_i will change for a fixed r_i . Also, if r_i is changed, x_i will need to be changed for a fixed β_i . In summary, the source coding rate x_i implicitly depends on r_i and β_i through the purpose of maintaining a quality goal and maximizing transmit bit rate utilization. Even more, each pair b_i has associated only one value $D_i(b_i)$. Therefore, D_i is a function of b_i . \square

The problem stated in (3.7) is analogous to the problem studied in [45] of allocating a bit quota R_b to an arbitrary set of quantizers. The problem is also analogous to the

one studied in [4] of allocating a fixed bandwidth among a number of users in a TDMA network. This analogy allows us to note that Ψ_i can be considered as the *equivalent bandwidth* assigned to user i out of the total $1 - \epsilon$. This definition is consistent to the definition of equivalent bandwidth in [8] and to the definition of effective interference in [48]. In fact the solution to (3.7) studied next can be considered as the solution to effective bandwidth assignment between real-time calls in a multiuser CDMA system, which is unsolved in [8]. Clearly, there is a direct analogy between problem (3.7) and statistical multiplexing. In essence, Proposition 5 states that the problem of modes and power allocation in the uplink of CDMA can be considered as performing statistical multiplexing in a multiuser CDMA setup. Furthermore, the formulation as presented is general but powerful enough that it allows including other related resource allocation problems in CDMA such as the ones studied in [26, 27]. The problem (3.7) differs from the one in [45] in that distortion now is a function of two variables, namely transmit bit rate and target SINR (as opposed to source bit rate only) and that the constraint function is the sum of functions of transmit bit rate and target SINR instead of just sum of allocated bits. We next extend the results in [45] and apply them to optimally solve (3.7).

Let $\mathcal{S}_r^{(i)}$ and $\mathcal{S}_\beta^{(i)}$ be the finite set of all user i 's possible transmit rates and target SINRs, respectively. Let $\mathcal{S}^{(i)}$ be the set of all user i 's possible allocation vectors $b_i = (r_i, \beta_i)$ and \mathcal{S} be the set of all possible allocations $B = \{b_1, b_2, \dots, b_N\}$. Let $H(B)$ be some real-valued function, called the objective function of B , defined for all $B \in \mathcal{S}$. Let $R(B)$ be some real-valued function, called the constraint function of B , defined for all $B \in \mathcal{S}$.

Theorem 2 *There exists a $\lambda \geq 0$ such that the optimal solution, $B^*(\lambda)$, to the constraint*

problem

$$\min_{B \in \mathcal{S}} H(B), \text{ subject to } R(B) \leq R_c,$$

with $R(B^*(\lambda)) = R_c$ is the solution to the unconstrained problem $\min_{B \in \mathcal{S}} \{H(B) + \lambda R(B)\}$ also.

Proof: The proof is in [45]. We next summarize the proof to emphasize that it still holds for the case of the problem under study here.

$$H(B^*) + \lambda R(B^*) \leq H(B) + \lambda R(B)$$

for all B in \mathcal{S} . Then, we have

$$H(B^*) - H(B) \leq \lambda R(B) - \lambda R(B^*),$$

which is true for all B in \mathcal{S} . Thus, (3.8) is true for all B in the subset of \mathcal{S} , $\mathcal{S}^* = \{B : R(B) \leq R(B^*)\}$. Since $\lambda \geq 0$

$$H(B^*) - H(B) \leq 0.$$

This means that B^* is the solution to the constrained problem with $R_c = R(B^*)$. \square

For the particular case of problem (3.7) we can say the following: Let $H(B)$ and $R(B)$ be of the form $H(B) = \sum_{i=1}^N D_i(b_i)$ and $R(B) = \sum_{i=1}^N \Psi_i(b_i)$, respectively.

Then, the unconstrained problem $\min_{B \in \mathcal{S}} \{H(B) + \lambda R(B)\}$, $\lambda \geq 0$, can be written as

$$\min_{B \in \mathcal{S}} \left\{ \sum_{i=1}^N D_i(b_i) + \lambda \sum_{i=1}^N \Psi_i(b_i) \right\}. \quad (3.8)$$

Note that the solution $B^*(\lambda) = \{b_1^*(\lambda), \dots, b_N^*(\lambda)\}$ can be obtained by minimizing each term of the sum in the unconstrained problem separately, i.e. $b_k^*(\lambda)$ solves

$$\min_{b_i \in \mathcal{S}^{(i)}} \{D_i(b_i) + \lambda \Psi_i(b_i)\}. \quad (3.9)$$

Theorem 3 Let $D_i(b_i)$ and $\Psi_i(b_i)$ be real-valued functions over some closed domain on the real line. Let $b_i^*(\lambda_1)$ be a solution to $\min_{b_i \in \mathcal{S}} \{D_i(b_i) + \lambda_1 \Psi_i(b_i)\}$ and let $b_i^*(\lambda_2)$ be a solution to $\min_{b_i \in \mathcal{S}} \{D_i(b_i) + \lambda_2 \Psi_i(b_i)\}$. Then for any function $D_i(b_i)$,

$$0 \leq (\lambda_2 - \lambda_1) \left(\Psi_i(b_i^*(\lambda_1)) - \Psi_i(b_i^*(\lambda_2)) \right).$$

Proof: Following [45], by definition of $b_i^*(\lambda_1)$ and $b_i^*(\lambda_2)$, we have

$$\begin{aligned} D_i(b_i^*(\lambda_2)) + \lambda_2 \Psi_i(b_i^*(\lambda_2)) &\leq D_i(b_i^*(\lambda_1)) + \lambda_2 \Psi_i(b_i^*(\lambda_1)) \\ D_i(b_i^*(\lambda_1)) + \lambda_1 \Psi_i(b_i^*(\lambda_1)) &\leq D_i(b_i^*(\lambda_2)) + \lambda_1 \Psi_i(b_i^*(\lambda_2)). \end{aligned}$$

From this we get,

$$\begin{aligned} D_i(b_i^*(\lambda_1)) - D_i(b_i^*(\lambda_2)) &\leq \lambda_1 [\Psi_i(b_i^*(\lambda_2)) - \Psi_i(b_i^*(\lambda_1))] \\ D_i(b_i^*(\lambda_2)) - D_i(b_i^*(\lambda_1)) &\leq \lambda_2 [\Psi_i(b_i^*(\lambda_1)) - \Psi_i(b_i^*(\lambda_2))]. \end{aligned}$$

Adding both sides of these inequalities proves the theorem. \square

Corollary 1 The solutions $\Psi_i(b_i^*(\lambda))$, for all i , and the corresponding constraint function $R^*(\lambda) = \sum_{i=1}^N \Psi_i(b_i^*(\lambda))$ are monotonically nonincreasing with λ , i.e. if $\lambda_2 \geq \lambda_1 > 0$, then $\Psi_i(b_i^*(\lambda_2)) \leq \Psi_i(b_i^*(\lambda_1))$, and $R^*(\lambda_2) \leq R^*(\lambda_1)$.

Proof: Theorem 3 says that as λ increases, the minimizing value for $b(\lambda)$ makes the resulting $\Psi_i(b_i^*(\lambda))$ either increase or stay the same. This proves the corollary for $\Psi_i(b_i^*(\lambda))$ and hence for the sum $R^*(\lambda) = \sum_{i=1}^N \Psi_i(b_i^*(\lambda))$. \square

Figure 3.3 shows a typical behavior of $\sum_{i=1}^N \Psi_i(b_i)$ as a function of λ . The curve in this figure was obtained from actual simulations, which are later detailed in this chapter.

We next use the theory just presented to develop two algorithms that solved problem (3.7) by optimally allocating resources among calls. We first describe the simpler, but

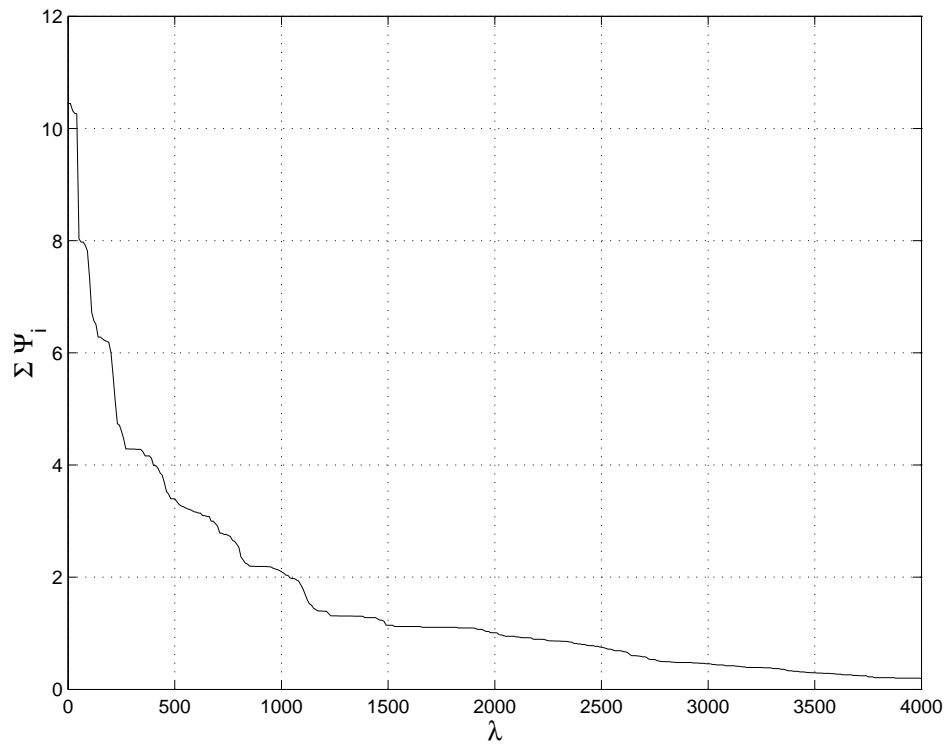


Figure 3.3: Total optimal equivalent bandwidth as a function of λ .

also important, case where the resource that is adapted is the transmit bit rate while the target SINR is kept unchanged. After this we discuss the general case where both transmit bit rate and target SINR are jointly adapted.

3.3.4 Flow Control by Transmit Bit Rate Adaptation

In this case, channel coding rate and, consequently, target SINR are assumed fixed, thus $b_i = r_i$. As discussed in Section 3.2 we assume that each call's D-R performance is known at the base station. Based on this information, the flow control protocol will choose, for each call, a transmit bit rate, $r_i = \hat{r}_i$, large enough so as to reach a target small distortion. If the network is lightly loaded and the total requests from all calls meets (3.4), all users are allocated resources so that they operate with best quality. If the network is congested ((3.4) fails) then a congestion resolution algorithm needs to solve (3.7). Because the problem in Proposition 5 is analogous to the ones studied in [4] and [45], we relied on these works to find a low-complexity greedy but optimal solution.

Given a finite set of available transmit rates $\mathbf{r} = \{r_{t_1}, r_{t_2}, \dots, r_{t_M}\}$, we define $\Delta_i^{(j)}$, the i^{th} incremental distortion associated with call j , as the distortion reduction caused by increasing the transmit rate one discrete step $\Delta q = r_{t_{i+1}} - r_{t_i}$, i.e.,

$$\Delta_i^{(j)} = D_j(\Psi_j(r_{t_{i+1}})) - D_j(\Psi_j(r_{t_i})).$$

The algorithm is based on a table associated with the incremental distortions. The table stores all pairs of indices (j, i) in increasing order of their associated incremental distortions, while also respecting each user's rate reduction order. The pair (j_1, i_1) precedes in the table the pair (j_2, i_2) if $i_1 > i_2$ and $j_1 = j_2$, or if $\Delta_{i_1}^{(j_1)} < \Delta_{i_2}^{(j_2)}$ and $j_1 \neq j_2$. There is a "0" in the first location of the table. In the second location of the table, there is a pair (j, i) corresponding to the smallest possible incremental distortion. There is a pointer p that addresses a location in the table.

The overflow resolution algorithm proceeds through iterations that we denote with the index m . For the m^{th} iteration, we let $r_i^{(m)}$ and $\Psi_i^{(m)}$ denote transmit rate and effective bandwidth associated with the i^{th} call, respectively. Also, let $\mathbb{S}^{(m)} = \sum_{i=1}^N \Psi_i^{(m)}$ denote the total effective bandwidth assigned at the m^{th} iteration. The overflow resolution algorithm is described next.

Algorithm 1:

1. Initialize the variables: $m = 0$, $r_i^{(0)} = \hat{r}_i$, $\Psi_i^{(0)}(\hat{r}_i)$ for $i = 1, 2, \dots, N$, and $\mathbb{S}^{(0)} = \sum_{i=1}^N \Psi_i^{(0)}$. Set the pointer $p = 1$ to the first entry in the table.
2. If $\mathbb{S}^{(m)} \leq 1 - \epsilon$ then there is no overflow, and so GO TO STEP 4; else, an overflow has occurred, and so GO TO STEP 3.
3. Set $p \leftarrow p + 1$. If p exceeds the table length, it is not possible to perform allocation subject to the minimum per-user distortion constraint, then EXIT and report OUTAGE. Else, the p^{th} entry of the table is a pair (j, i) , indicating that to optimally resolve the overflow, the transmit rate of the j^{th} user need to be updated, i.e. for $r_j^{(m)} = r_{t_{i+1}}$, $r_j^{(m+1)} \leftarrow r_{t_i}$ and $r_l^{(m+1)} \leftarrow r_l^{(m)}$ for $l \neq j$. Update $\Psi_i^{(m+1)}$ and $\mathbb{S}^{(m+1)}$. GO TO STEP 2 and proceed with next iteration.
4. EXIT STEP: if $\mathbb{S}^{(m)} < 1 - \epsilon$, it means that the network is not fully loaded. Also, if $p \neq 1$, it means that overflow has occurred and has been resolved.

The following theorem demonstrates that this greedy algorithm is optimal.

Theorem 4 *If the D-R functions $D_i(r_i)$ are convex and decreasing (as is commonly the case), then the proposed greedy algorithm for overflow resolution provides the optimal rate assignment minimizing the average distortion per call.*

Proof: We establish this assertion by mathematical induction. Since initialization assures the minimum absolute average distortion, the claim holds true for $m = 1$. Assuming optimal assignment at the m^{th} iteration, we prove that at iteration $m + 1$ the algorithm is optimal too. To do this, it suffices to show that the rate reduction in the optimal assignment in the m^{th} step will not be required for the optimal assignment in the $m + 1$ st step. In other words, the optimal rate assigned to each call at the m^{th} step is no less than the optimal rate assigned to each call at the $m + 1$ st step.

If β_i is fixed, so is the amount of error protection (channel coding rate) necessary to achieve the quality goal. Also, in the most general setting it is possible to use an unequal error protection (UEP) scheme where different source-encoded bits receive different error protection. In this case, the source encoding rate is the result of an affine mapping $x_i = \sum_k x_{ik} = \sum_k R_{ik} r_{ik}$, where R_{ik} is the fixed channel coding rate for each group of source bits and r_{ik} is the corresponding transmit bit rate with $r_i = \sum_k r_{ik}$. Also, it is assumed that the functions $h_k(r_i)$ that specifies how a change in r_i affects each r_{ik} are of the form $r_{ik} = h_k r_i + r_{ik}^o$, where $r_{ik}^o \geq 0$, $h_k \geq 0$ and $\sum_k h_k = 1$. Then, $D_i(r_i) = \alpha_i 2^{-k_i(r_i \sum_k h_k R_{ik} + \sum_k r_{ik}^o R_{ik})}$ is a convex function, decreasing in r_i . We have asserted that the solution to the constrained problem (3.7) can be obtained by minimizing each term of the sum separately, i.e.,

$$\min_{r_i} \{D_i(r_i) + \lambda \Psi_i(r_i)\}, \quad i = 1, 2, \dots, N. \quad (3.10)$$

It is important to notice that the same λ appears for all the terms independently of i and that $\Psi_i(r_i)$ is increasing. At the m^{th} step of the algorithm, the network obtains the optimal solution for a total assigned equivalent bandwidth $S^{(m)}$. Equivalently, from Theorem 2, at the m^{th} step of the algorithm, there exists a positive $\lambda^{(m)}$ corresponding to $\min \left\{ \sum_{i=1}^N D_i(r_i^{(m)}) + \lambda^{(m)} \sum_{i=1}^N \Psi_i(r_i^{(m)}) \right\}$. At the next step, if overflow persists, the rate assignment to at least one of the calls has to decrease so as to lower its equiv-

alent bandwidth. Because, from Corollary 1, the optimal equivalent bandwidth for this rate assignment is a nonincreasing function of the value of $\lambda^{(m+1)}$, this suggests that $\lambda^{(m+1)} > \lambda^{(m)}$. However, $\lambda^{(m+1)}$ is the same for all calls, independently of i . Therefore, as the algorithm proceeds, the Lagrange multiplier coefficient increases or remains unchanged, and the optimal rates (and equivalent bandwidths) for all calls decrease or remain unchanged. As a result, to achieve the optimal solution in a given step, the algorithm never needs to increase back the rate assignment to a call whose rate was reduced in previous steps; therefore, the proposed iterative greedy algorithm provides the optimal solution. \square

3.3.5 Flow Control by Transmit Bit Rate and Target SINR Adaptation

This case requires a different algorithm than the one just described because it cannot be asserted that $D(b_i)$ is convex and decreasing on the pair b_i . Nevertheless, based on the theory discussed above, we describe next an iterative algorithm to optimally allocate the pairs $b_i = (r_i, \beta_i)$ to all calls, where we use again m as the iteration index.

Algorithm 2:

1. Initialize $\lambda^{(0)}$ with some positive number.
2. Solve each of the N unconstrained problems

$$\min_{b_i(\lambda^{(m)})} \{D_i(b_i) + \lambda^{(m)}\Psi_i(b_i)\}, \quad (3.11)$$

and update $\mathbb{S}^{(m)} = \sum_{i=1}^N \Psi_i(b_i(\lambda^{(m)}))$.

3. (a) If $\mathbb{S}^{(m)} > 1 - \epsilon$ and stopping criteria is false then update $\lambda^{(m+1)}$ such that $\lambda^{(m+1)} > \lambda^{(m)}$, so that the effective interferences, $\Psi_i(b_i)$, will be reduced. GO TO STEP 2.

(b) If $\mathbb{S}^{(m)} < 1 - \epsilon$ and stopping criteria is false then update $\lambda^{(m+1)}$ such that $\lambda^{(m+1)} < \lambda^{(m)}$, so that the effective interferences, $\Psi_i(b_i)$, will be increased. GO TO STEP 2.

4. *Stopping criteria:* Iterations stops whenever one of the following occurs: (a) $\mathbb{S}^{(m)}$ is sufficiently close to, but less than, $1 - \epsilon$, this is optimal allocation in the presence of congestion; (b) $\mathbb{S}^{(m)} < 1 - \epsilon$ and all allocations b_i s correspond to the threshold minimum target distortion, this corresponds to a lightly loaded network; (c) $\mathbb{S}^{(m)} > 1 - \epsilon$ and all allocations b_i s correspond to the maximum allowable distortion, this corresponds to an outage condition and could be avoided with high probability by proper admission control.

Note that in practice both the number of available transmit bit rates and target SINRs are finite and typically small. Therefore, each minimization in step 2 is easily solved by exhaustive search, where each possible $\Psi_i(b_i)$ could be calculated offline and each $D_i(b_i)$ is essentially the D-R performance information communicated to the base station. For example, if there are eight possible channel coding rates and target SINRs each, the problem reduces to choosing the smallest element in a matrix resulting from adding two 8-by-8 matrices. The updates of λ in steps 3a and 3b can be done following any of the methods suggested in the literature ([45]), in our case we used a simple bisection.

Finally, the following algorithm discusses optimality of algorithm 2.

Theorem 5 *Algorithm 2 is optimum.*

Proof: The proof is straightforward by noticing that the algorithm performs an iterative search for the allocation that meets the optimality criteria in Theorem 2. \square

3.3.6 Rate-Distortion Data Overhead

As highlighted in Section 3.2, the centralized algorithms just described need an added communication overhead to learn each call D-R performance information. This overhead is clearly largest in the case of Algorithm 2 because it needs the distortion values corresponding to each possible pair b_i of transmit bit rate and target SINR. This also adds to the complexity of mobile terminal's encoder since it needs to compute each of these distortion values (this is a problem in a distributed algorithm also). We solved both problems by summarizing the D-R performance. Instead of sending the distortion and rate data for each operating mode, each mobile sends the following: one reference encoding bit rate and three distortion values at predefined bit rate points. These three predefined bit rate points are separated from the reference encoding bit rate (the one that is sent) by fixed bit rate values which are suitable picked to represent high, medium and low distortion values. The base station use the transmitted data to first calculate the encoding rate of the three distortion points and then approximates two curves of the form $f(x) = \alpha 2^{-kx}$, one for the high distortion section of the D-R performance using the high and medium bit rate distortion points and another for the low distortion section using the low and medium bit rate distortion points. The rest of the distortion-rate points are calculated by interpolation using the approximate D-R curves. As we shall see in Section 3.5, this scheme allows a representation for the D-R performance that has low-overhead, involves computing only three distortion points at the mobile station and that does not degrades performance of the overall allocation algorithm .

3.4 Analysis for Dynamic Call Traffic and Admission Control

So far we have considered a static network model where the number of calls N is fixed. In reality, this number is a random variable that depends on the traffic in the cell under consideration. Therefore, we want next to study the proposed scheme when the number of calls dynamically changes over time and, based on this, address the problem of admission control.

We assume that calls enter the cell at a rate ν following a Poisson arrival process and that the random calls duration follow an exponential distribution with mean $1/\mu$. As discussed in chapter 2 one possible approach is to model the CDMA network as an $M/M/\infty$ queue and base admission control on the outage probability, where outage occurs when the system exceeds some operational parameter. In particular, the failure of (3.4) has been typically considered as an outage condition [43, 52]. One key feature of our system is that it prevents condition (3.4) from failing at the cost of a smooth increased in end-to-end distortion. Therefore, the relevant operation for the call admission control is to limit the maximum number of calls to a maximum N_L , where N_L is set so that $\bar{D}_{N_L} = D_M$, D_M being the maximum tolerable expected distortion and \bar{D}_N the expected distortion per call when there are N calls. Then, considering the operational principles of the proposed scheme, for the purpose of call admission control it is more pertinent to model the network as a $M/M/N_L/N_L$ blocking system. Note, again, that this model better represents the problem of call admission control from the network operator's viewpoint, because it rejects new calls once a maximum number has been reached so as to maintain quality for the existing calls.

This queue model can be represented in the form of a state transition diagram as in

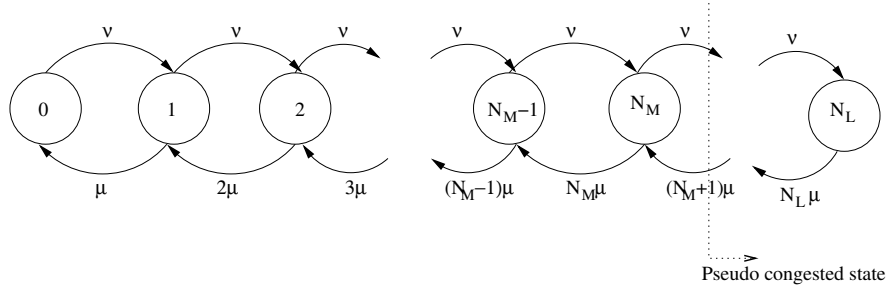


Figure 3.4: Markov chain representation of the $M/M/N_L/N_L$ traffic model.

Figure 3.4. From queuing theory, the steady-state probability that there are N calls in the network is, [3],

$$q_n \triangleq P[n = N] = \frac{\phi_N}{\sum_{i=0}^{N_L} \phi_i},$$

where $\phi_i = \rho^i / i!$ and $\rho = \nu / \mu$ is the offered load. Therefore, assuming ergodic processes, over a sufficiently large period of time the average distortion per call as a function of the offered load will be

$$E[\bar{D}_n] = \frac{\sum_{n=0}^{N_L} \bar{D}_n \phi_n}{\sum_{i=0}^{N_L} \phi_i}. \quad (3.12)$$

In contrast to our study in chapter 2, the fact that we are considering arbitrary source encoders, coupled with the use of our algorithms to allocate resources, makes it difficult to obtain a close form solution for \bar{D}_N . In our simulations we address this issue by estimating \bar{D}_N from Monte Carlo simulations.

