

GSM

Architecture, Protocols
and Services

3RD EDITION

JÖRG EBERSPÄCHER HANS-JÖRG VÖGEL
CHRISTIAN BETTSTETTER CHRISTIAN HARTMANN

 WILEY

EBERSPÄCH
VÖGEL
BETTSTETT
HARTMAN

GSM

Architecture, Protocols and Services

3RD EDITION



GSM – Architecture, Protocols and Services

Third Edition

GSM – Architecture, Protocols and Services

Third Edition

Jörg Eberspächer

Technische Universität München, Germany

Hans-Jörg Vögel

BMW Group Research & Technology, Germany

Christian Bettstetter

University of Klagenfurt, Austria

Christian Hartmann

Technische Universität München, Germany



A John Wiley and Sons, Ltd, Publication

This English language edition first published 2009
© 2009 John Wiley & Sons Ltd

Originally published in the German language by B.G. Teubner GmbH as "Jörg Eberspächer/Hans-Jörg Vögel/Christian Bettstetter: GSM Global System for Mobile Communication. 3. Auflage (3rd edition)."
© B.G. Teubner GmbH, Stuttgart/Leipzig/Wisbaden 2001

Registered office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ,
United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data

Eberspaecher, Joerg.

GSM, Global System for Mobile Communication. English

GSM : architecture, protocols and services / Joerg Eberspaecher ... [et al.]. – 3rd ed.

p. cm.

Prev. ed.: GSM switching, services, and protocols, 2001.

ISBN 978-0-470-03070-7 (cloth)

1. Global system for mobile communications. I. Eberspaecher, J. (Joerg) II. Title.

TK5103.483.E2413 2008

621.3845'6–dc22

2008034404

A catalogue record for this book is available from the British Library.

ISBN 978-0-470-03070-7 (H/B)

Set in 10/12pt Times by Sunrise Setting Ltd, Torquay, UK.



Contents

Preface	xi
1 Introduction	1
1.1 The idea of unbounded communication	1
1.2 The success of GSM	3
1.3 Classification of mobile communication systems	3
1.4 Some history and statistics of GSM	5
1.5 Overview of the book	7
2 The mobile radio channel and the cellular principle	9
2.1 Characteristics of the mobile radio channel	9
2.2 Separation of directions and duplex transmission	12
2.2.1 Frequency Division Duplex	13
2.2.2 Time Division Duplex	13
2.3 Multiple access	13
2.3.1 Frequency Division Multiple Access	14
2.3.2 Time Division Multiple Access	15
2.3.3 Code Division Multiple Access	17
2.3.4 Space Division Multiple Access	18
2.4 Cellular principle	22
2.4.1 Definitions	23
2.4.2 Carrier-to-interference ratio	24
2.4.3 Formation of clusters	25
2.4.4 Traffic capacity and traffic engineering	26
2.4.5 Sectorization of cells	28
2.4.6 Spatial filtering for interference reduction (SFIR)	31
3 System architecture and addressing	43
3.1 System architecture	43
3.2 The SIM concept	45
3.3 Addressing	46
3.3.1 International mobile station equipment identity	46
3.3.2 International mobile subscriber identity	47
3.3.3 Mobile subscriber ISDN number	47
3.3.4 Mobile station roaming number	48

3.3.5	Location area identity	49
3.3.6	Temporary mobile subscriber identity	49
3.3.7	Other identifiers	50
3.4	Registers and subscriber data	50
3.4.1	Location registers (HLR and VLR)	50
3.4.2	Security-related registers (AUC and EIR)	51
3.4.3	Subscriber data	52
3.5	Network interfaces and configurations	53
3.5.1	Interfaces	54
3.5.2	Configurations	55
4	Air interface – physical layer	57
4.1	Logical channels	57
4.1.1	Traffic channels	57
4.1.2	Signaling channels	58
4.1.3	Example: connection setup for incoming call	61
4.1.4	Bit rates, block lengths and block distances	61
4.1.5	Combinations of logical channels	62
4.2	Physical channels	62
4.2.1	Modulation	63
4.2.2	Multiple access, duplexing and bursts	65
4.2.3	Optional frequency hopping	69
4.2.4	Summary	70
4.3	Synchronization	70
4.3.1	Frequency and clock synchronization	71
4.3.2	Adaptive frame synchronization	73
4.4	Mapping of logical onto physical channels	75
4.4.1	26-frame multiframe	77
4.4.2	51-frame multiframe	77
4.5	Radio subsystem link control	80
4.5.1	Channel measurement	81
4.5.2	Transmission power control	86
4.5.3	Disconnection due to radio channel failure	87
4.5.4	Cell selection and operation in power conservation mode	89
4.6	Channel coding, source coding and speech processing	91
4.7	Source coding and speech processing	92
4.8	Channel coding	96
4.8.1	External error protection: block coding	98
4.8.2	Internal error protection: convolutional coding	103
4.8.3	Interleaving	107
4.8.4	Mapping onto the burst plane	113
4.8.5	Improved codecs for speech services: half-rate codec, enhanced full-rate codec and adaptive multi-rate codec	115
4.9	Power-up scenario	118

5	Protocols	121
5.1	Protocol architecture planes	121
5.2	Protocol architecture of the user plane	123
5.2.1	Speech transmission	123
5.2.2	Transparent data transmission	126
5.2.3	Nontransparent data transmission	127
5.3	Protocol architecture of the signaling plane	130
5.3.1	Overview of the signaling architecture	130
5.3.2	Transport of user data in the signaling plane	139
5.4	Signaling at the air interface (Um)	140
5.4.1	Layer 1 of the MS-BTS interface	140
5.4.2	Layer 2 signaling	142
5.4.3	Radio resource management	146
5.4.4	Mobility management	152
5.4.5	Connection management	156
5.4.6	Structured signaling procedures	160
5.4.7	Signaling procedures for supplementary services	161
5.4.8	Realization of SMS	165
5.5	Signaling at the A and Abis interfaces	166
5.6	Security-related network functions: authentication and encryption	173
5.6.1	Protection of subscriber identity	173
5.6.2	Verification of subscriber identity	173
5.6.3	Generating security data	175
5.6.4	Encryption of signaling and payload data	176
5.7	Signaling at the user interface	179
6	Roaming and handover	183
6.1	Mobile application part interfaces	183
6.2	Location registration and location update	184
6.3	Connection establishment and termination	188
6.3.1	Routing calls to MSs	188
6.3.2	Call establishment and corresponding MAP procedures	191
6.3.3	Call termination	195
6.3.4	MAP procedures and routing for short messages	195
6.4	Handover	197
6.4.1	Overview	197
6.4.2	Intra-MSC handover	199
6.4.3	Decision algorithm for handover timing	199
6.4.4	MAP and inter-MSC handover	205
7	Services	211
7.1	Classical GSM services	211
7.1.1	Teleservices	211
7.2	Popular GSM services: SMS and MMS	212
7.2.1	SMS	212
7.2.2	EMS	213

7.2.3	MMS	213
7.3	Overview of GSM services in Phase 2+	214
7.4	Bearer and teleservices of GSM Phase 2+	215
7.4.1	Advanced speech call items	215
7.4.2	New data services and higher data rates: HSCSD, GPRS and EDGE	220
7.5	Supplementary services in GSM Phase 2+	221
7.5.1	Supplementary services for speech	221
7.5.2	Location service	221
7.6	Service platforms	222
7.6.1	CAMEL: GSM and INs	223
7.6.2	Service platforms on the terminal side	224
7.7	Wireless application protocol	226
7.7.1	Wireless markup language	226
7.7.2	Protocol architecture	227
7.7.3	System architecture	230
7.7.4	Services and applications	231
8	Improved data services in GSM: GPRS, HSCSD and EDGE	233
8.1	GPRS	233
8.1.1	System architecture of GPRS	234
8.1.2	Services	237
8.1.3	Session management, mobility management and routing	238
8.1.4	Protocol architecture	242
8.1.5	Signaling plane	247
8.1.6	Interworking with IP networks	249
8.1.7	Air interface	250
8.1.8	Authentication and ciphering	257
8.1.9	Summary of GPRS	259
8.2	HSCSD	260
8.2.1	Architecture	261
8.2.2	Air interface	261
8.2.3	HSCSD resource allocation and capacity issues	263
8.3	EDGE	264
8.3.1	The EDGE concept	264
8.3.2	EDGE physical layer, modulation and coding	265
8.3.3	EDGE: effects on the GSM system architecture	266
8.3.4	ECSD and EGPRS	267
8.3.5	EDGE Classic and EDGE Compact	268
9	Beyond GSM and UMTS: 4G	269
	Appendices	271
A	Data communication and networking	273
A.1	Reference configuration	273
A.2	Overview of data communication	274

CONTENTS

ix

A.3	Service selection at transitions between networks	277
A.4	Bit rate adaptation	277
A.5	Asynchronous data services	280
A.5.1	Transparent transmission in the mobile network	280
A.5.2	Nontransparent data transmission	284
A.5.3	PAD access to public packet-switched data networks	286
A.6	Synchronous data services	288
A.6.1	Overview	288
A.6.2	Synchronous X.25 packet data network access	289
A.7	Teleservices: fax	291
B	Aspects of network operation	295
B.1	Objectives of GSM NM	295
B.2	Telecommunication management network	297
B.3	TMN realization in GSM networks	300
C	GSM Addresses	305
D	List of Acronyms	307
	References	313
	Index	317

Preface

The GSM family (GSM, GPRS, EDGE) has become one of the most successful technical innovations in history. As of June 2008, more than 2.9 billion subscribers were using GSM, corresponding to a market share of more than 81%, and its story continues, even now, despite the introduction and development of next-generation systems such as IMT-2000 or UMTS (3G) and even systems beyond 3G, dubbed IMT-Advanced.

At the same time, wireless local area networks have substantially expanded the wireless market, sometimes drawing market share from GPRS and 3G (e.g. in public WiFi hotspots), sometimes coexisting (e.g. in UMTS home routers used as a replacement for fixed wire connections). However, these are used typically for low mobility applications. Mobile communication with all of its features and stability has become increasingly important: cellular and GSM technology, plus, of course, lately 3G, GSMs sister technology, so-to-say.

Another impressive trend has emerged since our last edition: the permanent evolution in the handheld market, producing fancy mobile phones with cameras, large memory, MP3 players, Email clients and even satellite navigation. These features enable numerous nonvoice or multimedia applications, from which, of course, only a subset is or will be successful on the market.

In this third edition, we concentrate again on the architecture, protocols and operation of the GSM network and outline and explain the innovations introduced in recent years. The main novelties in this book are the presentation of capacity enhancement methods such as sectorization, the application of adaptive antennas for Spatial Filtering for Interference Reduction (SFIR) and Space Division Multiple Access (SDMA), a detailed introduction to HSCSD and EDGE for higher data rates, and an update of the available GSM services, specifically introducing the Multimedia Messaging Service (MMS).

We are happy to have received, over the past few years, many constructive comments, and a lot of praise and encouragement. The book has obviously been successfully used by professionals (especially people beginning careers in the cellular network business) but also by students including our own who use it as a textbook enhancing their course material.

Our author team has been enlarged with the addition of Dr. Christian Hartmann, an assistant professor at Technische Universität München, who took most of the load for this edition.

We thank all of the involved staff from Wiley who convinced us to prepare this updated version of a book that will hopefully be as successful over the next few years as in the past.

Jörg Eberspächer
Hans-Jörg Vögel
Christian Bettstetter
Christian Hartmann
Munich

1

Introduction

1.1 The idea of unbounded communication

Communication everywhere, with everybody, and at any time – that was the dream and goal of researchers, engineers and users, since the advent of the first wireless communication systems. Today it feels like we have almost reached that goal. Digitalization of communication systems, enormous progress in microelectronics, computers and software technology, the invention of efficient algorithms and procedures for compression, security and processing of all kinds of signals, as well as the development of flexible communication protocols have all been important prerequisites for this progress. Today, technologies are available that enable the realization of high-performance and cost-effective communication systems for many application areas.

Using current wireless communication systems, the most popular of which is GSM (Global System for Mobile Communication), we see that we have the freedom to not only roam within a network, but also between different networks, and that we can in fact communicate (almost) everywhere (unless we are in one of the rare spots still without GSM coverage today), with (almost) everybody (unless our desired communication partner is in one of the rare spots mentioned above or chooses not to be reachable), and at (almost) any time (unless we forgot to pay our last phone bill and the operator decides to lock us out). If there is one major aspect still missing in order to make our wireless experience flawless, it is the large (albeit diminishing) gap between data rates available through wireless services and those available through wired services, such as Digital Subscriber Line (xDSL). This and the limited capability of data representation at the mobile terminal (mostly due to the limited size of mobile phones) is one of the main challenges for future developments in wireless communication.

Let us now briefly take a look at the functionalities, which enable us to move and roam so freely in GSM systems: terminal mobility and personal mobility.

In the case of terminal mobility, the subscriber is connected to the network in a wireless way – via radio- or light-waves – and can move with their terminal freely, even during a

communication connection. The degree of mobility depends on the type of mobile radio network. The requirements for a cordless in-house telephone are much less critical than for a mobile telephone that can be used in a car or train. If mobility is to be supported across the whole network (or country) or even beyond the network (or national) boundaries, additional switching technology and administrative functions are required, to enable the subscribers to communicate in wireless mode outside of their home areas.

Such extended network functions are also needed to realize personal mobility and universal reachability. This is understood to comprise the possibility of location-independent use of all kinds of telecommunication services, including fixed and wireless networks. The user identifies themselves (the person), e.g. by using a chip card, at the place where they are currently staying and have access to the network. There, the same communication services can be used as at home, limited only by the properties of the local network or terminal used. A worldwide unique and uniform addressing system is an important requirement for personal mobility.

In the digital mobile communication system GSM, which is the subject of this book, terminal mobility is the predominant issue. Wireless communication has become possible with GSM in any town, any country and even on any continent.

GSM technology contains the essential intelligent functions for the support of personal mobility, especially with regards to user identification and authentication, and for the localization and administration of mobile users. Here it is often overlooked that in mobile communication networks by far the largest part of the communication occurs over the fixed network part, which interconnects the radio stations (base stations). Therefore, it is no surprise that in the course of further development and evolution of the telecommunication networks, a lot of thought has been given to the convergence of fixed and mobile networks.

In the beginning, GSM was used almost exclusively for speech communication; however, the Short Message Service (SMS) soon became extremely popular with GSM users: several billion text messages are being exchanged between mobile users each month. In the mean time, additional data services have been realized, most notably the High Speed Circuit Switched Data (HSCSD) and the General Packet Radio Service (GPRS), which enable improved data rate performance by allowing for more than one GSM timeslot to be used by a terminal for a service at a time. The driving factor for new (and higher bandwidth) data services obviously is wireless access to the Internet. To this end, the Wireless Application Protocol (WAP) is also explained in this book. These additions are already working towards closing the gap between wireless and fixed networks that we discussed above.

A further step was the introduction of third-generation (3G) mobile communication networks. The 3G networks, known as the Universal Mobile Telecommunication System (UMTS) in Europe and as the International Mobile Telecommunication System 2000 (IMT-2000) worldwide, have already been introduced. However, the implementation of such 3G wireless technologies has not so far stretched much beyond busy city centers. In fact, GSM is still the major technology for providing full coverage, while 3G technology is applied to cover hot-spot areas, mainly those with very high user densities. Thanks to multi-mode terminals, which can handle both standards (GSM and UMTS), wireless network users usually do not even realize which technology they are currently using while making a call or using other wireless services. Regarding the relevance of GSM technology, it is important to note that most network providers who have implemented UMTS are using basically the same fixed backbone infrastructure architecture as used for GSM and GPRS together.

1.2 The success of GSM

GSM is now in more countries than McDonalds.

(Mike Short, Chairman MoU Association 1995–1996)

The relevance of the GSM standard today becomes obvious when we take a brief look at the success story of GSM so far and keeping in mind that many countries are still working towards full wireless coverage, mainly by deploying GSM. GSM was initially designed as a pan-European mobile communication network, but shortly after the successful start of the first commercial networks in Europe, GSM systems were also deployed on other continents (e.g. in Australia, Hong Kong and New Zealand). In the meantime, as of May 2008, 670 networks in 208 countries are in operation according to GSM world.

In addition to GSM networks that operate in the 900 MHz frequency band, other so-called Personal Communication Networks (PCNs) and Personal Communication Systems (PCSs) are in operation. They use frequencies around 1800 MHz, or around 1900 MHz in North America. Apart from the peculiarities that result from the different frequency range, PCNs/PCSs are full GSM networks without any restrictions, in particular with respect to services and signaling protocols. International roaming among these networks is possible based on the standardized interface between mobile equipment and the Subscriber Identity Module (SIM) card, which enables personalization of equipment operating in different frequency ranges (SIM card roaming). Now that UMTS technology has been integrated by most wireless providers into their networks, roaming not only between providers but also between different technologies is already state of the art. To this end, multi-band and multi-standard terminals have been developed and are considered commonplace today. Users of state-of-the-art terminals with a SIM card from one of the major providers in Europe can use their terminals in different frequency ranges as well as in GSM and UMTS networks, without having to configure or select anything. The terminals roam between different networks and technologies automatically.

1.3 Classification of mobile communication systems

This book deals almost exclusively with GSM; however, GSM is only one of many facets of modern mobile communication.

For the bidirectional – and hence genuine – communication systems, the simplest variant is the cordless telephone with very limited mobility (particularly the Digital Enhanced Cordless Telecommunications (DECT) standard in Europe). This technology is also employed for the expansion of digital Private Branch Exchanges (PBXs) with mobile extensions.

Local Area Networks (LANs) have also been augmented with mobility functions: Wireless LANs (WLANs) have been standardized and are now offered by several companies. WLANs offer Internet Protocol (IP)-based, wireless data communication with very high bit rates but limited mobility. WLANs have been installed, for example, in office environments and airports, as a supplement or alternative to wired LANs, but also in universities, cafes, restaurants, etc. WLAN access points, however, are also very popular in private homes as access technology. In fact, in urban areas the coverage of IEEE 802.11 type access points is impressive and could theoretically be used for roaming while using WLAN by applying

Mobile IP enhanced routing for mobility support. This, however, is hindered by the fact that each WLAN cell is typically managed by someone else, in effect making it impossible to form a large network. Another aspect is that most WLAN cells are of course encrypted and cannot therefore be used by just anyone. A little different are campus-type WLAN networks, operated by companies or universities, for instance. The IEEE 802.11 type WLAN standards are continuously being amended. The IEEE 802.11n standard for high data rates enables data rates in the 100 Mbit/s range by applying multiple antennas and using multiple in multiple out (MIMO) technology. Even though standardization is not complete for IEEE 802.11n, so-called draft-n devices are already commercially available and promise data rates close to 100 Mbit/s.

Another emerging class of wireless networks are being used for short-range communication. Bluetooth, for example, replaces cables by enabling direct wireless information exchange between electronic devices (e.g. between cellular phones, Personal Digital Assistants (PDAs), computers and peripherals). These networks are also called Body Area Networks or Personal Area Networks. Unlike the mobile technologies mentioned above, they are not based on a fixed network infrastructure (e.g. base stations). The possibility of building up such networks in a spontaneous and fast way gave them the name *ad hoc* networks. WLAN technologies also include the capability for peer-to-peer *ad hoc* communication (in addition to the classical client-to-base station transmission modus).

GSM and UMTS belong to the class of cellular networks that are used predominantly for public mass communication. These had an early success with analog systems such as the Advanced Mobile Phone System (AMPS) in America, the Nordic Mobile Telephone (NMT) in Scandinavia, or the *C-Netz* in Germany. Founded on the digital system GSM (with its variants for 900, 1800 and 1900 MHz), a market with millions of subscribers worldwide was generated, and it represents an important economic force. A strongly contributing factor to this rapid development of markets and technologies has been the deregulation of the telecommunication markets, which allowed the establishment of new network operators.

Another competing or supplementary technology is satellite communication based on Low Earth Orbiting (LEO) or Medium Earth Orbiting (MEO) satellites, which also offer global, and in the long term even broadband, communication services. Trunked radio systems – in digital form with the European standard *Trans European Trunked Radio* (TETRA) – are used for business applications such as fleet control. They offer private services that are only accessible by closed user groups.

In addition to bidirectional communication systems, there also exists a variety of unidirectional systems, where subscribers can only receive but not send data. With unidirectional message systems (paging systems) users may receive short text messages. A couple of years ago, paging systems were very popular, since they offered a cost-effective reachability with wide-area coverage. Today, the SMS in GSM has basically replaced the function of paging systems. Some billion SMS messages are being exchanged between mobile GSM users each month. Digital broadcast systems, such as Digital Audio Broadcast (DAB) and Digital Video Broadcast (DVB), are very interesting for wireless transmission of radio and television stations as well as for audio- and video-on-demand and broadband transmission of Internet pages.

GSM and its enhancements (including UMTS air interfaces), however, will remain the technological base for mobile communication for many years, and will continue to open up new application areas.

1.4 Some history and statistics of GSM

In 1982 the development of a pan-European standard for digital cellular mobile radio was started by the Groupe Spécial Mobile of the CEPT (Conférence Européenne des Administrations des Postes et des Télécommunications) (see Table 1.1). Initially, the acronym GSM was derived from the name of this group. After the founding of the European standardization institute ETSI (European Telecommunication Standards Institute), the GSM group became a technical committee of ETSI in 1989. After the rapid worldwide proliferation of GSM networks, the name has been reinterpreted as Global System for Mobile Communication.

After a series of incompatible analog networks had been introduced in parallel in Europe, e.g. Total Access Communication System (TACS) in the UK, NMT in Scandinavia and the C-Netz in Germany, work on the definition of a European-wide standard for digital mobile radio was started in the early 1980s. The GSM was founded, which developed a set of technical recommendations and presented them to ETSI for approval. These proposals were produced by the Special Mobile Group (SMG) in working groups called Sub Technical Committees (STCs), with the following division of tasks: service aspects (SMG 01), radio aspects (SMG 02), network aspects (SMG 03), data services (SMG 04) and network operation and maintenance (SMG 06). Further working groups were mobile station testing (SMG 07), integrated circuit card aspects (SMG 09), security (SMG 10), speech aspects (SMG 11) and system architecture (SMG 12) (ETSI, 2008). SMG 05 dealt with future networks and was responsible for the initial standardization phase of the next generation of the European mobile radio system, the UMTS. Later, SMG 05 was closed, and UMTS became an independent project and technical body of ETSI. The Third Generation Partnership Project (3GPP) has been founded in cooperation with other standardization committees worldwide. Its goal was the composition of the technical specifications for UMTS. Finally, in July 2000, ETSI announced the closure of the SMG which has been responsible for setting GSM standards for the last 18 years. Their remaining and further work has been transferred to groups inside and outside ETSI; most of the ongoing work has been handed over to the 3GPP.

After the official start of the GSM networks during the summer of 1992, the number of subscribers increased rapidly such that during the fall of 1993 already more than one million subscribers had made calls in GSM networks, more than 80% of them in Germany. On a global scale, the GSM standard also received very fast recognition, as evident from the fact that at the end of 1993 several commercial GSM networks started operating outside Europe, in Australia, Hong Kong and New Zealand. Afterwards, GSM was introduced in Brunei, Cameroon, Iran, South Africa, Syria, Thailand, USA and United Arab Emirates. Whereas the majority of the GSM networks operate in the 900 MHz band (GSM900), there are also networks operating in the 1800 MHz band (GSM1800) – PCN and Digital Communication System (DCS1800) – and in the United States in the 1900 MHz band (GSM1900) – PCS. These networks use almost completely identical technology and architecture; they differ essentially only in the radio frequencies used and the pertinent high-frequency technology, such that synergy effects can be taken advantage of, and the mobile exchanges can be constructed with standard components.

In parallel to the standardization efforts of ETSI, in 1987 the then existing prospective GSM network operators and the national administrations formed a group whose members signed a common Memorandum of Understanding (MoU). The MoU Association was supposed to form a base for allowing the transnational operation of mobile stations using

Table 1.1 Time history – milestones in the evolution of GSM.

Year	Event
1982	Groupe Spécial Mobile established by the CEPT.
1986	Reservation of the 900 MHz spectrum band for GSM agreed in the EC Telecommunications Council. Trials of different digital radio transmission schemes and different speech codes in several countries.
1987	Basic parameters of the GSM standard agreed in February.
1988	Completion of first set of detailed GSM specifications for infrastructure.
1989	Groupe Spéciale Mobile (transferred to an ETSI technical committee) defines the GSM standard as the internationally accepted digital cellular telephony standard.
1990	GSM adaptation work started for the DCS1800 band.
1991	First GSM call made by Radiolinja in Finland.
1992	First international roaming agreement signed between Telecom Finland and Vodafone (UK). First SMS sent.
1993	Telstra Australia becomes the first non-European operator. Worlds first DCS1800 (later GSM1800) network opened in the UK.
1994	GSM Phase 2 data/fax bearer services launched. GSM MoU membership surpasses 100 operators. GSM subscribers hit one million.
1995	117 GSM networks on air. The number of GSM subscribers worldwide exceeds 10 million. Fax, data and SMS services started, video over GSM demonstrated. The first North American PCS 1900 (now GSM 1900) network opened.
1996	First GSM networks in Russia and China go live. Number of GSM subscribers hits 50 million.
1997	First tri-band handsets launched.
1998	Number of GSM subscribers worldwide over 100 million.
1999	WAP trials begin in France and Italy.
2000	First commercial GPRS services launched. First GPRS handsets enter the market. Five billion SMS messages sent in one month.
2001	First 3GSM (W-CDMA) network goes live. Number of GSM subscribers exceed 500 million worldwide.
2003	First EDGE networks go live. Membership of GSM Association breaks through 200-country barrier. Over half a billion handsets produced in a year.
2008	GSM surpasses three billion customer threshold.

internationally standardized interfaces. As of April 2008, the GSM MoU has 747 members which operates 670 GSM networks in 200 countries.

1.5 Overview of the book

The remainder of this book is organized as follows. In Chapter 2, we give an introduction to radio channel characteristics and the cellular principle. The understanding of duplex and multiple access schemes serves as the basis for understanding GSM technology. We also describe some measures to increase the capacity in GSM systems, sectorization, as applied by most GSM networks already today, and Spatial Filtering for Interference Reduction (SFIR). Chapter 3 introduces the GSM system architecture and addressing. It explains the basic structure and elements of a GSM system and their interfaces as well as the identifiers of users, equipment and system areas. Next, Chapter 4 deals with the physical layer at the air interface (how are speech and data transmitted over the radio channel?). Among other things, it describes GSM modulation, multiple access, duplexing, frequency hopping, the logical channels and synchronization. Also we discuss GSM coding (source coding, speech processing and channel coding). In Chapter 5, the entire protocol architecture of GSM (payload transport and signaling) is covered. For example, communication protocols for radio resource management, mobility management, connection management at the air interface are explained as well as mechanisms for authentication and encryption. Chapter 6 describes in detail three main principles that are needed for roaming and switching: location registration and update (i.e. how does the network keep track of the user and find them when there is an incoming call?), connection establishment and termination and handover (i.e. how is a call transferred between cells?). Chapter 8 is on enhanced data services in GSM. It explains in detail GPRS which can be used for wireless Internet access. In addition this chapter includes HSCSD and Enhanced Data Rates for Global Evolution (EDGE). Chapter 7 contains the major GSM services and, finally, Chapter 9 gives a brief outlook on future mobile network developments. Appendix A covers basic GSM data services and Appendix B describes network operation and management.

The mobile radio channel and the cellular principle

Many measures, functions and protocols in digital mobile radio networks are based on the properties of the radio channel and its specific qualities, in contrast to information transmission through guided media. For the understanding of digital mobile radio networks it is therefore helpful to know a few related basic principles. For this reason, the most important fundamentals of the radio channel and of cellular and transmission technology are presented and briefly explained in the following. For a more detailed treatment, see, for example, Bertsekas and Gallager (1987), Lee (1989), Proakis (1995) and Steele and Hanzo (1999).

2.1 Characteristics of the mobile radio channel

The electromagnetic wave of the radio signal propagates under ideal conditions in free space in a radial-symmetric pattern. The received power P_r decreases with the square of the distance L from the transmitter. Specifically, the received power P_r can be described according to the free-space model as a function of the transmit power P_t , the distance L and the wavelength of the radio signal λ as

$$P_r = P_t \cdot g_t \cdot g_r \cdot \left(\frac{\lambda}{4\pi L} \right)^2, \quad (2.1)$$

where g_t and g_r are the transmit and receive antenna gains, respectively. While this model is appropriate, for instance, for inter-satellite as well as for Earth-to-satellite communication, it does not capture the effects of terrestrial radio propagation, where the signal is scattered and reflected by obstacles such as buildings, mountains, vegetation, the ground and water surfaces. At the receiver, direct and – potentially many – reflected signal components are superimposed. In effect, we can describe P_r as a linear function of P_t , g_t , g_r , and an overall channel gain g_c :

$$P_r = g_c \cdot g_t \cdot g_r \cdot P_t. \quad (2.2)$$

The channel gain g_c can be split into three components

$$g_c = g_d(L) \cdot g_s \cdot g_m \quad (2.3)$$

each capturing one of the main propagation effects.

- **Distance-dependent path gain $g_d(L)$:** This part of the channel gain is usually modeled as a deterministic function of the distance L between the transmitter and the receiver, such that $g_d(L) \cdot P_t$ gives the mean received power at distance L from the transmitter (assuming $g_t = g_r = 1$). A common model for the path gain is given by

$$g_d(L) = \left(\frac{\lambda}{4\pi L} \right)^2 \left(\frac{L_0}{L} \right)^{\gamma-2} \sim L^{-\gamma}, \quad (2.4)$$

where L_0 is a reference distance and $\gamma \geq 2$ is the attenuation exponent, depending on the propagation environment (Rappaport, 2002). Typical values for γ are between 3 and 5. In addition to the described model, specifically for modeling and planning of GSM networks, measurement-based models are available, such as the Okumura–Hata model (Hata, 1980; Okumura, 1968) for GSM900 networks and the COST-231 Hata model (Damosso, 1999) for GSM1800 networks. Those models are parameterized by the heights of transmit and receive antennas as well as by the propagation environment (rural, sub-urban or urban).

- **Shadowing gain g_s :** Shadowing describes the effect of fluctuations of the received power around the main value, as it is caused by obstacles such as buildings and vegetation. The severeness of the shadowing effect depends on the number and properties of obstacles between the transmitter and receiver. Changes in shadowing occur in the order of meters, e.g. when a user turns around a corner during a phone call. In accordance with measurement data, the most commonly used model for shadowing is a statistical model, describing the shadowing gain g_s as a log-normal distributed random variable. Therefore, the shadowing gain in decibels, i.e. $\chi = 10 \log_{10}(g_s)$, is distributed according to a Gaussian distribution given by

$$f_\chi(\chi) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\chi^2/2\sigma^2}. \quad (2.5)$$

The standard deviation σ defines the severeness of the shadowing and depends on the environment to be modeled. According to measurements, typical values for σ are between 5 and 10 dB (Geng and Wiesbeck, 1998).

- **Multipath fading gain g_m :** Another source of received power fluctuations around the mean value is caused by multipath propagation. In urban environments, in particular, multiple copies of the transmitted signal arrive at the receiver through different propagation paths. The superposition of many such copies of the transmitted signal, arriving at the receiver from different directions and with different delays, causes a wave field around the receiver. The received signal strength within this wave field changes severely in the order of the signal wavelength between places where destructive and constructive superposition occurs. The resulting amplitude variations

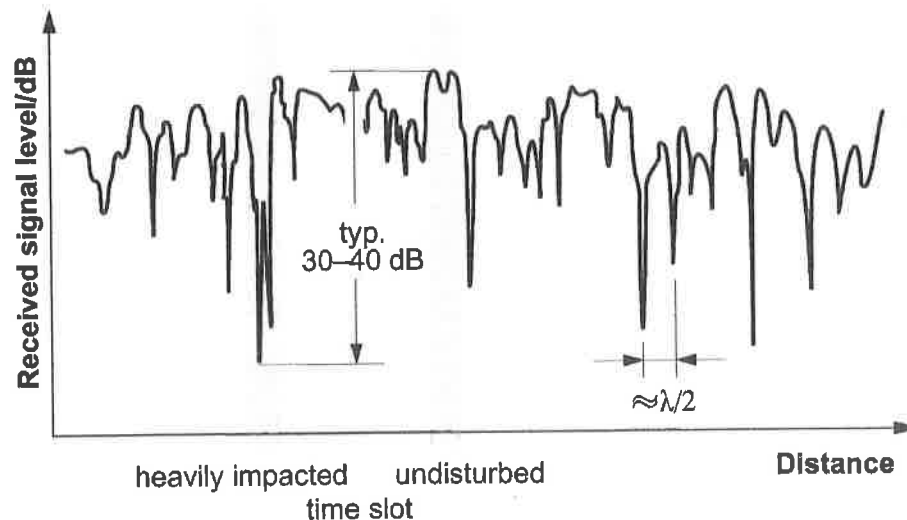


Figure 2.1 Typical signal in a channel with Rayleigh fading.

are modeled by a random variable a , such that

$$g_m = a^2. \quad (2.6)$$

The distribution of the random variable a depends on the propagation environment. If no direct line of sight between sender and receiver is present, a is assumed to be Rayleigh distributed, while an additional line of sight can be taken into consideration if a Rice distribution is applied. Figure 2.1 shows typical channel fluctuations according to Rayleigh fading for a receiver traveling through the wave field. It can be shown that if a is Rayleigh distributed, the multipath fading gain $g_m = a^2$ will be exponentially distributed (Schwartz, 2005).

The signal level observed at a specific location is determined by the phase shift of the multipath signal components. This phase shift depends on the wavelength of the signal, and thus the signal level at a fixed location is also dependent on the transmission frequency. Therefore, the fading phenomena in radio communication are also frequency specific. If the bandwidth of the mobile radio channel is small (narrowband signal), then the whole frequency band of this channel is subject to the same propagation conditions, and the mobile radio channel is considered frequency-nonselective. On the other hand, if the bandwidth of a channel is large (broadband signal), the individual frequencies suffer from different degrees of fading (Figure 2.2) in which case we speak of a frequency-selective channel (David and Benkner, 1996; Steele, 1992). Signal breaks because of frequency-selective fading along a signal path are much less frequent for a broadband signal than for a narrowband signal, because the fading holes only shift within the band and the received total signal energy remains relatively constant (Bossert, 1991).

In addition to frequency-selective fading, the different propagation times of the individual multipath components also cause time dispersion on their propagation paths. Therefore, signal distortions can occur due to interference of one symbol with its neighboring symbols ('intersymbol interference'). These distortions depend first on the spread experienced by a

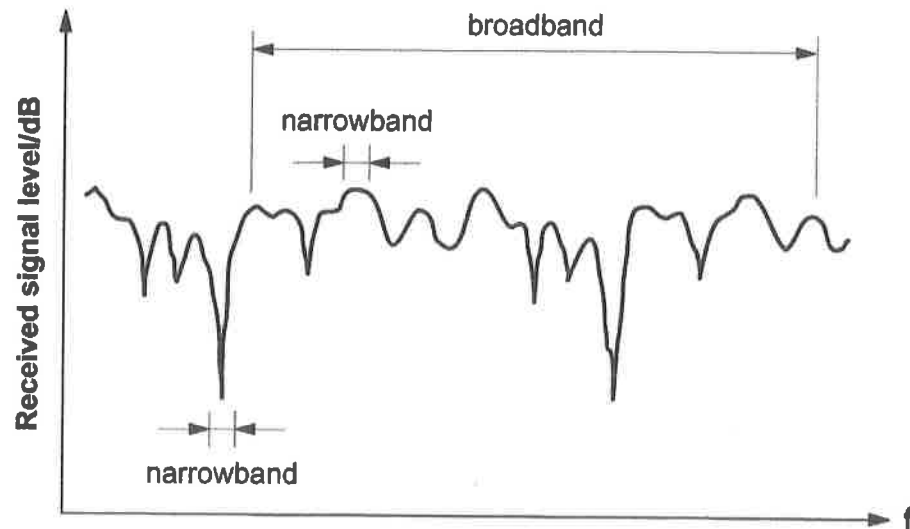


Figure 2.2 Frequency selectivity of a mobile radio channel.

pulse on the mobile channel, and second on the duration of the symbol or of the interval between symbols. Typical multipath channel delays range from $0.5 \mu\text{s}$ in urban areas to about 16 to $20 \mu\text{s}$ in hilly terrain, i.e. a transmitted pulse generates several echoes which reach the receiver with delays of up to $20 \mu\text{s}$. In digital mobile radio systems with typical symbol durations of a few microseconds, this can lead to smearing of individual pulses over several symbol durations.

Owing to the described effects of the wireless channel, mobile information transport requires additional, often very extensive measures, which compensate for the effects of multipath propagation. First, an equalizer is required, which attempts to eliminate the signal distortions caused by intersymbol interference. The operational principle of such an equalizer for mobile radio is based on the estimation of the channel pulse response to periodically transmitted, well-known bit patterns, known as the training sequences (Bertsekas and Gallager, 1987; Watson, 1993). This allows the time dispersion of the channel and its compensation to be determined. The performance of the equalizer has a significant effect on the quality of the digital transmission. On the other hand, for efficient transmission in digital mobile radio, channel coding measures are indispensable, such as forward error correction with error-correcting codes, which allows the effective bit error ratio to be reduced to a tolerable value (about 10^{-5} to 10^{-6}). Further important measures are transmitter power control and algorithms for the compensation of signal interruptions in fading, which may be of such a short duration that a disconnection of the call would not be appropriate.

2.2 Separation of directions and duplex transmission

The most frequent form of communication is the bidirectional communication which allows simultaneous transmitting and receiving. A system capable of doing this is called full-duplex. One can also achieve full-duplex capability if sending and receiving do not occur simultaneously but switching between both phases is done so fast that it is not noticed

by the user, i.e. both directions can be used quasi-simultaneously. Modern digital mobile radio systems are always full-duplex capable. Essentially, two basic duplex procedures are employed: Frequency Division Duplex (FDD) using different frequency bands in each direction, and Time Division Duplex (TDD) which periodically switches the direction of transmission.

2.2.1 Frequency Division Duplex

The frequency duplex procedure has been used already in analog mobile radio systems and is also used in digital systems. For communication between a mobile and a base station, the available frequency band is split into two partial bands, to enable simultaneous sending and receiving. One partial band is assigned for *uplink* (from mobile to base station) transmissions and the other partial band is assigned for *downlink* (from base station to mobile) transmissions.

- Uplink band: transmission band of the mobile *and* receiving band of the base station.
- Downlink band: receiving band of the mobile *and* transmission band of the base station.

To achieve good separation of both directions, the partial bands must be a sufficient frequency distance apart, i.e. the frequency pairs of a connection assigned to uplink and downlink must have this distance band between them. Usually, the same antenna is used for sending and receiving. A duplexing unit is then used for the directional separation, consisting essentially of two narrowband filters with steep flanks (Figure 2.3). These filters, however, cannot be integrated, so pure frequency duplexing is not appropriate for systems with small compact equipment (David and Benkner, 1996).

2.2.2 Time Division Duplex

Time duplexing is therefore a good alternative, especially in digital systems with time division multiple access. In this case, the transmitter and receiver operate only quasi-simultaneously at different points in time, i.e. the directional separation is achieved by switching in time between transmission and reception, and thus no duplexing unit is required. Switching occurs frequently enough that the communication appears to be over a quasi-simultaneous full-duplex connection. However, out of the periodic interval T available for the transmission of a time slot only a small part can be used, so that a time duplex system requires more than twice the bit rate of a frequency duplex system.

2.3 Multiple access

The radio channel is a communication medium shared by many subscribers in one cell. Mobile stations compete with one another for the frequency resource to transmit their information streams. Without any other measures to control simultaneous access of several users, collisions can occur (multiple access problem). Since collisions are very undesirable for a connection-oriented communication like mobile telephony, the individual subscribers/mobile stations must be assigned dedicated channels on demand. In order to divide the available physical resources of a mobile system, i.e. the frequency bands, into voice channels, special multiple access procedures are used which are presented in the following (Figure 2.4).

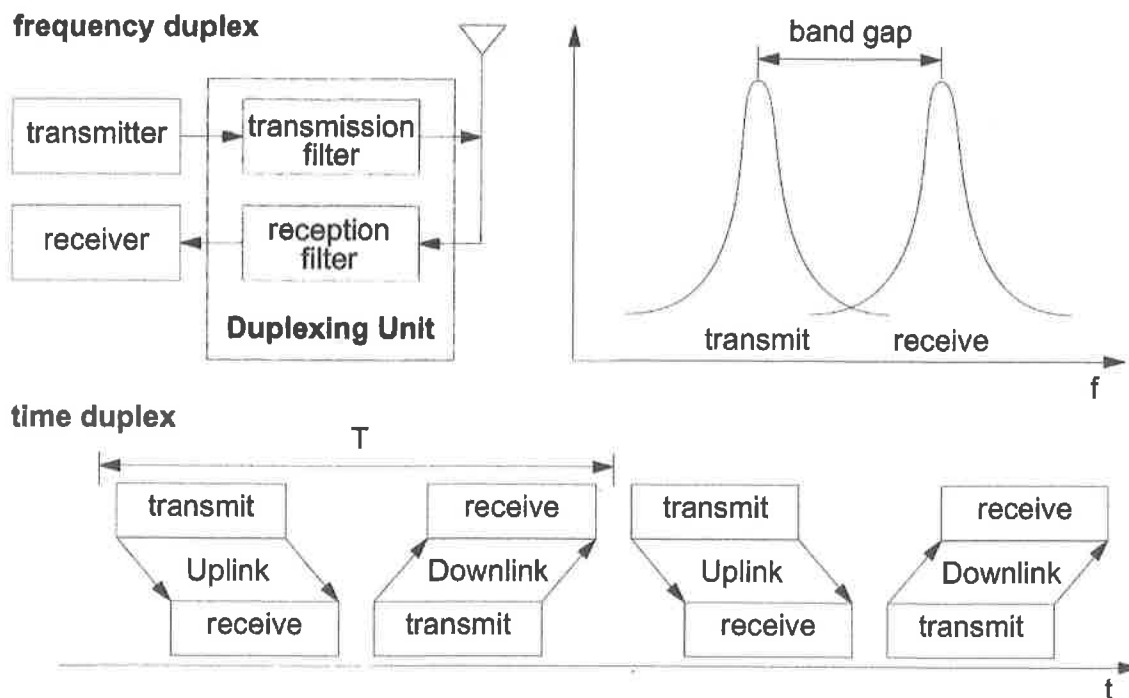


Figure 2.3 Frequency and time duplex.

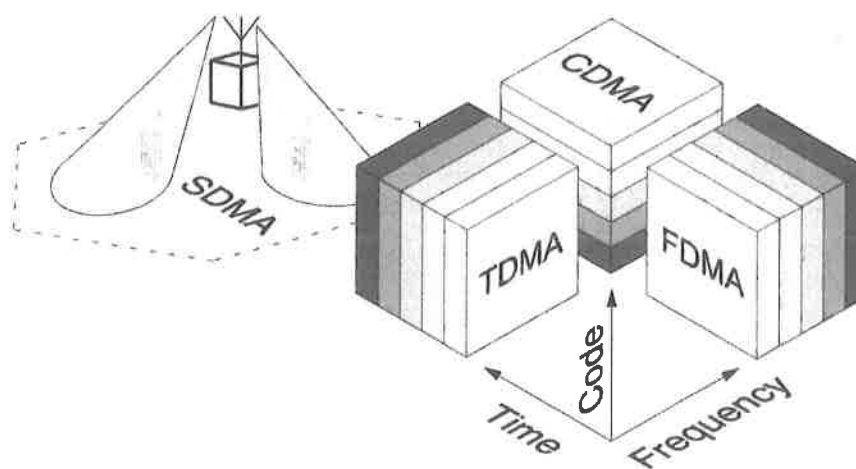


Figure 2.4 Multiple access procedures.

2.3.1 Frequency Division Multiple Access

Frequency Division Multiple Access (FDMA) is one of the most common multiple access procedures. The frequency band is divided into channels of certain bandwidth such that each conversation is carried on a different frequency (Figure 2.5). The effort in the base station to realize an FDMA system is very high. Even though the required hardware components are

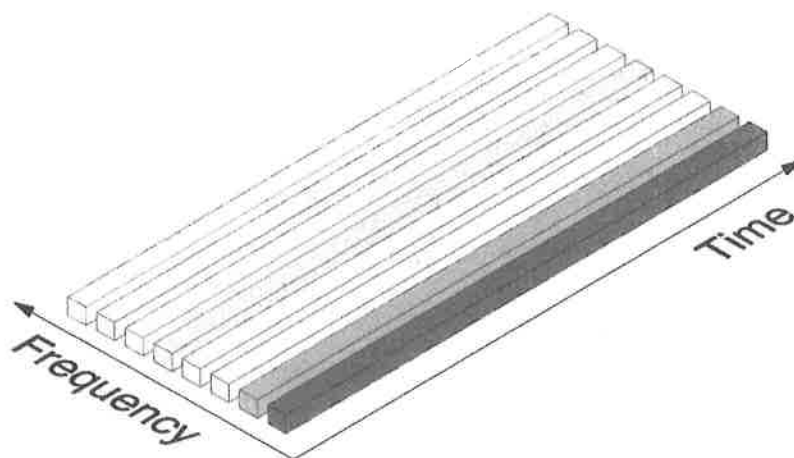


Figure 2.5 Channels of an FDMA system.

relatively simple, each channel needs its own transceiving unit. Furthermore, the tolerance requirements for the high-frequency networks and the linearity of the amplifiers in the transmitter stages of the base station are quite high, since a large number of channels need to be amplified and transmitted together (David and Benkner, 1996; Steele, 1992). One also needs a duplexing unit with filters for the transmitter and receiver units to enable full-duplex operation, which makes it hard to build small, compact mobile stations, since the required narrowband filters can hardly be realized with integrated circuits.

2.3.2 Time Division Multiple Access

Time Division Multiple Access (TDMA) is used in digital mobile radio systems. The individual mobile stations are cyclically assigned a frequency for exclusive use only for the duration of a time slot, which obviously requires frame synchronization between transmitter and receiver. Furthermore, in most cases the whole system bandwidth for a time slot is not assigned to one station, but the system frequency range is subdivided into subbands, and TDMA is used for multiple access to each subband. The subbands are known as carrier frequencies, and the mobile systems using this technique are designated as multicarrier systems (not to be confused with multicarrier modulation). GSM employs such a combination of FDMA and TDMA; it is a multicarrier TDMA system. The available frequency range is divided into frequency channels of 200 kHz bandwidth each (with guard bands between to ease filtering), with each of these frequency channels containing eight TDMA conversation channels.

Thus, the sequence of time slots assigned to a mobile station represents the physical channels of a TDMA system. In each time slot, the mobile station transmits a data burst. The period assigned to a time slot for a mobile station thus also determines the number of TDMA channels on a carrier frequency. The time slots of one period are combined into a so-called TDMA frame. Figure 2.6 shows five channels in a TDMA system with a period of four time slots and three carrier frequencies.

The TDMA signal transmitted on a carrier frequency in general requires more bandwidth than an FDMA signal; this is because with multiple time use, the gross data rate has to be

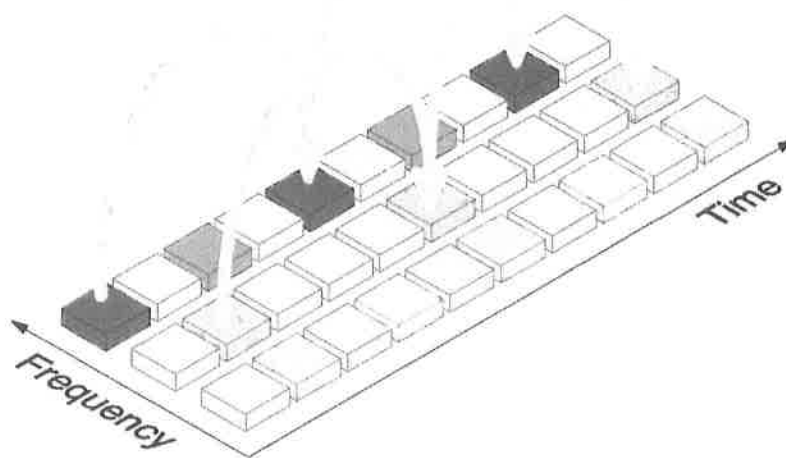


Figure 2.6 TDMA channels on multiple carrier frequencies.

correspondingly higher. For example, GSM systems employ a gross data rate (modulation data rate) of 271 kbit/s on a subband of 200 kHz, which amounts to 33.9 kbit/s for each of the eight time slots.

Narrowband systems are particularly susceptible to frequency-selective fading (Figures 2.1 and 2.2) as already mentioned, such that a single channel might be in a deep fade while switching to another channel might result in a significantly better reception. Furthermore, there are also frequency-selective co-channel interferences, which can contribute to the deterioration of the transmission quality. To this end a TDMA system offers very good opportunities to attack and drastically reduce such frequency-selective interference by introducing a frequency hopping technique. With this technique, each burst of a TDMA channel is transmitted on a different frequency (Figure 2.7).

In this technique, selective interference on one frequency at worst hits only every i^{th} time slot, if there are i frequencies available for hopping. Thus, the signal transmitted by a frequency hopping technique uses frequency diversity. Of course, the hopping sequences must be orthogonal, i.e. one must ascertain that two stations transmitting in the same time slot do not use the same frequency. Since the duration of a hopping period is long compared with the duration of a symbol, this technique is called slow frequency hopping. With fast frequency hopping, the hopping period is shorter than a time slot and is of the order of a single symbol duration or even less. This technique belongs to the family of spread spectrum techniques. As mentioned above, for TDMA, synchronization between a mobile and base station is necessary. This synchronization becomes even more complex due to the mobility of the subscribers, because they can stay at varying distances from the base station and their signals thus incur varying propagation times. First, the basic problem is determining the exact moment when to transmit. This is typically achieved by using one of the signals as a time reference, such as the signal from the base station (downlink, Figure 2.8). On receiving the TDMA frame from the base station, the mobile can synchronize and transmit a time slot

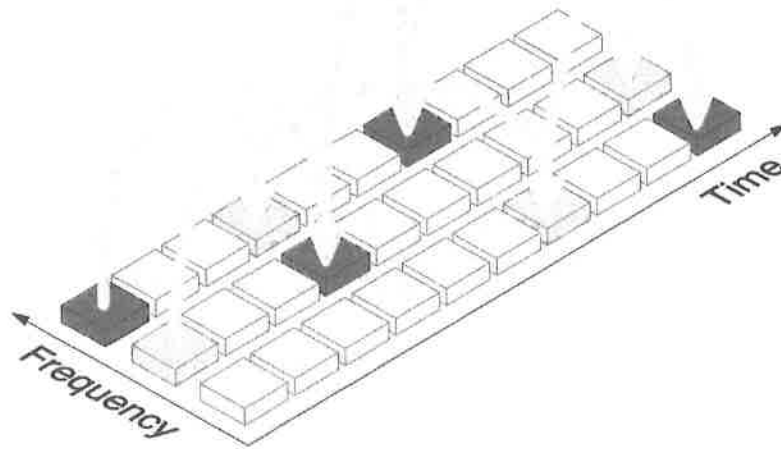


Figure 2.7 TDMA with the use of a frequency hopping technique.

synchronously with an additional time offset (e.g. three time slots in Figure 2.8). Another problem is the propagation time of the signals, ignored up to now. It also depends on the variable distance of the mobile from the base station. These propagation times are the reason why the signals that arrive on the uplink are not frame-synchronized at the base, but have variable delays. If these delays are not compensated, collisions of adjacent time slots can occur (Figure 2.8). In principle, the mobile stations must therefore advance the offset time between reception and transmission, i.e. the start of sending, so that the signals arrive in a frame-synchronous manner at the base station.

2.3.3 Code Division Multiple Access

Systems with Code Division Multiple Access (CDMA) are broadband systems, in which each subscriber uses the whole system bandwidth (similar to TDMA) for the complete duration of the connection (similar to FDMA). However, usage is not exclusive, i.e. all of the subscribers in a cell use the same frequency band simultaneously. To separate the signals, the subscribers are assigned orthogonal codes. The basis of CDMA is a band-spreading or spread spectrum technique. The signal of one subscriber is spread spectrally over a multiple of its original bandwidth. Typically, spreading factors are between 10 and 1000; they generate a broadband signal for transmission from the narrowband signal, and this is less sensitive to frequency-selective interference and disturbances. Furthermore, the spectral power density is decreased by band spreading, and communication is even possible below the noise threshold (David and Benkner, 1996).

Direct sequence CDMA

A common spread-spectrum procedure is the direct sequence technique (Figure 2.9). Here, each bit of the original data sequence is multiplied directly – before modulation – with a

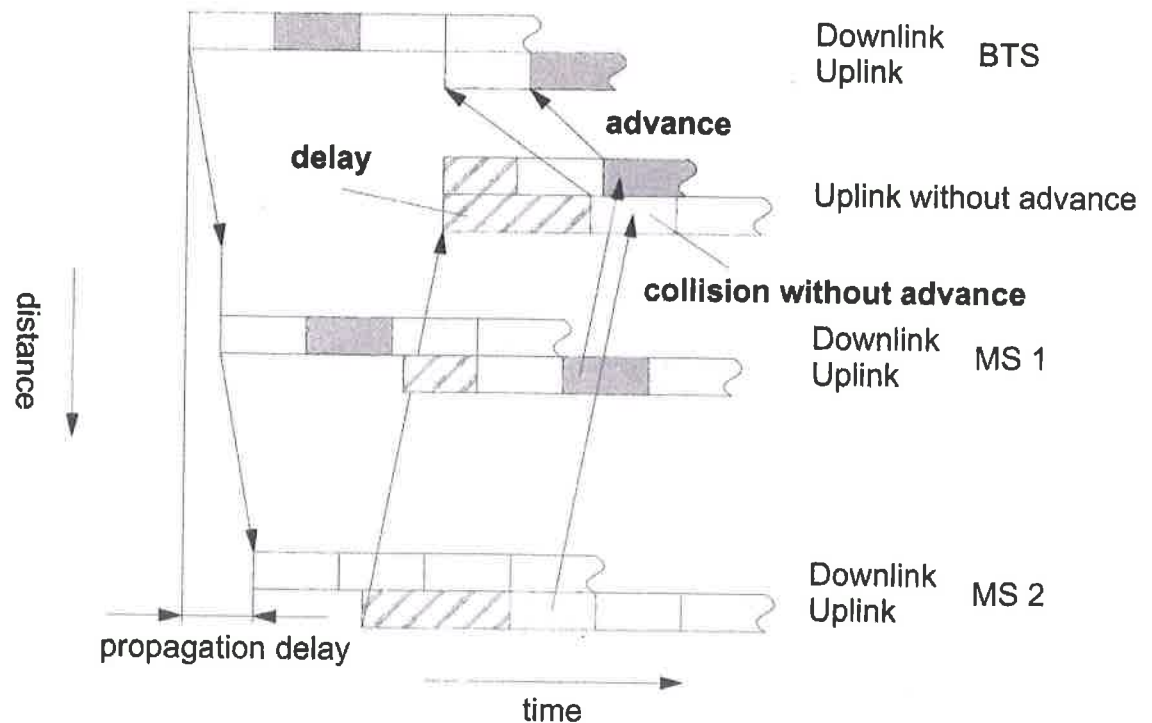


Figure 2.8 Differences in propagation delays and synchronization in TDMA systems.

spreading sequence or part of a spreading sequence to generate the spread spectrum signal. The length of the spreading sequence with respect to the duration of an original data bit defines the Spreading Factor (SF). Consequently, the transmission rate after spreading, which is called the chip rate, is larger than the original data bit rate (by the SF). In a multiple access system, the spread signals of many users are superimposed and at the receiver, the composite signal is correlated with the user specific spreading sequences in order to retrieve the original user data sequence (Figure 2.10).

Thus, if direct sequence spreading is used, the multiple access scheme is called Direct Sequence CDMA (DS-CDMA).

Frequency hopping CDMA

Another possibility for spreading the band is the use of a fast frequency hopping technique. If one changes the frequency several times during one transmitted data symbol, a similar spreading effect occurs as in the case of the direct sequence procedure. If the frequency hopping sequence is again controlled by orthogonal code sequences, another multiple access system can be realized, the Frequency Hopping CDMA (FH-CDMA).

2.3.4 Space Division Multiple Access

An essential property of the mobile radio channel is multipath propagation, which leads to frequency-selective fading phenomena. Furthermore, multipath propagation is the cause of another significant property of the mobile radio channel, the spatial fanning-out of signals.

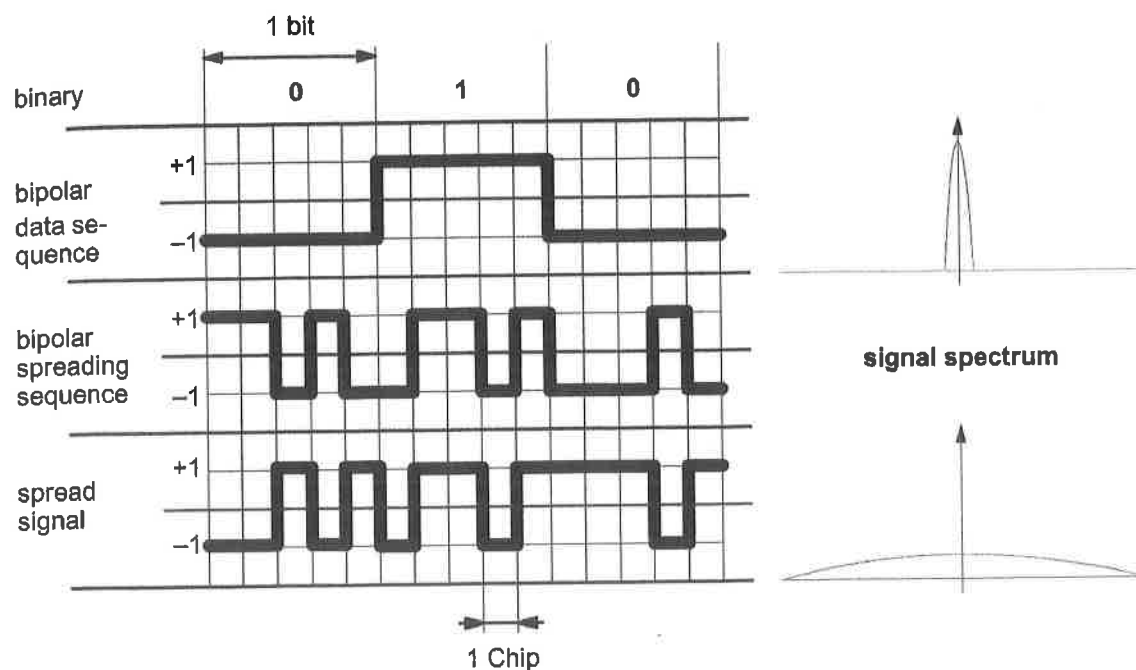


Figure 2.9 Principle of spread spectrum technique for direct sequence CDMA.

