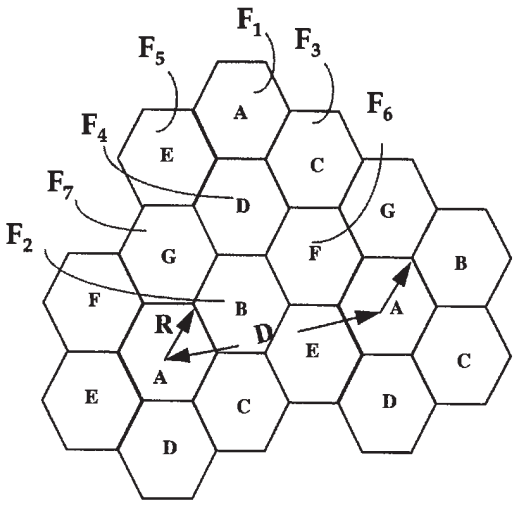


FIGURE 23.1 Fixed channel assignment.

the number of channels per cell N by

$$P_B = \frac{A^N/N!}{\sum_{i=0}^{N} A^i/i!}$$

This applies to the case of blocked calls cleared. If calls are delayed, the grade of service becomes the probability of calls being delayed P_Q and is given by the erlang-C formula [6],

$$P_Q = \frac{\dfrac{A^N}{N!(1-A/N)}}{\sum_{i=0}^{N-1} A^i/i! + \dfrac{A^N}{N!(1-A/N)}}$$

FCA is used in almost all existing cellular mobile networks employing FDMA or TDMA. To illustrate resource assignment in FCA, we describe the FCA schemes used in global system for mobile communications (GSM or DCS) [17]. Here, the mobile terminal continuously monitors the received signal strength and quality of a broadcast channel along with the identity of the transmitting base station. This mechanism allows both the mobile and the network to keep track of the subscriber movements throughout the service area. When the mobile terminal tries and eventually succeeds in accessing the network (see Fig. 23.2), a two way control channel is assigned that allows for both authentication and ciphering mode establishment. On completion of these two phases, the setup phase initiates and a traffic channel is eventually assigned to the mobile terminal. Here, the terminal only knows the identity of the serving cell; the (control and traffic) radio channel assignment is a responsibility of the network and fulfills the original carrier to cell association, while simplifying the control functions during normal system operation.

23.4 Enhanced Fixed Channel Assignment

FCA has the advantage of having simple realization. Since the frequency sets are preassigned to cells based on long-term traffic demand, however, it cannot adapt to traffic variation

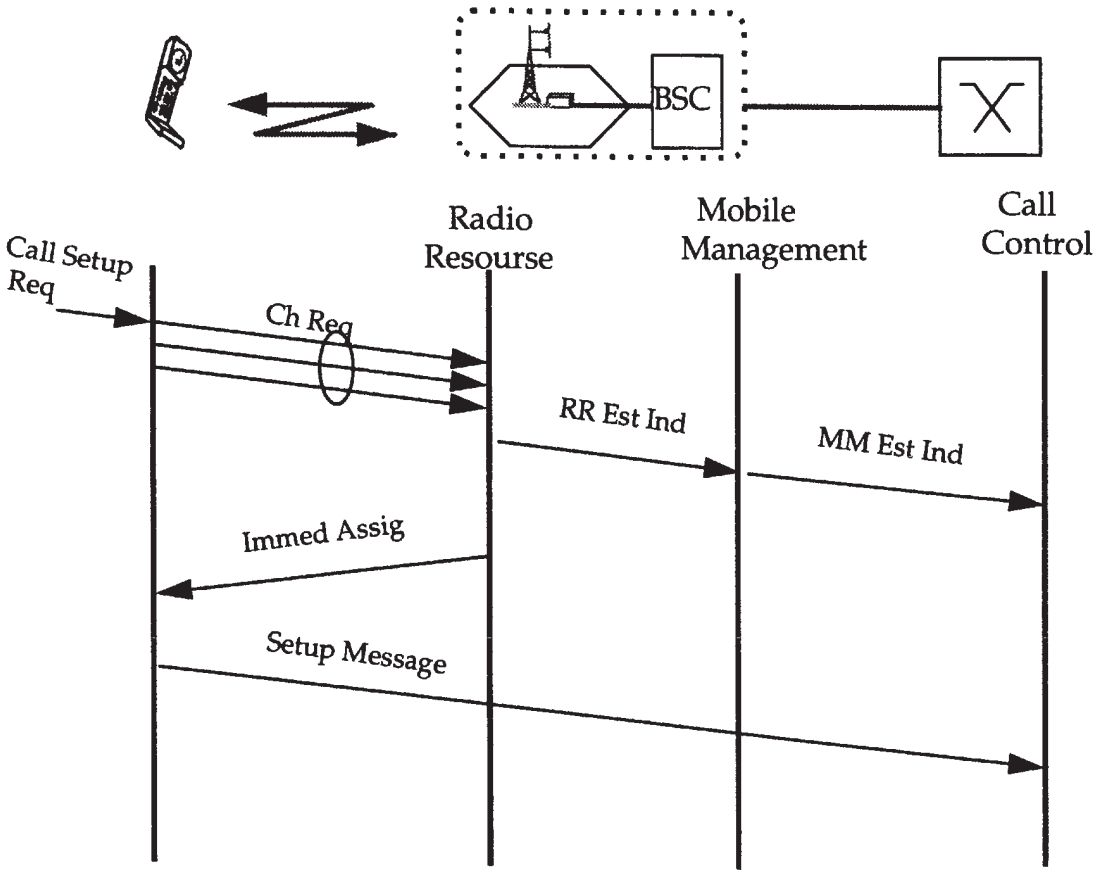


FIGURE 23.2 Radio resource assignment in GSM/DCS. *Source*: Jabbari, B., et al., Network issues for wireless personal communications. *IEEE Commun. Mag.*, 33(1), 1995.

across cells and, therefore, FCA will result in poor bandwidth utilization. To improve the utilization while maintaining the implementation simplicity, various strategies have been proposed as enhancements to FCA and deployed in existing networks. Two often used methods are *channel borrowing* and *directed retry*, which are described here.

In the channel borrowing strategy [11, 12, 18, 29], channels that are not in use in their cells may be borrowed by adjacent cells with high offered traffic on a call-by-call basis. Borrowing of channels allows the arriving calls to be served in their own cell. This implies that there will be further restrictions in using the borrowed channels in other cells. Various forms of borrowing have been surveyed in Tekinay and Jabbari [23].

In directed retry [10, 13, 14], a call to or from a mobile subscriber may try other cells with channels with sufficient signal strength meeting the C/I constraint if there are no channels available in its own cell to be served. In some cases it may be necessary to direct some of the calls in progress in a given congested cell to adjacent lightly loaded cells (those calls that can be served by adjacent cells) in order to accommodate the new calls in that given cell. This is referred to as directed handoff [13, 20]. The combination of these two capabilities provides a significant increase in bandwidth utilization.

23.5 Dynamic Channel Assignment

In dynamic channel assignment [2, 3, 4, 7, 8, 9, 12, 21], the assignment of channels to cells occurs based on the traffic demand in the cells. In other words, channels are pooled together and assignments are made and modified in real time. Therefore, this assignment scheme has the potential to achieve a significantly improved bandwidth utilization when there are temporal or spatial traffic variations.

In DCA, the interference constraints are ensured by a real-time evaluation of the most suitable (less interfered) channels that can be activated in a given cell (reassignment may be needed). That is, the system behaves as if the frequency plan was dynamically changing to meet the actual radio link quality and traffic loads, realizing an implicit sharing of the frequency band under interference constraints.

The implementation of DCA generally is more complex due to the requirement for system-wide state information where the state refers to which channel in which cell is being used. Obviously, it is impractical to update the system state in a large cellular network at any time, especially those based on microcells, as the controller will be overloaded or call set delay will be unacceptable. Therefore, methods have been devised based on a limited state space centralized control [15] or based on a distributed control to perform the necessary updating. A method referred to as maximum packing suggested in Everitt and Macfadyen [15], records only the number of channels in use. Distributed control schemes, however, where channel assignments are made at mobile stations or base stations may be attractive. These schemes are particularly suitable for a mobile controlled resource assignment where each mobile measures the actually perceived interference and decides to utilize a radio resource in a completely decentralized way.

Design of DCA algorithms is critical to achieve the potential advantages in efficiency and robustness to traffic heterogeneity throughout cells as compared to FCA and enhanced FCA. Poor DCA algorithms, however, might lead to an uncontrolled global situation, i.e., the locally selected channel might be very good for the specific mobile terminal and at the same time very poor for the interference level induced to other traffic sources.

Two classes of *regulated DCA* and *segregation DCA* are discussed here due to their importance. In a regulated DCA, appropriate thresholds tend to maintain the current channel, avoiding useless handoffs that can tie up channels; in a segregation DCA, channel acquisition obeys priorities assigned to each channel. In the latter, the channels successfully activated in a cell have a higher assignment probability than those found to have a high interference in the cell; of course, priorities change in time depending on changes in the system status. These types of assignment lead to a (frequency) plan still changing dynamically but more slowly and, in general, only because of substantial load imbalance. In steady-state conditions, the plan either tends to be confirmed or fluctuates around a basic configuration, proven to be suboptimal in terms of bandwidth efficiency.

Both digital European cordless telecommunications (DECT) and cordless technology known as CT2 are employing the DCA technique [25], and the next generation cellular systems are foreseen to deploy it. In CT2 the handset and the base station jointly determine a free suitable channel to serve the call. In DECT, the mobile terminal (portable handset) not only recognizes the visited base station (radio fixed part) through a pilot signal but continuously scans all of the system channels and holds a list of the less interfered ones. The potentially available channels are ordered by the mobile terminal with respect to the measured radio parameters, namely, the radio signal strength indicator, measuring and combining cochannel, adjacent-channel, and intermodulation interference. When a connection has to be established, the best channel is used to communicate on the radio interface.

It is possible to have a hybrid of DCA and FCA in a cellular network in which a fraction of channels are fixed assigned and the remainder are allocated based on FCA. This scheme has less system implementation complexity than the DCA scheme but provides performance improvement (lower probability of call blocking) depending on the DCA–FCA channel partitioning [19].

In general, DCA schemes cannot be considered independently of the adopted power control mechanism because the transmitted power from both mobile terminal and base station substantially affects the interference within the network. For a detailed discussion of DCA and power control, the readers are referred to [5]. Despite the availability of several power control algorithms, much work is needed to identify realizable power and channel control algorithms that maximize bandwidth efficiency. Nevertheless, realization of power control mechanisms would require distribution of the system status information and information exchange between mobile terminal and network entities. This in turn will involve overhead and terminal power usage.

The performance of DCA depends on the algorithm implementing this capability [13, 16] In general, due to interactions between different cells the performance of the system will involve modeling the system as a whole, as opposed to in FCA where cells are treated independently. Therefore, mathematical modeling and performance evaluation of DCA becomes quite complex. Simplifying assumptions may, therefore, be necessary to obtain approximate results [15, 22]. Simulation techniques have been widely used in evaluation of DCA performance. For a representative performance characteristics of DCA and a comparison to enhanced FCA schemes the readers are referred to [14].

23.6 CDMA Systems

Now we turn our attention to systems based on direct sequence spread spectrum CDMA. Examples include the second generation system based on IS-95 and the proposed third generation system based on Wideband CDMA (W-CDMA) [28]. Due to the simultaneous availability and reuse of the entire frequency spectrum in every cell, opportunity for dynamic channel assignment exists inherently. However, the problem of dynamic channel allocation turns into the problem of allocation of resources such as spreading codes, equipment limitations, intracell and intercell interference. Added to this is the practical implication of the soft capacity, which allows temporary degradation of the links. Furthermore, the use of voice activity compression technique allows increase in channel utilization while adding new performance measures such as speech freeze-out fraction.

The capacity of the CDMA system in terms of Erlang for both forward and reverse directions has been derived in [27]. The effect of imperfect power control, other cell interference has been taken into account for traffic sources with a given user activity. This capacity reflects the average traffic load the system can carry under a specified performance measure. This measure, although it is treated similar to the probability of blocking, in actuality is somewhat different from those commonly used in orthogonal time and frequency systems and indeed reflects the outage probability. The performance measure has been defined in [27] as the ratio of background noise density to total interference density, N_0/I_0, being greater than a specified value (for example 0.1). For the simple equally loaded cells case where the limiting direction is the reverse link, with a bandwidth of W, source data rate of R bits/second, required energy to interference ratio of E_b/I_0, and ratio of other cells to

own cell (sector) interference of f, the condition to have an outage is given by [27]:

$$P_{\text{outage}} = \Pr\left[\sum_{i=1}^{K(1+f)} v_i \left(E_b/I_0\right)_i > (W/R)\left(1 - N_0/I_0\right)\right]$$

In the above formula, v_i represents the event the bi-state source is in the active state and K is the number of sources in a single cell. Note that $(E_b/I_0)_i$ is the energy per bit-to-interference ratio of source i as received at the observed base station. Additional scenarios and some approximations for special cases have been further carried out and discussed in [26]. Such analysis can be used to evaluate the system capacity for when traffic is nonuniformly distributed across the cells/sectors.

23.7 Conclusion

In this chapter we have classified and reviewed channel assignment techniques. We have emphasized the advantages of DCA schemes over FCA in terms of bandwidth utilization in a heterogeneous traffic environment at the cost of implementation complexity. The DCA schemes are expected to play an essential role in future cellular and microcellular networks.

CDMA inherently provides the DCA capability. However, the system capacity depends on factors including power control accuracy and other cell interference. We presented a brief discussion of capacity of CDMA.

References

[1] Abramson, N., Multiple access techniques for wireless networks. *Proc. of IEEE,* 82(9), 1994.

[2] Anderson, L.G., A simulation study of some dynamic channel assignment algorithms in a high capacity mobile telecommunications system. *IEEE Trans. on Comm.,* COM-21(11), 1973.

[3] Beck, R. and Panzer, H., Strategies for handover and dynamic channel allocation in micro-cellular mobile radio systems. *Proceedings of the IEEE Vehicular Technology Conference,* 1989.

[4] Chuang, J.C.-I., Performance issues and algorithms for dynamic channel assignment. *IEEE J. on Selected Areas in Comm.,* 11(6), 1993.

[5] Chuang, J.C.-I., Sollenberger, N.R., and Cox, D.C., A pilot-based dynamic channel assignment schemes for wireless access TDMA/FDMA systems. *Internat. J. of Wireless Inform. Networks,* Jan. 1994.

[6] Cooper, R.B.,*Introduction to Queueing Theory,* 3rd ed., CEEPress Books, 1990.

[7] Cox, D.C. and Reudnik, D.O., A comparison of some channel assignment strategies in large-scale mobile communications systems. *IEEE Trans. on Comm.,* COM-20(2), 1972.

[8] Cox, D.C. and Reudnik, D.O., Increasing channel occupancy in large-scale mobile radio environments: dynamic channel reassignment. *IEEE Trans. on Comm.,* COM-21(11), 1973.

[9] Dimitrijevic, D. and Vucetic, J.F., Design and performance analysis of algorithms for channel allocation in cellular networks. *IEEE Trans. on Vehicular Tech.,* 42(4), 1993.

[10] Eklundh, B., Channel utilization and blocking probability in a cellular mobile telephone system with directed retry. *IEEE Trans. on Comm.,* COM 34(4), 1986.

[11] Elnoubi, S.M., Singh, R., and Gupta, S.C., A new frequency channel assignment algorithm in high capacity mobile communications. *IEEE Trans. on Vehicular Techno.,* 31(3), 1982.

[12] Engel, J.S. and Peritsky, M.M., Statistically-optimum dynamic server assignment in systems with interfering servers. *IEEE Trans. on Comm.,* COM-21(11), 1973.

[13] Everitt, D., Traffic capacity of cellular mobile communications systems. *Computer Networks and ISDN Systems,* ITC Specialist Seminar, Sept. 25–29, 1989, 1990.

[14] Everitt, D., Traffic engineering of the radio interface for cellular mobile networks. *Proc. of IEEE,* 82(9), 1994.

[15] Everitt, D.E. and Macfadyen, N.W., Analysis of multicellular mobile radiotelephone systems with loss. *BT Tech. J.,* 2, 1983.

[16] Everitt, D. and Manfield, D., Performance analysis of cellular mobile communication systems with dynamic channel assignment. *IEEE J. on Selected Areas in Comm.,* 7(8), 1989.

[17] Jabbari, B., Colombo, G., Nakajima, A., and Kulkarni, J., Network issues for wireless personal communications. *IEEE Comm. Mag.,* 33(1), 1995.

[18] Jakes, W.C. Ed. *Microwave Mobile Communications,* Wiley, New York, 1974, reissued by IEEE Press, 1994.

[19] Kahwa, T.J. and Georganas, N.D., A hybrid channel assignment scheme in Large scale, cellular-structured mobile communications systems. *IEEE Trans. on Comm.,* COM-26(4), 1978.