As in the study in Section 2.4, it is possible to recognize three different operating conditions in our system. When the system is lightly loaded and it is possible to allocate resources to all users such that they all meet their target distortion, the system would be operating in a congestion-free situation. New calls arriving when there are N_L calls in the network are denied serviced and dropped from the system. This corresponds to

network congestion. The third operating condition is that where new calls are accepted but at least one call cannot be granted resources to operate at the target distortion. We called this state as *pseudo congestion state*. Note that there is a fundamental difference between the pseudo congestion state in this case and the one in chapter 2. Because now we are dealing with an arbitrary set of source coders, in the pseudo congestion state some calls may still be operating at target distortion (while the rest not). In contrast, because the setup in chapter 2 assumed all source encoders sharing the same state and D-R performance, in the pseudo congestion state all calls were operating above target distortion. Based on this discussion, consider the following definitions

Definition 3 We call pseudo congestion state the operational state when one or more users are operating at a source rate lower than the one corresponding to the target distortion goal. With

$$P_{out} \triangleq P \left[\sum_{i=1}^N \Psi_i(\hat{b}_i) > 1 - \epsilon \right], \quad (3.13)$$

the steady state probability of this state is given by

$$P_{sc} = \sum_{N=0}^{N_L} P_{out} q_n = \frac{\sum_{N=0}^{N_L} P_{out} \phi_N}{\sum_{i=0}^{N_L} \phi_i}, \quad (3.14)$$

where \hat{b}_i is user' i allocation such that it meets its target distortion.

Definition 4 We call the operational state when new incoming calls need to be blocked as congestion state. This corresponds to the situation when $N = N_L$ and a new call arrives. The probability of this event is queuing theory's blocking probability. From the PASTA property [3], this probability is given by $P_b = q_{N_L}$.

From these definitions we see that the maximum number of users and the blocking probability can be determined from the maximum tolerable expected distortion. Also

the probability of operating in the pseudo congestion state depends on the traffic load and each user's target distortion and source encoder D-R performance. Note that in (3.14), P_{out} corresponds to the outage probability in $M/M/\infty$ models, [43, 52]. The outage states in [43, 52] have now been divided into a a congestion state and a pseudo congestion state, where new calls are still admitted, with the probability of at least one call operating with a distortion larger than the minimum target given by P_{sc} . We can see that the system studied in this chapter performs statistical multiplexing in such a way that it avoids congestion (or outage) by smoothly increasing source distortion up to the point where a maximum expected distortion is reached. Further discussion of this issue, for the specific case of real time MPEG4 FGS video, follows in the next section.

3.5 Performance Evaluation

We evaluated the performance of the designs studied in this chapter through Monte Carlo simulations. As discussed in Section 3.2.2, a CDMA system carrying video calls is an application that matches very well our problem setup. Because of this, we based the simulations on this application case. We designed the simulation so as to have a system that could support a reasonable number of video calls at good quality. Roughly half of the calls used the "Foreman" sequence and the rest used "Akiyo", both with the same characteristics as described in Section 3.2.2. To assure that all user's sequences were desynchronized with respect to the others, each sequence started at a random frame and were assumed to be a circular loop, i.e. the first frame followed the last frame once the end of a sequence was reached. The sequences were encoded using a MPEG4 FGS coder [36]. In order for the source encoded sequence to be transmitted over a noisy channel we divided, as error resiliency feature, the bit stream into packets. Those

packets for which errors were detected at the receiver after the error control coding block were discarded and replaced using an error concealment scheme. The error concealment replaced lost packets by using the corresponding correctly received previously packet and then applying motion compensation as necessary. For variable-rate channel coder we choose an RCPC code with mother code rate $1/4$, $K=9$ and puncturing period 8 [14]. We assumed system bandwidth equal to 40 Mhz. with available transmit rates of 5000, 2500, 1250, 625, 312.5, 156.25, 78.125 and 39.0625 Kbps. This corresponds to a choice for possible variable spreading factors similar to the OVVSF (Orthogonal Variable Spreading Factor) chosen for the UMTS standard [54]. Each user requested resources so as to achieve a target PSNR of at least 36 dB, corresponding to a reasonable good quality for both high and low motion sequences. As explained in Section 3.2.2, we set a target BER of 5×10^{-6} for the base layer and 10^{-5} for the enhancement layer, which corresponds to a simple form of unequal error protection. For the solutions that could adapt the target SINR, the possible values were 1.93, 1.76, 1.63, 1.47, 1.35, 1.16, 1 and 0.81 dBs. For these target SINRs and in order to guarantee the target channel BER, the corresponding available channel coding rates were $8/16$, $8/17$, $8/18$, $8/20$, $8/21$, $8/24$, $8/27$ and $8/32$ for the base layer and $8/16$, $8/17$, $8/18$, $8/19$, $8/20$, $8/23$, $8/26$ and $8/31$ for the enhancement layer. Other simulations parameters were $\sigma^2 = 10^{-6}$ and $\epsilon = .1$, unless otherwise noted.

We evaluated three different systems: one where the calls request resources so as to achieve some quality level but cannot perform any adaptation, another where calls can change transmit rate by changing source encoding rate with a fixed target SINR using Algorithm 1 and a third system where both transmit rate and target SINR are adapted and are allocated using Algorithm 2. We denoted the three systems as “No Adaptation”, “Transmit Rate Adaptation” and “Full Adaptation”, respectively. For both

the system with no adaptation and the one with only transmit rate adaptation we used the highest possible channel coding rate. Figure 3.5 shows the simulation result, where we gradually increased the number of users in the system and we measured the PSNR averaged over all calls and frames. Note how the performance of the system with no adaptation rapidly degrades as the number of calls increases. This behavior, which is justified by the inability of this system to perform any adaptation, have been already observed in [8]. Here, as more users are admitted it eventually becomes impossible to set the powers at a level such that the target SINRs and condition (3.4) are all met. Thus, each mobile user becomes constrained by the power amplifier dynamic range limit, their SINRs decrease and quality degrades due to the increase in BER. Simulations for this system stopped at 19 calls because at that point the degradation was so severe that the error concealment scheme essentially kept reproducing a frozen frame with no change in the sequence. Also, we can see in Figure 3.5 that both adapted systems are able to achieve both the target SINR and condition (3.4) for a larger number of users and that as more user are admitted into the system the distortion increases smoothly allowing an increase of roughly three times in the number of calls. In addition, we can see that the “Full Adaptation” system outperforms the “Transmit Rate Adaptation” one by roughly 0.8 dB in most of the operating points.

In Section 3.3.6 we addressed how to reduce the communication overhead necessary to send each call’s source encoder D-R data. Figure 3.5 also includes, labeled as “Compressed D-R data”, the simulations results for the same system with full adaptation but with the D-R performance information sent using the low overhead scheme described in Section 3.3.6. In this case the required overhead is equal to only five bytes per frame, one byte for each of the three distortion values and two to represents the number of bits in the base layer (rate data). In contrast, without implementing any scheme it would have

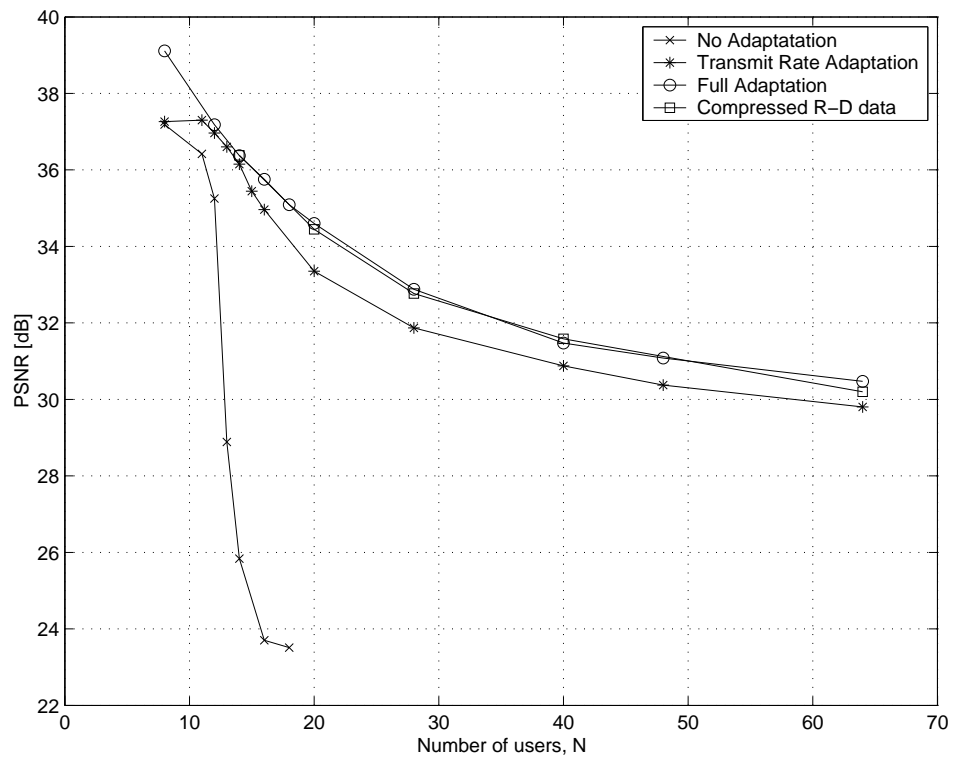


Figure 3.5: Comparison of three CDMA systems supporting video calls

been necessary to send a total of 64 D-R values per frame because in our setup there are eight possible channel coding rates and eight possible target SINRs. During simulations we noted that although in some cases the error in estimating distortion when using the reduced D-R representation were as high as 10 %, the algorithm was robust enough that there was no performance loss. Also, we noted that using extra data compression techniques could further reduce the overhead.

Similarly to the simpler scheme studied in chapter 2, we can make the important observation that both adapted systems (those performing statistical multiplexing) present the extra advantage that the increase in distortion is smooth and controllable. This is because channel-induced errors are kept at a perceptually acceptable and small value while distortion mostly follows the predictable rate-distortion function. This manifests mostly as a gradual blurring of the images. This is not the case for the system with no adaptation. In this case, the increase in distortion is a consequence of the uncontrolled increase in the BER and the associated random effects from increased channel-induced errors which are subjectively more annoying (mostly appearing as noticeable blocking artifacts and freezing of frames sections). This observation is further illustrated in Figures 3.6-3.11. These figures show results from frames that are representative of the results. Figures 3.6-3.8 show results corresponding to the system with no adaptation. We can see how the increase in BER and channel errors creates artifacts that, at the very least, are clearly noticeable, and in many cases affect understandability of the frame content. As expected, we observed that these artifacts were more frequent as the number of ongoing calls increased. Figures 3.9 and 3.11 show results when using algorithm 2. In Figure 3.9 the blurring associated with the smooth increase in distortion starts to become noticeable, especially in the region of the eyes and eyebrows. Nevertheless, there is clearly no annoying artifacts or important loss of understandability of the frame content. In 3.11



Figure 3.6: A frame from the sequence ‘Foreman’ when the network operates with the scheme with no adaptation and there are 13 ongoing calls in the network.

we can see how the blurring effect increases with the number of calls. Finally, in Figure 3.10 we can see the result when using algorithm 1 and the setup being the same as for Figure 3.9. As expected from the objective measurements, there is some degradation in the performance of algorithm 1 when compared to algorithm 2, but still many of the behavioral properties still hold.

As already discussed, since the problem under study focuses on a real-time communication network, it is not enough to evaluate results by fixing the number of calls. It is also important to consider a dynamic system and evaluate performance as a function of the traffic load. The results in figures 3.12 through 3.15 focuses on this evaluation approach, following the system setup described in Section 3.4 with $\mu = 3$ min. and $D_M = 30$ dB. Figure 3.12 shows the expected distortion per call as a function of the offered load. We can see that the “Full Adaptation” system can support an offered load



Figure 3.7: A frame from the sequence 'Foreman' when the network operates with the scheme with no adaptation and there are 14 ongoing calls in the network.



Figure 3.8: A frame from the sequence 'Foreman' when the network operates with the scheme with no adaptation and there are 18 ongoing calls in the network.



Figure 3.9: A frame from the sequence 'Foreman' when the network operates with using algorithm 2 and there are 30 ongoing calls in the network.



Figure 3.10: A frame from the sequence 'Foreman' when the network operates with using algorithm 1 and there are 30 ongoing calls in the network.



Figure 3.11: A frame from the sequence ‘Foreman’ when the network operates with using algorithm 2 and there are 60 ongoing calls in the network.

roughly 50 % larger than the “No Adaptation” system. Figure 3.13 shows the probability of pseudo congestion as a function of the offered load for different values of ϵ . We can see here that there is a range of offered loads where the probability of pseudo congestion transition from 0 to 1. This is the typical region of focus when studying Erlang capacity of $M/M/\infty$ CDMA networks [43, 52]. For larger offered loads we can see that although the probability of pseudo congestion is 1 (equivalently the outage probability in [43, 52]) the system is still able to accept more calls. Figure 3.14 highlights the fact that the extension of operation into the pseudo congestion state is achieved at the cost of a smooth and controlled degradation of quality by showing expected distortion as a function of probability of pseudo congestion. Finally, Figure 3.15 shows blocking probability as a function of the offered load when call admission control for both the “Full Adaptation” and the “No Adaptation” systems is performed so that the

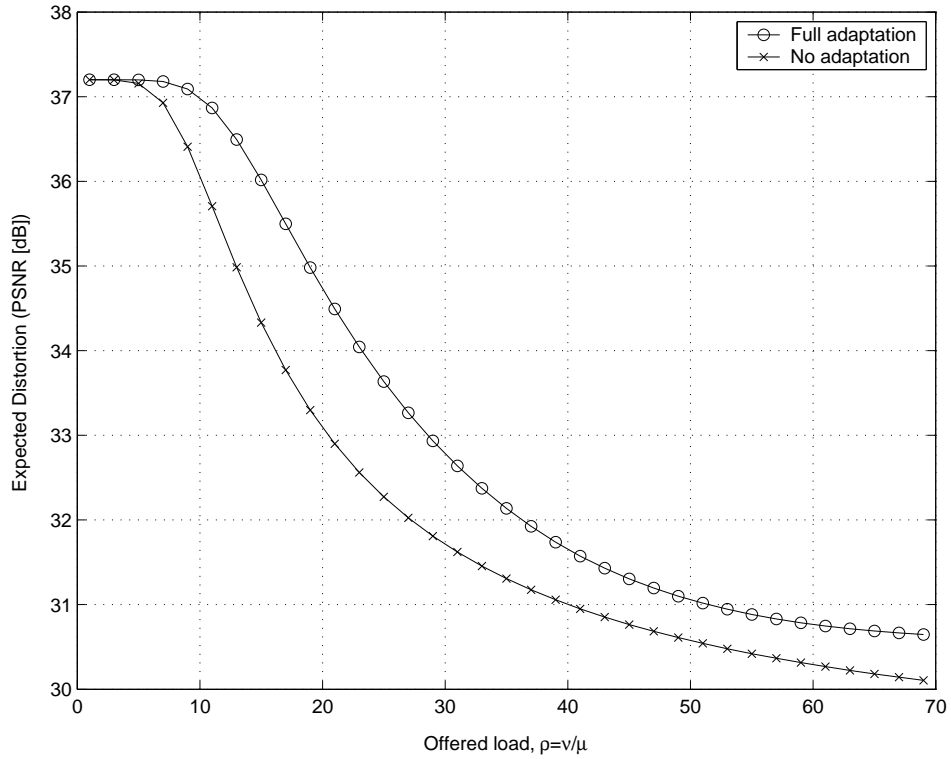


Figure 3.12: Expected distortion (PSNR in dBs) as a function of the offered load.

average distortion when the number of calls is maximum does not exceeds 30 dB. For this performance measure the difference between a smooth increase in distortion, as is the case for the “Full Adaptation” system, and the steep increase, as is the case for the “No Adaptation” system, translates into the “Full Adaptation” system supporting for the same limiting blocking probability more than 5 times the offered load than the “No Adaptation”.

3.6 Conclusions

In this chapter we have studied the solution to the problem of optimal adaptation to resolve interference-generated congestion for an arbitrary set of real-time source encoders

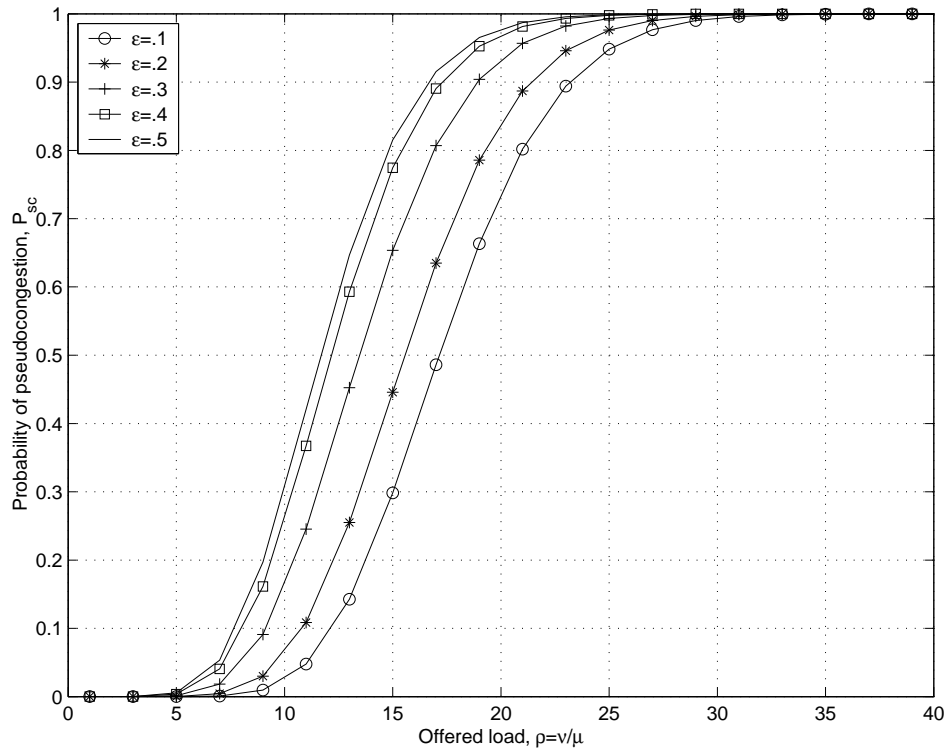


Figure 3.13: Probability of pseudo congestion as a function of the offered load.

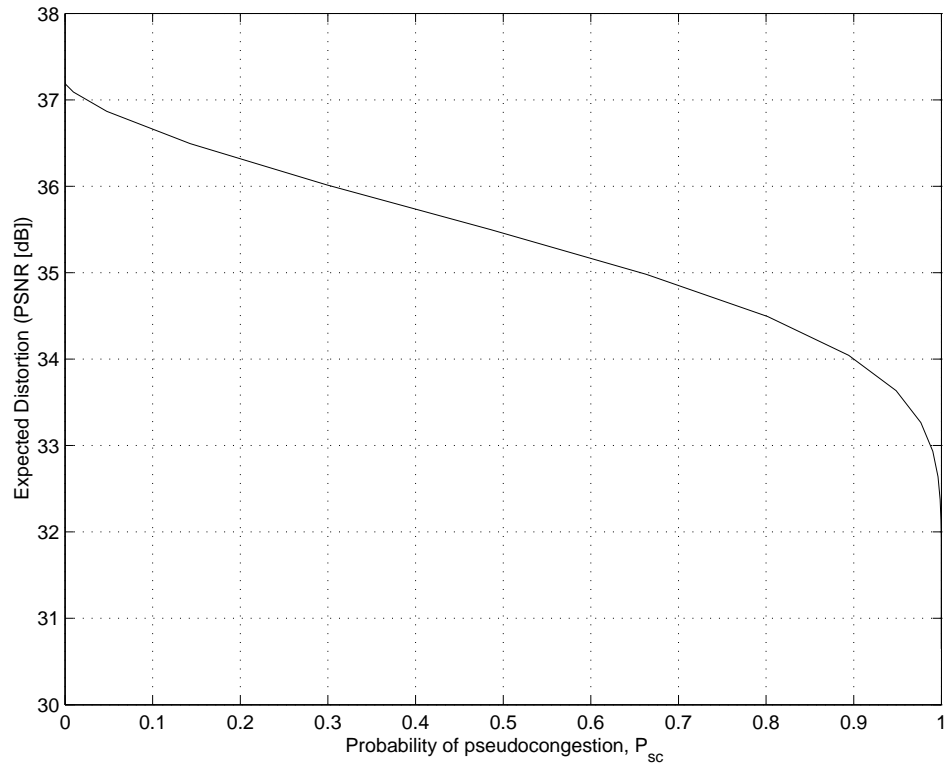


Figure 3.14: Expected distortion (PSNR in dBs) as a function of probability of pseudo congestion.

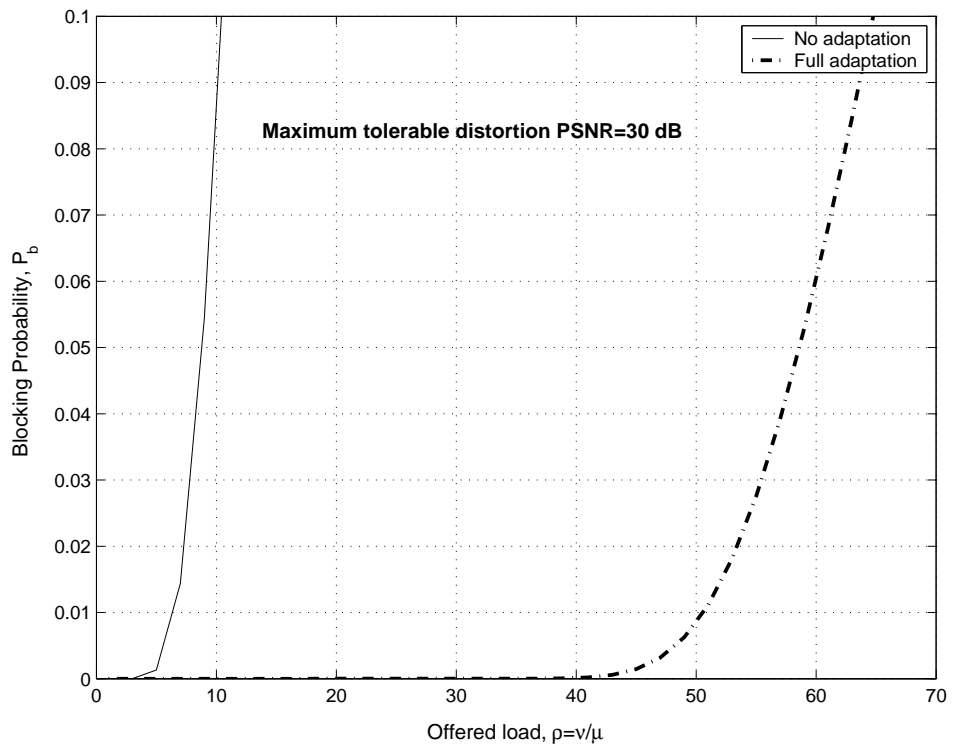


Figure 3.15: Blocking probability as a function of the offered load.

in the uplink of a CDMA network carrying real-time multimedia calls. From this study we have developed several key contributions. First, we showed that the problem (allocation of spreading factor, source coding rate and channel coding rate to minimize average distortion) could be considered as the optimal source-controlled statistical multiplexing in multimedia CDMA. In this case, an statistical multiplexer needs to perform resource allocation so as to assign an *equivalent bandwidth*, which depends on target SINR and transmit bit rate, among calls in such a way that average distortion is minimized. This viewpoint is a powerful abstraction to the studied problem and present many applications. Based on this, we then presented two solutions to the source-controlled statistical multiplexing problem and we showed them to be optimal. One of the solution keeps the target SINR (or equivalently, the channel coding rate) fixed and only transmit bit rate is changed through the source encoder rate adaptation. The other solution allows both the transmit rate and the target SINR to be adapted.

Also, we solved the practical implementation issue of communication overhead related with our centralized algorithm and we studied the behavior of our system when the number of calls changes dynamically. We recognized three possible operating conditions for our system: congestion-free, pseudo congestion and congestion. From here we showed that our system is able to extend operations beyond what has been normally considered an outage region at the cost of a smooth increase in distortion. We also advocated a call admission control based on rejecting new calls once a maximum number has been reached. This maximum number of calls is determined with the goal that the expected distortion per call when the number of calls is maximum does not exceed a tolerable limit.