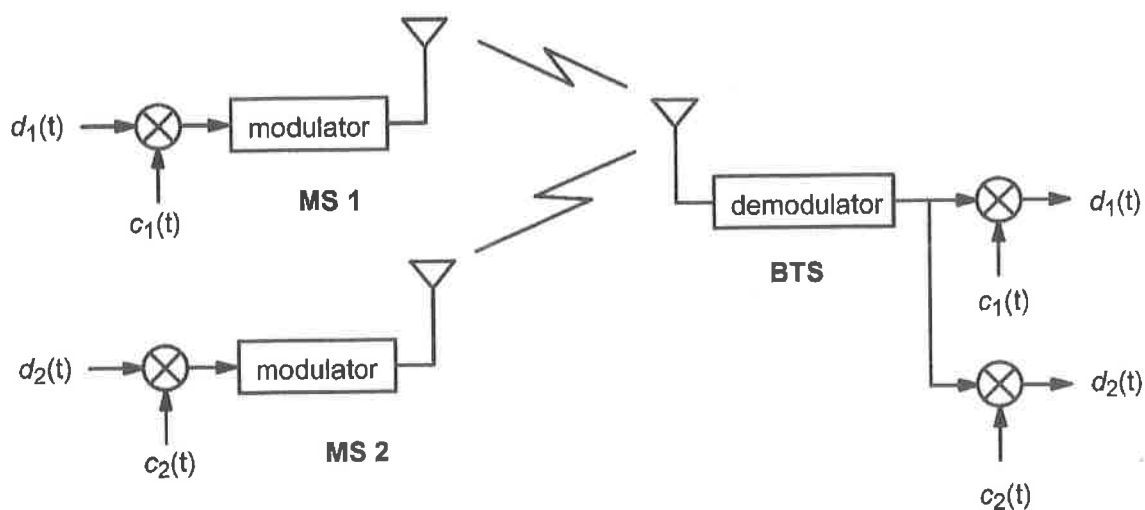


Figure 2.10 Simplified scheme of CDMA (uplink).

This causes the received signal to be a summation signal, which is not only determined by the Line of Sight (LOS) connection but also by an undetermined number of individual paths caused by refractions, infractions and reflections. In principle, the directions of incidence of these multipath components could therefore be distributed arbitrarily at the receiver. In particular, on the uplink from the mobile station to the base station, there is, however, in most

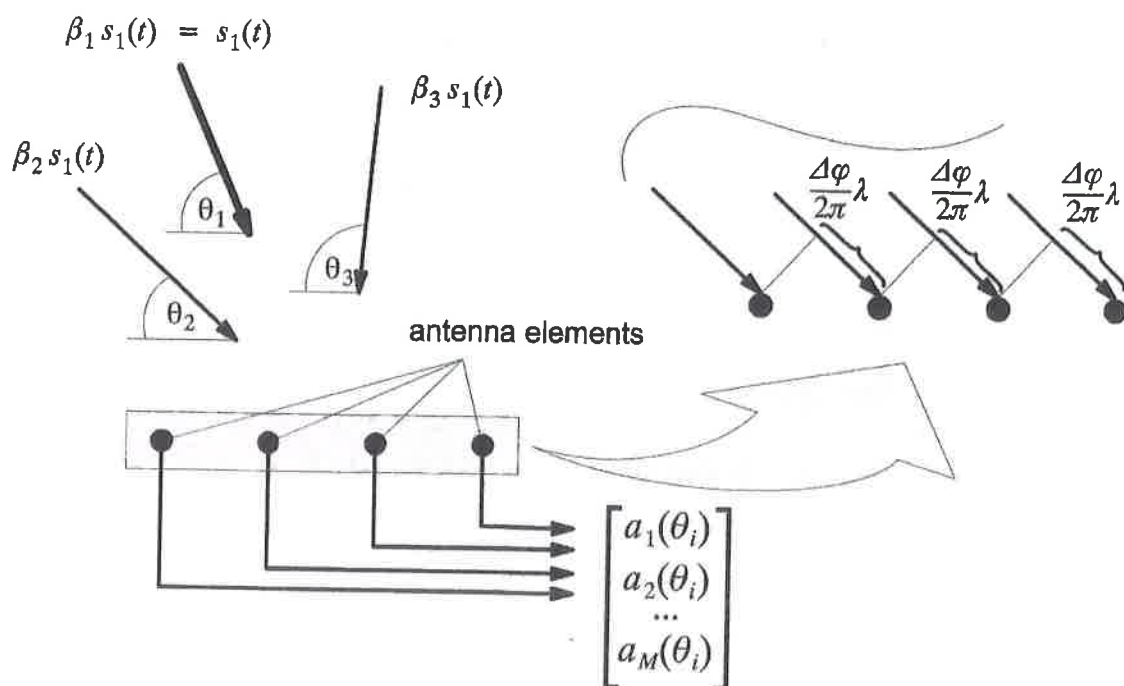


Figure 2.11 Multipath signal at an antenna array.

cases a main direction of incidence (usually LOS), about which the angles of incidence of the individual signal components are scattered in a relatively narrow range. Frequently, the essential signal portion at the receiver is distributed only over an angle of a few degrees. This is because base stations are installed wherever possible as free-standing units, and there are no interference centers in the immediate neighborhood. This directional selectivity of the mobile radio channel, which exists in spite of multipath propagation, can be exploited by using antenna arrays. These generate a directional characteristic by controlling the phases of the signals from the individual antenna elements. This allows the receiver to adjust the antenna selectively to the main direction of incidence of the received signal, and conversely to transmit selectively in one direction. This principle can be illustrated easily with a simple model (Figure 2.11).

The individual multipath components $\beta_i s_1(t)$ of a transmitted signal $s_1(t)$ propagate on different paths such that the multipath components incident at an antenna under the angle θ_i differ in amplitude and phase. If one considers an array antenna with M elements ($M = 4$ in Figure 2.11) and a wave front of a multipath component incident at angle θ_i on this array antenna, then the received signals at the antenna elements differ mainly in their phase – each shifted by $\Delta\phi$ (Figure 2.11) – and amplitude.

In this way, the response of the antenna to a signal incident at angle θ_i can be characterized by the complex response vector $\mathbf{a}(\theta_i)$ which defines amplitude gain and phase of each antenna element relative to the first antenna element ($a_1 = 1$):

$$\mathbf{a}(\theta_i) = [a_1(\theta_i) \ a_2(\theta_i) \ \dots \ a_M(\theta_i)]^T = [1 \ a_2(\theta_i) \ \dots \ a_M(\theta_i)]^T, \quad (2.7)$$

where the superscript T denotes the transpose of a vector. The N_m multipath components ($N_m = 3$ in Figure 2.11) of a signal $s_1(t)$ generate, depending on the incidence angle θ_i , a received signal vector $\mathbf{x}_1(t)$ which can be written with the respective antenna response vector and the signal of the i th multipath $\beta_i s_1(t)$ shifted in amplitude and phase against the direct path $s_1(t)$ as

$$\mathbf{x}_1(t) = \mathbf{a}(\theta_1)s_1(t) + \sum_{i=2}^{N_m} \mathbf{a}(\theta_i)s_1(t) = \mathbf{a}_1 s_1(t). \quad (2.8)$$

In this case, the vector $\mathbf{a}_1$ is also designated the spatial signature of the signal $s_1(t)$ which remains constant as long as the source of the signal does not move and the propagation conditions do not change (Xu and Li, 1994). In a multi-access situation, there are typically several (N_q) sources. This yields the following result for the total signal at the array antenna, neglecting noise and interferences

$$\mathbf{x}(t) = \sum_{j=1}^{N_q} \mathbf{a}_j s_j(t). \quad (2.9)$$

From this summation signal, the signals of the individual sources are separated by weighting the received signals of the individual antenna elements with a complex factor (weight vector $\mathbf{w}_i$), which yields

$$\mathbf{w}_i^H \mathbf{a}_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \quad (2.10)$$

For the weighted summation signal (Xu and Li, 1994) we obtain

$$\mathbf{w}_i^H \mathbf{x}(t) = \sum_{j=1}^{N_q} \mathbf{w}_i^H \mathbf{a}_j s_j(t) = s_i(t). \quad (2.11)$$

Under ideal conditions, i.e. neglecting noise and interference, the signal $s_i(t)$ of a single source i can be separated from the summation signal of the array antenna by using an appropriate weight vector during signal processing. The determination of the optimal weight vector, however, is a nontrivial and computation-intensive task. Owing to the considerable processing effort and also because of the mechanical dimensions of the antenna field, array antennas are predominantly used in base stations. So far only the receiving direction has been considered. The corresponding principles, however, can also be used for constructing the directional characteristics of the transmitter. Assume symmetric propagation conditions in the sending and receiving directions, and assume the transmitted signals $s_i(t)$ are weighted with the same weight vector $\mathbf{w}_i$ as the received signal, before they are transmitted through the array antenna; then one obtains the following summation signal radiated by the array antenna:

$$\mathbf{y}(t) = \sum_{j=1}^{N_q} \mathbf{w}_j s_j(t) \quad (2.12)$$

and for the signal received on the i th opposite side, respectively:

$$\hat{s}_i(t) = \mathbf{a}_i^H \mathbf{y}(t) = \sum_{j=1}^{N_q} \mathbf{a}_i^H \mathbf{w}_j s_j(t) = s_i(t). \quad (2.13)$$

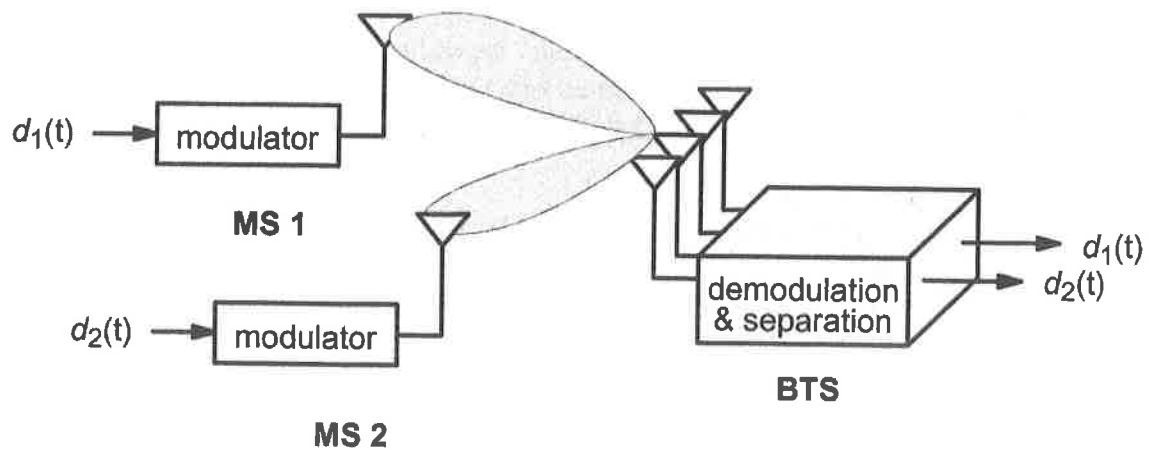


Figure 2.12 Schematic representation of spatial multiple access (uplink).

Thus, by using array antennas, one can separate the simultaneously received signals of spatially separated subscribers by exploiting the directional selectivity of the mobile radio channel. Owing to the use of intelligent signal processing and corresponding control algorithms, such systems are also known as systems with intelligent antennas. The directional characteristics of the array antenna can be controlled adaptively such that a signal is only received or transmitted in exactly the spatial segment where a certain mobile station is currently staying. On the one hand, one can thus reduce co-channel interference in other cells, and on the other hand, the sensitivity against interference can be reduced in the current cell. Furthermore, because of the spatial separation, physical channels in a cell can be reused, and the lobes of the antenna diagram can adaptively follow the movement of mobile stations. In this case, yet another multiple access technique (Figure 2.12) is defined and known as Space Division Multiple Access (SDMA). SDMA systems are currently the subject of intensive research. The SDMA technique can be combined with each of the other multiple access techniques (FDMA, TDMA, CDMA). This enables intracellular spatial channel reuse, which again increases the network capacity (Hartmann and Vögel, 1999). This is especially attractive for existing networks which can use an intelligent implementation of SDMA by selectively upgrading base stations with array antennas, appropriate signal processing and respective control protocols.

2.4 Cellular principle

Owing to the very limited frequency bands, a mobile radio network only has a relatively small number of speech channels available. For example, the GSM system has an allocation of 25 MHz bandwidth in the 900 MHz frequency range, which amounts to a maximum of 125 frequency channels each with a carrier bandwidth of 200 kHz. Within an eightfold time multiplex for each carrier, a maximum of 1000 channels can be realized. This number is further reduced by guardbands in the frequency spectrum and the overhead required for signaling. In order to be able to serve several hundreds of thousands or millions of subscribers in spite of this limitation, frequencies must be spatially reused, i.e. deployed repeatedly in a

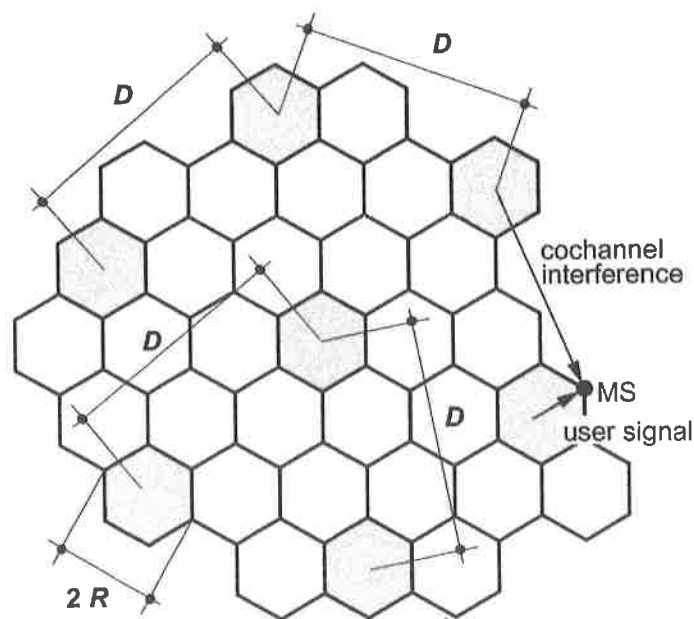


Figure 2.13 Model of a cellular network with frequency reuse.

geographic area. In this way, services can be offered with a cost-effective subscriber density and acceptable blocking probability.

2.4.1 Definitions

This spatial frequency reuse concept led to the development of the cellular principle, which allowed a significant improvement in the economic use of frequencies. The essential characteristics of the cellular network principle are as follows.

- The area to be covered is subdivided into *cells* (radio zones). These cells are often modeled in a simplified way as hexagons (Figure 2.12). The base station is located at the center of each cell.
- To each cell i a subset of the frequencies S_i is assigned from the total set (bundle) which is assigned to the respective mobile radio network. In the GSM system, the set S_i of frequencies assigned to a cell is called the Cell Allocation (CA). Two neighboring cells must never use the same frequencies, since this would lead to severe co-channel interference from the adjacent cells.
- Only at distance D (the frequency reuse distance) can a frequency from the set S_i be reused (Figure 2.13), i.e. cells with distance D to cell i can be assigned one or all of the frequencies from the set S_i belonging to cell i . When designing a mobile radio network, D must be chosen sufficiently large, such that the co-channel interference remains small enough not to affect speech quality.

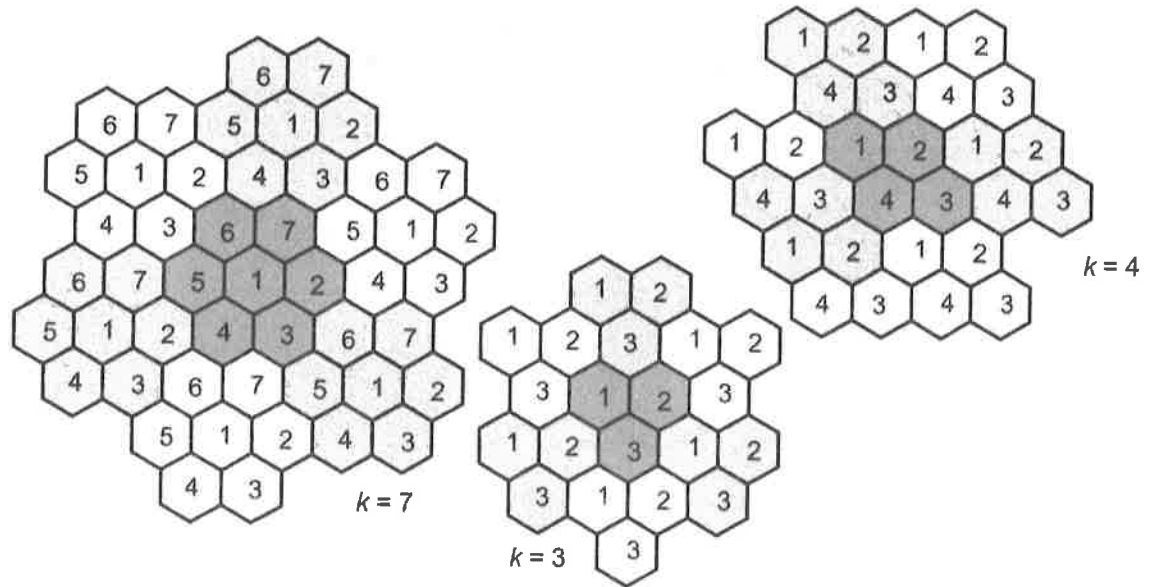


Figure 2.14 Frequency reuse and cluster formation.

- When a mobile station moves from one cell to another during an ongoing conversation, an automatic channel/frequency change occurs (*handover*), which maintains an active speech connection over cell boundaries.

The spatial repetition of frequencies is done in a regular systematic way, i.e. each cell with the cell allocation S_i sees its neighbors with the same frequencies again at a distance D (Figure 2.13). Therefore, there exist exactly six such next neighbor cells. Independent of form and size of the cells – not only in the hexagon model – the first ring in the frequency set contains six co-channel cells (see also Figure 2.14).

2.4.2 Carrier-to-interference ratio

The signal quality of a connection is measured as a function of received useful signal power and interference power received from co-channel cells and is given by the Carrier-to-Interference Ratio (CIR or C/I):

$$\frac{C}{I} = \frac{\text{Useful signal power}}{\text{Disturbing signal power}} = \frac{\text{Useful signal power}}{\text{Interference power from other cells}}. \quad (2.14)$$

The intensity of the interference is essentially a function of co-channel interference depending on the frequency reuse distance D . From the viewpoint of a mobile station, the co-channel interference is caused by base stations at a distance D from the current base station. A worst-case estimate for the CIR of a mobile station at the border of the covered area at distance R from the base station can be obtained, subject to propagation losses, by assuming that all six neighboring interfering transmitters operate at the same power and are approximately equally far apart (a distance D that is large compared with the cell radius R)

(Lee, 1989):

$$\frac{C}{I} = \frac{P_t R^{-\gamma}}{\sum_{i=1}^6 P_i} \approx \frac{P_t R^{-\gamma}}{\sum_{i=1}^6 P_i D^{-\gamma}} = \frac{P_t R^{-\gamma}}{6 P_i D^{-\gamma}}, \quad (2.15)$$

where it is implied, that the mean received power according to equation (2.4) is applied, and P_i is the power received from the i th interfering cell. Finally, we find the worst-case CIR as a function of the cell radius R , the reuse distance D and the attenuation exponent γ as

$$\frac{C}{I} = \frac{R^{-\gamma}}{6 D^{-\gamma}} = \frac{1}{6} \left(\frac{R}{D} \right)^{-\gamma}. \quad (2.16)$$

Therefore, in a given radio environment, the CIR depends essentially on the ratio R/D . From these considerations it follows that for a desired or required CIR value at a given cell radius, one must choose a minimum distance for the frequency reuse, above which the co-channel interference falls below the required threshold.

2.4.3 Formation of clusters

The regular spatial repetition of frequencies results in a clustering of cells. The cells within a cluster must each be assigned different sets of channels, while cells belonging to neighboring clusters can reuse the channels in the same spatial pattern. The size of a cluster is characterized by the number of cells per cluster k , which determines the frequency reuse distance D , when the cell radius R is given. Figure 2.14 shows some examples of clusters. The numbers designate the respective frequency sets $\mathcal{S}_i$ used within the single cells.

For each cluster the following holds.

- A cluster can contain all of the frequencies of the mobile radio system.
- Within a cluster, no frequency can be reused. The frequencies of a set $\mathcal{S}_i$ may be reused at the earliest in the neighboring cluster.
- The larger a cluster is, the larger the frequency reuse distance and the larger the CIR. However, the larger the values of k , the smaller the number of channels and the number of supportable active subscribers per cell.

The frequency reuse distance D can be derived geometrically from the hexagon model depending on k and the cell radius R :

$$D = R\sqrt{3k}. \quad (2.17)$$

The CIR is then given by

$$\frac{C}{I} = \frac{R^{-\gamma}}{6 D^{-\gamma}} = \frac{R^{-\gamma}}{6 (R\sqrt{3k})^{-\gamma}} = \frac{1}{6} (3k)^{\gamma/2}. \quad (2.18)$$

Allying this result, we can now determine the optimal cluster size given a required signal quality $(C/I)_{\min}$ according to the following formula:

$$\min_{i,j \geq 0} \left\{ k = i^2 + ij + j^2 \mid (C/I)_{\min} \leq \frac{1}{6} (3k)^{\gamma/2} \right\}. \quad (2.19)$$

According to measurements one can assume that, for good speech understandability, a CIR of about $(C/I)_{\min} = 18$ dB is sufficient. Assuming an approximate propagation coefficient of $\gamma = 4$, this yields the minimum cluster size

$$10 \log \left(\frac{C}{I} \right)_{\min} = 18 \text{ dB} \iff \left(\frac{C}{I} \right)_{\min} = 63.1, \quad (2.20)$$

$$\frac{1}{6}(3k)^{\gamma/2} \geq \left(\frac{C}{I} \right)_{\min} = 63.1 \implies k \geq 6.5 \implies k = 7. \quad (2.21)$$

These values are also confirmed by computer simulations, which have shown that for $(C/I)_{\min} = 18$ dB a reuse distance $D = 4.6R$ is required (Lee, 1989). In practically implemented networks, one can find other cluster sizes, e.g. $k = 3$ or 12.

The cellular models mentioned so far are very idealized for illustration and analysis. In reality, cells are neither circular nor hexagonal; rather they possess very irregular forms and sizes because of variable propagation conditions. An example of a possible cellular plan for a real network is shown in Figure 2.15, where one can easily recognize the individual cells with the assigned channels and the frequency reuse. The different cell sizes, which depend on whether it is an urban, suburban or rural area, are especially obvious. Figure 2.15 gives an impression of the approximate contours of equal signal power around the individual base stations. In spite of this representation, the precise fitting of signal power contours remains an idealization. The cell boundaries are, after all, blurred and defined by local thresholds, beyond which the neighboring base station's signal is received stronger than the current signal.

2.4.4 Traffic capacity and traffic engineering

As already mentioned, the number of channels and thus the maximal traffic capacity per cell depends on the cluster size k . The following relation holds:

$$n_F = \frac{B_t}{B_c k}, \quad (2.22)$$

where n_F is the number of frequencies per cell, B_t is the total bandwidth of the system and B_c is the bandwidth of one channel (including a margin for guard bands).

The number n of channels per cell in FDMA systems equals the number of frequency channels resulting from the channel and system bandwidth:

$$n = n_F. \quad (2.23)$$

The number of channels per cell in a TDMA system is the number of frequency channels multiplied by the number of time slots per channel:

$$n = m n_F, \quad (2.24)$$

where m is the number of time slots per frame.

A cell can be modeled as a traffic-theoretical loss system with n servers (channels), assuming a call arrival process with exponentially distributed interarrival times (Poisson process), and another Poisson process as a server process. Arrival and server processes are also called Markov processes, hence such a system is known as an M/M/ n loss system

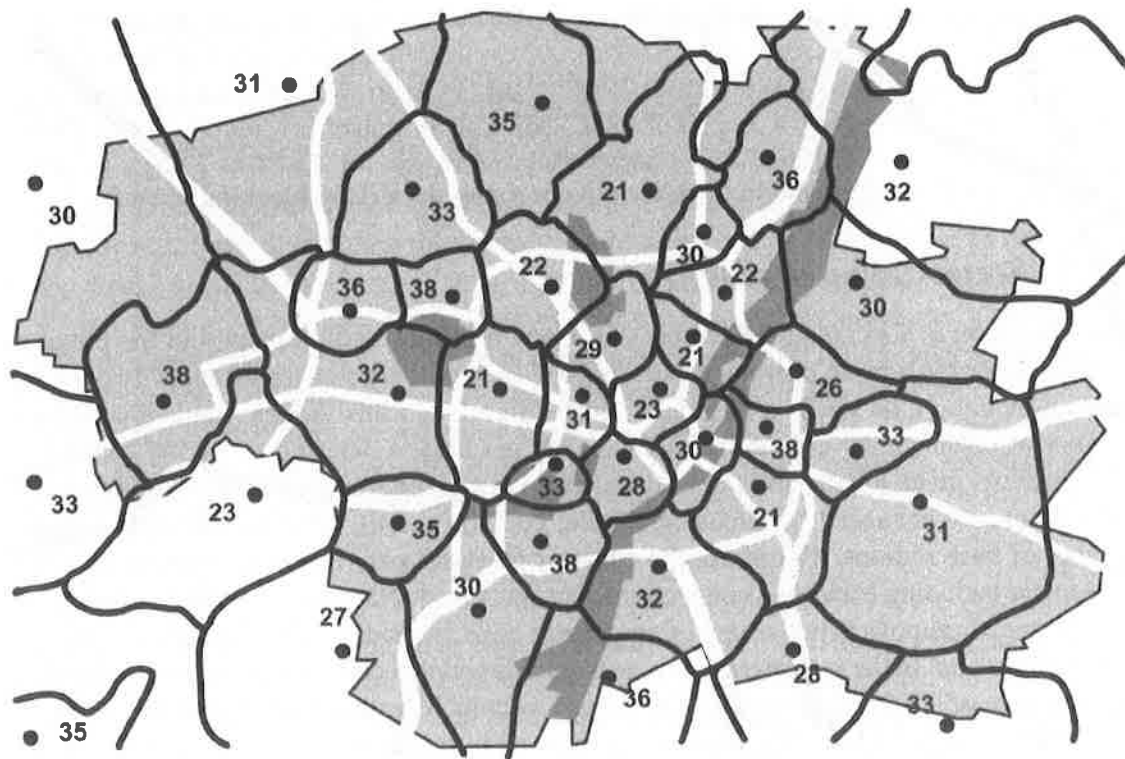


Figure 2.15 Realistic cell shapes.

(Kleinrock, 1975). For a given blocking probability B , a cell serves a maximum offered load $A_{\max}$ during the busy hour:

$$A_{\max} = f(B, n) = \lambda_{\max} T_m, \quad (2.25)$$

where $\lambda_{\max}$ is the busy hour call attempts (BHCA) and T_m is the mean call holding time. The relation between offered load A and blocking probability B with the total number of channels n is given by the Erlang blocking formula (see Kleinrock (1975) and Tran-Gia (1996) for more details and traffic tables):

$$B = \frac{A^n / n!}{\sum_{i=0}^n A^i / i!}. \quad (2.26)$$

However, these approximations are valid only for macrocellular environments, in which the number of users per cell is sufficiently large with regard to the number of available channels, such that the call arrival rate may be considered as approximately constant. For micro- and picocellular systems these assumptions usually no longer hold. Here, the traffic-theoretical dimensioning must be performed with Engset models, since the number of participants does not differ very much from the number of available channels. This results in a call arrival rate that is no longer constant. The probability that all channels are busy results from the number of users M per cell and the offer a of a non-busy user at

$$P_n = \frac{\binom{M}{n} a^n}{\sum_{i=0}^n \binom{M}{i} a^i}. \quad (2.27)$$

In this case, the probability that a call arrives when no free channels are available (blocking probability) is

$$P_B = \frac{\binom{M-1}{n} a^n}{\sum_{i=0}^n \binom{M-1}{i} a^i}. \quad (2.28)$$

For $M \rightarrow \infty$ the Engset blocking formula becomes the Erlang blocking formula.

2.4.5 Sectorization of cells

The discussion of CIR, required cluster size k and resulting traffic capacity in the previous sections indicated how an operator can design a system with a given traffic capacity per area: Since a required CIR yields a minimum cluster size k and this in turn yields a maximum number of channels per cell, the traffic capacity per area is determined by the cell radius R . Therefore, given a fixed amount of resources, i.e. a given frequency spectrum, the traffic capacity can be enhanced by choosing smaller cell sizes. This, however, requires a higher number of base stations. Therefore, the operator must not only invest in more base station equipment including backbone connections, but must also rent more base station locations. An alternative approach is the use of sectorized cells instead of omnidirectional cells, as considered so far. In the sectorized case, each omnidirectional cell is split into typically either three sectors of 120° or six sectors of 60° , each of which is supported by a sector antenna. Owing to the angular restriction of transmission and reception of each sector antenna, the number of interfering co-channel cells shrinks from six to only two co-channel cells in the case of three sectors and to just one co-channel cell in the case of six sectors. Therefore, we obtain an improved CIR of

$$\left(\frac{C}{I}\right)_{120^\circ} \approx 3 \cdot \left(\frac{C}{I}\right)_{\text{omni}} = \frac{1}{2} \left(\frac{R}{D}\right)^{-\gamma} = \frac{1}{2} (3k)^{\gamma/2} \quad (2.29)$$

in the case of three sectors per cell and

$$\left(\frac{C}{I}\right)_{60^\circ} \approx 6 \cdot \left(\frac{C}{I}\right)_{\text{omni}} = \left(\frac{R}{D}\right)^{-\gamma} = (3k)^{\gamma/2} \quad (2.30)$$

in the case of six sectors per cell. For a given required CIR, we can thus reduce the cluster size k to k^{120° and k^{60° for the case of three and six sectors, respectively, according to the following formulas:

$$\min_{i,j \geq 0} \left\{ k^{120^\circ} = i^2 + ij + j^2 \mid \left(\frac{C}{I}\right)_{\min} \leq \frac{1}{2} (3k)^{\gamma/2} \right\} \quad (2.31)$$

and

$$\min_{i,j \geq 0} \left\{ k^{60^\circ} = i^2 + ij + j^2 \mid \left(\frac{C}{I}\right)_{\min} \leq (3k)^{\gamma/2} \right\}. \quad (2.32)$$

Consequently we obtain a higher number of channels per cell. However, it is important to note that the channels assigned to each cell must now be divided among the sectors per cell. In particular, if the overall number of available channels is relatively low, this confinement of the channels to be used only within a sector leads to a reduced trunking efficiency, which can limit the capacity gain. In fact, depending on the parameters, sectorization might in extreme

cases even *decrease* the capacity. The effect of sectorization on capacity is discussed in the following, using some examples.

Assuming a required CIR of $(C/I)_{\min} = 18$ dB, a cluster size of $k^{\text{omni}} = 7$ is necessary (as discussed above), according to equation (2.19) with $\gamma = 4$. Using equation (2.31), we obtain $(C/I)_{k=7}^{120^\circ} = 23$ dB for the same cluster size $k = 7$ and 120° sectorization. In the case of a reduced cluster size of $k = 4$, the resulting value is $(C/I)_{k=4}^{120^\circ} = 18.5$ dB, which is still sufficient. Further reducing the cluster size to $k = 3$ yields $(C/I)_{k=3}^{120^\circ} = 16$ dB, which is too low for our requirements. Thus, the cluster size can be reduced to $k^{120^\circ} = 4$, if 120° sectorization is applied. Introducing 60° sectorization even allows for a cluster size of $k^{60^\circ} = 3$, yielding $(C/I)_{k=3}^{60^\circ} = 19$ dB according to equation (2.32). In order to discuss the capacity improvement, achievable with sectorization, we derive the blocking probabilities. The blocking probabilities are obtained from the Erlang-B formula, which we denote as $B(a, n)$, where a is the offered traffic in Erlang and n the number of servers, i.e. channels. In case of omnidirectional cells, the blocking probability becomes

$$P_b^{\text{omni}} = B(A, N^{\text{sys}}/k^{\text{omni}}), \quad (2.33)$$

where A is the offered traffic to a cell in Erlang and N^{sys} is the number of channels available to the system. In the case of sectorization, it must be accounted for the reduced trunking efficiency. Therefore, the number of channels of a single sector as well as the traffic offered to each sector must be considered. Thus, the blocking probability in the case of 120° sectorization can be expressed as

$$P_b^{120^\circ} = B(A/3, N^{\text{sys}}/k^{120^\circ}/3) \quad (2.34)$$

and for 60° sectorization we obtain

$$P_b^{60^\circ} = B(A/6, N^{\text{sys}}/k^{60^\circ}/6). \quad (2.35)$$

Using the appropriate values of k in either case, we can print the respective blocking probabilities given the number of system channels N^{sys} . Results are depicted in Figure 2.16 for $N^{\text{sys}} = 84$ and in Figure 2.17 for $N^{\text{sys}} = 252$, where in Figure 2.16 it has been considered that in the case of 60° sectorization, two sectors have four channels available, while the other four sectors can use five channels.

From the results in Figure 2.16 ($N^{\text{sys}} = 84$) it can be observed that in the case of low traffic values 60° sectorization actually performs worse than 120° sectorization and even worse than the omnidirectional system, owing to the loss in trunking efficiency with only four or five channels available to each sector. With increasing traffic load, however, 60° sectorization starts to perform better than the other two options. Assuming a desired blocking performance of $P_{b,\min} = 1\%$, the capacity gain of sectorization can be obtained, which can be seen in Table 2.1. Interestingly, the gain of 60° sectorization in this case is lower than that of 120° sectorization. If the number of channels is increased ($N^{\text{sys}} = 252$), the gains in both cases increase significantly and the highest capacity is now obtained with 60° sectorization.

As a conclusion, it can be said that sectorization can increase the system capacity, however, the number of sectors should be chosen carefully depending on the number of channels available. The higher the number of channels in the system, the more we can benefit from sectorization in terms of capacity.

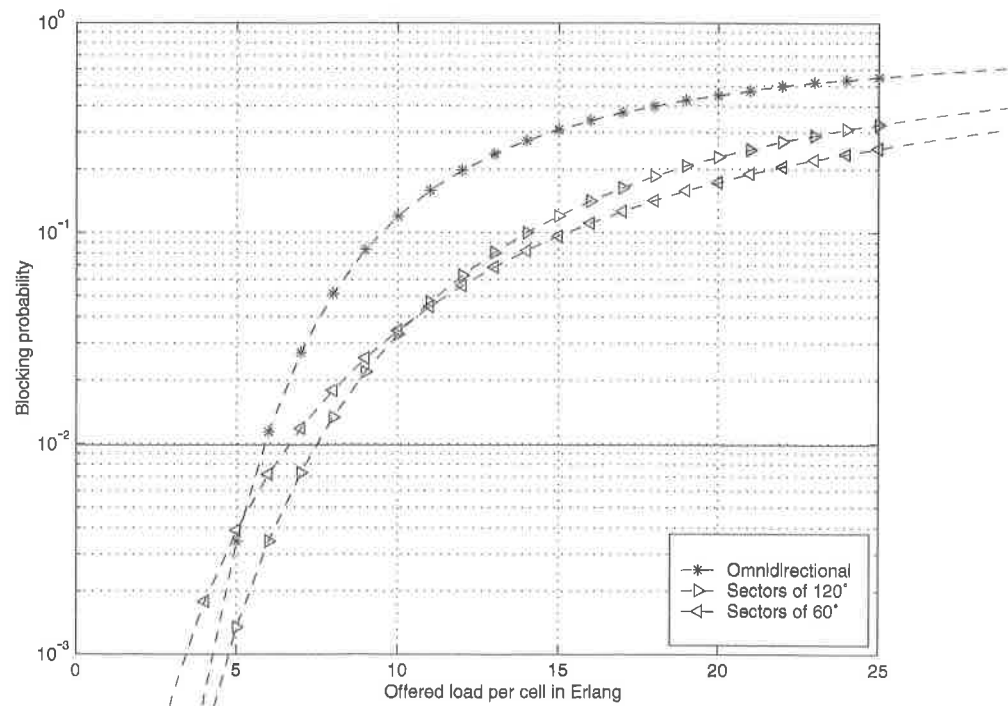


Figure 2.16 Blocking probabilities for omnidirectional and sectorized systems with 84 channels.

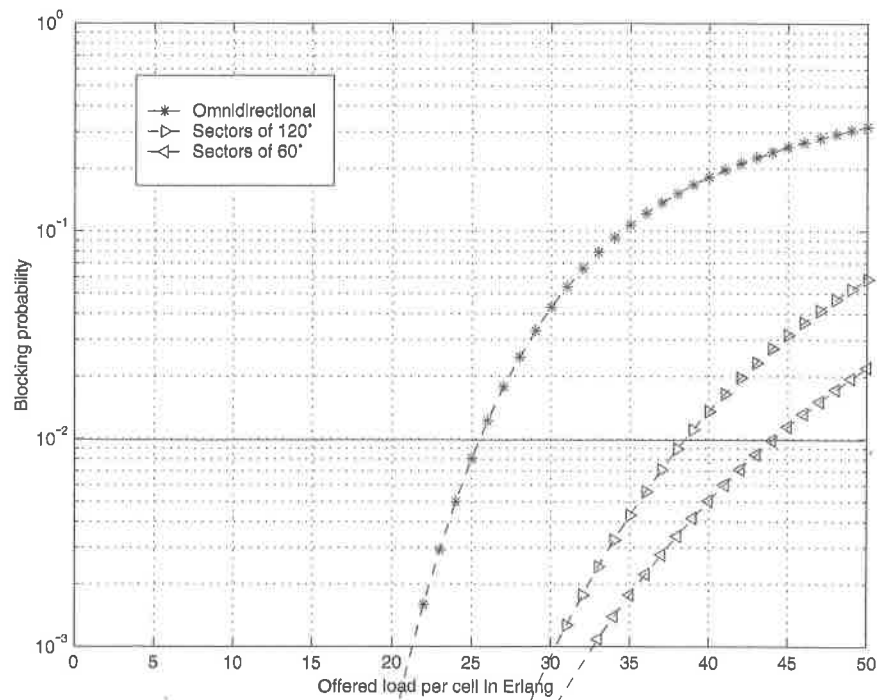


Figure 2.17 Blocking probabilities for omnidirectional and sectorized systems with 252 channels.

Table 2.1 Sectorization capacity gain with $(C/I)_{\min} = 18$ dB.

Sectors	$N^{\text{sys}} = 84$	$N^{\text{sys}} = 252$
120°	27.6%	50.6%
60°	17.8%	72.9%

2.4.6 Spatial filtering for interference reduction (SFIR)

Smart antennas at the base station can be used to perform adaptive beamforming for each individual user, yielding directional transmission on the downlink as well as directional reception on the uplink. Through this concept, the radiated and received interference power can be significantly reduced in a cellular system. One approach to apply adaptive beamforming for capacity enhancement is SDMA, as described in section 2.3.4. Another way to transform the interference reduction into actual traffic capacity gains is SFIR. While SDMA allows for intra-cellular channel reuse and has a very high potential for capacity improvement, it also demands for new advanced Radio Resource Management (RRM) methods, specifically for channel allocation and handover in order to maintain the orthogonality of spatial channels (Hartmann and Eberspächer, 2001). In contrast, SFIR does not allow for channel reuse within the cell and can thus be introduced without any modification of protocols or RRM algorithms. In an SFIR system, adaptive beams are steered towards each user; however, each user within one cell will be assigned a different traffic channel, i.e. a different FDMA or TDMA channel. This means that the user signals are separated by the orthogonal traffic channels, such that a spatial separation is not required. For this reason, intra-cell beam collisions do not have to be resolved by means of handovers. Additional RRM methods are not needed, since each traffic channel is just used once within the cell. However, how can we achieve a capacity gain with SFIR when channels are not reused within each cell? The answer is a tighter inter-cell channel reuse: owing to the reduced co-channel interference received from neighboring cells in an SFIR system, a lower reuse distance can be applied, tightening the frequency reuse and resulting in a smaller required cluster size. The lower cluster size will present the operator with a higher number of traffic channels per cell, which gives us the traffic capacity gain that we were looking for. To this end SFIR is similar to sectorization. However, there are two major differences: the advantage of SFIR as compared with sectorization is that there is no loss in terms of trunking efficiency, since channels assigned to a cell can be used, independent of the position of a user, everywhere within the cell. However, it must be considered that the interference reduction of SFIR is a statistical property. Thus, the average interference will be reduced, however, actual interference values at random time instants *may* be as high as without SFIR. The latter is the case if by chance all co-channel users are located within their cells such that their main beam pattern is steered towards the user in the center cell. In fact, the interference will fluctuate around an average value and cause outage when it surpasses a threshold value above the average value. Therefore, in order to assess the capacity of an SFIR system, we have to consider two aspects:

- average C/I improvement; and
- outage probability.

***C/I* improvement and blocking probability**

Let us discuss the average *C/I* improvement first. We assume a hexagonal cell layout with six co-channel cells in the first co-channel tier. The SFIR interference scenario is depicted in Figure 2.18 for the downlink reception of a user in the center cell¹. The antenna is modeled by a *brickwall* antenna pattern, which is defined by two parameters (Hartmann and Eberspächer, 2001): the angular width of the main lobe, denoted as β , and a constant sidelobe attenuation, denoted as D . Within the main lobe the antenna gain G is assumed to be constant as $G_0 = 1$. The probability that a user in the center cell is within the reception area of the main beam that is steered towards a user from one of the interfering cells is $p_i = \beta/2\pi$. Thus, on average, a user in the center cell will be in the main beam of a co-channel user of one of the co-channel cells $p_i \cdot 100\%$ of the time, only. Consequently, with probability $1 - p_i$ the user will not be in the main beam of a co-channel user of one of the co-channel cells. In the latter case the received interference will be attenuated by the sidelobe attenuation D produced by the beam pattern. Thus, the received mean interference in the center cell will be reduced by the factor $(p_i + (1 - p_i)/D)$. Therefore, the respective *C/I* value in the case of SFIR becomes

$$\left(\frac{C}{I}\right)^{\text{SFIR}} = \left(\frac{C}{I}\right)^{\text{omni}} \cdot \frac{1}{p_i + (1 - p_i)/D}, \quad (2.36)$$

where $(C/I)^{\text{omni}} = \frac{1}{6}(3k)^{\gamma/2}$ is the worst-case *C/I* of an omnidirectional cellular system with cluster size k and attenuation exponent γ (Eberspächer *et al.*, 2001). It is the worst-case *C/I* since it is based on the assumption that a channel is used in all co-channel cells at once. Likewise, we call $(C/I)^{\text{SFIR}}$, as defined in equation (2.36), the average worst-case *C/I* of an SFIR system. It is worst case in the sense that it is based on the assumption that a channel is used in all co-channel cells of the first co-channel tier in parallel. However, equation (2.36) represents the *C/I* averaged with respect to all possible beam directions of the co-channel users. With a given target $(C/I)_{\min}$ we can now determine the required cluster size of our SFIR system according to

$$\min_{i,j \geq 0} \left\{ k = i^2 + ij + j^2 \mid \left(\frac{C}{I}\right)_{\min} \leq \frac{\frac{1}{6}(3k)^{\gamma/2}}{p_i + (1 - p_i)/D} \right\}. \quad (2.37)$$

Depending on the beamforming properties β and D , the resulting cluster size k^{SFIR} will be smaller than in the omnidirectional case (k^{omni}). Re-assigning the channels accordingly will result in a higher capacity. Given N^{sys} channels available to the system we have $N^{\text{omni}} = N^{\text{sys}}/k^{\text{omni}}$ channels per cell in the omnidirectional case and $N^{\text{SFIR}} = N^{\text{sys}}/k^{\text{SFIR}} = N^{\text{omni}}(k^{\text{omni}}/k^{\text{SFIR}})$ channels per cell in the SFIR case. Using the Erlang-B formula (denoted by $B(a, n)$ where a is the traffic load in Erlang offered to the cell and n the number of channels per cell), we can compute the blocking probability in either case: $P_b^{\text{omni}} = B(A, N^{\text{omni}})$ and $P_b^{\text{SFIR}} = B(A, N^{\text{SFIR}})$. However, in contrast to sectorization, the blocking probability is not sufficient to assess the SFIR capacity. In addition, the outage probability, i.e. the probability that the actual (C/I) value will be worse than the target $(C/I)_{\min}$, must be considered.

¹We consider the downlink case here. The uplink case is less problematic, because the base station will receive at most one interfering user as long as $\beta < 60^\circ$ and at most two interferers if $60^\circ \leq \beta < 120^\circ$.

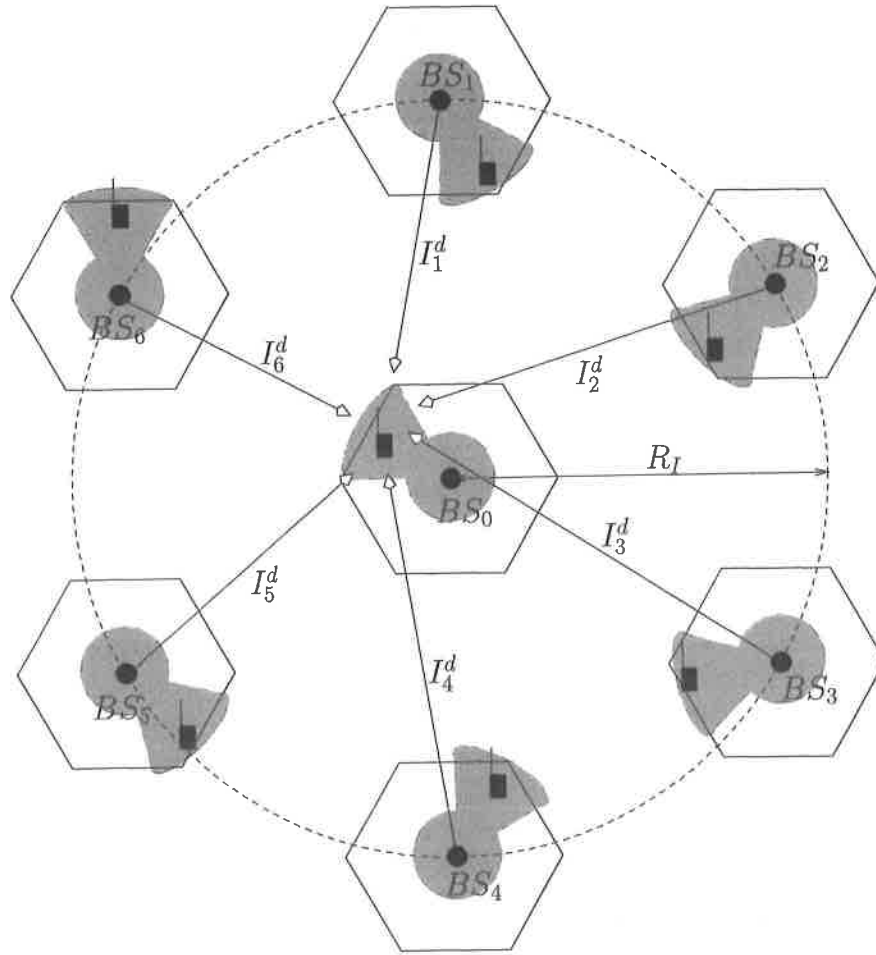


Figure 2.18 SFIR interference scenario downlink.

Outage probability

In order to determine the outage probability, we first determine a number $1 \leq n_{\text{out}} \leq 6$ such that the interference threshold of the center cell user will be surpassed, if at least n_{out} of the six co-channel users have the center cell user within their main beam – after the cluster size has been reduced. In other words, n_{out} is the lowest integer value n_i , for which the C/I value in the center cell becomes lower than the threshold value $(C/I)_{\text{min}}$ in

$$\left(\frac{C}{I}\right)^{\text{omni,new}} \cdot \frac{1}{n_i/6 + (1 - n_i/6)/D} \leq \left(\frac{C}{I}\right)_{\text{min}}, \quad (2.38)$$

with the worst case omnidirectional C/I after cluster size reduction: $(C/I)^{\text{omni,new}} = \frac{1}{6}(3k^{\text{SFIR}})^{\nu/2}$. Thus, with the ceiling function $\lceil \cdot \rceil$, n_{out} becomes

$$n_{\text{out}} = \left\lceil 6 \cdot \frac{(C/I)^{\text{omni,new}} / (C/I)_{\text{min}} - 1/D}{1 - 1/D} \right\rceil. \quad (2.39)$$

With n_{out} from equation (2.39) and $p_i = \beta/2\pi$ we can now determine the outage probability in case a channel is used in all six co-channel cells:

$$p(\text{out} | n_a = 6) = \sum_{n=n_{\text{out}}}^6 \binom{6}{n} p_i^n (1 - p_i)^{6-n}, \quad (2.40)$$

where n_a denotes the number of active co-channel cells. Likewise, we can determine the outage probability in the more general case that a channel is used in $n_a \leq 6$ cells of the co-channel cells:

$$p(\text{out} | n_a) = \begin{cases} \sum_{n=n_{\text{out}}}^{n_a} \binom{n_a}{n} p_i^n (1 - p_i)^{n_a-n} & n_{\text{out}} \leq n_a, \\ 0 & n_{\text{out}} > n_a. \end{cases} \quad (2.41)$$

With the probabilities $p(n_a)$ that n_a co-channel cells are using the same channel as the mobile in question in the target cell, we obtain the general outage probability

$$p(\text{out}) = \sum_{n_a=0}^6 p(\text{out} | n_a) \cdot p(n_a). \quad (2.42)$$

The values of $p(n_a)$ depend on the traffic load in the co-channel cells as well as on the number of channels available in each cell. However, they also depend on the channel assignment strategy used. Even if we assume Fixed Channel Allocation (FCA), the question remains whether channels are assigned in the same order in each cell or not. Let us assume that channels are assigned randomly, with no specific order. Now, what is the probability that a specific channel $i \in \{1, \dots, N\}$ is active in a cell? Therefore, we define the random variable i_a which is 1 if channel i is active in the cell and 0 else. Since channels are assigned randomly within a cell, the probability of a channel being used in a cell is independent of the channel i :

$$p(i_a = 1) = \sum_{j=1}^N p(j \text{ channels active}) \cdot \frac{j}{N}. \quad (2.43)$$

Therefore, the probability $p(j \text{ channels active})$ that $j \leq N$ channels are active in a cell is given by (Kleinrock, 1975)

$$p(j \text{ channels active}) = \frac{A^j / j!}{\sum_{k=0}^N (A^k / k!)}, \quad (2.44)$$

where A is the traffic load of the given cell. Thus, we obtain

$$p(i_a = 1) = \sum_{j=1}^N \frac{A^j / j!}{\sum_{k=0}^N (A^k / k!)} \cdot \frac{j}{N}. \quad (2.45)$$

We can now compute the probability $p(n_a)$ that a specific channel is in use in n_a co-channel cells as

$$p(n_a) = \binom{6}{n_a} p(i_a = 1)^{n_a} (1 - p(i_a = 1))^{6-n_a}, \quad (2.46)$$

where we assume that the traffic load is statistically equal in each co-channel cell.

Table 2.2 SFIR capacity gain with $(C/I)_{\min} = 14$ dB and $k^{\text{SFIR}} = 3$.

β	$(C/I)_{k=3}^{\text{SFIR}}$ (dB)	Capacity gain (%)	Limitation
90°	15.88	147.4	Outage
60°	17.02	268.4	Outage
45°	17.73	284.2	Blocking
30°	18.57	284.2	Blocking
20°	19.24	284.2	Blocking
15°	19.61	284.2	Blocking

With equations (2.42), (2.41), (2.46) and (2.45), we can now compute the outage probability $P_{\text{out}} = p(\text{out})$ of a user in the target cell which is using a specific channel. Given our assumption that channels are assigned in random order within each cell, this same measure is the outage probability for any active mobile in the target cell if the traffic load in each co-channel cell is A .

Numerical results

For a better understanding of the capacity of sectorized cellular networks, we discuss some numerical results in the following. We assume a target C/I value of $(C/I)_{\min} = 18$ dB, yielding a cluster size of $k^{\text{omni}} = 7$. Choosing the antenna parameters $D = 10$ dB and $\beta = 30^\circ$, we obtain $(C/I)_{k=7}^{\text{SFIR}} = (C/I)_{k=7}^{\text{omni}} + 7.6$ dB = 25.6 dB, when SFIR is introduced. To translate the improved C/I into capacity gain, the cluster size must be reduced. For $k = 3$ we obtain C/I values of $(C/I)_{k=3}^{\text{omni}} = 11$ dB and $(C/I)_{k=3}^{\text{SFIR}} = (C/I)_{k=3}^{\text{omni}} + 7.6$ dB = 18.6 dB, respectively. Since the resulting value of $(C/I)_{k=3}^{\text{SFIR}}$ is even slightly higher than the original value of $(C/I)_{k=7}^{\text{omni}}$, we could conclude that through SFIR and cluster size reduction with $k^{\text{SFIR}} = 3$ we can benefit from the capacity gain that can be achieved by channel re-arrangement. This channel re-arrangement yields $N^{\text{SFIR}} = 7/3 \cdot N^{\text{omni}}$ channels per cell if N^{omni} is the number of channels in the omnidirectional case, and for the blocking probabilities we have

$$P_b^{\text{omni}} = B(A, N^{\text{omni}}) \quad (2.47)$$

and

$$P_b^{\text{SFIR}} = B(A, N^{\text{omni}} \cdot k^{\text{omni}}/k^{\text{SFIR}}) = B(A, N^{\text{SFIR}}). \quad (2.48)$$

For $N^{\text{omni}} = 6$ channels per omnidirectional cell, yielding $N^{\text{SFIR}} = 7/3 \cdot N^{\text{omni}} = 14$ channels per SFIR cell, the respective blocking probabilities are plotted as a function of the offered traffic load per cell in Erlang in Figures 2.19, 2.20 and 2.21. As discussed above, in order to determine the SFIR capacity, also the outage probability must be considered. In Figures 2.19, 2.20 and 2.21 the outage probability is plotted for three different values of the target C/I , $(C/I)_{\min} \in \{14 \text{ dB}, 16 \text{ dB}, 18 \text{ dB}\}$ and for various values of β in each case.

Let us for the moment focus on $\beta = 30^\circ$, as chosen above. For a given target blocking probability of $P_b = 1\%$, we can obtain the carried traffic load $L_{\text{cell}}^{\text{car,omni}}$ and $L_{\text{cell}}^{\text{car,SFIR}}$ per cell in the omnidirectional case and in the SFIR case, respectively. We can now determine

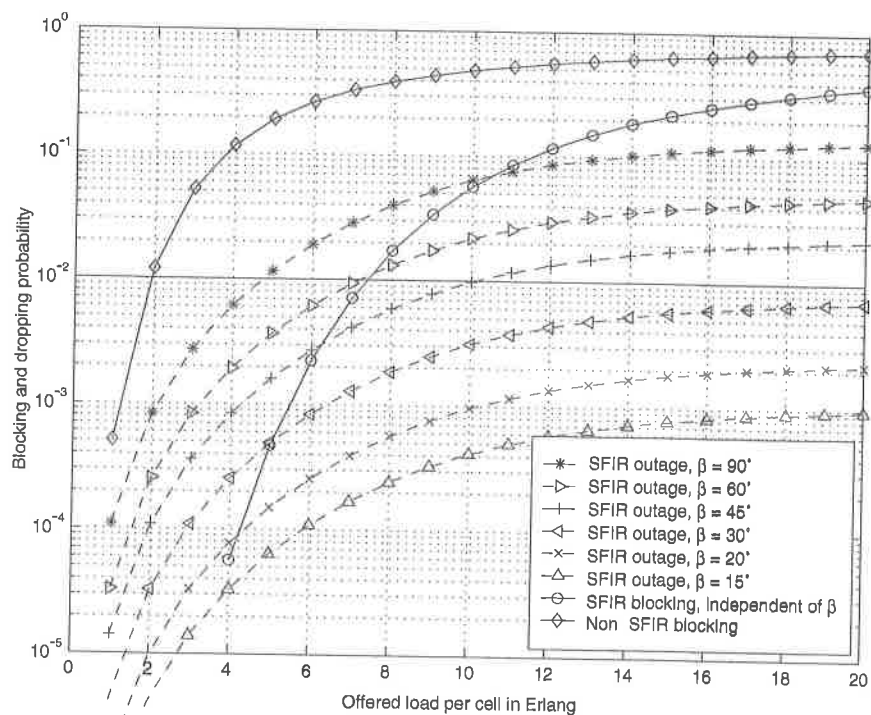


Figure 2.19 SFIR outage and blocking with $(C/I)_{\min} = 14$ dB and $k^{\text{SFIR}} = 3$.