[20] Karlsson, J. and Eklundh, B., A cellular mobile telephone system with load sharing—an enhancement of directed retry. *IEEE Trans. on Comm.,* COM 37(5), 1989.

[21] Panzer, H. and Beck, R., Adaptive resource allocation in metropolitan area cellular mobile radio systems. *Proceedings of the IEEE Vehicular Technology Conference,* 1990.

[22] Prabhu, V. and Rappaport, S.S., Approximate analysis for dynamic channel assignment in large systems with cellular structure. *IEEE Trans. on Comm.,* COM-22(10), 1974.

[23] Tekinay, S. and Jabbari, B., Handover and channel assignment in mobile cellular networks. *IEEE Comm. Mag.,* 29(11), 1991.

[24] Telecommunications Industry Association. TIA Interim Standard IS-95, CDMA Specifications, 1993.

[25] Tuttlebee, W.H.W., Cordless personal communications. *IEEE Comm. Mag.,* Dec. 1992.

[26] Viterbi, A.J., *CDMA—Principles of Spread Spectrum Communication,* Addison Wesley, Reading, MA, 1995.

[27] Viterbi, A.J. and Viterbi, A.J., Erlang capacity of a power controlled CDMA system. *IEEE JSAC,* 11(6), 892–900, 1993.

[28] W-CDMA. Feature topic. *IEEE Communications Magazine,* Sept. 1998.

[29] Zhang, M. and Yum, T.-S.P., Comparison of channel-assignment strategies in cellular mobile telephone systems. *IEEE Trans. on Vehicular Tech.,* 38(4), 1989.

24

Radiolocation Techniques

24.1 Introduction 24-1
24.2 Description of Radiolocation Methods 24-2
Angle of Arrival • Signal Strength • Time-Based Location
24.3 Location Algorithms 24-4
Problem Formulation • Location Solutions
24.4 Measures of Location Accuracy 24-6
Cramér-Rao Lower Bound • Circular Error Probability • Geometric Dilution of Precision
24.5 Location in Cellular Systems 24-8
24.6 Sources of Location Error 24-9
Multipath • NLoS Propagation • Multiple-Access Interference
24.7 Summary and Conclusions 24-11
Defining Terms 24-11
References 24-12
Further Information 24-12

Gordon L. Stüber
Georgia Institute of Technology

James J. Caffery, Jr.
Georgia Institute of Technology

24.1 Introduction

Several location technologies have been developed and commercially deployed for locating wireless radios including Decca, Loran, Omega, the Global Positioning System (GPS), and the Global Navigation Satellite System (GLONASS). GPS, originally developed for military use, is perhaps the most popular commercial location system today, providing location accuracies to within 100 m.

All of the above systems use a location technique known as radiolocation. A radiolocation system operates by measuring radio signals traveling between a mobile station (MS) and a set of fixed stations (FSs). The measurements are then used to determine the length and/or direction of the radio paths, and the MS's position is derived from known geometrical relationships. In general, measurements from $n + 1$ FSs are necessary to locate an MS in n dimensions. To achieve high accuracy in a radiolocation system, it is necessary that a line-of-sight (LoS) exist between the MS and FSs. Otherwise, large errors are likely to be incurred.

With the above mentioned radiolocation technologies, the MS formulates its own position by using signals received from the FSs. This form of location is often referred to as *self-positioning*. In these systems, a special receiver is placed in the MS to calculate the MS's position. Alternatively, the position of the MS could be calculated at a remote location by using signals received at the FSs. This form of radiolocation is known as *remote-positioning* and requires a transmitter for the MS.

0-8493-8597-0/99/$0.00+$.50

© 1999 by CRC Press LLC

Over the last several years, there has been increased interest in developing location services for wireless communications systems. An array of applications for such technology exists including location sensitive billing, fraud detection, cellular system design and resource management, fleet management, and Intelligent Transportation Services (ITS) [8]. The greatest driving force behind location system development in wireless systems has been the FCC's Emergency-911 (E-911) requirements, where a wireless E-911 caller must be located within an *rms* accuracy of 125 m in 67% of the cases [4]. Adding GPS to the handset is not a universal solution because of the large pool of existing handsets. Remote-positioning radiolocation is a natural choice since it requires no modification of existing MS handsets and most, if not all, of the complexity could be incorporated into the network side.

In this chapter, techniques for locating wireless users using measurements from radio signals are described. A few algorithms for location estimation are developed along with a discussion of measures of accuracy. The radiolocation methods that are appropriate for wireless location and the major sources of error in various mobile cellular networks are discussed.

24.2 Description of Radiolocation Methods

Radiolocation systems can be implemented that are based on angle-of-arrival (AoA), signal strength, time-of-arrival (ToA), time-difference-of-arrival (TDoA), or their combinations. These are briefly discussed below.

24.2.1 Angle of Arrival

AoA techniques estimate the MS location by first measuring the arrival angles of a signal from a MS at several FSs (Fig. 24.1). The AoA can be determined through the use of directive antennas or antenna arrays. Simple geometric relationships are then used to determine the location by finding the intersections of the lines-of-position (see Fig. 24.1). To generate a position fix, the AoA method requires that the signal transmitted from the MS be received by a minimum of two FSs.

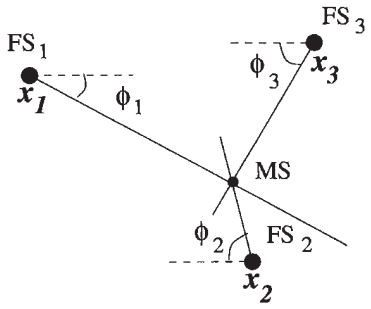


FIGURE 24.1 The measured angles, ϕ_i, determine the position of the MS for a given FS geometry.

24.2.2 Signal Strength

Radiolocation based on signal strength measurements uses a known mathematical model describing the **path loss** attenuation with distance. Since measurement of signal strength

provides a distance estimate between a MS and FS, the MS must lie on a circle centered at the FS. Hence, for signal strength based radiolocation, the lines-of-position are defined by circles. By using measurements from multiple FSs, the location of the MS can be determined from the intersection of circles (Fig. 24.2).

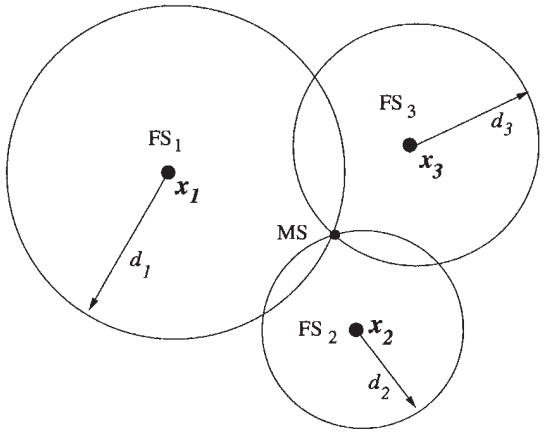


FIGURE 24.2 For signal strength and ToA based radiolocation systems the location of the MS is at the intersection of circles of radius d_i.

A second method makes use of premeasured signal strength contours around each FS. Received signal strength measured at multiple FSs can be mapped to a location by overlaying the contours for each FS. This technique can be used to combat **shadowing**, as discussed in Section 24.6.1.

24.2.3 Time-Based Location

The final class of radiolocation techniques are those based on estimating the ToAs of a signal transmitted by the MS and received at multiple FSs or the TDoAs of a signal received at multiple pairs of FSs. In the ToA approach, the distance between a MS and FS is measured by finding the one-way propagation time between the MS and FS. Geometrically, this provides a circle, centered at the FS, on which the MS must lie. Given a ToA at FS i, the equation for the circle is given by

$$\tau_i = D_i(\boldsymbol{x}_s)/c \tag{24.1}$$

where $D_i(\boldsymbol{x}_s) = \|\boldsymbol{x}_i - \boldsymbol{x}_s\|$, $\boldsymbol{x}_i$ is the position of ith FS, $\boldsymbol{x}_s$ is the position of the MS and c is the speed of light. By using at least three base stations to resolve ambiguities, the MS's position is given by the intersection of circles (Fig. 24.2). Since the ToA and path loss based signal strength methods are based on distance measurements between the MS and FSs, they are often referred to as *ranging* systems.

In the TDoA approach, time differences of arrival are used. Hence, the time of signal transmission need not be known. Since the hyperbola is a curve of constant time *difference* of arrival for two FSs, a TDoA measurement defines a line-of-position as a hyperbola, with the foci located at one of the two FSs. For FSs i and j, the equation of the hyperbola, $\rho_{i,j}$,

for a given TDoA is

$$\rho_{i,j} = \frac{D_i(\boldsymbol{x}_s) - D_j(\boldsymbol{x}_s)}{c}. \tag{24.2}$$

The location of the MS is at the intersection of the hyperbolas (Fig. 24.3). In general, for N FSs receiving the signal from the MS, $N-1$ non-redundant TDoA measurements can be made. Thus, a MS can be located in $N-1$ dimensions.

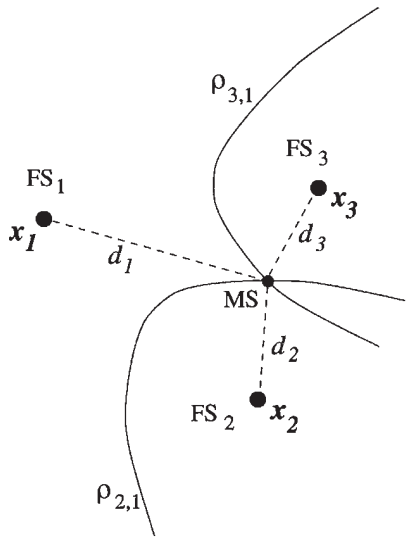


FIGURE 24.3 For TDoA based radiolocation the position of the MS is at the intersection of hyperbolas. The curves represent constant differences in distance to the MS with respect to the first FS, $\rho_{i,1} = d_i - d_1$.

24.3 Location Algorithms

When there are no measurement errors, the lines-of-position intersect at a point and a geometric solution can be obtained for the MS's location by finding the intersection of the lines of position. However, in practice, measurement errors occur and the lines of position do not intersect at a point. Consequently, other solution methods must be utilized. In the following, a solution approach is developed for two-dimensional location systems where the MS is located at $\boldsymbol{x}_s = [x_s, y_s]^T$ and the FSs are located at $\boldsymbol{x}_i = [x_i, y_i]^T$ for $i = 1, \ldots, N$.

24.3.1 Problem Formulation

In general, the $N \times 1$ vector of noisy measurements, $\boldsymbol{r}$, from a set of N FSs can be modeled by

$$\boldsymbol{r} = \boldsymbol{C}(\boldsymbol{x}_s) + \boldsymbol{n} \tag{24.3}$$

where $\boldsymbol{n}$ is an $N \times 1$ measurement noise vector, generally assumed to have zero mean and an $N \times N$ covariance matrix $\boldsymbol{\Sigma}$. The system measurement model $\boldsymbol{C}(\boldsymbol{x}_s)$ depends on the

location method used:

$$\boldsymbol{C}(\boldsymbol{x}_s) = \begin{cases} \boldsymbol{D}(\boldsymbol{x}_s) & \text{for ToA} \\ \boldsymbol{\mathcal{R}}(\boldsymbol{x}_s) & \text{for TDoA} \\ \boldsymbol{\Phi}(\boldsymbol{x}_s) & \text{for AoA} \end{cases} \tag{24.4}$$

where

$$\boldsymbol{D}(\boldsymbol{x}_s) = [\tau_1, \tau_2, \ldots, \tau_N]^T \tag{24.5}$$

$$\boldsymbol{\mathcal{R}}(\boldsymbol{x}_s) = [\rho_{2,1}, \rho_{3,1}, \ldots \rho_{N,1}]^T \tag{24.6}$$

$$\boldsymbol{\Phi}(\boldsymbol{x}_s) = [\phi_1, \phi_2, \ldots, \phi_N]^T \ . \tag{24.7}$$

The terms τ_i and $\rho_{i,1}$ are the ToAs and TDoAs defined in Eqs. (24.1) and (24.2), respectively, where without loss of generality, the TDoAs are referenced to the first FS. If the time of transmission τ_s is needed to form the ToA estimates, it can be incorporated into $\boldsymbol{x}_s$ as a parameter to be estimated along with x_s and y_s. The unknown parameter vector can then be modified to $\boldsymbol{x}_s = [x_s, y_s, \tau_s]^T$, while the system measurement model becomes $\boldsymbol{C}(\boldsymbol{x}_s) = \boldsymbol{D}(x_s, y_s) + \tau_s \mathbf{1}$.

The AoAs are defined by

$$\phi_i = \tan^{-1}\left(\frac{y_i - y_s}{x_i - x_s}\right) \ . \tag{24.8}$$

Although not explicitly shown in the above equations, τ_i, $\rho_{i,1}$ and ϕ_i are nonlinear functions of $\boldsymbol{x}_s$.

A well-known approach for determining an estimate from a noisy set of measurements is the method of **least squares** (LS) estimation. The weighted least squares (WLS) solution is formed as the vector $\hat{\boldsymbol{x}}_s$ that minimizes the cost function

$$\mathcal{E}(\hat{\boldsymbol{x}}_s) = [\boldsymbol{r} - \boldsymbol{C}(\hat{\boldsymbol{x}}_s)]^T \boldsymbol{W} [\boldsymbol{r} - \boldsymbol{C}(\hat{\boldsymbol{x}}_s)] \ . \tag{24.9}$$

LS methods can achieve the **maximum likelihood** (ML) estimate when the measurement noise vector is Gaussian with $\mathrm{E}[\boldsymbol{n}] = 0$ and equal variances, i.e., $\boldsymbol{\Sigma} = \sigma_n^2 \boldsymbol{I}$. For unequal variances, WLS with $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$ gives the ML estimate. In the following, $\boldsymbol{W} = \boldsymbol{I}$ will be assumed.

24.3.2 Location Solutions

As Eq. (24.4) indicates, $\boldsymbol{C}(\boldsymbol{x}_s)$ is a nonlinear function of the unknown parameter vector $\boldsymbol{x}_s$ so that the LS problem is a nonlinear one. One straightforward approach is to iteratively search for the minimum of the function using a **gradient descent method**. With this approach, an initial guess is made of the MS location and successive estimates are updated according to

$$\hat{\boldsymbol{x}}_s^{(k+1)} = \hat{\boldsymbol{x}}_s^{(k)} - \boldsymbol{\nu} \, \nabla \mathcal{E}\left(\hat{\boldsymbol{x}}_s^{(k)}\right) \tag{24.10}$$

where the matrix $\boldsymbol{\nu} = \mathrm{diag}(\nu_x, \nu_y)$ is the step size, $\hat{\boldsymbol{x}}_s^{(k)}$ is the estimate at time k, and $\nabla = \partial / \partial \boldsymbol{x}$ denotes the gradient vector with respect to the vector $\boldsymbol{x}$.

In order to mold the problem into a linear LS problem, the nonlinear function $\boldsymbol{C}(\boldsymbol{x}_s)$ can be linearized by using a Taylor series expansion about some reference point $\boldsymbol{x}_0$ so that

$$\boldsymbol{C}(\boldsymbol{x}_s) \approx \boldsymbol{C}(\boldsymbol{x}_0) + \boldsymbol{H}(\boldsymbol{x}_s - \boldsymbol{x}_0) \tag{24.11}$$

where $\boldsymbol{H}$ is the Jacobian matrix of $\boldsymbol{C}(\boldsymbol{x}_s)$. Then the LS solution can be formed as

$$\hat{\boldsymbol{x}}_s = \boldsymbol{x}_0 + \left(\boldsymbol{H}^T \boldsymbol{H}\right)^{-1} \boldsymbol{H}^T [\boldsymbol{r} - \boldsymbol{C}(\boldsymbol{x}_0)] \ . \tag{24.12}$$

This approach can be performed iteratively, with each successive estimate being closer to the final estimate. A key drawback to this approach is that an initial guess, $\boldsymbol{x}_0$, must be made of the MS's position.