Simulations results using MPEG4 FGS video show that the contributing solutions significantly outperform systems with no adaptation (which represents current imple-

mentations). In particular while the system with no adaptation reaches a break down point where the network is congested and no operation is possible, ours can support more than 3 times that number of calls while the quality degrades gracefully. Also, when the number of calls changes dynamically, our system can support a 50 % increase in offered load for the same level of expected distortion. The smooth increase in distortion also increases more than five times the offered load supported for the same level of blocking probability and maximum tolerable distortion.

Similarly as the observations made in chapter 2, we noticed that the mechanisms by which distortion increases in the statistical multiplexed systems is subjectively much more preferable than the mechanisms followed by the system with no adaptation. This does not only pertains to the fact that in the system with no adaptation the distortion increases rapidly beyond congestion, while the statistical multiplexed systems this increase is smooth. The mechanism to increase distortion in the statistical multiplexed system is preferable because it always meets the quality goal (a target BER in this case), limiting the channel errors to acceptable limits and increasing distortion by the subjectively more acceptable adaptation of the source encoders, which is a fully predictable mechanism. In the case of the system with no adaptation, there is no mechanism to meet the quality goal beyond congestion, thus the distortion increases due to the subjectively more annoying channel errors.

Chapter 4

Real-time and Data Traffic Integration

4.1 Introduction

In this chapter we apply the ideas and concepts developed in chapter 3 to study the integration of real-time calls with data traffic. The integration of these two traffic classes is based on considering the idea of allocating a portion of the total equivalent bandwidth to the real-time traffic and the remainder to data.

The chapter is divided in three main parts: the study of the data subsection, the study of the real-time subsection and the study of the integration of real-time and data traffic. In the study of the data subsection, we consider a system with a number of data calls, each associated with an infinite-size waiting queue and with packet retransmission when received in error. The main contribution of this part is the study of the influence of the equivalent bandwidth assignment on the data subsection performance. The main observation here is that small absolute changes in equivalent bandwidth assignment to the data section generate important changes in the data subsection performance. In the case of the real-time subsection study, the main focus is on the relation between distortion and total equivalent bandwidth. The main observation here is that distortion does not change much when modifying the portion of equivalent bandwidth assigned to

the real-time subsection. The analysis in this part revisits the study in chapter 2, to study resource allocation for uniform sources when the processing could change for each call. Also, we derive formulation that justifies our approximation of the relation between target SINR and source coding (or equivalently channel coding rate with the setup in chapter 2) to be exponential. Finally, we argue that although generally the data traffic is considered to be not delay-limited, there are many practical cases where this is not the case, thus it is important to implement methods that resolve congestion and control delay. An integrated environment that services both real-time and data traffic, such as the one being studied, allows for a transient reduction in the equivalent bandwidth assigned to real-time calls so as to notably increase the traffic capacity for the data section at the cost of a small increase of distortion for the real-time calls. We finish this chapter by summarizing the main conclusions and contributions.

4.2 Real-time and Data Integration Through Equivalent Bandwidth Allocation

An important observation from the previous chapter is that the problem of resource allocation in CDMA that meets constraints on system stability and power amplifiers dynamic range could be viewed as statistical multiplexing a total equivalent bandwidth. Mathematically, $\sum_{i=1}^K \Psi_i = 1 - \epsilon = \Omega$, where K is the total number of calls. In our current formulation, this total equivalent bandwidth is $\Omega = 1 - \epsilon$. From (3.3), we note that the equivalent bandwidth assigned to each call depends on both the call's processing gain and target SINR.

A likely requirement for a practical system would be to be able to effectively integrate both real-time and non-real-time (data) traffic. This implies that resource sharing

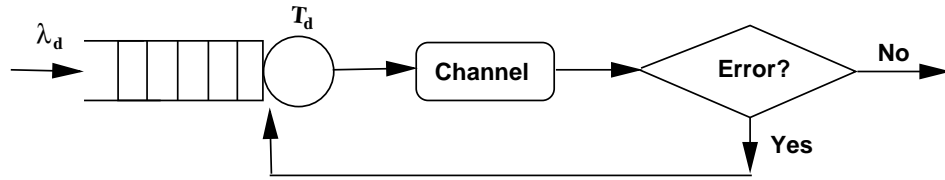


Figure 4.1: Block diagram of the data section call.

needs to be implemented across the two types of traffic. Following our approach, this means that the total equivalent bandwidth will need to be divided into a portion used by the real-time traffic and a portion used by the data traffic, i.e. we have $\Omega = \Omega_d + \Omega_r$, where Ω_d is the equivalent bandwidth assigned to the data section and Ω_r is the equivalent bandwidth assigned to the real-time section.

4.3 Data Subsection

We are going to focus next on the data subsection design. Our approach is based on [28], but in our case the goal is to assign the data section equivalent bandwidth among the data users when integrated with real-time calls. We assume that there are N_d data users present, all communicating over the uplink of a DS-CDMA system and all receiving the same service. This means that we will assume that the chip rate and processing gain is the same for all users. Figure 4.1 illustrates the operation of the data section. Each data user generate packets of fixed length L following a Poisson arrival process with average rate λ_d . These packets are error protected with a fixed-rate error control coder with rate π_r and placed in a FIFO buffer. The buffer content is then sent to the base station using a processing gain Z_d . If a packet is received with errors that cannot be corrected by the error protection scheme, a request for retransmission is sent back to the transmitter using a feedback channel that is assumed error free and with no delay.

If p_r is the data packet retransmission probability, the packet average service time (packet average transfer time) S_d is,

$$S_d = \frac{LZ_d}{W(1 - p_r)}. \quad (4.1)$$

As before, W is the system bandwidth. Note that S_d not only depends on the processing gain, but it also depends on the call SINR. This is because p_r depends on the SINR. If P_r is the total received power of real-time users, the target SINR for data calls, β_d , can be specified as,

$$\beta_d = \frac{Z_d P_d}{\sigma^2 + P_r + I P_d}, \quad (4.2)$$

where I is a random variable that represents the number of active interfering calls (not the number of calls). The probability that each data call is active is equal to the traffic load of that data call $\rho_d = \lambda_d S_d$, [35]. Therefore, the probability mass function of I is

$$P[I = j] = \binom{N_d - 1}{j} \rho_d^j (1 - \rho_d)^{N_d - 1 - j}. \quad (4.3)$$

Since the total received power of real-time users is

$$P_r = \frac{\sigma^2}{\epsilon} \sum_{i=1}^{N_r} \Psi_{r_i} = \frac{\sigma^2 \Omega_r}{\epsilon}, \quad (4.4)$$

where Ψ_{r_i} is the equivalent bandwidth for real-time users only, the target SINR for data calls can be written as

$$\beta_d = \frac{Z_d P_d}{\sigma^2 \left(1 + \frac{\Omega_r}{\epsilon}\right) + I P_d}. \quad (4.5)$$

The packet error probability not only depends on the target SINR β_d but it also depends on the number of active data calls I (by creating interference). If p_{bd} is the average bit error probability and $p_{bd}(I = j)$ is the bit error probability given that there are j active calls, the packet error probability is

$$p_r = 1 - [1 - p_{bd}]^{L\pi_r} = 1 - \left[1 - \sum_{j=0}^{N_d - 1} p_{bd}(I = j) P[I = j]\right]^{L\pi_r}. \quad (4.6)$$

Also, using (4.1), we have

$$\rho_d = \lambda_d S_d = \frac{\lambda_d L Z_d}{W(1 - p_r)}. \quad (4.7)$$

Both equations (4.7) and (4.6) form a system of two equations that need to be solved to find ρ_d and p_r . This solution is equivalent to finding the set of operating parameters Z_d and β_d . It was shown in [28] that the solutions to this system may represent three possible system states, named “phases”. Of the possible phases, only one corresponds to a stable system. Thus, in order to operate in the correct phase, it was shown in [28] that Z_d should be chosen larger than a threshold Z_d^* . This threshold can be computed from (4.6) and (4.7) by noting that Z_d^* corresponds to the situation when $\rho_d = 1$ both p_r from (4.6) and (4.7) are the same. This means that Z_d^* can be found by solving for

$$1 - [1 - p_{bd}(I = N_d - 1)]^{L\pi} = \frac{\lambda_d L Z_d^*}{W}. \quad (4.8)$$

Note that $p_{bd}(I = N_d - 1)$, the bit error probability when there are $N_d - 1$ interfering users, depends on the target SINR and, thus, depends implicitly on the processing gain. Due to the system stability and power amplifiers dynamic range constraints already discussed, the processing gain assigned to active calls is related to the I' active calls and the data subsection equivalent bandwidth Ω_d by,

$$\frac{I'}{1 + Z_d/\beta_d} = \Omega_d. \quad (4.9)$$

Since $I' = I + 1$, the target SINR that determines $p_{bd}(I = N_d - 1)$ corresponds to the case when $I' = N_d$. This target SINR should be related to the processing gain by,

$$\beta_d = \frac{Z_d \Omega_d}{N_d - \Omega_d}. \quad (4.10)$$

In general there might be at most three solutions to (4.8). Of these, the optimal choice to minimize delay is to pick the $Z_d > Z_d^*$ equal to $\lceil \hat{Z}_d^* \rceil$, where \hat{Z}_d^* is the smallest of the solutions to (4.8) that satisfies $1 < Z_d^* \leq \lfloor W/(\lambda_d L) \rfloor$.

Once the assignment for processing gain has been computed, the average data packet delay is given by $D_d = W_d + S_d$, where W_d is the average waiting time in queue. W_d can be found using the theory for M/G/1 systems. From [35], with Δ_D being the packet transfer time, W_d is given by

$$W_d = \frac{\lambda_d E[\Delta_d^2]}{2(1 - \rho_d)}. \quad (4.11)$$

Δ_D is a random variable with probability mass function equal to $P[\Delta_D = jT_d] = p_r^{j-1}(1 - p_r)$, where $T_d = LZ_d/W$ is the single packet transmission time. Then

$$E[\Delta_d^2] = \frac{T_d^2(1 + p_r)}{(1 - p_r)^2}. \quad (4.12)$$

Using this result we have

$$\begin{aligned} D_d &= \frac{\lambda_d T_d^2 (1 + p_r)}{2(1 - p_r)^2 (1 - \rho_d)} + \frac{LZ_d}{W(1 - p_r)} \\ &= \frac{LZ_d (2 - \lambda_d LZ_d/W)}{2W(1 - p_r - \lambda_d LZ_d/W)} \end{aligned} \quad (4.13)$$

4.4 Influence of Equivalent Bandwidth Assignment on the Data Subsection

In a typical setup where data is integrated with real-time calls, the data calls will be assigned the portion of the total equivalent bandwidth left unused by real-time calls. Also, it is possible that a minimum portion of the total equivalent bandwidth is reserved for data traffic. It is of interest to use the equations in the previous section, (4.1)-(4.13) to study the performance of the data subsection as its total equivalent bandwidth, Ω_r , is changed.

We first consider the case where the number of data calls is fixed. For 30 data calls ($N_d = 30$), Figure 4.2 shows the average data packet delay from (4.13) as a function

of the per-call average packet arrival rate as the data section total equivalent bandwidth is changed from 0.1 to 0.9. For each data call, the packet size was kept constant with $L = 2400$ bits. For channel error control coding we used the rate $1/2$, memory 8, convolutional encoder used in the IS-95 standard, [40]. The generator functions for this coder are 753 (octal) and 561 (octal). Also, we choose a system bandwidth $W = 40\text{MHz}$ (same as in chapter 3) with the processing gain taking values between 1 and 1024. Figure 4.2 shows the expected behavior, where the delay increases with decreasing data section total equivalent bandwidth. Also, it is clear from the figure that there is a maximum average packet arrival rate that the system can support before it becomes unstable and delay becomes infinite.

Figure 4.3 shows the optimum processing gain, computed as described in the previous section, as a function of the per-call average packet arrival rate. The setup for this figure is the same as for Figure 4.2. We can see that for the larger data section total equivalent bandwidth, the assignment does not change much. As the data section total equivalent bandwidth becomes smaller the optimum processing gain increases with the per-call average packet arrival rate. This means that the solution for optimum processing gain that minimizes average delay tends to do so by increasing the target SINR to reduce the probability of retransmission rather than by increasing the transmit bit rate. Also, we can see that, for any fixed per-call average packet arrival rate, processing gain increases with the decrease in data section total equivalent bandwidth. Note that the change in processing gain becomes larger as the data section total equivalent bandwidth becomes smaller. Here, again, the implication is that the solution for optimum processing gain that minimizes average delay tends to do so by increasing the target SINR to reduce the probability of retransmission rather than by increasing the transmit bit rate.

As noted in the discussion of the results in Figure 4.2, there is a maximum average

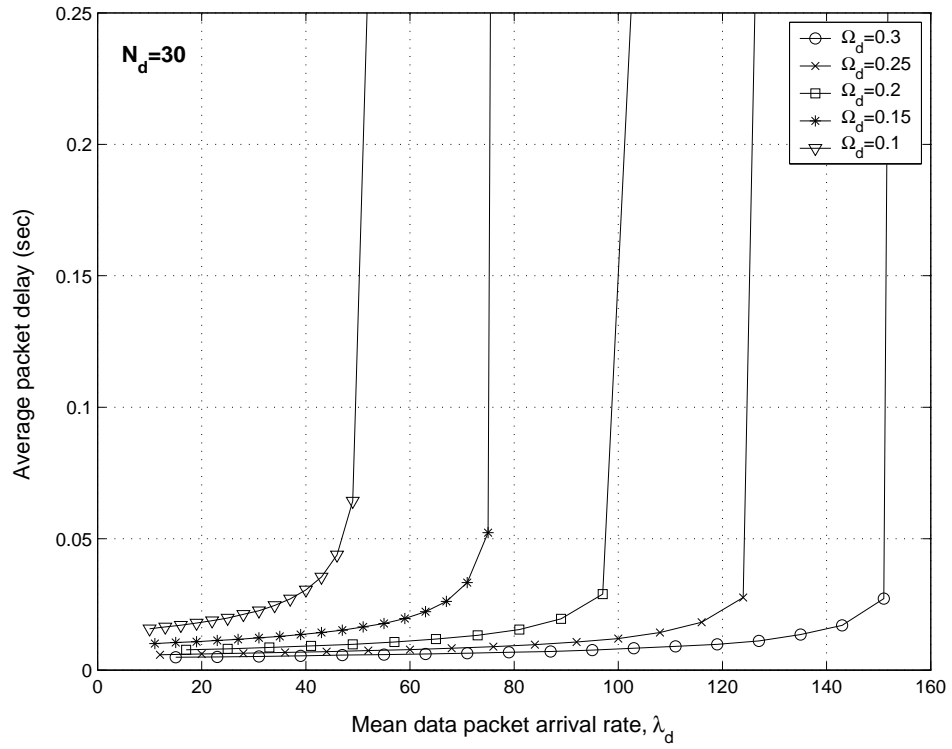


Figure 4.2: Average data packet delay as a function of average packet arrival rate for a single data call. The total number of data calls was assumed $N_d = 30$.

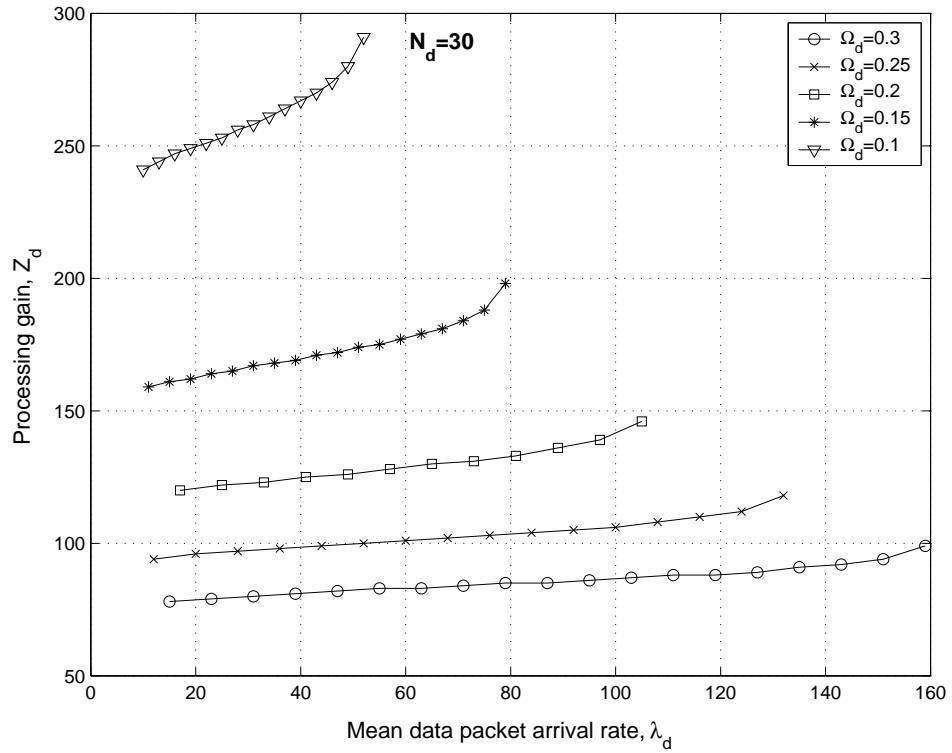


Figure 4.3: Optimum processing gain as a function of average packet arrival rate for a single data call. The total number of data calls was assumed $N_d = 30$.

packet arrival rate that the system can support before it becomes unstable. This maximum average packet arrival rate changes with the data section total equivalent bandwidth. With the same setup as Figure 4.2 and considering three different number of data calls ($N_d = 10$, $N_d = 30$ and $N_d = 50$), Figure 4.4 shows the maximum average packet arrival rate per data call as a function of the total equivalent bandwidth assigned to the data section. Interestingly, the figure shows that the relation is practically linear, independently of the number of data calls. Also, the figure shows that the slope of this linear relation decreases with the growth in number of data calls. This is because, in terms of effect on the whole system, a change in the per-call packet arrival rate is multiplied by the number of calls present.

Next, we consider the case where the per-call average packet arrival rate is fixed and we let the number of data calls change. For an average arrival rate of 100 packets per second and per call, Figure 4.5 shows the average data packet delay as a function of the number of data calls. In this case, the observations related to performance are conceptually the same as for Figure 4.2, i.e. the delay increases with the number of data calls in the system and there is a maximum number of data calls that the system can support before it becomes unstable and delay becomes infinite. This maximum number of data calls depends on the data section total equivalent bandwidth.

Figure 4.6 shows the optimum processing gain as a function of the number of data calls. The setup for this figure is the same as for Figure 4.5. We can see that the optimum processing gain increases with the number of data calls. Also, for a fixed number of data calls, note that the processing gain increases with the decrease in data section total equivalent bandwidth.

As noted in the discussion of the results in Figure 4.5, there is a maximum number of data calls that the system can support before it becomes unstable. This maximum number

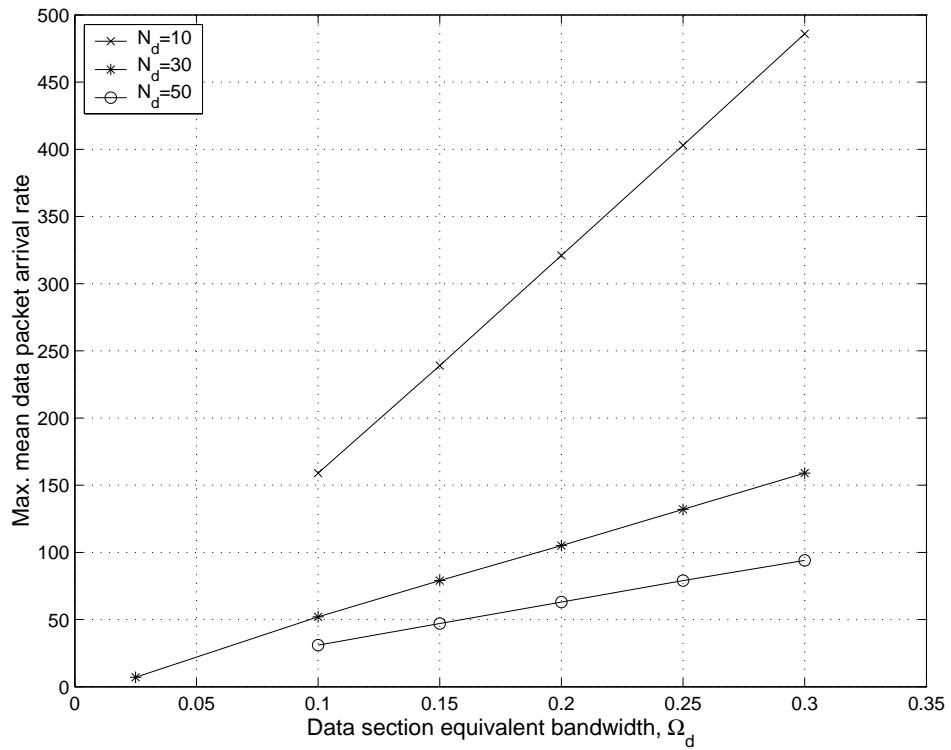


Figure 4.4: Maximum average packet arrival rate per data call as a function of the total equivalent bandwidth assigned to the data section for total number of data calls $N_d = 10$, $N_d = 30$ and $N_d = 50$.

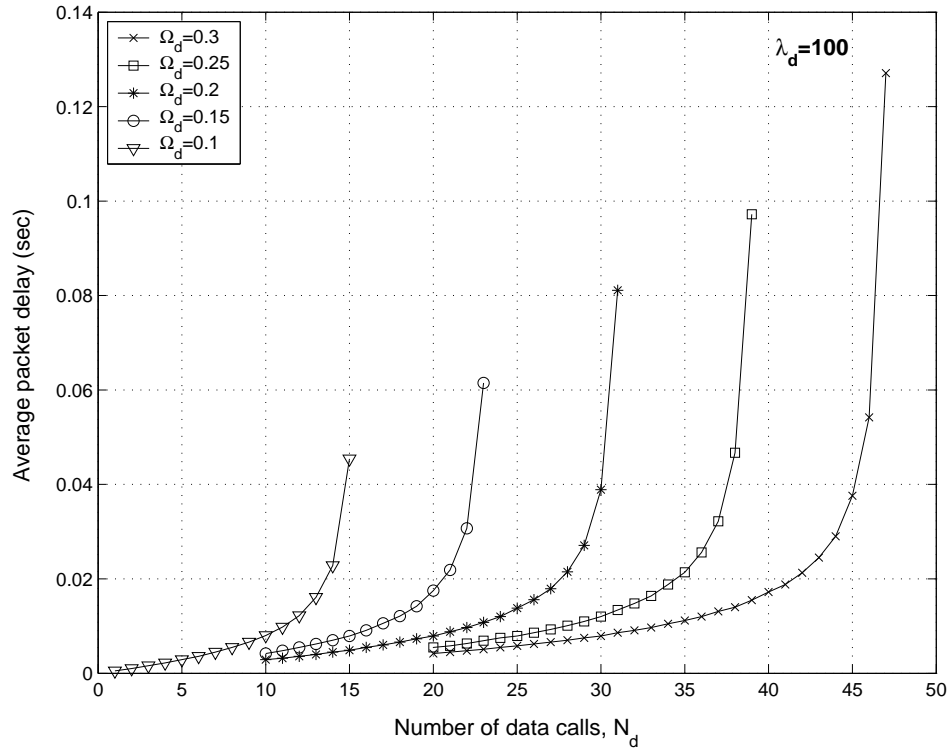


Figure 4.5: Average data packet delay as a function of the number of data calls. The average packet arrival rate per data call was assumed $\lambda_d = 100$.