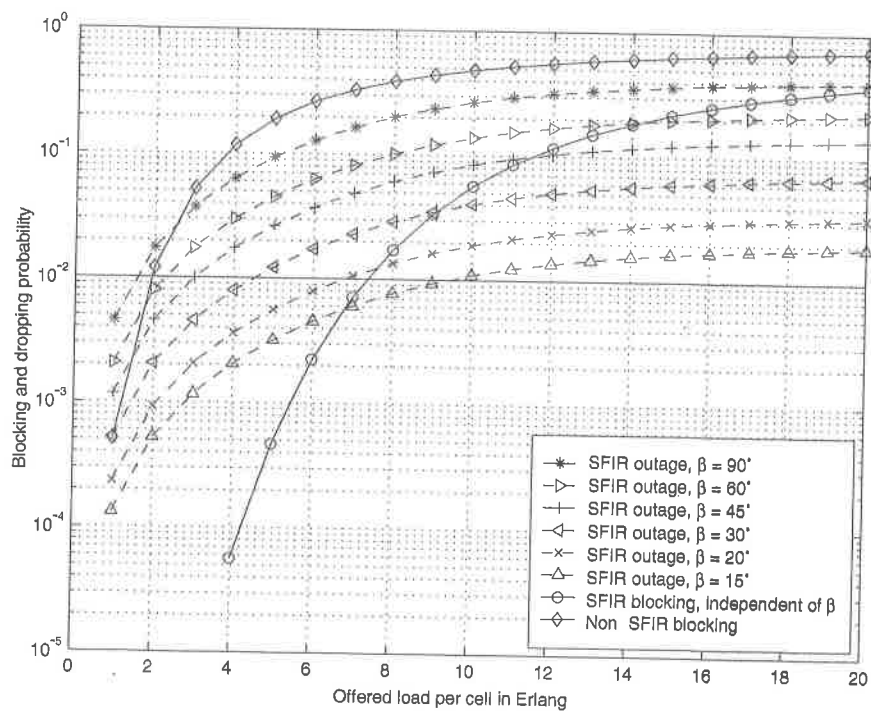


Figure 2.20 SFIR outage and blocking with $(C/I)_{\min} = 16$ dB and $k^{\text{SFIR}} = 3$.

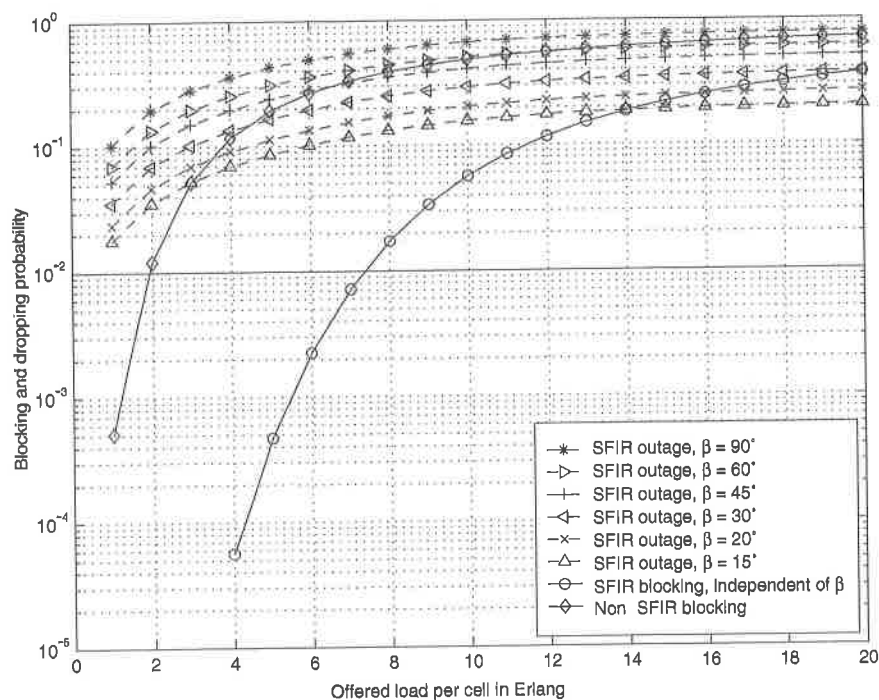


Figure 2.21 SFIR outage and blocking with $(C/I)_{\min} = 18$ dB and $k^{\text{SFIR}} = 3$.

Table 2.3 SFIR capacity gain with $(C/I)_{\min} = 16$ dB and $k^{\text{SFIR}} = 3$.

β	$(C/I)_{k=3}^{\text{SFIR}}$ (dB)	Capacity gain (%)	Limitation
90°	15.88	-21.1	Outage
60°	17.02	21.1	Outage
45°	17.73	57.9	Outage
30°	18.57	131.6	Outage
20°	19.24	257.9	Outage
15°	19.61	284.2	Blocking

Table 2.4 SFIR capacity gain with $(C/I)_{\min} = 18$ dB and $k^{\text{SFIR}} = 3$.

β	$(C/I)_{k=3}^{\text{SFIR}}$ (dB)	Capacity gain	Limitation
90°	15.88	<0	Outage
60°	17.02	<0	Outage
45°	17.73	<0	Outage
30°	18.57	<0	Outage
20°	19.24	<0	Outage
15°	19.61	<0	Outage

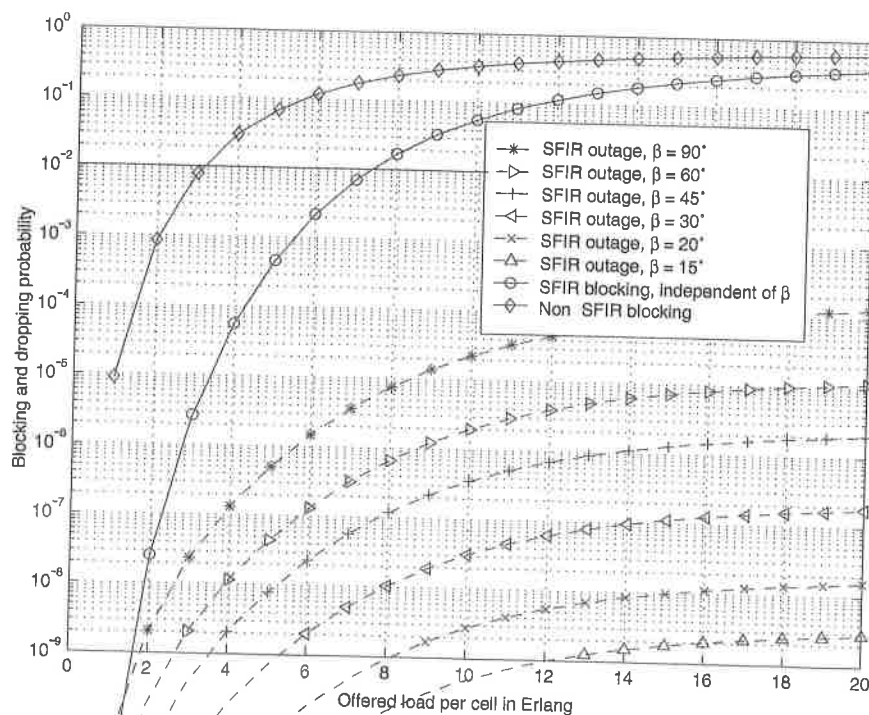


Figure 2.22 SFIR outage and blocking with $(C/I)_{\min} = 14$ dB and $k^{\text{SFIR}} = 4$.

Table 2.5 SFIR capacity gain with $(C/I)_{\min} = 14$ dB and $k^{\text{SFIR}} = 4$.

β	$(C/I)_{k=4}^{\text{SFIR}}$ (dB)	Capacity gain (%)	Limitation
90°	18.68	128.1	Blocking
60°	19.82	128.1	Blocking
45°	20.53	128.1	Blocking
30°	21.37	128.1	Blocking
20°	22.04	128.1	Blocking
15°	22.42	128.1	Blocking

the SFIR capacity gain $G^{\text{SFIR}} = (L_{\text{cell}}^{\text{car,SFIR}} / L_{\text{cell}}^{\text{car,omni}} - 1) \cdot 100\%$ and we obtain $G^{\text{SFIR}} = 284.2\%$, as long as we are only concerned with the blocking probability ($L_{\text{cell}}^{\text{car,SFIR}}$ and $L_{\text{cell}}^{\text{car,omni}}$ are the carried traffic load in the SFIR case and in the omni case, respectively). Now we have to choose a target outage probability. Note that this kind of outage, as opposed to outage caused by fast fading, will last in the order of seconds at least, depending on the user mobility. Since this can easily cause a call to be dropped, we choose the target outage probability to be $P_{\text{out}} = 1\%$. To see whether the outage probability constraint will put a more severe limit on the carried traffic load as the blocking probability, we observe from the plots at which traffic load the outage probability crosses the 1% line. From Figure 2.19 we see

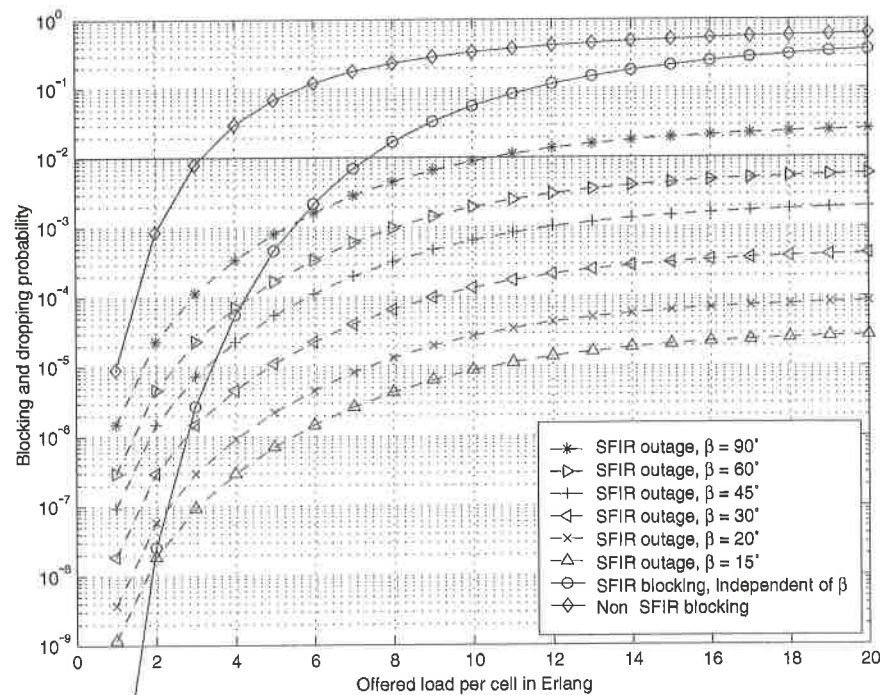


Figure 2.23 SFIR outage and blocking with $(C/I)_{\min} = 16$ dB and $k^{\text{SFIR}} = 4$.

Table 2.6 SFIR capacity gain with $(C/I)_{\min} = 16$ dB and $k^{\text{SFIR}} = 4$.

β	$(C/I)_{k=4}^{\text{SFIR}}$ (dB)	Capacity gain (%)	Limitation
90°	18.68	128.1	Blocking
60°	19.82	128.1	Blocking
45°	20.53	128.1	Blocking
30°	21.37	128.1	Blocking
20°	22.04	128.1	Blocking
15°	22.42	128.1	Blocking

that in the case of $(C/I)_{\min} = 14$ dB the outage constraint becomes a limiting factor for a higher load than the blocking constraint. Thus, in this case the SFIR capacity is blocking limited and the above capacity gain of 284.2% can be fully utilized. However, moving on to the case of $(C/I)_{\min} = 16$ dB, we see from Figure 2.20 that the outage constraint becomes more severe than the blocking constraint, which reduces the carried traffic load and thus the capacity gain shrinks to 131.6%. Finally, looking at Figure 2.21 we observe that the outage constraint leads to a lower traffic load than in the omnidirectional case. Thus, in the case of $(C/I)_{\min} = 18$ dB we actually obtain a reduced capacity through SFIR, unless we accept a higher outage probability.

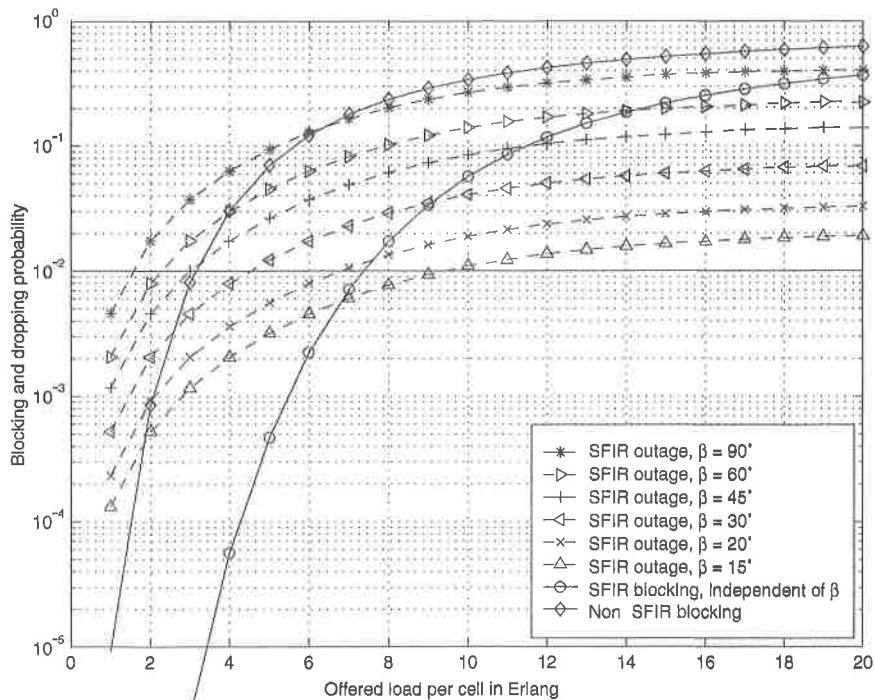


Figure 2.24 SFIR outage and blocking with $(C/I)_{\min} = 18$ dB and $k^{\text{SFIR}} = 4$.

Table 2.7 SFIR capacity gain with $(C/I)_{\min} = 18$ dB and $k^{\text{SFIR}} = 4$.

β	$(C/I)_{k=4}^{\text{SFIR}}$ (dB)	Capacity gain (%)	Limitation
90°	18.68	-50.0	Outage
60°	19.82	-28.1	Outage
45°	20.53	-6.3	Outage
30°	21.37	40.6	Outage
20°	22.04	112.5	Outage
15°	22.42	128.1	Blocking

Tables 2.2, 2.3 and 2.4 contain the achieved capacity gains for a variety of values of β for each of the target C/I values. We can conclude that the capacity gain decreases with growing requirements on the signal quality (growing $(C/I)_{\min}$), and the capacity gain increases as the antenna main lobe becomes narrower. It is also indicated in the tables whether the capacity is rather limited by the blocking constraint or by the outage constraint. Once the system becomes blocking limited, the capacity is independent of the beamwidth.

Looking at the results in Table 2.4, we see that for $(C/I)_{\min} = 18$ dB, the capacity actually decreases with SFIR for all considered values of β . This is remarkable, since according to the averaged worst-case C/I , which is $(C/I)_{k=3}^{\text{SFIR}} = 18.6$ dB for $\beta = 30^\circ$, as computed above, a cluster size reduction from $k = 7$ down to $k = 3$ appears to be feasible.

However, the outage constraint limits the capacity so severely, that it becomes actually lower than in the omnidirectional case. Therefore, we include the results for the case that the cluster size is reduced to $k^{\text{SFIR}} = 4$ instead of $k^{\text{SFIR}} = 3$. Also the number of available channels is altered to $N^{\text{omni}} = 8$ resulting in $N^{\text{SFIR}} = 7/4 \cdot N^{\text{omni}} = 14$ channels per cell after channel reassignment. The results are shown in Figures 2.22, 2.23 and 2.24 as well as in Tables 2.5, 2.6 and 2.7. The results shown in Table 2.7 show that now, with a cluster size of $k = 4$, a capacity gain of 40.6% is achieved for $\beta = 30^\circ$ and $(C/I)_{\min} = 18$ dB. Since the reuse distance is higher in this case, the outage constraint does not limit the capacity as strongly.

System architecture and addressing

3.1 System architecture

The fundamental components of a GSM network are shown in Figure 3.1. A user carries a Mobile Station (MS), which can communicate over the air with a base station, called Base Transceiver Station (BTS) in GSM. The BTS contains transmitter and receiver equipment, such as antennas and amplifiers, as well as a few components for signal and protocol processing. For example, error protection coding is performed in the BTS, and the link-level protocol for signaling on the radio path is terminated here. In order to keep the base stations small, the essential control and protocol intelligence resides in the Base Station Controller (BSC). It contains, for example, protocol functions for radio channel allocation, channel setup and management of handovers. Typically, several BTSs are controlled by one BSC. In practice, the BTS and BSC are connected by fixed lines or point-to-point radio links. BTS and BSC together form the radio access network.

The combined traffic of the users is routed through a switch, called the Mobile Switching Center (MSC). It performs all of the switching functions of a switching node in a fixed telephone network, e.g., in an Integrated Services Digital Network (ISDN). This includes path search, data forwarding and service feature processing. The main difference between an ISDN switch and an MSC is that the MSC also has to consider the allocation and administration of radio resources and the mobility of the users. The MSC therefore has to provide additional functions for location registration of users and for the handover of a connection in the case of changing from cell to cell. A cellular network can have several MSCs with each being responsible for a part of the network (e.g., a city or metropolitan area). Calls originating from or terminating in the fixed network are handled by a dedicated Gateway MSC (GMSC). The interworking of a cellular network and a fixed network (e.g., PSTN, ISDN) is performed by the Interworking Function (IWF). It is needed to map the protocols of the cellular network onto those of the respective fixed network. Connections to other mobile or international networks are typically routed over the International Switching Center (ISC) of the respective country.

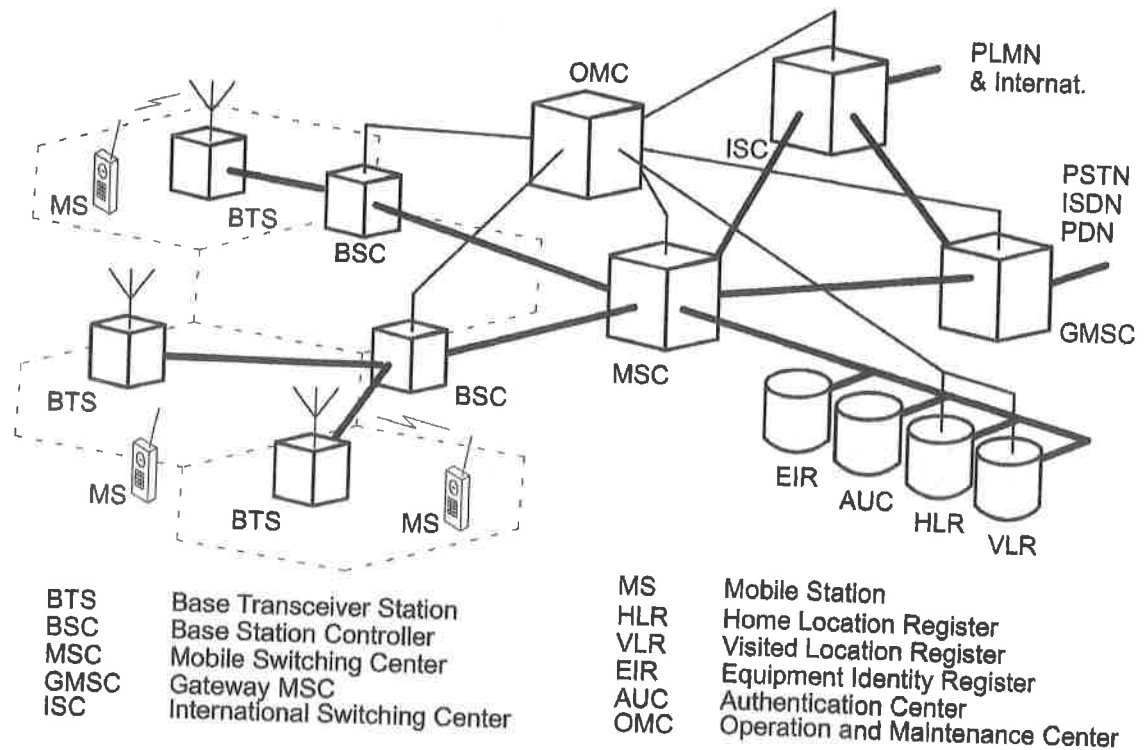


Figure 3.1 GSM system architecture.

A GSM network also contains several types of databases. The Home Location Register (HLR) and the Visited Location Register (VLR) store the current location of a mobile user. This is needed since the network must know the current cell of a user to establish a call to the correct base station. In addition, these registers store the profiles of users, which are required for charging and billing and other administrative issues. Two further databases perform security functions: the Authentication Center (AUC) stores security-related data such as keys used for authentication and encryption; the Equipment Identity Register (EIR) registers equipment data rather than subscriber data.

The network management is organized from a central place, the Operation and Maintenance Center (OMC). Its functions include the administration of subscribers, terminals, charging data, network configuration, operation, performance monitoring and network maintenance. The operation and maintenance functions are based on the concept of the Telecommunication Management Network (TMN) which is standardized in the ITU-T series M.30.

In summary, a GSM network can be divided into three subnetworks: the radio access network, the core network and the management network. These subnetworks are called subsystems in the GSM standard. The respective three subsystems are called the Base Station Subsystem (BSS), the Network Switching Subsystem (NSS) and the Operation and Maintenance Subsystem (OMSS).

Figure 3.2 summarizes the hierarchical relationship between the network components MSC, BSC and BTS. The entire network is divided into MSC regions. Each of these is composed of at least one Location Area (LA), which in turn consists of several cell groups.

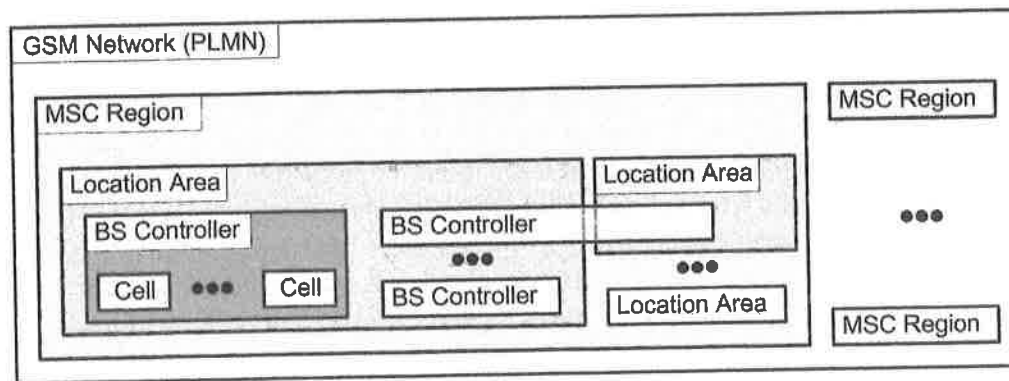


Figure 3.2 GSM system hierarchy.



Figure 3.3 Mobile equipment personalization with the SIM.

Each cell group is assigned to a BSC. For each LA there exists at least one BSC, but cells of one BSC may belong to different LAs. The exact partitioning of the network area with respect to LAs, BSCs and MSCs is not, however, uniquely determined and is left to the network operator who thus has many possibilities for optimization.

3.2 The SIM concept

Each GSM user owns a personal chip card, the Subscriber Identity Module (SIM). As illustrated in Figure 3.3, it can be plugged into a piece of mobile equipment. In fact, only the SIM of a subscriber turns a piece of mobile equipment into a complete mobile station with network usage privileges, which can be used to make calls or receive calls.

This concept allows us to distinguish between equipment mobility and subscriber mobility. The subscriber can register to the locally available network with their SIM card on different mobile stations, or the SIM card could be used as a normal telephone card in the fixed telephone network. This enables international roaming independent of mobile equipment

and network technology, provided that the interface between SIM and end terminal is standardized.

Beyond that, the SIM can store short messages and charging information, and it has a telephone book function and short list of call numbers storing names and telephone numbers for efficient and fast number selection. These functions, in particular, contribute to a genuine personalization of a mobile terminal, since the subscriber can use their normal 'environment' plus telephone list and short message archive with any piece of mobile equipment. In addition to subscriber-specific data, the SIM can also store network-specific data, e.g., lists of carrier frequencies used by the network to broadcast system information periodically. Use of the SIM and thus of the whole MS can be protected with a Personal Identification Number (PIN) against unauthorized access.

The SIM also takes over security functions: all of the cryptographic algorithms to be kept confidential are realized on the SIM, which implements important functions for the authentication and user data encryption based on the subscriber identity and secret keys.

3.3 Addressing

As in each communication network, the entities of a GSM network must be assigned certain addresses or identities. These serve to identify, authenticate and localize the network entities. The most commonly known GSM address is the telephone number of a user. In addition to telephone numbers, several other identifiers have been defined; they are needed for the management of user mobility and for addressing all remaining network elements.

GSM distinguishes explicitly between a user and their equipment. Hence, there are specific address types for users and specific address types for MSs. The user identities are stored on the SIM; the equipment identities on the mobile equipment. In addition, GSM distinguishes between user identity and their telephone number. This leaves some scope for development of services when each subscriber may be called personally, independent of reachability or type of connection (mobile or fixed). In addition to the personal identifier, each GSM subscriber is assigned one or several ISDN numbers. The following sections explain the most important addresses and identifiers used in GSM.

3.3.1 International mobile station equipment identity

The International Mobile Station Equipment Identity (IMEI) uniquely identifies a mobile station internationally and gives clues about its manufacturer and the date of manufacturing. It is a kind of serial number. The IMEI is allocated by the equipment manufacturer and registered by the network operator, who stores it in the EIR.

By means of the IMEI one recognizes obsolete, stolen or nonfunctional equipment and can deny service if required. For this purpose, the IMEI is assigned to one or more of the following three categories within the EIR.

- The white list is a register of all equipment.
- The black list contains all suspended equipment. This list is periodically exchanged among network operators.

- Optionally, an operator may maintain a gray list, in which malfunctioning equipment or equipment with obsolete software versions is registered. Such equipment has network access, but its use is reported to the operating personnel.

The IMEI is usually requested by the network at registration, but it can be requested repeatedly. It is a hierarchical address, containing the following parts:

- Type Approval Code (TAC), six digits, centrally assigned;
- Final Assembly Code (FAC), six digits, assigned by the manufacturer;
- Serial Number, six digits, assigned by the manufacturer;
- Spare, one digit.

3.3.2 International mobile subscriber identity

When registering for service with a mobile network operator, each subscriber receives a unique identifier, the International Mobile Subscriber Identity (IMSI). This IMSI is stored in the SIM. A mobile station can only be operated if a SIM with a valid IMSI is inserted into equipment with a valid IMEI, since this is the only way to correctly bill the associated subscriber. The IMSI uses a maximum of 15 decimal digits and consists of three parts:

- Mobile Country Code (MCC), three digits, internationally standardized;
- Mobile Network Code (MNC), two digits, for unique identification of mobile networks within a country;
- Mobile Subscriber Identification Number (MSIN), maximum of 10 digits, identification number of the subscriber in their mobile home network.

The IMSI is a GSM-specific addressing concept and is different from the ISDN numbering plan. A three-digit MCC has been assigned to each of the GSM countries, and two-digit MNCs have been assigned within countries (e.g., 262 as MCC for Germany; and MNC 01, 02 and 07 for the networks of T-Mobile, Vodafone, and O2, respectively). Whereas the MCC is defined internationally, the National Mobile Subscriber Identity (NMSI = MNC + MSIN) is assigned by the operator of the home network.

3.3.3 Mobile subscriber ISDN number

The 'real telephone number' of a mobile user is called the Mobile Subscriber ISDN Number (MSISDN). It is assigned to the subscriber (their SIM), such that a mobile station can have several MSISDNs depending on the SIM. With this concept, GSM was the first mobile system to distinguish between subscriber identity and the number to call. The separation of call number (MSISDN) and subscriber identity (IMSI) primarily serves to protect the confidentiality of the IMSI. In contrast to the MSISDN, the IMSI need not be made public. With this separation, one cannot derive the subscriber identity from the MSISDN, unless the association of IMSI and MSISDN as stored in the HLR has been made public. It is the rule that the IMSI used for subscriber identification is not known, and thus the faking of a false identity is significantly more difficult.

In addition, a subscriber can hold several MSISDNs for selection of different services. Each MSISDN of a subscriber is reserved for specific service (voice, data, fax, etc.). In order to realize this service, service-specific resources have to be activated in the MS as well as in the network. The service desired and the resources needed for the specific call can be derived from the MSISDN. Thus, an automatic activation of service-specific resources is already possible during the setup of a connection. The MSISDN categories follow the international ISDN numbering plan, having the following structure:

- Country Code (CC), up to three digits;
- National Destination Code (NDC), typically two or three digits;
- Subscriber Number (SN), a maximum of 10 digits.

The CCs are internationally standardized, complying with the ITU-T recommendation E.164. There are country codes with one, two, or three digits, e.g. the country code for the USA is 1, for the UK it is 44 and for Finland it is 358. The national operator or regulatory administration assigns the NDC as well as the SN, which may have variable length. The NDC of the mobile networks in Germany have three digits (e.g., 170, 171, 172). The MSISDN is stored centrally in the HLR.

3.3.4 Mobile station roaming number

The Mobile Station Roaming Number (MSRN) is a temporary location-dependent ISDN number. It is assigned by the locally responsible VLR to each MS in its area. Calls are routed to the MS by using the MSRN. On request, the MSRN is passed from the HLR to the GMSC. The MSRN has the same structure as the MSISDN:

- CC of the visited network;
- NDC of the visited network;
- SN in the current mobile network.

The components CC and NDC are determined by the network visited and depend on the current location. The SN is assigned by the current VLR and is unique within the mobile network. An MSRN is assigned in such a way that the currently responsible switching node MSC in the visited network can be determined from the subscriber number, which allows routing decisions to be made.

The MSRN can be assigned in two different ways by the VLR: either at each registration when the MS enters a new LA or each time when the HLR requests it for setting up a connection for incoming calls to the MS.

In the first case, the MSRN is also passed on from the VLR to the HLR, where it is stored for routing. In the case of an incoming call, the MSRN is first requested from the HLR of this MS. In this way the currently responsible MSC can be determined, and the call can be routed to this switching node. Additional localization information can be obtained there from the responsible VLR.

In the second case, the MSRN cannot be stored in the HLR, since it is only assigned at the time of call setup. Therefore, the address of the current VLR must be stored in the tables

of the HLR. Once routing information is requested from the HLR, the HLR itself goes to the current VLR and uses a unique subscriber identification (IMSI and MSISDN) to request a valid roaming number MSRN. This allows further routing of the call.

3.3.5 Location area identity

Each LA of a cellular network has its own identifier. The Location Area Identifier (LAI) is also structured hierarchically and internationally unique, with LAI again consisting of an internationally standardized part and an operator-dependent part:

- CC, three digits;
- MNC, two digits;
- Location Area Code (LAC), a maximum of five digits or a maximum of 2×8 bits, coded in hexadecimal.

This LAI is broadcast regularly by the base station on the Broadcast Control Channel (BCCH). Thus, each cell is identified uniquely on the radio channel as belonging to an LA, and each MS can determine its current location through the LAI. If the LAI that is 'heard' by the MS changes, the MS notices this LA change and requests an update to its location information in the VLR and HLR (location update). The significance for GSM networks is that the MS itself rather than the network is responsible for monitoring the local conditions of signal reception, to select the base station that can be received best, and to register with the VLR of that LA which the current base station belongs to. The LAI is requested from the VLR if the connection for an incoming call has been routed to the current MSC using the MSRN. This determines the precise location of the MS where the mobile can be subsequently paged. When the MS answers, the exact cell and therefore also the base station become known; this information can then be used to switch the call through.

3.3.6 Temporary mobile subscriber identity

The VLR responsible for the current location of a subscriber can assign a Temporary Mobile Subscriber Identity (TMSI), which has only local significance in the area handled by the VLR. It is used in place of the IMSI for the definite identification and addressing of the MS. In this way nobody can determine the identity of the subscriber by listening to the radio channel, since this TMSI is only assigned during the presence of the MS in the area of one VLR, and can even be changed during this period (ID hopping). The MS stores the TMSI on the SIM card. The TMSI is stored on the network side only in the VLR and is not passed to the HLR. A TMSI may therefore be assigned in an operator-specific way; it can consist of up to 4×8 bits, but the HEX value FFFF FFFF is excluded, because the SIM marks empty fields internally with logical 1.

Together with the current location area, a TMSI allows a subscriber to be identified uniquely, i.e., for the ongoing communication the IMSI is replaced by the 2-tuple (TMSI, LAI).

3.3.7 Other identifiers

The VLR can assign an additional searching key to each MS within its area to accelerate database access; this is the Local Mobile Station Identity (LMSI). The LMSI is assigned when the MS registers with the VLR and is also sent to the HLR. The LMSI is no longer used by the HLR, but each time messages are sent to the VLR concerning a MS, the LMSI is added, so the VLR can use the short searching key for transactions concerning this MS. This kind of additional identification is only used when the MSRN is newly assigned with each call. In this case, fast processing is very important to achieve short times for call setup. Like the TMSI, an LMSI is also assigned in an operator-specific way, and it is only unique within the administrative area of a VLR. An LMSI consists of four octets (4×8 bits).

Within an LA, the individual cells are uniquely identified with a Cell Identifier (CI), a maximum of 2×8 bits. Together with the global CI cells are thus also internationally defined in a unique way.

In order to distinguish neighboring base stations, these receive a Base Transceiver Station Identity Code (BSIC) which consists of two components:

- Network Color Code (NCC), a color code within a mobile network (3 bits);
- Base Transceiver Station Color Code (BCC), a BTS color code (3 bits).

The BSIC is broadcast periodically by the base station. Directly adjacent mobile networks must have different NCCs, and neighboring base stations of a mobile network must have different BCCs.

MSCs and location registers (HLR, VLR) are addressed with ISDN numbers. In addition, they may have a Signaling Point Code (SPC) within a mobile network, which can be used to address them uniquely within the Signaling System Number 7 network (SS#7). The number of the VLR in whose area a MS is currently roaming must be stored in the HLR data for this MS, if the MSRN distribution is on a call-by-call basis; thus the MSRN can be requested for incoming calls and the call can be switched through to the MS.

3.4 Registers and subscriber data

3.4.1 Location registers (HLR and VLR)

The GSM standard defines two database types for the management of user data and location: the HLR and the VLR. These databases are queried by the network for user registration and localization.

The HLR has a record for all subscribers registered with a network operator. It stores, for example, each user's telephone number, service subscriptions, permissions and authentication data. In addition to this permanent administrative data, it also contains temporary data, such as the current location of a subscriber. In the case of incoming traffic to a mobile user, the HLR is queried to determine the user's current location. This enables the gateway to route the traffic to the appropriate MSC. The MS must inform the network about its current location area; to do so it sends a location update message to the network whenever it changes its location area. The full list of subscriber data stored in the HLR is given in Table 3.1.

A VLR is responsible for a group of location areas and stores the data of all users that are currently located in this area. The data includes part of the permanent user data, which is

Table 3.1 Mobile subscriber data in the HLR.

Subscriber and subscription data:	– IMSI
	– MSISDN
	– Service subscriptions
	– Service restrictions (e.g., roaming restrictions)
	– Information on the subscriber's equipment (if available)
	– Authentication data (subject to implementation)
Tracking and routing information:	– Mobile Station Roaming Number (MSRN)
	– Current VLR address (if available)
	– Current MSC address (if available)
	– Local Mobile Subscriber Identity (LMSI) (if available)

Table 3.2 Mobile subscriber data in the VLR.

Subscriber and subscription data:	– IMSI
	– MSISDN
	– Parameters for supplementary services
	– Information on the subscriber's equipment (if available)
	– Authentication data (subject to implementation)
Tracking and routing information:	– MSRN
	– TMSI
	– LMSI (if available)
	– LAI of LA where the MS was registered (used for paging and call setup)

copied from the HLR to the VLR for fast access. In addition, the VLR may also assign and store local data, such as temporary identifiers. A user may either be registered with a VLR of their home network or a foreign network. Upon a location update, the MSC forwards the user's identity and current location to the VLR, which subsequently updates its database. If the user has not been registered with this VLR before, the HLR is informed about the current VLR of the user. This process enables incoming calls to be routed to this MS. Table 3.2 summarizes the subscriber data stored in the VLR.

Typically, there is one central HLR per network and one VLR for each MSC. This organization depends on the number of subscribers, the processing and storage capacity of the switches and the structure of the network.

3.4.2 Security-related registers (AUC and EIR)

Two additional databases are responsible for various aspects of system security. System security of GSM networks is based primarily on the verification of equipment and subscriber

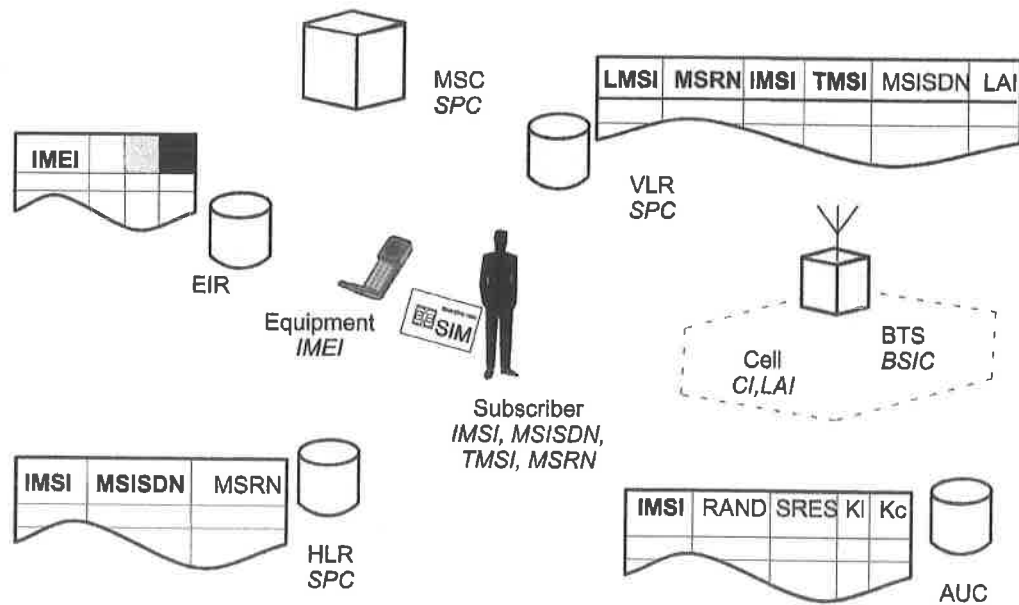


Figure 3.4 GSM databases and addresses.

identity; therefore, the databases serve for subscriber identification and authentication and for equipment registration. Confidential data and keys are stored or generated in the AUC. The keys serve for user authentication and authorize the respective service access. The EIR stores the serial numbers (supplied by the manufacturer) of the terminals (IMEI), which makes it possible to check for MSs with obsolete software or to block service access for MSs reported as stolen.

3.4.3 Subscriber data

Service-specific data are used to parameterize and personalize supplementary services. Finally, contracts with subscribers can define different service levels, e.g., booking of special supplementary services or subscriptions to data or teleservices. The contents of such contracts are stored in appropriate data structures in order to enable correct realization or provision of these services.

The association of the most important identifiers and their storage locations is summarized in Figure 3.4. Subscriber-related addresses are stored on the SIM and in the HLR and VLR as well. These data (IMSI, MSISDN, TMSI, MSRN) serve to address, identify and localize a subscriber or a MS. Whereas IMSI and MSISDN are permanent data items, TMSI and MSRN are temporary values, which change according to the current location of the subscriber. Of the other data items defined for user or network equipment elements (such as IMEI, LAI or SPCs), only some are used (LAI, SPC) for localizing or routing. IMEI and BSIC/CI hold a special position by being used only for identification of network elements.

Security-relevant subscriber data are stored in the AUC, which also calculates identifiers and keys for cryptographic processing functions. Each set of data in the AUC contains the IMSI of the subscriber as a search key. For identification and authentication of a subscriber,

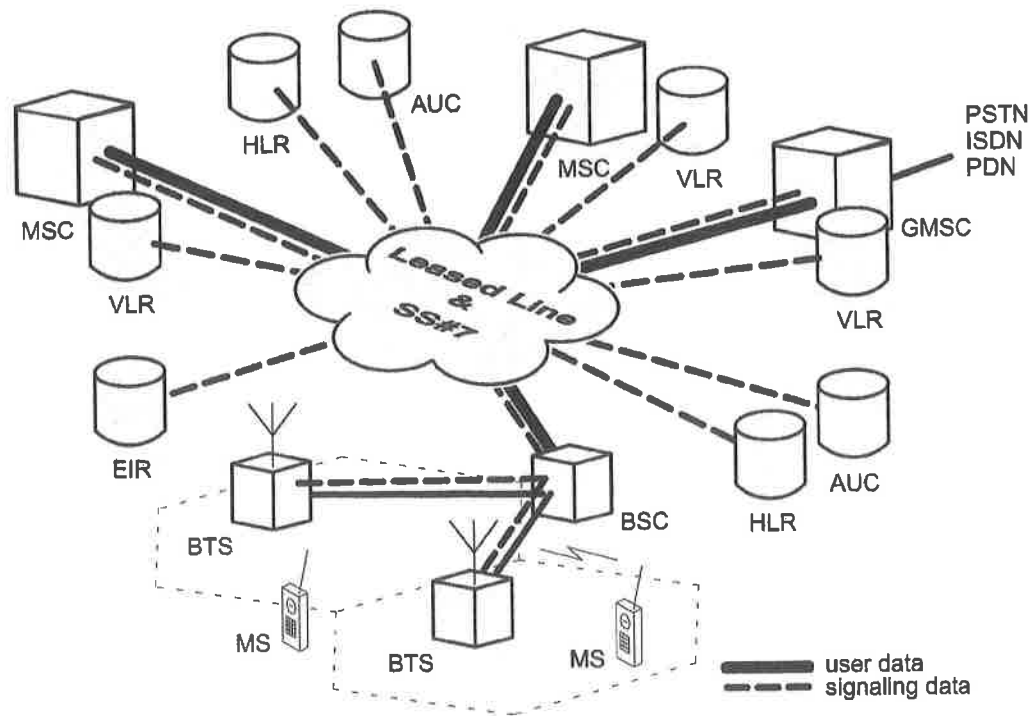


Figure 3.5 User data transport and signaling in a GSM network.

the AUC stores the subscriber's secret key K_i from which a pair of keys RAND/SRES are precalculated and stored. Once an authentication request occurs, this pair of keys is queried by the VLR to conduct the identification/authentication process properly. The key K_c for user data encryption on the radio channel is also calculated in advance in the AUC from the secret key K_i and is requested by the VLR at connection setup.

Above all, the HLR contains the permanent data about the subscriber's contractual relationship, e.g., information about subscribed bearer and teleservices (data, fax, etc.), service restrictions and parameters for supplementary services. Beyond that, the registers also contain information about equipment used by the subscriber (IMEI). Depending on the implementation of the authentication center and the security mechanisms, data and keys used for subscriber authentication and encryption can also be stored there. The search keys used for retrieving subscriber information (such as IMSI, MSISDN, MSRN, TMSI and LMSI) from a register are indicated in boldface (Figure 3.4).

3.5 Network interfaces and configurations

Figure 3.5 shows the GSM system architecture with corresponding user data and signaling links between the network components.

Signaling has two fundamentally different parts: The core network employs the SS#7, which is well known from fixed networks. In order to setup, manage and release calls, the SS#7 protocol called ISDN User Part (ISUP) is used. In order to perform signaling that

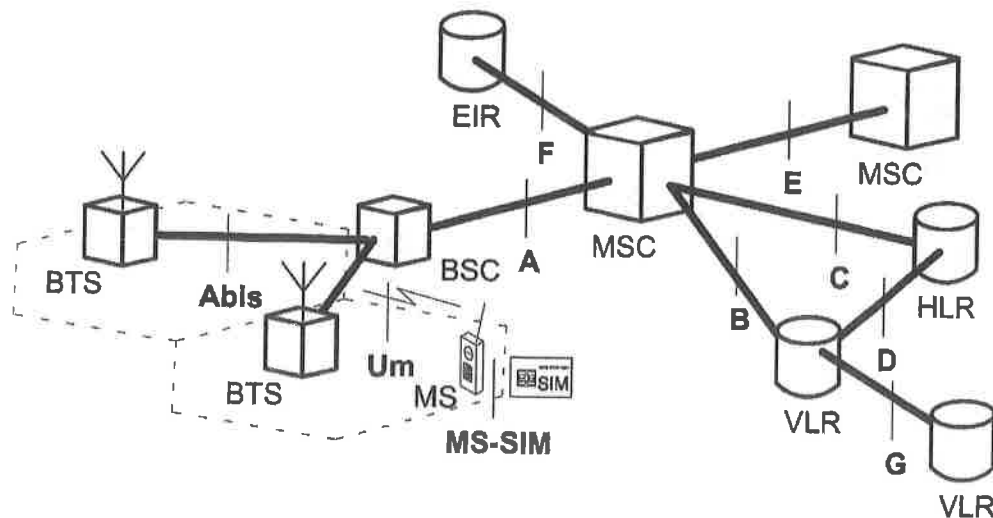


Figure 3.6 Interfaces in a GSM network.

is specific to mobile networking, an extension to SS#7 has been developed, the so-called Mobile Application Part (MAP). It is implemented in the MSC, HLR and VLR.

The radio access network (including the air interface) does not employ the SS#7 protocol, but uses a GSM-specific protocol. Signaling between the radio access network and the MSC uses the Base Station System Application Part (BSSAP).

3.5.1 Interfaces

The communication relationships between the GSM network components are formally described by a number of standardized interfaces (Figure 3.6).

The A interface between BSS and MSC is used for the transfer of data for BSS management, for connection control and for mobility management. Within the BSS, the Abis interface between BTS and BSC and the air interface Um have been defined.

An MSC which needs to obtain data about an MS staying in its administrative area, requests the data from the VLR responsible for this area over the B interface. Conversely, the MSC forwards to this VLR any data generated at location updates by MSs. If the subscriber reconfigures special service features or activates supplementary services, the VLR is also informed first, which then updates the HLR.

This updating of the HLR occurs through the D interface. The D interface is used for the exchange of location-dependent subscriber data and for subscriber management. The VLR informs the HLR about the current location of the mobile subscriber and reports the current MSRN. The HLR transfers all of the subscriber data to the VLR that is needed to give the subscriber their usual customized service access. The HLR is also responsible for giving a cancellation request for the subscriber data to the old VLR once the acknowledgement for the location update arrives from the new VLR. If, during location updating, the new VLR needs data from the old VLR, it is directly requested over the G interface. Furthermore, the identity of subscriber or equipment can be verified during a location update; for requesting and checking the equipment identity, the MSC has an interface F to the EIR.

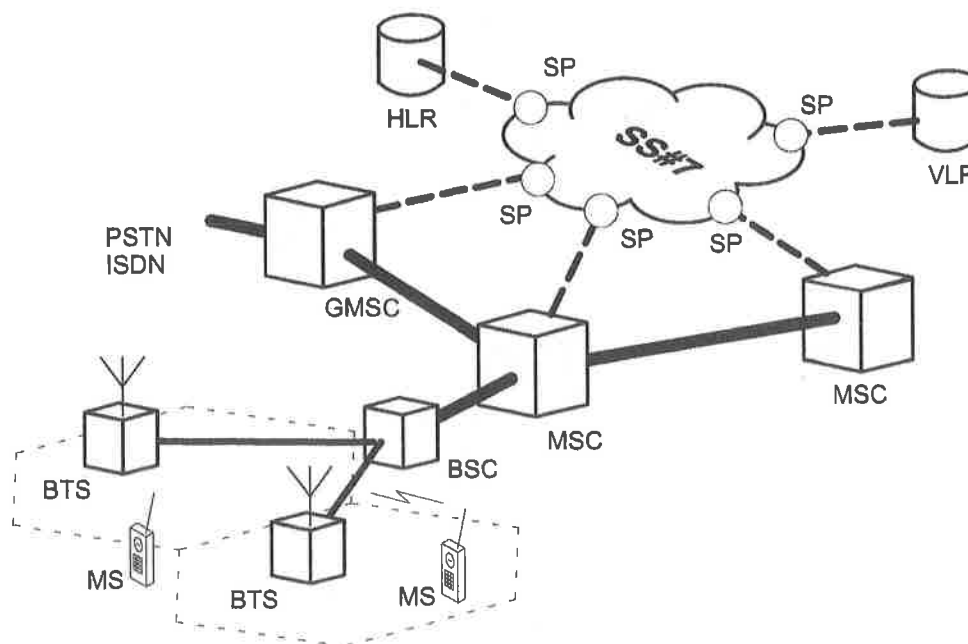


Figure 3.7 Basic configuration of a GSM network.

An MSC has two more interfaces in addition to the A and B interfaces, namely the C and E interfaces. Charging information can be sent over the C interface to the HLR. In addition to this, the MSC must be able to request routing information from the HLR during call setup, for calls from the mobile network as well as for calls from the fixed network. In the case of a call from the fixed network, if the fixed network's switch cannot interrogate the HLR directly, initially it routes the call to a GMSC, which then interrogates the HLR. If the mobile subscriber changes during a conversation from one MSC area to another, a handover needs to be performed between these two MSCs, which occurs across the E interface.

3.5.2 Configurations

As mentioned above, the configuration of a mobile network is largely left to the network operator. Figure 3.7 shows a basic configuration of a GSM network. This configuration contains a central HLR and a central VLR. All database transactions (updates, inquiries, etc.) and handover transactions between the MSC are performed with the help of the MAP over the SS#7 network. For this purpose, each MSC and register is known as a Signaling Point (SP) and is identified by its Signaling Point Code (SPC) within the SS#7 network. Whenever an MS changes its location area, the location information in the VLR must be updated. Furthermore, the VLR has to be interrogated: the MSC needs subscriber parameters in addition to location data for successful connection setup, such as service restrictions and supplementary services to be activated. Thus, there is a significant message traffic between MSC and VLR, which constitutes an ensuing load on the signaling network.

Hence, these two functional units can be combined into one physical unit, i.e. the entire VLR is implemented in distributed form and a VLR is associated with each MSC

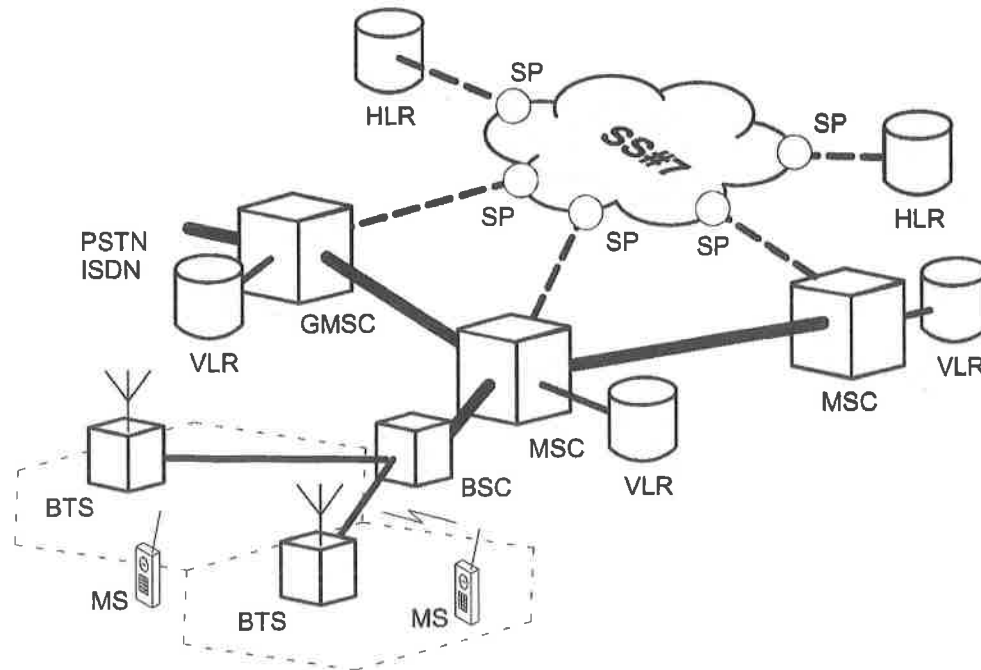


Figure 3.8 Configuration of a GSM network with a VLR for each MSC.

(see Figure 3.8). The traffic between MSC and VLR no longer needs to be transported through the SS#7 network.

We can go one step further and also distribute the database of the HLR, thus introducing several HLRs into a mobile network. This is especially interesting for a growing pool of subscribers, since a centralized database leads to a high traffic load for this database. If there are several HLRs in a network, the network operator has to define an association rule between MSISDNs and HLRs, such that for incoming calls the routing information to an MSISDN can be derived from the associated HLR. One possible association is geographic partitioning of the whole subscriber identification space (SN field in the MSISDN), where, for example, the first two digits of the SN indicate the region and the associated HLR. In extreme cases, the HLR can be realized with the VLR in a single physical unit. In this case, an HLR would also be associated with each MSC.

Air interface – physical layer

The GSM physical layer, which resides on the first of the seven layers of the Open Systems Interconnection (OSI) Reference Model (Tanenbaum, 1996), contains very complex functions. The physical channels are defined here by a TDMA scheme. On top of the physical channels, a series of logical channels are defined, which are transmitted in the time slots of the physical channels. Logical channels perform a multiplicity of functions, such as payload transport, signaling, broadcast of general system information, synchronization and channel assignment.

The structure of this chapter is as follows: In section 4.1, we describe the logical channels. This serves as a foundation for understanding the signaling procedures at the air interface. The realization of the physical channels, including GSM modulation, multiple access, duplexing and frequency hopping follows in section 4.2. Next, section 4.3 covers synchronization. The mapping of logical onto physical channels follows in section 4.4, where the higher-level multiplexing of logical channels into multiframes is also covered. Section 4.5 contains a discussion of the most important control mechanisms for the air interface (channel measurement, power control, disconnection and cell selection). The conclusion of the chapter is a power-up scenario with the sequence of events occurring, from when a MS is turned on to when it is in a synchronized state ready to transmit (section 4.9).

4.1 Logical channels

On Layer 1 of the OSI Reference Model, GSM defines a series of logical channels, which are made available either in an unassigned random access mode or in a dedicated mode assigned to a specific user. Logical channels are divided into two categories (Table 4.1): traffic channels and signaling (control) channels.

4.1.1 Traffic channels

The Traffic Channels (TCHs) are used for the transmission of user payload data (speech, data). They do not carry any control information of Layer 3. Communication over a TCH can be circuit-switched or packet-switched. In the circuit-switched case, the TCH provides a transparent data connection or a connection that is specially treated according to the carried

Table 4.1 Classification of logical channels in GSM.

Group		Channel	Function	Direction
Traffic channel	TCH	TCH/F, Bm	Full-rate TCH	MS ↔ BSS
		TCH/H, Lm	Half-rate TCH	MS ↔ BSS
Signaling channels (Dm)	BCH	BCCH	Broadcast control	MS ← BSS
		FCCH	Frequency correction	MS ← BSS
		SCH	Synchronization	MS ← BSS
	CCCH	RACH	Random access	MS → BSS
		AGCH	Access grant	MS ← BSS
		PCH	Paging	MS ← BSS
		NCH	Notification	MS ← BSS
	DCCH	SDCCH	Stand-alone dedicated control	MS ↔ BSS
		SACCH	Slow associated control	MS ↔ BSS
		FACCH	Fast associated control	MS ↔ BSS

service (e.g. telephony). For the packet-switched mode, the TCH carries user data of OSI Layers 2 and 3 according to the recommendations of the X.25 standard or similar standard packet protocols.

A TCH may either be fully used (full-rate TCH, TCH/F) or be split into two half-rate channels (half-rate TCH, TCH/H), which can be allocated to different subscribers. Following ISDN terminology, the GSM traffic channels are also designated as Bm channel (mobile B channel) or Lm channel (lower-rate mobile channel, with half the bit rate). A Bm channel is a TCH for the transmission of bit streams of either 13 kbit/s of digitally coded speech or of data streams at 14.5, 12, 6 or 3.6 kbit/s. Lm channels are TCH channels with less transmission bandwidth than Bm channels and transport speech signals of half the bit rate (TCH/H) or bit streams for data services with 6 or 3.6 kbit/s.

4.1.2 Signaling channels

The control and management of a cellular network demands a very high signaling effort. Even when there is no active connection, signaling information (for example, location update information) is permanently transmitted over the air interface. The GSM signaling channels offer a continuous, packet-oriented signaling service to MSs in order to enable them to send and receive messages at any time over the air interface to the BTS. Following ISDN terminology, the GSM signaling channels are also called Dm channels (mobile D channel). They are further divided into Broadcast Channel (BCH), Common Control Channel (CCCH) and Dedicated Control Channel (DCCH) (see Table 4.1).

The unidirectional BCHs are used by the BSS to broadcast the same information to all MSs in a cell. The group of BCHs consists of three channels.

- Broadcast Control Channel (BCCH): On this channel, a series of information elements is broadcast to the MSs which characterize the organization of the radio network, such as radio channel configurations (of the currently used cell as well as of the neighboring

cells), synchronization information (frequencies as well as frame numbering) and registration identifiers (LAI, CI, BSIC). In particular, this includes information about the structural organization (formats) of the CCCH of the local BTS. The BCCH is broadcast on the first frequency assigned to the cell (the so-called BCCH carrier).

- Frequency Correction Channel (FCCH): On the FCCH, information about correction of the transmission frequency is broadcast to the MSs; see section 4.2.2 (frequency correction burst).
- Synchronization Channel (SCH): The SCH broadcasts information to identify a BTS, i.e. BSIC; see Chapter 3. The SCH also broadcasts data for the frame synchronization of a MS, i.e. Reduced Frame Number (RFN) of the TDMA frame; see section 4.3.1.

FCCH and SCH are only visible within protocol Layer 1, since they are only needed for the operation of the radio subsystem. There is no access to them from Layer 2. In spite of this fact, the SCH messages contain data which are needed by Layer 3 for the administration of radio resources. These two channels are always broadcast together with the BCCH.