The Taylor series approach introduces error when the linearized function $\boldsymbol{C}(\boldsymbol{x}_s)$ does not accurately approximate the nonlinear function. Other approaches have been developed for TDoA that avoid linearization by transforming the TDoA measurements into "pseudo-measurements." The pseudo-measurements are given by [5],

$$\boldsymbol{\varphi} = \boldsymbol{\Delta}\boldsymbol{x}_s + D_1(\boldsymbol{x}_s)\,\boldsymbol{r} \tag{24.13}$$

where

$$\boldsymbol{\Delta} = \begin{bmatrix} (\boldsymbol{x}_2 - \boldsymbol{x}_1)^T \\ \vdots \\ (\boldsymbol{x}_N - \boldsymbol{x}_1)^T \end{bmatrix} \qquad \boldsymbol{\varphi} = \frac{1}{2}\begin{bmatrix} \|\boldsymbol{x}_2\|^2 - \|\boldsymbol{x}_1\|^2 - \rho_{2,1}^2 \\ \vdots \\ \|\boldsymbol{x}_N\|^2 - \|\boldsymbol{x}_1\|^2 - \rho_{N,1}^2 \end{bmatrix}. \tag{24.14}$$

The term $D_1(\boldsymbol{x}_s)$ is nonlinear in the unknown vector $\boldsymbol{x}_s$ and can be removed by using a projection matrix that has $\boldsymbol{r}$ in its null space. A suggested projection is $\boldsymbol{P} = (\boldsymbol{I} - \boldsymbol{Z})[\mathrm{diag}(\boldsymbol{r})]^{-1}$ where $\boldsymbol{Z}$ is a circular shift matrix [5]. Projecting (24.13) with $\boldsymbol{P}$, the following linear equation results:

$$\boldsymbol{P}\boldsymbol{\varphi} = \boldsymbol{P}\boldsymbol{\Delta}\boldsymbol{x}_s \tag{24.15}$$

which leads to the following linear LS solution for the location of the MS

$$\hat{\boldsymbol{x}}_s = \left(\boldsymbol{\Delta}^T\boldsymbol{P}^T\boldsymbol{P}\boldsymbol{\Delta}\right)^{-1}\boldsymbol{\Delta}^T\boldsymbol{P}^T\boldsymbol{P}\boldsymbol{\varphi}\,. \tag{24.16}$$

24.4 Measures of Location Accuracy

To evaluate the performance of a location method, several benchmarks have been proposed. A common measure of accuracy is the comparison of the mean-squared-error (MSE) of the location estimate with the Cramér-Rao lower bound (CRLB) [10]. The concepts of circular error probability (CEP) [9] and geometric dilution of precision (GDOP) [6] have also been used as accuracy measures.

24.4.1 Cramér-Rao Lower Bound

For location in M dimensions, the MSE of the position estimate is given by

$$\mathrm{MSE} = \sqrt{\mathrm{E}\left[(\boldsymbol{x}_s - \hat{\boldsymbol{x}}_s)^T(\boldsymbol{x}_s - \hat{\boldsymbol{x}}_s)\right]} \tag{24.17}$$

where $\mathrm{E}[\,\cdot\,]$ denotes expectation. The calculated MSE is often compared to the theoretical minimum MSE given by the CRLB which sets a lower bound on the variance of any unbiased estimator. The CRLB is the inverse of the information matrix $\boldsymbol{J}$ defined as [10]

$$\boldsymbol{J} = \mathrm{E}\left[\left(\frac{\partial p(\boldsymbol{r}|\boldsymbol{x})}{\partial \boldsymbol{x}}\right)\left(\frac{\partial p(\boldsymbol{r}|\boldsymbol{x})}{\partial \boldsymbol{x}}\right)^T\right]\Bigg|_{\boldsymbol{x}=\boldsymbol{x}_s} \tag{24.18}$$

where $\boldsymbol{r}$ is the vector of TDoA, ToA, or AoA estimates and $p(\boldsymbol{r}|\boldsymbol{x})$ is the probability density function of $\boldsymbol{r}$ conditioned on the parameter vector $\boldsymbol{x}$. Assuming Gaussian measurement

noise, $p(\boldsymbol{r}|\boldsymbol{x})$ is Gaussian with mean $\boldsymbol{r}_0$ and covariance matrix $\boldsymbol{Q}$, and the CRLB reduces to

$$\text{CRLB} = \boldsymbol{J}^{-1} = c^2 \left(\frac{\partial \boldsymbol{r}_o^T}{\partial \boldsymbol{x}} \boldsymbol{Q}^{-1} \frac{\partial \boldsymbol{r}_0}{\partial \boldsymbol{x}^T} \right)^{-1} \Bigg|_{\boldsymbol{x}=\boldsymbol{x}_s} . \tag{24.19}$$

24.4.2 Circular Error Probability

A simple measure of accuracy is the CEP which is defined as the radius of the circle that has its center at the mean and contains half the realizations of a random vector. The CEP is a measure of the uncertainty in the location estimator $\hat{\boldsymbol{x}}_s$ relative to its mean $\mathrm{E}[\,\hat{\boldsymbol{x}}_s\,]$. If the location estimator is unbiased, the CEP is a measure of the estimator uncertainty relative to the true MS position. If the magnitude of the bias vector is bounded by B, then with a probability of one-half, a particular estimate is within a distance of B+CEP from the true position.

Because it is difficult to derive an exact expression for the CEP, an approximation that is accurate to within 10% is often used. The approximation for CEP is given as [9]

$$\text{CEP} \approx 0.75 \sqrt{\mathrm{E}\left[(\hat{\boldsymbol{x}}_s - \hat{\boldsymbol{\mu}})^T (\hat{\boldsymbol{x}}_s - \hat{\boldsymbol{\mu}}) \right]} \tag{24.20}$$

$$= 0.75 \sqrt{\sum_{i=1}^{M} \sigma^2_{\hat{x}_{s,i}}} \tag{24.21}$$

where $\hat{\boldsymbol{\mu}} = \mathrm{E}[\hat{\boldsymbol{x}}_s]$ is the mean location estimate and $\sigma^2_{\hat{x}_{s,i}}$ is the variance of the ith estimated coordinate, $i = 1, \ldots, M$.

24.4.3 Geometric Dilution of Precision

The GDOP provides a measure of the effect of the geometric configuration of the FSs on the location estimate. It is defined as the ratio of the *rms* position error to the *rms* ranging error [9, 6]. Hence, for an unbiased estimator, the GDOP is given by

$$\text{GDOP} = \frac{\sqrt{\mathrm{E}\left[(\hat{\boldsymbol{x}}_s - \hat{\boldsymbol{\mu}})^T (\hat{\boldsymbol{x}}_s - \hat{\boldsymbol{\mu}}) \right]}}{\sigma_r} \tag{24.22}$$

where σ_r denotes the fundamental ranging error for ToA and TDoA systems. For AoA, σ_r^2 is the average variance of the distance between each FS and a reference point near the true position of the MS.

The GDOP is an indicator of the extent to which the fundamental ranging error is magnified by the geometric relation between the MS and FSs. Furthermore, comparing (24.21) and (24.22), we find that the CEP and GDOP are related by

$$\text{CEP} \approx (0.75\sigma_r)\,\text{GDOP} . \tag{24.23}$$

The GDOP serves as a useful criterion for selecting the set of FSs from a large set to produce the minimum location error. In addition, it may aid cell site planning for cellular networks which plan to provide location services to their users.

24.5 Location in Cellular Systems

The location requirements set forth by the FCC [4] must be met not only by the new digital cellular systems, but the older analog system as well. In cellular networks, the BSs serve the role of the FSs in the algorithms of Section 24.3.1. With several different wireless systems on the market (**AMPS**, IS-54/136 **TDMA**, **GSM**, IS-95 **CDMA**), different methods may be necessary to implement location services in each of those systems. The signal strength method is often not implemented for cellular systems because of the large variability of received signal strength resulting from shadowing and **multipath fading** (see Section 24.6.1). AoA requires the placement of antenna arrays at the BSs which may be extremely costly. The AoA measurements can be obtained from array signal processing and is not dependent on the type of cellular system deployed. Unlike AoA, the ToA and TDoA methods require that timing information be obtained from the signals transmitted by a MS which may be implemented in different ways for each cellular system. The time-based methods may also require strict synchronization of the BSs, especially the TDoA approach. The remainder of this section discusses implementation strategies for the ToA and TDoA location methods in current and future generation cellular systems.

The most straightforward approach for obtaining timing information for ToA or TDoA location is the use of signal correlation methods. Specifically, maximizing cross-correlations between the signals received at pairs of BSs will provide an estimate of the TDoAs for each pair of BSs. Of course, this approach requires that the BSs be synchronized. These techniques are necessary for implementing a location system in AMPS since no system message parameters provide useful radiolocation information.

For CDMA, different methods can be used for the uplink (MS to BS) and downlink (BS to MS). On the uplink, the timing information for ToA or TDoA can be obtained using correlation techniques. Since the BSs in IS-95 are synchronized to a GPS time reference, the time of detection of the signal from the MS can serve as a ToA time stamp. Similarly, segments of the detected signal can be sent to a central processing office for cross-correlation in order to determine the set TDoAs for the BSs. The signals for the ToA/TDoA measurements can come from the reverse traffic channel or the access channel. The reverse traffic channel could be used for E-911 calls, for example, since a voice call must be initially made. For other location applications, the location may be desired when the MS is not actively transmitting. In these cases, the MS could be prompted to transmit messages on the access channel in response to commands from its serving BS on the paging channels. Unfortunately, it may be impossible to detect the MS transmissions at other BSs due to the **near–far effect** (see Section 24.6.3), although this problem can be alleviated by having the MS power-up to its maximum power for a short time. However, the use of the power up function must be limited to emergencies (such as E-911) in order to avoid excessive interference to other users.

An alternative for location in CDMA is to utilize pilot monitoring in the MS on the downlink. To assist in the handoff process, the MS monitors the strongest pilots from the surrounding BSs. The serving BS can send a pilot measurement request order (PMRO) causing the BS to respond with a message which includes the magnitudes of the pilots in the candidate set as well as the code phase of each pilot relative to its serving BS [2]. Hence, it is possible to construct TDoA estimates from these system messages. The accuracy of the TDoA estimates is dependent on the resolution of the code phase and the synchronization of the BSs. Fortunately, for IS-95, the BSs are synchronized to a GPS time reference. However, the code phase resolution is limited to a chip time, T_c, which implies a TDoA resolution of approximately 244 m. Finally, the **soft handoff**, during which the MS communicates with

nearby BSs during a handoff, can be used for location in CDMA systems as long as at least three BSs are in a soft handoff with the MS.

The TDMA-based systems also provide timing information in their system messages that can be used for ToA or TDoA location. The time alignment parameter in IS-54/136 and timing advance in GSM (both abbreviated TA) are used by each of those networks to ensure that the transmissions of MSs arrive at their serving BSs in the appropriate time slots. Each BS sends the MSs a TA value which is the amount the MS must advance or retard the timing of its transmissions. Additionally, the TA serves as a measure of the propagation time between the MS and BS. By artificially forcing the MS to handoff to two or more BSs, the location of the MS could be found using the ToA method. A primary consideration is the accuracy of the TA. For IS-54, the timing of MS transmissions are advanced or retarded in units of $T_b/2$, where $T_b = 20.6\,\mu s$ is the bit duration [1]. Hence, the TAs are accurate to $T_b/4$, or 1543 m. For GSM, the TA messages are reported in units of bits, with $T_b = 3.7\,\mu s$, which gives a TA resolution of $T_b/2$, or 554 m, in GSM [3].

An alternative for GSM is to use the observed time difference (OTD) measurements which are made at the MS without forcing additional handoffs. The OTDs are used to facilitate handoffs by estimating the amount the timing of the MS would have to be advanced or retarded if it were to be handed over to another BS. With a synchronized network, the OTDs could be used to implement a TDoA location system. Unfortunately, the GSM standard does not require that the network be synchronized. Additionally, the OTD measurements are made to the same accuracy of the TA measurements, 554 m.

Because of the high chip rate and good correlation properties of the spreading code sequences used in CDMA systems, these systems have greater potential than the other systems for accurate location estimates. It is apparent that the resolution of the timing parameters in the system messages needs to be improved in order to provide more accurate estimates of location.

24.6 Sources of Location Error

In all cellular systems, several factors can introduce error in the location estimates. Sources of error that are common to all cellular systems include multipath propagation, non-line-of-sight (NLoS) propagation and multiple-access interference (MAI). However, MAI poses a more significant problem in CDMA systems because of **power control** and the near–far effect. These effects are described below.

24.6.1 Multipath

Multipath propagation can introduce error in signal strength, AoA, ToA, and TDoA measurements. For signal-strength-based location systems, multipath fading and shadowing cause variations in the signal strength that can be as great as 30–40 dB over distances in the order of a half wavelength. Signal strength averaging can help, but low mobility MSs may not be able to average out the effects of multipath fading and there will still be the variability due to shadowing. The errors due to shadowing can be combated by using pre-measured signal strength contours that are centered at the BSs. However, this approach assumes a constant physical topography since shadows will change with the tree foliage, construction/destruction of structures, etc.

For AoA-based systems, scattering near and around the MS and BS will affect the measured AoA. Multipath will interfere with the angle measurement even when a LoS component is present. For macrocells, scattering objects are primarily within a small distance of

the MS and the BSs are usually elevated well above the local terrain. Consequently, the signals arrive with a relatively narrow AoA spread at the BSs. For microcells, the BSs may be placed below roof top level. Consequently, the BSs will often be surrounded by local scatterers such that the signals arrive at the BSs with a large AoA spread. Thus, the AoA method may be impractical for microcells.

In time-based location systems, the ToA or TDoA estimates can be in error even when there is a LoS path between the MS and BS. Conventional delay estimators, which are usually based on correlation techniques, are influenced by the presence of multipath fading. The result is a shift in the peak of the correlation away from the true value. Conventional delay estimators will detect a delay in the vicinity of these later arriving rays.

24.6.2 NLoS Propagation

With NLoS propagation, the signal transmitted from the MS (or BS) is reflected or diffracted and takes a path that is longer than the direct path or received at a different angle (Fig. 24.4). Obviously, the effect on an AoA system can be disastrous if the received AoAs are in a much different direction than the true AoAs. For the time-based systems, the measured distances can be considerably greater than true distances. For instance, for ToA location in the GSM system, the typical ranging error introduced by NLoS propagation can average 400–700 meters [7]. NLoS propagation will corrupt the ToA or TDoA measurements even when high resolution timing techniques are employed and even if there is no multipath fading.

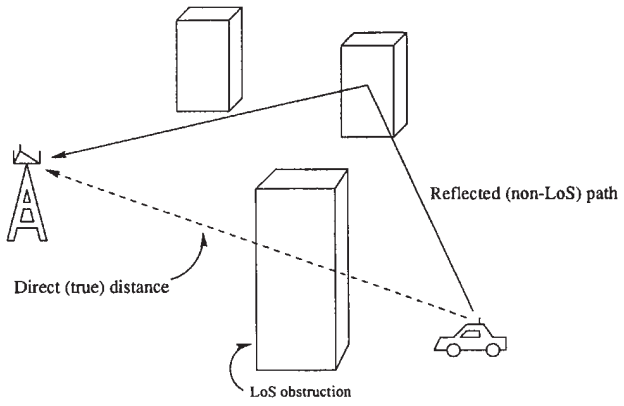


FIGURE 24.4 Propagation in a NLoS environment where signals are received from a reflection rather than a direct (LoS) path.

24.6.3 Multiple-Access Interference

All cellular systems suffer from cochannel interference. The transmissions of other users interfere with the signal of the desired user reducing the accuracy with which location measurements can be made. The problem is most evident in CDMA systems where the users share the same frequency band. As a result, signals from higher-powered users may mask the signals of the lower-powered users, a phenomenon known as the near–far effect. To combat the near–far effect, power control is used. However, for a location system where

multiple BSs must receive the transmission from the MS, the near–far problem still exists because the MS is not power-controlled to the other BSs. Consequently, the signal from the MS may not be detected at enough BSs to form a location estimate. As mentioned in Section 24.5, it may be possible, for instance in E-911 situations, for the MS to power up to maximum level and, therefore, mitigate the near–far effect. A further possibility is to take advantage of soft handoffs. However, the MS must be in position for a three-way soft handoff to be located.

24.7 Summary and Conclusions

This chapter has provided a brief introduction to radiolocation techniques for wireless systems. Several algorithms were developed for locating a MS using AoA, ToA, and TDoA, and measures of locator accuracy were described. For location services in mobile cellular networks, many possibilities exist. However, it is apparent that none has the current capability of providing high accuracy location estimates to meet the FCC requirements.

Defining Terms

AMPS (advanced mobile phone service): Analog cellular system in North America.

CDMA (code-division multiple-access): A technique for spread-spectrum multiple-access digital communications that separates users through the use of unique code sequences.

Gradient descent method: A minimization technique which searches for the minimum of an error surface by taking steps along the direction of greatest slope.

GSM (global system for mobile communications): Pan-European digital cellular standard.

Least squares estimation: A method whose estimate is chosen as the value that minimizes the sum of squares of the measured error.

Maximum likelihood estimation: A method whose estimate is chosen as the parameter value from which the observed data was most likely to come.

Multipath fading: Rapid fluctuation of the complex envelope of the received signal caused by reception of multiple copies of the transmitted signal, each with different amplitude, phase, and delay.

Near–far effect: A phenomenon that arises from unequal received power levels from the MSs. Stronger signals mask the weaker signals.

Path loss: Description of the attenuation of signal power with distance from a transmitter.

Power control: System for controlling the transmission power of the MS. Used to reduce cochannel interference and mitigate the near–far effect on the uplink.

Shadowing: Slow variation in the mean envelope over a distance corresponding to tens of wavelengths.

Soft handoff: Reception and transmission of radio signals between an MS and two or more BSs to achieve a macrodiversity gain.