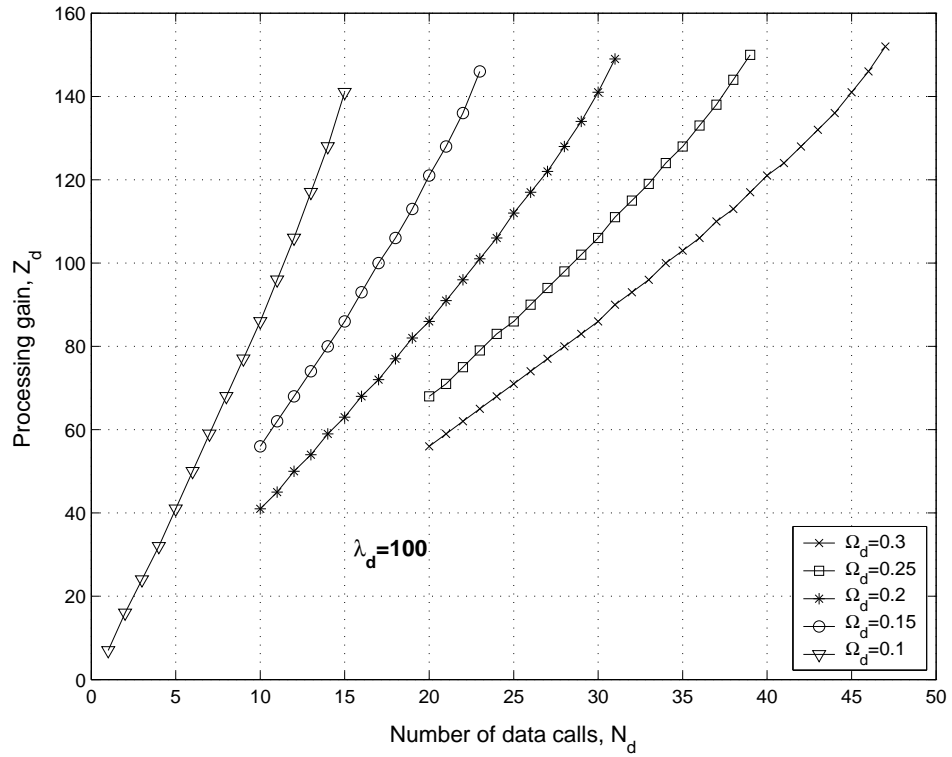


Figure 4.6: Optimum processing gain as a function of the number of data calls. The average packet arrival rate per data call was assumed $\lambda_d = 100$.

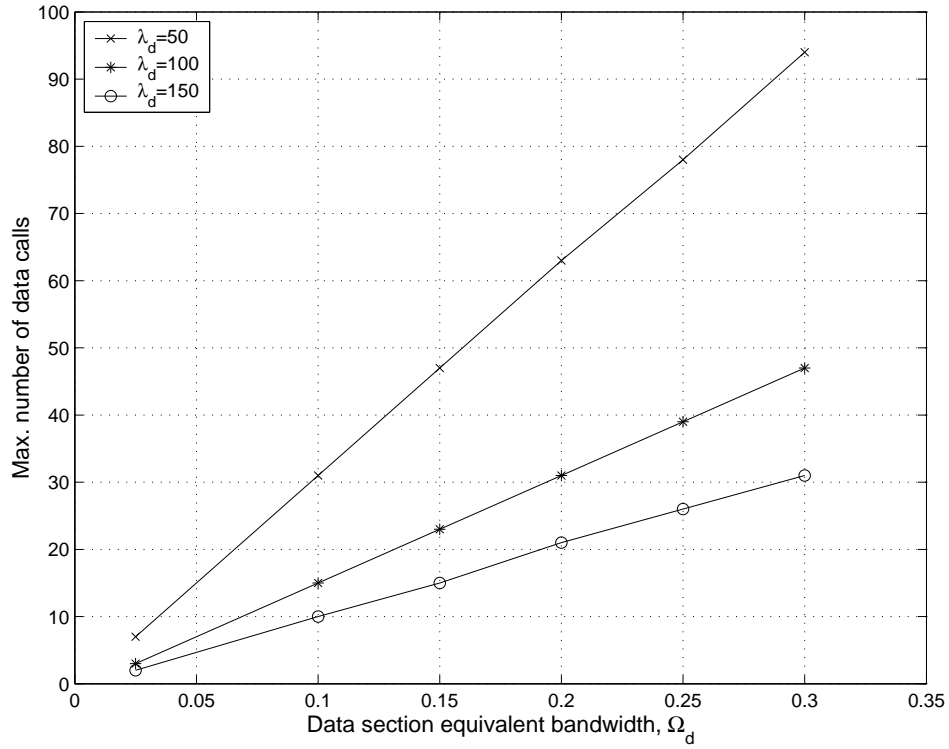


Figure 4.7: Maximum number of data calls that can be supported as a function of the total equivalent bandwidth assigned to the data section.

of calls changes with the data section total equivalent bandwidth. With the same setup as Figure 4.5 and considering three data packets arrival rates ($\lambda_d = 50$, $\lambda_d = 100$, $\lambda_d = 150$), Figure 4.7 shows the maximum number of data calls as a function of the total equivalent bandwidth assigned to the data section. Here again, the results shows the same interesting behavior as in Figure 4.4 with the relation between the maximum number of calls and the data section total equivalent bandwidth being practically linear. Also, similarly to Figure 4.4, the slope of this linear relation decreases with higher per-call packet arrival rates.

4.5 Real-time Traffic Subsection Dependence on Total Assigned Equivalent Bandwidth

In the previous chapters we have studied how to allocate resources, performing statistical multiplexing, to a number of calls carrying real-time traffic given the total equivalent bandwidth assigned to this traffic subsection. Next, we want to study the relation between distortion and total equivalent bandwidth.

Consider again the same system setup for real-time communication as the one described in chapter 3, i.e a system where the processing gain and both the source and the channel encoder can be externally controlled. Let's assume that all calls have the same distortion-rate performance. This is a condition necessary to maintain this study mathematical tractable. In this case we know from chapter 2 that all calls will be operating at the same operating mode.

Clearly, with this setup we have that the transmit bit rate at the input of the spreader is $x/\pi_r = W/Z_r$, where x is the source encoding rate, π_r is the channel coding rate, W is the system bandwidth (or, equivalently the chip rate in this case) and Z_r the real-time calls processing gain. Here, to minimize average distortion if the rate-distortion performance is convex and decreasing, the source encoding rate $x = \pi_r W/Z_r$ needs to be chosen as large as possible.

As discussed in chapter 3 the channel coding rate needs to be chosen so as to maintain the BER at a value small enough so that channel-induced distortion remain negligible. We want to find next an expression for this assignment. As in previous chapters, let's assume that the variable rate channel coder is implemented through an Rate-Compatible Punctured Convolutional (RCPC) code [19]. From [19], the bit error

probability after decoding can be upper bounded by

$$p_{RCPC_b} \leq \frac{1}{P_T} \sum_{d=d_{free}}^{\infty} c_d P_d, \quad (4.14)$$

where P_T is the puncturing period, d_{free} is the code's free distance, c_d is information error weight and P_d is the probability that the wrong path at a distance d is selected. When the SNR in the communication channel is large or, equivalently, operation is at a low BER, we may approximate (4.14) as

$$p_{RCPC_b} \leq \frac{1}{P_T} \sum_{d=d_{free}}^{\infty} c_d P_d \approx \frac{1}{P_T} c_{d_{free}} P_{d_{free}} \quad (4.15)$$

In the case of communication over an AWGN channel, we have [39],

$$P_d = Q\left(\sqrt{2d\beta_r}\right), \quad (4.16)$$

where β_r is the SNR of a real-time source, or SINR if the CDMA interference is approximated as a white Gaussian noise process. In this equation Q is the complementary error function:

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-z^2/2} dz < \frac{1}{\sqrt{4\pi}} e^{-x^2/2}. \quad (4.17)$$

Combining (4.15), (4.16) and (4.17) we get

$$p_{RCPC_b}(\beta_r, \pi_r) \leq \frac{1}{P_T} c_{d_{free}}(\pi_r) \frac{e^{-d_{free}(\pi_r)\beta_r}}{\sqrt{4\pi}}. \quad (4.18)$$

Defining $C_1 = 1/(P_T\sqrt{4\pi})$, we make the approximation,

$$p_{RCPC_b}(\beta_r, \pi_r) \approx C_1 c_{d_{free}}(\pi_r) e^{-d_{free}(\pi_r)\beta_r}. \quad (4.19)$$

Tables that describes values for $c_{d_{free}}$ and d_{free} for different RCPC codes and coding rates π_r can be found in [19] and [14]. From these tables we noticed that $c_{d_{free}}(\pi_r)$ does not present any definite pattern, other than taking values of roughly the same order of

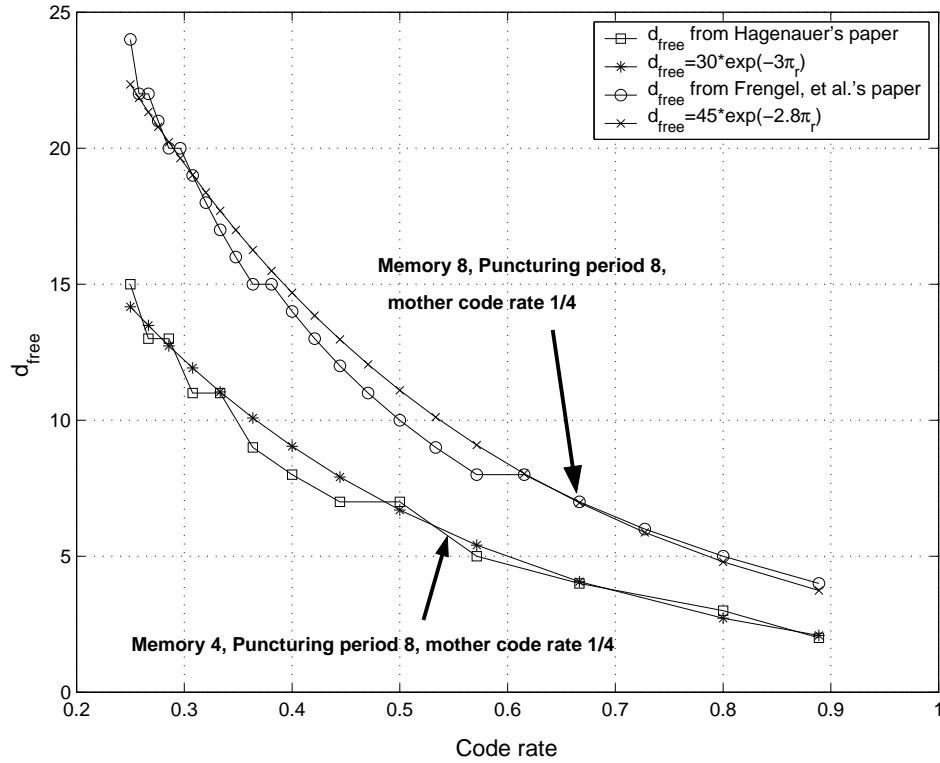


Figure 4.8: Free distance as a function of code rate for two different RCPC codes.

magnitude. Then we make the approximation $C_1 c_{d_{free}}(\pi_r) \approx C$, C being a constant. In Figure 4.8 we plot the free distance as a function of code rate for two different RCPC codes: a memory 4, puncturing period 8, mother code rate 1/4 from [19] and a memory 8, puncturing period 8, mother code rate 1/4 from [14]. The figure shows also an approximation for each of these curves. It is clear from the figure that the free distance as a function of code rate can be accurately represented by the expression

$$d_{free}(\pi_r) \approx C' e^{-A\pi_r}. \quad (4.20)$$

Thus, we have

$$P_{RCPC_b}(\beta_r, \pi_r) \approx C e^{-C' e^{-A\pi_r} \beta_r}. \quad (4.21)$$

If the target SINRs are designed so as to meet some constant target BER (as was the case in chapter 3) that we denote by B_T , we have that

$$B_T = C e^{-C' e^{-A\pi_r} \beta_r}. \quad (4.22)$$

From this equation we can see that

$$\beta_r = \frac{e^{A\pi_r}}{C'} \ln(C/B_T). \quad (4.23)$$

Note that this result is consistent with the observations and approximation for the target SINR discussed in chapter 2. Also,

$$\pi_r = \frac{1}{A} (C'' + \ln \beta_r), \quad (4.24)$$

where

$$C'' = \ln \left[\frac{C'}{\ln(C/B_T)} \right] = \ln \left[\frac{C'}{\ln \left(\frac{C_{dfree}}{P_T B_T \sqrt{4\pi}} \right)} \right]. \quad (4.25)$$

Therefore, using (4.24) the goal of choosing $x = \pi_r W / Z_r$ as large as possible to minimize average distortion now becomes choosing Z_r and β_r to maximize $x = W (C'' + \ln \beta_r) / (Z_r A)$. Z_r and β_r are constrained by a relation similar to (4.10). In this case we have that

$$\frac{Z_r}{\beta_r} \geq \frac{N_r - \Omega_r}{\Omega_r}. \quad (4.26)$$

Then, the optimization problem is

$$\max_{Z_r, \beta_r} \frac{W (C'' + \ln \beta_r)}{Z_r A} \quad \text{s.t.} \quad Z_r \geq \beta_r \frac{N_r - \Omega_r}{\Omega_r}. \quad (4.27)$$

Using Lagrange multipliers the solution to this optimization problem is

$$\beta_r^* = e^{1-C''}. \quad (4.28)$$

$$Z_r^* = e^{1-C''} \frac{N_r - \Omega_r}{\Omega_r}. \quad (4.29)$$

This solution is useful to study the relation between distortion and total equivalent bandwidth. If we assume a distortion-rate function of the form $D(x) = \alpha 2^{-kx}$, using with the optimum solution for β_r and Z_r , we will have that $x^* = W(C'' + \ln \beta_r^*) / (Z_r^* A)$ and the distortion-rate performance is

$$D^* = \alpha 2^{-\frac{kW\Omega_r e^{C''-1}}{A(N_r - \Omega_r)}}. \quad (4.30)$$

For the purpose of this analysis we will consider that the real-time source is a video sequence. We will assume that the distortion-rate performance of the video codec does not change between users or from frame to frame. This assumption is to some extent artificial but it becomes necessary to maintain mathematical tractability of the present study. Figure 4.9 shows the distortion-rate performance of the MPEG4-FGS codec averaged over all frames. The results shown are for the two sequences used in chapter 3: ‘Foreman’ and ‘Akiyo’. Also included in the same figure are the two approximations to the distortion-rate curve of the form $D(x) = \alpha 2^{-kx}$. Note for this setup the assumed distortion-rate performance matches well the one obtained from measurements.

We used the parameters from the approximations in Figure 4.9 to evaluate the distortion as a function of the real-time section total equivalent bandwidth, given by (4.30). The results are shown in Figure 4.10. In the figure we measure distortion as the mean squared error normalized to the minimum measured such value. The figure not only shows results for ‘Foreman’ and ‘Akiyo’, but it also shows result when using as parameters the average of those corresponding to the two sequences.

In a practical situation, the distortion-rate performance of the source codec changes both between users and from frame to frame. To study the impact that the change of the real-time traffic subsection total equivalent bandwidth has on the statistical multiplexer system studied in chapter 3 we measured the quality at the receiver (using PSNR in decibels) of a conversational video system as a function of the number of calls in the

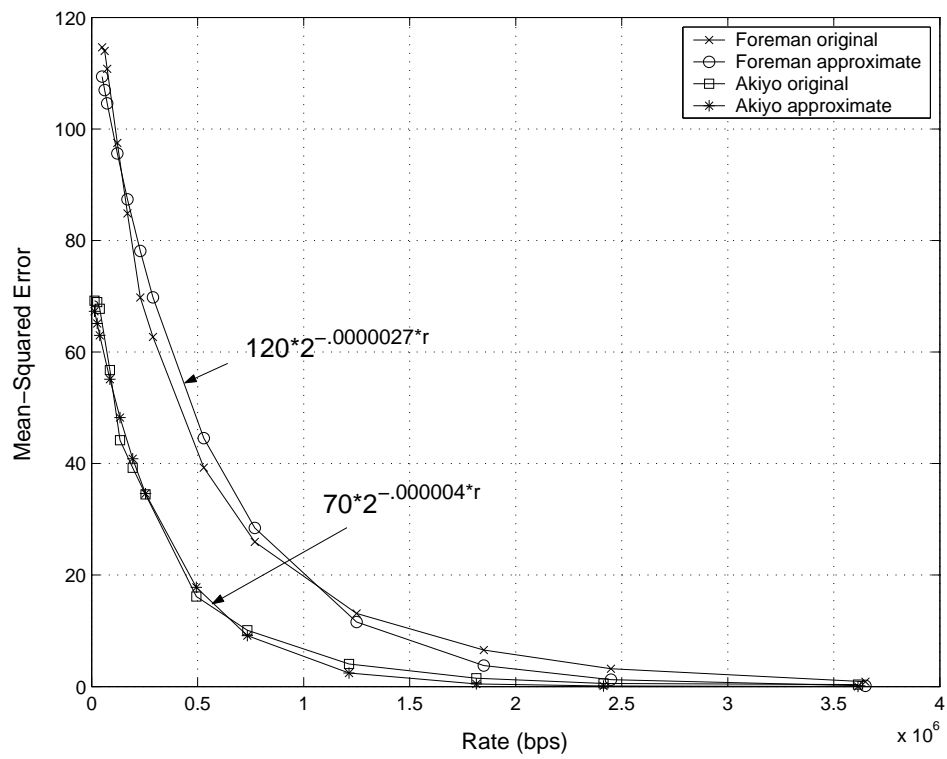


Figure 4.9: Averaged distortion-rate performance for two video sequences.

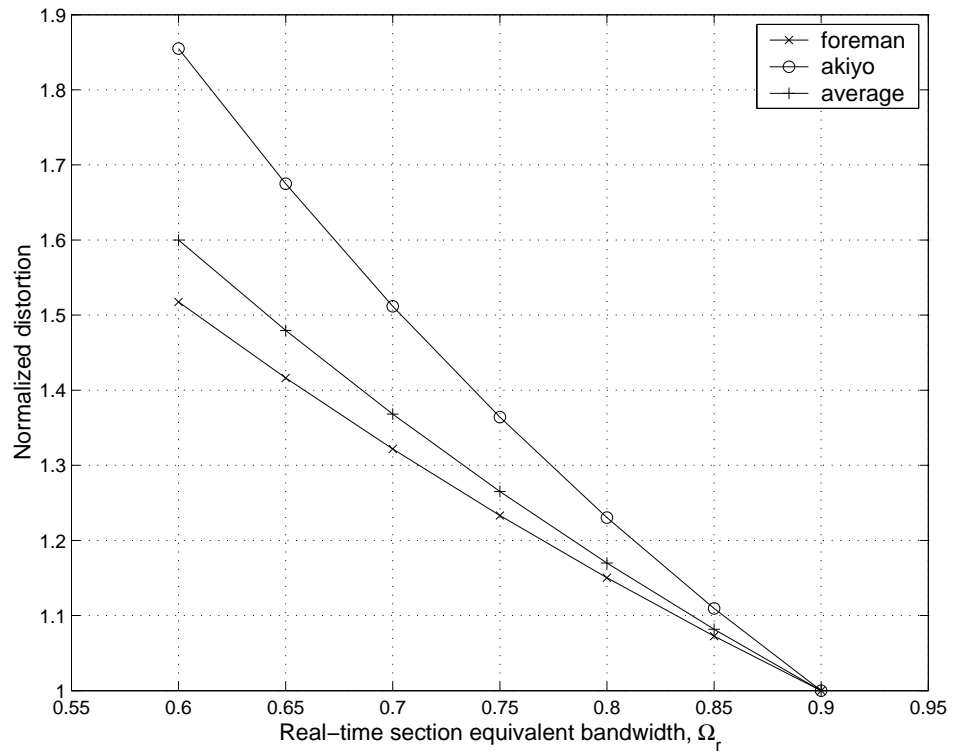


Figure 4.10: Normalized distortion as a function of the real-time traffic subsection total equivalent bandwidth.

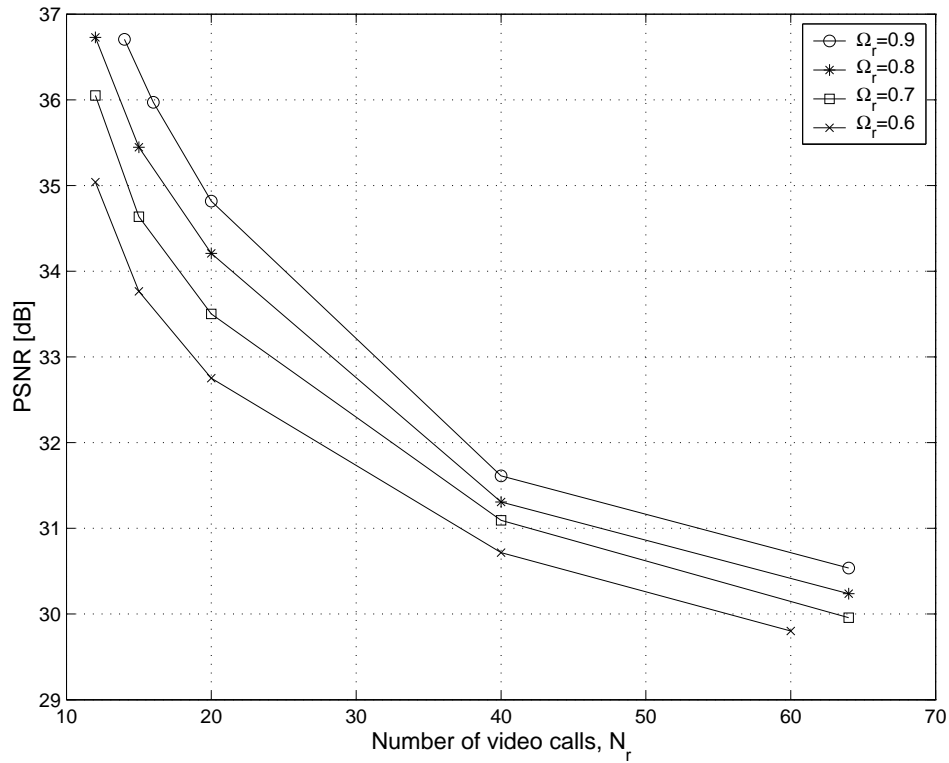


Figure 4.11: Quality at the receiver (using PSNR in dBs) of a conversational video system as a function of the number of calls in the system, using the real-time traffic subsection total equivalent bandwidth Ω_r as parameter.

system. In this measurement we used the real-time traffic subsection total equivalent bandwidth as a parameter. The setup for this measurement was the same as the simulations in chapter 3. Figure 4.11 shows these results. Clearly, the quality is better the more equivalent bandwidth is assigned to the real-time traffic subsection.

Figure 4.12 summarizes the results from Figures 4.10 and 4.11. The quality at the receiver is measured as the PSNR loss, in decibels, with respect to the best quality measured (that is the one with maximum total equivalent bandwidth). From the analytical we can see that ‘Foreman’ is the sequence that presents the largest increase in distortion

for a reduction in real-time section total equivalent bandwidth. In this case the loss is of approximately 2.7 dB in PSNR for a 33 % reduction in total equivalent bandwidth. Under the same conditions ‘Akiyo’ presents a loss of approximately 1.8 dB for a 33 % reduction in equivalent bandwidth. The results obtained from simulations show a smaller quality loss than any analytical results. This is due to the approximations and bounds (such as (4.15)) used to derive the analytical results but also due to the fact that the system in the simulation takes advantage of the video sequences changes in resource requirement from frame to frame to perform a statistical multiplexing that efficiently assigns resources. Overall, we can conclude that the analytical results used in Figure 4.12 work well to represent an approximation of the system behavior as an upper bound to the quality loss if the total equivalent bandwidth for the real-time subsection is reduced.