The CCCH is a point-to-multipoint signaling channel to deal with access management functions. This includes the assignment of dedicated channels and paging to localize a MS. It comprises the following.

- Random Access Channel (RACH): The RACH is the uplink portion of the CCCH. It is accessed from the mobile stations in a cell without reservation in a competitive multiple-access mode using the principle of slotted Aloha (Bertsekas and Gallager, 1987), to ask for a dedicated signaling channel for exclusive use by one MS for one signaling transaction.
- Access Grant Channel (AGCH): The AGCH is the downlink part of the CCCH. It is used to assign an SDCCH or a TCH to a MS.
- Paging Channel (PCH): The PCH is also part of the downlink of the CCCH. It is used for paging to find specific MSs.
- Notification Channel (NCH): The NCH is used to inform MSs about incoming group and broadcast calls.

The last type of signaling channel, the DCCH is a bidirectional point-to-point signaling channel. An Associated Control Channel (ACCH) is also a dedicated control channel, but it is assigned only in connection with a TCH or an SDCCH. The group of Dedicated/Associated Control Channels (D/ACCH) comprises the following.

- Stand-alone Dedicated Control Channel (SDCCH): The SDCCH is a dedicated point-to-point signaling channel (DCCH) which is not tied to the existence of a TCH ('stand-alone'), i.e. it is used for signaling between a MS and the BSS when there is no active connection. The SDCCH is requested from the MS via the RACH and assigned via the AGCH. After the completion of the signaling transaction, the SDCCH is released and can be reassigned to another MS. Examples of signaling transactions which use an SDCCH are the updating of location information or parts of the connection setup until the connection is switched through (see Figure 4.1).

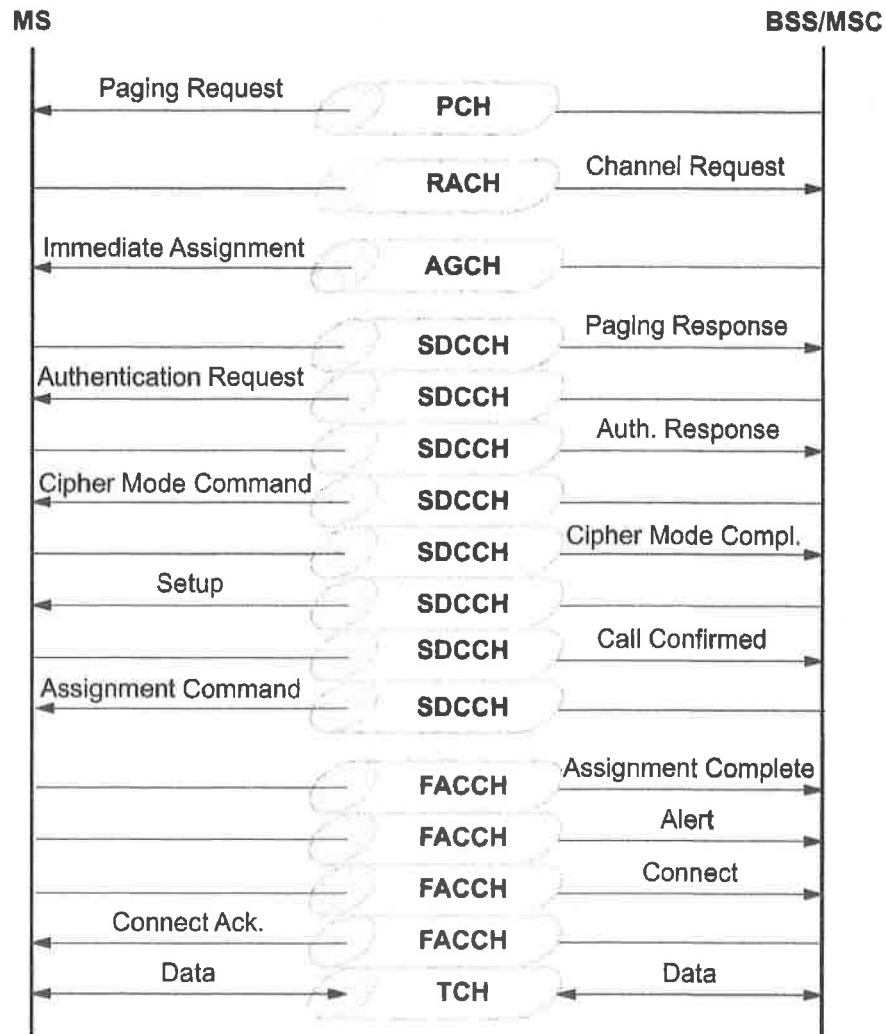


Figure 4.1 Logical channels and signaling (connection setup for an incoming call).

- **Slow Associated Control Channel (SACCH):** An SACCH is always assigned and used with a TCH or an SDCCH. The SACCH carries information for the optimal radio operation, e.g., commands for synchronization and transmitter power control and reports on channel measurements (section 4.5). Data must be transmitted continuously over the SACCH since the arrival of SACCH packets is taken as proof of the existence of the physical radio connection (section 4.5.3). When there is no signaling data to transmit, the MS sends a measurement report with the current results of the continuously conducted radio signal level measurements (section 4.5.1).
- **Fast Associated Control Channel (FACCH):** By using dynamic preemptive multiplexing on a TCH, additional bandwidth can be made available for signaling. The signaling channel created this way is called FACCH. It is only assigned in connection with a TCH, and its short-time usage goes at the expense of the user data transport.

Table 4.2 Logical channels of GSM protocol Layer 1.

Channel type	Net data throughput (kbit/s)	Block length (bits)	Block distance (ms)
TCH (full-rate speech)	13.0	182 + 78	20
TCH (half-rate speech)	5.6	95 + 17	20
TCH (data, 14.4 kbit/s)	14.5	290	20
TCH (data, 9.6 kbit/s)	12.0	60	5
TCH (data, 4.8 kbit/s)	6.0	60	10
TCH (data, up to 2.4 kbit/s)	3.6	72	10
FACCH full rate	9.2	184	20
FACCH half rate	4.6	184	40
SDCCH	598/765	184	3060/13
SACCH (with TCH)	115/300	168 + 16	480
SACCH (with SDCCH)	299/765	168 + 16	6120/13
BCCH	598/765	184	3060/13
AGCH	$n \times 598/765$	184	3060/13
NCH	$m \times 598/765$	184	3060/13
PCH	$p \times 598/765$	184	3060/13
RACH	$r \times 27/765$	8	3060/13
CBCH	598/765	184	3060/13

In addition to these channels, a Cell Broadcast Channel (CBCH) is defined, which is used to broadcast the messages of the Short Message Service Cell Broadcast (SMSCB). The CBCH shares a physical channel with the SDCCH.

4.1.3 Example: connection setup for incoming call

Figure 4.1 shows an example for an incoming call connection setup at the air interface. It is illustrated how the various logical channels are used in principle. The MS is called via the PCH and requests a signaling channel on the RACH. It obtains the SDCCH through an IMMEDIATE ASSIGNMENT message on the AGCH. Then follow authentication, start of ciphering and start of setup over the SDCCH. An ASSIGNMENT COMMAND message gives the traffic channel to the MS, which acknowledges its receipt on the FACCH of this traffic channel. The FACCH is also used to continue the connection setup.

4.1.4 Bit rates, block lengths and block distances

Table 4.2 gives an overview of the logical channels of Layer 1, the available bit rates, block lengths used and the intervals between transmission of blocks. The 14.4 kbit/s data service has been standardized in further GSM standardization phases. Note that the logical channels can suffer from substantial transmission delays depending on the respective use of forward error correction (channel coding and interleaving, see section 4.8 and Table 4.17 below).

Table 4.3 Channel combinations offered by the base station.

	B1	B2	B3	B4	B5	B6	B7	B8	B9
TCH/F									
TCH/H									
TCH/H									
BCCH									
FCCH									
SCH									
CCCH									
SDCCH									
SACCH									
FACCH									

4.1.5 Combinations of logical channels

Not all logical channels can be used simultaneously at the radio interface. They can only be deployed in certain combinations and on certain physical channels. GSM has defined several channel configurations, which are realized and offered by the base stations (Table 4.3). As already mentioned before, an SACCH is always allocated either with a TCH or with an SDCCH, which accounts for the attribute 'associated'.

Depending on its current state, a MS can only use a subset of the logical channels offered by the base station. It uses the channels only in the combinations indicated in Table 4.4. The combination M1 is used in the phase when no physical connection exists, i.e. immediately after the power-up of the MS or after a disruption due to unsatisfactory radio signal conditions. Channel combinations M2 and M3 are used by active MSs in standby mode. In phases requiring a dedicated signaling channel, a MS uses the combination M4, whereas M5 to M8 are used when there is a traffic channel up. M8 is a multislot combination (a MS transmits on several physical channels), where n denotes the number of bidirectional channels, and m denotes the number of unidirectional channels ($n = 1, \dots, 8, m = 0, \dots, 7, n + m = 1, \dots, 8$).

4.2 Physical channels

After discussing the logical channels and their tasks, we now deal with the physical channels, which transport the logical channels via the air interface. We first describe the GSM modulation technique (section 4.2.1), followed by the multiplexing structure (section 4.2.2): GSM is a multicarrier TDMA system, i.e. it employs a combination of FDMA and TDMA for multiple access. This section also covers the explanation of the radio bursts. Finally, section 4.2.3 briefly describes the (optional) frequency hopping technique, which has been standardized to reduce interference.

Table 4.4 Channel combinations used by the mobile station.

	M1	M2	M3	M4	M5	M6	M7	M8
TCH/F								<i>n+m</i>
TCH/H								
TCH/H								
BCCH								
CCCH								
SDCCH								
SACCH								<i>n+m</i>
FACCH								



Figure 4.2 Steps involved in GSM digital modulation.

4.2.1 Modulation

The modulation technique used on the radio channel is Gaussian Minimum Shift Keying (GMSK). GMSK belongs to a family of continuous-phase modulation procedures, which have the special advantages of a narrow transmitter power spectrum with low adjacent channel interference, on the one hand, and a constant amplitude envelope, on the other hand, which allows use of simple amplifiers in the transmitters without special linearity requirements (class C amplifiers). Such amplifiers are especially inexpensive to manufacture, have high degree of efficiency and therefore allow longer operation on a battery charge (David and Benkner, 1996; Watson, 1993).

The digital modulation procedure for the GSM air interface comprises several steps for the generation of a high-frequency signal from channel-coded and enciphered data blocks (Figure 4.2).

The data d_i arrives at the modulator with a bit rate of $1625/6 = 270.83$ kbit/s (gross data rate) and are first differential-coded:

$$\hat{d}_i = (d_i + d_{i-1}) \bmod 2, \quad d_i \in (0; 1).$$

From this differential data, the modulation data are formed, which represents a sequence of Dirac pulses:

$$a_i = 1 - 2\hat{d}_i.$$

This bipolar sequence of modulation data is fed into the transmitter filter – also called a frequency filter – to generate the phase $\varphi(t)$ of the modulation signal. The impulse response $g(t)$ of this linear filter is defined by the convolution of the impulse response $h(t)$ of a

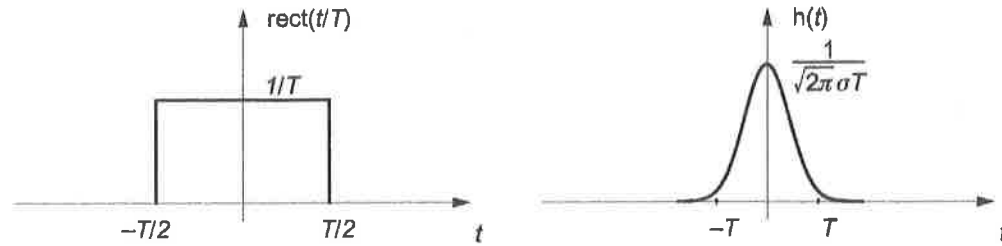


Figure 4.3 Impulse responses for the building blocks of the GSM transmitter filter.

Gaussian low-pass with a rectangular step function:

$$g(t) = h(t) * \text{rect}(t/T),$$

$$\text{rect}(t/T) = \begin{cases} 1/T & \text{for } |t| < T/2, \\ 0 & \text{for } |t| \geq T/2, \end{cases}$$

$$h(t) = \frac{1}{\sqrt{2\pi}\sigma T} \exp\left(\frac{-t^2}{2\sigma^2 T^2}\right), \quad \sigma = \frac{\sqrt{\ln 2}}{2\pi BT}, \quad BT = 0.3.$$

In the equations above, B is the 3 dB bandwidth of the filter $h(t)$ and T the bit duration of the incoming bit stream. The rectangular step function and the impulse response of the Gaussian lowpass are shown in Figure 4.3, and the resulting impulse response $g(t)$ of the transmitter filter is given in Figure 4.4 for some values of BT . Note that with decreasing BT the impulse response becomes broader. For $BT \rightarrow \infty$ it converges to the $\text{rect}()$ function.

In essence, this modulation consists of a Minimum Shift Keying (MSK) procedure, where the data are filtered through an additional Gaussian lowpass before Continuous Phase Modulation (CPM) with the rectangular filter (David and Benkner, 1996). Accordingly it is called GMSK. The Gaussian lowpass filtering has the effect of additional smoothing, but also of broadening the impulse response $g(t)$. This means that, on the one hand, the power spectrum of the signal is made narrower, but, on the other hand, the individual impulse responses are 'smeared' across several bit durations, which leads to increased intersymbol interference. This partial-response behavior has to be compensated for in the receiver by means of an equalizer (David and Benkner, 1996).

The phase of the modulation signal is the convolution of the impulse response $g(t)$ of the frequency filter with the Dirac impulse sequence a_i of the stream of modulation data:

$$\varphi(t) = \sum_i a_i \pi \eta \int_{-\infty}^{t-iT} g(u) du$$

with the modulation index at $\eta = 1/2$, i.e. the maximal phase shift is $\pi/2$ per bit duration. Accordingly, GSM modulation is designated as 0.3-GMSK with a $\pi/2$ phase shift. The phase $\varphi(t)$ is now fed to a phase modulator. The modulated high-frequency carrier signal can then be represented by the following expression, where E_c is the energy per bit of the modulated data rate, f_0 the carrier frequency and φ_0 is a random phase component staying constant during a burst:

$$x(t) = \sqrt{\frac{2E_c}{T}} \cos(2\pi f_0 t + \varphi(t) + \varphi_0).$$

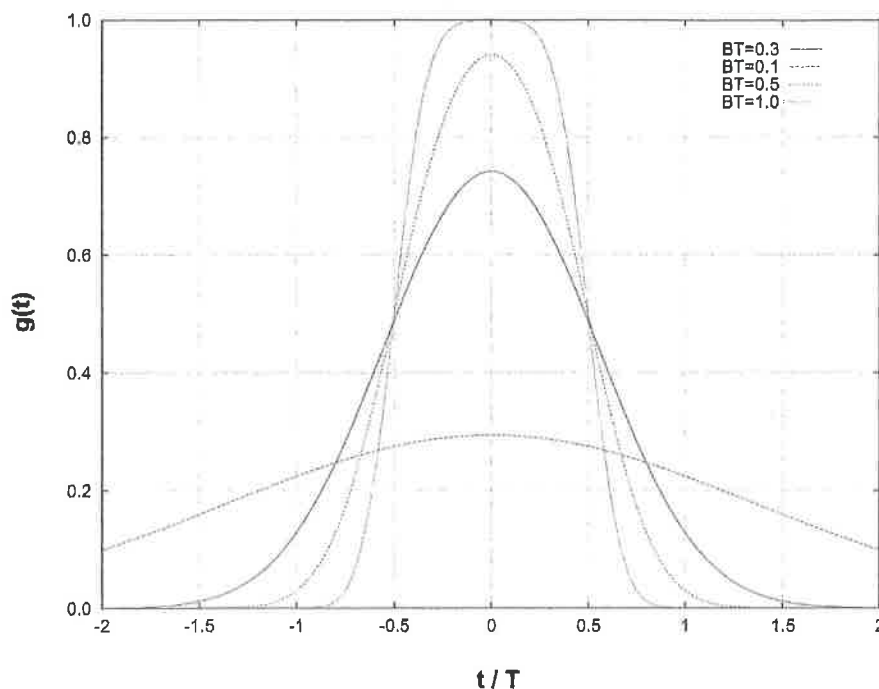


Figure 4.4 Impulse response $g(t)$ of the frequency filter (transmitter filter).

4.2.2 Multiple access, duplexing and bursts

At the physical layer (OSI Layer 1), GSM uses a combination of FDMA and TDMA for multiple access. Two frequency bands 45 MHz apart have been reserved for GSM operation (Figure 4.5): 890–915 MHz for transmission from the MS, i.e. uplink, and 935–960 MHz for transmission from the base station, i.e. downlink. Each of these bands of 25 MHz width is divided into 124 single carrier channels of 200 kHz width. This variant of FDMA is also called Multi-Carrier (MC). In each of the uplink/downlink bands there remains a guardband of 200 kHz. Each Radio Frequency Channel (RFCH) is uniquely numbered, and a pair of channels with the same number form a duplex channel with a duplex distance of 45 MHz (Figure 4.5).

A subset of the frequency channels, the Cell Allocation (CA), is allocated to a base station, i.e. to a cell. One of the frequency channels of the CA is used for broadcasting the synchronization data (FCCH and SCH) and the BCCH. Therefore, this channel is also called the BCCH carrier (see section 4.4). Another subset of the cell allocation is allocated to a MS, the Mobile Allocation (MA). The MA is used among others for the optional frequency hopping procedure (section 4.2.3). Countries or areas which allow more than one mobile network to operate in the same area of the spectrum must have a licensing agency which distributes the available frequency number space (e.g. the Federal Communication Commission in the USA or the 'Bundesnetzagentur' in Germany), in order to avoid collisions and to allow the network operators to perform independent network planning. Here is an example for a possible division: Operator A uses RFCH 2–13, 52–81 and 106–120, whereas operator B receives RFCH 15–50 and 83–103, in which case RFCH 1, 14, 51, 82, 104, 105 and 121–124 are left unused as additional guard bands.

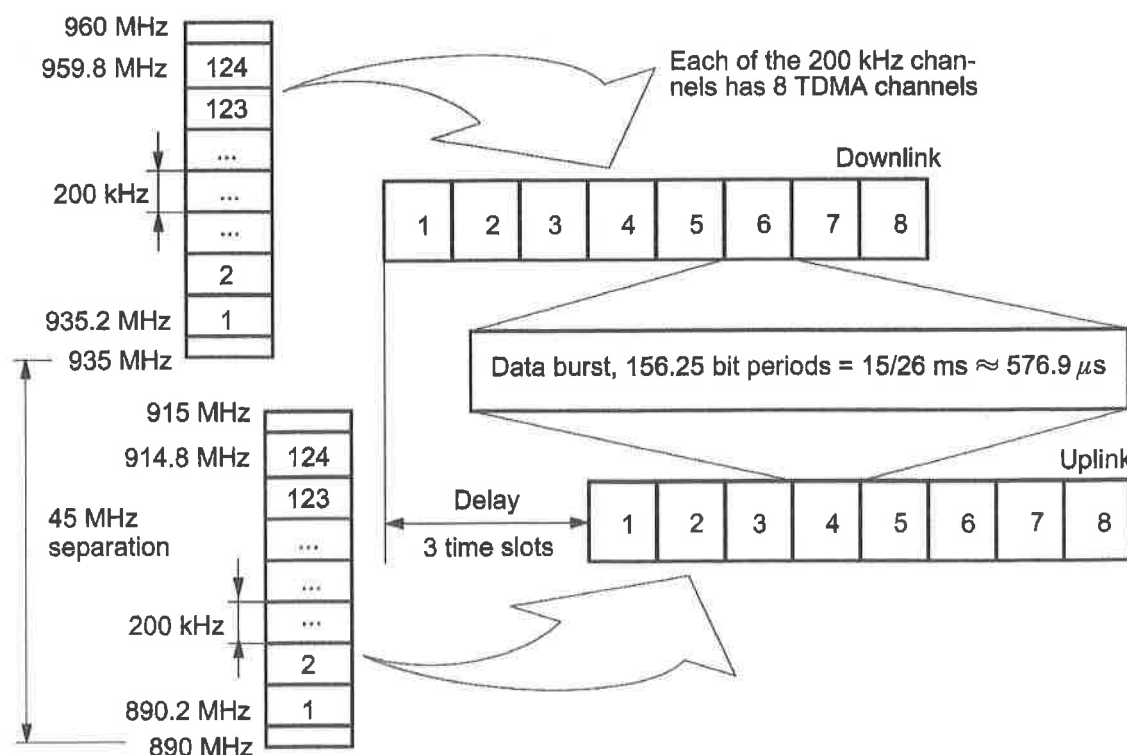


Figure 4.5 Carrier frequencies, duplexing and TDMA frames.

Each of the 200 kHz channels is divided into eight time slots and thus carries eight TDMA channels. The eight time slots together form a TDMA frame (Figure 4.5). The TDMA frames of the uplink are transmitted with a delay of three time slots with regard to the downlink (see Figure 4.6). A MS uses the same time slots in the uplink as in the downlink, i.e. the time slots with the same number (TN). Owing to the shift of three time slots, a MS does not have to send at the same time as it receives, and therefore does not need a duplex unit. This reduces the high-frequency requirements for the front end of the mobile and allows it to be manufactured as a less expensive and more compact unit.

So in addition to the separation into uplink and downlink bands – FDD with a distance of 45 MHz – the GSM access procedure contains a TDD component. Thus, the MS does not need its own high-frequency duplexing unit, which again reduces cost as well as energy consumption.

Each time slot of a TDMA frame lasts for a duration of 156.25 bit periods and, if used, contains a data burst. The time slot lasts $15/26 \text{ ms} = 576.9 \mu\text{s}$; so a frame takes 4.615 ms. The same result is also obtained from the GMSK procedure, which realizes a gross data transmission rate of 270.83 kbit/s per carrier frequency.

There are five kinds of burst (Figure 4.7):

- Normal Burst (NB): The NB is used to transmit information on traffic and control (except RACH) channels. The individual bursts are separated from each other by guard periods during which no bits are transmitted. At the start and end of each burst are three tail bits which are always set to logical '0'. These bits fill a short time span

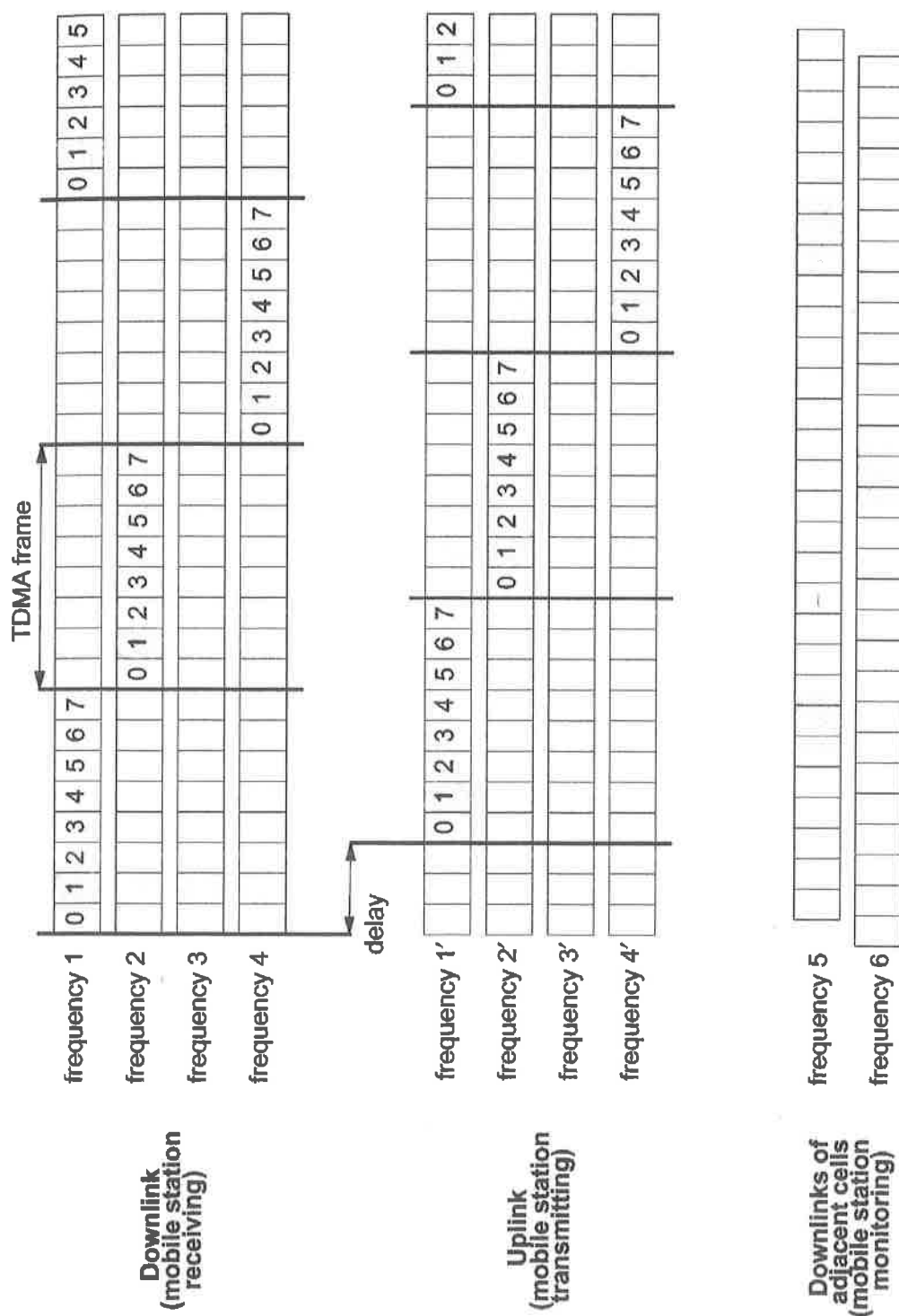


Figure 4.6 GSM full-rate traffic channel with frequency hopping.

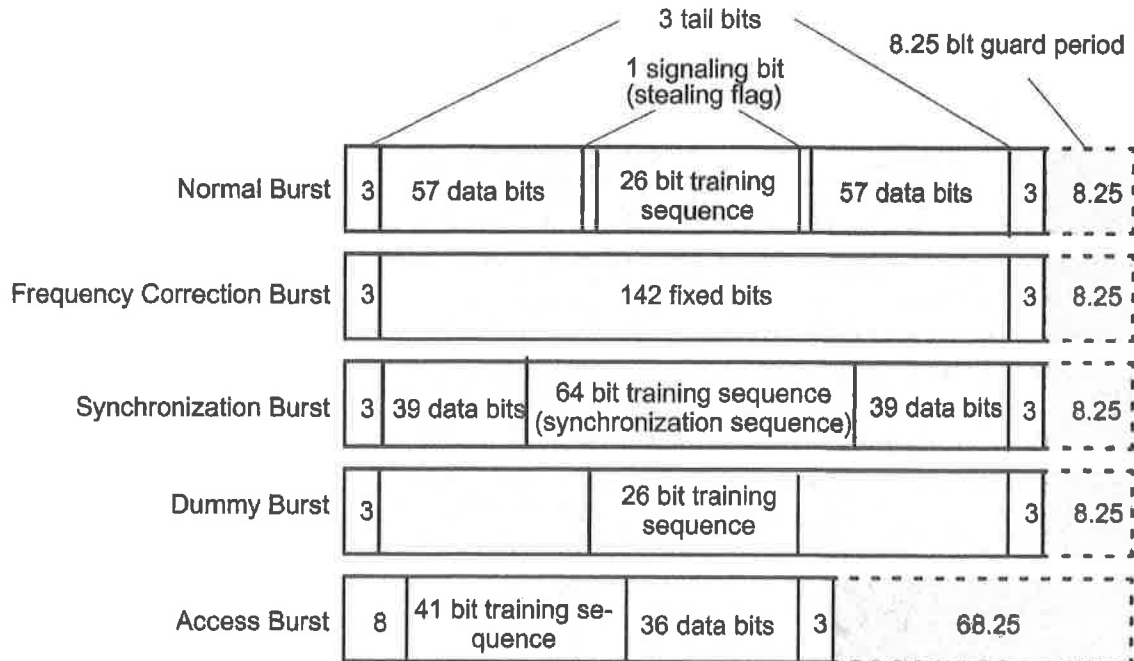


Figure 4.7 Bursts of the GSM TDMA procedure.

during which transmitter power is ramped up or ramped down and during which no data transmission is possible. Furthermore, the initial zero bits are also needed for the demodulation process. The Stealing Flags (SFs) are signaling bits which indicate whether the burst contains traffic data or signaling data. They are set to allow use of single time slots of the TCH in preemptive multiplexing mode, e.g. when, during a handover, fast transmission of signaling data on the FACCH is needed. This causes a loss of user data, i.e. these time slots are 'stolen' from the traffic channel, hence the name SF. In addition to the synchronization and signaling bits (Figure 4.7), the NB also contains two blocks of 57 bits each of error-protected and channel-coded user data separated by a 26-bit midamble. This midamble consists of predefined, known bit patterns, the training sequences, which are used for channel estimation to optimize reception with an equalizer and for synchronization. With the help of these training sequences, the equalizer eliminates or reduces the intersymbol interferences which are caused by propagation time differences of the multipath propagation. Time differences of up to 16 μ s can be compensated for. Eight different training sequences are defined for the NB which are designated by the Training Sequence Code (TSC). Initially, the TSC is obtained when the BCC is obtained, which is transmitted as part of the BSIC (see Chapter 3). Beyond that, training sequences can be individually assigned to mobile stations. In this case the TSC is contained in the Layer 3 message of the channel assignment (TCH or SDCCH). In this way the base station tells a MS which training sequence it should use with NBs of a specific traffic channel.

- **Frequency Correction Burst (FB):** This burst is used for the frequency synchronization of a MS. The repeated transmission of FBs is also called the FCCH. Tail bits as well as

data bits are all set to 0 in the FB. Owing to the GSM modulation procedure (0.3-GMSK) this corresponds to broadcasting an unmodulated carrier with a frequency shift of 1625/24 kHz above the nominal carrier frequency. This signal is periodically transmitted by the base station on the BCCH carrier. It allows time synchronization with the TDMA frame of a MS as well as the exact tuning to the carrier frequency. Depending on the stability of its own reference clock, the mobile can periodically resynchronize with the base station using the FCCH.

- **Synchronization Burst (SB):** This burst is used to transmit information which allows the MS to synchronize time-wise with the BTS. In addition to a long midamble, this burst contains the running number of the TDMA frame, the RFN and the BSIC; the RFN is covered in section 4.3. Repeated broadcasting of SBs is considered as the SCH.
- **Dummy Burst (DB):** This burst is transmitted on one frequency of the CA, when no other bursts are to be transmitted. The frequency channel used is the same as that which carries the BCCH, i.e. it is the BCCH carrier. This ensures that the BCCH transmits a burst in each time slot which enables the MS to perform signal power measurements of the BCCH, a procedure also known as quality monitoring.
- **Access Burst (AB):** This burst is used for random access to the RACH without reservation. It has a guard period significantly longer than the other bursts. This reduces the probability of collisions, since the MSs competing for the RACH are not (yet) time-synchronized.

A single user gets one-eighth or 33.9 kbit/s of the gross data rate of 270.83 kbit/s. Considering a normal burst, 9.2 kbit/s are used for signaling and synchronization, i.e. tail bits, SFs and training sequences, including guard periods. The remaining 24.7 kbit/s are available for the transmission of (raw) user or control data on the physical layer.

4.2.3 Optional frequency hopping

Mobile radio channels suffer from frequency-selective interferences, e.g. frequency-selective fading due to multipath propagation phenomena. This selective frequency interference can increase with the distance from the base station, especially at the cell boundaries and under unfavorable conditions. Frequency hopping procedures change the transmission frequencies periodically and thus average the interference over the frequencies in one cell. This leads to a further improvement in the Signal-to-Noise Ratio (SNR) to a high enough level for good speech quality, so that conversations with acceptable quality can be conducted. GSM systems achieve a good speech quality with a SNR of about 11 dB. With frequency hopping a value of 9 dB is sufficient. GSM provides for an optional frequency hopping procedure which changes to a different frequency with each burst; this is known as slow frequency hopping. The resulting hopping rate is about 217 changes per second, corresponding to the TDMA frame duration. The frequencies available for hopping, the hopping assignment, are taken from the CA. The principle is illustrated in Figure 4.6, showing the time slot allocations for a full-rate TCH. The exact synchronization is determined by several parameters: the MA, a Mobile Allocation Index Offset (MAIO), a Hopping Sequence Number (HSN) and the TDMA Frame Number (FN); see section 4.3. The use of frequency hopping is an option left to the network operator, which can be decided on an individual cell basis. Therefore, a MS

must be able to switch to frequency hopping if a base station notices adverse conditions and decides to activate frequency hopping.

4.2.4 Summary

A physical GSM channel is defined by a sequence of frequencies and a sequence of TDMA frames. The RFCH sequence is defined by the frequency hopping parameters, and the temporal sequence of time slots of a physical channel is defined as a sequence of frame numbers and the time slot number within the frame. Frequencies for the uplink and downlink are always assigned as a pair of frequencies with a 45 MHz duplex separation.

As shown above, GSM uses a series of parameters to define a specific physical channel of a base station. Summarizing, these parameters are:

- MAIO;
- HSN;
- TSC;
- Time Slot Number (TN);
- MA, also known as RFCH Allocation;
- type of logical channel carried on this physical channel;
- the number of the logical subchannel (if used) – the Subchannel Number (SCN).

Within a logical channel, there can be several subchannels (e.g. subrate multiplexing of the same channel type). The TDMA frame sequence can be derived from the type of the channel and the logical subchannel if present.

4.3 Synchronization

For the successful operation of a mobile radio system, synchronization between MSs and the base station is necessary. Two kinds of synchronization are distinguished: frequency synchrony and time synchrony of the bits and frames.

Frequency synchronization is necessary so that transmitter and receiver frequencies agree. The objective is to compensate for tolerances of the less-expensive and, therefore, less-stable oscillators in the mobile stations by obtaining an exact reference from the base station and to follow it.

Bit and frame synchrony are important in two regards for TDMA systems. First, the propagation time differences of signals from different MSs have to be adjusted, so that the transmitted bursts are received synchronously with the time slots of the base station and that bursts in adjacent time slots do not overlap and interfere with each other. Second, synchrony is needed for the frame structure since there is a higher-level frame structure superimposed on the TDMA frames for multiplexing logical signaling channels onto one physical channel. The synchronization procedures defined for GSM are explained in the following section.

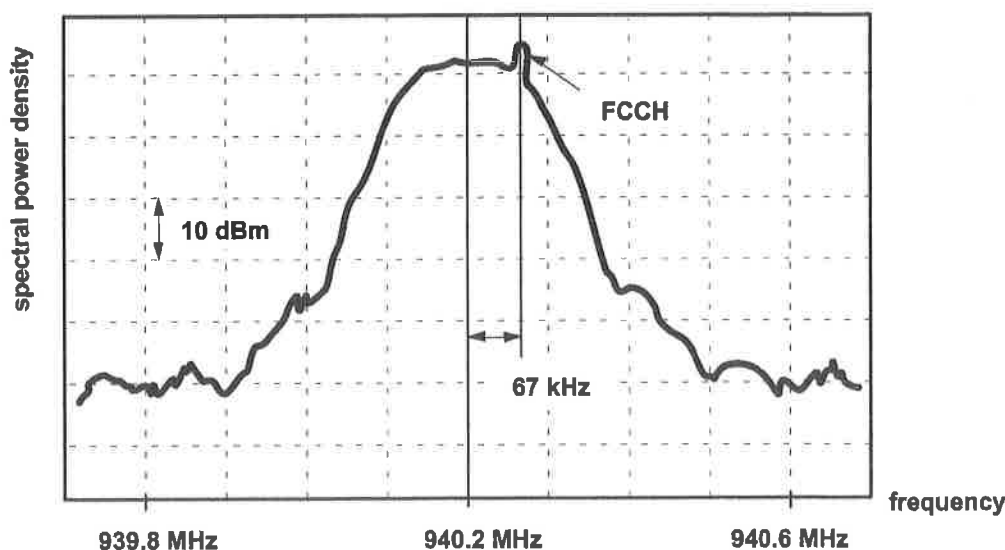


Figure 4.8 Typical power spectrum of a BCCH carrier.

4.3.1 Frequency and clock synchronization

A GSM base station transmits signals on the frequency carrier of the BCCH which allow a MS to synchronize with the base station. Synchronization means on the one hand the time-wise synchronization of the MS and base station with regards to bits and frames, and on the other hand tuning the MS to the correct transmitter and receiver frequencies.

For this purpose, the BTS provides the following signals (Figure 4.7):

- SCH with SBs and extra long training sequences, which facilitate synchronization;
- FCCH with FBs.

Owing to the 0.3-GMSK modulation procedure used in GSM, a data sequence of logical '0' generates a pure sine wave signal, i.e. broadcasting of the FB corresponds to an unmodulated carrier (frequency channel) with a frequency shift of $1625/24 \text{ kHz} (\approx 67.7 \text{ kHz})$ above the nominal carrier frequency (Figure 4.8). In this way, the MS can keep exactly synchronized by periodically monitoring the FCCH. On the other hand, if the frequency of the BCCH is still unknown, the MS can search for the channel with the highest signal level. This channel is with all likelihood a BCCH channel, because DBs must be transmitted on all unused time slots in this channel, whereas not all time slots are always used on other carrier frequencies. Using the FCCH sine wave signal allows identification of a BCCH and synchronization of a MSs oscillator.

For the time synchronization, TDMA frames in GSM are cyclically numbered *modulo* 2 715 648 ($= 26 \times 51 \times 2^{11}$) with the FN. One cycle generates the so-called hyperframe structure which comprises 2 715 648 TDMA frames. This long numbering cycle of TDMA frames is used to synchronize the ciphering algorithm at the air interface (see section 5.6). Each base station BTS periodically transmits the RFN on the SCH. With each SB the mobiles thus receive information about the number of the current TDMA frame. This enables each MS to be time-synchronized with the base station.

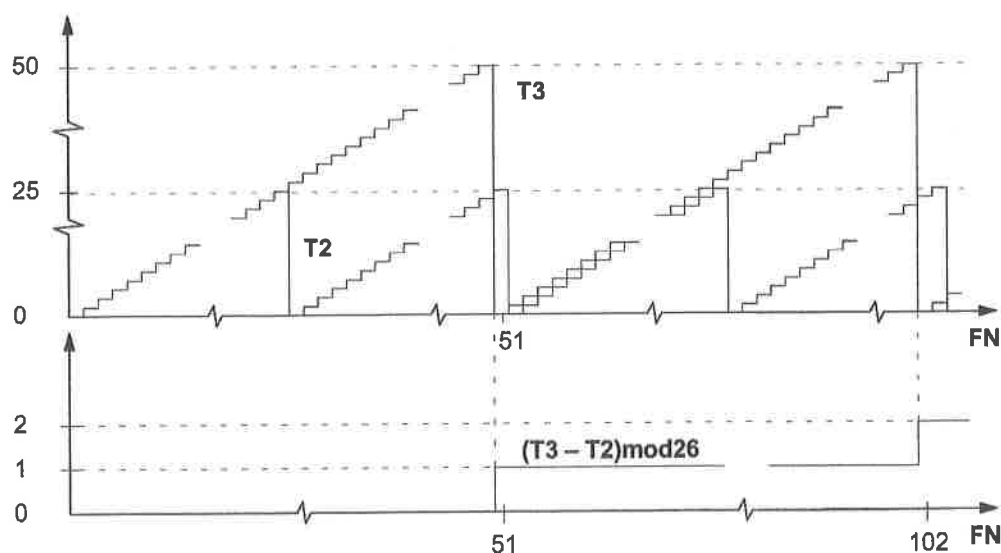


Figure 4.9 Values T2 and T3 for the calculation of RFN.

The RFN has a length of 19 bits. It consists of three fields: T1 (11 bits), T2 (5 bits) and T3' (3 bits). These three fields are defined by (with div designating integer division)

$$T1 = FN \text{ div } (26 \times 51) [0-2047],$$

$$T2 = FN \text{ mod } 26 [0-25],$$

$$T3' = (T3 - 1) \text{ div } 10 [0-4], \quad \text{with } T3 = FN \text{ mod } 51 [0-50].$$

The sequences of running values of T2 and T3 are illustrated in Figure 4.9. The value crucial for the reconstruction of the FN is the difference between the two fields. The time synchronization of a MS and its time slots, TDMA frames and control channels is based on a set of counters which run continuously, independent of MS or base station transmission. Once these counters have been started and initialized correctly, the MS is in a synchronized state with the base station. The following four counters are kept for this purpose:

- quarter bit counter counting the Quarter Bit Number (QN);
- bit counter counting the Bit Number (BN);
- time slot counter counting the TN;
- frame counter counting the FN.

Owing to the bit and frame counting, these counters are of course interrelated, namely in such a way that the subsequent counter counts the overflows of the preceding counter. The following principle is used (Figure 4.10): QN is incremented every 12 or 13 μs ; BN is obtained from this by integer division ($BN = QN \text{ div } 4$). With each transition from 624 to 0 TN is incremented, and each overflow of TN increments the frame counter FN by 1.

The timers can be reset and restarted when receiving an SB. The quarter bit counter is set by using the timing of the training sequence of the burst, whereas the TN is reset to 0 with

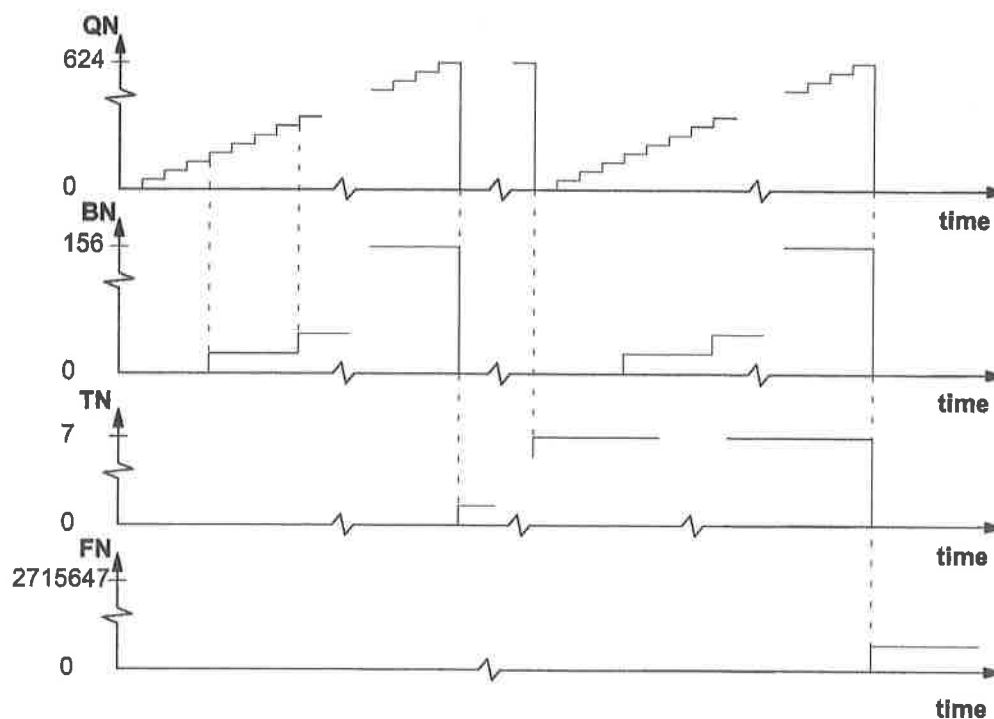


Figure 4.10 Synchronization timers, simplified: the TDMA frame duration is 156.25 bit times.

the end of the burst. The FN can then be calculated from the RFN transmitted on the SCH:

$$FN = 51 \times ((T3 - T2) \bmod 26) + T3 + 51 \times 26 \times T1, \quad \text{with } T3 = 10 \times T3' + 1.$$

It is important to recalculate $T3$ from $T3'$ although, because of the binary representation, only the integer part of the division by 10 is taken into account.

If the optional frequency hopping procedure is used (see section 4.2.3), an additional mapping of the TDMA frame number onto the frequency to be used is required in addition to the evaluation of the synchronization signals from the FCCH and SCH. One has to obtain the index number of the frequency channel on which the current burst has to be transmitted from the MA table. This process uses a predefined RFNTABLE, the FN and a HSN (Figure 4.11). The MA holds N frequencies, with a maximum value of 64 for N . With this procedure, every burst is sent on a different frequency in a cyclic way.

4.3.2 Adaptive frame synchronization

The MS can be anywhere within a cell, which means the distance between MS and base station may vary. Thus, the signal propagation times between MS and base station vary. Owing to the mobility of the subscribers, the bursts received at the base would be offset. The TDMA procedure cannot tolerate such time shifts, since it is based on the exact synchronization of transmitted and received data bursts. Bursts transmitted by different MSs in adjacent time slots must not overlap when received at the base station by more than the guard period (Figure 4.7), even if the propagation times within the cell are very different.

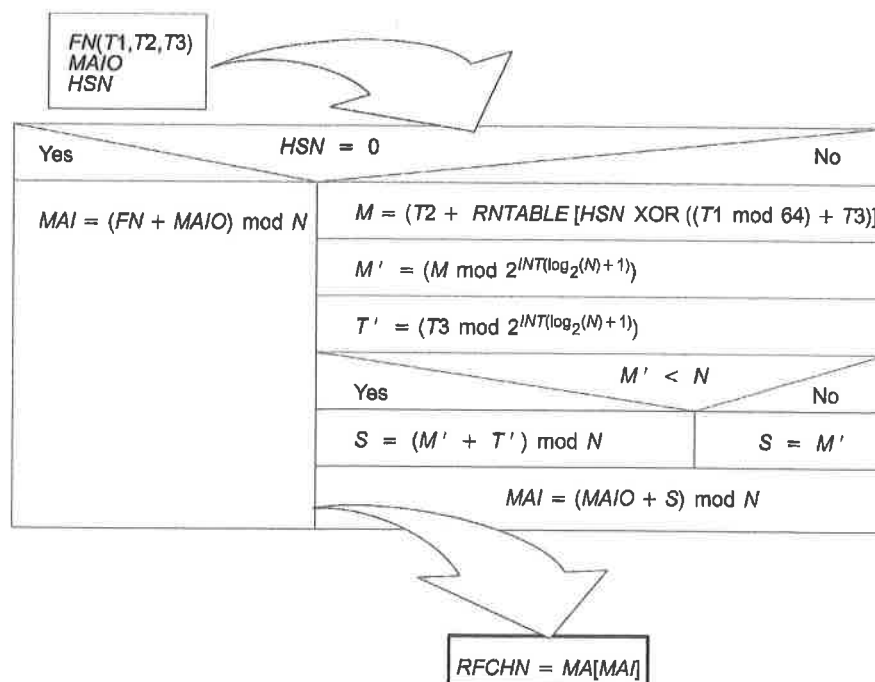


Figure 4.11 Generation of the GSM frequency hopping sequence.

To avoid such collisions, the start of transmission time from the MS is advanced in proportion to the distance from the base station. The process of adapting the transmissions from the MSs to the TDMA frame is called adaptive frame alignment.

For this purpose, the parameter Timing Advance (TA) in each SACCH Layer 1 protocol block is used (section 4.5). The MS receives the TA value it must use from the base station on the SACCH downlink; it reports the actually used value on the SACCH uplink. There are 64 steps for the timing advance which are coded as 0 to 63. One step corresponds to one bit period. Step 0 means no timing advance, i.e. the frames are transmitted with a time shift of three slots or 468.75 bit durations with regards to the downlink. At step 63, the timing of the uplink is shifted by 63 bit durations, such that the TDMA frames are transmitted on the uplink only with a delay of 405.75 bit durations. So the required adjustment always corresponds to twice the propagation time or is equal to the round-trip delay (Figure 4.12). In this way, the available range of values allows a compensation over a maximum propagation time of 31.5 bit periods ($\approx 113.3 \mu\text{s}$). This corresponds to a maximum distance between mobile and base station of 35 km. A GSM cell may therefore have a maximum diameter of 70 km. The distance from the base station or the currently valid TA value for a MS is therefore an important handover criterion in GSM networks (section 6.4.3).

The adaptive frame alignment technique is based on continuous measurement of propagation delays by the base station and corresponding timing advance activity by the MS. In the case of an (unreserved) random access to the RACH, a channel must first be established. The base station has in this case not yet had the opportunity to measure the distance to the MS and to transmit a corresponding timing advance command. If a MS transmits an access burst in the current time slot, it uses a timing advance value of 0 or a default value. To minimize collisions

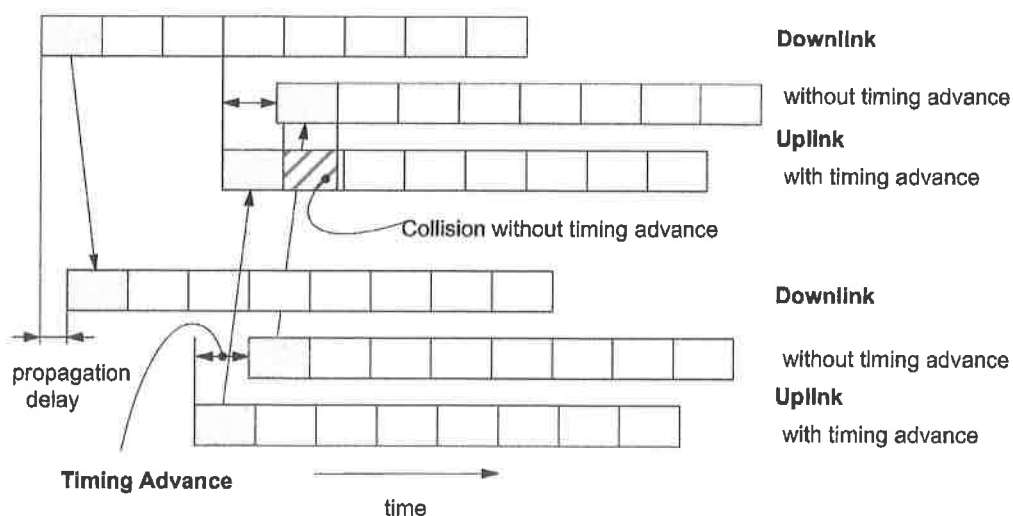


Figure 4.12 Operation of timing advance.

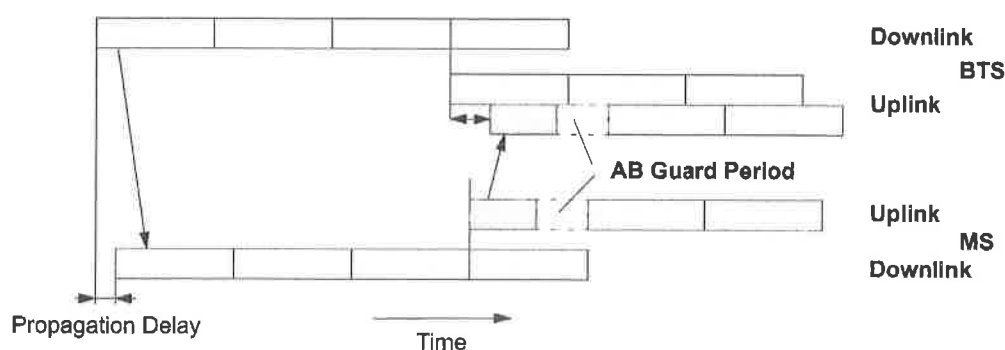


Figure 4.13 Timing for RACH random multiple access.

with subsequent time slots at the base station, the access burst (AB) has to be correspondingly shorter than the time slot duration (Figure 4.13). This explains the long guard period of the AB of 68.25 bit periods, which can compensate for the propagation delay if a MS sends an AB from the boundary of a cell of 70 km diameter.

4.4 Mapping of logical onto physical channels

The mapping of logical channels onto physical channels has two components: mapping in frequency and mapping in time. The mapping of a logical channel onto a physical channel in the frequency domain is based on the FN, the frequencies allocated to base stations and MSs – CA and MA – and the rules for the optional frequency hopping (see section 4.2.3).

In the time domain, logical channels are transported in the corresponding time slots of the physical channel. They are mapped onto physical channels in certain time-multiplexed combinations, where they can occupy a complete physical channel or just part of a physical

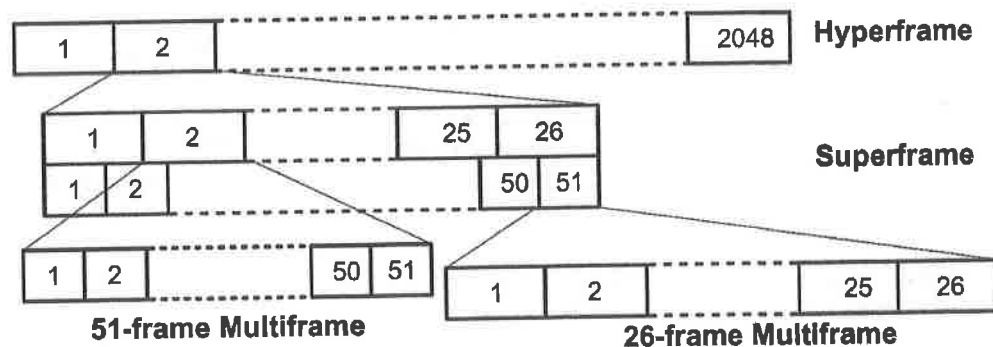


Figure 4.14 GSM frame structures.

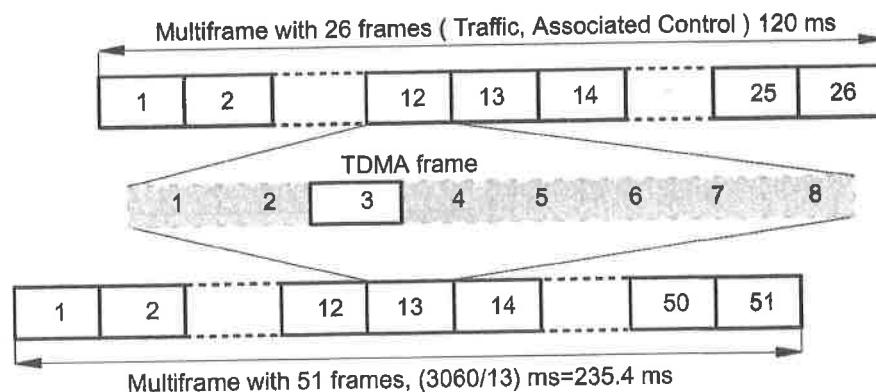


Figure 4.15 GSM multiframes.

channel. Whereas user payload data is allocated a dedicated full-rate or half-rate channel, logical signaling (control) channels have to share a physical channel.

The logical channels are organized by the definition of complex superstructures on top of the TDMA frames, forming so-called multiframes, superframes and hyperframes (Figure 4.14). For the mapping of logical onto physical channels, we are interested in the multiframe domain. These multiframes allow us to map (logical) subchannels onto physical channels. Two kinds of multiframes are defined (Figure 4.15): a multiframe consisting of 26 TDMA frames (predominantly payload – speech and data – frames) and a multiframe of 51 TDMA frames (predominantly signaling frames).

Each hyperframe is divided into 2048 superframes. With its long cycle period of 3 h 28 min 53.760 s, it is used for the synchronization of user data encryption. A superframe consists of 1326 consecutive TDMA frames which therefore lasts for 6.12 s, the same length as 51 multiframes of 26 TDMA frames or 26 multiframes of 51 TDMA frames. These multiframes are again used to multiplex the different logical channels onto a physical channel as shown below.

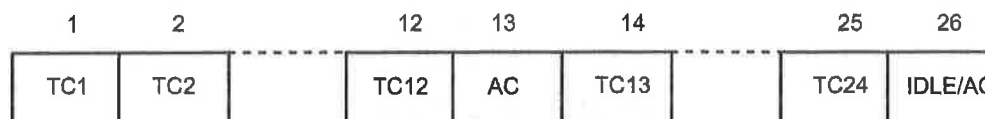


Figure 4.16 Channel organization in a 26-frame multiframe.

4.4.1 26-frame multiframe

Each 26 subsequent TDMA frames form a multiframe which multiplexes two logical channels, a TCH and the SACCH, onto the physical channel (Figure 4.16). This process uses only one time slot per TDMA frame for the corresponding multiframe (e.g. time slot 3 in Figure 4.15), since a physical channel consists of just one time slot per TDMA frame. In addition to the 24 TCH frames for user data, this multiframe also contains an Associated Control (AC) frame for signaling data (SACCH data). One frame (the 26th) remains unused in the case of a full-rate TCH (IDLE/AC); it is reserved for the introduction of two half-rate TCHs; then the 26th frame will be used to carry the SACCH channels of the other half.

The data of the FACCH is transmitted by occupying one-half of the bits in eight consecutive bursts, by 'stealing' these bits from the TCH. For this purpose, the SFs of the normal bursts are set (Figure 4.7).

A subscriber has available a gross data rate of $271/8 = 33.9$ kbit/s (section 4.2). Of this budget, 9.2 kbit/s are for signaling, synchronization and guard periods of the burst. Of the remaining 24.7 kbit/s, in the case of the 26-frame multiframe, 22.8 kbit/s are left for the coded and enciphered user data of a full-rate channel and 1.9 kbit/s remain for the SACCH and IDLE.

4.4.2 51-frame multiframe

For the transmission of the control channels which are not associated with a TCH (all except FACCH and SACCH), a multiframe is formed consisting of 51 consecutive TDMA frames (Figure 4.7). According to the channel configuration (section 4.1), the multiframe is used differently. In each case, multiframes of 51 TDMA frames serve the purpose of mapping several logical channels onto a physical channel.

Furthermore, some of these control channels are unidirectional, which results in different structures for uplink and downlink. For some configurations, two adjacent multiframes are required to map all of the logical channels. Some examples are illustrated in Figure 4.17. They correspond to the combinations B2, B3 and B4 in Table 4.3 whereas for channels SDCCH and SACCH some four or eight logical subchannels have been defined (D0, D1, ..., A0, A1, ...). One of the frequency channels of the CA of a base station is used to broadcast synchronization data (FCCH and SCH) and the BCCH. Since the base station has to transmit in each time slot of the BCCH carrier to enable a continuous measurement of the BCCH carrier by the MS, a DB is transmitted in all time slots with no traffic.