TDMA (time-division multiple access): A form of multiple access giving each user a different time slot for transmission and reception of signals.

References

[1] EIA/TIA Interim Standard, IS-54., Cellular System Dual Mode Mobile Station–Land Station Compatibility Specifications, May 1990.

[2] EIA/TIA Interim Standard IS-95., Mobile Station–Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System, May 1992.

[3] ETSI-SMG., GSM 05.10 Technical Specification, Version 4.0.1, Apr. 1992.

[4] FCC Docket No.96-254., Report and Order and Further Notice of Proposed Rulemaking in the Matter of Revision of the Commission's Rules to Ensure Compatibility with Enhanced 911 Emergency Calling Systems, Jun. 12, 1996.

[5] Friedlander, B., A passive location algorithm and its accuracy analysis, *IEEE Journal of Oceanic Engineering,* 234–244, Jan. 1987.

[6] Massatt, P. and Rudnick, K., Geometric formulas for dilution of precision calculations, *Journal of the Institute of Navigation,* 37, 379–391, Winter, 1991.

[7] Silventoinen, M. and Rantalainen, T., Mobile station emergency locating in GSM, in *Proceedings of the IEEE International Conference on Personal Wireless Communications,* 232–238, 1996.

[8] Stilp, L., Carrier and end-user applications for wireless location systems. *Proc. SPIE,* 119–126, 1996.

[9] Torrieri, D.J., Statistical theory of passive location systems, *IEEE Trans. on Aerospace and Electronic Systems,* AES-20, 183–197, Mar. 1984.

[10] Van Trees, H.L., *Detection, Estimation and Modulation Theory, Part I,* John Wiley & Sons, New York, Ch. 2, 1968.

Further Information

More information on radiolocation can be found in the April, 1998, issue of *IEEE Communications Magazine,* which provided several articles regarding location service issues for wireless communications networks.

Articles discussing location techniques and algorithms can be found in many IEEE journals including *Transactions on Vehicular Technology, Transactions on Aerospace and Electronic Systems,* and the *Journal of Oceanic Engineering.*

Proceedings of various IEEE conferences such as the *Vehicular Technology Conference* document some of the latest developments for location in cellular networks.

25

Power Control

Roman Pichna
Nokia Telecommunications

Qiang Wang
Nikko Research Center

25.1 Introduction 25-1
25.2 Cellular Systems and Power Control 25-2
25.3 Power Control Examples 25-7
25.4 Summary 25-11
Defining Terms 25-12
References 25-12
Further Information 25-13

25.1 Introduction

The growing demand for mobile communications is pushing the technological barriers of wireless communications. The available spectrum is becoming crowded and the old analog **FDMA** (frequency division multiple access) cellular systems no longer meet the growing demand for new services, higher quality, and spectral efficiency. A second generation of digital cellular mobile communication systems are being deployed all around the world. The second generation systems are represented by three major standards: the **GSM,** IS-136, and IS-95. The first two are TDMA-based digital cellular systems and offer a significant increase in spectral efficiency and quality of service as compared to the first generation systems, e.g., **AMPS, NMT,** and **TACS.** IS-95 is based on **DS/CDMA** technology. The standardization of the third generation systems, IMT-2000, (formerly known as FPLMTS) is being pursued at ITU. Similar efforts are being conducted at regional standardization bodies.

The channel capacity of any cellular system is significantly influenced by the cochannel interference. To minimize the cochannel interference, several techniques are proposed: frequency reuse patterns, which ensure that the same frequencies are not used in adjacent cells; efficient **power control,** which minimizes the transmitted power; cochannel interference cancellation techniques; and orthogonal signalling (time, frequency, or code). All of these are being intensively researched, and some have already been implemented.

This chapter provides a short overview of power control. Since power control is a very broad topic, it is not possible to exhaustively cover all facets associated with power control. The interested reader can find additional information in the recommended reading that is appended at the end of this chapter.

The following section (Section 25.2) provides a brief introduction into cellular networks and demonstrates the necessity of power control. The various types of power control are presented in this section. The next section (Section 25.3) illustrates some applications of power control employed in various systems such as analog AMPS, GSM, DS/CDMA cellular

0-8493-8597-0/99/$0.00+$.50
© 1999 by CRC Press LLC

standard IS-95, and digital cordless telephone standard **CT2**. A glossary of definitions is provided at the end of the chapter.

25.2 Cellular Systems and Power Control

In cellular communication systems, the service area is divided into cells, each covered by a single base station. If, in the **forward link** (base station to mobile), all users served by all base stations share the same frequency, each communication between a base station and a particular user would also reach all other users in the form of cochannel interference. However, the greater the distance between the mobile and the interfering transmitter, the weaker the interference becomes due to the propagation loss. To ensure a good quality of service throughout the cell, the received signal in the fringe area of the cell must be strong. Once the signal has crossed the boundary of a cell, however, it becomes interference and is required to be as weak as possible. Since this is difficult, the channel frequency is usually not reused in adjacent cells in most of the cellular systems. If the frequency is reused, the cochannel interference impairs the signal reception in the adjacent cell, and the quality of service severely degrades unless other measures are taken to mitigate the interference. Therefore, a typical reuse pattern reuses the frequency in every seventh cell (frequency reuse factor = 1/7). The only exception is for CDMA-based systems where the users are separated by codes, and the allocated frequency may be shared by all users in all cells.

Even if the frequency is reused in every seventh cell, there is still some cochannel interference arriving at the receiver. It is, therefore, very important to maintain a minimal transmitted level at the base station to keep the cochannel interference low, frequency reuse factor high, and therefore the capacity of the system and quality of service high.

The same principle applies in the **reverse link** (mobile to base station)—the power control maintains the minimum necessary transmitted power for reliable communication. Several additional benefits can be gained from this strategy. The lower transmitted power conserves the battery energy allowing the mobile terminal (the portable) to be lighter and stay on the air longer. Furthermore, recent concerns about health hazards caused by the portable's electromagnetic emissions are also alleviated.

In the reverse link, the power control also serves to alleviate the near–far effect. If all mobiles transmitted at the same power level, the signal from a near mobile would be received as the strongest. The difference between the received signal strength from the nearest and the farthest mobile can be in the range of 100 dB, which would cause saturation of the weaker signals' receivers or an excessive amount of adjacent channel interference. To avoid this, the transmitted power at the mobile must be adjusted inversely proportional to the effective distance from the base station. The term effective distance is used since a closely located user in a propagation shadow or in a deep fade may have a weaker signal than a more distant user having excellent propagation conditions.

In a DS/CDMA system, power control is a vital necessity for system operation. The capacity of a DS/CDMA cellular system is interference limited since the channels are separated neither in frequency nor in time, and the cochannel interference is inherently strong. A single user exceeding the limit on transmitted power could inhibit the communication of all other users.

The power control systems have to compensate not only for signal strength variations due to the varying distance between base station and mobile but must also attempt to compensate for signal strength fluctuations typical of a wireless channel. These fluctuations are due to the changing propagation environment between the base station and the user as

the user moves across the cell or as some elements in the cell move. There are two main groups of channel fluctuations: slow (i.e., **shadowing**) and fast **fading.**

As the user moves away from the base station, the received signal becomes weaker because of the growing propagation attenuation with the distance. As the mobile moves in uneven terrain, it often travels into a propagation shadow behind a building or a hill or other obstacle much larger than the wavelength of the frequency of the wireless channel. This phenomenon is called shadowing. Shadowing in a land-mobile channel is usually described as a stochastic process having log-normal distributed amplitude. For other types of channels other distributions are used, e.g., Nakagami.

Electromagnetic waves transmitted from the transmitter may follow multiple paths on the way from the transmitter to the receiver. The different paths have different delays and interfere at the antenna of the receiver. If two paths have the same propagation attenuation and their delay differs in an odd number of half-wavelengths (half-periods), the two waves may cancel each other at the antenna completely. If the delay is an even multiple of the half-wavelengths (half-periods), the two waves may constructively add, resulting in a signal of double amplitude. In all other cases (nonequal gains, delays not a multiple of half-wavelength), the resultant signal at the antenna of the receiver is between the two mentioned limiting cases. This fluctuation of the channel gain is called fading. Since the scattering and reflecting surfaces in the service area are randomly distributed (buildings, trees, furniture, walls, etc.), the amplitude of the resulting signal is also a random variable. The amplitude of fading is usually described by a Rayleigh, Rice, or Nakagami distributed random variable.

Since the mobile terminal may move at the velocity of a moving car or even of a fast train, the rate of channel fluctuations may be quite high and the power control has to react very quickly in order to compensate for it.

The performance of the **reverse link** of DS/CDMA systems is most affected by the near–far effect and, therefore, very sophisticated power control systems in the reverse link that attempt to alleviate the effects of channel fluctuations must be used. Together with other techniques, such as micro- and macrodiversity, interleaving and coding, interference cancellation, multiuser detection, and adaptive antennae, the DS/CDMA cellular system is able to cope with the wireless channel extremely well.

The effective use of the **power control** in DS/CDMA cellular system enables the frequency to be reused in every cell, which in turn enables features such as the soft hand-off and base station diversity. All together, these help enhance the capacity of the system.

In the **forward link** of a DS/CDMA system, power control may also be used. It may vary the transmitted power to the mobile, but the dynamic range is smaller due to the shared spectrum and, thus, shared interference.

We can distinguish between two kinds of power control, the open-loop power control and the closed-loop power control. The open-loop power control estimates the channel and adjusts the transmitted power accordingly but does not attempt to obtain feedback information on its effectiveness. Obviously, the open-loop power control is not very accurate, but since it does not have to wait for the feedback information it may be relatively fast. This can be advantageous in the case of a sudden channel fluctuation, such as a mobile driving from behind a big building or in case it should provide only the initial or rough transmitted power setting.

The principle operation of open-loop power control is shown in Fig. 25.1. The open-loop power control must base its action on the estimation of the channel state. In the reverse link it estimates the channel by measuring the received power level of the pilot from the base station in the forward link and sets the transmitted power level inversely proportional to it. Estimating the power of pilot is, in general, more reliable than estimating the power of the

voice (or data) channel since the pilot is usually transmitted at higher power levels. Using the estimated value for setting the transmitted power ensures that the average power level received from the mobile at the base station remains constant irrespective of the channel variations. However, this approach assumes that the forward and the reverse link signal strengths are closely correlated. Although forward and reverse link may not share the same frequency and, therefore, the fading is significantly different, the long-term channel fluctuations due to shadowing and propagation loss are basically the same.

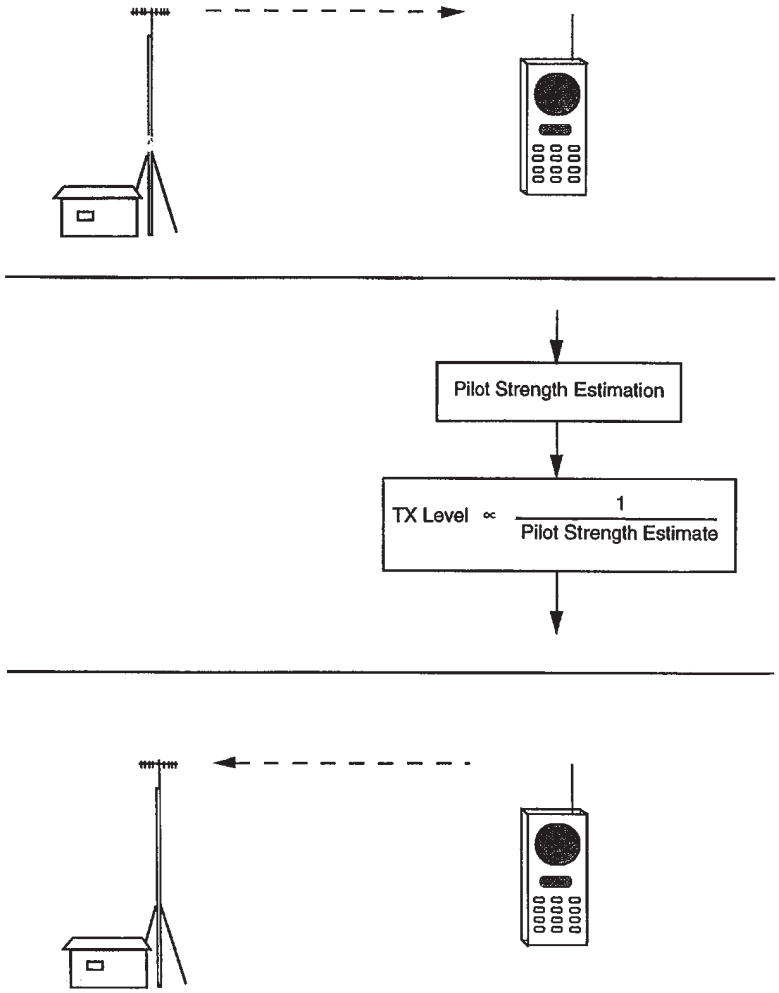


FIGURE 25.1 Reverse link open-loop power control.

The closed-loop power control system (Fig. 25.2a) may base its decision on an actual communication link performance metric, e.g., received signal power level, received signal-to-noise ratio, received bit-error rate, or received frame-error rate, or a combination of them. In the case of the reverse link power control, this metric may be forwarded to the mobile as a base for an autonomous power control decision, or the metric may be evaluated at the base station and only a power control adjustment command is transmitted to the mobile. If the reverse link power control decision is made at the base station, it may be based on

the additional knowledge of the particular mobile's performance and/or a group of mobiles' performance (such as mobiles in a sector, cell, or even in a cluster of cells). If the power control decision for a particular mobile is made at the base station or at the switching office for all mobiles and is based on the knowledge of all other mobile's performance, it is called a centralized power control system. A centralized power control system may be more accurate than a distributed power control system, but it is much more complex in design, more costly, and technologically challenging.

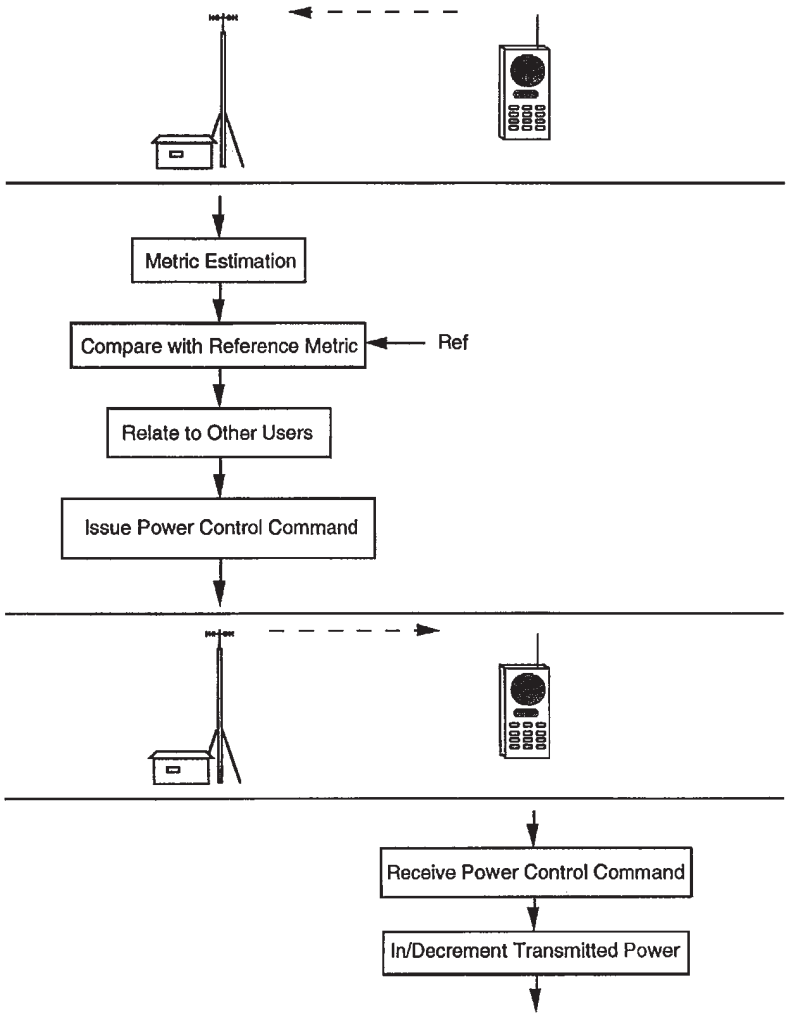


Figure 25.2a Reverse link closed-loop power control.

In principle, the same categorization may be used for the power control in the forward link (Fig. 25.2b) except that in the reverse link pilots from the mobiles are usually unavailable and only closed-loop power control is applied.