4.6 Real-time/Data Traffic Integrated Congestion Relief

Historically, the prevalent approach to integration of real-time traffic and data is to assign to the data traffic only those resources that the real-time traffic is not using. The justification for this approach is that data traffic is not delay sensitive, thus data packets can wait in a queue as long as necessary whenever the resources for the data section become scarce. In reality, there are many classes of data services that have different degrees of sensitivity to delay. In fact, even an activity such as downloading a file becomes sensitive if the delay is sufficiently large. This is not only due to the end-user needs and expectations but also because excessive delay could create resource starvation, especially at the mobile. As a justification to this, consider that the data traffic is carried using a transport protocol such as TCP. In this case, the transmitter will need to keep in

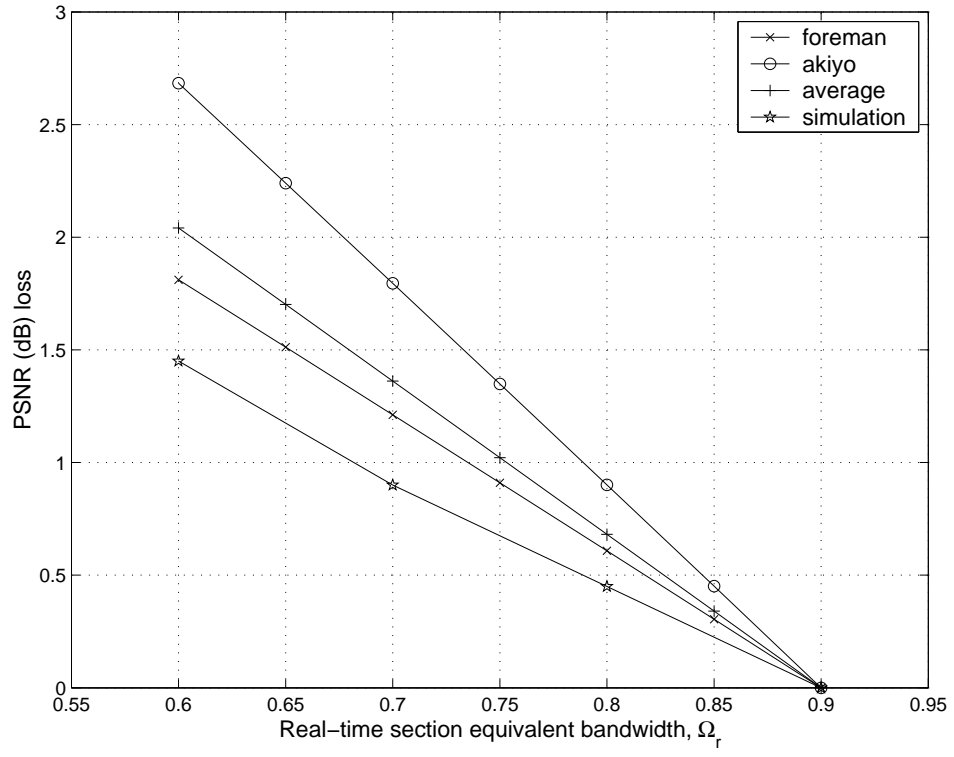


Figure 4.12: Quality loss (in PSNR loss) versus the data section total equivalent bandwidth.

a buffer all transmitted packet that has not been acknowledged yet. Therefore, the larger the delay, the more packets that need to be kept in the buffer. This may create buffer overruns since the mobiles are typically resource limited, including memory. Therefore, there is a clear need to control the delay of the data traffic and to implement mechanisms to reduce delay when necessary so as to avoid congestion.

The results just presented show that a small change in the total effective bandwidth assigned to the real-time subsection only generates a small degradation in the overall quality. From our previous study of the data subsection, we can see that, conceptually, the effect is the opposite, i.e. a small increase in the data subsection total equivalent bandwidth allows for a large increase in either the maximum number of data calls or the maximum per-call mean packet arrival rate (see Figures 4.4 and 4.7). Therefore, the results obtained so far in this chapter suggest that it is possible to implement an integrated congestion control mechanism for the real-time and data subsections. The main tools used by the integrated congestion control system are the statistical multiplexing of real-time sources and an adaptive distribution of equivalent bandwidth between the real-time and the data subsections. The general idea here is that whenever the data subsection is supporting a large number of calls or a high mean packet arrival rate the real-time subsection could undergo a transient and small degradation in quality by re-assigning more equivalent bandwidth from the real-time section to the data section. As soon as the traffic in the data section is alleviated, the equivalent bandwidth is reassigned to the real-time section. Therefore, there is a constant balance between the equivalent bandwidth assigned to the real-time section and its distortion.

To illustrate the interaction between real-time traffic and data when adapting the equivalent bandwidth assignment, we measured the relation between the maximum number of data calls or the maximum data mean packet arrival time versus real-time

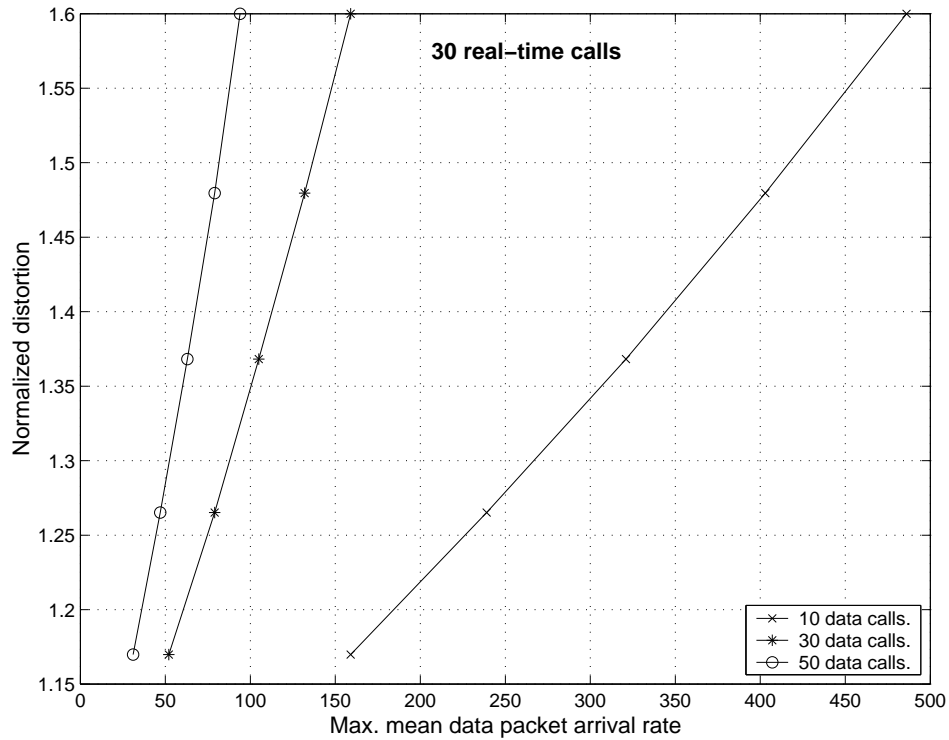


Figure 4.13: Normalized distortion of real time calls versus the maximum mean data packet arrival rate.

normalized distortion when $\Omega_r + \Omega_d = \Omega = 1 - \epsilon = 0.9$. In the measurement we used the same setup as in previous experiments in this same chapter. Figure 4.13 shows the normalized distortion of real time calls versus the maximum mean data packet arrival rate, assuming that there are 30 real-time calls and having the number of data calls as a parameter. We can see that the relation is approximately linear in all cases and that for a 15 % increase in distortion (0.6 dB reduction in PSNR) it is possible to increase the maximum mean data packet arrival rate approximately 60 % when there are 50 data calls, 70 % when there are 30 data calls and 50 % when there are 10 data calls.

Figure 4.14 shows the normalized distortion of real time calls versus the maximum number of data calls, assuming that there are 30 real-time calls and having the mean

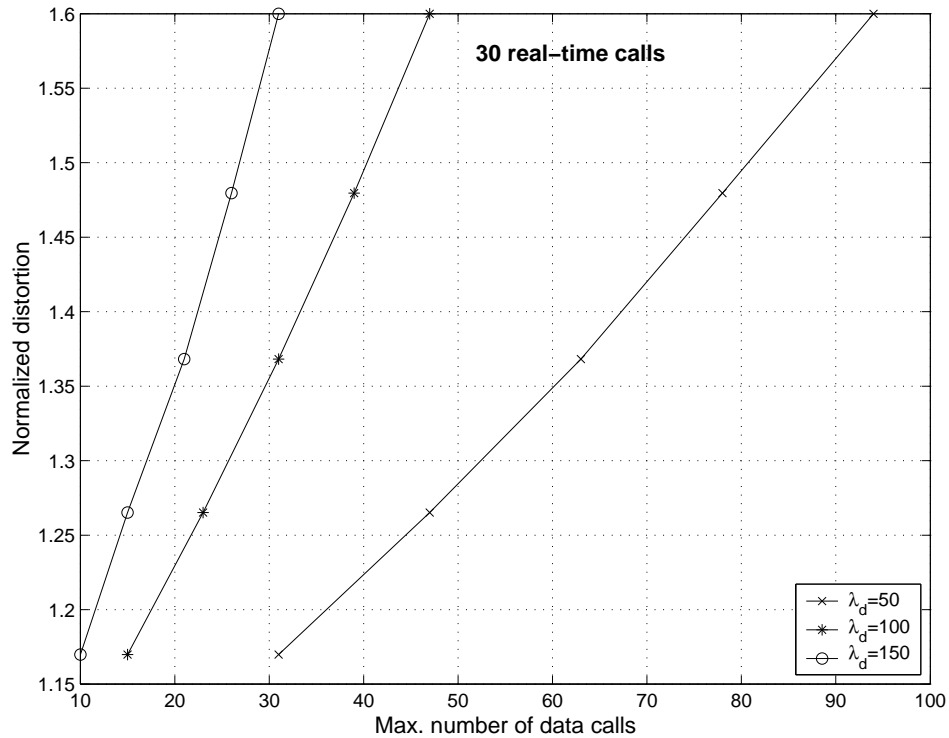


Figure 4.14: Normalized distortion of real time calls versus the maximum number of data calls.

per-call data packet arrival rate, λ_d , as a parameter. We can see that the relation is again approximately linear in all cases and that for a 15 % increase in distortion (0.6 dB reduction in PSNR) it is possible to increase the maximum mean data packet arrival rate approximately 60 % when $\lambda_d = 150$, 50 % when $\lambda_d = 100$ and 55 % when $\lambda_d = 50$.

Clearly, both Figures 4.13 and 4.14 show that a small increase in distortion for the real-time sources allow for an increase in the data subsection traffic capacity large enough that the congestion would be rapidly resolved.

We next consider the case when the number of real-time calls (assumed to be conversation video as in Chapter 3) changes dynamically over time. For this purpose we assume the same real-time traffic behavior as in Section 3.4, i.e. calls enter the cell

following a Poisson arrival process and that the random calls duration follow an exponential distribution. All other setting for the real-time calls were the same as in Chapter 3. We denote the real-time traffic offered load as ρ_r . For the data section we kept the number of data calls fixed at 50 and we assumed a mean packet arrival rate equal to $\lambda_d = 30$. Other parameters for the data subsection were the same as the ones used so far in this chapter. By performing Monte Carlo simulations on the real-time section, we measured the quality at the receiver for real time calls and the mean data packet delay as a function of the real-time subsection offered load and maximum allowed total equivalent bandwidth. In the cases where the real-time traffic did not used all its assigned equivalent bandwidth, the unused portion was allocated to the data section. Since we assumed real-time maximum allowed total equivalent bandwidth values from 0.6 to 0.8 while ϵ is still 0.9, the data subsection was always guaranteed a minimum equivalent bandwidth in the range 0.1 to 0.3. Figure 4.15 shows the result for the quality at the receiver for real time calls Figure 4.16 shows the mean data packet delay.

Table 4.1 tabulates together the results from Figure 4.15 and 4.16. The results confirm our previous observations when the number of real-time calls was assumed fixed, i.e. for relative small increases in distortion of the real-time traffic due to a reduction in its maximum allowed total equivalent bandwidth, the mean delay of the data section could be significantly reduced. Being the nature of the reduction in the real-time maximum allowed total equivalent bandwidth transient it is to expect that the overall effect on the perceptual quality should be small.

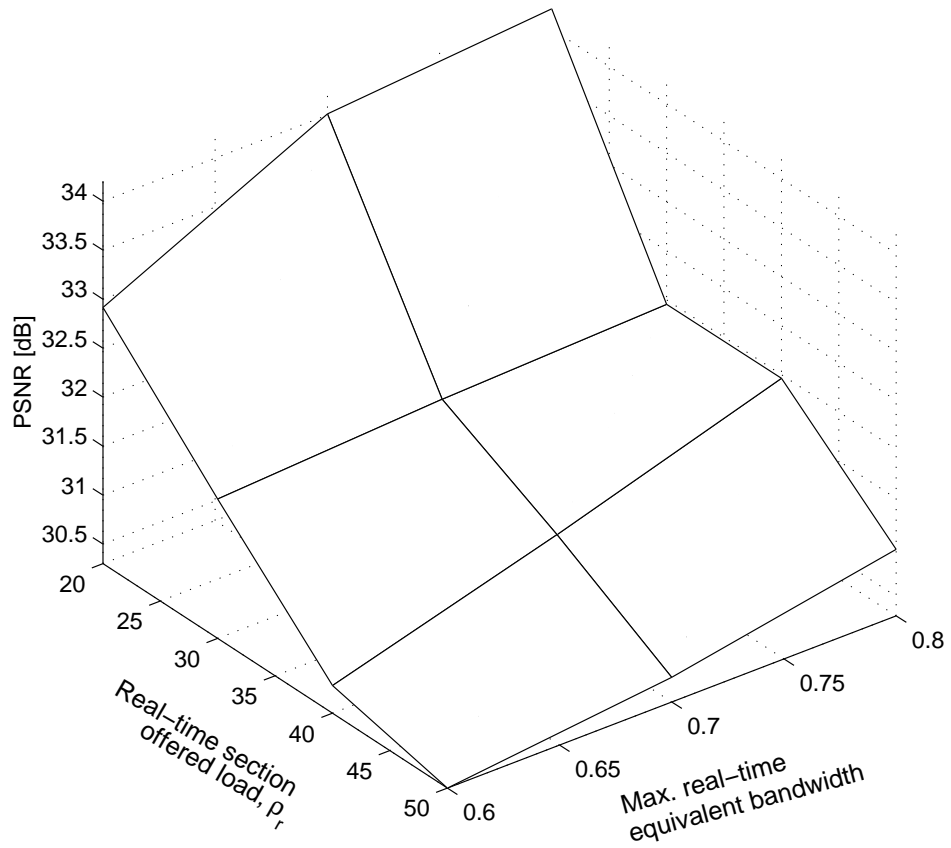


Figure 4.15: Quality at the receiver (using PSNR in dBs) of a conversational video system as a function of the real-time subsection offered load and maximum allowed total equivalent bandwidth.

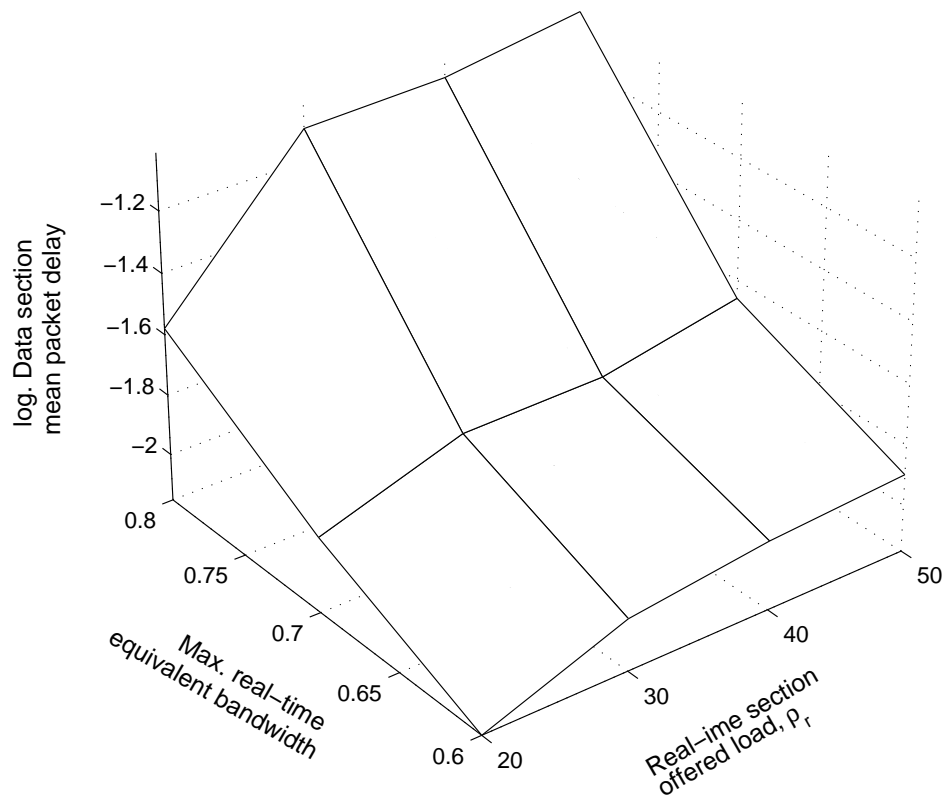


Figure 4.16: Data subsection mean packet delay as a function of the real-time subsection offered load and maximum allowed total equivalent bandwidth.

Table 4.1: Comparison of results from Figures 4.15 and 4.16

ρ_r	Change in real time eq. bandwidth	PSNR loss	Delay diff. (%)
50	0.8 to 0.6	0.68	87
50	0.8 to 0.7	0.43	73
40	0.8 to 0.6	1.38	86
40	0.8 to 0.7	0.72	78
30	0.8 to 0.6	0.23	87
30	0.8 to 0.7	0.09	78
20	0.8 to 0.6	1.29	73
20	0.8 to 0.7	0.19	53

4.7 Conclusions

In this chapter we have studied the integration of real-time and data calls in a CDMA system. The concept used in the integration is the ‘equivalent bandwidth’ discussed in the previous chapter. The study focus on a system where real-time and data calls are divided in two traffic subsection; both subsection are integrated by distribution the total equivalent bandwidth.

The chapter is divided in three main parts. In the first part we considered the data subsection. Our contribution here is the study on how the data section performance depends on its total assigned equivalent bandwidth. Here we noticed that small absolute changes in equivalent bandwidth assignment to the data section generate important changes in the data subsection performance.

In the second part of this chapter we considered the real-time subsection with a focus on studying the dependence of performance on the total equivalent bandwidth assigned to this section. We first revisited the study in chapter 2 to analyze resource allocation for uniform sources when the processing gain can also change. Also, we derive a formulation that justifies our approximation of the relation between target SINR and source coding (or equivalently channel coding rate with the setup in chapter 2) to be exponential. The main observation here is that distortion does not change much when modifying the portion of equivalent bandwidth assigned to the real-time subsection.

In the third part we argue that although generally the data traffic is considered to be not delay-limited, there are many practical cases where this is not the case, thus it is important to implement methods that resolve congestion and control delay. Our contribution here shows our proposed integrated environment allows for a transient reduction in the equivalent bandwidth assigned to real-time calls so as to notably increase the traffic capacity for the data section at the cost of a small increase of distortion for the real-time calls.

Chapter 5

Other Related Work

5.1 Introduction

In this chapter we conclude our study of cross-layer designs for multimedia CDMA by discussing other related works. These works focus on problems similar to the ones discussed so far, the most notable difference being that they considering the downlink. Despite the fact that in these problems the link is synchronous, inter-user interference is still an issue due to multipath propagation. Also, in these cases the constraining resources are the total transmit power at the base station and the spreading codes. It is important to remark here that the contributions in this chapter were developed in the context of collaborative work with members of the University of Maryland's Communication and Signal Processing Laboratory and were presented in [20, 33, 46].

The first part of this chapter will discuss a design that adapts a real-time source encoder to the channel and traffic conditions in the downlink of a CDMA system. The goal of the adaptation is to minimize the average distortion subject to constraints on the total power, each user distortion and a quality goal that limits the channel-induced distortion to a small proportion of the total.

The second part of this chapter will discuss an algorithm for resource allocation in

the downlink of multicode CDMA network. This means that among other resources, there is a pool of spreading codes that need to be allocated among real-time calls. The number of codes allocated to a call is directly proportional to its transmit bit rate. The design in this section is aimed at systems that use a real-time source encoder that generates a layered and embedded bit stream. This certainly applies to the MPEG4 FGS that is the center of the application example discussed in chapter 3, but is also applicable in other sources.

The chapter concludes with a summary of the main results and contributions.

5.2 Resource Allocation in the Downlink of CDMA by Real-Time Source Encoder Adaptation

Consider the downlink of a single cell CDMA system with N users. Fig. 5.1 shows the block diagram of the proposed cross layer design where it is possible to control the users' source encoding rates, channel coding rates and transmitted powers. A protocol located at the base station performs the allocation function for all calls.

As before, in the proposed system, the real time source encoder has the key property that the output rate can be externally controlled. We assume the source encoder have output rate $x_i = R_i r$ bits/s, where R_i is the variable channel coding rate and r is the transmit bit rate, assumed to be fixed. This means that, as was the case in chapter 2 source and channel rate allocation is determined so that any reduction in source encoding rate is matched by an increase in error protection in such a way that the bit rate at the input of the spreader remains constant. Also, this means that once either the source coding rate or the channel coding rate has been specified, the other is automatically determined. We use BPSK modulation with power control in the modulator. Also,

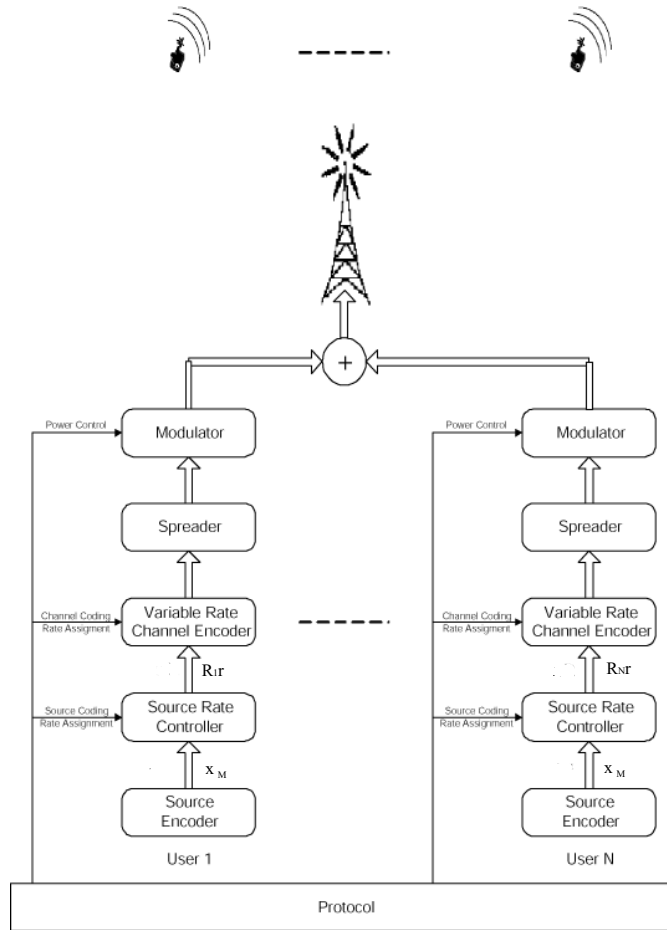


Figure 5.1: Proposed system block diagram

for simplicity, we assume that all the transmitted bits are equally important for error protection purposes.

As was the case for the designs in previous chapters, here again we will have as design condition a *quality goal*. In this case, this would be that channel induced errors would account for a small proportion of the overall end-to-end distortion. Thus, the design will be constrained by the condition of meeting a target SINR that achieves the desired small channel-induced distortion. Note again that the target SINR required to satisfy the quality goal is a function of the source rate. Thus, by reducing some or all

calls' source rate, it is possible to lower transmit power and interference at the cost of higher source encoding distortion but without increasing channel induced errors.

We assume the system is synchronous and each user is assigned a unique spreading code within each cell. Also, we assume a multipath channel. Because of this environment, the orthogonality between codes could not be maintained and each mobile user is subject to interferences from other users in the cell [2]. Under these conditions, the SINR of mobile i is given by:

$$\beta_i = \frac{W}{r} \frac{P_i G_i}{G_i \sum_{\substack{k=1 \\ k \neq i}}^N \theta_{ki} P_k + \sigma^2} \quad (5.1)$$

where W is the total bandwidth, P_i is the transmitted power from the base station to mobile i , G_i is the corresponding path loss, θ_{ki} is orthogonality factor and σ^2 is the background AWGN variance. W/r is the processing gain. θ_{ki} is the orthogonality factor between mobile k and mobile i and represents the fraction of the received downlink power that is converted by multipath into the intra-cell interference. We assume that all fading profiles are the same and $\theta = \theta_{ki}, \forall i, k$.