On time slot 0 of the BCCH carrier, only two combinations of logical channels may be transmitted, the combinations B2 or B3 from Table 4.3: (BCCH + CCCH + FCCH + SCH + SDCCH + SACCH or BCCH + CCCH + FCCH + SCH). No other time slot of the CA must carry this combination of logical channels.

Table 4.5 Mapping of the frame number onto a BCCH message.

TC	System information message
0	Type 1
1	Type 2
2, 6	Type 3
3, 7	Type 4
4, 5	Any (optional)

As one can see in Figure 4.17, in the time slot 0 of the BCCH carrier of a base station (downlink) the frames 1, 11, 21, ... are FCCH frames, and the subsequent frames 2, 12, 22, ... form SCH frames. Frames 3, 4, 5 and 6 of the 51-frame BCCH multiframe transport the appropriate BCCH information, whereas the remaining frames may contain different combinations of logical channels. Once the MS has synchronized by using the information from FCCH and SCH, it can determine from the information in the FCCH and SCCH how the remainder of the BCCH is constructed. For this purpose, the base station radio resource management periodically transmits a set of messages to all MSs in this cell.

These system information messages comprise six types, of which only types 1–4 are of interest here. Using FN, one can determine which type is to be sent in the current time slot by calculating a Type Code (TC):

$$TC = (FN \text{ div } 51) \bmod 8.$$

Table 4.5 shows how the TC determines the type of the system information message to be sent within the current multiframe.

Of the parameters contained in such a message, the following are of special interest. BS_CC_CHANS determines the number of physical channels which support a CCCH. The first CCCH is transmitted in time slot 0, the second in time slot 2, the third in time slot 4 and the fourth in time slot 6 of the BCCH carrier. Another parameter, BS_CCCH_SDCCH_COMB, determines whether the DCCHs SDCCH(0-3) and SACCH(0-3) are transmitted together with the CCCH on the same physical channel. In this case, each of these dedicated control channels consists of four subchannels.

Each of the CCCHs of a base station is assigned a group CCCH_GROUP of MSs. MSs are allowed random access (RACH) or receive paging information (PCH) only on the CCCH assigned to this group. Furthermore, a MS needs only to listen for paging information on every N th block of the PCH. The number N is determined by multiplying the number of paging blocks per 51-frame multiframe of a CCCH with the parameter BS_PA_MFRMS designating the number of multiframes between paging frames of the same paging group (PAGING_GROUP).

In cells with high traffic, in particular, the CCCH and paging groups serve to subdivide traffic and to reduce the load on the individual CCCHs. For this purpose, there is a simple algorithm which allows each MS to calculate its respective CCCH_GROUP and PAGING_GROUP from its IMSI and parameters BS_CC_CHANS, BS_PA_MFRMS and N .

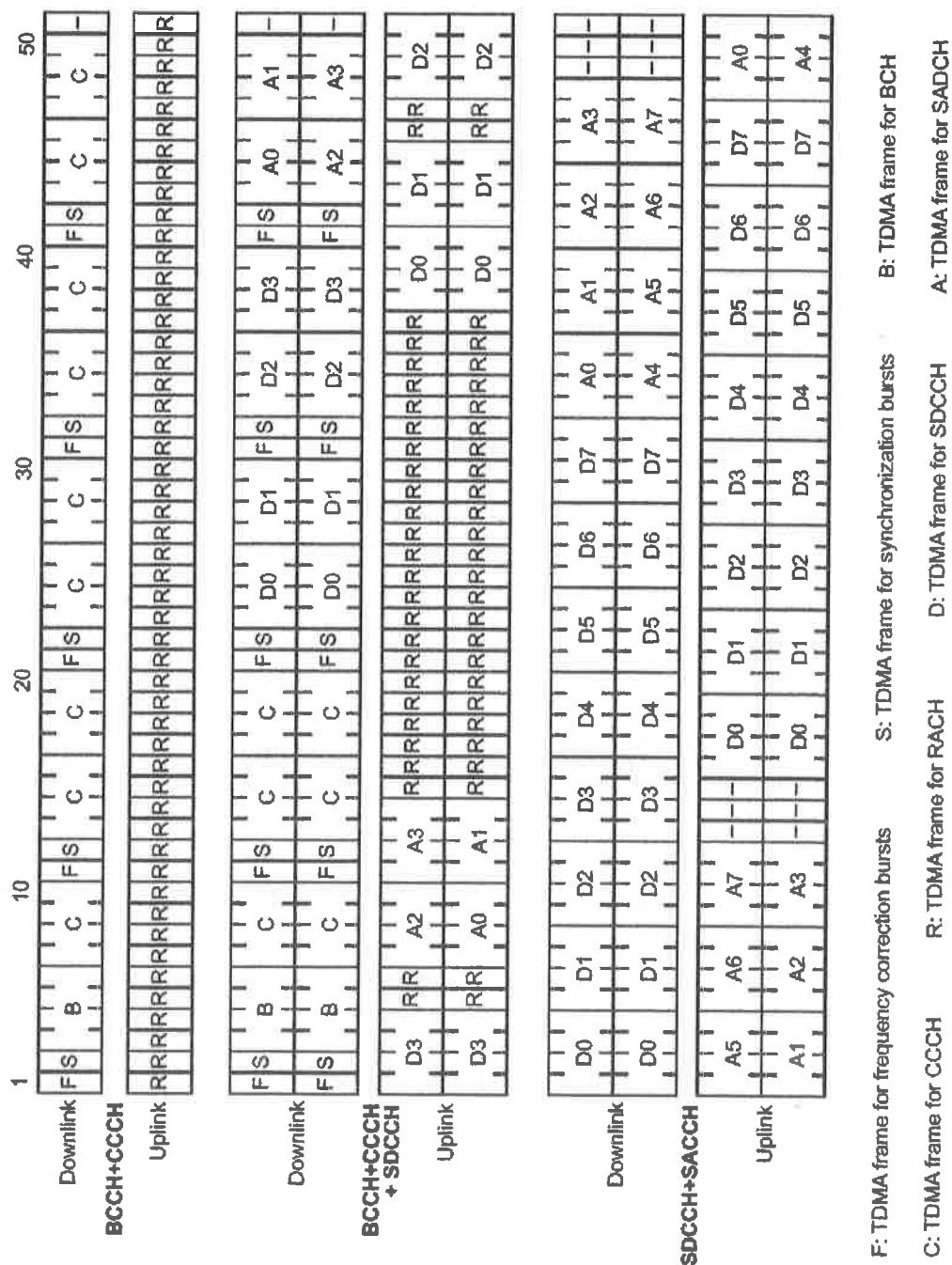


Figure 4.17 Channel organization in a 51-frame multiframe.

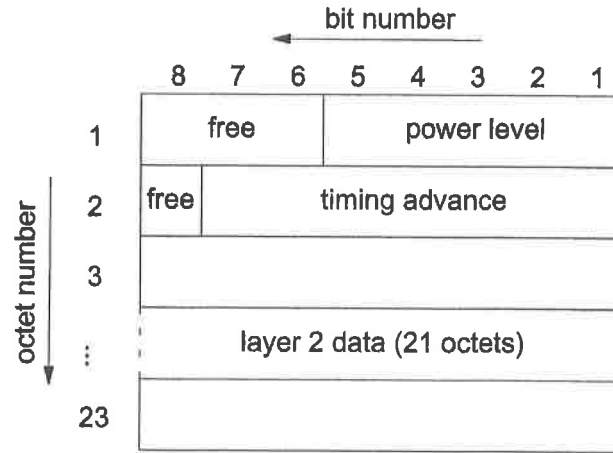


Figure 4.18 SACCH block format.

4.5 Radio subsystem link control

The radio interface is characterized by another set of functions of which we discuss only the most important in the following. One of these functions is the control of the radio link: radio subsystem link control, with the main activities of received-signal quality measurement (quality monitoring) for cell selection and handover preparation, and of transmitter power control.

If there is no active connection, i.e. if the MS is at rest, the BSS has no tasks to perform. The MS, however, is still committed to continuously observing the BCCH carrier of the current and neighboring cells, so that it would be able to select the cell in which it can communicate with the highest probability. If a new cell needs to be selected, a location update may become necessary.

During a connection (TCH or SDCCH), the functions of channel measurement and power control serve to maintain and optimize the radio channel; this also includes adaptive frame alignment (section 4.3.1) and frequency hopping (section 4.2.3). Both need to be done until the current base can hand over the current connection to the next base station.

These link control functions are performed over the SACCH channel. Two fields are defined in an SACCH block (Figure 4.18) for this purpose, the power level and the TA. On the downlink, these fields contain values as assigned by the BSS. On the uplink, the MS inserts its currently used values. The quality monitoring measurement values are transmitted in the data part of the SACCH block.

The following illustrates the basic operation of the radio subsystem link control at the BSS side for an existing connection; the detailed explanation of the respective functions is given later. In principle, the radio link control can be subdivided into three tasks: measurement collection and processing, transmitter power control, and handover control.

In the example of Figure 4.19, the process BSS_Link_Control starts at initialization the processes BSS_Power_Control and BSS_HO_Control and then enters a measurement loop, which is only left when the connection is terminated. In this loop, measurement data are periodically received (every 480 ms) and current mean values are calculated. At first, these measurement data are supplied to the transmitter power control to adapt the power of MS and

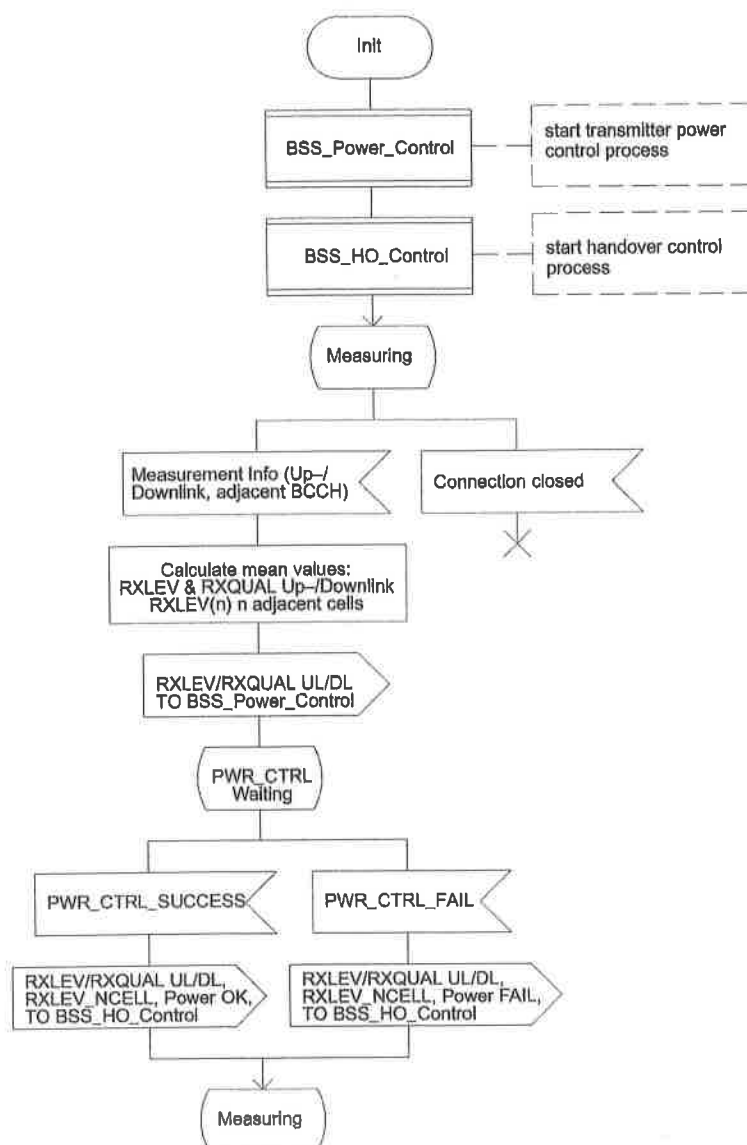


Figure 4.19 Principal operation of the radio subsystem link control.

BSS to a new situation if necessary. Thereafter, the measurement data and the result of the power control activity are supplied to the handover process, which can then decide whether a handover is necessary or not.

4.5.1 Channel measurement

The task of radio subsystem link control in the MS includes identification of the reachable base stations and measurement of their respective received signal level and channel quality (quality monitoring task). In idle mode, these measurements serve to select the current base

Table 4.6 Measurement range of the received signal level.

Level	Received signal level (dBm)	
	From	To
RXLEV_0	—	–110
RXLEV_1	–110	–109
⋮	⋮	⋮
RXLEV_62	–49	–48
RXLEV_63	–48	—

station, whose PCH is then periodically examined and on whose RACH desired connections can be requested.

During a connection, i.e. on a TCH or SDCCH with respective SACCH/FACCH, this measurement data are transmitted on the SACCH to the base station as a measurement report/measurement info. These reports serve as inputs for the handover and power control algorithms.

The measurement objects are, on the one hand, the uplink and downlink of the current channel (TCH or SDCCH) and, on the other hand, the BCCH carriers which are continuously broadcast with constant power by all BTSs in all time slots. It is especially important to keep the transmitter power of the BCCH carriers constant to allow comparisons between neighboring base stations. A list of neighboring base station's BCCH carrier frequencies, called the BCCH Allocation (BA) is supplied to each mobile by its current BTS, to enable measurement of all cells which are candidates for a handover. The cell identity is broadcast as the BSIC on the BCCH. Furthermore, up to 36 BCCH carrier frequencies and their BSICs can be stored on the SIM card. In principle, GSM uses two parameters to describe the quality of a channel: the Received Signal Level (RXLEV), measured in dBm, and the Received Signal Quality (RXQUAL), measured as bit error ratio as a percentage before error correction (Tables 4.6 and 4.7). The received signal power is measured continuously by MSs and base stations in each received burst within a range of –110 to –48 dBm. The respective RXLEV values are obtained by averaging.

The bit error ratio before error correction can be determined in a variety of ways. For example, it can be estimated from information obtained from channel estimation for equalization from the training sequences, or the number of erroneous (corrected) bits can be determined through repeated coding of the decoded, error-corrected data blocks and comparison with the received data. Since the data before error correction is presented as blocks of 456 bits (section 4.8), the bit error ratio can only be given with a quantizing resolution of 2×10^{-3} . Again, the value of RXQUAL is determined from this information by averaging.

Channel measurement during idle mode

In idle mode (see also Figure 5.17) the MS must always stay aware of its environment. The main purpose is to be able to assign a MS to a cell, whose BCCH carrier it can decode reliably. If this is the case, the MS is able to read system and paging information. If there is a

Table 4.7 Measurement range of the bit error ratio.

Level	Bit error ratio (%)	
	From	To
RXQUAL_0	–	0.2
RXQUAL_1	0.2	0.4
RXQUAL_2	0.4	0.8
RXQUAL_3	0.8	1.6
RXQUAL_4	1.6	3.2
RXQUAL_5	3.2	6.4
RXQUAL_6	6.4	12.8
RXQUAL_7	12.8	–

desire to set up a connection, the MS can most likely communicate with the network. There are two possible starting situations:

- the MS has no *a priori* knowledge about the network at hand, especially which BCCH carrier frequencies are in use;
- the MS has a stored list of BCCH carriers.

In the first case, the more unfavorable of the two, the mobile has to search through all of the 124 GSM frequencies, measure their signal power level and calculate an average from at least five measurements. The measurements of the individual carriers should be evenly distributed over an interval of 3–5 s. After at most 5 s, a minimum of 629 measurement values are available that allow the 124 RXLEV values to be determined. The carriers with the highest RXLEV values are very likely BCCH carriers, since continuous transmission is required on them. Final identification occurs with the FB of the FCCH. Once the received BCCH carriers have been found, the MS starts to synchronize with each of them and reads the system information, beginning with the BCCH with the highest RXLEV value.

This orientation concerning the current location can be accelerated considerably, if a list of BCCH carriers has been stored on the SIM card. Then the MS tries first to synchronize with some known carrier. Only if it cannot find any of the stored BCCH carrier frequencies, it does start with the normal BCCH search. A MS can store several lists for the recently visited networks.

Channel measurement during a connection

During a traffic (TCH) or signaling (SDCCH) connection, the channel measurement of the MS occurs over an SACCH interval, which comprises 104 TDMA frames in the case of a TCH channel (480 ms) or 102 TDMA frames (470.8 ms) in the case of an SDCCH channel.

For the channel at hand, two parameters are determined: the received signal level RXLEV and the signal quality RXQUAL. These two values are averaged over an SACCH interval (480 or 470.8 ms) and transmitted to the base station on the SACCH as a measurement report/measurement info. In this way the downlink quality of the channel assigned to the MS

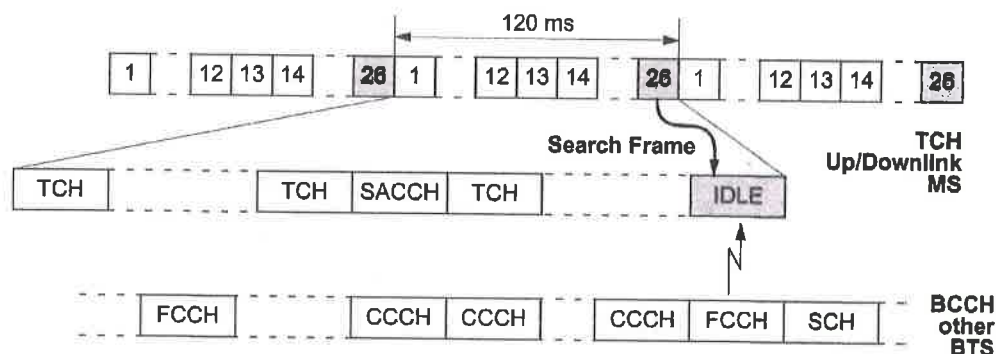


Figure 4.20 Synchronization with adjacent cells during a call.

can be judged. In addition to these measurements of the downlink by the MS, the base station also measures the RXLEV and RXQUAL values of the respective uplink.

In order to make a handover decision, information about possible handover targets must be available. For this purpose, the MS has to observe continuously the BCCH carriers of up to six neighboring base stations. The RXLEV measurements of the neighboring BCCH carriers are performed during the MSs unused time slots (see Figure 4.6). The BCCH measurement results of the six strongest signals are included in the measurement report transmitted to the BSS.

However, the received signal power level and the frequency of a BCCH carrier alone are not a sufficient criterion for a successful handover. Owing to the frequency reuse in cellular networks, and especially in the case of small clusters, it is possible that a cell can receive the same BCCH carrier from more than one neighboring cell, i.e. there exist several neighboring cells which use the same BCCH carrier. It is therefore necessary, to also know the identity (BSIC) of each neighboring cell. Simultaneously with the signal level measurement, the MS has to synchronize with each of the six neighboring BCCHs and read at least the SCH information.

For this purpose, one must first search for the FCCH burst of the BCCH carrier; then the SCH can be found in the next TDMA frame. Since the FCCH/SCH/BCCH is always transmitted in time slot 0 of the BCCH carrier, the search during a conversation for FCCHs can only be conducted in unused frames, i.e. in the case of a full-rate TCH in the IDLE frame of the multiframe (frame number 26 in Figures 4.16 and 4.20).

These free frames are therefore also known as search frames. Therefore, there are exactly four search frames within an SACCH block of 480 ms (four 26-frame multiframes of 120 ms). The MS has to examine the surrounding BCCH carriers for FCCH bursts, in order to synchronize with them and to decode the SCH. However, how can one search for synchronization points exactly within these frames during synchronized operation?

This is possible because the actual traffic channel and the respective BCCH carriers use different multiframe formats. Whereas the traffic channel uses the 26-frame multiframe format, time slot 0 of the BCCH carrier with the FCCH/SCH/BCCH is carried on a 51-frame multiframe format. This ratio of the different multiframe formats has the effect that the relative position of the search frames (frame 26 in a TCH multiframe) is shifting with regard to the BCCH multiframe by exactly one frame each 240 ms (Figure 4.21).

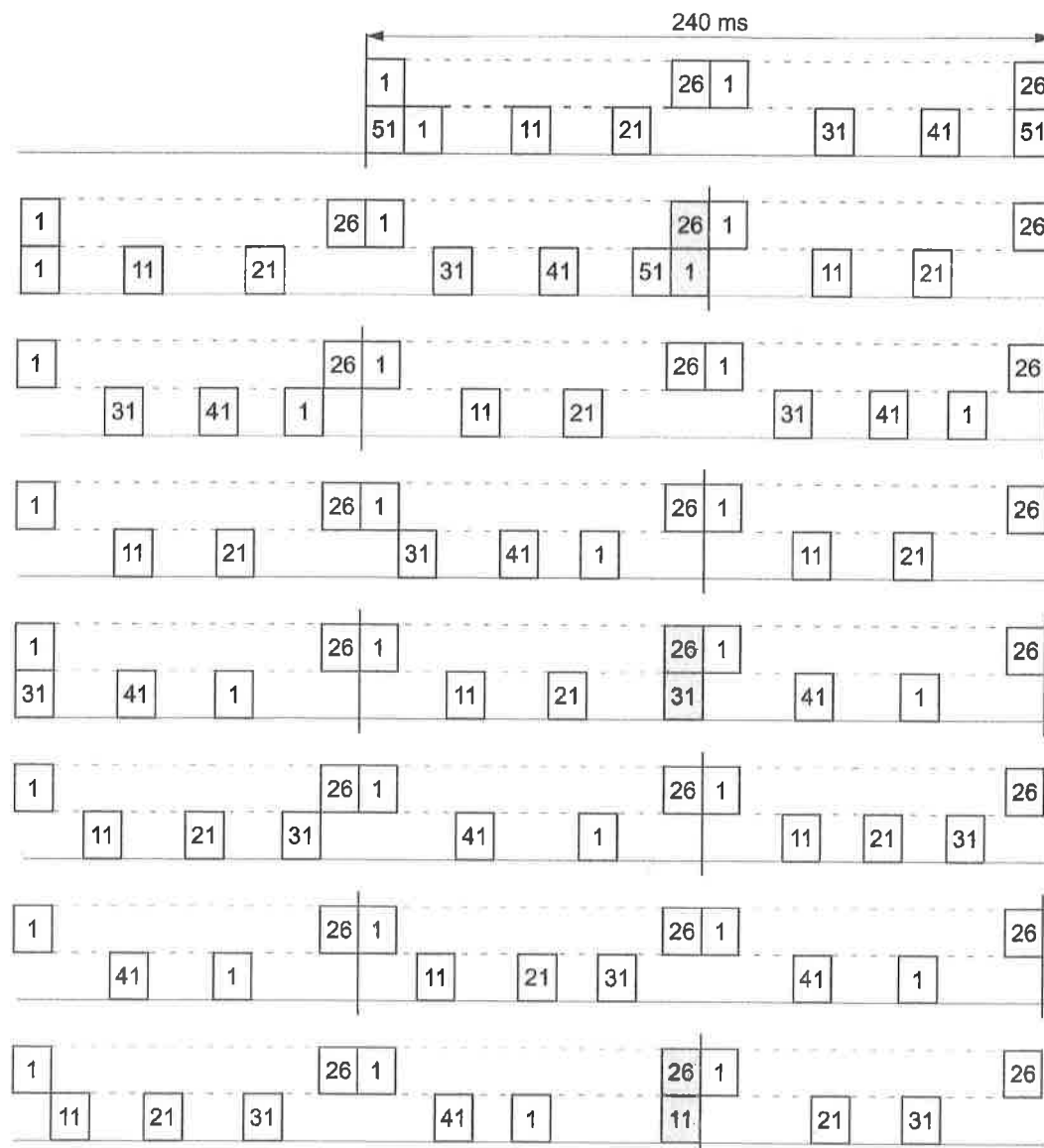


Figure 4.21 Principle of FCCH search during the search frame.

Figuratively speaking, the search frame is travelling along the BCCH multiframe in such a way that at most after 11 TCH multiframe (= 1320 ms) a frequency correction burst of a neighboring cell becomes visible in a search frame.

In this way, the MS is able to determine the BSIC for the respective RXLEV measurement value. Only BCCH carrier measurements whose identity can be established without doubt are included in the measurement report to the base station.

The base station can now make a handover decision based on these values, on the distance of the MS, and on the momentary interference of unused time slots.

The algorithm for handover decisions has not been included in the GSM standard. The network operators may use algorithms which are optimized for their network or the local

Table 4.8 GSM power classes.

Power class	Maximum peak transmission power (W)	
	Mobile station (dBm)	Base station
1	20 (43)	320
2	<8 (39)	160
3	<5 (37)	80
4	<2 (33)	40
5	<0.8 (29)	20
6	–	10
7	–	5
8	–	2.5

situation. GSM only gives a basic proposal which satisfies the minimum requirements for a handover decision algorithm. This algorithm defines threshold values, which must be violated in one or the other direction to arrive at a safe handover decision and to avoid so-called ping-pong handovers, which oscillate between two cells. Although the decision algorithm is part of radio subsystem link control, its discussion is postponed and it is treated together with handover signaling (see section 6.4.3).

4.5.2 Transmission power control

Power classes (Table 4.8) are used for classification of base stations and MSs. The transmission power can also be controlled adaptively. As part of the radio subsystem link control, the MS's transmitter power is controlled in steps of 2 dBm.

The GSM transmitter power control has the purpose of limiting the MS's transmitter power to the minimum necessary level, in such a way that the base station receives signals from different MSs at approximately the same power level. Sixteen power control steps are defined for this purpose: step 0 (43 dBm = 20 W) to step 15 (13 dBm). Starting with the lowest, step 15, the base station can increment the transmitter power of the MS in steps of 2 dBm up to the maximum power level of the respective power class of the MS. Similarly, the transmitter power of the base station can be controlled in steps of 2 dBm, with the exception of the BCCH carrier of the base station, which must remain constant to allow comparative measurements of neighboring BCCH carriers by the MSs.

Transmission power control is based on the measurement values RXLEV and RXQUAL, for which one has defined upper and lower thresholds for uplink and downlink (Table 4.9). Network management defines the adjustable parameters P and N . If the values of P for the last N calculated mean values of the respective criterion (RXLEV or RXQUAL) are above or below the respective threshold value, the BSS can adjust the transmitter power (Figure 4.22).

If the thresholds $U_{xx_UL_P}$ of the uplink are exceeded, the transmission power of the MS is reduced; in the other case, if the signal level is below the threshold $L_{xx_UL_P}$, the MS is ordered to increase its transmitter power. In an analogous way, the transmitter power of the base station can be adjusted, when the criteria for the downlink are exceeded in either direction.

Table 4.9 Thresholds for transmitter power control.

Threshold parameter	Typical value (dBm)	Meaning
L_RXLEV_UL_P	–103 to –73	Threshold for raising of transmission power in uplink or downlink
L_RXLEV_DL_P	–103 to –73	
L_RXQUAL_UL_P	–	
L_RXQUAL_DL_P	–	
U_RXLEV_UL_P	–	Threshold for reducing of transmission power in uplink or downlink
U_RXLEV_DL_P	–	
U_RXQUAL_UL_P	–	
U_RXQUAL_DL_P	–	

Even if the MS or base station signal levels stay within the thresholds, the current RXLEV/RXQUAL values can cause a change to another channel of the same or another cell based on the handover thresholds (Table 6.1). For this reason, checking for transmitter thresholds is immediately followed by a check of the handover thresholds as the second part of the radio subsystem link control (Figures 4.19 and 6.17). If one of the threshold values is exceeded in either direction and the transmitter power cannot be adjusted accordingly, i.e. the respective transmitter power has reached its maximum or minimum value, this is an overriding cause for handover (PWR_CTRL_FAIL; Table 6.2) which the BSS must communicate immediately to the MSC (section 6.4).

4.5.3 Disconnection due to radio channel failure

The quality of a radio channel can vary considerably during an existing connection, or it can even fail in the case of shadowing. This should not lead to immediate disconnection, since such failures are often of short duration. For this reason GSM has a special algorithm within the Radio Subsystem Link Control which continuously checks for connectivity. It consists of recognizing a radio link failure by the inability to decode signaling information on the SACCH. This connectivity check is done both in the MS as well as in the base station. The connection is not immediately terminated, but is delayed so that only repeated consecutive failures (erroneous messages) represent a valid disconnect criterion. On the downlink, the MS must check the frequency of erroneous, nondecodable messages on the SACCH. The error protection on the SACCH has very powerful error correction capabilities and thus guarantees a very low probability of 10^{-10} for nonrecognized, wrongly corrected bits in SACCH messages.

In this way, erroneous SACCH messages supply a measure for the quality of the downlink, which is already quite low when errors on the SACCH cannot be corrected any more. If a consecutive number of SACCH messages is erroneous, the link is considered bad and the connection is terminated. For this purpose, a counter S has been defined which is incremented by 2 with each arrival of an error-free message, and decremented by 1 for each erroneous SACCH message (Figure 4.23). When the counter reaches the value $S = 0$, the downlink is considered as failing, and the connection is terminated. This failure is signaled to the

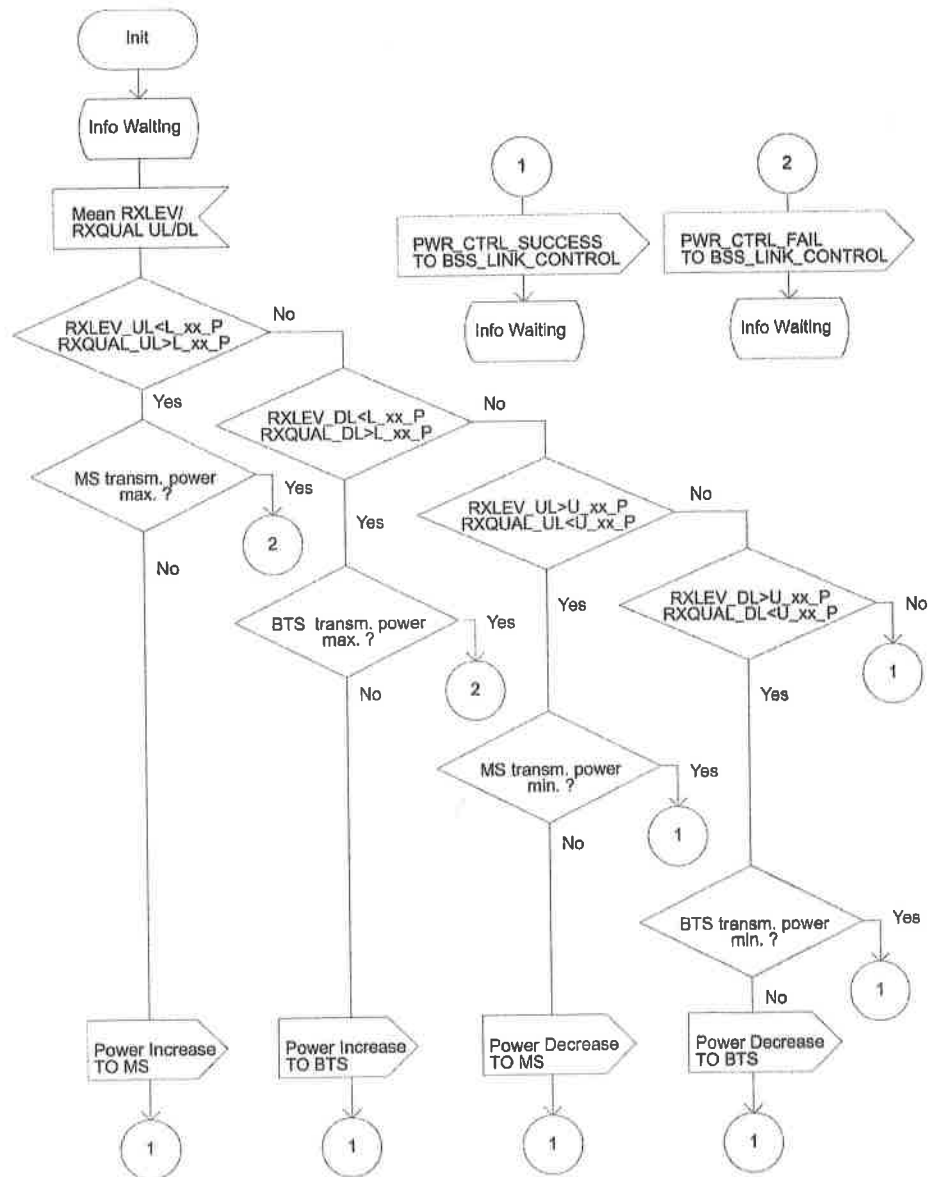


Figure 4.22 Schematic operation of transmitter power control.

upper layers, Mobility Management (MM), which can start a call reestablishment procedure. The maximum value `RADIO_LINK_TIMEOUT` for the counter *S* therefore determines the interval length during which a channel has to fail before a connection is terminated. After assignment of a dedicated channel (TCH or SDCCH), the MS starts the checking process by initializing the counter *S* with this value (Figure 4.23), which can be set individually per cell and is broadcast on the BCCH.

The corresponding checks are also conducted on the uplink. In both cases, however, this requires continuous transmission of data on the SACCH, i.e. when no signaling data have to be sent, filling data are transmitted. On the uplink, current measurement reports

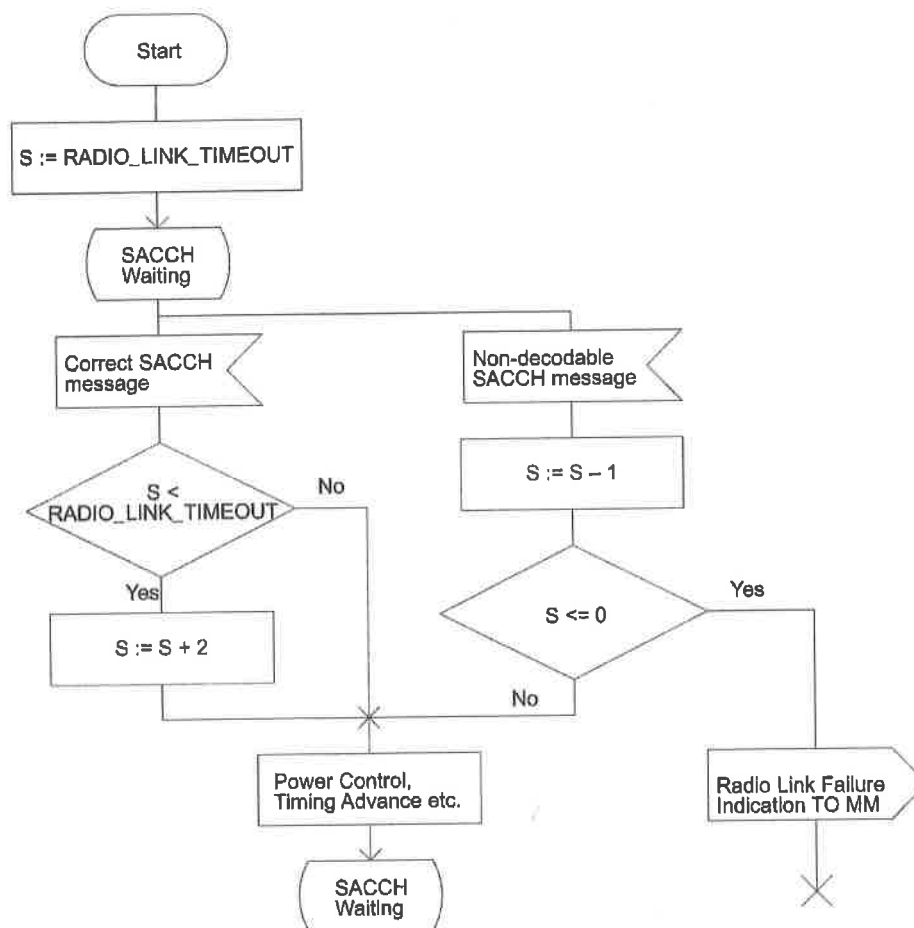


Figure 4.23 MS disconnect procedure.

are transmitted, whereas the downlink carries system information of Type 5 and Type 6 (section 5.4.3).

4.5.4 Cell selection and operation in power conservation mode

Cell selection and cell reselection

A MS in idle mode must periodically measure the receivable BCCH carriers of the base stations in the area and calculate mean values $RXLEV(n)$ from this data (section 4.5.1). Based on these measurements, the MS selects a cell, namely that with the best reception, i.e. the MS is committed to this cell. This is called 'camping' on this cell. In this state, accessing a service becomes possible, and the MS listens periodically to the PCH. Two criteria are defined for the automatic selection of cells: the path loss criterion C1 and the reselection criterion C2. The path loss criterion serves to identify cell candidates for camping. For such cells, C1 has to be greater than zero. At least every 5 s, a MS has to recalculate C1 and C2 for the current and neighboring cells. If the path loss criterion of the current cell falls below zero, the path loss to the current base station has become too large. A new cell has to be

selected, which requires use of the criterion C2. If one of the neighboring cells has a value of C2 greater than zero, it becomes the new current cell. The cell selection algorithm uses two further threshold values, which are broadcast on the BCCH:

- the minimum received power level $RXLEV_ACCESS_MIN$ (typically -98 to -106 dBm) required for registration into the network of the current cell;
- the maximum allowed transmitter power MS_TXPWR_CCH (typically 31 – 39 dBm) allowed for transmission on a control channel (RACH) before having received the first power control command.

In consideration of the maximal transmitter power P of a mobile station, the path loss criterion C1 is now defined using the minimal threshold $RXLEV_ACCESS_MIN$ for network access and the maximal allowed transmitter power $MS_TXPWR_MAX_CCH$:

$$C1(n) = (RXLEV(n) - RXLEV_ACCESS_MIN \\ - \text{maximum}(0, (MS_TXPWR_MAX_CCH - P))).$$

The values of the path loss criterion C1 are determined for each cell for which a value $RXLEV(n)$ of a BCCH carrier can be obtained. The cell with the lowest path loss can thus be determined using this criterion. It is the cell for which $C1 > 0$ has the largest value. During cell selection, the MS is not allowed to enter power conservation mode (Discontinuous Transmission (DTX), section 4.5.4).

A prerequisite for cell selection is that the cell considered belongs to the home Public Land Mobile Network (PLMN) of the MS or that access to the PLMN of this cell is allowed. Beyond that, a limited service mode has been defined with restricted service access, which still allows emergency calls if nothing else. In limited service mode, a MS can be camping on any cell but can only make emergency calls. Limited service mode exists when there is no SIM card in the MS, when the IMSI is unknown in the network or the IMEI is barred from service, but also if the cell with the best value of C1 does not belong to an allowed PLMN.

Once a MS is camping on a cell and is in idle mode, it should keep observing all of the BCCH carriers whose frequencies, the BA, are broadcast on the current BCCH. Having left idle mode, e.g. if a TCH has been assigned, the MS monitors only the six strongest neighboring BCCH carriers. A list of these six strongest neighboring BCCH carriers has already been prepared and kept up to date in idle mode. The BCCH of the camped-on cell must be decoded at least every 30 s. At least once every 5 min, the complete set of data from the six strongest neighboring BCCH carriers has to be decoded, and the BSIC of each of these carriers has to be checked every 30 s. This allows the MS to stay aware of changes in its environment and to react appropriately. In the worst case, conditions have changed so much that a new cell to camp on needs to be selected (cell reselection).

For this cell reselection, a further criterion C2, the reselection criterion, has been defined:

$$C2(n) = C1(n) + CELL_RESELECT_OFFSET \\ - (\text{TEMPORARY_OFFSET} \times H(\text{PENALTY_TIME} - T)),$$

$$\text{with } H(x) = \begin{cases} 0 & \text{for } x < 0, \\ 1 & \text{for } x \geq 0. \end{cases}$$

The interval T in this criterion is the time passed since the MS observed the cell n for the first time with a value of $C1 > 0$. It is set back to 0 when the path loss criterion $C1$ falls to $C1 < 0$. The parameters `CELL_RESELECT_OFFSET`, `TEMPORARY_OFFSET`, and `PENALTY_TIME` are announced on the BCCH. However, as a default, they are set to 0. Otherwise, the criterion $C2$ introduces a time hysteresis for cell reselection. It tries to ensure that the MS is camping on the cell with the highest probability of successful communication.

One exception for cell reselection is the case when a new cell belongs to another location area. In this case $C2$ must not only be larger than zero, but $C2 > \text{CELL_RESELECT_HYSTERESIS}$ to avoid too frequent location updates.

Discontinuous reception

To limit power consumption in idle mode and thus increase battery life in standby mode, the MS can activate the Discontinuous Reception (DRX) mode. In this mode, the receiver is turned on only for the phases of receiving paging messages and is otherwise in the power conservation mode which still maintains synchronization with BCCH signals through internal timers. In this DRX mode, measurement of BCCH carriers is performed only during unused time slots of the paging blocks.

4.6 Channel coding, source coding and speech processing

The previous sections explained the basic functions of the physical layer at the air interface, e.g. the definition of logical and physical channels, modulation, multiple access techniques, duplexing, and the definition of bursts. The following sections discuss several additional functions that are performed to transmit the data in an efficient, reliable way over the radio channel: source coding and speech processing (section 4.7), channel coding and burst mapping. Security related functions, such as encryption and authentication are discussed in Chapter 5.

Figure 4.24 gives a schematic overview of the basic elements of the GSM transmission chain. The stream of sampled speech data is fed into a source encoder, which compresses the data by removing unnecessary redundancy (section 4.7). The resulting information bit sequence is passed to the channel encoder (section 4.8). Its purpose is to add, in a controlled manner, some redundancy to the information sequence. This redundancy serves to protect the data against the negative effects of noise and interference encountered in the transmission through the radio channel. On the receiver side, the introduced redundancy allows the channel decoder to detect and correct transmission errors. GSM uses a combination of block and convolutional coding. Moreover, an interleaving scheme is used to deal with burst errors that occur over multipath and fading channels. Next, the encoded and interleaved data are encrypted to guarantee secure and confident data transmission. The encryption technique as well as the methods for subscriber authentication and secrecy of the subscriber identity are explained in section 5.6. The encrypted data are subsequently mapped to bursts (section 4.8.4), which are then multiplexed as explained in previous sections. Finally the stream of bits is differential coded and modulated.

After transmission, the demodulator processes the signal, which was corrupted by the noisy channel. It attempts to recover the actual signal from the received signal. The next steps are demultiplexing and decryption. The channel decoder attempts to reconstruct the

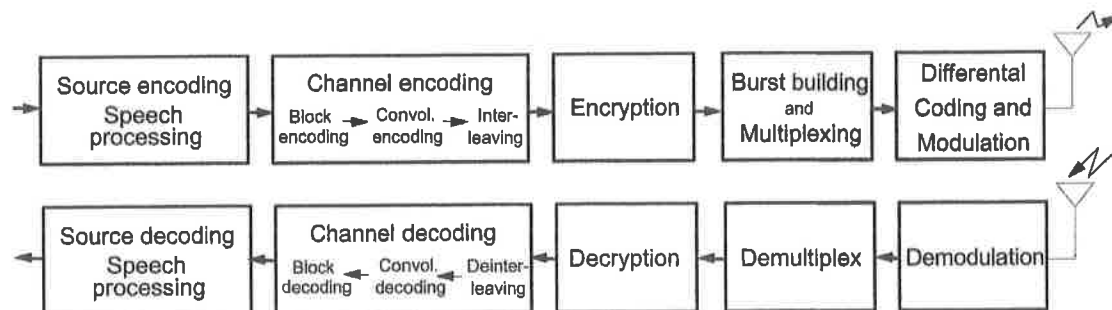


Figure 4.24 Basic elements of GSM transmission chain on the physical layer at the air interface.

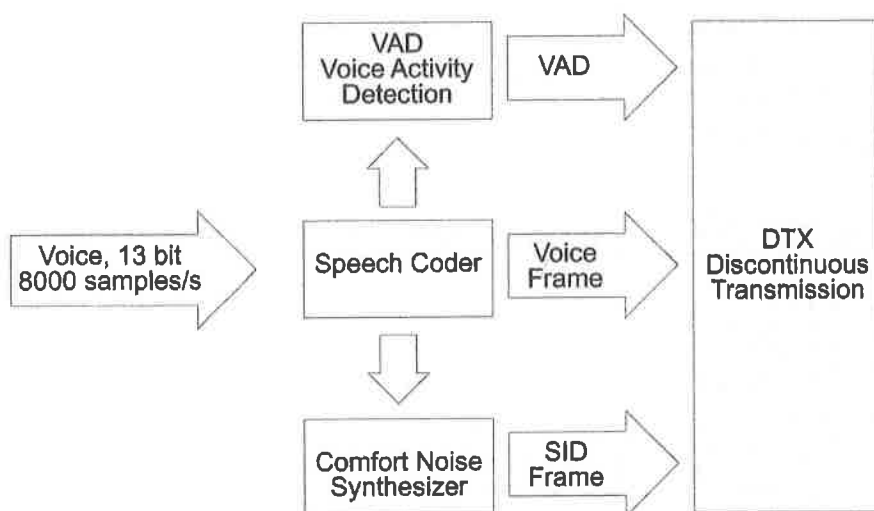


Figure 4.25 Schematic representation of speech functions at the transmitter.

original information sequence and, as a final step, the source decoder tries to reconstruct the original source signal.

4.7 Source coding and speech processing

Source coding reduces redundancy in the speech signal and thus results in signal compression, which means that a significantly lower bit rate is achieved than needed by the original speech signal. The speech coder/decoder is the central part of the GSM speech processing function, both at the transmitter (Figure 4.25) as well as at the receiver (Figure 4.26). The functions of the GSM speech coder and decoder are usually combined in one building block called the codec (COder/DECoder).

The analog speech signal at the transmitter is sampled at a rate of 8000 samples per second and the samples are quantized with a resolution of 13 bits. This corresponds to a bit rate of 104 kbit/s for the speech signal. At the input to the speech codec, a speech frame containing

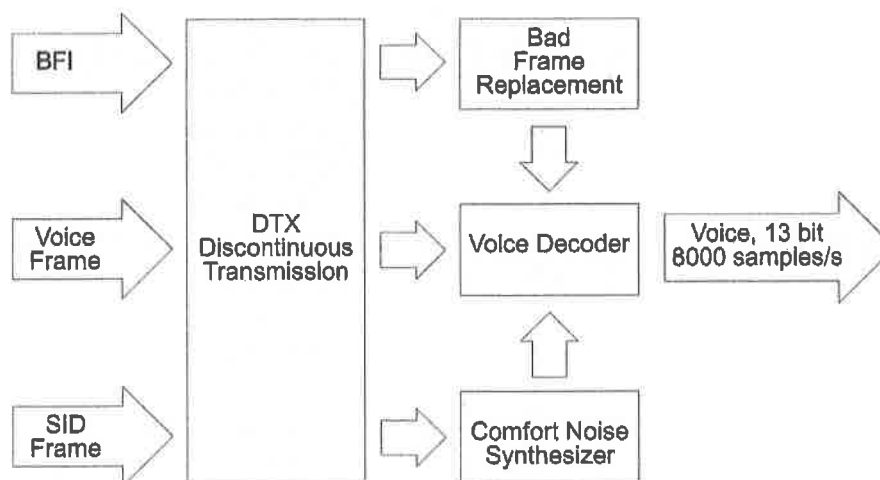


Figure 4.26 Schematic representation of speech functions at the receiver.

160 samples of 13 bits arrives every 20 ms. The speech codec compresses this speech signal into a source-coded speech signal of 260-bit blocks at a bit rate of 13 kbit/s. Thus, the GSM speech coder achieves a compression ratio of 1 to 8. The source coding procedure is briefly explained in the following; detailed discussions of speech coding procedures are given in Steele (1992).

A further ingredient of speech processing at the transmitter is the recognition of speech pauses, called Voice Activity Detection (VAD). The voice activity detector decides, based on a set of parameters delivered by the speech coder, whether the current speech frame (20 ms) contains speech or a speech pause. This decision is used to turn off the transmitter amplifier during speech pauses, under control of the Discontinuous Transmission (DTX) block.

The discontinuous transmission mode takes advantage of the fact, that during a normal telephone conversation, both parties rarely speak at the same time, and thus each directional transmission path has to transport speech data only half of the time. In DTX mode, the transmitter is only activated when the current frame indeed carries speech information. This decision is based on the VAD signal of speech pause recognition. The DTX mode can reduce the power consumption and, hence, prolong the battery life. In addition, the reduction of transmitted energy also reduces the level of interference and thus improves the spectral efficiency of the GSM system. The missing speech frames are replaced at the receiver by a synthetic background noise signal called comfort noise (Figure 4.26). The parameters for the comfort noise synthesizer are transmitted in a special silence descriptor (SID) frame.

This silence descriptor is generated at the transmitter from continuous measurements of the (acoustic) background noise level. It represents a speech frame which is transmitted at the end of a speech burst, i.e. at the beginning of a speech pause. In this way, the receiver recognizes the end of a speech burst and can activate the comfort noise synthesizer with the parameters received in the SID frame. The generation of this artificial background noise in DTX mode prevents the audible background noise transmitted with normal speech bursts from suddenly dropping to a minimal level at a speech pause. This modulation of the background noise would have a very disturbing effect on the human listener and would

significantly deteriorate the subjective speech quality. Insertion of comfort noise is a very effective countermeasure to compensate for this so-called noise-contrast effect.

Another loss of speech frames can occur, when bit errors caused by a noisy transmission channel cannot be corrected by the channel coding protection mechanism, and the block is received at the codec as a speech frame in error, which must be discarded. Such bad speech frames are flagged by the channel decoder with the Bad Frame Indication (BFI). In this case, the respective speech frame is discarded and the lost frame is replaced by a speech frame which is predictively calculated from the preceding frame. This technique is called error concealment. Simple insertion of comfort noise is not allowed. If 16 consecutive speech frames are lost, the receiver is muted to acoustically signal the temporary failure of the channel.

The speech compression takes place in the speech coder. The GSM speech coder uses a procedure known as Regular Pulse Excitation Long-Term Prediction (RPE-LTP). This procedure belongs to the family of hybrid speech coders. This hybrid procedure transmits part of the speech signal as the amplitude of a signal envelope, a pure wave form encoding, whereas the remaining part is encoded into a set of parameters. The receiver reconstructs these signal parts through speech synthesis (vocoder technique). Examples of envelope encoding are Pulse Code Modulation (PCM) or Adaptive Delta Pulse Code Modulation (ADPCM). A pure vocoder procedure is Linear Predictive Coding (LPC). The GSM procedure RPE-LTP as well as Code Excited Linear Predictive Coding (CELP) represent mixed (hybrid) approaches (David and Benkner, 1996; Natwig, 1998; Steele, 1992).

A simplified block diagram of the RPE-LTP coder is shown in Figure 4.27. Speech data generated with a sampling rate of 8000 samples per second and 13-bit resolution arrive in blocks of 160 samples at the input of the coder. The speech signal is then decomposed into three components: a set of parameters for the adjustment of the short-term analysis filter (LPC) also called reflection coefficients; an excitation signal for the RPE part with irrelevant portions removed and highly compressed; and finally a set of parameters for the control of the LTP long-term analysis filter. The LPC and LTP analyses supply 36 filter parameters for each sample block, and the RPE coding compresses the sample block to 188 bits of RPE parameters. This results in the generation of a frame of 260 bits every 20 ms, equivalent to a 13 kbit/s GSM speech signal rate.

The speech data preprocessing of the coder (Figure 4.27) removes the DC portion of the signal if present and uses a preemphasis filter to emphasize the higher frequencies of the speech spectrum. The preprocessed speech data are run through a nonrecursive lattice filter (LPC filter; Figure 4.27) to reduce the dynamic range of the signal. Since this filter has a 'memory' of about 1 ms, it is also called short-term prediction filter. The coefficients of this filter, called reflection coefficients, are calculated during LPC analysis and transmitted in a logarithmic representation as part of the speech frame, Log Area Ratios (LARs).

Further processing of the speech data is preceded by a recalculation of the coefficients of the long-term prediction filter (LTP analysis in Figure 4.27). The new prediction is based on the previous and current blocks of speech data. The resulting estimated block is finally subtracted from the block to be processed, and the resulting difference signal is passed on to the RPE coder.

After LPC and LTP filtering, the speech signal has been redundancy reduced, i.e. it already needs a lower bit rate than the sampled signal; however, the original signal can still be reconstructed from the calculated parameters. The irrelevance contained in the speech

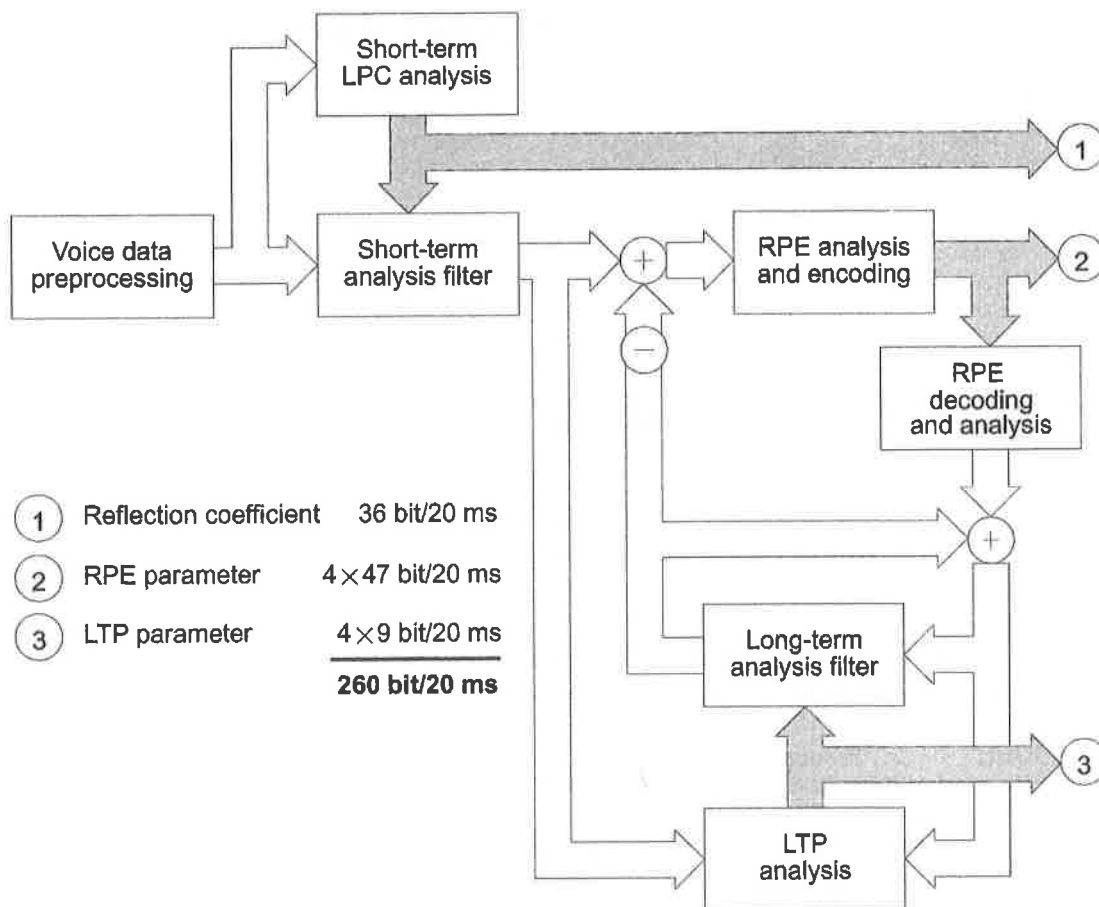


Figure 4.27 Simplified block diagram of the GSM speech coder.

signal is reduced by the RPE coder. This irrelevance represents speech information that is not needed for the understandability of the speech signal, since it is hardly noticeable to human hearing and thus can be removed without loss of quality. On the one hand, this results in a significant compression (factor $160 \times 13/188 \approx 11$); on the other hand, it has the effect that the original signal cannot be reconstructed uniquely. Figure 4.28 summarizes the reconstruction of the speech signal from RPE data, as well as the long-term and short-term synthesis from LTP and LPC filter parameters. In principle, at the receiver site, the functions performed are the inverse of the functions of the encoding process.

The irrelevance reduction only minimally affects the subjectively perceived speech quality, since the main objective of the GSM codec is not just the highest possible compression but also good subjective speech quality. To measure the speech quality in an objective manner, a series of tests were performed on a large number of candidate systems and competing codecs.

The base for comparison used is the Mean Opinion Score (MOS), ranging from MOS = 1, meaning quality is very bad or unacceptable, to MOS = 5, meaning quality is very good or fully acceptable. A series of coding procedures were discussed for the GSM system; they were examined in extensive hearing tests for their respective subjective speech quality (Natwig, 1998). Table 4.10 gives an overview of these test results; it includes as reference

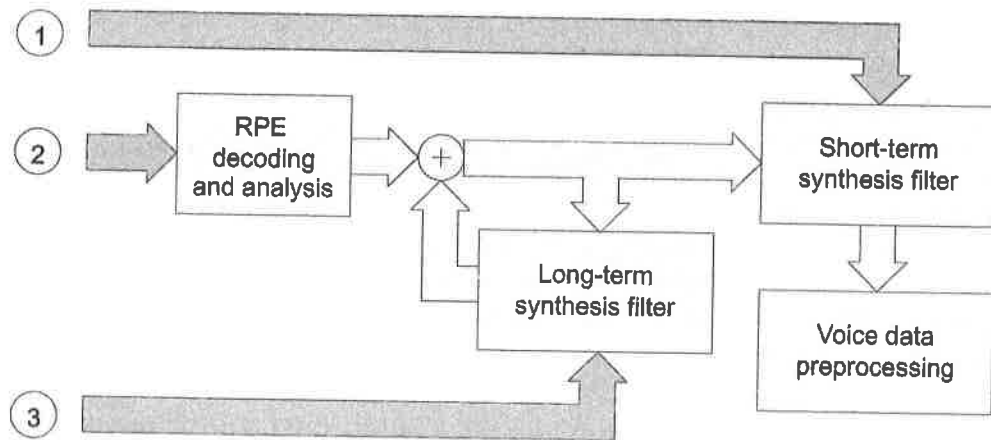


Figure 4.28 Simplified block diagram of the GSM speech decoder.