A special case for the design of power control are **TDD**-based systems [3]. In TDD systems, the forward and reverse link are highly correlated, and therefore a very good estimate of the channel gain in the forward link can be obtained from the estimate of

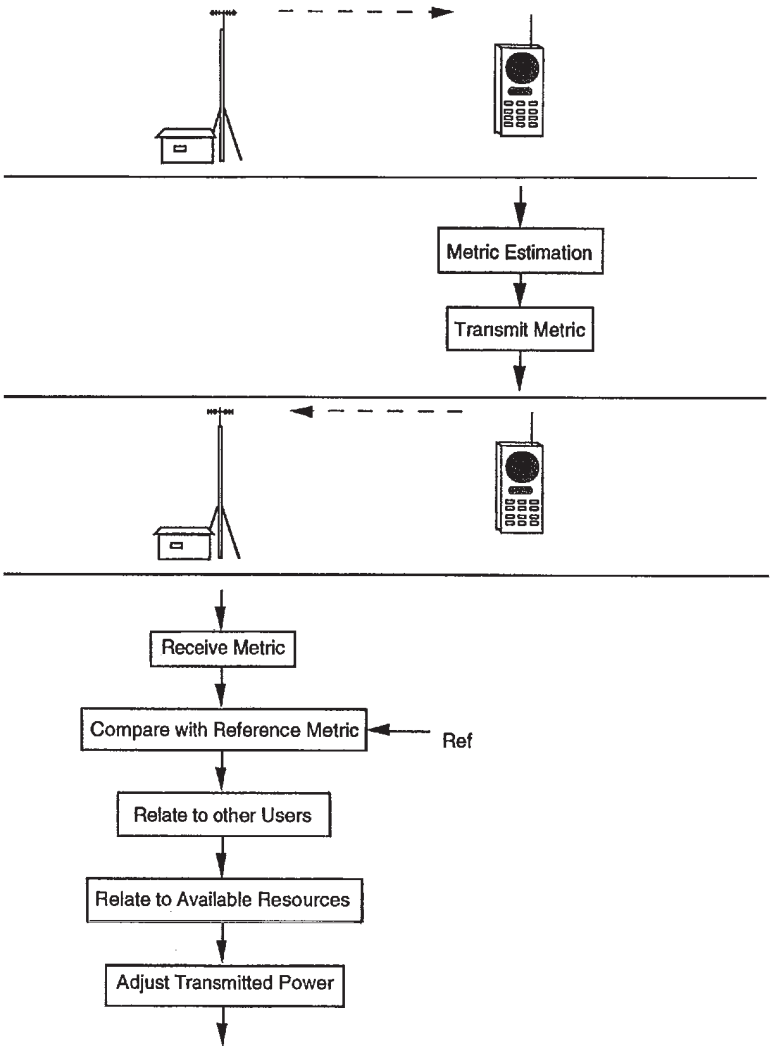


Figure 25.2b Forward link closed-loop power control.

the reverse link gain and vice versa. An open-loop power control then performs with the precision of a closed-loop power control but much faster since no feedback information has to be transmitted.

In the ideal case, power control compensates for the propagation loss, shadowing, and fast fading. However, there are many effects that prevent the power control from becoming ideal. Fast fading rate, finite delays of the power control system, nonideal channel estimation, error in the power control command transmission, limited dynamic range, etc., all contribute to degrading the performance of the power control system. It is very important to examine the performance of power control under nonideal conditions since the research done has shown that the power control system is quite sensitive to some of these conditions [11]. Kudoh [5] simulated a nonideal closed-loop power control system. Errors in the system were represented by a log-normal distributed control error with standard deviation σ_E (dB). Some results on capacity reduction are presented in Table 25.1.

TABLE 25.1 Capacity Reduction versus Power Control Error

	$\sigma_E = 0.5$ dB, %	$\sigma_E = 1$ dB, %	$\sigma_E = 2$ dB, %	$\sigma_E = 3$ dB, %
Forward link	10	29	64	83
Reverse link	10	31	61	81

Source: Kudoh, E., On the capacity of DS/CDMA cellular mobile radios under imperfect transmitter power control. *IEICE Trans. Commun.*, E76-B, 886–893, Apr. 1993.

The authors have also studied the effects of Doppler and delay and feedback errors in power control loop on power control [8].

25.3 Power Control Examples

In the following section, several applications of power control of analog and digital cellular systems are presented.

In the analog networks we may see power control implemented in both the reverse link and forward link [6]. Power control in the reverse link:

- reduces the chance of receiver saturation by a closely located mobile
- reduces the cochannel interference and thus increases the frequency-reuse factor and capacity, and
- reduces the average transmitted power at the mobile thus conserving battery energy at the mobile.

The power control in the forward link:

- reduces cochannel interference and thus increases the frequency reuse factor and capacity and
- reduces adjacent-channel interference and improves the quality of service.

One example of a power control system shown by Lee [7] was of an air-to-ground communication system. The relevant airspace is divided into six zones based on the aircraft altitude. The transmitted power at the aircraft is then varied in six steps based on the zone in which the aircraft is located. The power control system exhibits a total of approximately 28 dB of dynamic range. This reduces the cochannel interference and, due to the excellent propagation conditions in the free air, has a significant effect on the capacity of the system.

Another example of a power control system in an analog wireless network is in the analog part of the TIA standard IS-95 [10]. IS-95 standardizes a dual-mode FDMA/CDMA cellular system compatible with the present day AMPS analog FDMA cellular system.

The analog part of IS-95 divides the mobiles into three classes according to nominal **ERP** (effective radiated power with respect to half-wave dipole) at the mobile. For each class, the standard specifies eight power levels. Based on the propagation conditions, the mobile station may receive a power control command that specifies at what power level the mobile should transmit. The maximum change is 4 dB per step. (See Table 25.2).

IS-95 supports further discontinuous transmission. This feature allows the mobile to vary its transmitted power between two states: low and high. These two states must be at least 8 dB apart.

As for the power control in a digital wireless system, three examples will be shown: GSM [1], CT2/CT2PLUS standard [2] for digital cordless telephones of second generation, and the IS-95 standard for digital cellular DS/CDMA system [10].

TABLE 25.2 Nominal ERP of the Mobile

	Nominal ERP (dBW) of mobile		
Power level	I	II	III
0	6	2	−2
1	2	2	−2
2	−2	−2	−2
3	−6	−6	−6
4	−10	−10	−10
5	−14	−14	−14
6	−18	−18	−18
7	−22	−22	−22

Source: Telecommunications Industry Association/Electronic Industries Association. Mobile Station-Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System, *TIA/EIA/IS-95 Interim Standard,* Jul. 1993.

GSM is a Pan-European digital cellular system that was introduced in many countries during the 1992–1993 period. GSM is a digital TDMA system with a frequency hopping feature. The power control in GSM ensures that the mobile station uses only the minimum power level necessary for reliable communication with the base station. GSM defines eight classes of base stations and five classes of mobiles according to their power output, as shown in Table 25.3.

TABLE 25.3 GSM Transmitter Classes

Power Class	Base Station Power (W)	Mobile Station power (W)
1	320	20
2	160	8
3	80	5
4	40	2
5	20	0.8
6	10	
7	5	
8	2.5	

Source: Balston, D.M. and Macario, R.C.V., *Cellular Radio Systems,* Artech House, Norwood, MA, 1993.

The transmitted power at the base station is controlled, nominally in 2-dB steps. The adjustment of the transmitted power reduces the intercell interference and, thus, increases the frequency reuse factor and capacity. The transmitted power at the base station may be decremented to a minimum of 13 dBm.

The power control of the mobile station is a closed-loop system controlled from the base station. The power control at the mobile sets the transmitted power to one of 15 transmission power levels spaced by 2 dB. Any change can be made only in steps of 2 dB during each time slot. Another task for the power control in GSM is to control graceful ramp-on and ramp-off of the TDMA bursts since too steep slopes would cause spurious frequency emissions.

The dynamic range of the received signal at the base station may be up to 116 dB [1] and, thus, the near–far problem may also by experienced, especially if the problem occurs in adjacent time slots. In addition to power control, a careful assignment of adjacent slots can also alleviate the near–far effect.

The CT2PLUS standard [2] is a Canadian enhancement of the **ETSI** CT2 standard. Both these standards allow power control in the forward and in the reverse link. Due to the expected small cell radius and relatively slow signal level fluctuation rate, given by the fact that the user of the portable is a pedestrian, the power control specifications are relatively simple. The transmission at the portable can have two levels: *normal* (full) and *low*. The low–normal difference is up to 20 dB.

The IS-95 standard represents a second generation digital wireless cellular system using direct-sequence code division multiple access (DS/CDMA). Since in a DS/CDMA system all users have the same frequency allocation, the cochannel interference is crucial for the performance of the system [4]. The near–far effect may cause the received signal level to change up to 100 dB [12]. This considerable dynamic range is disastrous for a DS/CDMA where the channels are separated by a finite correlation between spreading sequences. This is further aggravated by the shadowing and the fading. The fading may have a relatively high rate since the mobile terminal is expected to move at the speed of a car. Therefore, the power control system must be very sophisticated. Power control is employed in both the reverse link and in the forward link.

The reverse link power control serves to do the following:

- Equalizes the received power level from all mobiles at the base station. This function is vital for system operation. The better the power control performs, the more it reduces the cochannel interference and, thus, increases the capacity. The power control compensates for the near–far effect, shadowing, and partially for slow fading.
- Minimizes the necessary transmission power level to achieve good quality of service. This reduces the cochannel interference, which increases the system capacity and alleviates health concerns. In addition, it saves the battery power. Viterbi [12] has shown up to 20–30 dB average power reduction compared to the AMPS mobile user as measured in field trials.

The forward link power control serves to:

- Equalize the system performance over the service area (good quality signal coverage of the worst-case areas),
- Provide load shedding between unequally loaded cells in the service areas (e.g., along a busy highway) by controlling the intercell interference to the heavy loaded cells, and
- Minimize the necessary transmission power level to achieve good quality of service. This reduces the cochannel interference in other cells, which increases the system capacity and alleviates health concerns in the area around the base station.

The reverse link power control system is composed of two subsystems: the closed-loop and the open-loop. The system operates as follows. Prior to the application to access, closed-loop power control is inactive. The mobile estimates the mean received power of the received pilot from the base station and the open-loop power control estimates the mean output power at the access channel [10]. The system then sets the closed-loop probing and

estimates the mean output power,

$$\begin{aligned}\text{mean output power (dBm)} = &-\text{mean input power (dBm)}\\ &-73\\ &+\text{NOM\_PWR (dB)}\\ &+\text{INIT\_PWR (dB)}\end{aligned} \tag{25.1}$$

where NOM_PWR and INIT_PWR are parameters obtained by the mobile prior to transmission. Subsequent probes are sent at increased power levels in steps until a response is obtained. The initial transmission on the reverse traffic channel is estimated as

$$\begin{aligned}\text{mean output power (dBm)} = &-\text{mean input power (dBm)}\\ &-73\\ &+\text{NOM\_PWR (dB)}\\ &+\text{INIT\_PWR (dB)}\\ &+\text{the sum of all access probe}\\ &\quad\text{corrections (dB)}\end{aligned} \tag{25.2}$$

Once the first closed-loop power control bit is received the mean output power is estimated as

$$\begin{aligned}\text{mean output power (dBm)} = &-\text{mean input power (dBm)}\\ &-73\\ &+\text{NOM\_PWR (dB)}\\ &+\text{INIT\_PWR (dB)}\\ &+\text{the sum of all access probe corrections (dB)}\\ &+\text{the sum of all closed-loop power control}\\ &\quad\text{corrections (dB)}\end{aligned} \tag{25.3}$$

The ranges of the parameters NOM_PWR and INIT_PWR are shown in Table 25.4.

TABLE 25.4 NOM_PWR and INIT_PWR Parameters

	Nominal value, dB	Range, dB
NOM_PWR	0	−8–7
INIT_PWR	0	−16–15

Source: Telecommunications Industry Association/Electronic Industries Association. Mobile Station-Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System. *TIA/EIA/IS-95 Interim Standard*, Jul. 1993.

The closed-loop power control command arrives at the mobile every 1.25 ms (i.e., 800 b/s). Therefore, the base station estimates the received power level for approximately 1.25 ms. A closed-loop power control command can have only two values: 0 to increase the power level and 1 to decrease the power level. The mobile must respond to the power control

command by setting the required transmitted power level within 500 μs. The total range of the closed-loop power control system is ± 24 dB. The total supported range of power control (closed-loop and open-loop) must be at least ± 32 dB.

The behavior of the closed-loop power control system while the mobile receives base station diversity transmissions is straightforward. If all diversity transmitting base stations request the mobile to increase the transmitted power (all power control commands are 0), the mobile increases the power level. If at least one base station requests the mobile to decrease its power, the mobile decreases its power level.

The system also offers a feature of gated transmitted power for variable rate transmission mode. The gate-off state reduces the output power by at least 20 dB within 6 μs. This reduces the interference to the other users at the expense of transmitted bit rate. This feature may be used together with variable rate voice encoder or voice activated keying of the transmission.

The forward link power control works as follows. The mobile monitors the errors in the frames arriving from the base station. It reports the frame-error rate to the base station periodically. (Another mode of operation may report the error rate only if the error rate exceeds a preset threshold.) The base station evaluates the received frame-error rate reports and slightly adjusts its transmitting power. In this way, the base station may equalize the performance of the forward links in the cell or sector.

A system conforming with the standard has been field tested, and the results show that the power control is able to combat the channel fluctuation (together with other techniques such as RAKE reception) and achieve the bit energy to interference power density (E_b/I_0) necessary for a reliable service [12].

Power control together with soft handoff determines the feasibility of the DS/CDMA cellular system and is crucial to its performance. QUALCOMM, Inc. has shown on field trials that their system conforms with the theoretical predictions and surpasses the capacity of other currently proposed cellular systems [12].

25.4 Summary

We have shown the basic principles of power control in wireless cellular networks and have presented some examples of power control systems employed in some networks.

In a wireless channel the channel transmission or channel gain is a random variable. If all transmitters in the system transmitted at equal and constant power levels, the received powers would be random.

In the reverse link (mobile to base station) each user has its own wireless channel, generally uncorrelated with all other users. The received signals at the base station are independent and random. Furthermore, since the users are randomly distributed over the cell, the distance between the mobiles and the base station may vary and so does the propagation loss. The differences between the strongest and the weakest received signal level may approach the order of 100 dB. This power level difference may cause saturation of the receivers at the base station even if they are allocated a different frequency or time slot. This phenomenon is called the near–far effect.

The near–far effect is especially detrimental for a DS/CDMA system where the frequency band is shared by all users and, for any given user, all other users' transmissions form the cochannel interference. Therefore, for the DS/CDMA system it is vitally important to efficiently mitigate the near–far effect.

The most natural way to mitigate the near–far effect is to power control the transmission in such a way that the transmitted power counterfollows the channel fluctuations and

compensates for them. Then the received signal at the base station arrives at a constant amplitude.

The use of power control is not limited to the reverse link, but is also employed in the forward link. The controlled transmission maintaining the transmitted level at the minimum acceptable level reduces the cochannel interference, which translates into an increased capacity of the system.

Since the DS/CDMA systems are most vulnerable to the near–far effect they have a very sophisticated power control system. In giving examples, we have concentrated on the DS/CDMA cellular system. We have also shown the power control used in other systems such as GSM, AMPS, and CT2.

Although there are more techniques available for mitigation of the near–far effect, power control is the most efficacious. As such, power control forms the core in the effort in combatting the near–far effect and channel fluctuations in general [12].

Defining Terms

AMPS: Advanced Mobile Phone Service. Analog cellular system in North America.

CT2: Cordless Telephone, Second Generation. A digital FDMA/TDD system.

DS/CDMA: Direct Sequence Code Division Multiple Access.

DOC: Department of Communications.

ERP: Effective Radiated Power.

ETSI: European Telecommunications Standard Institute.

Fading: Fast varying fluctuations of the wireless channel mainly due to the interference of time-delayed multipaths.

FDMA: Frequency Division Multiple Access.

Forward link: Link from the base (fixed) station to the mobile (user, portable).

GSM: Groupe Spéciale Mobile, recently referred to as the **Global System for Mobility**. An ETSI standard for digital cellular and microcellular systems.

NMT: Nordic Mobile Telephone. A cellular telephony standard used mainly in Northern Europe.

Power control: Control system for controlling the transmission power. Used to reduce the cochannel interference and mitigate the near–far effect in the reverse link.

Reverse link: Link from the mobile (user, portable) to the base (fixed) station.

Shadowing: Slowly varying fluctuations of the wireless channel due mainly to the shades in propagation of electromagnetic waves. Often described by log-normal probability density function.

TACS: Total Access Communication System. An analogue cellular system used mainly in UK.

TDD: Time Division Duplex.

TDMA: Time Division Multiple Access.

References

[1] Balston, D.M. and Macario, R.C.V., *Cellular Radio Systems,* Artech House, Norwood, MA, 1993.

[2] Department of Communications. ETI Interim Standard # I-ETS 300 131, Annex 1, Issue 2, Attachment 1. In *CT2PLUS Class 2: Specification for the Canadian Common Air Interface for Digital Cordless Telephony, Including Public Access Services, RS-130.* Communications Canada, Ottawa, ON, 1993.

[3] Esmailzadeh, R., Nakagawa, M., and Sourour, E.A., Time-division duplex CDMA communications, *IEEE Personal Comm.*, 4(3), 51–56, Apr. 1997.