Let $f_i(x_i)$ be the distortion-rate performance function of the i^{th} user's source coder encoding at rate $x_i = R_i r$. Generally, for most well designed encoders, f_i is a convex and decreasing function. The minimum distortion occurs at maximum source rate x_M . Assuming $f_i(x_i) = \alpha 2^{-k x_i}$, in chapter 2, the source encoder distortion-rate performance function can be expressed as:

$$f_i = \delta 2^{2k(x_M - R_i r)} \quad (5.2)$$

where δ is the minimum distortion and k is a parameter depending on the encoder. Note again that this approximation can effectively represent the end-to-end behavior because the quality goal keeps the contribution of channel induced errors to the overall distortion

within negligible values. Define $D = 2^{2kx_M}/\delta$, the normalized distortion is given by:

$$D_i(R_i r) = \frac{f_i}{\delta} = D 2^{-2kR_i r}. \quad (5.3)$$

Due to the multipath environment, the system under consideration still exhibits inter-user interference. Therefore, as the system admits more users, the increase in interference will prevent allocating resources so that all users operate at their target (minimum) distortion. This may occur even when the base station uses all the available power during the transmission. Therefore, in this situation, some users will have to operate at an increased distortion. Thus, when designing the algorithm that allocates resources and decides operating parameters for each call, the problem is to decide which user will be configured to operate at non-minimum distortion and how these users will increase their distortions.

The basic effect of the adaptation that solves this allocation problem is to choose for each user a target SINR. From the discussion in the previous chapters, we have seen that each target SINR is associated to a distinct source encoding rate, or equivalently, a channel coding rate R_i . Furthermore, we have seen that the target SINR to achieve the quality goal can be approximated as a function of channel coding rate, when transmit rate is fixed, by

$$\beta_i = 2^{AR_i+B}. \quad (5.4)$$

Recall that A and B are parameters of the error control coding scheme.

Therefore, the adaptation goal would be to find the channel coding rate for each user that minimize the overall system distortion, under the constraint that each user's distortion is smaller than a maximum acceptable value and that the total transmitted power from the base station, $P_{sum} = \sum_{i=1}^N P_i$, does not exceed a maximum. The

problem is formulated as:

$$\min_{R_i} \sum_{i=1}^N D_i \quad (5.5)$$

$$\text{subject to } \begin{cases} \text{Distortion Range:} & 1 \leq D_i \leq D_{max}, \forall i, \\ \text{Transmitted Power:} & P_{sum} \leq P_{max}, \\ \text{Meet quality goal.} \end{cases}$$

In this problem formulation, P_{max} is the maximum transmit power available at the base station and D_{max} is the maximum acceptable distortion that we assume is the same for all users. Note that R_i is implicitly constrained by the combination of equation (5.3) with the above distortion range constraint.

The problem in (5.5) is a nonlinear nonconvex problem with possibly many local minima. Methods such as Lagrangian or nonlinear integer programming do not appear to yield a solution. Moreover, the computation complexity will grow exponential as the number of users increases. We next a suboptimal algorithmic solution developed by Z. Han. [33] that is fast and exhibits near-optimum performance.

5.2.1 Resource Allocation Algorithm

As noted, when the system is lightly crowded, each user could obtain a share of resources so as to operate at the minimum distortion and the necessary total transmitted power could still be less than the maximum available from the base station. As the system admits more calls, it may become not possible for all calls to operate at minimum distortion. In this operational condition, there is a need to have a graceful distortion control so that those users in bad channel condition or who introduce too much interference to others may sacrifice their performance slightly and in a controlled way so as to allow an optimal resource allocation. Note that the constraint on the channel-induced errors (the quality goal) will allow the increase in distortion to be smooth, controllable,

and predictable. This is because the dominant process in increasing distortion is the reduction in source encoding rate, thus the system behavior follows the rate-distortion curve. Also, this will keep the random and subjectively more annoying channel-induced distortion at a negligible value. Channel induced distortion is kept at a sufficiently small value by appropriately setting the rates and powers.

In order to decide the calls that will be operating at a larger distortion we consider the effect that a change in each user's distortion have on the total transmit power. For this, we derive a simple approximation for P_{sum} . Define

$$T_i = \frac{2^{AR_i+B_r}}{W} = \frac{P_i G_i}{G_i \sum_{k \neq i} \theta_{ki} P_k + \sigma^2}. \quad (5.6)$$

If the processing gain is large, i.e., W/r is large, T_i is small. We know that $\theta_{ki} < 1$, thus $\theta_{ki} T_i$ is also small. Therefore, we have

$$P_{sum} = \mathbf{1}^T [\mathbf{I} - \mathbf{F}]^{-1} \mathbf{u} \approx \mathbf{1}^T [\mathbf{I} + \mathbf{F}] \mathbf{u}, \quad (5.7)$$

where $\mathbf{1} = [1 \dots 1]^T$, $\mathbf{u} = [u_1, \dots, u_N]^T$ with $u_i = \sigma^2 T_i / G_i$, and

$$[\mathbf{F}]_{ji} = \begin{cases} 0 & \text{if } j = i, \\ \theta_{ji} T_i & \text{if } j \neq i. \end{cases}$$

Therefore,

$$P_{sum} \approx \sum_{i=1}^N \frac{\sigma^2 T_i}{G_i} + \sum_{i=1}^N \sum_{j \neq i}^N \frac{\sigma^2 \theta_{ji} T_i T_j}{G_j}, \quad (5.8)$$

We find the gradient of the overall transmitted power with respect to each user's distortion, g_i . The gradient can be written as a function of three differentials, as follows:

$$g_i = \frac{\partial P_{sum}}{\partial D_i} = \frac{\partial P_{sum}}{\partial T_i} \frac{\partial T_i}{\partial r_i} / \frac{\partial D_i}{\partial r_i}, \quad (5.9)$$

where

$$\frac{\partial P_{sum}}{\partial T_i} = \frac{\sigma^2}{G_i} + \sum_{j \neq i}^N \frac{\sigma^2 \theta_{ji} T_j}{G_j}, \quad (5.10)$$

$$\frac{\partial T_i}{\partial R_i} = \frac{Ar 2^{AR_i+B} \ln 2}{W}, \quad (5.11)$$

$$\frac{\partial D_i}{\partial R_i} = -2k D r 2^{-2k R_i r} \ln 2. \quad (5.12)$$

Therefore, the final gradient can be written as:

$$g_i = C 2^{(A+2kr)R_i} \left(\frac{1}{G_i} + \sum_{j \neq i}^N \frac{\theta_{ji} T_j}{G_j} \right) \quad (5.13)$$

where C is a negative constant. The absolute value of g_i is determined by three factors: current rates (the term before the parentheses), channel gain (the first term inside the parentheses), and interferences to others (the second term inside the parentheses).

If P_{max} is large enough for every user in the cell to operate at minimum distortion, the algorithm assigns $D_i = 1$ to all calls. If the system is lightly loaded, there might be some total transmit power left from the maximum available.

If P_{max} is not large enough for every user to operate at minimum distortion, the algorithm initially assigns $D_i = D_{max}$ to all calls. If there is not enough power for this allocation, i.e. if there is not enough power to satisfy the maximum distortion constraints, the system is in an outage condition. If allocation corresponding to $D_i = D_{max}$, $\forall i$ requires less power than the maximum available, the unused extra power could be used to reduce distortion. The user that will reduce its distortion is chosen by determining the gradient $\partial P_{sum} / \partial D_i$. If the absolute value of the gradient is small, it means that a reduction in distortion will have a small effect on the total consumed power. From (5.9), for this user, the current rates are low (i.e. the distortion is high), channel gain is good, or interferences to others are small. In other words, this user can reduce

Table 5.1: Downlink CDMA Resource Allocation Algorithm

<p>1. Initialization:</p> <p>If all calls can be configured to D_{min}, then allocate the powers and stop; else allocate D_{max} to all calls.</p> <p>If $P_{sum} > P_{max}$, report outage.</p>
<p>2. Repeat:</p> <ul style="list-style-type: none"> • Calculate g_i • Increase the rate of the user with smallest g_i to the next available discrete rate, unless the rate is 1/2 already. • If $P_{sum} > P_{max}$, return the previous rate allocation and break.
<p>3. Rate and Power Assignment.</p>

its end-to-end distortion while creating the smallest strain on the available resources. Consequently, this user is a good choice to reduce its distortion and the algorithm assigns a higher R_i to this user to let the distortion become small. After a user has been assigned a higher R_i , the operation involving estimated the gradient and increasing the rate to a call is repeated until reaching an allocation that does not allow increasing the total power beyond. Note that, in practice, there are only a finite number of possible values for R_i . Therefore, the assignment of a higher R_i is done in incremental discrete steps.

On the whole, the adaptive resource allocation algorithm is given in Table 5.1. As we have mentioned before, (5.5) is extremely difficult to solve by traditional methods in which the complexity grows fast with the number of users N . In the proposed algorithm, the complexity lies in computing the gradients in (5.13) and calculating the overall transmitted power in (5.26). So the complexity is $O(N^2)$ and can be easily implemented in practice.

5.2.2 Performance Bound

In order to evaluate the performance of the proposed algorithm in Table 5.1, we provide a computable performance upper bound. This bound might have a better performance than the optimal, but not viable to implement in practice, solution to the problem (5.5). If the proposed algorithm has a similar performance as the bound, we can conclude that the proposed algorithm is at least near optimal.

Assuming the transmit rate as fixed and the channel coding rate as a continuous real variable, the modified problem definition from (5.5) can be expressed as:

$$\begin{aligned} & \min_{R_i} \sum_{i=1}^N 2^{-2kR_i r_i} & (5.14) \\ & \text{subject to } \begin{cases} R_i^{min} \leq R_i \leq R_i^{max}, \forall i, \\ P_{sum} \leq P_{max}, \end{cases} \end{aligned}$$

where R_i is a real number. From (5.6) and (5.26), the power constraint is a nonlinear function of R_i . If we assume the channel coding as a continuous variable, the problem in (5.5) becomes the nonlinear constrained problem (5.14). To solve this case there exist useful nonlinear optimization methods. Therefore, we implement an algorithm that combines the barrier and Newton methods [5] that calculates the performance upper bound. The basic idea for the barrier method is to add barrier functions to the optimization goal such that the constrained optimization problem becomes an unconstrained optimization problem. The sum of optimization goal and barrier functions approaches infinity if the constraints are not satisfied. On the other hand, if the constraint is satisfied, the barrier function does not affect the optimization goal. The barrier function is commonly approximated by logarithmic functions [5]. In the case of the nonlinear constrained problem (5.14), the barrier function is given by:

$$I_{constraint} \approx \Phi_1 + \Phi_2 + \Phi_3, \quad (5.15)$$

where

$$\Phi_1 = \begin{cases} -\sum_{i=1}^N \ln(R_i - R_{min}), & R_i > R_{min}, \\ \infty, & \text{otherwise,} \end{cases} \quad (5.16)$$

$$\Phi_2 = \begin{cases} -\sum_{i=1}^N \ln(R_{max} - R_i), & R_{max} > R_i, \\ \infty, & \text{otherwise,} \end{cases} \quad (5.17)$$

and

$$\Phi_3 = \begin{cases} -\ln(P_{max} - P_{sum}), & P_{max} > P_{sum}, \\ \infty, & \text{otherwise.} \end{cases} \quad (5.18)$$

Φ_1 and Φ_2 correspond to the channel coding rate range and Φ_3 to the overall power.

As said, the barrier method approach solves the constrained optimization problem by solving a sequence of iterated unconstrained problems, each initialized by the results in the previous iteration. Rewrite (5.14) as:

$$\min_{R_i} f = \tilde{t} \sum_{i=1}^N 2^{-2kR_i r_i} + I_{constraint} \quad (5.19)$$

where \tilde{t} is a value that increases from iteration to iteration. Because, the barrier functions become more and more like the ideal barrier function as \tilde{t} increases, the solution becomes increasingly optimal. Within each iteration, the Newton method [5] is used to solve the unconstrained optimization problem. The complete algorithm is given in Table 5.2, where $\mathbf{R} = [R_1 \dots R_N]^T$, m is the iteration number for barrier method, e_p determines the accuracy of the proposed algorithm, t' is the optimal step for the Newton method, t_0 is the initial value for barrier function, whose value determines the convergence rate of the first iteration, and $C_t > 1$ is the constant that multiplies \tilde{t} in each iteration.

The performance bound determined by algorithm in Table 5.2 cannot be implemented in practice. This is because the rate is assumed to be continuous, which is

Table 5.2: Barrier Method for Performance Bound

1. Initial:

\mathbf{R} = any feasible value, $\tilde{t} = t_0 > 0$, $C_t > 1$, $e_p > 0$.

2. Repeat:

- Compute new \mathbf{R} by minimizing f , using

Newton Method:

1. Compute Newton step \mathbf{v}_{nt} and decrement λ^2 .

$$\mathbf{v}_{nt} = -\nabla^2 f^{-1} \nabla f$$

$$\lambda^2 = \nabla f^T \nabla^2 f^{-1} \nabla f$$

2. quit if λ^2 is stable.

3. Line search: compute step size t' by

backtracking line search.

4. Update: $\mathbf{R} = \mathbf{R} + t' * \mathbf{v}_{nt}$.

- if $m/\tilde{t} < e_p$, return \mathbf{R} .

- $\tilde{t} = C_t \tilde{t}$.

not true in practical channel coding coder. Because of this assumption, the algorithm will find a performance upper bound with a performance better than the optimal solution in (5.5). Also, the algorithm in Table 5.2 cannot be implemented in practice because its complexity is much higher than the proposed algorithm in Table 5.1. The complexity lies in that to find the solution, one iteration is needed for the Newton method and another iteration is needed for the barrier method. Yet another reason why the algorithm in Table 5.2 cannot be implemented in practice is that the problem in (5.14) is non-linear and non-convex with possibly many local optima. Multiple initializations or even annealing is necessary to find the global optimum. Thus, despite being hard to implement in practice, the algorithm in Table 5.2 can be used to judge the performance of the proposed fast implementable algorithm in Table 5.1. We will see in the simulation results in the next section the proposed algorithm has a similar performance as the performance bound. Consequently, the proposed algorithm is at least near optimal.

5.2.3 Performance Evaluation

The algorithm in Table 5.1 is applicable to any real-time source. To evaluate its performance we use an approach similar to the one in chapter 2. Then, we evaluated the performance of the proposed algorithm using speech at source. Similarly to the setup in chapter 2 we used the same eighteen speech sequences from the NIST speech corpus [47] and we used as source encoder the GSM AMR (Advance Multi-Rate) Narrow-band Speech Encoder [12]

To determine the end-to-end distortion, we used the same perceptually weighted log-spectral distortion measure discussed in chapter 2:

$$SD(\hat{A}(f), A(f)) = \sqrt{\int |W_B(f)|^2 \left| 10 \log \frac{|\hat{A}(f)|^2}{|A(f)|^2} \right|^2 df} \quad (5.20)$$

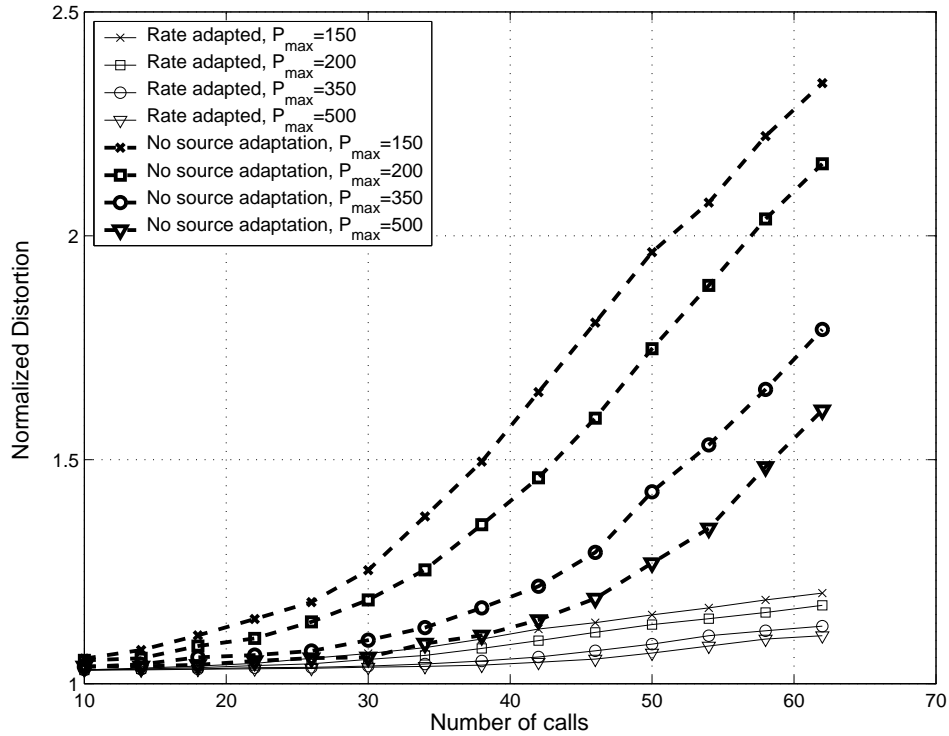


Figure 5.2: Normalized distortion vs. number of calls

where $A(f)$ and $\hat{A}(f)$ are the FFT-approximated spectra of the original and the reconstructed speech frames, respectively, and $W_B(f)$ is the subjective sensitivity weighting function [9]:

$$W_B(f) = \frac{1}{25 + 75(1 + 1.4(f/1000)^2)^{0.69}}. \quad (5.21)$$

For the proposed system, we used of the eight possible source encoding rates: 12.2, 10.2, 7.95, 7.4, 6.7, 5.9, 5.15 and 4.75 kbps, we used only the six highest ones. We also used a memory 4, puncturing period 8, mother code rate 1/4 RCPC code, [19], decoded with a soft Viterbi decoder. We assumed BPSK modulation and a channel affected by normalized Rayleigh fading (average power loss equal to 1) and normalized path loss (with propagation constants assumed equal to 1), with a path loss exponent equal to

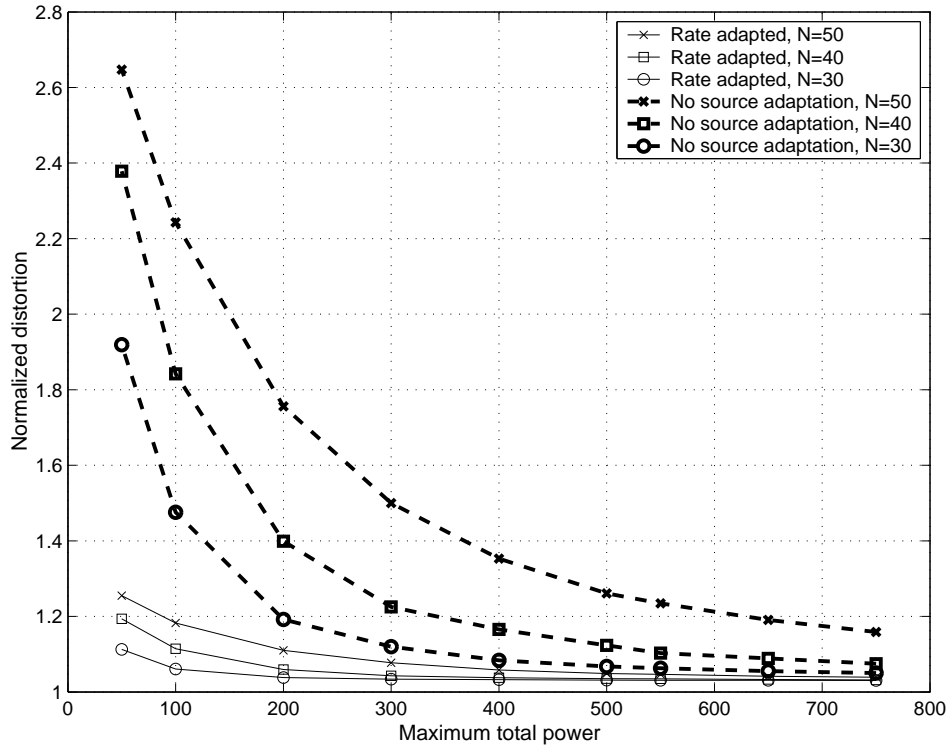


Figure 5.3: Normalized distortion vs. P_{max}

3 (typical of urban areas). The total bandwidth was 1.5616MHz. Users are randomly located in a cell with radius 500 m. Background noise level was assumed equal to 10^{-6} . $k = 3.3 \times 10^{-5}$. $R_{max} = 12.2$ kbps. Finally, we fixed the transmit rate at 24.4kbps and processing gain at 64.

To evaluate the proposed system, we analyzed its performance as the system becomes increasingly loaded. In Fig. 5.2, we show the normalized distortion versus the number of calls with the total transmit power as a parameter. The figure also includes, for comparison purposes, results for an equivalent CDMA system that shares the same configuration as the proposed scheme but is forced to operate without adaptation. For the case of this equivalent system, all calls operate at a source encoding rate equal to 12.2 Kbps and channel coding rate 1/2. From these results we can draw several con-

clusions. When the number of users is small, all the schemes show approximately the same performance. This is because there is enough power for every call to be configured for minimal distortion. When the number of users increases, the proposed scheme can reduce the normalized distortion relative to the traditional system setup. This is because the proposed scheme controls the distortion smoothly by adapting the source and channel coding rates. In particular, if example $P_{max} = 350$, the proposed system can support 30 users with 6 % less distortion, 40 with 12 % and 50 users with 37 % less distortion. When the ransmitted power is increased, the distortion is also reduced. In Fig. 5.3, we compared the normalized distortion as a function of the maximal available power for a fixed number of users in the system ($N = 30$, $N = 40$, and $N = 50$) that represents different network loading conditions. It shows the proposed system can deliver the same level of average end-to-end distortion by a much lower maximum transmitted power.

We also compared the performance of the proposed algorithm with the upper bound in Table 2 by measuring the relative difference. We define the relative difference as the average distortion of the proposed algorithm minus the average distortion obtained from the upper bound divided by the average distortion of the upper bound. In Fig. 5.4, we show the relative difference versus the number. To obtain the global optimization using the bound algorithm, we applied multiple initializations. Since the channel rate is assumed continuous for the algorithm in Table 5.2, the global optimum is always better than the one defined in the problem (5.5). Although the proposed algorithm is suboptimal, the difference in performance between the proposed algorithm and the upper bound is very small. This shows that the proposed algorithm is at least near optimal. Note that the performance of the proposed algorithm gets worse when the number of users increases. This can be justified by the fact that, in practice, the adaptation changes in the proposed algorithm are limited to a discrete number of possibilities; as the system

becomes increasingly loaded it becomes harder to find adaptations that would reduce distortion while not exceeding the constraints. This is not the case for the upper bound since the coding rates are assumed continuous.

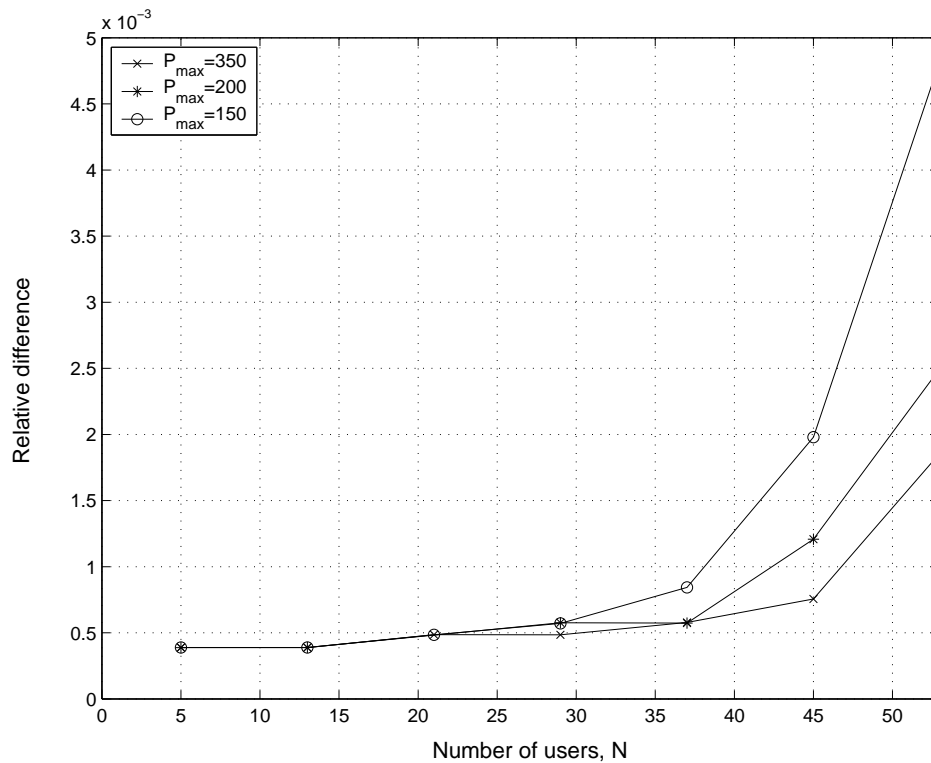


Figure 5.4: Evaluation of the proposed algorithm as compared to the performance bound.