Table 4.10 MOS results of codec hearing tests (Natwig, 1998).

CODEC	Process	Bit rate (in kbit/s)	MOS
FM	Frequency Modulation	—	1.95
SBC-ADPCM	Subband-CODEC – Adaptive Delta-PCM	15	2.92
SBC-APCM	Subband-CODEC – Adaptive PCM	16	3.14
MPE-LTP	Multi-Pulse Excited LPC-CODEC – Long-Term Prediction	16	3.27
RPE-LPC	Regular-Pulse Excited LPC-CODEC	13	3.54
RPE-LTP	Regular Pulse Excited LPC-CODEC – Long-Term Prediction	13	≈4
ADPCM	Adaptive Delta Modulation	32	≥4

also ADPCM and frequency-modulated analog transmission. The GSM codec with the RPE-LTP procedure generates a speech quality with an MOS value of about 4 for a wide range of different inputs.

4.8 Channel coding

The heavily varying properties of the mobile radio channel (see Chapter 2) often result in a very high bit error ratio, of the order of 10^{-3} to 10^{-1} . The highly compressed, redundancy-reduced source coding makes speech communication with acceptable quality almost impossible; moreover, it makes reasonable data communication impossible. Suitable error correction procedures are therefore necessary to reduce the bit error probability into an acceptable range of about 10^{-5} to 10^{-6} . Channel coding, in contrast to source coding, adds redundancy to the data stream to enable detection and correction of transmission errors.

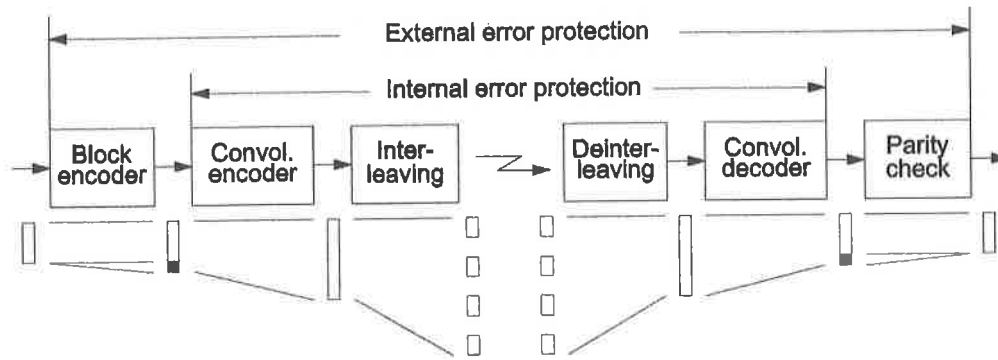


Figure 4.29 Stages of channel coding.

It is the modern high-performance coding and error correction techniques which essentially enable the implementation of a digital mobile communication system.

The GSM system uses a combination of several procedures: in addition to a block code, which generates parity bits for error detection, a convolutional code generates the redundancy needed for error correction. Furthermore, sophisticated interleaving of data over several blocks reduces the damage done by burst errors. The individual steps of channel coding are shown in Figure 4.29:

- calculation of parity bits (block code) and addition of fill bits;
- error protection coding through convolutional coding;
- Interleaving.

Finally, the coded and interleaved blocks are enciphered, distributed across bursts, modulated and transmitted on the respective carrier frequencies.

The sequence of data blocks that arrives at the input of the channel encoder is combined into blocks, partially supplemented by parity bits (depending on the logical channel) and then complemented to a block size suitable for the convolutional encoder. This involves appending zero bits at the end of each data block, which allow a defined resetting procedure of the convolutional encoder (zero termination) and thus a correct decoding decision. Finally, these blocks are run through the convolutional encoder. The ratio of uncoded to coded block length is called the *rate* of the convolutional code. Some of the redundancy bits generated by the convolutional encoder are deleted again for some of the logical channels. This procedure is known as *puncturing*, and the resulting code is a punctured convolutional code (Begin and Haccoun, 1994; Hagenauer *et al.*, 1990; Kallel, 1995). Puncturing increases the rate of the convolutional code, so it reduces the redundancy per block to be transmitted, and lowers the bandwidth requirements, such that the convolution-encoded signal fits into the available channel bit rate. The convolution-encoded bits are passed to the interleaver, which shuffles various bit streams. At the receiving site, the respective inverse functions are performed: deinterleaving, convolutional decoding and parity checking. Depending on the position within the transmission chain (Figure 4.29), one distinguishes between external error protection (block code) and internal protection (convolutional code).

In the following, the GSM channel coding is presented according to these stages. section 4.8.1 explains the block coding, section 4.8.2 deals with convolutional coding and,

finally, section 4.8.3 presents the interleaving procedures used in GSM. The error protection measures have different parameters depending on channel and type of transported data. Table 4.11 gives an overview. (Note that the tail bits indicated in the second column are the fill bits needed by the decoding process; they should not be confused with the tail bits of the bursts (section 4.2).)

The basic unit for all coding procedures is the data block. For example, the speech coder delivers a sequence of data blocks to the channel encoder. Depending on the logical channel, the length of the data block is different; after convolutional coding at the latest, data from all channels are transformed into units of 456 bits. Such a block of 456 bits transports a complete speech frame or a protocol message in most of the signaling channels, except for the RACH and SCH channels. The starting points are the blocks delivered to the input of the channel encoder from the protocol processing in higher layers (Figure 4.30).

Speech traffic channels

One block of the full-rate speech codec consists of 260 bits of speech data, i.e. each block contains 260 information bits, which must be encoded. They are graded into two classes (class I, 182 bits; class II, 78 bits) which have different sensitivity against bit errors. Class I includes speech bits that have a greater impact on speech quality and hence must be better protected. Speech bits of class II, however, are less important. They are therefore transmitted without convolutional coding, but are included in the interleaving process. The individual sections of a speech frame are therefore protected to differing degrees against transmission errors (Unequal Error Protection (UEP)). In the case of a half-rate speech codec, data blocks of 112 information bits are input into the channel encoder. Of these, 95 bits belong to class I and 17 bits belong to class II. Again, one data block corresponds to one speech frame.

Data traffic channels

Blocks of traffic channels for data services have a length of N_0 bits, the value of N_0 being a function of the data service bit rate. We take for example the 9.6 kbit/s data service on a full-rate traffic channel (TCH/F9.6). Here, a bit stream organized into blocks of 60 information bits arrives every 5 ms at the input of the encoder. Four subsequent blocks are combined for the encoding process.

Signaling channels

The data streams of most of the signaling channels are constructed of blocks of 184 bits each; with the exception of the RACH and SCH which supply blocks of length P_0 to the channel coder. The block length of 184 bits results from the fixed length of the protocol message frames of 23 octets on the signaling channels. The channel coding process maps pairs of subblocks of 57 bits onto the bursts such that it can fill a normal data burst N_B (Figure 4.7).

4.8.1 External error protection: block coding

The block coding stage in GSM has the purpose of generating parity bits for a block of data, which allow the detection of errors in this block. In addition, these blocks are supplemented by fill bits (tail bits) to a block length suitable for further processing. Since block coding is the

Table 4.11 Error protection coding and interleaving of logical channels.

Channel type	Abbreviation	Block distance (ms)	Bits per block			Convolution code rate	Encoded bits per block	Interleaver depth
			Data	Parity	Tail			
TCH, full rate, speech	TCH/FS	20	260				456	8
Class I			182	3	4	1/2	378	
Class II			78	0	0	–	78	
TCH, half rate, speech	TCH/HS	20	112				228	4
Class I			95	3	6	104/211	211	
Class II			17	0	0	–	17	
TCH, full rate, 14.4 kbit/s	TCH/F14.4	20	290	0	4	294/456	456	19
TCH, full rate, 9.6 kbit/s	TCH/F9.6	5	4 × 60	0	4	244/456	456	19
TCH, full rate, 4.8 kbit/s	TCH/F4.8	10	60	0	16	1/3	228	19
TCH, half rate, 4.8 kbit/s	TCH/H4.8	10	4 × 60	0	4	244/456	456	19
TCH, full rate, 2.4 kbit/s	TCH/F2.4	10	2 × 36	0	4	1/6	456	8
TCH, half rate, 2.4 kbit/s	TCH/H2.4	10	2 × 36	0	4	1/3	228	19
FACCH, full rate	FACCH/F	20	184	40	4	1/2	456	8
FACCH, half rate	FACCH/H	40	184	40	4	1/2	456	6
SDCCH, SACCH			184	40	4	1/2	456	4
BCCH, NCH, AGCH, PCH		235	184	40	4	1/2	456	4
RACH		235	8	6	4	1/2	36	1
SCH			25	10	4	1/2	78	1
CBCH		235	184	40	4	1/2	456	4

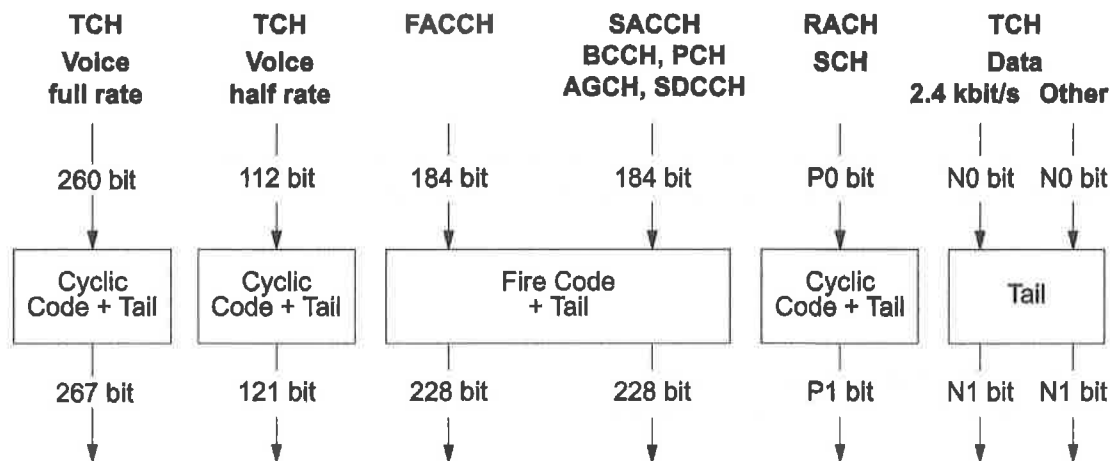


Figure 4.30 Overview of block coding for logical channels (also see Table 4.11).

first or external stage of channel coding, the block code is also known as external protection. Figure 4.30 gives a brief overview showing which codes are used for which channels. In principle, only two kinds of codes are used: a Cyclic Redundancy Check (CRC) and a fire code.

Block coding for speech traffic channel

As mentioned above, speech data occurs on the TCH in speech frames (blocks) of 260 bits for TCH/F and 112 bits for TCH/H, respectively. The bits belonging to class I are error-protected, whereas the bits of class II are not protected. A 3-bit CRC code is calculated for the first 50 bits of class I (in the case of TCH/F). The generator polynomial for this CRC is

$$G_{\text{CRC}}(x) = x^3 + x + 1.$$

In the case of a TCH/H speech channel, the most significant 22 bits of Class I are protected by 3 parity bits, using the same generator polynomial.

We now explain the block coding process in more detail with focus on the TCH/F speech codec. Since cyclic codes are easily generated with a feedback shift register, they are often defined directly with this register representation. Figure 4.31 shows such a shift register with storage locations (delay elements) and modulo-2 adders. For initialization, the register is primed with the first three bits of the data block. The other data are shifted bitwise into the feedback shift register; after the last data bit has been shifted out of the register, the register contains the check sum bits, which are then appended to the block.

The operation of this shift register can be easily explained if the bit sequences are also represented as polynomials as with the generating function. The first 50 bits of a speech frame $D_0, D_1, \dots, D_{49}$ are denoted as

$$D(x) = D_{49}x^{49} + D_{48}x^{48} + \dots + D_1x + D_0.$$

If this data sequence is shifted through the register of Figure 4.31, after the register was primed with D_{47}, D_{48}, D_{49} followed by 50 shift operations, then the check sum bits

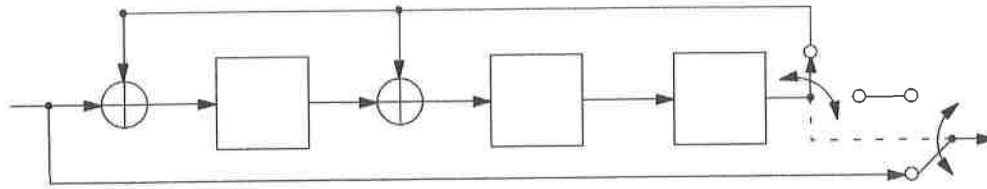


Figure 4.31 Feedback shift register for CRC.

$R(x)$ correspond to the remainder, which is left by dividing the data sequence $x^3 D(x)$ (supplemented by three zero bits) by the generator polynomial:

$$R(x) = \text{Remainder} \left[\frac{x^3 D(x)}{G_{\text{CRC}}(x)} \right].$$

In the case of error-free transmission, the codeword $C'(x) = x^3 D(x) + R(x)$ is therefore divisible by $G_{\text{CRC}}(x)$ without remainder. However, since the check sum bits $R(x)$ are transmitted in inverted form, the division yields a remainder:

$$S(x) = \text{Remainder} \left[\frac{C(x)}{G_{\text{CRC}}(x)} \right] = \text{Remainder} \left[\frac{x^3 D(x) + \bar{R}(x)}{G_{\text{CRC}}(x)} \right] = x^2 + x + 1.$$

This is equivalent to shifting the whole codeword $C(x)$ through an identical shift register on the decoder side, after priming it with C_{50} , C_{51} , C_{52} . After shifting in the last check sum bit (50 shift operations), this register should contain a 1. If this is not the case, the block contains erroneous bits. Inversion of the parity bits avoids the generation of null code-words, i.e. bursts which contain only zeros cannot occur on the traffic channel.

The speech data $d(k)$ ($k = 1, \dots, 182$) of class I of a block are combined with the parity bits $p(k)$ ($k = 1, 2, 3$) and fill bits to form a new block $u(k)$ ($k = 1, \dots, 189$)

$$u(k) = \begin{cases} d(2k) & k = 1, \dots, 90, \\ d(2 \times (184 - k) + 1) & k = 94, \dots, 184, \\ p((k - 91) + 1) & k = 91, 92, 93, \\ 0 & k = 185, \dots, 189. \end{cases}$$

The bits in even or odd positions are shifted to the upper or lower half of the block, respectively, and separated by the three check sum bits; in addition, the order of the odd bits is reversed. Finally, the block is filled to 189 bits. Combination with the speech bits of class II yields a block of 267 bits, which serves as input to the convolutional coder.

This enormous effort is taken because of the high compression rate and sensitivity against bit errors of the speech data. A speech frame in which the bits of class I have been recognized as erroneous can therefore be reported as erroneous to the speech codec using the BFI; see section 4.7. In order to maintain a constantly good speech quality, speech frames recognized as faulty are discarded, and the last correctly received frame is repeated, or an extrapolation of received speech data is performed.

Table 4.12 Block formation for data traffic channels.

Data channel	N0	Tail bits		N1
TCH/F14.4	290	+	4	= 294
TCH/F9.6	4 × 60	+	4	= 244
TCH/F4.8	(2 ×)60	+	(2 ×)16	= (2 ×)76
TCH/H4.8	4 × 60	+	4	= 244
TCH/F2.4	2 × 36	+	4	= 76
TCH/H2.4	(2 ×)2 × 36	+	(2 ×)4	= (2 ×)76

Block coding for data traffic channels

Block coding of traffic channels is somewhat simpler for data services. In this case, no parity bits are determined. Blocks of length N0 arriving at the input of the encoder are supplemented by fill bits to a size of N1 suitable for further coding. Table 4.12 gives an overview of the different block lengths, which depend on the data rate and channel type, i.e. whether the channel is a full-rate (TCH/Fxx) or half-rate (TCH/Hxx) channel.

The 9.6 kbit/s data service is only offered on a full-rate traffic channel. The data comes in blocks of 60 bits to the channel encoder (every 5 ms). Four blocks each are combined and supplemented by four appended tail bits (zero bits). In the case of a nontransparent data service, these four blocks make up exactly one protocol frame of the Radio Link Protocol (RLP; 240 bits). The procedures for other data services are similar. As shown in Table 4.11, for the 4.8 and 2.4 kbit/s services, blocks of 60 or 36 bit length arrive every 10 ms. Subsequent blocks are combined and are then supplemented with tail bits (zero bits) to form blocks of 76 or 244 bits, respectively. The bit stream for the 14.4 kbit/s data service (TCH/F14.4) is offered to the encoder in blocks of 290 information bits every 20 ms. Here, four tail bits are added, resulting in 294 bits (see Table 4.12).

Block coding for signaling channels

The majority of the signaling channels (SACCH, FACCH, SDCCH, BCCH, PCH and AGCH) use an extremely powerful block code for error detection. This is a so-called fire code, i.e. a shortened binary cyclic code which appends 40 redundancy bits to the 184-bit data block. Its pure error-detection capability is sufficient to let undetected errors go through only with a probability of 2^{-40} . (A fire code can also be used for error correction, but here it is used only for error detection.) Error detection with the fire code in the SACCH channel is used to verify connectivity (Figure 4.23), and is used, if indicated, for decisions regarding breaking a connection. The fire code can be defined in the same way as the CRC by way of a generator polynomial:

$$TG_F(x) = (x^{23} + 1)(x^{17} + x^3 + 1).$$

The check sum bits $R_F(s)$ of this code are calculated in such a way that a 40-bit remainder $S_F(x)$ is left after dividing the codeword $C_F(x)$ by the generator polynomial $G_F(x)$. In the

Table 4.13 Block lengths for the RACH and SCH channels.

Data channel	P0	Parity bits	Tail bits	P1
RACH	8	+6	+4	= 18
SCH	25	+10	+4	= 39

case of no errors, the remainder contains only '1' bits:

$$S_F(x) = \text{Remainder} \left[\frac{C_F(x)}{G_F(x)} \right] = \text{Remainder} \left[\frac{x^{40} D_F(x) + R_F(x)}{G_F(x)} \right]$$

$$= x^{39} + x^{38} + \dots + x^2 + x + 1.$$

The codeword generated with the redundancy bits of the fire code is supplemented with '0' bits to a total length of 228 bits, which are then delivered to the convolutional coder.

Another approach has been used for error detection in the RACH channel. The very short random access burst in the RACH allows only a data block length of $P0 = 8$ bits, which is supplemented in a cyclic code by six redundancy bits. The corresponding generator polynomial is

$$G_{\text{RACH}}(x) = x^6 + x^5 + x^3 + x^2 + x + 1.$$

In the AB, the MS also has to indicate a target base station. The BSIC of the respective base station is used for this purpose. The six bits of the BSIC are added to the six redundancy bits modulo-2, and the resulting sequence is inserted as the redundancy of the data block. The total codeword to be convolution-coded for the RACH thus has a length of 18 bits; i.e. four fill bits ('0') are also added in the RACH to this block. In exactly the same way, block coding is performed for the handover access burst, which is in principle also a random access burst.

The SCH channel, as an important synchronization channel, uses somewhat more elaborate error protection than the RACH channel. The SCH data blocks have a length of 25 bits and receive, in addition to the fill bits, another 10 bits of redundancy for error detection through a cyclic code with somewhat better error detection capability than on the RACH:

$$G_{\text{SCH}}(x) = x^{10} + x^8 + x^6 + x^5 + x^4 + x^2 + 1.$$

Thus the length of the codewords delivered to the channel coder in the SCH channel is 39 bits. Table 4.13 summarizes the block parameters of the RACH and SCH channels. Table 4.14 presents an overview of the cyclic codes used in GSM.

4.8.2 Internal error protection: convolutional coding

After block coding has supplemented the data with redundancy bits for error detection (parity bits), added fill bits and thus generated sorted blocks, the next stage is the calculation of additional redundancy for error correction to correct the transmission errors caused by the radio channel. The internal error correction of GSM is based exclusively on convolutional codes.

Convolutional codes (Johannesson and Zigangirov, 1999) can also be defined using shift registers and generator polynomials. Figure 4.32 illustrates a possible convolutional encoder

Table 4.14 Cyclic codes used for block coding in GSM.

Channel	Polynomial
TCH/FS	$x^3 + x + 1$
DCCH and CCCH (part)	$(x^{23} + 1)(x^{17} + x^3 + 1)$
RACH	$x^6 + x^5 + x^3 + x^2 + x + 1$
SCH	$x^{10} + x^8 + x^6 + x^5 + x^4 + x^2 + x + 1$

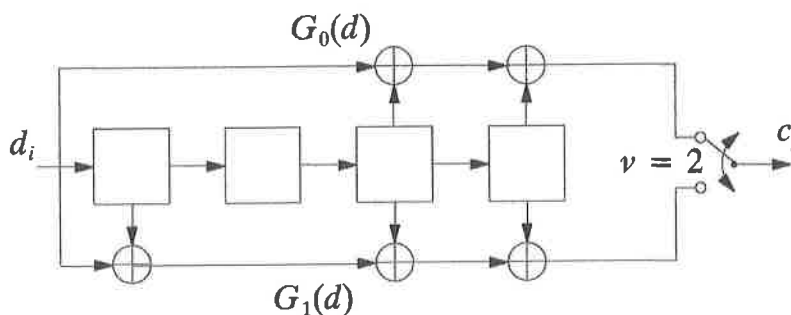


Figure 4.32 Principle of a convolutional encoder.

realization. It basically consists of a shift register with modulo-2 adders and K storage locations (here $K = 4$). One data/information symbol d_i is read into the shift register per tact interval. A symbol consists of k (here $k = 1$) data/information bits, each of which is moved into the shift register. A data symbol could also consist of more than one bit ($k > 1$), but this is not implemented in GSM. The symbol read is combined with up to K of its predecessor symbols $d_{i-1}, \dots, d_{i-K}$ in several modulo-2 additions. The results of these operations are given to the interleaver as coded user payload symbols c_j . The value K determines the number of predecessor symbols to be combined with a data symbol and is therefore also called the memory of the convolutional encoder. The number v of combinatorial rules (here $v = 2$) determines the number of coded bits in a code symbol c_j generated for each input symbol d_i . In Figure 4.32, the combinatorial results are scanned from top to bottom to generate the code symbol c_j . The combinatorial rules are defined by the generator polynomial $G_i(d)$. It is important to note that a specific convolutional code can be generated by various encoders. Thus, it must be carefully distinguished between code properties and encoder properties.

As mentioned in section 4.8.1, block coding appends at least four zero bits to each block. These bits not only serve as fill bits at the end of a block, but they are also important for the channel coding procedure. Shifted at the end of each block into the encoder, these bits serve to reset the encoder into the defined starting position (zero termination of the encoder), such that in principle adjacent data blocks can be coded independently of each other.

The rate r of a convolutional code indicates how many data (information) bits are processed for each coded bit. Consequently, $1/r$ is the number of coded bits per information bit. This rate is the essential measure of the redundancy produced by the code and, hence, its

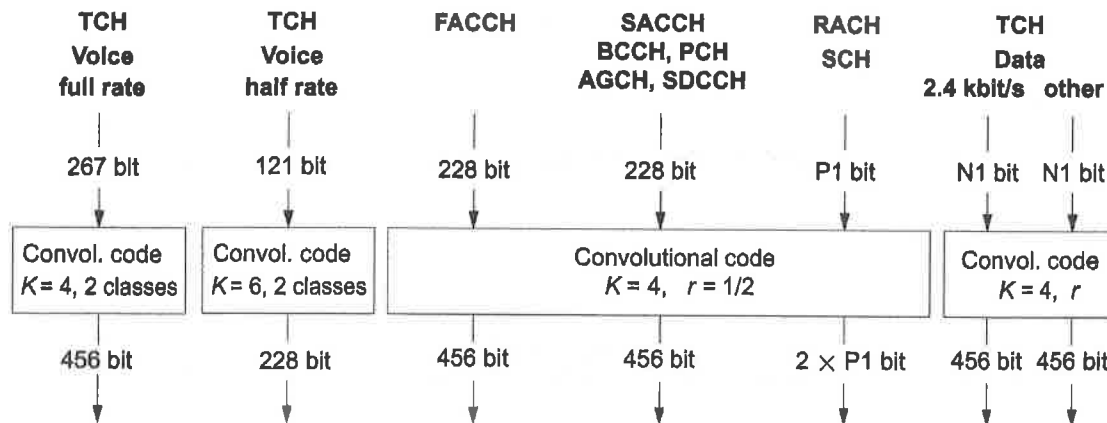


Figure 4.33 Overview of convolutional coding of logical channels (also see Table 4.11).

error correction capability:

$$r = \frac{k}{v}, \quad \text{here: } r = \frac{1}{v} = \frac{1}{2}.$$

The code rate is therefore determined by the number of bits k per input data symbol and the number of combinatorial rules v which are used for the calculation of a code symbol. In combination with the memory K , the code rate r determines the error-correction capability of the code. In a simplified way: with decreasing r and increasing K , the number of corrigible errors per codeword increase, and, thus, the error-correction capabilities of the code are improved. The encoding procedure is expressed in the combinatorial operations (modulo-2 additions). These coding rules can be described with polynomials. In the case of the convolutional encoder of Figure 4.32, the two generator polynomials are

$$G_0(d) = d^4 + d^3 + 1,$$

$$G_1(d) = d^4 + d^3 + d + 1.$$

These give a compact representation of the encoding procedure. The maximal exponent of a generator polynomial is known as its constraint length. The maximal of all constraint lengths (i.e. the maximal exponent of all polynomials) defines the memory K of the convolutional encoder. The number of polynomials determines the rate r . The exponents represent how an input symbol d_i is processed in the encoder. For example, in the upper path of the encoder (represented by $G_0(d)$), an input symbol d_i is immediately forwarded to the output (exponent '0'), and it is processed again in the third and fourth tact interval.

GSM defines different convolutional codes for the different logical channels (see Figure 4.33). Table 4.15 lists the seven generator polynomials ($G_0, \dots, G_6$) used in different combinations. The convolutional encoder used for half-rate speech channels (TCH/HS) has memory 6. All other encoders have memory 4, but they differ in the code rate and the polynomials used.

Table 4.16 gives an overview of the uses and combinations of generator polynomials. Most logical channels use a convolutional code of rate 1/2 based on polynomials G_0 and G_1 .

Table 4.15 Generator polynomials for convolutional codes.

Type	Polynomial
G0	$1 + d^3 + d^4$
G1	$1 + d + d^3 + d^4$
G2	$1 + d^2 + d^4$
G3	$1 + d + d^2 + d^3 + d^4$
G4	$1 + d^2 + d^3 + d^5 + d^6$
G5	$1 + d + d^4 + d^6$
G6	$1 + d + d^2 + d^3 + d^4 + d^6$

Table 4.16 Usage of generator polynomials.

Channel type	Generator polynomial						
	G0	G1	G2	G3	G4	G5	G6
TCH, full rate, speech							
Class I							
Class II							
TCH, half rate, speech							
Class I							
Class II							
TCH, full rate, 14.4 kbit/s							
TCH, full rate, 9.6 kbit/s							
TCH, full rate, 4.8 kbit/s							
TCH, half rate, 4.8 kbit/s							
TCH, full rate, 2.4 kbit/s							
TCH, half rate, 2.4 kbit/s							
FACCHs							
SDCCHs, SACCHs							
BCCH, AGCH, PCH							
RACH							
SCH							

Speech traffic channels

Convolutional coding of class I speech bits on the full-rate speech channel generates $1/r \times (182 + 3 + 4)$ bits = 378 bits. The 78 bits of class II are not encoded at all, which results in a total of 456 bits. This is the uniform block size needed for mapping these data blocks onto the bursts with 114-bit payload. In a similar way, for most of the remaining channels, two coded blocks of 228 bits are combined.

In the case of a half-rate speech traffic channel (TCH/HS), the class I bits are encoded using a punctured version of a rate-1/3 convolutional encoder defined by G4, G5 and G6. Including puncturing, the net rate of the encoder is $r' = 104/211 \approx 1/2$. The 95 information bits, 3 parity bits and 6 tail bits are thus mapped to 211 bits. The 17 class II bits are not convolutional encoded, which results in a total number of 228 bits.

Data traffic channels

The 4.8 kbit/s data service on a full-rate channel (TCH/F4.8) and the 2.4 kbit/s data service on a half-rate channel (TCH/H2.4) use a code of rate 1/3 based on polynomials G1, G2 and G3. The 2.4 kbit/s data service on a full-rate channel (TCH/F2.4) uses these three polynomials twice in a row to generate a convolutional code of rate 1/6.

The data on a TCH/F9.6 and TCH/H4.8 is encoded using the rate-1/2 code defined by the polynomials G0 and G1. At the input of the convolution encoder, blocks of 244 bits arrive which the encoder maps to blocks of 488 bits. These blocks are reduced to 456 bits by removing (puncturing) every 15th bit beginning with the 11th bit, i.e. a total of 32 bits are punctured. On the one hand, puncturing cuts the block size to a length suitable for further processing; on the other hand, puncturing removes redundancy. The resulting net code rate of $r' = 244/456$ is therefore somewhat higher than the rate of 1/2 for the convolutional encoder. Thus, the code has slightly lower error-correction capability. Puncturing cuts down the convolutional coded blocks of the TCH/F9.6 and TCH/H4.8 channels to the standard format of 456 bits. Thus, blocks of these channels can also be processed in a standardized way (interleaving, etc.), and the amount of redundancy contained in a block is also matched to the bit rate available for transmission.

For the encoding of the TCH/F14.4, again a punctured version of the (G0, G1) convolutional encoder is employed. The 294 bits at the input of the encoder are mapped to 588 bits, followed by a puncturing of 132 bits.

Convolutional decoding

In most cases, the decoding of a convolutional code employs the Viterbi algorithm. It uses a suitable metric to determine the data sequence that most likely equals the transmitted data (maximum likelihood decoding) (Bossert, 1999). Using the knowledge of the generator polynomials, the decoder can determine the original data sequence.

4.8.3 Interleaving

The decoding result of the convolutional code is highly dependent on the frequency and grouping of bit errors that occur during transmission. In particular, burst errors during long and deep fading periods, i.e. a series of erroneous sequential bits, have a negative impact on error correction. In such cases, the channel is not a binary channel without memory, rather the single-bit errors have statistical dependence, which diminishes the result of the error-correction procedure of the convolutional code. To achieve good error correction results, the channel should have no memory, i.e. the bit errors should be statistically independent. Therefore, burst errors occurring frequently on the radio channel should be distributed uniformly across the transmitted codewords. This can be accomplished through the interleaving technique described in the following.

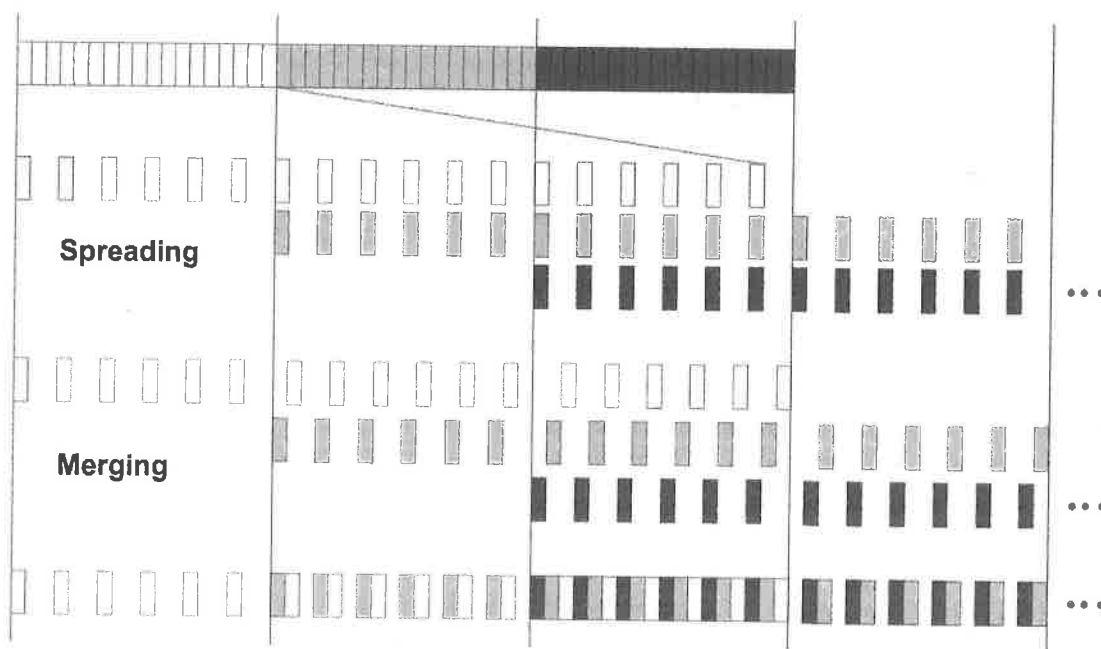


Figure 4.34 Interleaving: spreading and merging.

The interleaving approach is to distribute codewords from the convolutional encoder by spreading in time and merging them across several bursts for transmission. This principle is shown in Figure 4.34. By time spreading, each of the codewords is distributed across a threefold length. Merging the bit sequences generated in this way has the effect that the individual bits from each of the three codewords are sorted into alternate bursts; this way each codeword is transmitted as distributed over a total of three bursts, and two bits of a data block are never transmitted adjacent to each other.

This kind of interleaving is also known as diagonal interleaving. The number of bursts over which a codeword is spread is called the interleaving depth; a spreading factor can be defined analogously. A burst error is therefore distributed uniformly over several subsequently transmitted codewords because of the distribution of the data over several bursts. This generates bit error sequences which are less dependently distributed in the data stream, hence it improves the success of the error-correction process.

Figure 4.35 shows an example. During the third burst of transmission, severe fading of the signal leads to a massive burst error. This burst is now heavily affected by a total of six single-bit errors. In the process of deinterleaving (inversion of merging, despreading) these bit errors are distributed across three data blocks, corresponding to the bit positions which were sorted into the respective bursts during interleaving. The number of errors per data block is now only two, which can be much more easily corrected.

Another kind of interleaving is block interleaving. In this principle, codewords are written line by line into a matrix (Figure 4.36), which is subsequently read out column by column. The number of lines of the interleaving matrix determines the interleaving depth. As long as the length of a burst error is shorter than the interleaving depth, the burst error generates only single-bit errors per codeword if block interleaving is used (Bossert, 1999; Steele, 1992).

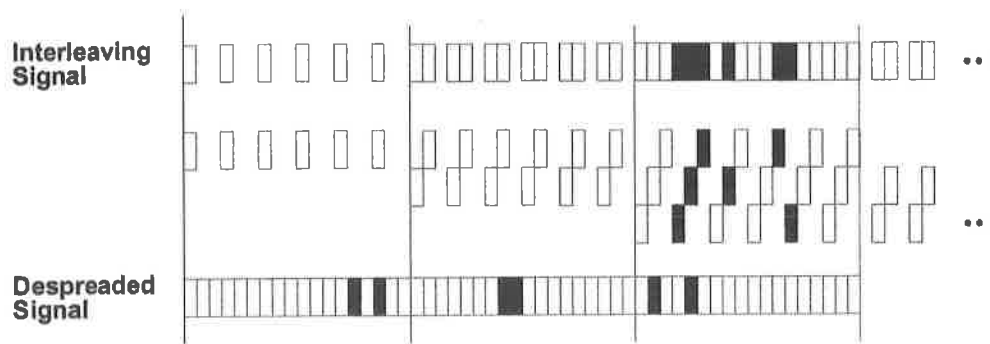


Figure 4.35 Distributing bit errors through deinterleaving.

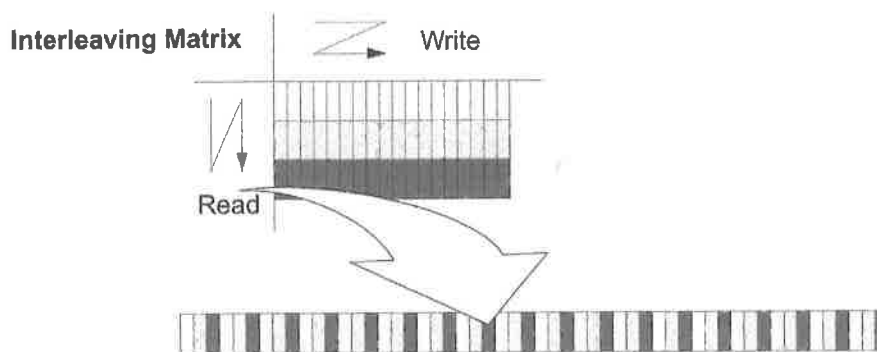


Figure 4.36 Principle of block interleaving.

However, the great advantage of interleaving, to alleviate the effect of burst errors for optimal error correction with a convolutional code, is traded for a significant disadvantage for speech and data communication. As evident from Figures 4.34 and 4.36, the bits of a codeword are spread across several bursts (three in the example here). For a complete reconstruction of a codeword, one has to wait for the complete transmission of three bursts. This forces a transmission delay, which is a function of the interleaving depth.

In GSM, both methods of interleaving are used (Figure 4.37), blockwise as well as bitwise. With a maximal interleaving depth of 19, this can lead to delays of up to 360 ms (Table 4.17).

Full-rate speech channel, TCH/F2.4 and FACCH

The speech channel TCH/FS in GSM uses block-diagonal interleaving. The 456 bits of a codeword are distributed across eight interleaving blocks, where one interleaving block has 114 bit positions. The exact interleaving rule for mapping the coded bits $c(n, k = 0, \dots, 455)$ of the n th codeword, onto bit position $i(b, j = 0, \dots, 114)$ of the b th interleaving block, is

$$i(b, j) = c(n, k),$$

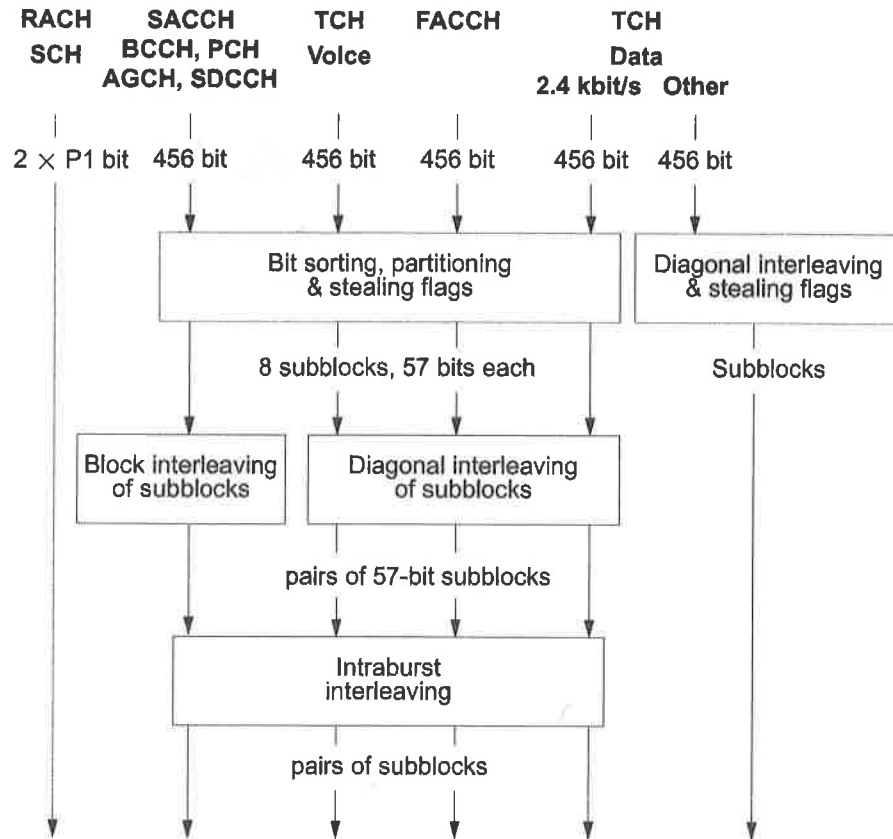


Figure 4.37 Overview of the interleaving of (full-rate) logical channels.

with

$$\begin{cases} n = 0, 1, 2, \dots, N, N + 1, \dots, \\ k = 0, 1, 2, \dots, 455, \\ b = b_0 + 4n + (k \bmod 8), \\ j = 2((49k) \bmod 57) + ((k \bmod 8) \operatorname{div} 4). \end{cases}$$

The bits of the n th codeword (data block n in Figure 4.38) are distributed across eight interleaving blocks, beginning with block $B = b_0 + 4n$. To do so, the coded bits are mapped to the even bits of the first four interleaving blocks ($B + 0, \dots, B + 3$) and to the odd bits of the other four interleaving blocks ($B + 4, \dots, B + 7$). The even bits of the last four interleaving blocks ($B + 4, \dots, B + 7$) are occupied by data from codeword $n + 1$. Each interleaving block thus contains 57 bits of the current codeword n and 57 bits of the following codeword $n + 1$ or the preceding codeword $n - 1$, respectively. In this way, a new codeword is started after each fourth merged interleaving block.

The individual bits of codeword n are alternatively distributed across the interleaving blocks, e.g., every eighth bit is in the same interleaving block according to the term $(k \bmod 8)$, whereas bit position j within an interleaving block $b = B + 0, B + 1, \dots, B + 7$ is determined by two terms: the term $(k \bmod 8) \operatorname{div} 4$ is used to determine the even/odd bit positions; and the term $2((49k) \bmod 57)$ determines the offset within the interleaving

Table 4.17 Transmission delay caused by interleaving.

Channel type	Interleaving depth	Transmission delay (ms)
TCH, full rate, voice	8	38
TCH, half rate, voice	4	
TCH, full rate, 14.4 kbit/s	19	93
TCH, full rate, 9.6 kbit/s	19	93
TCH, full rate, 4.8 kbit/s	19	93
TCH, half rate, 4.8 kbit/s	19	185
TCH, full rate, 2.4 kbit/s	8	38
TCH, half rate, 2.4 kbit/s	19	185
FACCH, full rate	8	38
FACCH, half rate	8	74
SDCCH	4	14
SACCH/TCH	4	360
SACCH/SDCCH	4	14
BCCH, AGCH, PCH	4	14

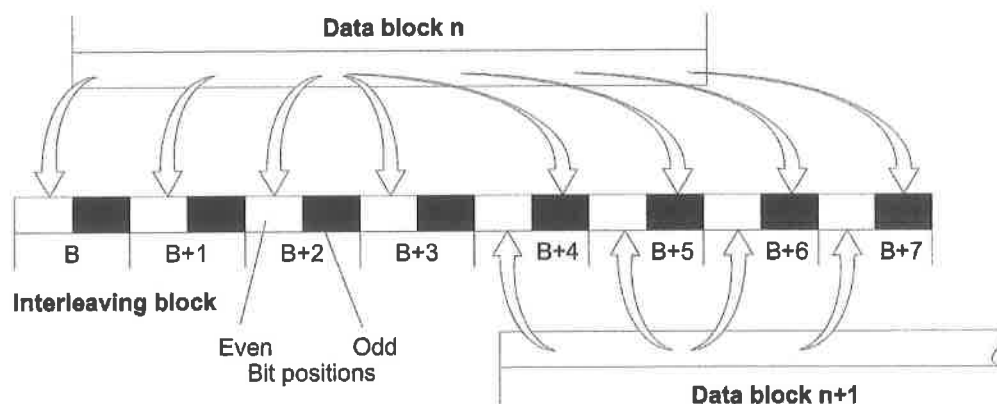


Figure 4.38 Interleaving TCH/FS: block mapping.

block. The first interleaving block B derived from codeword n thus contains bit numbers 0, 8, 16, ..., 448, 456 of this codeword.

The placement of these bits for the first block B in the interleaving block is illustrated in Figure 4.39. This placement is chosen in such a way that no two directly adjacent bits of the interleaving block belong to the same codeword. In addition, the mapped bits are combined into groups of eight bits each, which are distributed as uniformly as possible across the entire interleaving block. This achieves additional spreading of error bursts within a data block. Therefore, the interleaving for the TCH/FS is block-diagonal interleaving with additional merging of data bits within the interleaving block. This is also called intraburst interleaving.

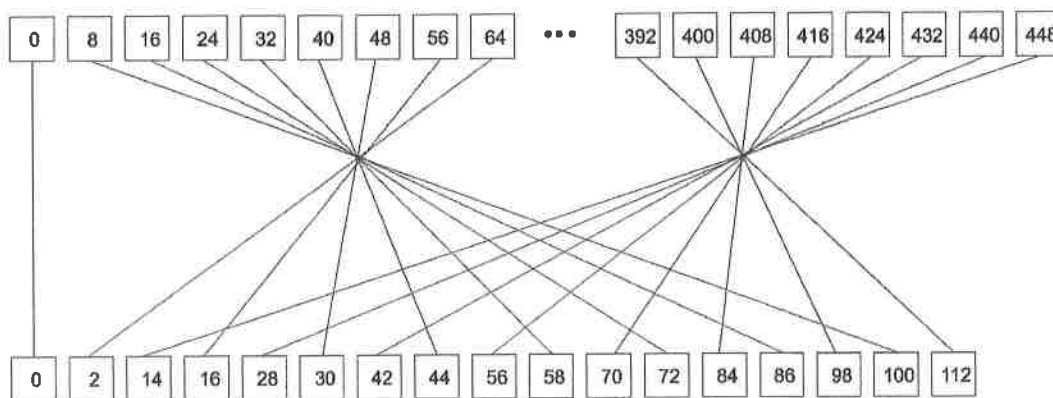


Figure 4.39 Mapping of codeword n onto interleaving block B for a TCH/FS.

(Figure 4.37). The data channel TCH/F2.4 and the FACCH in GSM use the same interleaving methods as the TCH/FS.

Other data traffic channels

For the other data services in the traffic channel (TCH/F14.4, TCH/F9.6, TCH/F4.8, TCH/H4.8 and TCH/H2.4) the interleaving is somewhat simpler. A pure bitwise diagonal interleaving with an interleaving depth of 19 is used. In this case, the interleaving rule is

$$i(b, j) = c(n, k),$$

with

$$\begin{cases} n = 0, 1, 2, \dots, N, N + 1, \dots, \\ k = 0, 1, 2, \dots, 455, \\ b = b_0 + 4n + (k \bmod 19) + k, \\ j = k \bmod 19 + 19((k \bmod 6) \div 114). \end{cases}$$

The bits of a data block (n, k) are distributed in groups of 114 bits across 19 interleaving blocks, whereby groups of six bits are distributed uniformly over one interleaving block.

With this diagonal interleaving, each interleaving block also starts a new 114-bit block of data. A closer look at this interleaving rule reveals that the input to the interleaver consists of blocks of 456 coded data bits as codewords. The whole codeword is therefore really spread across 22 interleaving blocks; the nominal interleaving depth of 19 results historically from 114-bit block interleaving.

Half-rate speech channel

The interleaving rule for the half-rate speech channel (TCH/HS) is given by

$$i(b, j) = c(n, k),$$

with

$$\begin{cases} n = 0, 1, 2, \dots, N, N + 1, \dots, \\ k = 0, 1, 2, \dots, 227, \\ b = b_0 + 2n + (k \bmod 4), \end{cases}$$

and j according to a table in the GSM standard. The 228 bits of a codeword n are distributed over 4 blocks. Beginning with interleaving block $B = b_0 + 2n$, it occupies the even-numbered bits of the first two interleaving blocks ($B + 0, B + 1$) and the odd-numbered bits of the other two blocks ($B + 2, B + 3$). Consequently, the following codeword $n + 1$ uses the even-numbered bits of the blocks $B + 2 (= b_0 + 2(n + 1) + 0)$ and $B + 3 (= b_0 + 2(n + 1) + 1)$ as well as the odd-numbered bits of the interleaving blocks $b_0 + 2(n + 1) + 2$ and $b_0 + 2(n + 1) + 3$. As with the TCH/FS, one interleaving block contains 57 bits from codeword n and 57 bits from codeword $n + 1$ or $n - 1$. In summary, a new codeword starts every second interleaving block.

Signaling channels

Most signaling channels use an interleaving depth of 4, such as SACCH, BCCH, PCH, AGCH and SDCCH. The interleaving scheme is almost identical to that used for the TCH/FS; however, the codewords $c(n, k)$ are spread across four rather than eight interleaving blocks:

$$i(b, j) = c(n, k),$$

with

$$\begin{cases} n = 0, 1, 2, \dots, N, N + 1, \dots, \\ k = 0, 1, 2, \dots, 455, \\ b = b_0 + 4n + (k \bmod 4), \\ j = 2((49k) \bmod 57) + ((k \bmod 8) \div 4). \end{cases}$$

With this kind of interleaving, there are also eight blocks generated, just as in the case of the TCH/FS; however, at the same time a block of 57 even bits is combined with a block of 57 odd bits to form a complete interleaving block. This has the consequence that consecutive coded signaling messages are not block-diagonally interleaved, but that each four consecutive interleaving blocks are fully occupied with the data of just one, and only one, codeword. Also, a new codeword starts after every four interleaving blocks. Therefore, this interleaving of GSM signaling messages is in essence also a block interleaving procedure. This is especially important for signaling channels to ensure the transmission of individual protocol messages independent of preceding or succeeding messages. This also enables some kind of asynchronous communication of signaling information. The signaling data of the RACH and SCH must each be transmitted in single data bursts; no interleaving occurs.

4.8.4 Mapping onto the burst plane

After block encoding, convolutional encoding and interleaving, the data are available in the form of 114-bit interleaving blocks. This corresponds exactly to the amount of data which can be carried by a NB (Figure 4.7). Each interleaving block is mapped directly onto one burst (Figure 4.40). After setting the stealing flags, the bursts can be composed and passed to the

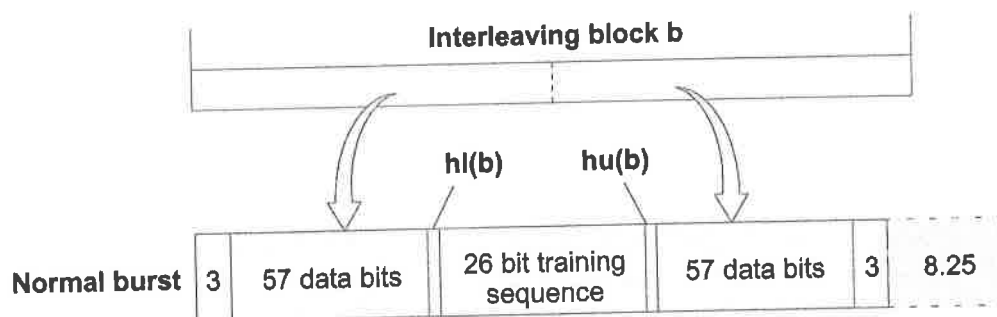


Figure 4.40 Mapping onto a burst.

modulator. The stealing flags indicate whether high-priority signaling messages are present (FACCH messages), which must be transmitted as fast as possible, instead of the originally planned data of the traffic channel.

An essential component of GSM channel coding is the correct treatment of FACCH signaling messages which are multiplexed in a preemptive way into the traffic channel. At the burst level, each FACCH codeword displaces a codeword of the current TCH/FS traffic channel, i.e. the codewords must be tied into the interleaving structure instead of regular data blocks of the traffic channel. The interleaving rule for the FACCH is the same as for the TCH/FS: from an FACCH codeword the even positions are occupied in one set of four interleaving blocks, and the odd bit positions are occupied in another set of interleaving blocks; in addition, the bit positions within the interleaving blocks are shuffled (intra-burst interleaving).

When an FACCH message needs to be transmitted (e.g. a handover command), the current data block n is replaced by the convolutional coded FACCH message, and it is interleaved in a block-diagonal way with the data blocks $(n - 1)$ and $(n + 1)$ of the traffic channel (Figure 4.41). In the eight blocks involved in this procedure, $B, B + 1, \dots, B + 7$, the respective stealing flags $hl(b)$ and $hu(b)$ have to be set (Figure 4.40). If neither flag is set, the burst contains data of the traffic channel. If the even bits of the burst are occupied by FACCH data, $hu(b)$ is set; in the case of the odd bits being used for FACCH, $hl(b)$ is set (Figure 4.41).

If the current burst is not available for traffic channel data, the data block n has to be discarded. Bits 'stolen' in this way have various effects.

- A complete speech frame of the TCH/FS is lost (20 ms speech).
- In the case of TCH/F9.6 and TCH/H4.8 channels, 3 bits are stolen from each of the 8 interleaving blocks, which belong to the same data block, such that a maximum of 24 coded data bits are interfered with.
- In the case of TCH/F4.8 and TCH/H2.4 channels, 6 bits are stolen from each of the 8 interleaving blocks, which belong to the same data block, such that a maximum of 48 coded data bits are interfered with.
- In the case of TCH/F2.4 channels, the same interleaving rules as in the TCH/FS are used, such that a complete data block is displaced by the FACCH.

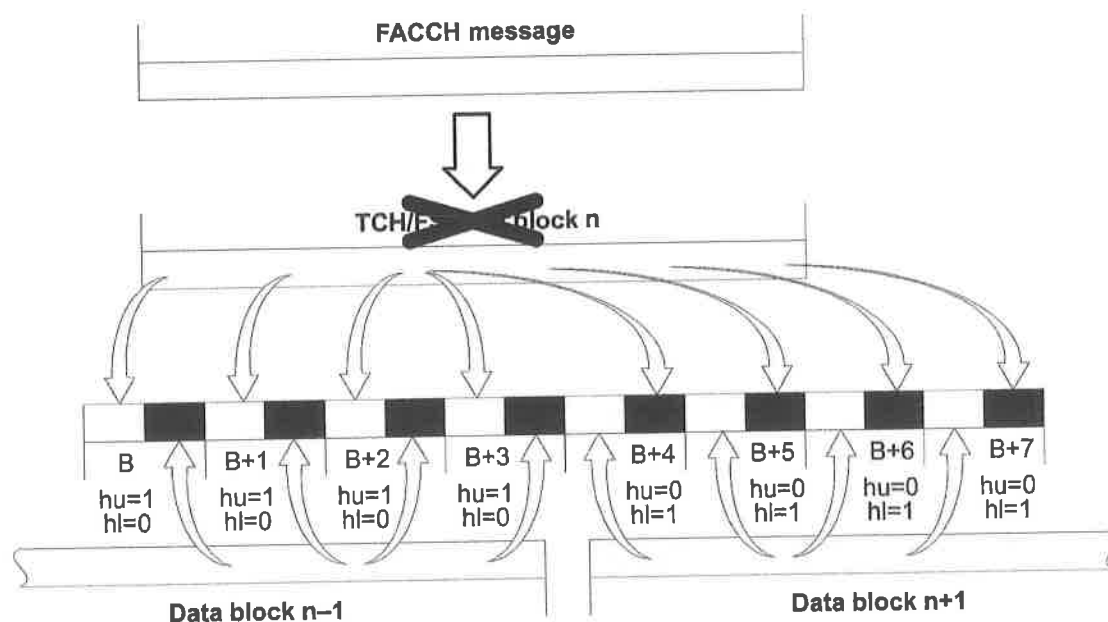


Figure 4.41 Insertion of an FACCH message into the TCH/FS data stream.

In summary, the FACCH signaling needed for fast reactions causes data losses or bit errors in the accompanying traffic channel, and they have to be totally or partially corrected by the convolutional code.

4.8.5 Improved codecs for speech services: half-rate codec, enhanced full-rate codec and adaptive multi-rate codec

One of the most important services in GSM is (of course) voice service. Thus, it is obvious that the voice service has to be further improved. The priority is the development of new speech codecs with two competing objectives:

- better utilization of the frequency bands assigned to GSM; and
- improvement of speech quality to a level that is similar to that offered by ISDN networks, which is the primary request of professional users.

Half-rate codec

The reason for improved bandwidth utilization is to increase the network capacity and the spectral efficiency (i.e. traffic carried per cell area and frequency band). Early plans were already in place to introduce a *half-rate speech codec*. Under good channel conditions, this codec achieves, in spite of the half bit rate, almost the same speech quality as the full-rate codec used so far. However, quality loss occurs in particular for mobile-to-mobile communication, since in this case (due to the ISDN architecture) one has to go twice through the GSM speech coding/decoding process. These multiple, or tandem, conversions degrade speech quality. The end-to-end transmission of GSM-coded speech is intended to avoid

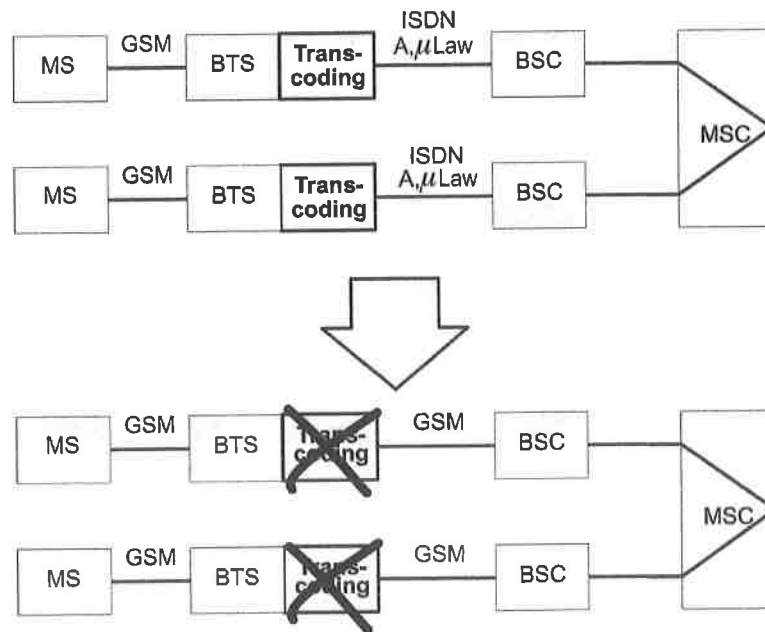


Figure 4.42 Through-transport of GSM-coded speech in Phase 2+ for mobile-to-mobile connections (TFO).

multiple unnecessary transcoding and the resulting quality loss (Figure 4.42) (Mouly and Pautet, 1995). This technique has been passed under the name Tandem Free Operation (TFO) in GSM Release 98.