[4] Gilhousen, K.S., Jacobs, I.S., Padovani, R, Viterbi, A.J., Weaver, L.A., and Wheatley III, C.E., On the capacity of cellular CDMA system. *IEEE Trans. Veh. Tech.*, 40, 303–312, May 1991.

[5] Kudoh, E., On the capacity of DS/CDMA cellular mobile radios under imperfect transmitter power control. *IEICE Trans. Commun.*, E76-B, 886–893, Apr. 1993.

[6] Lee, W.C.L., *Mobile Cellular Telecommunications Systems,* McGraw-Hill, New York, 1989.

[7] Lee, W.C.L., *Mobile Communications Design Fundamentals,* 2nd ed., John Wiley & Sons, New York, 1993.

[8] Pichna, R., Kerr, R., Wang, Q., Bhargava, V.K., and Blake, I.F., *CDMA Cellular Network Analysis Software.* Final Rep. Ref. No. 36-001-2-3560/01-ST, prepared for Department of Communications, Communications Research Centre, Ottawa, ON, Mar. 1993.

[9] Simon, M.K., Omura, J.K., Scholtz, R.A., and Levitt, B.K., *Spread Spectrum Communication Handbook,* McGraw-Hill, New York, 1994.

[10] Telecommunications Industry Association/Electronic Industries Association. Mobile Station-Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System. *TIA/EIA/IS-95 Interim Standard,* Jul. 1993.

[11] Viterbi, A.J. and Zehavi, E., Performance of power-controlled wideband terrestrial digital communication. *IEEE Trans. Comm.*, 41, 559–569, Apr. 1993.

[12] Viterbi, A.J., The orthogonal-random waveform dichotomy for digital mobile personal communication. *IEEE Personal Comm.*, 1(1st qtr.), 18–24, 1994.

Further Information

For general information see the following overview books:

[1] Balston, D.M. and Macario, R.C.V., *Cellular Radio Systems,* Artech House, Norwood, MA, 1993.

[2] Simon, M.K., Omura, J.K., Scholtz, R.A., and Levitt, B.K., *Spread Spectrum Communication Handbook,* McGraw-Hill, New York, 1994.

For more details on power control in DS/CDMA systems consult the following:

[3] Gilhousen, K.S., Jacobs, I.S., Padovani, R., Viterbi, A.J., Weaver, L.A., and Wheatley C.E., III., On the capacity of cellular CDMA system. *IEEE Trans. Veh. Tech.*, 40, 303–312, May 1991.
or:

[4] Viterbi, A.J. and Zehavi, E., Performance of power-controlled wideband terrestrial digital communication. *IEEE Trans. Comm.*, 41, 559–569, Apr. 1993.

Readers deeply interested in power control are recommended the *IEEE Transactions on Communications, IEEE Transactions on Vehicular Technology,* and relevant issues of *IEEE Journal on Selected Areas in Communications.*

26

Enhancements in Second Generation Systems

26.1 Introduction 26-1
26.2 Overview of Second Generation Systems 26-1
26.3 Capacity Enhancement 26-2
Capacity Enhancement Through Increase in Number of Carriers and/or Voice Circuits • Improved Interference Management Techniques • Novel Cellular Configurations
26.4 Quality Enhancement 26-8
Quality Aspects and Definitions • Speech Quality Enhancements • Coverage Quality Enhancements
26.5 High Bit Rate Data Transmission 26-13
Circuit Mode Techniques • Packet Mode Techniques • New Modulation Schemes
26.6 Conclusion 26-16
Defining Terms 26-16
References 26-17
Further Information 26-18

Marc Delprat
Alcatel Mobile Communication Division

Vinod Kumar
Alcatel Mobile Communication Division

26.1 Introduction

Present digital **cellular** and **cordless** systems were optimized for voice services. At the very best, they can provide low and medium bit rate information services. The development of enhanced versions of these radio interfaces is motivated by the following:

- Provision of additional services, including high bit rate circuit-switched and packet-switched services that meet the short-term needs of mobile multimedia services.
- Improvements in the radio coverage of existing cellular systems. An extension to allow cordless coverage in the home is also being considered.
- Even more efficient utilization of the available frequency spectrum, which is a valuable but limited resource.

This chapter deals mainly with the air interface of second generation systems. The subject of enhancement in system performance is addressed from two directions. On the one hand, system features like adaptive multirate coders, packet transmission, which have been explicitly included in the standard, are discussed in detail. On the other hand, methods

0-8493-8597-0/99/$0.00+$.50

© 1999 by CRC Press LLC

of equipment design or network design possible with the new or already existing features of the air interface are presented.

26.2 Overview of Second Generation Systems

Initially the need of a pan-European system to replace a large variety of disparate analog cellular systems was the major motivating factor behind the creation of the Global System for Mobile communications (GSM). In North America and Japan, where unique analog systems existed, the need to standardize respectively IS-54, IS-95, and Personal Digital Cellular (PDC) for digital cellular applications arose from the lack of spectrum to serve the high traffic density areas [3]. Additionally, some of the second generation systems like Digital European Cordless Telecommunications (DECT) and Personal Handy Phone Systems (PHS) are the result of a need to offer wireless services in residential and office environments with low cost subscriber equipment [16].

The physical layer characteristics of all these systems offer robust radio links paired with good spectral efficiency. The network related functionalities have been designed to offer secure communication to authenticated users even when roaming between various networks based on the same system. Table 26.1 provides the essential characteristics of the second generation systems as initially designed in the late eighties and early nineties. Since then several additions have been made to those standards. Enhancements of air interface as well as network subsystem functionalities have been incorporated.

26.3 Capacity Enhancement

The **capacity** of a mobile network can be defined as the Erlangs throughput by a cell, a cluster of cells, or by a portion of a network. For a given radio interface, the achievable capacity is a function of the robustness of the physical layer, the effectiveness of the medium access control (MAC) layer and the multiple access technique. Moreover, it is strongly dependent on the radio spectrum available for network planning.
If we define:

BW	Available radio spectrum
Wc	Spectrum occupied by a single radio carrier
Nc	Number of circuits handled by a single carrier (e.g., number of time slots per frame in a TDMA system)
Cs	Permissible cluster size which guarantees a good quality for a vast majority of active calls (like 90 to 95%)

The number of circuits per cell is given by (BW / Wc) $\times$ (Nc / Cs).

The Erlang capacity can be derived from this expression after consideration of the signalling overhead required by the pilot channel, signalling and traffic channel overheads for handover, and the specified call blocking rate.

This definition of capacity is applicable in case of TDMA/FDMA air interfaces and when the networks are designed using "deterministic frequency allocation." A slightly different approach to the problem is necessary in systems like DECT which use dynamic channel selection and for DS-CDMA systems. IS-95 networks pretend to use a cluster size (Cs) of one. However, the number of good quality circuits in a cell is a function of the available noninterfered spreading codes.
Capacity enhancement can be obtained through:

TABLE 26.1 Air Interface Characteristics of Second Generation Systems

	CELLULAR				CORDLESS	
Standard	GSM	IS-54	IS-95	PDC	DECT	PHS
Frequency band (MHz)	Europe	USA	USA	Japan	Europe	Japan
Uplink	890–915 (1710–1785)	824–849	824–849	940–956 (1429–1441)	1880–1900	1895–1907
Downlink	935–960 (1805–1880)	869–894	869–894	810–826 (1477–1489)		
Duplex spacing (MHz)	45 (95)	45	45	130 (48)	—	—
Carrier spacing (kHz)	200	30	1250	25	1728	300
Number of radio channels in the frequency band	124 (374)	832	20	640 (480)	10	77
Multiple access	TDMA	TDMA	CDMA	TDMA	TDMA	TDMA
Duplex mode	FDD	FDD	FDD	FDD	TDD	TDD
Number of channels per carrier	8 (half rate: 16)	3 (half rate: 6)	128	3 (half rate: 6)	12	4
Modulation	GMSK	Π/4 DQPSK	QPSK BPSK	Π/4 DQPSK	GFSK	Π/4 DQPSK
Carrier bit rate (kb/s)	270.8	48.6	1288	42	1152	384
Speech coder (full rate)	RPE-LTP	VSELP	QCELP	VSELP	ADPCM	ADPCM
Net bit rate (kb/s)	13	7.95	(var.rate: 8, 4, 2, 0.8)	6.7	32	32
Channel coder for speech channels	1/2 rate convolutional + CRC	1/2 rate convolutional + CRC	1/2 (downlink), 1/3 (uplink) convolutional + CRC	1/2 rate convolutional + CRC	no	no
Gross bit rate speech+channel coding (kb/s)	22.8	13	var. rate 19.2, 9.6, 4.8, 2.4	11.2	—	—
Frame size (ms)	4.6	40	20	20	10	5
MS transmission power (W)	Peak 8 2 (1) / Aver. 1 0.25 (0.125)	Peak 9 4.8 1.8 / Aver. 3 1.6 0.6	0.6	Peak 2 / Aver. 0.66	Peak 0.25 / Aver. 0.01	Peak 0.08 / Aver. 0.01
Power control MS	Y	Y	Y	Y	N	Y
Power control BS	Y	Y	Y	Y	N	Y
Operational C/I(dB)	9	16	6	17	21	26
Equalizer	needed	needed	Rake receiver	option	option	no
Handover	Y	Y	Soft handoff	Y	Y	Y

- An increase in number of radio carriers and/or the number of traffic channels (e.g., voice circuits)
- Improved interference management techniques
- Novel cellular configurations.

26.3.1 Capacity Enhancement Through Increase in Number of Carriers and/or Voice Circuits

This family of relatively simple solutions can be subdivided in three categories related to the availability of spectrum in the same band, or in a different band where the same radio interface can be utilized or the situation in which the number of voice circuits can be increased by the introduction of multiple speech codecs.

Increased Spectrum Availability

In such a fortunate situation, extra carriers can be added to the already installed base stations (BS). The existing Cs is maintained and the increase in capacity is according to the added carriers/circuits. An additional gain in Erlang capacity is available due to increased trunking efficiency.

Supplementary Spectrum Availability (in Different Frequency Bands)

Most of the second generation systems were implemented in the 800 and 900 MHz frequency bands. Their application has been extended to the 1800 and 1900 MHz bands too (e.g., GSM 1800 and PCS1900). Carriers from both the bands can be used at the same BSs—either according to a common or independent frequency reuse scheme. Increase in traffic throughput can be maximized by using dual band mobile stations (MSs) and by providing adequate handover mechanisms between the carriers of two bands. Also, due to difference in propagation for carriers from two widely separated bands, certain adjustments related to coverage planning might be required when the carriers are co-sited.

Multiple Speech Codecs for Increased Number of Voice Circuits

Originally, only full rate (FR) speech codecs were used in second generation systems. Hence, a one-to-one correspondence between the physical channels on the air interface and the available voice circuits was established. If the number of installed carriers in a cell is kept unchanged, the introduction of half rate (HR) codecs will double the number of voice circuits in the cell and a more than two-fold increase in Erlang capacity in the cell will be achievable. A similar possibility is offered by the adaptive multirate codec (AMR). The output bit rates of speech codec and channel codec can be adapted in order to minimize the carrier occupancy time necessary to offer a predetermined call quality. The statistical multiplexing gain thus obtained can be exploited to enhance the traffic throughput.

Such capacity enhancement methods can be implemented only if corresponding MSs or MSs with multiple codec capabilities are commercialized. Moreover, every cell site in a network will have to maintain a certain number of FR voice circuits necessary to offer downwards compatibility. Table 26.2 provides an applicability matrix related to the above mentioned capacity enhancement methods.

TABLE 26.2 Applicability Matrix of Methods for Capacity Enhancements Based on Increased Spectrum Availability

Second Generation System	Method for Capacity Enhancement		
	Additional Spectrum in Same Band	Additional Spectrum in Another Band	Multiple Speech Codecs
GSM Family	Available	Under trial in real networks	Under trial in real networks
IS-54/IS-136	Available	Applicable	Applicable
PDC	Available	Applicable	Applicable
IS-95 Family using DS-CDMA	Some of these solutions are applicable. However, their implementation shall adversely effect the functioning of soft hand over which is an essential feature of DS-CDMA networks.		
DECT	Available	Under Consideration	Under Consideration
PHS	Available	Not Applicable	Available

26.3.2 Improved Interference Management Techniques

Initially, the design of a cellular network is based on some simplifying assumptions like uniformly distributed subscriber density and traffic patterns or homogeneous propagation conditions. Usually, such design results in a worst case value of Cs. The situation can be improved through better managemnt of cochannel interference by implementing one or a combination of the following:

- Slow Frequency Hopping (SFH)
- Voice Activity Detection (VAD) and Discontinuous Transmission (DTx)
- Transmit Power Control (PC)
- Antenna beam forming

Slow Frequency Hopping

Every call is "spread" over all the carrier frequencies available in the cell. To avoid intracell interference, orthogonal (random or pseudo random) frequency hopping laws are used for different calls. Also, the worst-case intercell interference situation can last only one hop and an averaging out of cochannel interference in the network occurs. Statistically, the distribution of carrier-to-cochannel interference ratio in the network is more compact with frequency hopping than without it, and this can be exploited to reduce the cluster size Cs and increase capacity.

Voice Activity Detection (VAD) and Discontinuous Transmission (DTx)

Collection of statistics related to telephone conversations has demonstrated that the duty cycle of voice sources is around 40%. With DTx, the radio signal is transmitted according to the activity of voice source. An overall reduction in the averaged cochannel interference experienced by the calls can thus be observed. This offers the possibility of implementing smaller Cs. Also, saving energy with DTx in the up-link (UL) results in a prolonged autonomy for the MS.

Transmit Power Control

Usually, the full available transmit power is necessary for the initial access and for a short duration after call establishment. For the rest of the call, both DL and UL transmit powers can be reduced to a level necessary to maintain a good link quality. The overall

improvement in radio interference in the network thus obtained is helpful for the reduction of cluster size. Like VAD/DTx, the DL transmit power control is not permitted for pilot carriers if the mobile assisted handover is implemented in the system.

Antenna Beamforming

Interference related to every call can be individually managed by "dynamic cell sectorization." Actually, the base station receiver captures the up-link signal in a narrow antenna beam dynamically "placed" around the MS. Similarly, the down-link signal is transmitted in a beam focused towards the MS. This sort of spatial filtering of cochannel interference is useful for implementing very compact frequency reuse schemes. Generally, both the down-link and up-link beamforming capabilities are placed at the BS where an antenna array and signal processing algorithms related to direction finding, signal source separation, and beam synthesis have to be implemented. Table 26.3 provides an applicability matrix for the second generation systems.

TABLE 26.3 Applicability Matrix of Methods for Capacity Enhancements Based on Improved Interference Management

Second Generation System	Method for Capacity Enhancement			
	Slow Frequency Hopping	VAD/DTx	Power Control	Antenna Beam Forming
GSM Family	AAP	AAP	AAP	APP
IS-95 (DS-CDMA)	Not Applicable	Essential requirements for satisfactory system operation and not capacity enhancement features		APP
IS-54/IS-136	ANP	ANP	AAP	APP
PDC	ANP	ANP	AAP	APP
DECT	ANP	ANP	ANP	APP
PHS	ANP	ANP	ANP	APP

Note: AAP Applicable and already provided by the standard.
ANP Applicable but not explicitly provided by the standard.
APP Applicable depending on BS equipment design.

Capacity enhancements of 200% or more have been reported [10, 13] through the implementation of combinations of SFH, VAD/DTx, and PC in GSM networks. With SFH and antenna beam forming, Cs of three is achievable for GSM networks [1]. Since VAD/DTx and PC cannot be applied to the pilot carriers, most of the operational networks deploy a dual cluster scheme where Cs for pilot carriers is slightly higher than Cs for traffic carriers. Moreover, some other issues related to pilot channels/carriers have somewhat impeded the introduction of antenna beam forming in GSM or IS-95 (DS-CDMA) networks for cellular applications. However, in wireless local loop applications, substantial capacity gains have been reported for IS-95 [12].

26.3.3 Novel Cellular Configurations

The traffic capacity throughput by a regular grid of homogeneous cells can be increased by cell splitting. Theoretically speaking, if the cell size is divided by two by adding base stations in the middle of existing ones and the current frequency reuse scheme is maintained the achievable traffic capacity is multiplied by four. However, a reduction in cell

size beyond a limit leads to excessive overlap between cells due to increased difficulty of coverage prediction. Moreover, since the dwelling time of MS in a cell is reduced, the average number of handovers per call increases. Conventional cellular organizations can no longer meet the requirements for good quality of service due to the failure of handover mechanisms. For high capacity coverage, the following novel cellular configurations have been suggested/implemented:

- Microcells and the associated hierarchical network organization
- Concentric cells
- Frequency reuse of one through generalized slow frequency hopping.