5.3 Resource Allocation in the Downlink of Multicode CDMA for Layered and Embedded Real-Time Video

Consider the downlink of a single-cell Multicode CDMA (MC-CDMA) system with N users. In this system different transmit rates for each call are achieved by assigning a

number of spreading codes to the call. Each code is essentially used as a communication channel, thus the transmit bit rate of a given call is equal to the number of assigned codes times the transmit bit rate of each code. We assume that there are a total of C codes to be used.

Figure 5.5 shows a block diagram of the proposed design to transmit a layered and embedded real-time source, MPEG 4 FGS conversation video from now on, over the downlink of multicode CDMA. An allocation algorithm, implemented at the base station, allocates resources among each call with the goal of minimizing average distortion. These system resources are spreading codes and transmit power. The algorithm allocates resources based on rate-distortion (R-D) information received from each call. The algorithm assigns a variable number of spreading codes to each call according to its resource demand (based on the time-varying characteristic of each video sequence, as discussed in chapter 3) and channel condition. Also, according estimations of the downlink channel, the algorithm assigns channel coding rates and power allocations to the data transmitted over each code. In allocating resources, the design goal is to maintain good received quality, even when transmitting through a noisy channel with interference. As was the case of all designs presented in this work, we will have a *quality goal* to limit the effects of channel errors. In this case, the approach is similar to the one in chapter 3, since we will define the quality goal in terms of guaranteeing a target BER. Because the number of spreading codes and the total transmit power are limited, the challenge for the proposed algorithm is to efficiently allocate these resources while minimizing average distortion and meeting the quality goal.

As discussed in Section 5.2, because of the multipath environment, the orthogonality between codes could not be maintained and each mobile user is subject to interferences from other users in the cell. If the i^{th} code is assigned to user j , the received SINR when

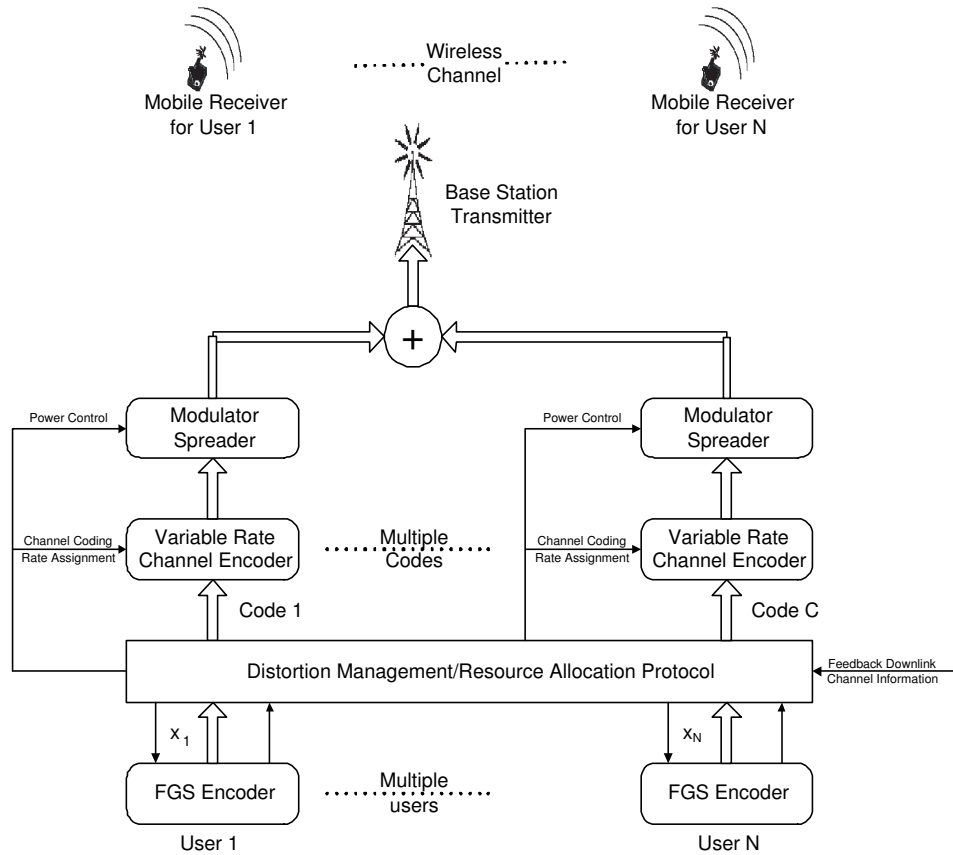


Figure 5.5: Block diagram for the proposed desing.

decoding data spreaded with this code is:

$$\beta_i = \frac{W}{r} \frac{P_i G_j}{G_j \sum_{k=1, k \neq i}^C \theta_{ki} P_k + \sigma^2} \quad (5.22)$$

where W is the total bandwidth and is fixed, r the transmit bit rate associated with a code, P_i the transmitted power from the base station for code i , θ_{ki} is the orthogonality factor between codes, G_j is the j^{th} user's path loss, and σ^2 the thermal noise level that is assumed to be the same at all mobile receivers. The ratio W/r is the processing gain.

To meet the quality goal, the received SINR should be no less than a target SINR, which is a function of channel coding rate. Here, again, we use the approximation for the target SINR as a function of channel coding rate

$$\beta_i = 2^{AR_i+B} \quad (5.23)$$

where β_i is the required targeted SINR. A and B are parameters of the error control coder, and R_i is the channel coding rate with range $[R_{min}, R_{max}]$.

In this muticode system, we denote a_{ij} as an indicator to specify whether the i^{th} code is assigned to user j . The maximum total power is P_{max} , and a valid allocation requires that each user's source data throughput, x_{T_j} , should be larger than the base layer bit rate x_j^0 (to guarantee a baseline quality) and smaller than the maximum source rate x_j^P .

The design optimization problem is, then,

$$\begin{aligned} & \min_{R_i, a_{ij}} \sum_{j=1}^N D_j \quad (5.24) \\ & \text{subject to } \left\{ \begin{array}{l} \sum_{j=1}^N a_{ij} \leq 1, a_{ij} \in \{0, 1\}, \forall i; \\ P_{sum} = \sum_{i=1}^C P_i \leq P_{max}; \\ x_j^0 \leq x_{T_j} = r \sum_{i=1}^C a_{ij} R_i \leq x_j^P, \forall j. \end{array} \right. \end{aligned}$$

The challenge in solve (5.24) lies on the power and the code constraint. Each call could reduce its operating distortion by increasing its allocated transmit power. Because different users experience different channel conditions, users will need different increases of power to have the same distortion reduction. However, the total transmit power at the base station is limited. Therefore, it is important to allocate power to each user efficiently. Moreover, as discussed in chapter 3, the bit rate necessary to transmit a video frame at a target quality level changes practically from one frame to the next. Then it is necessary to allocate codes based on each user's transmit bit rate needs and when doing so, the resulting effect will be different reductions of the average distortion for each call. Next, we present an algorithm to efficiently allocate the limited power and codes to reduce the overall distortion.

5.3.1 Distortion Management Algorithm

The algorithm is divided in two stages. The first stage allocates resources to guarantee delivery of the base layer data (so as to provide a baseline video quality for each user). The second stage allocates resources deciding the number of bits from the enhancement (FGS) layer to be delivered to reduce the average distortion. The goal for the second stage adjustment to each call total bit rate is to fully utilize the codes and power resources and to avoid exhausting only one resource first while having the other resource left underutilized, which leads to a local optimal solution. The proposed algorithm can overcome the problem mentioned above by keeping a balance between code and power allocation during the process of resource allocation.

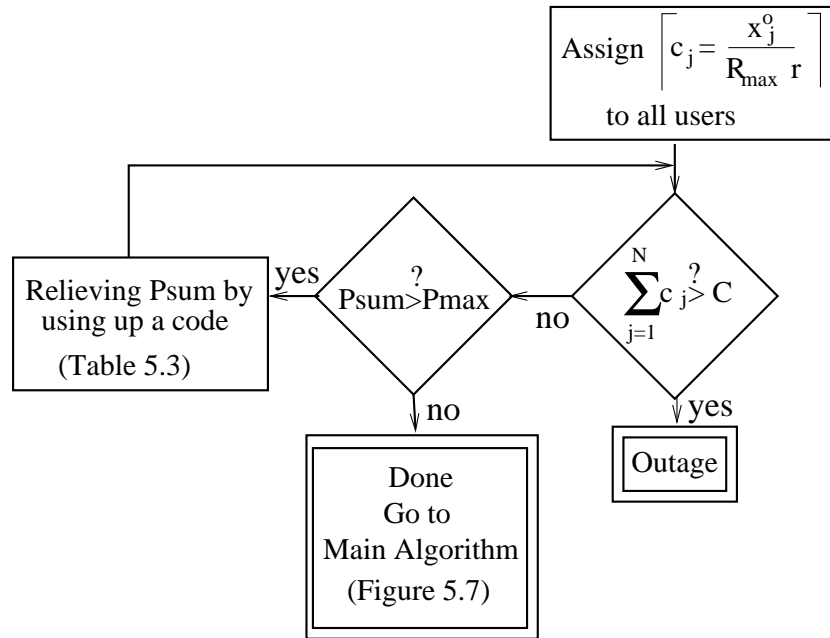


Figure 5.6: Base layer initialization algorithm.

Base Layer

Figure 5.6 shows the initialization procedure, where codes, channel coding rates, and power are assigned to each user so that all of them can transmit the base layer with bit rate x_j^0 , while the power constraint is satisfied. First, the algorithm allocates only the maximum channel coding rate R_{max} and computes the number of codes c_j necessary for each user to transmit the base layer. If there are no codes left, an outage is reported, indicating that there are too many users in the system and there are no resources even for accommodating the base layers only. If there are enough codes for the base layer, the algorithm evaluates whether the power constraint is met or not. If yes, the initialization stage is done and the algorithm proceeds to the second stage that allocates resource for the FGS layer. Otherwise, the algorithm reacts by “relieving P_{sum} ”. Relieving P_{sum} means that a previously unassigned spreading code is allocated to a user to reduce its

Table 5.3: P_{sum} Relieve Algorithm

1. For hypothesis $j = 1$ to N :
 - Assign a candidate code to user j .
 - Calculate the optimal R_i for all codes assigned to user j including the candidate code, such that P_{sum} is reduced most, while x_{T_j} is unchanged.
2. Pick the call with the largest reduced P_{sum} and assign it a real code.

transmit power P_{sum} . The unassigned code is allocated to the user that can reduce the power most when assigned an extra code while the distortion is kept. This operation is repeated until the power constraint is satisfied.

To derive the P_{sum} relieve algorithm, we first need to find an approximate expression for P_{sum} . Depending on which user a code is assigned to, we define

$$T_i = \begin{cases} 0, & \text{if code is not assigned;} \\ \frac{2^{AR_i+B_r}}{W} = \frac{P_i G_j}{G_j \sum_{k \neq i} \theta_{ki} P_k + \sigma^2}, & \text{for user } j. \end{cases} \quad (5.25)$$

Since the processing gain W/R is large and T_i is small, P_{sum} can be approximated as:

$$\begin{aligned} P_{sum} &= \mathbf{1}^T [\mathbf{I} - \mathbf{F}]^{-1} \mathbf{u} \approx \mathbf{1}^T [\mathbf{I} + \mathbf{F}] \mathbf{u} \\ &= \sum_{i=1}^C \frac{\sigma^2 T_i}{G_i} + \sum_{i=1}^C \sum_{k \neq i}^C \frac{\sigma^2 \theta_{ki} T_i T_k}{G_k}, \end{aligned} \quad (5.26)$$

where $\mathbf{1} = [1 \dots 1]^T$, $\mathbf{u} = [u_1, \dots, u_C]^T$ with $u_i = \sigma^2 T_i / G_i$, and $[\mathbf{F}]_{ij} = 0$ if $j = i$; $[\mathbf{F}]_{ij} = \theta_{ji} T_i$ if $j \neq i$. Note that this simplification is similar to (5.8), but now for the multicode scheme, each code, not a user, appear as an individual interfering call.

The P_{sum} relieve algorithm is shown in Table 5.25 Before assigning an actual code to a specific user, the algorithm makes N hypotheses. For the j^{th} hypothesis, it assigns a *candidate code* to the j^{th} user while keeping the settings for the other users unchanged. The j^{th} call will keep its source coding rate, x_j unchanged and will redistribute this bit rate evenly among the assigned codes (including the candidate code). Therefore, the channel coding rates corresponding to those codes assigned to user j are reduced which allows for a reduction in target SINR and total transmit power. In performing this operation the goal is to allocate the source coding rates and channel coding rates to the already assigned codes plus the candidate code such that P_{sum} is minimized and distortion is kept fixed. A simple solution for this problem is to use the water filling method. First, the candidate code is assigned a channel coding rate R_{max} , which means that the new throughput will be larger than x_j . Since T_i is a monotonic increasing function of R_i , P_{sum} can be reduced by searching for the code with the largest $|g_i^T = \frac{\partial P_{sum}}{\partial R_i}|$ and reducing its channel coding rate. The algorithm repeats the searching procedure until the throughput is equal to x_j . From all hypotheses, the user reducing P_{sum} the most is selected and assigned an actual spreading code.

FGS Layer

After initialization, the algorithm proceeds to the algorithm shown in Figure 5.7 that efficiently allocate resources to the enhancement layers so as to reduce average distortion. This algorithm starts by deciding whether all spreading codes have been used up. If this is the case, the algorithm uses the remaining transmit power (the difference between the current P_{sum} and P_{max} to reduce the distortion as described in Table 5.3.1. If there are unassigned spreading codes the allocation algorithm proceeds to one of two possible sub-algorithms that assigns a new code, one to reduce transmit power and the other to

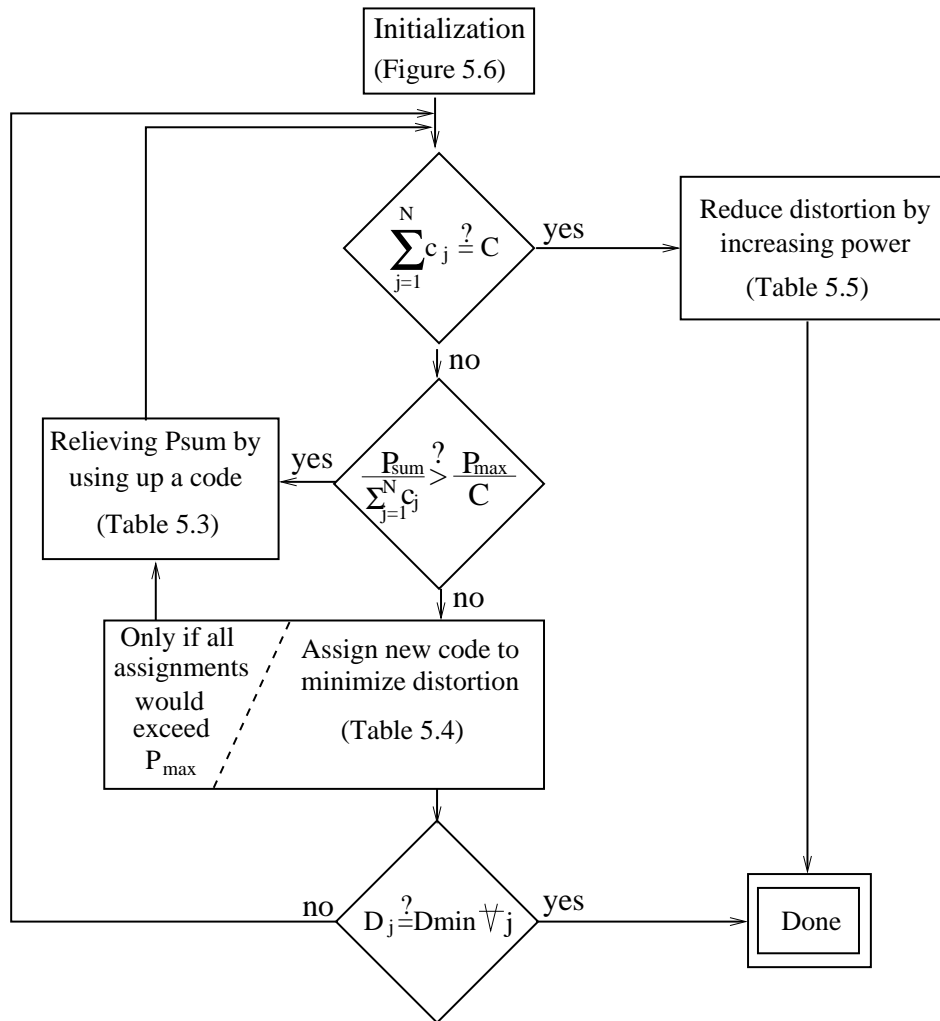


Figure 5.7: Resource allocation algorithm for enhancement (FGS) layer.

reduce distortion, depending on whether the system is power unbalanced. The whole algorithm is terminated under two conditions: First, all available codes are distributed and the power is within the acceptable range. This is the situation when the system is heavily loaded. Second, all the users have the minimal distortion. This is the lightly loaded situation.

The criterion to judge whether the system is power unbalanced is based on the average consumed power per assigned code. If the current P_{sum} over the total assigned codes $\sum_{j=1}^N c_j$ is greater than P_{max} over the total available number of codes C , the system is power unbalanced. Then, the algorithm in Table 5.3.1 is applied to reduce the power. Otherwise, the algorithm listed in 5.4 is used to reduce distortion.

In 5.4, the algorithm reduces the overall distortion by assigning one spreading code to a user at a time. Before an actual spreading code is assigned to a specific user, the algorithm makes N hypotheses. For the j^{th} hypothesis, a *candidate code* with associated channel coding rate R_{max} is assigned to the j^{th} user while keeping the settings of the other users unchanged. Then, the total power required for transmission, P_{sum} , and the source distortion reduction, are calculated using (5.26) and:

$$\Delta D_j = D_j(x_j) - D_j(x_j + rR_{max}), \quad (5.27)$$

respectively. In (5.27) x_j is the user's current source rate. If P_{sum} is smaller than P_{max} , this hypothesis is added into a candidate list. Then, the algorithm assigns an actual spreading code to the call among the candidate hypotheses list that can reduce the distortion by the largest magnitude. If the candidate list is empty, it means the average distortion cannot be further reduced without the corresponding required transmit power exceeding P_{max} . To achieve further distortion reduction in successive iterations, the power needs to be reduced using the algorithm in Table 5.3.1.

If all the spreading codes have been assigned and there is still some transmit power

Table 5.4: Code Assignment to Reduce Distortion

1. For hypothesis $j = 1$ to N :
 - Assign user j a candidate code, analyze $\Delta D_j, P_{sum}$
 - If $P_{sum} < P_{max}$, add hypothesis j to candidate list.
2. If there is no candidate user, do not assign the code and go to the algorithm in Table 5.3.1.
3. Among the candidates, choose the one with the largest ΔD_j and assign an actual spreading code to user j .

left, it is useful to reduce the distortion by increasing the transmit power. The algorithm is listed in Table 5.3.1. Here the algorithm makes C hypotheses, i.e. one for each spreading code. In hypothesis i , the algorithm checks whether the channel coding rate for the spreading code i , R_i , is less than R_{max} . If not, it checks the next hypothesis. Otherwise, it increases R_i by a discrete step ΔR_i while it keeps the settings of the rest $C-1$ codes unchanged. If this code belongs to the j^{th} user, the algorithm calculate the reduced distortion ΔD_j and the increase in total transmit power ΔP_{sum} . If the new total transmit power does not exceeds P_{max} this hypothesis is added to a candidate list. Among the candidates in the list, the algorithm picks the code with the largest $|\Delta D_j / \Delta P_{sum}|$ and set $R_i = R_i + \Delta R_i$. The above process is repeated until there is no transmit power left.

5.3.2 Performance Evaluation

We evaluated the performance of the proposed algorithm through simulations. The results shown here were obtained by G.-M. Su as part of our collaborative work [46].

The simulations are set up as follows. We assumed bandwidth equal to 7.5 MHz.

Table 5.5: Distortion Reduction by Increasing Power

1. For hypothesis $i = 1$ to C :
 - If R_i of code i is equal to R_{max} , do next hypothesis.
 - For code i , calculate the corresponding decrease in channel coding rate of one discrete step, ΔR_i .
 - Given ΔR_i , calculate ΔD_j , and ΔP_{sum} .
If $P_{sum} < P_{max}$, add hypothesis i to candidate list.
2. If no candidate left, exit. Otherwise, choose the code with the largest $|\Delta D_j / \Delta P_{sum}|$ and change the channel coding rate of the chosen code.
3. Empty candidate list. Go to step 1.

and spreading factor 64. We assumed a fading channel with path loss factor equal to 4 and delay profile typical of urban area. The noise power was assumed 10^{-9} Watts and maximum transmit power equal to 280 Watts. The mobiles are uniformly distributed within a cell at a distance from 20 m. to 700 m. For error protection we used the memory 4, puncturing period 8, and mother code rate 1/4 RCPC codes from [19]. Through simulations we found that in order to meet a target BER= 10^{-6} for the MPEG-4 FGS codec, the parameters A and B in (5.23) are $A = 4.4$ and $B = -1.4$.

As input sequence we concatenate 15 classic video sequences (*Akiyo*, *Carphone*, *Claire*, *Coastguard*, *Container*, *Foreman*, *Grandmother*, *Hall objects*, *Miss American*, *Mother and daughter*, *MPEG4 news*, *Salesman*, *Silent*, *Suzie*, and *Trevor*). To all sequences we applied a temporal down sampling factor equal to 2. This resulted in a video sequence with 2775 frames and video refresh rate of 15 frames per second. The base layer of the encoded video sequence was generated using an MPEG-4 encoder

with a fixed quantization step of 30 and a GOP pattern of 14 P frames between each I frame. All frames of the enhancement (FGS) layer have up to six bit planes. For each call we choose as input sequence a section from the main concatenated testing video sequence of 100 frames of length. For the i^{th} user this sequence corresponded to frames $173 \times (i-1) + 1$ to $173 \times (i-1) + 100$.

Figure 5.8 shows a simulation result for the convergence track of the total transmit power (solid line and scale on the right) and overall distortion $D_{sum} = \sum_{j=1}^N D_j$ (dashed-dotted line and scale on the right) as the spreading codes are gradually assigned using the proposed algorithm. After initialization (shown as point A), 17 codes have been assigned so as to accommodate the base layer of each user. The overall distortion (shown at point A') is large because only the base layer is to be transmitted so far. When the system is power unbalanced, i.e. the operating point is above the balanced resource allocation line (the situation at point A, for example), the algorithm uses the power relieve algorithm in Table 5.3.1 to reduce total transmit power while keeping the distortion fixed. When the system is not power unbalanced (such as the case at point B), the algorithm assigns new codes to reduce distortion (consequently, the required power is increased) using the algorithm in Table 5.4 until all the codes are used up. Finally, the algorithm uses the sub-algorithm in Table 5.3.1 to further reduce distortion until the transmit power has been use as much as possible (shown at point C).

To further evaluate performance we compare the proposed algorithm with the algorithm in Section 5.2, modified so that code assignment for the FGS layer follows a greedy approach. For each iteration, this greedy algorithm tries to assign a candidate code by calculating $|\Delta D_j / \Delta P_{sum}|$ for each call, and assigning the new code to the user with the largest such value. Figure 5.9 shows the PSNR results in a four-user system. The first three users receive video with a better quality using the proposed algorithm.