Enhanced full-rate codec

A very important concern is the improvement of speech quality. Speech quality that is close to what is found in fixed networks is particularly important for business applications and in cases where GSM systems are intended to replace fixed networks, e.g., for fast installation of telecommunication networks in areas with insufficient or missing telephone infrastructure.

Work on the Enhanced Full-Rate (EFR) codec was therefore considered as high priority. This EFR is a full-rate codec (net bit rate 12.2 kbit/s). Nevertheless, it achieves speech quality that is clearly superior to the previously used full-rate codec. It has been initially standardized and used in North American DCS1900 networks (Mouly and Pautet, 1995) and has been implemented in GSM with very good success. Instead of using the RPE-LTP coding scheme (section 4.7), a so-called Algebraic Code Excitation – Linear Prediction (ACELP) is employed.

The EFR speech coder delivers data blocks of 244 information bits to the channel encoder (compare with Table 4.11). In addition to grading the bits into important class I bits and less important class II bits, EFR further divides into class Ia bits and class Ib bits. A special preliminary channel coding is employed for the most significant bits: eight parity bits (generated by a CRC) and eight repetition bits are added to provide additional error detection. The resulting 260 bits are processed by the block encoder as described in section 4.8.1.

For convolutional coding of class I bits the convolutional encoder defined by the generator polynomials G0 and G1 is employed.

Adaptive multi-rate codec

The speech codecs mentioned before (full rate, half rate and EFR) all use a fixed source/information bit rate, which has been optimized for typical radio channel conditions. The problem with this approach is its inflexibility: whenever the channel conditions are much worse than usual, very poor speech quality will result, since the channel capacity assigned to the MS is too small for error free transmission. On the other hand, radio resources will be wasted for unneeded error protection if the radio conditions are better than usual.

To overcome these problems, a much more flexible codec has been developed and standardized: the Adaptive Multi-rate (AMR) codec. It can improve speech quality by adaptively switching between different speech coding schemes (with different levels of error protection) according to the current channel quality. To be more precise, AMR has two principles of adaptability (Bruhn *et al.*, 2000): channel mode adaptation and codec mode adaptation.

Channel mode adaptation dynamically selects the type of traffic channel that a connection should be assigned to: either a full-rate (TCH/F) or a half-rate traffic channel (TCH/H). The basic idea here is to adapt a user's gross bit rate in order to optimize the use of radio resources. If the traffic load in a cell is high, those connections using a TCH/F (gross bit rate 22.8 kbit/s) and having high channel quality should be switched to a TCH/H (11.4 kbit/s). On the other hand, if the load is low, the speech quality of several TCH/H connections can be improved by switching them to a TCH/F. The signaling information for this type of adaptation is done with existing protocols on GSM signaling channels; the switching between full-rate and half-rate channels is realized by an intracell handover.

The task of codec mode adaptation is to adapt the coding rate (i.e. the tradeoff between the level of error protection versus the source bit rate) according to the current channel conditions. When the radio channel is bad, the encoder operates at low source bit rates at its input and uses more bits for forward error protection. When the quality of the channel is good, less error protection is employed.

The AMR codec consists of eight different modes with source/information bit rates ranging from 12.2 to 4.75 kbit/s (see Table 4.18). All modes are scaled versions of a common ACELP basis codec.

From the results of link quality measures, an adaptation unit selects the most appropriate codec mode. Figure 4.43 illustrates the AMR encoding principle. Channel coding is performed using a punctured recursive systematic convolutional code. Since not all bits of the voice data are equally important for audibility, AMR also employs an UEP structure. The most important bits (class Ia; e.g. mode bits and LPC coefficients) are additionally protected by a CRC code with six parity bits. On the receiver side, the decoder will discard the entire speech frame if the parity check fails. Also the degree of puncturing depends on the importance of the bits. At the end of the encoding process, a block with a fixed number of gross bits results, which is subsequently interleaved to reduce the number of burst errors.

Since the channel conditions can change rapidly, codec mode adaptation requires a fast signaling mechanism. This is achieved by transmitting the information about the used codec

Table 4.18 AMR codec modes.

	Source data rate (kbit/s)							
	12.2	10.2	7.95	7.4	6.7	5.9	5.15	4.75
Information bits per block	244	204	159	148	134	118	103	95
Class 1a bits (CRC protected)	81	65	75	61	55	55	49	39
Class 1b bits (not CRC protected)	163	139	84	87	79	63	54	56
Rate R of convolutional encoder	1/2	1/3	1/3	1/3	1/4	1/4	1/5	1/5
Output bits from convolutional encoder	508	642	513	474	576	520	565	535
Punctured bits	60	194	65	26	128	72	117	87

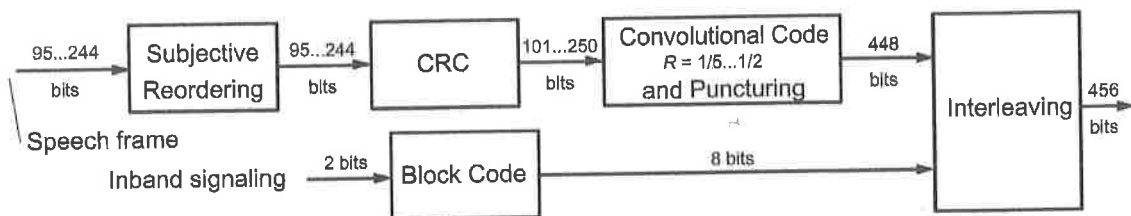


Figure 4.43 AMR channel encoding principle (bit numbers for TCH/F).

mode, link control, and DTX, etc. together with the speech data in the TCH, i.e. a special inband signaling is employed.

We give an example: the 12.2 kbit/s codec for a TCH/F operates with 244 source bits ($12.2 \text{ kbit/s} \times 20 \text{ ms}$), which are first rearranged to subjective importance. By adding six CRC bits for class 1a bits, we obtain 250 bits. The subsequent recursive convolutional encoder, defined by the two generators 1 and $G_1/G_0 = (d^4 + d^3 + d + 1)/(d^4 + d^3 + d)$, with rate $R \approx 1/2$, maps those bits to 508 bits. Next, 60 bits are punctured, which results in an output sequence of 448 bits. Together with the encoded inband signaling (8 bits) this block is interleaved and finally mapped to bursts. The resulting gross bit rate is thus $456 \text{ bits}/20 \text{ ms} = 22.8 \text{ kbit/s}$.

4.9 Power-up scenario

At this point, all of the functions, protocols and mechanisms of the GSM radio interface have been presented which are needed to illustrate a basic power-up scenario. The following describes the basic events that occur during a power up of the MS. The scenario can be divided into several steps.

1. Provided that a SIM card is present, immediately after turning on power, a MS starts the search for BCCH carriers. Normally, the station has a stored list of up

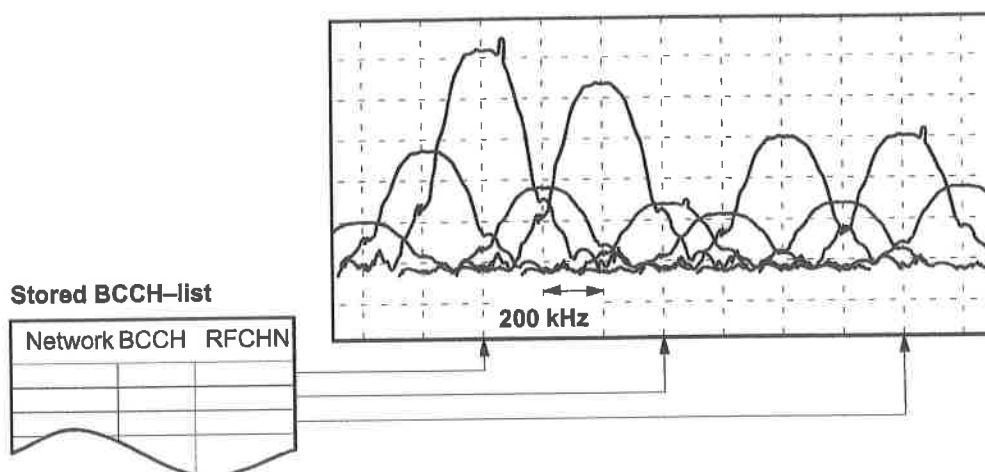


Figure 4.44 BCCH search in the power density spectrum (schematic).

to 32 carriers (Figure 4.44) of the current network. Signal level measurements are performed on each of these frequencies (RXLEV). Alternatively, if no list is available, all GSM frequencies have to be measured to find potential BCCH carriers. Using the path loss criterion C1 and the threshold values stored with the list of carriers (RXLEV_ACCESS_MIN, MS_TXPWR_MAX_CCH), a first ordering can be done.

2. After having found potential candidates based on the received signal level RXLEV, each carrier is investigated for the presence of an FCCH signal, beginning with the strongest signal. Its presence identifies the carrier as a BCCH carrier for synchronization. Using the sine wave signal allows coarse time synchronization as well as fine tuning of the oscillator.
3. The SB of the SCH in the TDMA frame immediately following the FCCH burst (Figure 4.17) has a long training sequence of 64 bits (Figure 4.7) which is used for fine tuning of the frequency correction and time synchronization. In this way the MS is able to read and decode synchronization data from the SCH, the BSIC and the RFN. This process starts with the strongest of all BCCH carriers. If a cell is identified using BSIC and path loss criterion C1, the cell is selected for camping on it.
4. The exact channel configuration of the selected cell is obtained from the BCCH data as well as the frequencies of the neighboring cells. The MS can now monitor the PCH of the current cell and measure the signal levels of the neighboring cells.
5. The MS must now prepare synchronization with the six cells with the strongest signal level (RXLEV) and read out their BCCH/SCH information, i.e. steps 1–4 above are to be performed continuously for the six neighboring cells with the best RXLEV values.
6. If significant changes are noticed using the path loss criterion C1 and the reselection criterion C2, the MS can start reselection of a new cell. Both criteria are determined periodically for the current BCCH and the six strongest neighbors.

To limit power consumption and to extend standby time of the battery, the mobile station can activate the DRX mode.

Protocols

5.1 Protocol architecture planes

The various physical aspects of radio transmission across the GSM air interface and the realization of physical and logical channels were explained in Chapter 4. According to the terminology of the OSI Reference Model, these logical channels are at the Service Access Point of Layer 1 (physical layer), where they are visible to the upper layers as transmission channels of the physical layer. The physical layer also includes the forward error correction and the encryption of user data.

The separation of logical channels into the two categories of control channels (signaling channels) and traffic channels (Table 4.1) corresponds to the distinction made in the ISDN Reference Model between user plane and control plane. Figure 5.1 shows a simplified reference model for the GSM User–Network Interface (UNI) Um, where the layer-transcending management plane is not elaborated on further in the following. In the user plane, protocols of the seven OSI layers are defined for the transport of data from a subscriber or a data terminal. User data are transmitted in GSM across the air interface over traffic channels TCH, which therefore belong to Layer 1 of the user plane (Figure 5.1).

Protocols in the signaling plane are used to handle subscriber access to the network and for the control of the user plane (reservation, activation, routing, switching of channels and connections). In addition, signaling protocols between network nodes are needed (network internal signaling). The Dm channels of the air interface in GSM are signaling channels and are therefore realized in the signaling plane (Figure 5.1).

Since signaling channels are physically present but mostly unused during an active user connection, it is obvious to use them also for the transmission of certain user data. In ISDN, packet-switched data communication is therefore permitted on the D channel, i.e. the physical D channel carries multiplexed traffic of signaling data (s-data) and user (payload) data (p-data). The same possibility also exists in GSM. Data transmission without allocation of a dedicated traffic channel is used for SMS by using free capacities on signaling channels. For this purpose, a separate SDCCH is allocated or, if a traffic connection exists, the SMS protocol data units are multiplexed onto the signaling data stream of the SACCH (Figure 5.2).

The control (signaling) and user plane can be defined and implemented separately, ignoring for the moment that control and user data have to be transmitted across the same

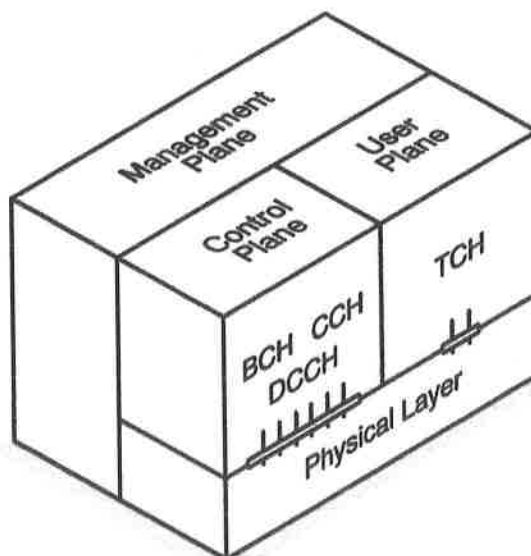


Figure 5.1 Logical channels at the air interface in the ISDN reference model.

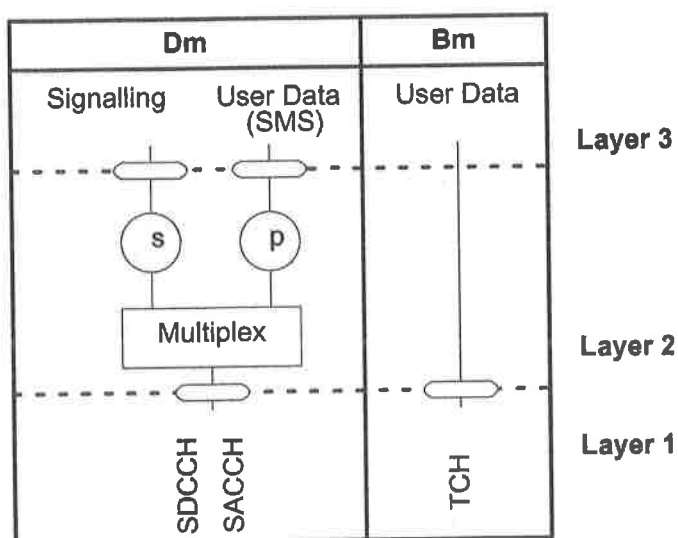


Figure 5.2 User data and control at the air interface.

physical medium at the air interface and that signaling procedures initiate and control activities in the user plane. Therefore, for each plane there exists a corresponding separate protocol architecture within the GSM system: the user data protocol architecture (section 5.2) and the signaling protocol architecture (section 5.3), with an additional separate protocol architecture for the transmission of p-data on the control (signaling) plane (section 5.3.2). A protocol architecture comprises not only the protocol entities at the radio interface Um but all protocol entities of the GSM network components.

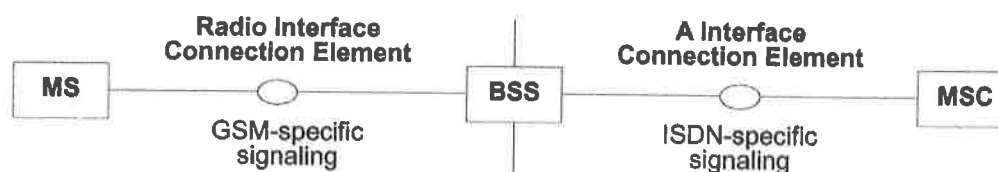


Figure 5.3 Connection elements.

5.2 Protocol architecture of the user plane

A GSM PLMN can be defined by a set of access interfaces (Appendix A.1) and a set of connection types used to realize the various communication services. A connection in GSM is defined between reference points. Connections are constructed from connection elements (Figure 5.3), and the signaling and transmission systems may change from element to element. Two elements therefore exist within a GSM connection: the radio interface connection element and the A interface connection element. The radio interface and the pertinent connection element are defined between the MS and the BSS, whereas the A interface connection element exists between BSS and MSC across the A interface. A GSM-specific signaling system is used at the radio interface, whereas ISDN-compatible signaling and payload transport are used across the A interface. The BSS is subdivided into BTS and BSC. Between them they define the Abis interface, which has no connection element defined; this is because it is usually transparent for user data.

A GSM connection type provides a way to describe GSM connections. Connection types represent the capabilities of the lower layers of the GSM PLMN. In the following section, the protocol models are presented as the basis for some of the connection types defined in the GSM standards. These are speech connections and transparent as well as nontransparent data connections. A detailed discussion of the individual connection types can be found in Appendix A with a description of how classical data services have been realized in GSM.

5.2.1 Speech transmission

The digital, source-coded speech signal of the MS is transmitted across the air interface in an error-protected and encrypted form. The signal is then deciphered in the BTS, and the error protection is removed before the signal is passed on. This specially protected speech transmission occurs transparently between the MS and a Transcoding and Rate Adaptation Unit (TRAU) which serves to transform the GSM speech-coded signals into the ISDN standard format (ITU-T A-law). A possible transport path for speech signals is shown in Figure 5.4, where the bit transport plane (encryption and TDMA/FDMA) has been omitted.

A simple GSM speech terminal (MT0, see also Figure A.1) contains a GSM Speech Codec (GSC) for speech coding. Its speech signals are transmitted to the BTS after channel coding (FEC) and encryption, where they are again deciphered, decoded and, if necessary, error-corrected. More than one GSM speech signal can be multiplexed onto an ISDN channel, with up to four GSM speech signals (at 13 kbit/s each) per ISDN B channel (64 kbit/s). Before they are passed to the MSC, speech signals are transcoded in the BSS from GSM format to ISDN format (ITU-T A-law).

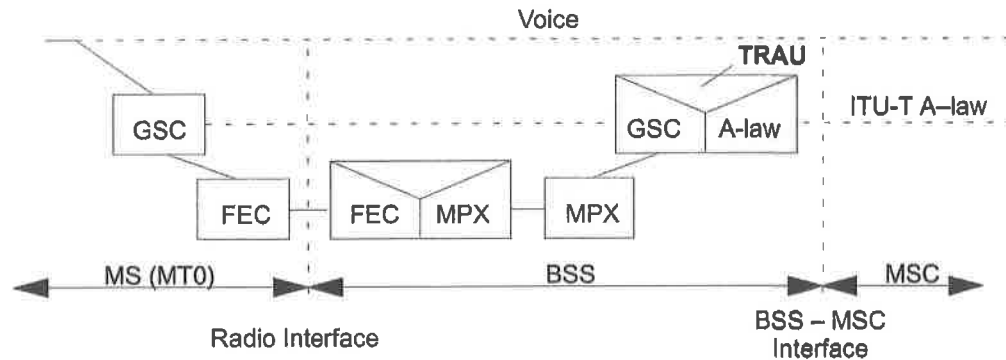


Figure 5.4 Speech transmission in GSM.

The BTSs are connected to the BSC over digital fixed lines, usually leased lines or microwave links, with typical transmission rates of 2048 kbit/s (in Europe), 1544 kbit/s (in the USA) or 64 kbit/s (ITU-T G.703, G.705, G.732). For speech transmission, the BSS implements channels of 64 or 16 kbit/s. The physical placement of the TRAU largely determines which kind of speech channel is used in the fixed network. The TRAU performs the conversion of speech data between GSM format (13 kbit/s) and ISDN A-law format (64 kbit/s). In addition, it is responsible for the adaptation of data rates, if necessary, for data services. There are two alternatives for the positioning of the TRAU: the TRAU can be placed in the BTS or outside the BTS in the BSC. An advantage of placing the TRAU outside the BTS is that up to four speech signals can be submultiplexed (MPX in Figure 5.4) onto an ISDN B channel, so that less bandwidth is required on the BTS-to-BSC connection. Beyond this consideration, placing the TRAU outside the BTS allows the TRAU functions to be combined for all BTSs of a BSS in one separate hardware unit, perhaps produced by a separate manufacturer. The TRAU is, however, always considered as part of the BSS and not as an independent network element.

Figure 5.5 shows some variants of TRAU placement. A BTS consists of a Base Control Function (BCF) for general control functions like frequency hopping, and several (at least one) Transceiver Function (TRX) modules which realize the eight physical TDMA channels on each frequency carrier. The TRX modules are also responsible for channel coding and decoding as well as encryption of speech and data signals. If the TRAU is integrated into the BTS, speech transcoding between GSM and ISDN formats is also performed within the BTS.

In the first case, TRAU within the BTS (BTS 1,2,3 in Figure 5.5), the speech signal in the BTS is transcoded into a 64 kbit/s A-law signal, and a single speech signal per B channel (64 kbit/s) is transmitted to the BSC/MSC. For data signals, the bit rates are adapted to 64 kbit/s, or several data channels are submultiplexed over one ISDN channel. The resulting user plane protocol architecture for speech transport is shown in Figure 5.6.

GSM-coded speech (13 kbit/s) is transmitted over the radio interface (Um) in a format that is coded for error protection and encryption. At the BTS site, the GSM signal is transcoded into an ISDN speech signal and transmitted transparently through the ISDN access network of the MSC.

In the second case, the TRAU resides outside the BTS (BTS 4 in Figure 5.5) and is considered as part of the BSC. However, physically it could also be located at the MSC site,

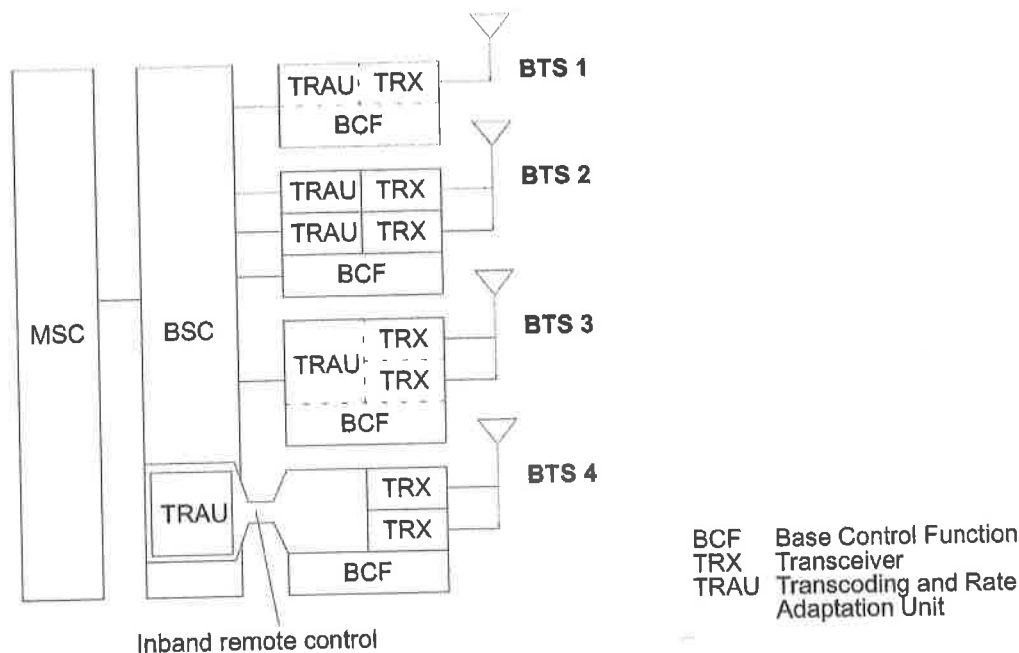


Figure 5.5 BTS architecture variations and TRAU placement.

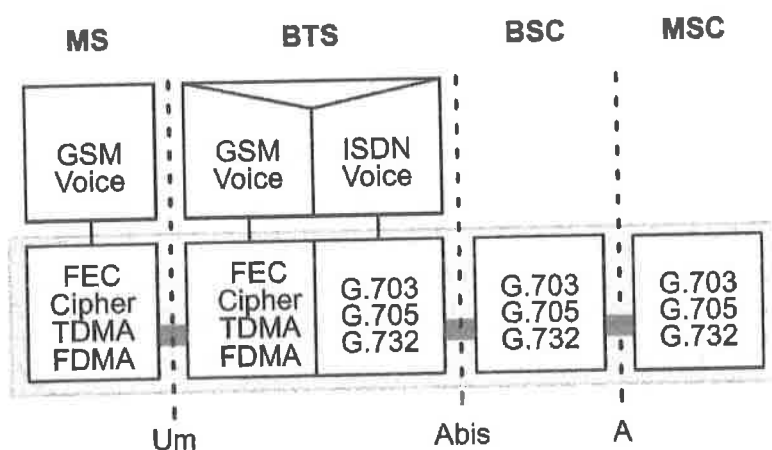


Figure 5.6 GSM protocol architecture for speech (TRAU at BTS site) Um.

i.e. at the MSC side of the BSC-to-MSC links (Figure 5.7). Channel coding/decoding and encryption are still performed in the TRX module of the BTS, whereas speech transcoding takes place in the BSC. For control purposes, the TRAU needs to receive synchronization and decoding information from the BTS, e.g. BFI for error concealment (section 4.7). If the TRAU does not reside in the BTS, it must be remotely controlled from the BTS by inband signaling. For this purpose, a subchannel of 16 kbit/s is reserved for the GSM speech signal on the BTS-to-BSC link, so an additional 3 kbit/s is made available for inband signaling. Alternatively, the GSM speech signal with added inband signaling could also be transmitted in a full ISDN B channel.

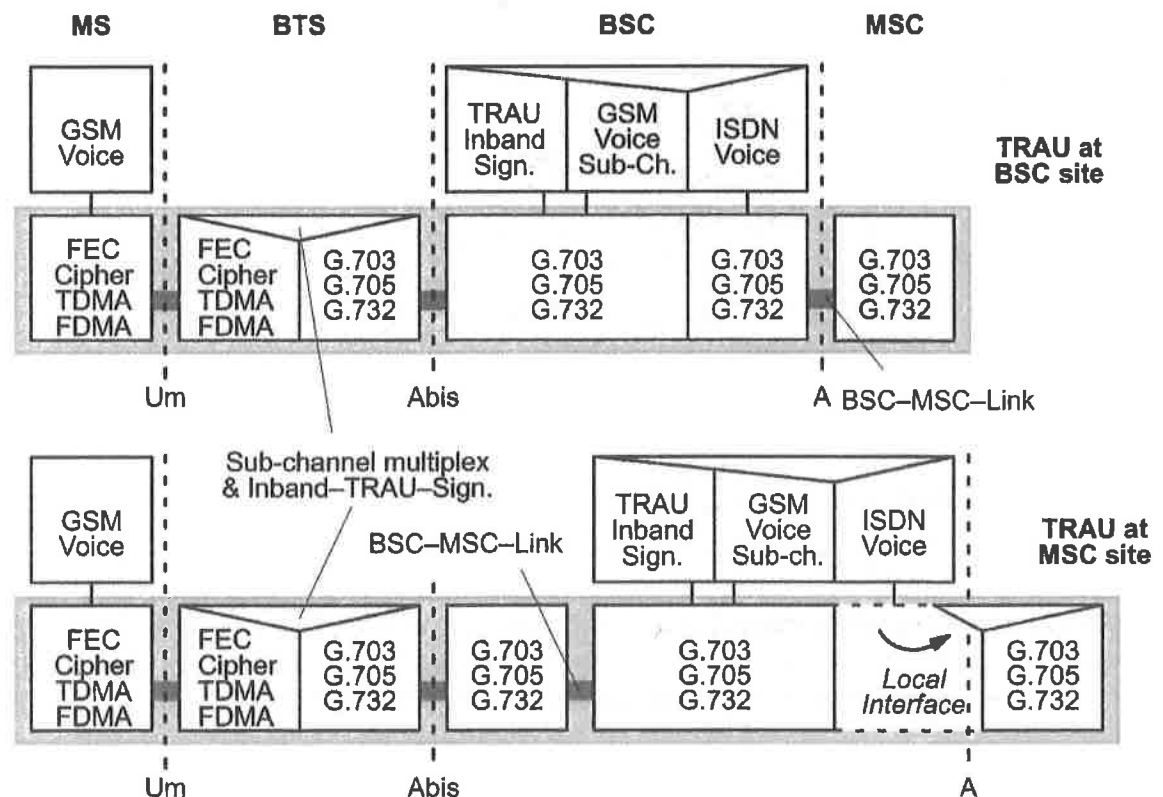


Figure 5.7 GSM protocol architecture for speech.

5.2.2 Transparent data transmission

The digital mobile radio channel is subject to severe quality variations and generates burst errors, which one tries to correct through interleaving and convolutional codes (section 4.8). However, if the signal quality is too low due to fading breaks or interference, the resulting errors cannot be corrected. For data transmission across the air interface U_m , a residual bit error ratio varying between 10^{-2} and 10^{-5} according to channel conditions can be observed (Vögel *et al.*, 1995). This kind of variable quality of data transmission at the air interface determines the service quality of transparent data transmission. Transparent data transmission defines a GSM connection type used for the realization of some basic bearer services. The pertinent protocol architecture is illustrated in Figure 5.8. The main aspect of the transparent connection type is that user data are protected against transmission errors by forward error correction only across the air interface. Further transmission within the GSM network to the next MSC with an IWF to an ISDN or a Public Switched Telephone Network (PSTN) occurs unprotected on digital line segments, which already have a very low bit error ratio in comparison with the radio channel. The transparent GSM data service offers a constant throughput rate and constant transmission delay; however, the residual error ratio varies with channel quality due to the limited correction capabilities of the forward error correction.

For example, take a data terminal communicating over a serial interface of type V.24. A transparent bearer service provides access to the GSM network directly at a MS or through

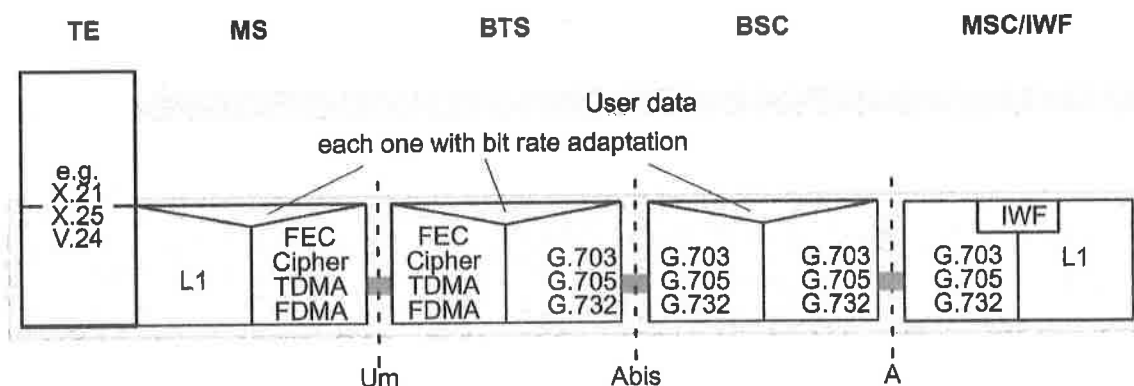


Figure 5.8 GSM protocol architecture for transparent data.

a terminal adapter (reference point R in Figure A.1). A data rate of up to 9.6 kbit/s can be offered based on the transmission capacity of the air interface and using an appropriate bit rate adaptation. The bit rate adaptation also performs the required asynchronous-to-synchronous conversion at the same time. This involves supplementing the tokens arriving asynchronously from the serial interface with fill data, since the channel coder requires a fixed block rate. In this way there is a digital synchronous circuit-switched connection between the terminal accessing the service and the IWF in the MSC, which extends across the air interface and the digital ISDN B channel inside the GSM network; this synchronous connection is completely transparent for the asynchronous user data of the Terminal Equipment (TE).

5.2.3 Nontransparent data transmission

Compared with the bit error ratio of the fixed network, which is of the order of 10^{-6} to 10^{-9} , the quality of the transparent data service is often insufficient for many applications, especially under adverse conditions. To provide more protection against transmission errors, more redundancy has to be added to the data stream. Since this redundancy is not always required, but only when there are residual errors in the data stream, forward error correction is inappropriate. Rather, an error-detection scheme with automatic retransmission of faulty blocks is used, Automatic Repeat Request (ARQ). Such an ARQ scheme which was specifically adapted to the GSM channel, is the RLP. The assumption for RLP is that the underlying forward error correction of the convolutional code realizes a channel with an average block error ratio of less than 10%, with a block corresponding to an RLP protocol frame of length 240 bits. Now the nontransparent channel experiences a constantly lower bit error ratio than the transparent channel, independent of the varying transmission quality of the radio channel; however, due to the RLP-ARQ procedure both throughput and transmission delay vary with the radio channel quality.

The data transmission between MS and interworking function of the next MSC is protected with the data link layer protocol RLP, i.e. the endpoints of RLP terminate in MS and IWF entities, respectively (Figure 5.9). At the interface to the data terminal TE, a Nontransparent Protocol (NTP) and an Interface Protocol (IFP) are defined, depending on the nature of the data terminal interface. Typically, a V.24 interface is used to carry character-oriented user data. These characters of the NTP are buffered and combined into blocks in the

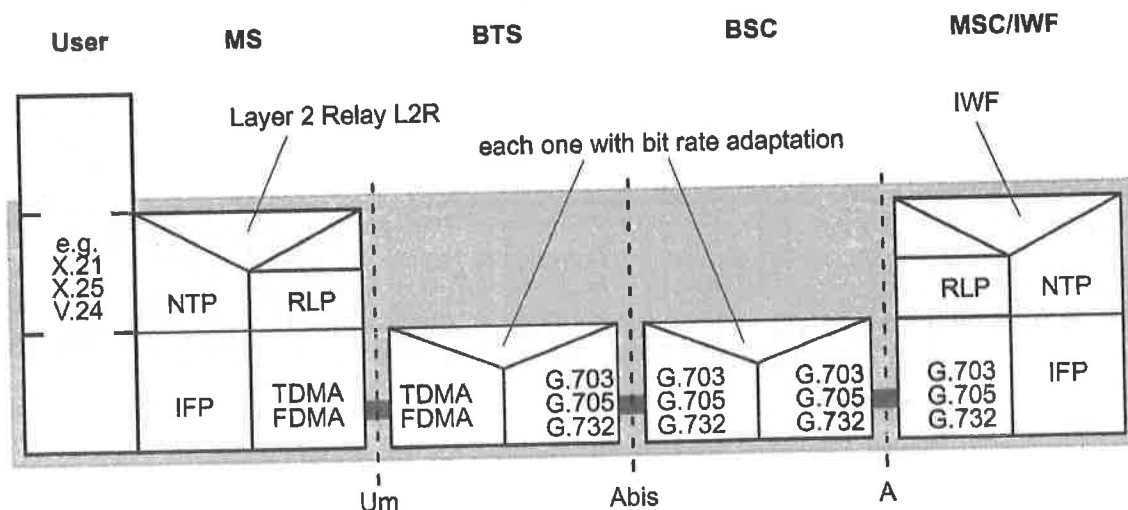


Figure 5.9 GSM protocol architecture for nontransparent data.

Layer 2 Relay (L2R) protocol, which transmits them as RLP frames. The data transport to and from the data terminal is flow-controlled. Therefore, transmission within the PLMN is no longer transparent for the data terminal. At the air interface, a new RLP frame is transmitted every 20 ms; thus L2R may have to insert fill tokens, if a frame cannot be completely filled at transmission time.

The RLP is very similar to the High-level Data Link Control (HDLC) protocol of ISDN (ISO/IEC 3309:1991) with regard to frame structure and protocol procedures, the main difference being the fixed frame length of 240 bits, in contrast to the variable length of HDLC. The frame consists of a 16-bit protocol header, 200-bit information field, and a 24-bit Frame Check Sequence (FCS); Figure 5.10. Owing to the fixed frame length, the RLP has no reserved flag pattern, and a special procedure to realize code transparency such as bit stuffing in HDLC is not needed. The very short – and hence less error prone – frames are exactly aligned with channel coding blocks. (The probability of frame errors increases with the length of the frame.)

RLP makes use of the services of the lower layers to transport its Protocol Data Units (PDUs). The channel offered to RLP therefore has the main characteristic of a 200 ms transmission delay, in addition to the possibly occurring residual bit errors. The delay is mostly caused by interleaving and channel coding, since the transmission itself takes only about 25 ms for a data rate of 9.6 kbit/s. This means it will take at least 400 ms until a positive acknowledgement is received for an RLP frame, and protocol parameters such as the transmission window and repeat timers need to be adjusted accordingly.

The RLP header is similar to that used in HDLC, with the difference that the RLP header contains no address information but only control information for which 16 bits are available. One distinguishes between supervisory frames and information frames. Whereas information frames carry user data, supervisory frames serve to control the connection (initialize, disconnect, reset) as well as the retransmission of information frames during data transfer. The information frames are labeled with a sequence number N(S) for identification, for which 6 bits are available in the RLP header (Figure 5.10). To conserve space, this

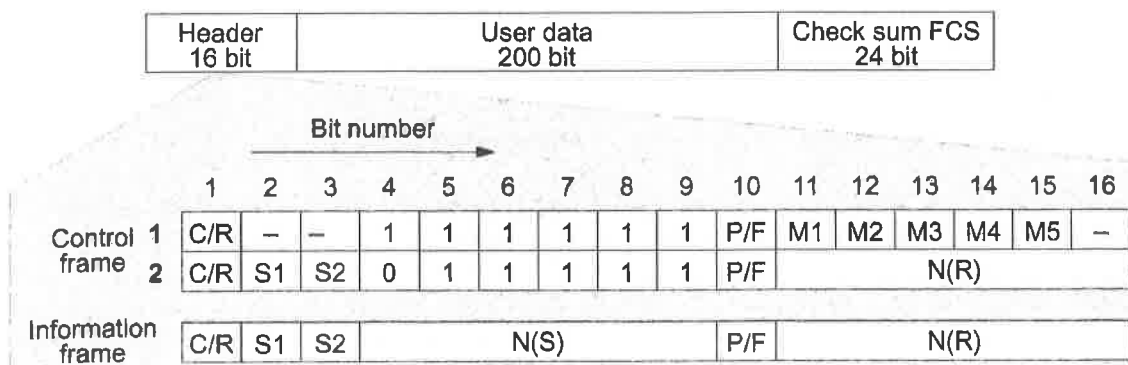


Figure 5.10 Frame structure of the RLP.

field is also used to code the frame type. Sequence number values smaller than 62 indicate that the frame carries user data in the information field (information frame). Otherwise the information field is discarded, and only the control information in the header is of interest (supervisory frame). These frames are marked with the reserved values 62 and 63 (Figure 5.10).

Owing to this header format, information frames can also carry (implicit) control information, a process known as piggybacking. The header information of the second variant can be carried completely within the header of an information frame. This illustrates further how RLP has been adapted to the radio channel, since it makes the transmission of additional control frames unnecessary during information transfer, which reduces the protocol overhead and increases the throughput.

Thus, the send sequence number is calculated modulo 62, which amounts to a window of 61 frames, allowing 61 outstanding frames without acknowledgement before the sender has to receive the acknowledgement of the first frame. Positive acknowledgement is used, i.e. the receiver sends an explicit supervisory frame as a receipt or an implicit receipt within an information frame. Such an acknowledgement frame contains a receive frame number $N(R)$ which designates correct reception of all frames, including send sequence number $N(S) = N(R) - 1$.

Each time the last information frame is sent, a timer $T1$ is started at the sender. If an acknowledgement for some or all sent frames is not received in time, perhaps because the acknowledging RLP frame had errors and was therefore discarded, the timer expires and causes the sender to request an explicit acknowledgement. Such a request may be repeated $N2$ times; if this still leads to no acknowledgement, the connection is terminated. If an acknowledgement $N(R)$ is obtained after expiration of timer $T1$, all sent frames starting from and including $N(R)$ are retransmitted. In the case of an explicitly requested acknowledgement, this corresponds to a modified Go-back- N procedure. Such a retransmission is also allowed only up to $N2$ times. If no receipt can be obtained even after $N2$ trials, the RLP connection is reset or terminated.

Two procedures are provided in RLP for dealing with faulty frames: selective reject, which selects a single information frame without acknowledgement; and reject, which causes retransmission with implicit acknowledgement. With selective reject, the receiving RLP entity requests retransmission of a faulty frame with sequence number $N(R)$, but this does

not acknowledge receipt of other frames. Each RLP implementation must at least include the reject method for requesting retransmission of faulty frames. With a reject, the receiver asks for retransmission of all frames starting with the first defective received frame with number $N(R)$ (Go-back-N). Simultaneously, this implicitly acknowledges correct reception of all frames up to and including $N(R) - 1$. Realization of selective reject is not mandatory in RLP implementations, but it is recommended. This is because Go-back-N causes retransmission of frames that may have been transmitted correctly and thus deteriorates the throughput that could be achieved with selective reject.

5.3 Protocol architecture of the signaling plane

5.3.1 Overview of the signaling architecture

Figure 5.11 shows the essential protocol entities of the GSM signaling architecture (control plane or signaling plane). Three connection elements are distinguished: the radio-interface connection element, the BSS-interface connection element and the A-interface connection element. This control plane protocol architecture consists of a GSM-specific part with the interfaces Um and Abis and a part based on SS#7 with the interfaces A, B, C, E (Figure 5.11). This change of signaling system corresponds to the change from radio interface connection element to A-interface connection element as discussed above for the user data plane (Figure 5.3).

The radio interface Um is defined between MS and BSS, more exactly between MS and BTS. Within the BSS, the BTS and the BSC cooperate over the Abis interface, whereas the A interface is located between BSC and MSC. The MSC has also signaling interfaces to VLR (B), HLR (C), to other MSCs (E) and to the EIR (F). Further signaling interfaces are defined between VLRs (G) and between VLR and HLR (D). Figure 3.3 gives an overview of the interfaces in a GSM PLMN.

Physical layer

In the control plane, the lowest layer of the protocol model at the air interface, the physical layer, implements the logical signaling channels (TDMA/FDMA, multiframe, channel coding, etc.; sections 4.7, 4.8 and 5.6). Like user data, signaling messages are transported over the Abis interface (BTS–BSC) and the A-interface (BSC–MSC) on digital lines with data rates of 2048 kbit/s (1544 kbit/s in the USA) or 64 kbit/s (ITU-T G.703, G.705, G.732).

Layer 2: LAPDm

On Layer 2 of the logical signaling channels across the air interface a data link protocol entity is implemented, the Link Access Procedure on Dm Channels (LAPDm). LAPDm is a derivative of Link Access Protocol Channel D (LAPD) which is specifically adapted to the air interface. This data link protocol is responsible for the protected transfer of signaling messages between MS and BTS over the air interface, i.e. LAPDm is terminated in MS and base station.

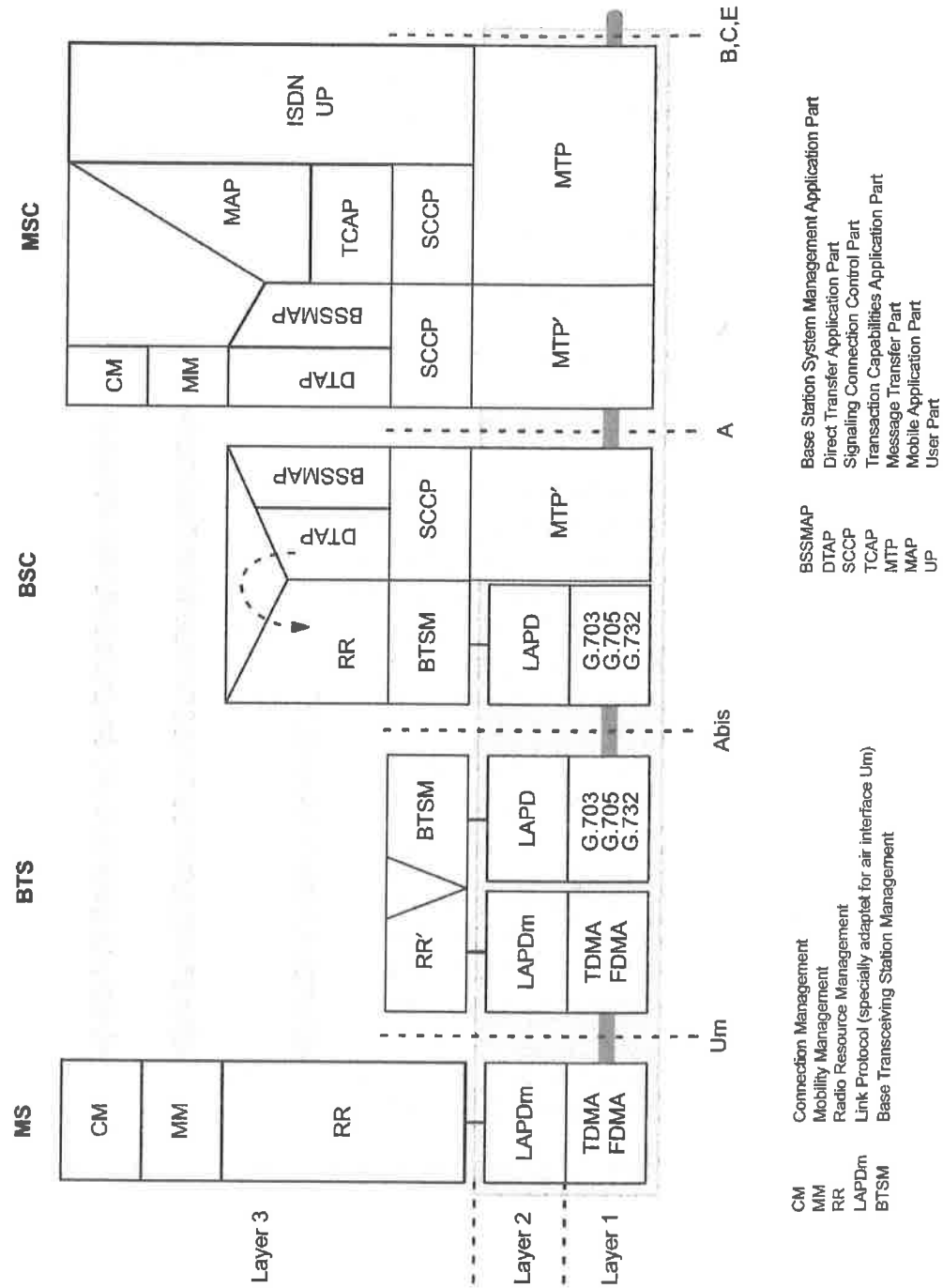


Figure 5.11 GSM protocol architecture for signaling.

In essence, LAPDm is a protocol similar to HDLC which offers a number of services on the various logical Dm channels of Layer 3: connection setup and teardown protected signaling data transfer. It is based on various link protocols used in fixed networks, such as LAPD in ISDN (Bocker, 1990). The main task of LAPDm is the transparent transport of messages between protocol entities of Layer 3 with special support for:

- multiple entities in Layer 3 and Layer 2;
- signaling for broadcasting (BCCH);
- signaling for paging (PCH);
- signaling for channel assignment (AGCH);
- signaling on dedicated channels (SDCCH).

A detailed discussion of LAPDm is presented in section 5.4.2.

Layer 3

In the MS, the LAPDm services are used at Layer 3 of the signaling protocol architecture. Layer 3 is divided into three sublayers: Radio Resource Management (RR), Mobility Management (MM) and Connection Management (CM). The protocol architecture formed by these three sublayers is shown in Figure 5.12. CM is further subdivided into three protocol entities: Call Control (CC), Supplementary Services (SS) and SMS. Additional multiplexing functions within Layer 3 are required between these sublayers.

The call-independent SS and the SMS are offered to higher layers at two Service Access Points (SAPs), MNSS and MNSMS. A more detailed look at the services offered by the RR, MM and CC protocol entities is given in the following.

Radio resource management

RR essentially handles the administration of the frequencies and channels. This involves the RR module of the MS communicating with the RR module of the BSC (Figure 5.11). The general objective of the RR is to set up, maintain and take down RR connections which enable point-to-point communication between MS and network. This also includes cell selection in idle mode and handover procedures. Furthermore, the RR is responsible for monitoring BCCH and CCCH on the downlink when no RR connections are active.

The following functions are realized in the RR module:

- monitoring of BCCH and PCH (readout of system information and paging messages);
- RACH administration: MSs send their requests for connections and replies to paging announcements to the BSS;
- requests for and assignments of data and signaling channels;
- periodic measurement of channel quality (quality monitoring);
- transmitter power control and synchronization of the MS;

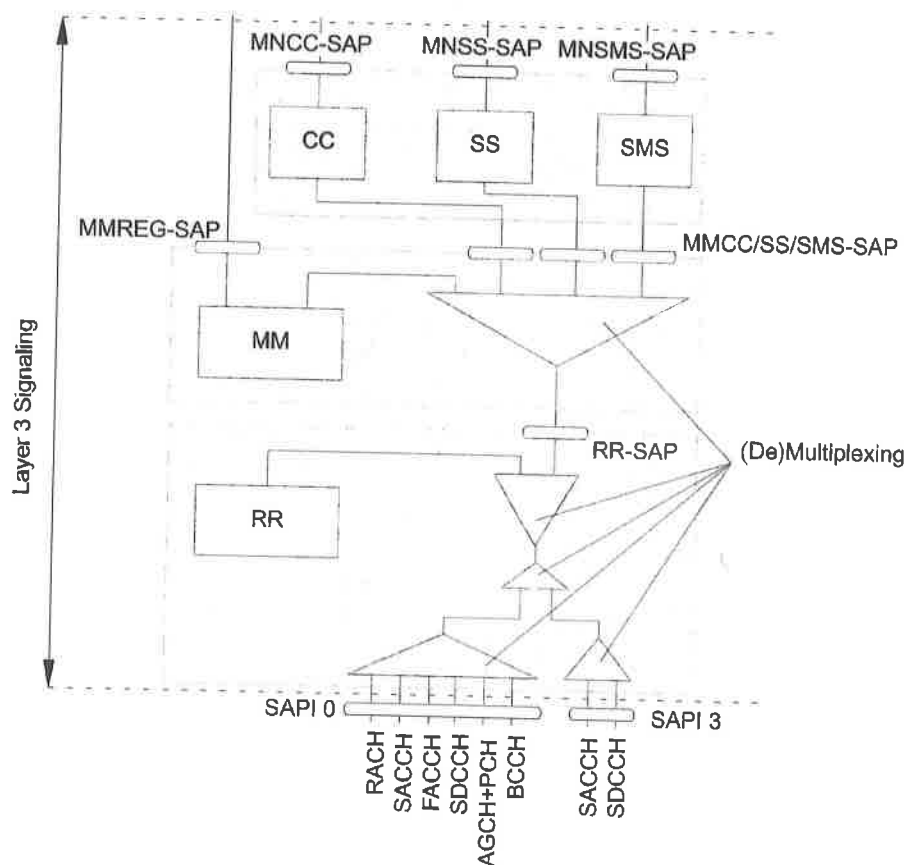


Figure 5.12 Layer 3 protocol architecture at the MS side.

- handover (part of which is sometimes erroneously attributed to roaming functions and MM), always initiated by the network;
- synchronization of encryption and decryption on the data channel.

The RR sublayer provides several services at the RR-SAP to the MM sublayer. These services are needed to set up and take down signaling connections and to transmit signaling messages.

Mobility management

MM encompasses all of the tasks resulting from mobility. The MM activities are exclusively performed in cooperation between MS and MSC, and they include:

- TMSI assignment;
- localization of the MS;
- location updating of the MS, sometimes known as roaming functions;
- identification of the MS (IMSI, IMEI);

Table 5.1 Distribution of functions between BTS, BSC and MSC (according to GSM Rec. 08.02, 08.52).

	BTS	BSC	MSC
<i>Terrestrial channel management</i>			
MSC-BSC-channels			X
Channel allocation		X	
Blocking indication			
BSC-BTS-Channels		X	
Channel allocation			
Blocking indication	X		
<i>Mobility management</i>			
Authentication			X
Location updating			X
<i>Call control</i>			
<i>Radio channel management</i>			
Channel coding/decoding	X		
Transcoding/rate adaptation	X		
Interworking function			X
Measurements			
Uplink measuring	X		X
Processing of reports from MS/TRX	X	X	X
Traffic measurements			X
Handover			
BSC internal, intracell		X	
BSC internal, intercell		X	
BSC external		X	
Recognition, decision, execution			X
HO access detection	X		
Paging			
Initiation		X	
Execution	X		
Channel configuration management		X	
Frequency hopping			
Management		X	
Execution	X		
TCH management			
Channel allocation		X	
Link supervision		X	
Channel release		X	X
Idle channel observation	X		
Power control determination	X	X	

Table 5.1 Continued.

	BTS	BSC	MSC
SDCCH management			
SDCCH allocation		X	
Link supervision		X	
Channel release		X	X
Power control determination	X	X	
BCCH/CCCH management			
Message scheduling management		X	
Message scheduling execution	X		
Random access detection	X		
Immediate assign		X	
Timing Advance			
Calculation	X		
Signaling to MS at random access		X	
Signaling to MS at handover/during call	X		
Radio resource indication			
Report status of idle channels	X		
LAPDm functions	X		
Encryption			
Management		X	
Execution	X		

- authentication of the MS;
- IMSI attach and detach procedures (e.g. at insertion or removal of SIM);
- ensuring confidentiality of subscriber identity.

Registration services for higher layers are provided by Layer 3 at the MMREG-SAP (Figure 5.12). Registration involves the IMSI attach and detach procedures which are used by the mobile to report state changes such as power-up or power-down, or SIM card removal or insertion.

The MM sublayer offers its services at the MMCC-SAP, MMSS-SAP and MMSMS-SAP to the CC, SS and SMS entities. This is essentially a connection to the network side over which these units can communicate.

Connection management

CM consists of three entities: CC, SS and SMS. CC handles all tasks related to setting up, maintaining and taking down calls. The services of CC are provided at the MNCC-SAP, and they encompass:

- establishment of normal calls (MS-originating and MS-terminating);
- establishment of emergency calls (only MS-originating);

- termination of calls;
- Dual-Tone Multifrequency (DTMF) signaling;
- call-related supplementary services;
- in-call modification: the service may be changed during a connection (e.g. speech and transparent/nontransparent data are alternating; or speech and fax alternate).

The service primitives at this SAP of the interface to higher layers report reception of incoming messages and effect the sending of messages, essentially ISDN user-network signaling according to Q.931.

RR messages are mainly exchanged between MS and BSS. In contrast, CM and MM functions are handled exclusively between MS and MSC; the exact division of labor between BTS, BSC and MSC is summarized in Table 5.1. As can be seen, RR messages have to be transported over the Um and Abis interfaces, whereas CM and MM messages need additional transport mechanisms across the A interface.

Message transfer part

From a conceptual viewpoint, the A interface in GSM networks is the interface between the MSCs, the ISDN exchanges with mobile network specific extensions, and the BSC, the dedicated mobile network specific control units. Here too is the reference point, where the signaling system changes from GSM-specific to the general ISDN-compatible SS#7. Message transport in the SS#7 network is realized through the Message Transfer Part (MTP). In essence, MTP comprises the lower three layers of the OSI Reference Model, i.e. the MTP provides routing and transport of signaling messages.

A slightly modified (reduced) version of the MTP, called MTP', has been defined for the protected transport of signaling messages across the A interface between BSC and MSC. At the ISDN side of the MSCs, the complete MTP is available. For signaling transactions between MSC and MS (CM, MM), it is necessary to establish and identify distinct logical connections. The Signaling Connection Control Part (SCCP) is used for this purpose to facilitate implementation with a slightly reduced range of functions defined in SS#7.

BSS application part

For GSM-specific signaling between MSC and BSC, the Base Station System Application Part (BSSAP) has been defined. The BSSAP consists of the Direct Transfer Application Part (DTAP) and the Base Station System Management Application Part (BSSMAP). The DTAP is used to transport messages between MSC and MS. These are the CC and MM messages. At the A interface, they are transmitted with DTAP and then passed transparently through the BSS across the Abis interface to the MS without interpretation by the BTS.

The BSSMAP is the protocol definition part which is responsible for all of the administration and control of the radio resources of the BSS. RR is one of the main functions of a BSS. Therefore, the RR entities terminate in the MS and the BTS or BSC, respectively. Some functions of RR, however, require involvement of the MSC (e.g. some handover situations, or release of connections or channels). Such actions should be initiated and controlled by the MSC (e.g. handover and channel assignment). This control is the responsibility of BSSMAP.

RR messages are mapped and converted within the BSC into procedures and messages of BSSMAP and vice versa. BSSMAP offers the functions which are required at the A interface between BSS and MSC for RR of the BSS. Accordingly, RR messages initiate BSSMAP functions, and BSSMAP functions control RR protocol functions.

BTS management

A similar situation exists at the Abis interface. Most of the RR messages are passed transparently by the BTS between MS and BSC. Certain RR information, however, must be interpreted by the BTS, e.g. in situations such as random access of the MS, the start of the ciphering process, or paging to localize a MS for connection setup. The Base Transceiver Station Management (BTSM) contains functions for the treatment of these messages and other procedures for BTS management. In addition a mapping occurs in the BTS from BTSM onto the RR messages relevant at the air interface (RR' , Figure 5.11).

Mobile application part

The MSC is equipped with the MAP, a mobile network specific extension of SS#7, for communication with the other components of the GSM network (the HLR and VLR registers, other MSCs) and other PLMNs. Among the MAP functions are all signaling functions among MSCs as well as between MSC and the registers (Figure 5.13). These functions include:

- updating of residence information in the VLR;
- cancellation of residence information in the VLR;
- storage of routing information in the HLR;
- updating and supplementing of user profiles in HLR and VLR;
- inquiry of routing information from the HLR;
- handover of connections between MSCs.

The exchange of MAP messages, e.g., with other MSCs, HLR or VLR, occurs over the transport and transaction protocol of the SS#7. The SS#7 transaction protocol is the Transaction Capabilities Application Part (TCAP). A connectionless transport service is offered by the SCCP.

The MAP functions require channels for signaling between different PLMNs which are provided by the international SS#7. Access to SS#7 occurs through the fixed ISDN.