Microcells and Associated Hierarchical Network Organization

Hot spot coverage in dense urban areas is realized in the form of isolated islands of microcells. This is implemented through micro-BS antennas placed below the roof tops of the surrounding buildings. Each antenna radiates very low power. Each island is covered by an umbrella cell which is a part of a continuous macrocellular network over a wider area. Traffic throughput is optimized by intelligent spectrum management performed either off-line or on-line. A set of carriers is assigned to the macrocellular layer organized in a conventional manner. The remaining carriers are repeatedly used in the islands of microcells. Cell selection parameters in the MS and the call admission control algorithms (e.g., Forced Directed Retry) in the base station controllers are adjusted such that a maximum of traffic in the hot spot is taken by the microcells. The umbrella cells are dimensioned to handle the spill-over traffic. This can be the fast moving MS which could generate too many handovers if kept with the microcells. An MS which experiences a sudden degradation of link budget with respect to its serving microcell and/or in the absence of a good target microcell can be temporarily handed over to the umbrella cell.

Despite its difficulty related to spectrum management, this technique has proved to be quite popular for the densification of parts of existing networks based on second generation systems using TDMA/FDMA. [6, 11] provide analysis and guidelines for spectrum management for optimized efficiency in hierarchical networks.

Concentric Cells

A concentric cell coverage is implemented by splitting the available traffic carriers in two groups. One group transmits at full power required to cover the complete cell and the other group transmits at a lower level thus providing the coverage of an inner zone concentric with the original cell. The pilot carrier is transmitted at full power. The localized transmission in the inner zone creates a lower level of interference for other cells. Hence, a smaller cluster size can be used for the frequencies of the inner zone leading to traffic capacity enhancement. Call admission control is designed to keep the MSs near the BS on the carriers of inner zone. Simple intracell handover mechanisms are used to ensure call continuity for the MSs moving across the boundary of the inner zone. Analysis and simulations have shown that optimized capacity enhancement is achieved if the inner zone is limited to 40% of the total cell area. Field trials of concentric cell networks have demonstrated 35% higher spectrum efficiency as compared to single cell networks.

Reuse of One Through Generalized Frequency Hopping and Fractional Loading

Micro- or picocellular networks with TDMA/FDMA systems can be deployed with frequency "reuse of one." Every base station has the capability of using all the available carriers by slow frequency hopping. The allocation of frequency hopping laws for MSs in

clusters of adjacent cells is managed by a centralized control.

During the steady state of operation in a loaded network, only a fraction of the total available bandwidth is in active use in every cell (fractional loading). The level of interference for active calls, for unused circuits and the availability of noninterfered frequency hopping laws is constantly monitored by the network. New calls in a cell are accepted according to the availability of interference-free circuits. In case of unevenly distributed traffic load, an average circuit occupancy in the network is maintained at a level necessary to keep the interference level for active calls below a predetermined threshold. Extreme situations where the same circuit and/or same frequency hopping law is used in two adjacent cells are remedied by intracell handover of one of the two calls. Very high capacity enhancement has been demonstrated in operational GSM networks by using this technique.

Table 26.4 provides the applicability matrix for novel cellular configurations.

TABLE 26.4 Applicability Matrix of Methods for Capacity Enhancements Based on Novel Cellular Configurations

Second Generation System	Method for Capacity Enhancement		
	Micro-cells and Hierarchical N/W	Concentric Cells	Reuse of One with GSFH
GSM Family	AAP	AAP	AAP
IS-54/IS-136	AAP	AAP	ANP
PDC	AAP	AAP	ANP
IS-95 Family using DS-CDMA	Not applicable due to the resulting near–far problem and the power control complexity		Inherent due to DS-CDMA
DECT	AAP	AAP	ANP. Reuse of one possible with DCS
PHS	AAP	AAP	ANP

Note: AAP Applicable and already provided by the standard.
ANP Applicable but not explicitly provided by the standard.
APP Applicable depending on BS equipment design.

26.4 Quality Enhancement

26.4.1 Quality Aspects and Definitions

The quality of service in a telecommunications network can simply be defined as the average performance perceived by the end user in setting up and maintaining a communication. However, its assessment is complex since it is influenced by many parameters, especially in digital wireless systems. In these systems the quality is primarily based on the end-to-end bit error rate and on the continuity of radio links between the two ends. Interference-free radio coverage with sufficient desired signal strength needs to be provided to achieve the above. Moreover, communication continuity has to be ensured between coverage areas with high traffic density (microcells) and low/medium traffic density (macrocells). Three main quality aspects will be distinguished in the following, namely the call handling quality, the communication quality, and the coverage quality.

For speech transmission, the communication quality strongly depends on the intrinsic performance of the speech coder, and its evaluation normally requires intensive listening tests. When it is comparable to the quality achieved on modern wire-line telephone networks, it is called "toll quality." But the **speech quality** is also influenced by other parameters linked to the communication characteristics like radio channel impairments (bit error rate), transmission delay, echo, background noise and tandeming (i.e., when several coding/decoding operations exist in the link).

For data transmission, the communication quality can be more easily quantified based on bit error rate and transmission delay. In synchronous circuit mode, delay is fixed and bit error rate depends on radio channel quality. In packet mode, bit error rate can be kept low thanks to retransmission mechanisms, but average delay increases and throughput decreases as the radio channel degrades.

The **coverage quality** is the percentage of the served area where a communication can be established. It is determined by the acceptable path loss of the radio link and by the propagation characteristics in the area. The radio link budget generally includes some margin depending on the type of terrain (for shadowing effects) and on operator's requirements (for indoor penetration). A coverage quality of 90% is a typical value for cellular networks.

The call handling quality mainly depends on the capacity of the mobile network. When the user is under coverage of the network, the call set-up performance is measured by the blocking rate which depends on the network load. Since in cellular networks mobility of the users is high, capacity requirements are less predictible than in fixed networks which results in a higher blocking rate (typically 1%). Requirements are even more stringent in cordless systems.

A last call-handling quality attribute specific to mobile networks is the success rate in maintaining the communication for mobile users. The handover procedure triggered in cellular networks when the user moves from one cell to another implies the establishment of a new link with some risk to lose the call. The performance is here given by the handover success rate. It must be noted that even if successful, a handover generally results in a short transmission break which degrades the communication quality.

In fact, capacity and quality are dual parameters and a compromise is needed when designing a mobile network. The call handling quality is strongly linked to the correct dimensioning of the network capacity. On the other hand, offering a high capacity in a mobile network implies an intensive use of the available radio spectrum and hence a high average interference level, which may in turn degrade the communication quality. In the following, some major enhancements in speech quality and coverage quality standardized or implemented in second generation systems are reviewed.

26.4.2 Speech Quality Enhancements

In digital cellular systems the first generation of speech coders (full rate) were standardized in the late 1980s to provide good communication quality at medium bit rate (6.7 to 13 kb/s). Operation at a fixed bit rate matched to the intended communication channel was the most important requirement in creating these standards. The main characteristics of the full rate speech coders standardized for second generation cellular systems are listed in Table 26.1 together with other air interface parameters.

With the recent advances in speech coding techniques, low delay toll quality at 8 kb/s and near-toll quality below 6 kb/s are now available [2]. This has enabled the standardization of half rate codecs in cellular systems, though with an increased complexity of implementation. In GSM/DCS, IS-54, and PDC this evolution was anticipated at an early stage, so the need

for a half rate speech channel was taken into account in the design of the TDMA frame structure. Half rate codecs provide a doubling of the systems capacity (in terms of number of voice circuits per cell site) while maintaining a speech quality comparable to that available from related full rate codecs. A 5.6 kb/s VSELP coder for GSM and a 3.45 kb/s PSI-CELP coder for PDC were standardized in 1993. However, they have not been widely introduced up to now because of their slightly lower quality in some conditions (e.g., background noise, tandeming) compared to their full rate counterpart.

On the contrary, operators have pushed towards speech quality enhancements, and a new generation of speech coders has emerged. In 1996 ETSI standardized the GSM EFR (enhanced full rate) coder, adopting without any competitive selection process the US1 coder defined for PCS1900 in the U.S. The EFR coder has the remarkable feature of keeping the same channel coding as for the full rate channel, hence simplifying its implementation in the infrastructure. The ACELP coder (using an algebraic codebook with some benefits in computational efficiency like in the the 8 kb/s G.729 ITU standard) has a net bit rate of 12.2 kb/s which leaves room for additional protection (CRC).

Similarly the IS-641 coder has been standardized for IS-54 cellular systems. Its structure is very similar to that of G.729 but with a frame size of 20 ms and a bit rate of 7.4 kb/s. Also a 13 kb/s QCELP coder has been standardized for IS-95. These new coders provide near toll quality in ideal transmission conditions.

Concerning cordless systems, both DECT and PHS use 32 kb/s ADPCM and, therefore, provide toll quality in normal conditions thanks to their relatively high operational C/I. For these systems the concern is rather to introduce lower rate coders with equivalent quality in some kind of half rate mode, allowing for capacity increase, but such coders have not been standardized yet.

With the emergence of variable bit rate (VBR) techniques, some further developments in speech coding are now taking place to satisfy the requirements of cellular operators for toll quality speech with better robustness to radio channel impairments, combined with the capacity increase achievable with half rate operation. One VBR approach is to adapt the bit rate according to the source requirements, taking advantage of silence and stationary segments in the speech signal. Another VBR approach is to adapt the bit rate according to the radio channel quality, either by varying the bit rate allocation between speech and channel coding within a constant gross bit rate (for better quality) or by reducing the gross bit rate in good transmission conditions (for higher capacity). Adaptation to the source characteristics is fast (on a frame-by-frame basis) and adaptation to the channel quality is slower (maximum, a few times per second). In all cases, the underlying idea is that current coders are dimensioned for worst-case operation so that the system quality and capacity can be increased by exploiting the large variations over time of the bit rate requirements (for a given quality).

VBR capabilities with source driven adaptation were introduced since the beginning in IS-95. In CDMA systems the resulting average bit rate reduction directly translates into increased capacity. The initial IS-96B QCELP coder supports four bit rates (8, 4, 2, and 0.8 kb/s) with a scalable CELP architecture, and bit rate adjustment is performed based on adaptive energy thresholds. In practice the two extreme rates are used most frequently. The achievable speech quality was estimated to be lower than that of the IS-54 VSELP coder. Subsequently an enhanced variable rate coder (EVR) was standardized as IS-127. It is based on the relaxed CELP (RCELP) coding technique and supports three rates (8.5, 4, and 0.8 kbit/s).

In Europe, ETSI has launched in 1997 the standardization of an Adaptive Multirate coder (AMR) for GSM systems. Its output bit rate is continuously adapted to radio channel conditions and traffic load. The objective is to find the best compromise between speech

quality and capacity by selecting an optimum combination of channel mode and codec mode. The AMR codec can operate in two channel modes, full rate and half rate. For each channel mode there are several possible codec modes (typically three) with different bit rate allocations between speech and channel coding. An adaptation algorithm tracks the variations in speech quality using specific metrics and decides upon the changes in codec mode (up to several times per second). Changes in codec mode is detected by the receiver either via in-band signalling or by automatic mode identification. As shown in Fig. 26.1, the multiplicity of codec modes gives significant performance improvement over any of the corresponding fixed rate codecs.

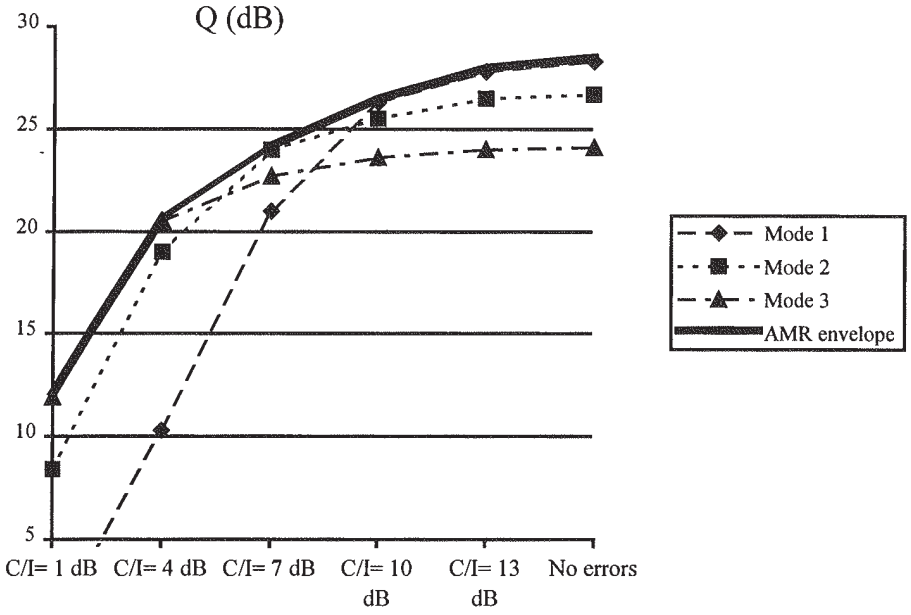


FIGURE 26.1 AMR codec quality as a function of C/I. (Typical results derived from the first AMR selection tests. Quality is given on the equivalent MNRU scale in dB).

AMR codec will also allow handovers between half rate and full rate channels using intracell handover mechanisms. The AMR codec will be selected by October 1998 and the full standard will be available in mid-1999. A potential extension is wideband coding, which could be added later to AMR as an option. The wideband option would extend the audio bandwidth from the current 300–3400 Hz to 50–5000 Hz or even 50–7000 Hz.

Speech quality enhancements in cellular networks do not concern only the speech coder itself. First, it is well known that tandeming (i.e., several cascaded coding/decoding operations) can be a source of significant degradation of speech quality. In the case of mobile-to-mobile calls there is no need to decode the speech signal in the network (assuming both mobiles use the same speech codec). Therefore, a tandem-free operation (TFO) will soon be introduced for GSM using in-band signalling between peer transcoders. The benefit of TFO, however, will largely depend on the percentage of intra-GSM mobile-to-mobile calls.

In CDMA systems like IS-95, the soft handoff feature enables a smooth transition from one cell to another. On the contrary in TDMA systems, handovers produce short breaks in the communication which locally degrade the speech quality in spite of the speech extrapolation

mechanism used in the decoder. In GSM the duration of the transmission break can be as long as 160 ms due to the necessary time alignment of the mobile in the new cell. The speech interruption can be reduced by 40 ms in a synchronized network, but in practice local synchronization is only offered for co-sited cells. The same performance can be more easily achieved with the presynchronized handover where the mobile receives an indication of the distance to the base station together with the handover command. Some improvement can also be obtained on the uplink by switching at the right point in time and on the downlink by broadcasting the speech information to both the old and the new cell. As a result of all these improvements, the interruption time can be reduced down to 60 ms.

Some advances have also been made concerning the robustness to radio channel impairments. Unequal error protection (UEP) is now used in most codecs designed for cellular systems and it is often based on rate-compatible punctured convolutional codes. UEP enables the adjustment of the protection rate as a function of the sensitivity of bits output by the speech coder. Besides, most linear predictive coders use a speech extrapolation procedure, replacing potentially corrupted information in the current frame with more reliable information from a previously received frame. More sophisticated error concealment techniques, using both reliability information at the output of the channel decoder and a statistical source model (*a priori* knowledge), have been reported to provide up to 3 dB improvement in E_b/N_0 under adverse channel conditions [5].

Robustness to background noise is another topic of interest. Speech quality improvements in medium- to low-bit rate coders have been obtained with coding algorithms optimized for speech signals. Such algorithms may produce poor results in the presence of background noise. Therefore, optional noise cancellation processing has been introduced in recent standards. This can be performed either in the time domain (Kalman filtering) as in JDC half rate or in the frequency domain (spectral subtraction) as in IS-127 EVR.

26.4.3 Coverage Quality Enhancements

Various second generation systems provide the possibility of implementing mechanisms like slow frequency hopping (as in GSM), antenna diversity (also called microdiversity), macrodiversity or multisite transmission, and dynamic channel selection (as in DECT). Such mechanisms are useful to alleviate the effects of radio transmission phenomena like shadowing, fading, and cochannel interference. They are, therefore, particularly interesting to enhance the coverage quality in interference-limited or strong multipath environments (e.g., urban areas). Optional equalizers have been defined for cordless systems (DECT and PHS) with the introduction of a suitable training sequence ("prolonged preamble" in DECT), allowing for channel estimation in the presence of longer delay spread and thus providing increased coverage and/or quality.

Operators also have to face network planning issues linked to natural or artificial obstacles. Various solutions have been designed for the cases where the coverage cannot be efficiently ensured with regular base stations. Radiating cables (leaky feeders) are typically used for the coverage of tunnels. Radio repeaters have been standardized for GSM and for DECT (Wireless Relay Station, WRS) and are useful to fill coverage holes and to provide outdoor-to-indoor or outdoor-to-underground coverage extensions. Microcells (and microbase stations) may also be used as "gap fillers."