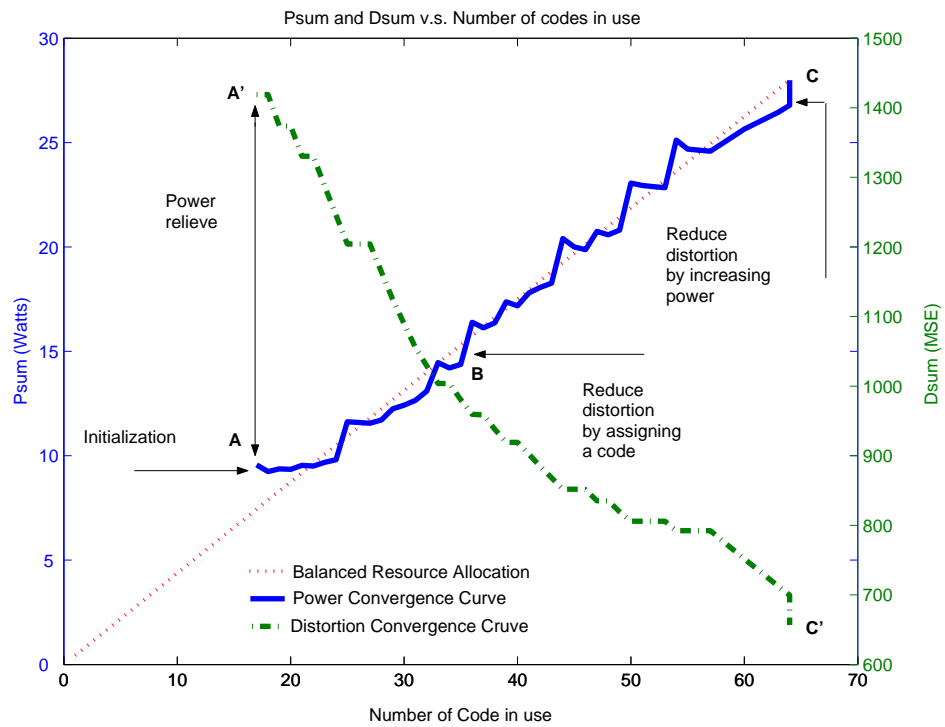


Figure 5.8: Power (solid line and scale on the left) and distortion (dashed-dotted line and scale on the right) as a function of the number of assigned codes

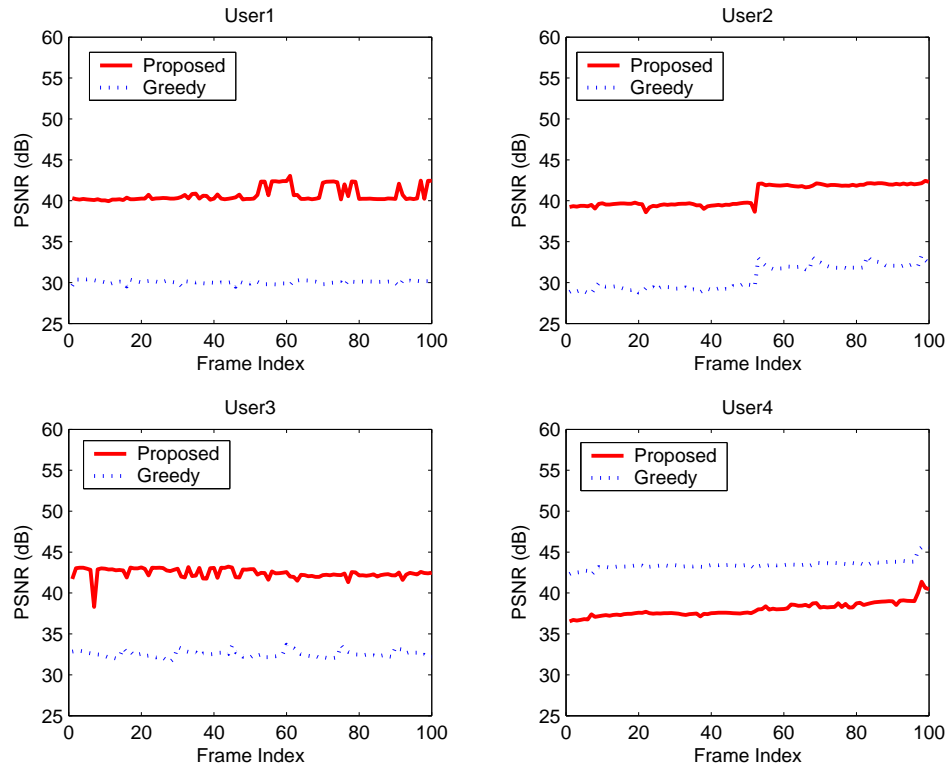


Figure 5.9: PSNR results for the sequences corresponding to 4 users.

Since the greedy algorithm assigns codes to the users who can use the least power to obtain the largest decreased distortion, which is the fourth user in this example, it cannot reduce the overall distortion much. Also, we can see that when using the proposed algorithm all users maintain good video quality (with PSNRs above 35 dB in all cases). This is not the case for the greedy algorithm, since some users (users 1 and 2 most notably) see their PSNR reach levels of approximately 30 dB. Figure 5.10 shows the number of users versus average total distortion over 100 frames from 50 different mobiles' locations. The simulation results demonstrate that the proposed algorithm performs an efficient resource allocation function since the average D_{sum} of the proposed algorithm outperforms that of the greedy algorithm by at least 45%.

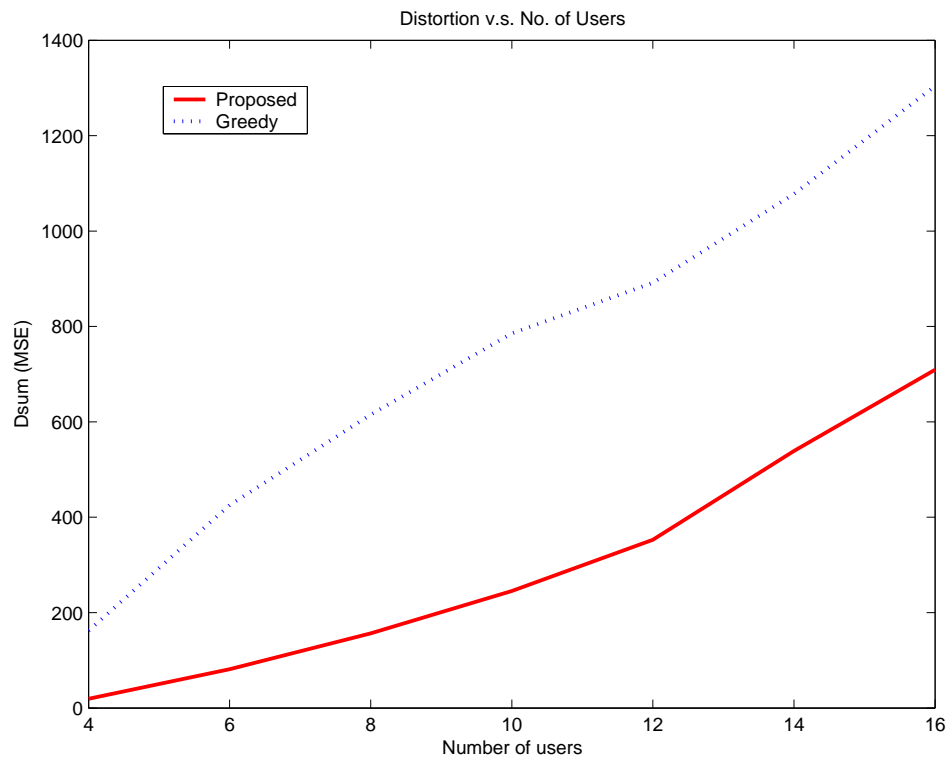


Figure 5.10: Performance comparison of the proposed and greedy scheme.

5.4 Conclusions

In this chapter we conclude our study of cross-layer designs for multimedia CDMA by discussing two designs, developed as collaborative work with members of the University of Maryland's Communication and Signal Processing Laboratory, that allocate resources in the down link. We started by noticing that due to the multipath propagation environment, inter-user interference is still an issue. We also noted that the constraining resources are the total transmit power at the base station and the spreading codes.

We then presented as contributions two algorithms. The first algorithm adapts the real-time source encoder to the channel and multiuser traffic conditions. In this case, we assume that the processing gain is fixed and that the parameters to be adapted are transmit power, source coding rate and channel coding rate. The algorithm performs this adaptation by minimizing the average distortion subject to constraints on the total power, each user distortion and a quality goal that limits the channel-induced distortion to a small proportion of the total. Simulations results shows the superior performance of this algorithm when compared to an equivalent system that cannot perform adaptation. Also, using a performance upper bound we showed that the proposed algorithm is near-optimum.

The second algorithm was designed to perform resource allocation in the downlink of multicode CDMA network. This means that, in addition to allocating transmit power, source coding rate and channel coding rate, the algorithm assigns to each real-time call spreading codes from a finite pool. This design was aimed at systems that use a real-time source encoder that generates a layered and embedded bit stream. During simulations, we used the MPEG-4 FGS video codec and we observed that the proposed algorithm both performs an efficient resource allocation and it outperforms an equivalent algorithm that assigns spreading codes using a greedy rule.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis we have study designs for resource allocation in CDMA networks carrying conversational-type calls. The designs are based on a cross-layer approach where the source encoder, the channel encoder and, in some cases, the processing gains are adapted. The primary focus of the study is on optimally multiplexing multimedia sources.

We first started by presenting the general model for our design. One of the characteristics of this model is that we constrain our design so that channel-distortion is kept at subjectively acceptable levels. The reason for this is that we noticed that, for the same distortion measure, channel-induced errors are perceptible more annoying than source encoding distortion because they manifest as artifacts that many times are perceptually evident, annoying and that in some cases might affect the understandability of the message. We called this goal the *quality goal*. The use of adaptable elements in our designs allow for an increase in the number of calls while meeting the quality goal. The tradeoff involved in this operation is that distortion also increases but now in a smooth and controllable way, following the source encoder distortion-rate performance. The result is a flexible system that sets an efficient tradeoff between end-to-end distortion and num-

ber of users and that clearly outperforms equivalent CDMA systems where capacity is increased in the traditional way, by allowing a reduction in SINR.

We first consider the design for only real-time sources in the uplink of the DS-SS-CDMA system. This study was divided into two parts, with chapter 2 dedicated to a simple model where all sources are encoded using the same single-state source encoder and processing gain is constant, and chapter 3 addressing the most general problem of optimal resource allocation to resolve interference-generated congestion for an arbitrary set of real-time variable-rate source encoders in a variable spreading factor multimedia CDMA network.

From the mathematical analysis in chapter 2 we concluded that the proposed system is able to significantly increase capacity at the cost of a moderate controlled smooth degradation of reconstructed source quality. We reach similar conclusions when we modified the design to consider a Rayleigh fading channel. When considering that the calls present in the system change over time, we found that the proposed design can support much larger offered loads as compared to a traditional equivalent CDMA system.

In chapter 3 we presented the important conclusion that the design problem (allocation of spreading factor, source coding rate and channel coding rate) could be further considered as the optimal source-controlled statistical multiplexing in multimedia CDMA. In this case, the statistical multiplexer needs to perform resource allocation so as to assign an *equivalent bandwidth*, which depends on target SINR and transmit bit rate, among calls in such a way that average distortion is minimized. Two solutions were discussed, one where the transmit bit rate is adapted through the source encoder rate and the other where both the transmit rate and the target SINR are adapted. We showed that both solutions are optimal in the sense of minimizing average distortion while meeting the quality goal and the system stability/power amplifiers dynamic range limits. We

also showed that the proposed system is able to extend operation beyond what has been normally considered an outage region at the cost of a smooth increase in distortion.

The analogy between the problem under study and statistical multiplexing prompts the design of a system that integrates both real-time and data traffic. Since our focus is the statistical multiplexing technique we studied in chapter 4 the effects that changing the equivalent bandwidth have on a system carrying real-time and data traffic. In this chapter we studied the sensibility of both the real-time and data traffic subsections to changes in their assigned total equivalent bandwidth. We observed that while the data subsection mean delay is clearly sensible to this changes, the real-time subsection quality remains relatively insensible. Finally, we used these results to propose an integrated congestion relief scheme that temporarily accepts small increases in real-time distortion to relieve the congestion in the data subsection.

Finally, in chapter 5 we have presented two algorithms to allocate resources in the downlink of the CDMA system. The first of the algorithms was aimed at being used for voice calls since only one spreading code is used per call. Because the second algorithm can allocate more than one spreading code per call, it is applicable to any type of real-time source that uses a layered and embedded source encoder. For these solutions we showed their near-optimality and efficiency in the allocation of resources.

6.2 Future Work

In this thesis we observed that the problem of resource (spreading factor, source coding rate and channel coding rate) allocation that minimizes average distortion could be considered as the optimal source-controlled statistical multiplexing in multimedia CDMA. This is a powerful abstraction for the studied problem and could be used as a frame-

work to solve many problems in this area. Some potential future research problems that can be derived from this thesis involves the design of protocols to efficiently integrate heterogeneous traffic in the CDMA network and the study of statistical multiplexing in networks that feature combined time-division and code-division multiple access.

An important application of the study in this thesis is the use of the proposed designs to extend operation of the CDMA network beyond a defined congestion operating point. As noted, this feature increases the CDMA cellular network resiliency in a way analogous to rerouting of traffic in a packet network where a router goes out of service. Future work in this area involves considering practical implementation issues beyond the academic study in this thesis.

BIBLIOGRAPHY

- [1] 3GPP2. 3gpp2 c.s0001-d, introduction to cdma2000 spread spectrum systems, version 1.0. February 2004.
- [2] F. Adachi. Effects of orthogonal spreading and rake combining on ds-cdma forward link in mobile radio. *IEICE Trans. on Communications*, (11):1703–1712, November 1997.
- [3] H. Akimaru and K. Kawashima. *Teletraffic, Second Edition*. Springer-Verlag, 1999.
- [4] M. Alasti and N. Farvardin. Seama: a source encoding assisted multiple access protocol for wireless communications. *IEEE Journ. Sel. Areas in Comm.*, 18(9):1682–1700, September 2000.
- [5] S. Boyd and L. Vandenberghe. Convex optimization. <http://www.stanford.edu/boyd/cvxbook.html>, Cambridge University Press 2003.
- [6] P. T. Brady. A model for on-off speech patterns in two-way conversation. *Bell System Technical Journal*, 48:2445–2472, 1969.

- [7] S. Chakravarty, R. Pankaj, and E. Esteves. An algorithm for reverse traffic channel rate control for cdma2000 high rate packet data systems. In *Global Telecommunications Conference, GLOBECOM '01*, volume 6, pages 3733–3737, 2001.
- [8] Y. S. Chan and J. W. Modestino. Transport of scalable video over cdma wireless networks: a joint source coding and power control approach. In *IEEE International Conference on Image Processing (ICIP)*, volume 2, pages 973–976, 2001.
- [9] J.S. Collura, A. McCree, and T.E. Tremain. Perceptually based distortion measurements for spectrum quantization. In *IEEE Workshop on Speech Coding for Telecommunications*, number 49–50, 1995.
- [10] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley Inc., 1991.
- [11] E. Dahlman, B. Gudmundson, M. Nilsson, and J. Skold. Umts/imt-2000 based in wideband cdma. *IEEE Communications Magazine*, 36:70–80, September 1998.
- [12] ETSI/GSM. Digital cellular telecommunications system (phase 2+); adaptive multi-rate (amr) speech transcoding (gsm 06.90 version 7.2.1 release 1998). In *Document ETSI EN 301 704 V7.2.1 (2000-04)*.
- [13] R. Fantacci and S. Nannicini. Multiple access protocol for integration of variable bit rate multimedia traffic in umts/imt-2000 based on wideband cdma. *IEEE Journ. Sel. Areas in Comm.*, 18(8):1441–1454, August 2000.
- [14] P. Frenger, P. Orten, T. Ottosson, and A. B. Svensson. Rate-compatible convolutional codes for multirate ds-cdma systems. *IEEE Trans. Comm.*, 47:828–836, June 1999.

- [15] Y. Gao, E. Shlomot, A. Benyassine, J. Thyssen, H. Su, and C. Murgia. The smv algorithm selected by tia and 3gpp2 for cdma applications. In *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, volume 2, pages 709–712, May 2001.
- [16] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic, 1991.
- [17] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, and C. E. Wheatley. On the capacity of a cellular cdma system. *IEEE Transactions on Vehicular Technology*, 40(2):303–312, 1991.
- [18] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard, and B. Ramamurthi. Packet reservation multiple access for local wireless communications. *IEEE Trans. Comm.*, 37(8):885–890, August 1989.
- [19] J. Hagenauer. Rate compatible punctured convolutional (RCPC) codes and their applications. *IEEE Trans. Comm.*, 36(4):389–399, April 1988.
- [20] Z. Han, A. Kwasinski, K. J. Ray Liu, and N. Farvardin. Pizza party algorithm for real time distortion management in downlink single-cell cdma systems. In *2003 Allerton Conference*, number 1858–1859, October 2003.
- [21] T. Holliday and A. Goldsmith. Optimal power control and source-channel coding for delay constrained traffic over wireless channels. In *IEEE International Conference on Communications (ICC)*, volume 2, pages 831–835, New York, New York, 2002.
- [22] M.L. Honig and B. K. Joon. Allocation of ds-cdma parameters to achieve multiple rates and qualities of service. In *Global Telecommunications Conference, GLOBECOM '96*, volume 3, pages 1974–1978, 1996.

- [23] Qualcomm Inc. An overview of the application of code division multiple access (cdma) to digital cellular systems and personal cellular. In *Document EX60-10010*, 1992.
- [24] J. M. Jacobsmeyer. Congestion relief on power-controlled cdma networks. *IEEE Journ. Sel. Areas in Comm.*, 14(9):1758–1761, December 1996.
- [25] S. A. Jafar and A. Goldsmith. Optimal rate and power adaptation for multirate cdma. In *IEEE 52nd Vehicular Technology Conference*, volume 3, pages 994–1000, Fall 2000.
- [26] I-M. Kim and H-M. Kim. A new medium access scheme for multimedia service in cdma systems. In *Proc. IEEE 51st. Vehicular Technology Conference, Spring 2000, Tokio*, number 2207–2211, May 2000.
- [27] I-M. Kim and H-M. Kim. An optimum power management scheme for wireless video service in cdma systems. *IEEE Trans. On Wireless Communications*, 2(1):81–91, January 2003.
- [28] J. B. Kim and M. L. Honig. Resource allocation for multiple classes of ds-cdma traffic. *IEEE Trans. on Vehicular Technology*, 49(2):506–519, March 2000.
- [29] J. B. Kim, M. L. Honig, and S. Jordan. Dynamic resource allocation for integrated voice and data traffic in ds-cdma. In *IEEE 54th Vehicular Technology Conference*, volume 1, pages 42–46, Fall 2001.
- [30] A. Kwasinski and N. Farvardin. Extending operation of a cdma network by cross-layer resource allocation for real-time traffic in fading channels. In *Allerton Conference*, Allerton House, Monticello, Illinois, October 2003.

- [31] A. Kwasinski and N. Farvardin. Resource allocation for cdma networks based on real-time rate adaptation. In *IEEE International Conference on Communications (ICC)*, Anchorage, Alaska, May 2003.
- [32] A. Kwasinski and N. Farvardin. Optimal resource allocation for cdma networks based on arbitrary real-time source coders adaptation with application to mpeg4 fgs. In *IEEE Wireless Communications and Networking Conference (WCNC)*, Atlanta, Georgia, March 2004.
- [33] A. Kwasinski, Z. Han, and K. J. Ray Liu. Dynamic real time distortion management over multimedia downlink cdma. In *2004 IEEE Wireless Communications and Networking Conference*, 2004.
- [34] E. A. Lee and D. G. Messerschmitt. *Digital Communications, Second Edition*. Kluwer Academic Publishers., 1994.
- [35] A. Leon-Garcia. *Probability and Random Processes for Electrical Engineering, Second Edition*. Addison-Wesley Pub., 1993.
- [36] W. Li. Overview of fine granularity scalability in mpeg-4 video standard. *IEEE Trans. On Circuits and Systems for Video Technology*, 11(3):301–317, March 2001.
- [37] S. Nanda, D. J. Goodman, and U. Timor. Performance of prma: A packet voice protocol for cellular systems. *IEEE Trans. on Vehicular Technology*, 40:585–598, August 1991.
- [38] A. Papoulis. *Probability, Random Variables, and Stochastic Processes, Third Edition*. WCB/McGraw-Hill, 1991.

- [39] J. G. Proakis. *Digital Communications*. McGraw-Hill Inc., 1994.
- [40] Inc. Qualcomm. *Proposed EIA/TIA Interim Standard, Wideband Spread Spectrum Digital Cellular System Dual-Mode Mobile Station-Base Station Compatibility Standard, document 80-7814 Rev. DCR 03567*. 1992.
- [41] L. Rabiner and B-H Juang. *Fundamentals of Speech Recognition*. Prentice Hall PTR., 1993.
- [42] A. Sampath, N. B. Mandayam, and J. M. Holtzman. Power control and resource management for a multimedia cdma wireless system. In *PIMRC'95*, Toronto, Canada, 1995.
- [43] A. Sampath, N. B. Mandayam, and J. M. Holtzman. Erlang capacity of a power controlled integrated voice and data cdma system. In *IEEE 47th Vehicular Technology Conference*, volume 3, pages 1557–1561, 1997.
- [44] Yoon seok Jung, Hyun cheol Jeon, Si mon Shin, Jang-Hoon Oh, Bum Kwon, and Jong-Tae Ihm. Coverage and capacity analysis in cdma2000 network for voice and packet data services. In *IEEE 54th. Vehicular Technology Conference, Fall*, volume 4, pages 2028–2032, October 2001.
- [45] Y. Shoham and A. Gersho. Efficient bit allocation for an arbitrary set of quantizers. *IEEE Trans. on Acoust. Speech, Signal Processing*, 36(9):1445–1453, September 1988.
- [46] G-M. Su, Z. Han, A. Kwasinski, M. Wu, and K. J. Ray Liu. Distortion management of real-time mpeg-4 fgs video over downlink multicode cdma networks. In *IEEE Intl. Conf. on Communications*, 2004.

- [47] DARPA TIMIT. Acoustic-phonetic continuous speech corpus cd-rom. In *Document NISTIR 4930, NIST Speech Disk 1-1.1*.
- [48] D. N. C. Tse and S. V. Hanly. Linear multiuser receiver: Effective interference, effective bandwidth and user capacity. *IEEE Trans. Info. Theory*, 45(2):641–657, March 1999.
- [49] S. Verdú. *Multiuser Detection*. Cambridge University Press, 1998.
- [50] S. Vishwanath, S. A. Jafar, and A. Goldsmith. Optimal power and rate allocation strategies for multiple access fading channels. In *IEEE 53rd Vehicular Technology Conference*, volume 4, pages 2888–2892, Spring 2001.
- [51] A. J. Viterbi. Error-correcting coding for cdma systems. In *IEEE Third International Symposium on Spread Spectrum Techniques and Applications, IEEE ISSSTA '94*, volume 1, pages 22–26, July 1994.
- [52] A. J. Viterbi. *CDMA, Principles of Spread Spectrum Communications*. Addison-Wesley Wireless Communications Series, 1995.
- [53] A. J. Viterbi and A. M. Viterbi. Erlang capacity of power controlled cdma systems. *IEEE Journ. Sel. Areas in Comm.*, 11(6):892–900, August 1993.
- [54] B. Walke, P. Seidenberg, and M. P. Althoff. *UMTS, The Fundamentals*. John Wiley and Sons, Ltd., 2003.
- [55] R. Yallapragada. Qos implementation in cdma2000. In *IEEE International Conference on Personal Wireless Communications*, pages 45–50, December 2002.
- [56] R. Yallapragada and V. Kripalani. Increments in voice capacity and impact on voice quality with new vocoders in gsm and cdma systems. In *IEEE Interna-*

tional Conference on Personal Wireless Communications, pages 100–104, December 2002.

- [57] N. Yin and M. Hluchyj. A dynamic rate control mechanism for source coded traffic in a fast packet network. *IEEE Journ. Sel. Areas in Comm.*, 9(7):1003–1012, September 1991.
- [58] L. C. Yun and D. G. Messerschmitt. Variable quality of service in cdma systems by statistical power control. In *IEEE International Conference on Communications (ICC)*, Seattle, Washington, 1995.
- [59] J. Zander and S.-L. Kim. *Radio Resource Management for Wireless Networks*. Artech House, 2001.
- [60] J. Zhang and E. K. P. Chong. Cdma systems in fading channels: Admissibility, network capacity and power control. *IEEE Trans. Info. Theory*, 46(3):962–981, May 2000.
- [61] Q. Zhang, Z. Ji, W. Zhu, and Y.-Q. Zhang. Power-minimized bit allocation for video communication over wireless channels. *IEEE Trans. on Circuits and Systems for Video Technology*, 12(6):398–408, June 2002.
- [62] S. Zhao, Z. Xiong, and X. Wang. Joint error control and power allocation for video transmission over cdma networks with multiuser detection. *IEEE Trans. on Circuits and Systems for Video Technology*, 12(6):425–437, June 2002.