Connection to the fixed network is typically done through leased lines; in the case of the German GSM network operators, it is through lines with a rate of 2 Mbit/s from Deutsche Telekom (Gottschalk, 1993). Often the majority of the MSC in a PLMN has such an access to the fixed network. On these lines, both user data and signaling data are transported. From the viewpoint of a fixed network, an MSC is integrated into the network in the same way as a normal ISDN exchange node. Outside of a PLMN, starting with the GMSC, calls for MSs are treated like calls for subscribers of the fixed network, i.e. the mobility of a subscriber with an MSISDN becomes 'visible' only beyond the GMSC. For CC, the MSC has the same interface as an ISDN switching node. Connection-oriented signaling of GSM networks is mapped at

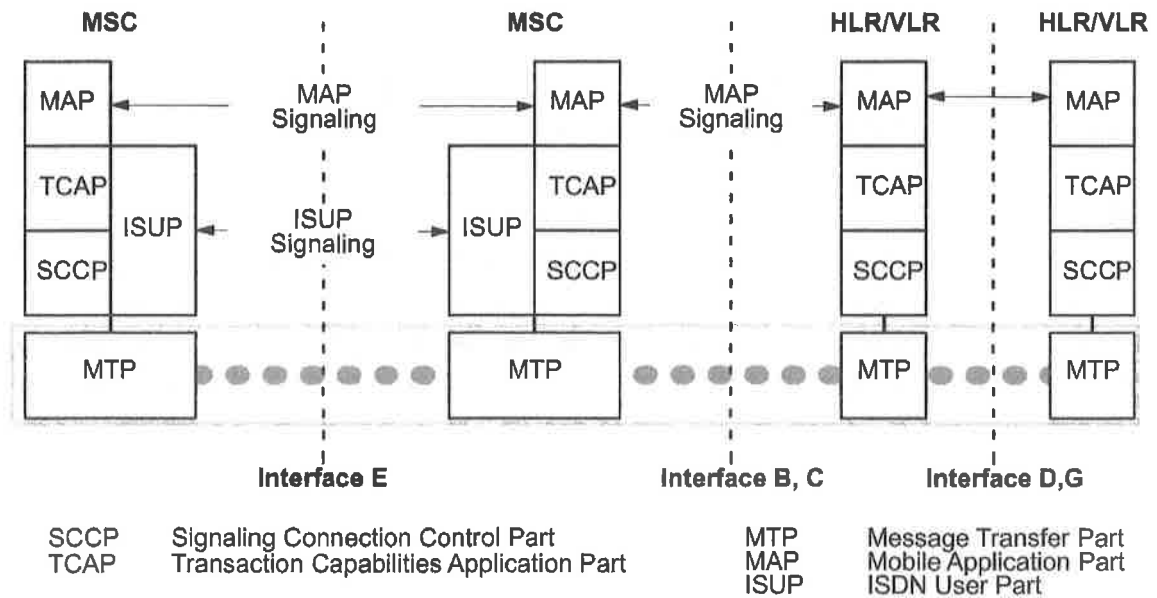


Figure 5.13 Protocol interfaces in the mobile network.

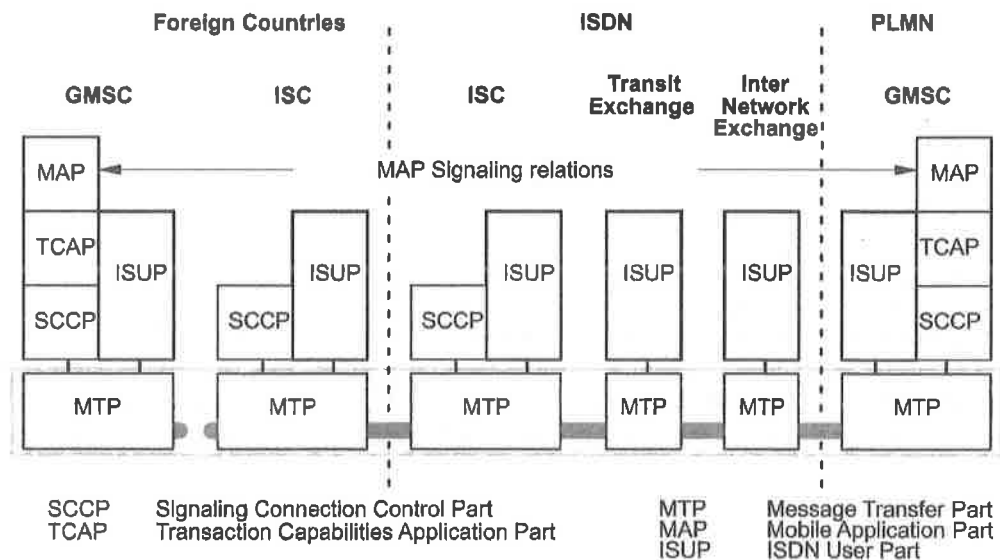


Figure 5.14 International signaling relations via ISDN (Gottschalk, 1993).

the fixed network side (interface to ISDN) into the ISDN User Part (ISUP) used to connect ISDN channels through the network (Figure 5.14). The mobile-specific signaling of the MAP is routed over a gateway of the PLMN (GMSC) and the ISC of the national ISDN network into the international SS#7 network (Gottschalk, 1993). In this way, transport of signaling data between different GSM networks is also guaranteed without problems.

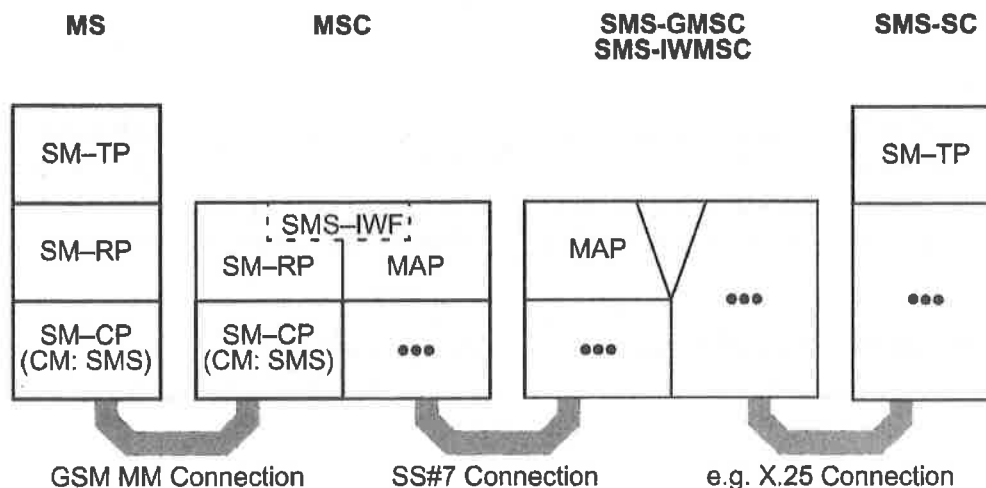


Figure 5.15 Protocol architecture for SMS transfer.

5.3.2 Transport of user data in the signaling plane

In the signaling plane (control plane) of the GSM architecture, one can also transport packet-oriented user data from or to MSs. This occurs for the point-to-point SMS. Short messages are always transmitted in store-and-forward mode through a Short Message Service Center (SMS-SC). The service center accepts these messages, which can be up to 160 characters long, and forwards them to the recipients (other MSs or fax, email, etc.). In principle, GSM defines a separate protocol architecture for the realization of this service.

Between MS and service center, short messages are transmitted using a connectionless transport protocol: Short Message Transport Protocol (SM-TP) which uses the services of the signaling protocols within the GSM network. Transport of these messages outside of the GSM network is not defined. For example, the SMS-SC could be directly connected to the gateway switching center (SMS-GMSC), or it could be connected to a Short Message Service Interworking MSC (SMS-IW MSC) through an X.25 connection (Figure 5.15). Within the GSM network between MSCs, a short message is transferred with the MAP and the lower layers of SS#7. Finally, between a MS and its local MSC, two protocol layers are responsible for the transfer of transport protocol units of SMS. First, there is the SMS entity in the CM sublayer of Layer 3 at the user-network interface (see Figure 5.12) which realizes the Short Message Control Protocol (SM-CP) and its connection-oriented service. Second, there is the relay layer, in which the Short Message Relay Protocol (SM-RP) is defined, which offers a connectionless service for transfer of SMS transport PDUs between MS and MSC. This, however, uses services at the service access point MMSMS-SAP (see Figure 5.12) and thus a connection of the MM sublayer.

In addition to the SM-CP, the relay protocol SM-RP was introduced above the CM sublayer (Figures 5.12 and 5.15) to realize an acknowledged transmission of short messages, but with minimal overhead for the radio channel. A short message sent by a MS is passed over the signaling network until it reaches the service center SMS-SC. If the service center determines the error-free reception of a message, an acknowledgement message is returned on the reverse path, which finally causes sending of an acknowledgement message from

the SM-RP entity in the MSC to the MS. Until this acknowledgement message arrives, the connection in the MM sublayer can be taken down, and thus also the reserved radio channel. In this way, radio resources across the air interface are only occupied during the actual transmission of SM-RP messages. Each successful transmission of an SM-CP PDU across the MM connection, which includes the error-prone air interface, is immediately acknowledged, or else errors are immediately reported to the sending SM-CP entity. So if a message is damaged at the radio interface, this avoids it being transmitted to the service center.

5.4 Signaling at the air interface (Um)

Signaling at the user-network interface in GSM is essentially concentrated in Layer 3. Layers 1 and 2 provide the mechanisms for the protected transmission of signaling messages across the air interface. In addition to the local interface, they contain functionality and procedures for the interface to the BTS. The signaling of Layer 3 at the user-network interface is very complex and comprises protocol entities in the MS and in all functional entities of the GSM network (BTS, BSC and MSC).

5.4.1 Layer 1 of the MS-BTS interface

Layer 1 of the OSI Reference Model (physical layer) contains all of the functions necessary for the transmission of bit streams over the physical medium, in this case the radio channel. GSM Layer 1 defines a series of logical channels based on the channel access procedures with their physical channels. The higher layer protocols access these services at the Layer 1 service interface. The three interfaces of Layer 1 are schematically illustrated in Figure 5.16.

LAPDm protocol frames are transmitted across the service mechanisms of the data link layer interface, and the establishment of logical channels is reported to Layer 2. The communication across this interface is defined by abstract physical layer service primitives. A separate SAP is defined for each logical control channel (BCCH, PCH + AGCH, RACH, SDCCH, SACCH, FACCH).

There is a direct interface between Layer 1 and the RR sublayer of Layer 3. The abstract service primitives exchanged at this interface mostly concern channel assignment and Layer 1 system information, including measurement results of channel monitoring. At the third Layer 1 interface, the traffic channels for user (payload) data are provided.

The SAPs of Layer 1 as defined in GSM are not genuine SAPs in the spirit of OSI. They differ from the PHY-SAPs of the OSI Reference Model insofar as these SAPs are controlled by Layer 3 RR sublayer (layer management, establishment and release of channels) rather than by control procedures in the link layer. Control of Layer 1 SAPs by RR comprises activation and deactivation, configuration, routing and disconnection of physical and logical channels. Furthermore, exchange of measurement and control information for channel monitoring occurs through service primitives.

Layer 1 services

Layer 1 services of the GSM user-network interface are divided into three groups:

- access capabilities;

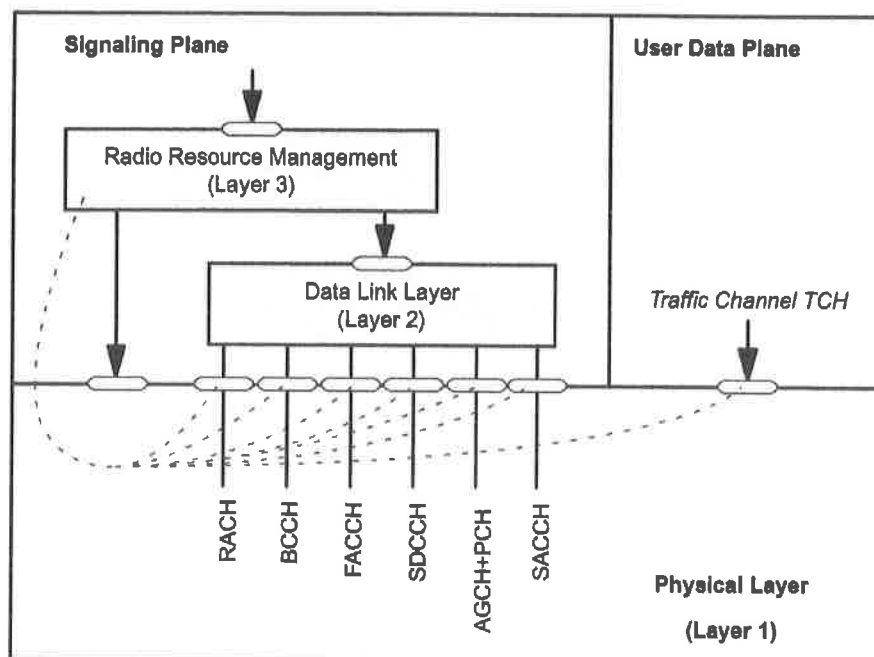


Figure 5.16 Layer 1 service interfaces.

- error detection;
- encryption.

Layer 1 provides a bit transport service for the logical channels. These are transmitted in multiplexed format over physical channels which consist of elements defined for the transmission on the radio channel (frequency, time slot, hopping sequence, etc.; section 5.1). Some physical channels are provided for common (shared) use (BCCH and CCCH), whereas others are assigned to dedicated connections with single MSs (dedicated physical channels). The combination of logical channels used on a physical channel can vary over time, e.g. TCH + SACCH/FACCH replaced by SDCCH + SACCH (see Table 4.4).

The GSM standard distinguishes explicitly between access capabilities for dedicated physical channels and for common physical channels BCCH/CCCHs. Dedicated physical channels are established and controlled by Layer 3 RR management. During the operation of a dedicated physical channel, Layer 1 continuously measures the signal quality of the used channel and the quality of the BCCH channels of the neighboring base stations. This measurement information is passed to Layer 3 in measurement service primitives Management Physical Headers (MPH). In idle mode, Layer 1 selects the cell with the best signal quality in cooperation with the RR sublayer based on the quality of the BCCH/CCCH (cell selection).

GSM Layer 1 offers an error-protected bit transport service and therefore also error detection and correction mechanisms. To do this, error-correcting and error-detecting coding mechanisms are provided (section 4.8). Frames recognized as faulty are not passed to Layer 2. Furthermore, security-relevant functions such as encryption of user data are implemented in Layer 1 (section 5.6).

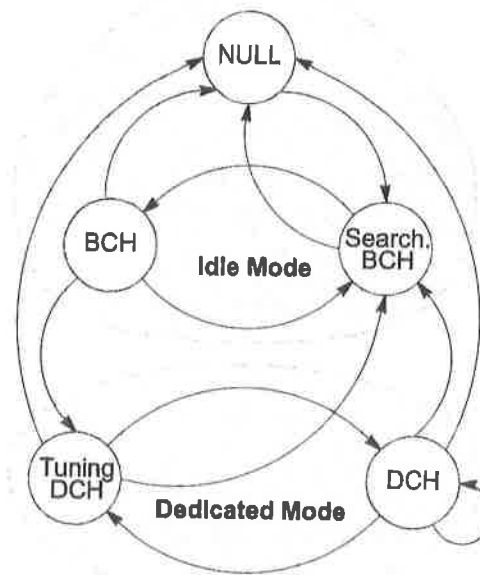


Figure 5.17 State diagram of the physical layer of MS.

Layer 1: Procedures and peer-to-peer signaling

GSM defines and distinguishes between two operational modes of a MS: idle mode and dedicated mode (Figure 5.17). In idle mode, the MS is either powered off (state NULL) or it searches for or measures the BCCH with the best signal quality (state SEARCHING BCH), or is synchronized to a specific base station's BCCH and ready to perform a random access procedure on the RACH for requesting a dedicated channel in state BCH (section 4.5.4).

In state TUNING DCH of the dedicated mode, the MS occupies a physical channel and tries to synchronize with it, which will eventually result in transition to state DCH. In this state, the MS is finally ready to establish logical channels and switch them through. The state transitions of Layer 1 are controlled by MPH service primitives of the RR interface, i.e. directly from the Layer 3 RR sublayer of the signaling protocol stack.

Layer 1 defines its own frame structure for the transport of signaling messages, which occur as LAPDm frames at the SAP of the respective logical channel. Figure 5.18 shows the format of an SACCH block as an example, which essentially contains 21 octets of LAPDm data.

Furthermore, the SACCH frame contains a kind of protocol header which carries the current power level and the value of the timing advance. This header is omitted in the other logical channels (FACCH, SDCCH, CCCH and BCCH) which contain only LAPDm PDUs.

5.4.2 Layer 2 signaling

The LAPDm protocol is the data link protocol for signaling channels at the air interface. It is similar to HDLC. It provides two operational modes:

- unacknowledged operation;
- acknowledged operation.

Table 5.2 Logical channels, operational modes and Layer 2 SAPIs.

Logical channel	SAPI = 0	SAPI = 3
BCCH	Unacknowledged	–
CCCH	Unacknowledged	–
SDCCH	Unacknowledged and acknowledged	Unacknowledged and acknowledged
SACCH associated with SDCCH	Unacknowledged	–
SACCH associated with TCH	Unacknowledged	Unacknowledged and acknowledged
FACCH	Unacknowledged and acknowledged	–

An LAPDm entity is established for each of the pertinent physical/logical channels. For some of the channel/SAPI combinations only a subset of the LAPDm protocol is needed (e.g. unacknowledged operation), and some channel/SAPI combinations are not supported (Table 5.2). These LAPDm entities perform the data link procedure, i.e. the functions of the L2 peer-to-peer communication as well as the service primitives between adjacent layers. Segmentation and reassembly of Layer 3 messages is also included.

Further Layer 2 procedures are the distribution procedure and the Random Access (RA) procedure. The distribution procedure is needed if multiple SAPs are associated with one physical/logical channel. It performs the distribution of the L2 frames received on one channel to the respective data link procedure or the priority-controlled multiplexing of L2 frames from multiple SAPs onto one channel. The random access procedure is used on the RACH; it deals with the random controlled retransmission of random access bursts, but it does not perform any error protection on the unidirectional RACH.

For certain aspects of RR, the protocol logic of Layer 3 has to have direct access to the services of Layer 1. In particular, this is needed for the functions of radio subsystem link control, i.e. for channel measurement, transmitter power control and timing advance.

A possible link layer configuration of a MS is shown in Figure 5.19. The base station has a similar configuration with one PCH + AGCH, SDCCH and SACCH/FACCH for each active MS.

Figure 5.20 shows the different types of protocol data frames used for communication between L2 peer entities in MS and BTS. Frame formats A and B are used on the SACCH, FACCH and SDCCH channels, depending upon whether the frame has an information field (type B) or not (type A). For unacknowledged operation (BCCH, PCH, AGCH), format types Abis and Bbis are used on channels with SAPI = 0. The Abis format is used when there is no information to be transmitted on the respective logical channel.

In contrast to HDLC, LAPDm frames have no flag to designate the beginning and end of a frame, rather the delineation of frames at the link level is performed as in RLP (section 5.2.3) through the fixed-length block structure of Layer 1. The maximum number of octets N201 per information field depends on the type of logical channel (Table 5.3). The end of the information field is given by a length indicator, a value of less than N201 indicates that the frame has to be supplemented with fill bits to the full length. In the case of an SACCH

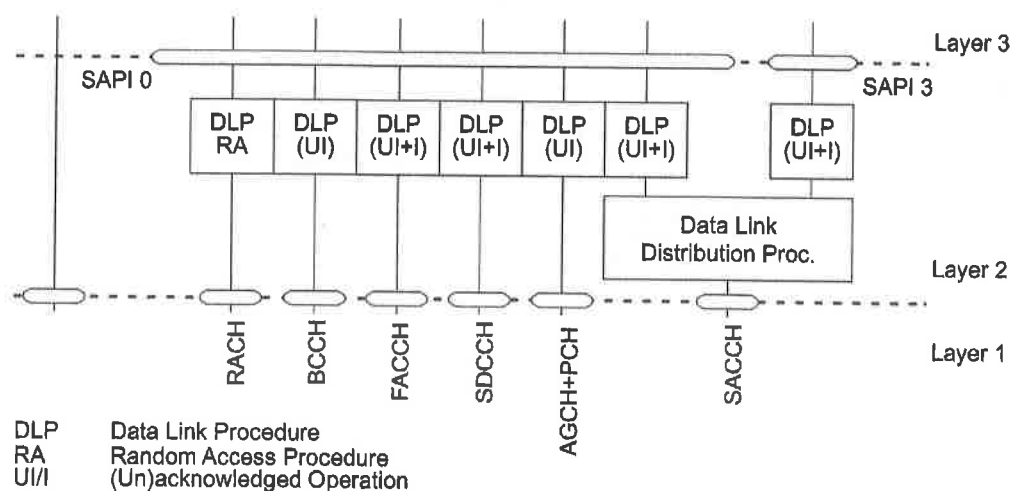


Figure 5.19 Sample configuration of the MS data link layer.

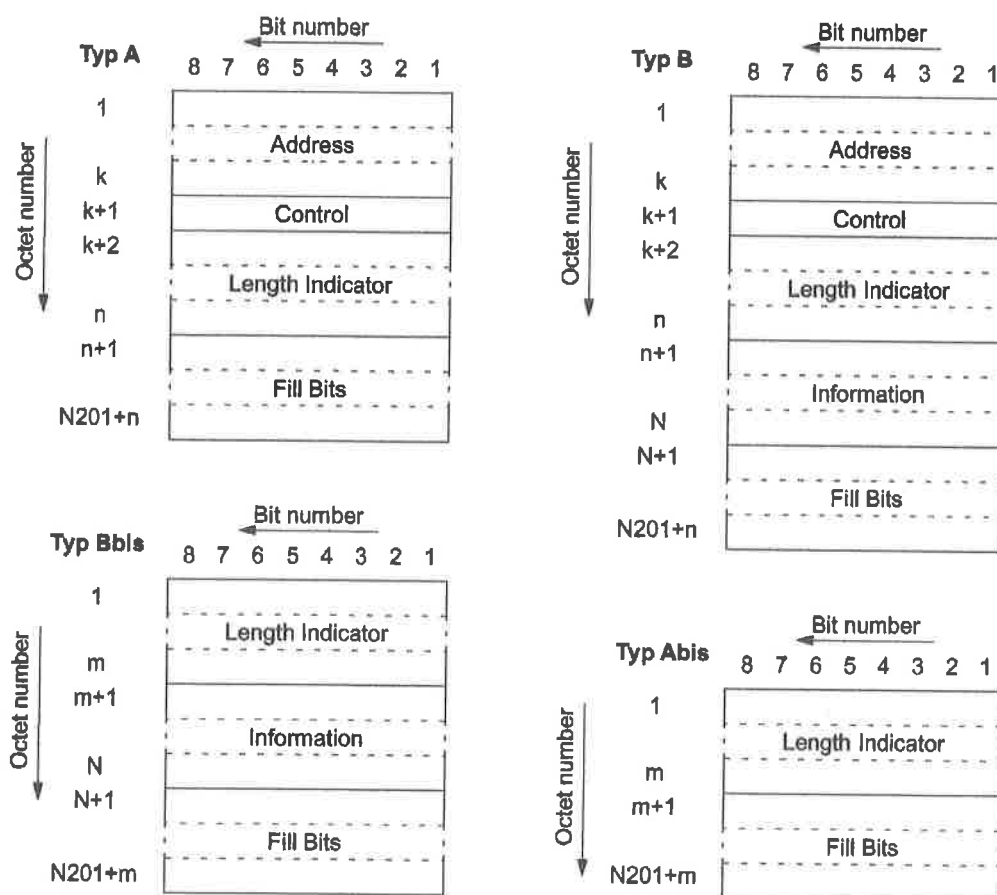


Figure 5.20 LAPDm frame formats.

Table 5.3 Logical channels and the maximum length of the LAPDm information field.

Logical channel	N201
SACCH	18 octets
SDCCH, FACCH	20 octets
BCCH, AGCH, PCH	22 octets

channel, for example, this yields a fixed-length LAPDm packet of 21 octets. Combined with the fields for transmitter power control and timing advance, an SACCH block of Layer 1 is thus 23 octets long.

The address field may have a variable length, however; for use on control channels it consists of exactly one octet. In addition to other fields, this octet contains a SAPI (3 bits) and the Command/Response (C/R) flag known from HDLC. In LAPDm, the coding of the control field with sending and receiving sequence numbers and the state diagram describing the protocol procedures are almost identical to HDLC (ISO/IEC 3309:1991, 2008). Some additional parameters are required at the service interface to Layer 3; for example, a parameter CEP designating the desired logical channel. Furthermore, the LAPDm protocol has some simplifications or peculiarities with regards to HDLC.

- The sending window size is restricted to $k = 1$.
- The protocol entities should be implemented in such a way that the state RECEIVER BUSY is never reached. Thus, RNR packets can be safely ignored. The HDLC polling procedure for state inquiry of the partner station need not be implemented in LAPDm.
- Connections to SAPI = 0 are always initiated by the MS.

In addition, the repetition timer T200 and the maximum number of allowed repetitions N200 have been adapted to the special needs of the mobile channel. In particular, they have their own value determined by the type of logical channel.

5.4.3 Radio resource management

The procedures for RR are the basic signaling and control procedures at the air interface. They handle the assignment, allocation and administration of radio resources, the acquisition of system information from broadcast channels (BCCH) and the selection of the cell with the best signal reception (see cell selection in section 4.5.4). Accordingly, the RR procedures and pertinent messages (Table 5.4) are defined for idle mode as well as for setting up, maintaining and taking down of RR connections.

Figure 5.21 shows the format of RR messages, which is uniform for all three Layer 3 signaling sublayers (CM, MM, RR). Each Layer 3 message contains a protocol discriminator in the first octet, which allows association of messages with the respective sublayer or service access point (Figure 5.12). The uppermost four bits of the first octet also contain a transaction ID, which enables a MS to perform several signaling transactions in parallel.

Table 5.4 RR messages.

Category	Message	Logical channel	Direction	MT code
Channel establishment	Additional assignment	DCCH	N → MS	00111011
	Immediate assignment	CCCH	N → MS	00111111
	Immediate assignment extended	CCCH	N → MS	00111001
	Immediate assignment rejected	CCCH	N → MS	00111010
Ciphering	Ciphering mode command	DCCH	N → MS	00110101
	Ciphering mode complete	DCCH	MS → N	00110010
Handover	Assignment command	DCCH	N → MS	00101110
	Assignment complete	DCCH	MS → N	00101001
	Assignment failure	DCCH	MS → N	00101111
	Handover access	DCCH	MS → N	—
	Handover command	DCCH	N → MS	00101011
	Handover complete	DCCH	MS → N	00101100
	Handover failure	DCCH	MS → N	00101000
	Physical information	DCCH	N → MS	00101101
	Channel release	DCCH	N → MS	00001101
	Partial release	DCCH	N → MS	00001010
	Partial release complete	DCCH	MS → N	00001111
	Paging request, type 1/2/3	PCH	N → MS	00100xxx
	Paging response	DCCH	MS → N	00100111
	System information, type 1/2/3/4	BCCH	N → MS	00011xxx
	System information, type 5/6	SACCH	N → MS	00011xxx
Miscellaneous	Channel mode modify	DCCH	N → MS	00010000
	Channel mode modify acknowledge	DCCH	MS → N	00010111
	Channel request	RACH	MS → N	—
	Classmark change	DCCH	MS → N	00010110
	Frequency redefinition	DCCH	N → MS	00010100
	Measurement report	SACCH	MS → N	00010101
	Synchronization channel information	SCH	N → MS	—
	RR status	DCCH	MS ↔ N	00010010

The Message Type (MT) is registered in the lower seven bits of the second octet (Tables 5.4–5.6). Otherwise, Layer 3 messages consist of Information Elements (IEs) of fixed or variable length; a Length Indicator (LI) is added for variable-length messages.

In idle mode, the MS is reading the BCCH information continuously and conducts periodic measurements of the signaling strength of the BCCH carriers in order to be able to select the current cell (section 4.5.4). In this state, there is no exchange of signaling messages with the network. The data required for RR and other signaling procedures are collected and stored: the list of neighboring BCCH carriers, thresholds for RR algorithms, CCCH configurations, information about the use of RACH and PCH, etc. This information is broadcast by the BSS on the BCCH (SYSTEM INFORMATION, types 1–4) and therefore

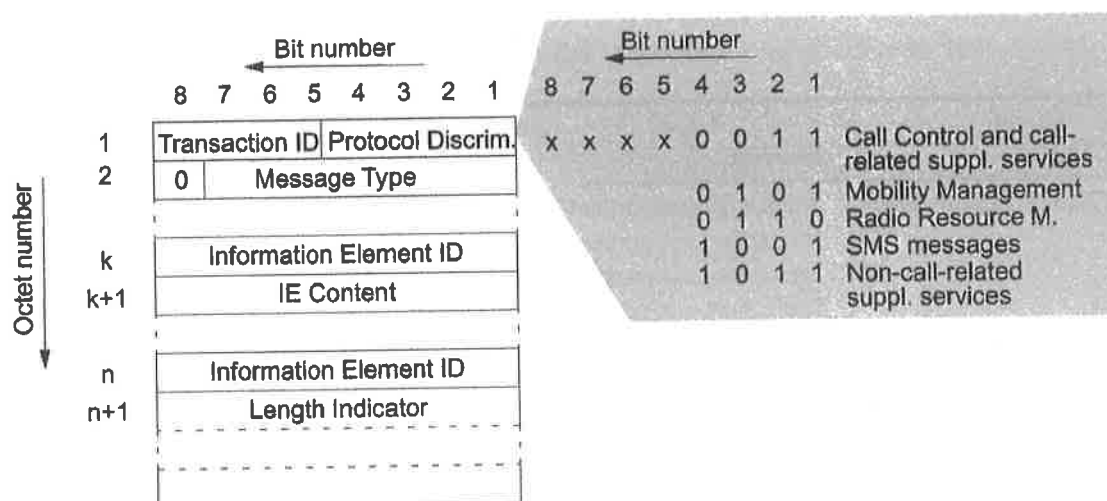


Figure 5.21 Format of a Um signaling message (Layer 3).

Table 5.5 MM messages.

Category	Message	Direction	MT
Registration	IMSI detach indication	MS → N	0x000001
	Location updating accept	N → MS	0x000010
	Location updating reject	N → MS	0x000100
	Location updating request	MS → N	0x001000
Security	Authentication reject	N → MS	0x010001
	Authentication request	N → MS	0x010010
	Authentication response	MS → N	0x010100
	Identity request	N → MS	0x001000
	Identity response	MS → N	0x001001
	TMSI reallocation command	N → MS	0x001010
	TMSI reallocation complete	MS → N	0x001011
Connection management	CM service accept	MS ↔ N	0x100001
	CM service reject	N → MS	0x100010
	CM service request	MS → N	0x100100
	CM reestablishment request	MS → N	0x101000
Miscellaneous	MM status	MS → N	0x110001

is available to all MSs currently in the cell. Also important is the periodic monitoring of the PCH so that paging calls are not lost. For this purpose, the BSS is sending on all paging channels of a cell continuously valid Layer 3 messages (PAGING REQUEST) which the MS can decode and recognize if its address is paged.

Table 5.6 CC messages for circuit-switched connections.

Category	Message	Direction	MT
Call establishment	Alerting	N → MS	0x000001
	Call confirmed	MS → N	0x001000
	Call proceeding	N → MS	0x000010
	Connect	N ↔ MS	0x000111
	Connect acknowledge	N ↔ MS	0x001111
	Emergency setup	MS → N	0x001110
	Progress	N → MS	0x000011
	Setup	N ↔ MS	0x000101
Call information phase	Modify	N ↔ MS	0x010111
	Modify complete	N ↔ MS	0x011111
	Modify reject	N ↔ MS	0x010011
	User information	N ↔ MS	0x010000
Call clearing	Disconnect	N ↔ MS	0x100101
	Release	N ↔ MS	0x101101
	Release complete	N ↔ MS	0x101010
Miscellaneous	Congestion control	N ↔ MS	0x111001
	Notify	N ↔ MS	0x111110
	Start DTMF	MS → N	0x110101
	Start DTMF acknowledge	N → MS	0x110010
	Start DTMF reject	N → MS	0x110111
	Status	N ↔ MS	0x111101
	Status enquiry	N ↔ MS	0x110100
	Stop DTMF	MS → N	0x110001
	Stop DTMF acknowledge	N → MS	0x110010

Connection setup and release

Each exchange of signaling messages with the network (BSS, MSC) requires a RR connection and the establishment of an LAPDm connection between MS and BTS. Setting up the RR connection can be initiated by the network or the MS (Figure 5.22). In either case, the MS sends a channel request (CHANREQUEST) on the RACH in order to get a channel assigned on the AGCH (immediate assignment procedure). There is also a procedure to deny a channel request (immediate assignment reject).

If the network does not immediately answer the channel request, the request is repeated using the Aloha method with a random number controlled timer (Figure 5.22). In the case of a network-initiated connection, this procedure is preceded by a paging call (PAGING REQUEST) to be answered by the MS (PAGING RESPONSE). After a RR connection has been successfully completed, the higher protocol layers (CM, MM) can receive and transmit signaling messages at SAPI 0.

In contrast to the setup of connections, the release is always initiated by the network (CHANNEL RELEASE). Reasons for the release of the channel could be end of the signaling

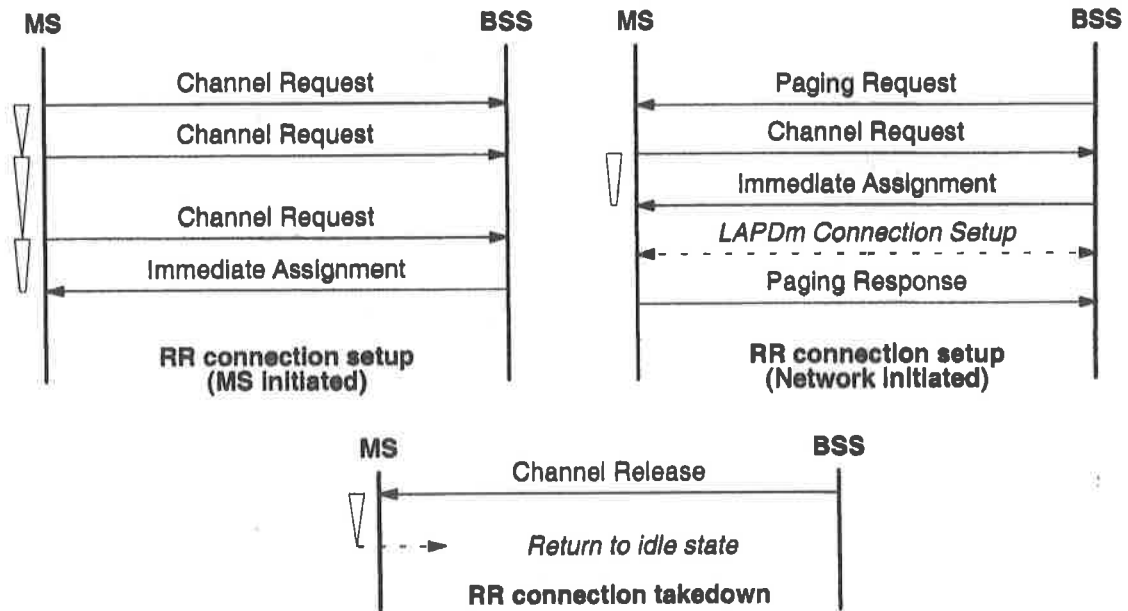


Figure 5.22 RR connection setup and takedown.

transaction, too many errors, removal of the channel in favor of a higher priority call (e.g. emergency call) or end of a call. After receiving the channel release command, the MS assumes the idle state following a brief waiting period (Figure 5.22).

Once a RR connection has been set up, the MS has either a SDCCH or a TCH with associated SACCH/FACCH available for exclusive bidirectional use. On the SACCH, data must be sent continuously (see also section 4.5.3), i.e. the MS keeps sending current channel measurements (MEASUREMENT REPORT; section 4.5.1) if no other signaling messages need to be sent. In the other direction, the BSS keeps sending system information (SYSTEM INFORMATION, alternating between type 5 and type 6). The information element with the coded measurement results contains the following among other data: RXLEV and RXQUAL of the serving cell as well as RXLEV and carrier frequency of up to six neighboring cells as well as their BSICs (Figure 5.23). The system information sent by the BSS on the SACCH contains first information about the neighbor cells and their BCCH (type 5), and second, information about the current cell (Type 6) such as CI and the current LAI.

Channel change

For established RR connections, a channel change within the cell can be performed (dedicated channel assignment; Figure 5.24) to change the configuration of the physical channel in use. Such a channel change can be requested by higher protocol layers, or it can be requested by the RR sublayer; however, it is always initiated by the network. When the MS receives an ASSIGNMENT COMMAND, the transmission of all signaling messages is suspended, the LAPDm connection is taken down, the traffic channel, if it exists, is switched off, and the old channel is deactivated. After activation of the new physical channel and a successful establishment of a new LAPDm connection (Layer 2), the held-back signaling messages can be transmitted.

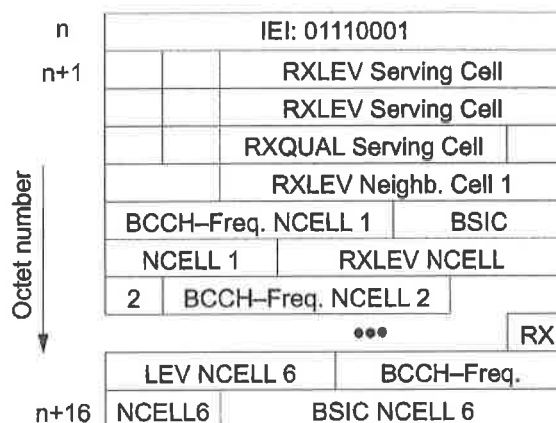


Figure 5.23 Measurement result (IE).

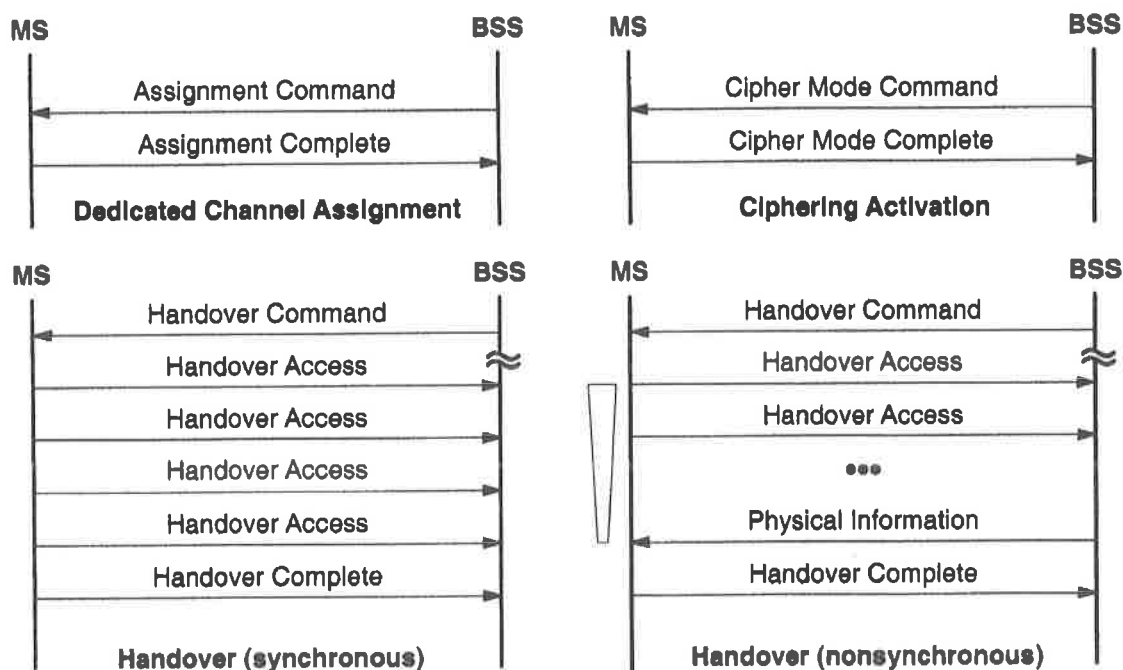


Figure 5.24 Channel change, encryption and handover.

Handover

A second signaling procedure to change the physical channel configuration of an established RR connection is the handover procedure, which is also initiated only from the network side and, for example, becomes necessary if the current cell is left. In contrast to ASSIGNMENT COMMAND, a HANDOVER COMMAND contains not only the new channel configuration but also information about the new cell (e.g. BSIC and BCCH frequency), the procedure variant

to establish a physical channel (asynchronous or synchronous handover; Figure 5.24), and a handover reference number.

Having received a **HANDOVER COMMAND** on the **FACCH**, the MS terminates the **LAPDm** connection on the old channel, interrupts the connection, deactivates the old physical channel, and finally switches over to the channel newly assigned in the **HANDOVER COMMAND**. On the main **DCCH** (in this case **FACCH**), the MS sends the unencrypted message **HANDOVER ACCESS** in an access burst (Figure 4.7, coding as on the **RACCH**, section 4.8) to the base station. Even though this is a message on the **FACCH**, an access burst is used because the MS at this time does not yet know the complete synchronization information. The eight data bits of the AB contain the handover reference of the handover command. The way in which these ABs are transmitted depends on whether both cells have synchronized their **TDMA** transmission or not.

In the case of existing synchronization, the AB (**HANDOVER ACCESS**) is sent in exactly four subsequent time slots of the main **DCCH** (**FACCH**). Thereafter, the MS activates the new physical channel in both directions, establishes an **LAPDm** connection, activates encryption and, finally, sends a message **HANDOVER COMPLETE** to the BSS. In the nonsynchronous case, the mobile station repeats the access burst until either a timer expires (handover failed) or until the base station answers with a **RR** message **PHYSICAL INFORMATION** which contains the currently needed timing advance and this way enables the establishment of the new **RR** connection.

Activation of ciphering

Another important **RR** procedure is the activation of ciphering. This is done by the BSS with the **CIPHER MODE COMMAND**, which also indicates that the **BTS** has activated its deciphering function. Having received the **CIPHER MODE COMMAND**, the MS activates ciphering as well as deciphering and sends the answer **CIPHER MODE COMPLETE** already in enciphered form. If the **BTS** is able to correctly decipher this message, the ciphering mode has been successfully established.

Other signaling procedures

In addition, there are a number of less significant signaling procedures defined, such as frequency redefinition, additional assignment, partial release or classmark change. The first concerns the change of the **MA**; section 4.2.3. The next two deal with a change of the physical channel configuration. With the last message, **CLASSMARK CHANGE**, the MS reports that it now belongs to a new power class (see Table 4.8), which can be achieved by installing a commercially available power booster kit, for example.

5.4.4 Mobility management

The main task of **MM** is to support the mobility of the MS; for example, by reporting the current location to the network or verifying the subscriber identity. Another task of the **MM** sublayer is to offer **MM** connections and associated services to the **CM** sublayer above. The message format for **MM** messages is the uniform Layer 3 signaling message format (Figure 5.21). **MM** has its own protocol discriminator, and the **MM** messages are marked with a type code (**MT**; Table 5.5).

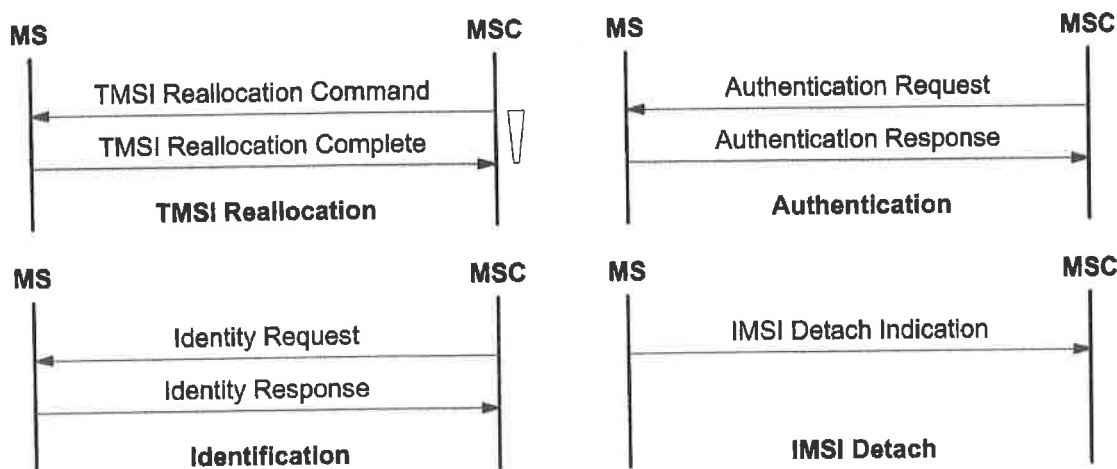


Figure 5.25 MM signaling procedures of category 'common'.

All MM procedures presume an established RR connection, i.e. a dedicated logical channel must be assigned with an established LAPDm connection in place, before MM transactions can be performed. These transactions occur between MS and MSC, i.e. messages are passed through the BSS transparently without interpretation and forwarded to the MSC with the DTAP transport mechanism. The MM procedures are divided into three categories: common, specific, and MM connection management. Whereas common procedures can always be initiated and executed as soon as a RR connection exists, specific procedures exclude one another, i.e. they cannot be processed simultaneously or during a MM connection. Conversely, a MM connection can only be set up if no specific procedure is running.

Common MM procedures

The common MM procedures are summarized in Figure 5.25. In addition to the IMSI detach procedure, they are all initiated from the network side. An important role for the protection of subscriber identity is held by the TMSI reallocation procedure. If the confidentiality of a subscriber's identity IMSI is to be protected (an optional network service), the signaling procedures across the air interface use the TMSI instead of the IMSI. This TMSI has only local significance within a location area and must be used together with the LAI for the unique identification of a subscriber.

For further protection, the TMSI can also be repeatedly reallocated (TMSI reallocation) which must be done at the latest when the location area changes. Otherwise this TMSI change is left as an option to the network operator, but it can be performed any time after an encrypted RR connection to the MS has been set up. The TMSI reallocation is either executed explicitly as a standalone procedure, or implicitly from other procedures using the TMSI, e.g. the location update. In the case of explicit TMSI reallocation, the network sends a TMSI REALLOCATION COMMAND with the new TMSI and the current LAI on an encrypted RR connection to the MS (Figure 5.25).

The MS stores the TMSI and LAI in nonvolatile SIM storage and acknowledges it with the message TMSI REALLOCATION COMPLETE. If this message reaches the MSC before

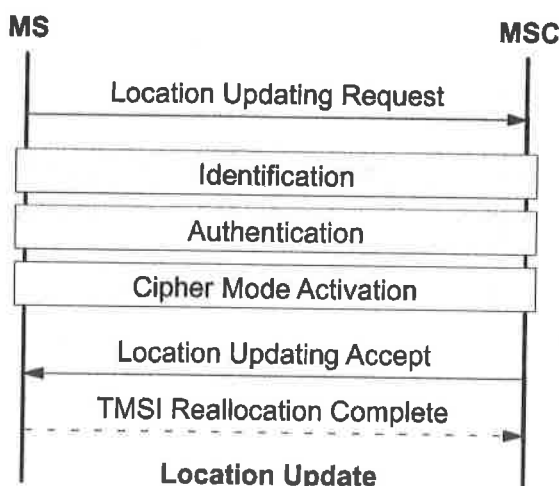


Figure 5.26 MM signaling procedures of category 'specific'.

the timer expires, the timer is cancelled and the TMSI is valid. However, if the timer expires before the acknowledgement arrives, the procedure is repeated. If it fails a second time, the old as well as the new TMSI are barred for a certain time interval and the IMSI is used for paging the MS. If the MS answers a paging call, TMSI reallocation is started again. Furthermore, the TMSI is assumed valid in spite of failed reallocation if it is used by the MS in subsequent transactions.

Two more common procedures are used for the identification of a MS or a subscriber (identification procedure) and for the verification of the respective identity (authentication procedure). For the identification of a MS, there is the equipment identity IMEI as well as the subscriber identity IMSI which is assigned to the MS through the SIM card. The network may request these two identity parameters at any time from the MS with an **IDENTITY REQUEST**. Therefore the MS must be able at any time to supply these identity parameters to the network with an **IDENTITY RESPONSE** message.

Authentication also assigns a new key for encryption of user payload data. This procedure is started from the network with an **AUTHENTICATION REQUEST** message. A mobile station must be able to process this request at any time during a RR connection. The MS calculates the new key K_c for the encryption of user data from the information obtained during authentication which is locally stored, and it also calculates authentication information to prove its identity without doubt. This authentication data are transmitted with an **AUTHENTICATION RESPONSE** message to the MSC which evaluates them. If the answer is not valid and the authentication has therefore failed, further processing depends on whether the IMSI or TMSI was used. In the case of TMSI, the network can start the identification procedure. If the implied IMSI is not identical to that associated with the TMSI by the network, the authentication is restarted with new correct parameters. If the two IMSIs agree, or the IMSI was used *a priori* by the MS, the authentication has failed, which is indicated to the MS with an **AUTHENTICATION REJECT** message. This forces the MS to cancel all of the assigned identity and security parameters (TMSI, LAI, K_c) and to enter idle mode, so that only simple cell selection and emergency calls are enabled.

If the MS is powered off or the SIM has been removed, the MS is not reachable because the MS does not monitor the paging channel, and calls cannot be delivered. In order to relieve the paging load on the BSS caused by unnecessary paging calls, a network operator can optionally request an explicit deregistration message from the MS, which is not normally required. This option is signaled by setting a flag on the BCCH (SYSTEM INFORMATION, type 3) and on the SACCH (SYSTEM INFORMATION, type 8). If the flag is set, the MS sends an IMSI DETACH INDICATION message when it powers off or when the SIM is removed, which allows the network to mark the MS as inactive. The IMSI detach procedure is the only common procedure that cannot be started at an arbitrary time even during a specific procedure: its start has to be delayed until any specific procedure has ended.

Specific MM procedures

In GSM systems, updating of current location information is the sole responsibility of the MS. Using the information broadcast on BCCH channels, it has to recognize any change in the current location area and report it to the network, so that the databases HLR and VLR can be kept up to date. The generic structure of a location update is shown in Figure 5.26: the MS requests to update its current location information in the network with a LOCATION UPDATING REQUEST. If this can be done successfully, the network acknowledges this with a message LOCATION UPDATING ACCEPT. In the course of a location update, the network can ask for the MS's identity and check it out (identification and authentication). If the service 'confidential subscriber identity' has been activated, a new TMSI assignment is a permanent component of the location update. In this case, enciphering of user data on the RR connection is activated, and the new TMSI is transmitted together with the message LOCATION UPDATING ACCEPT and is acknowledged with the message REALLOCATION COMPLETE. Periodic updating of location information can be used to indicate the presence of the MS in the network. For this purpose, the MS keeps a timer which periodically triggers a location update procedure. If this option is in use, the timer interval to be used is broadcast on the BCCH (SYSTEM INFORMATION, type 3). The procedure IMSI attach is the converse of the procedure IMSI detach (see Figure 5.25) and is executed as a special variant of the location update if the network requires this. However, the MS executes an IMSI detach only if the LAI broadcast on the BCCH agrees with the LAI stored in the MS. If the stored LAI and received LAI differ, a normal location update procedure is executed.

MM connection management

Finally, there is a third category of MM procedures which are needed for the establishment and the operation of MM connections (Figure 5.27). A MM connection is established on request from the CM sublayer above and serves for the exchange of messages between CM entities, where each CM entity has its own MM connection (Figure 5.12). The procedures for the setup of MM connections are different depending on whether initiation occurs from the network or the MS.

Setting up a MM connection from the side of the MS presumes the existence of a RR connection, but a single RR connection can be used by multiple MM connections. The MM connection can only be established if the MS has executed a successful location update in the current location area. An exception is an emergency call, which is possible at any time. If there is a request from the CM sublayer for a MM connection, it may be delayed or rejected

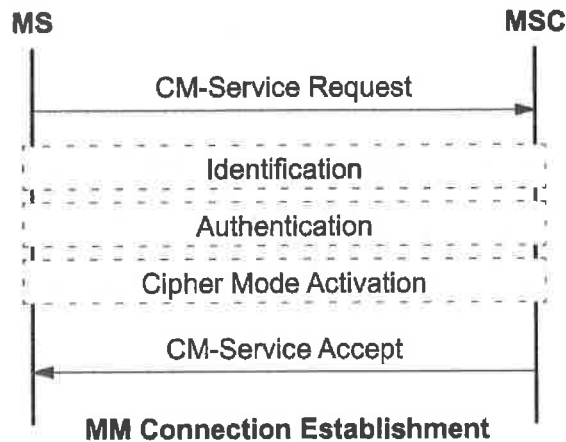


Figure 5.27 MM signaling procedures of category 'MM connection management'.

if there are specific procedures active, depending on implementation. If the MM connection can be established, the MS sends the message **CM-SERVICE REQUEST** to the network. This message contains information about the mobile subscriber (IMSI or TMSI) as well as information about the requested service (outgoing voice call, SMS transfer, activation or registration of a supplementary service, etc.). Depending on these parameters, the network can execute any common MM procedure (except IMSI detach) or activate enciphering of user data. If the MS receives the message **CM-SERVICE ACCEPT** or the local message from the RR sublayer that enciphering was activated, it treats this as an acceptance of the service request, and the requesting CM entity is informed about the successful setup of a MM connection. Otherwise, if the service request has been rejected by the network, the MS receives a message **CM-SERVICE REJECT** and the MM connection cannot be established.

The network-initiated setup of a MM connection does not require an exchange of CM service messages. After successful paging, a RR connection is established, and the sublayer on the network side executes one of the MM procedures if necessary (mostly location update) and requests from the RR sublayer the activation of user data encryption. If these transactions are successful, the service requesting CM entity is informed and the MM connection is established.

5.4.5 Connection management

CC is one of the entities of CM; the CM sublayer is shown in Figure 5.12. It comprises procedures to establish, control and terminate calls. Several parallel CC entities are provided, such that several parallel calls on different MM connections can be processed. For CC, finite state models are defined both on the mobile side as well as on the network side. The two entities at the MS and MSC sites each instantiate a protocol automaton, and these communicate with each other using the messages in Table 5.6 and the uniform Layer 3 signaling message format (Figure 5.21).

Parts of CC in the MS are presented schematically in Figures 5.28 and 5.29. They show the mobile-originating and mobile-terminating setup of a call and the mobile/network initiated call takedown. If there is a desire to call from the MS (mobile-originating call), the CC entity

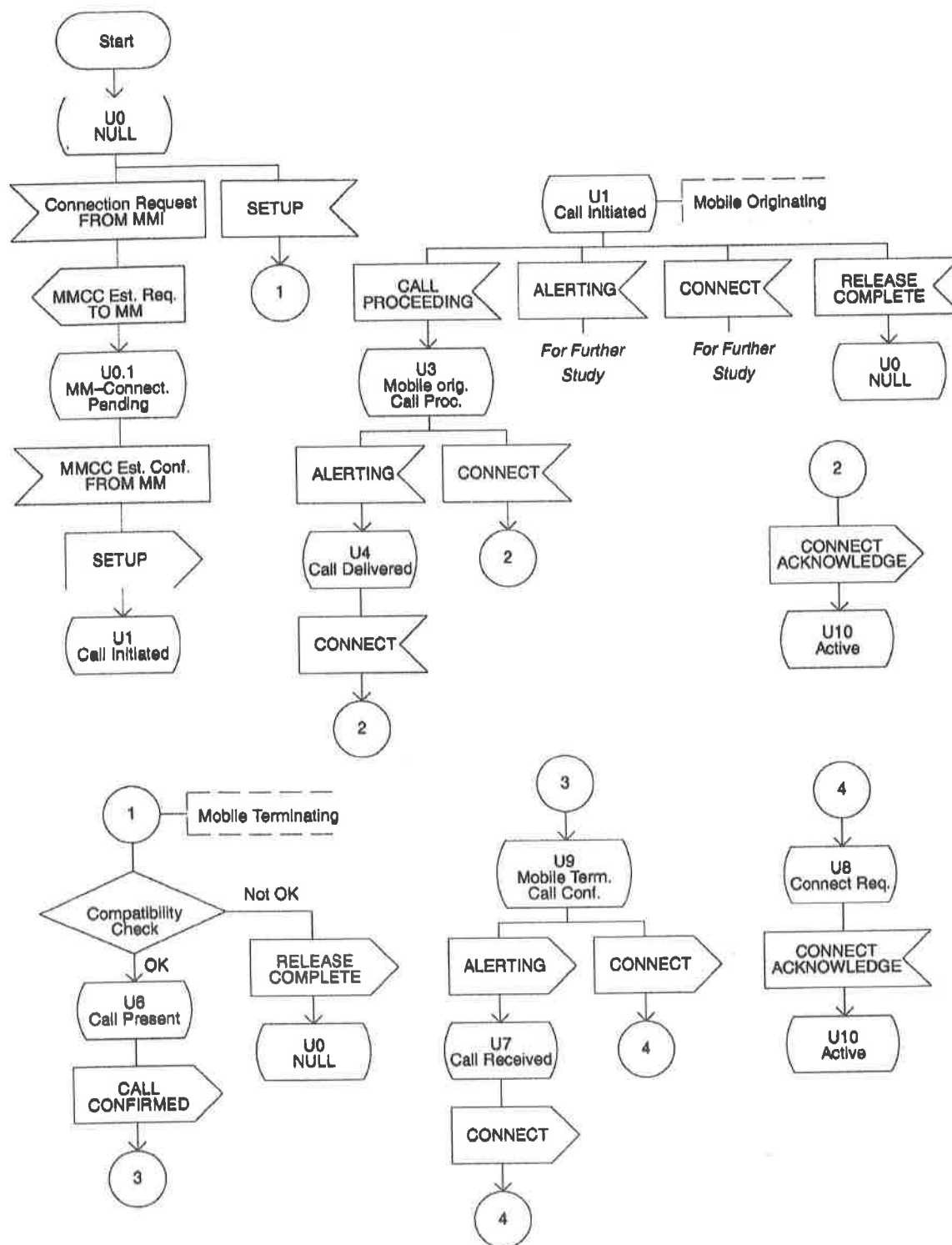


Figure 5.28 Call setup (MS): mobile-originating and mobile-terminating.