The use of smart antennas at cell sites helps to extend the cell range. In rural environments where traffic is low, the required number of cell sites is minimized by using antenna arrays and signal processing algorithms to ensure high sensitivity reception at the base station. Algorithms that implement either n-fold receive diversity or mobile station direction finding followed by beamforming for mobile station tracking have been shown to perform

well in such radio channels. Adaptive beamforming techniques with an M-element antenna array generally provide a directivity gain of M, e.g., 9 dB with $M = 8$ (plus some diversity gain depending on channel type). It results in a significant increase of the uplink range (typically by a factor > 1.5 with $M = 8$), but an increased range in the downlink is also needed to get an effective reduction of the number of cell sites. An increased downlink range can be achieved using adaptive beamforming (but with a much higher complexity compared to the uplink-only implementation), a multibeam antenna (i.e., a phased array doing fixed beamforming), or an increased transmit power of the base station. However, the success of smart antenna techniques for range extension applications in second generation systems has been slowed down by their complexity of implementation and by operational constraints (multiple feeders, large antenna panels).

26.5 High Bit Rate Data Transmission

26.5.1 Circuit Mode Techniques

All second generation wireless systems support circuit mode data services with basic rates typically ranging from 9.6 kb/s (in cellular systems) to 32 kb/s (in cordless systems) for a single physical radio resource. With the growing needs for higher rates, new services have been developed based on multiple allocation or grouping of physical resource.

In GSM, HSCSD (High Speed Circuit Switched Data) enables multiple Full Rate Traffic Channels (TCH/F) to be allocated to a call so that a mobile subscriber can use n times the transmission capacity of a single TCH/F channel (Fig. 26.2). The n full rate channels over which the user data stream is split are handled completely independently in the physical layer and for layer 1 error control. The HSCSD channel resulting from the logical combination of n TCH/F channels is controlled as a single radio link during cellular operations such as handover. At the A interface, calls will be limited to a single 64 kb/s circuit. Thus HSCSD will support transparent (up to 64 kb/s) and nontransparent modes (up to $4 \times 9.6 = 38.4$ kb/s and, later, $4 \times 14.4 = 57.6$ kb/s). The initial allocation can be changed during a call if required by the user and authorized by the network. Initially the network allocates an appropriate HSCSD connection according to the requested user bit rate over the air interface. Both symmetric and asymmetric configurations for bidirectional HSCSD operation are authorized. The required TCH/F channels are allocated over consecutive or nonconsecutive timeslots.

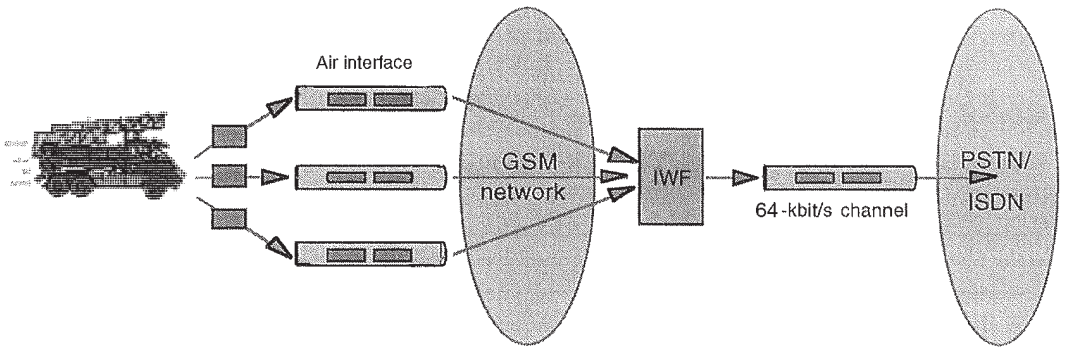


FIGURE 26.2 Simplified GSM network configuration for HSCSD.

Similar multislot schemes are envisaged or standardized for other TDMA systems. In IS-54 and PDC, where radio channels are relatively narrowband, no more than three time slots can be used per carrier and the achievable data rate is therefore limited to, say, 32 kb/s. On the contrary, in DECT up to 12 time slots can be used at 32 kb/s each, yielding a maximum data rate of 384 kb/s. Moreover, the TDD access mode of DECT allows asymmetric time slot allocation between uplink and downlink, thus enabling even higher data rates in one direction.

26.5.2 Packet Mode Techniques

There is a growing interest for packet data services in second generation wireless systems to support data applications with intermittent and bursty transmission requirements like the Internet, with a better usage of available radio resources, thanks to the multiplexing of data from several mobile users on the same physical channel. Cellular Digital Packet Data (CDPD) has been defined in the U.S. as a radio access overlay for AMPS or D-AMPS (IS-54) systems, allowing packet data transmission on available radio channels. However, CDPD is optimized for short data transmission and the bit rate is limited to 19.2 kb/s. A CDMA packet data standard has also been defined (IS-657) which supports CDPD and Internet protocols with a similar bit rate limitation but allowing use of the same backhaul as for voice traffic.

In Europe, ETSI has almost completed the standardization of GPRS (General Packet Radio Service) for GSM. A GPRS subscriber will be able to send and receive in an end-to-end packet transfer mode. Both point-to-point and point-to-multipoint modes are defined. A GPRS network coexists with a GSM PLMN as an autonomous network. In fact, the Serving GPRS Support Node (SGSN) interfaces with the GSM Base Station Controller (BSC), an MSC and a Gateway GPRS Service Node (GGSN). In turn, the GGSN interfaces with the GGSNs of other GPRS networks and with public Packet Data Networks (PDN). Typically, GPRS traffic can be set up through the common control channels of GSM, which are accessed in slotted ALOHA mode. The layer 2 protocol data units, which are about 2 kbytes in length, are segmented and transmitted over the air interface using one of the four possible channel coding schemes. The system is highly scaleable as it allows from one mobile using 8 radio time slots up to 16 mobiles per time slot, with separate allocation in up- and downlink. The resulting peak data rate per user ranges from 9 kb/s up to 170 kb/s. Time slot concatenation and variable channel coding to maximize the user information bit rate are envisaged for future implementations. This is indicated by the mobile station, which provides information concerning the desire to initiate in-call modifications and the channel coding schemes that can be used during the call set up phase. It is expected that use of the GPRS service will initially be limited and traffic growth will depend on the introduction of GPRS capable subscriber terminals. Easy scalability of the GPRS backbone (e.g., by introducing parallel GGSNs) is an essential feature of the system architecture (Fig. 26.3).

26.5.3 New Modulation Schemes

New modulation schemes are being studied as an option in several second generation wireless standards. The aim is to offer higher rate data services equivalent or close to the 2 Mb/s objective of the forthcoming third generation standards. Multilevel modulations (i.e., several bits per modulated symbol) represent a straightforward means to increase the carrier bit rate. However, it represents a significant change in the air interface characteristics, and the increased bit rate is achieved at the expense of a higher operational signal-to-noise plus interference ratio, which is not compatible with large cell dimensions. Therefore, the new

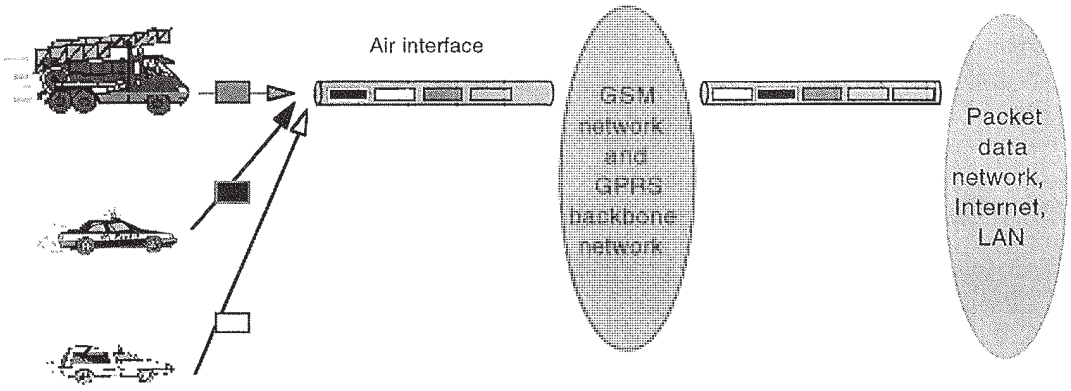


FIGURE 26.3 Simplified view of the GPRS architecture.

high bit rate data services are mainly targetting urban areas, and the effective bit rate allocated to data users will depend on the system load.

Such a new air interface option is being standardized for GSM under the name of EDGE (Enhanced Data rates for GSM Evolution). The selected modulation scheme is 8-PSK, suitable coding schemes are under study, whereas the other air interface parameters (carrier spacing, TDMA frame structure,...) are kept unchanged. Reusing HSCSD (for circuit data) and GPRS (for packet data) protocols and service capabilities, EDGE will provide similar ECSD and EGPRS services but with a three-fold increase of the user bit rate. The higher level modulation requires better radio link performances, typically a loss of 3 to 4 dB in sensitivity and a C/I increased by 6 to 7 dB. Operation will also be restricted to environments with limited time dispersion and limited mobile speed. Nevertheless, EGPRS will roughly double the mean throughput compared to GPRS (for the same average transmitted power). EDGE will also increase the maximum achievable data rate in a GSM system to 553.6 kb/s in multislot (unprotected) operation. Six different protection schemes are foreseen in EGPRS using convolutional coding with a rate ranging from 1/3 to 1 and corresponding to user rates between 22.8 and 69.2 kb/s per time slot. This is in addition to the four coding schemes already defined for GPRS. An intelligent link adaptation algorithm will dynamically select the most appropriate modulation and coding schemes, i.e., those yielding the highest throughput for a given channel quality. The first phase of EDGE standardization should be completed by end 1999. It should be noted that a similar EDGE option is being studied for IS-54/IS-136 (and their PCS derivatives). Initially, the 30 kHz channel spacing will be maintained and then extension to a 200 kHz channel will be provided in order to offer a convergence with its GSM counterpart.

A higher bit rate option is also under standardization for DECT. Here it is seen as an essential requirement to maintain backward compatibility with existing equipment so the new multilevel modulation will only affect the payload part of the bursts, keeping the control and signalling parts unchanged. This ensures that equipment with basic modulation and equipment with a higher rate option can efficiently share a common base station infrastructure. Only 4-level and 8-level modulations are considered and the symbol length, carrier spacing, and slot structure remain unchanged. The requirements on transmitter modulation accuracy need to be more stringent for 4- and 8-level modulation than for the current 2-level scheme. An increased accuracy can provide for coherent demodulation, whereby some (or most) of the sensitivity and C/I loss when using the multilevel mode can be regained. In combination with other new air interface features like forward error correction and double slots (with reduced overhead), the new modulation scheme will provide a wide range of

data rates up to 2 Mb/s. For instance using ($\Pi/4$-DQPSK modulation (a possible/suitable choice), an unprotected connection with two double slots in each direction gives a data rate of 384 kb/s. Asymmetric connections with a maximum of 11 double slots in one direction will also be supported.

26.6 Conclusion

Since their introduction in the early 1990s, most of the second generation systems have been enjoying exponential growth. With more than 100 million subscribers acquired worldwide in less than ten years of lifetime, the systems based on the GSM family of standards have demonstrated the most spectacular development. Despite a more regional implementation of other second generation systems, each one of those can boast a multimillion subscriber base in mobile or fixed wireless networks.

A variety of service requirements of third generation mobile communication systems are being already met by the upcoming enhancements of second generation systems. Two important trends are reflected by this:

- The introduction of third generation systems like Universal Mobile Telecommunication System (UMTS) or International Mobile Telecommunication-2000 (IMT-2000) might be delayed to a point in time where the evolutionary capabilities of second generation systems have been exhausted.
- The deployment of networks based on third generation systems will be progressive. Any new radio interface will be imposed worldwide if and only if it provides substantial advantages as compared to the present systems. Another essential requirement is the capability of downward compatibility to second generation systems.

Defining Terms

Capacity: In a mobile network it can be defined as the Erlangs throughput by a cell, a cluster of cells, or by a portion of a network. For a given radio interface, the achievable capacity is a function of the robustness of the physical layer, the effectiveness of the medium access control (MAC) layer and the multiple access technique. Moreover, it is strongly dependent on the radio spectrum available for network planning.

Cellular: Refers to public land mobile radio networks for generally wide area (e.g., national) coverage, to be used with medium- or high-power vehicular mobiles or portable stations and for providing mobile access to the Public Switched Telephone Network (PSTN). The network implementation exhibits a cellular architecture which enables frequency reuse in nonadjacent cells.

Cordless: These are systems to be used with simple low power portable stations operating within a short range of a base station and providing access to fixed public or private networks. There are three main applications, namely, residential (at home, for Plain Old Telephone Service, POTS), public-access (in public places and crowded areas, also called Telepoint), and Wireless Private Automatic Branch eXchange (WPABX, providing cordless access in the office environment), plus emerging applications like radio access for local loop.

Coverage quality: It is the percentage of the served area where a communication can be established. It is determined by the acceptable path loss of the radio link and by the propagation characteristics in the area. The radio link budget generally includes some margin depending on the type of terrain (for shadowing effects) and on operator's requirements (for indoor penetration). A coverage quality of 90% is a typical value for cellular networks.

Speech quality: It strongly depends on the intrinsic performance of the speech coder and its evaluation normally requires intensive listening tests. When it is comparable to the quality achieved on modern wire-line telephone networks, it is called "toll quality." In wireless systems it is also influenced by other parameters linked to the communication characteristics like radio channel impairments (bit error rate), transmission delay, echo, background noise, and tandeming (i.e., when several coding/decoding operations are involved in the link).

References

[1] Anderson, S., Antenna Arrays in Mobile Communication Systems, *Proc. Second Workshop on Smart Antennas in Wireless Mobile Communications,* Stanford University, Jul. 1995.

[2] Budagavi, M. and Gibson, J.D., Speech coding in mobile radio communications, *Proceedings of the IEEE,* 86(7), 1402–1412, Jul. 1998.

[3] Cox, D.C., Wireless network access for personal communications, *IEEE Communications Magazine,* 96–115, Dec. 1992.

[4] DECT, *Digital European Cordless Telecommunications Common Interface,* ETS-300-175, ETSI, 1992.

[5] Fingscheidt, T. and Vary, P., Robust Speech Decoding: A Universal Approach to Bit Error Concealment, *Proc. IEEE ICASSP,* 1667–1670, Apr. 1997.

[6] Ganz, A., et al., On optimal design of multitier wireless cellular systems, *IEEE Communications Magazine,* 88–93, Feb. 1997.

[7] GSM, *GSM Recommendations Series 01-12,* ETSI, 1990.

[8] IS-54, *Cellular System, Dual-Mode Mobile Station-Base Station Compatibility Standard,* EIA/TIA Interim Standard, 1991.

[9] IS-95, *Mobile Station-Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System,* EIA/TIA Interim Standard, 1993.

[10] Kuhn, A., et al., Validation of the Feature Frequency Hopping in a Live GSM Network, *Proc. 46th IEEE Vehic. Tech. Conf.,* 321–325, Apr. 1996.

[11] Lagrange, X., Multitier cell design, *IEEE Communications Magazine,* 60–64, Aug. 1997.

[12] Lee, D. and Xu, C., The effect of narrowbeam antenna and multiple tiers on system capacity in CDMA wireless local loop, *IEEE Communications Magazine,* 110–114, Sep. 1997.

[13] Olofsson, H., et al., Interference Diversity as Means for Increased Capacity in GSM, *Proc. EPMCC'95,* 97–102, Nov. 1995.

[14] PDC, *Personal Digital Cellular System Common Air Interface,* RCR-STD27B, 1991.

[15] PHS, *Personal Handy Phone System: Second Generation Cordless Telephone System Standard,* RCR-STD28, 1993.

[16] Tuttlebee, W.H.W., Cordless personal communications, *IEEE Communications Magazine,* 42–53, Dec. 1992.

Further Information

European standards (GSM, CT2, DECT, TETRA) are published by ETSI Secretariat, 06921 Sophia Antipolis Cedex, France.

U.S. standards (IS-54, IS-95, APCO) are published by Electronic Industries Association, Engineering Department, 2001 Eye Street, N.W., Washington D.C. 20006, U.S.A.

Japanese standards (PDC, PHS) are published by RCR (Research and Development Center for Radio Systems), 1-5-16, Toranomon, Minato-ku, Tokyo 105, Japan.

27

The Pan-European Cellular System

27.1 Introduction 27-1
27.2 Overview 27-2
27.3 Logical and Physical Channels 27-4
27.4 Speech and Data Transmission 27-5
27.5 Transmission of Control Signals 27-8
27.6 Synchronization Issues 27-11
27.7 Gaussian Minimum Shift Keying Modulation 27-12
27.8 Wideband Channel Models 27-13
27.9 Adaptive Link Control 27-15
27.10 Discontinuous Transmission 27-17
27.11 Summary 27-18
Defining Terms 27-19
References 27-22

Lajos Hanzo
University of Southampton

27.1 Introduction

Following the standardization and launch of the Pan-European digital mobile cellular radio system known as GSM, it is of practical merit to provide a rudimentary introduction to the system's main features for the communications practitioner. Since GSM operating licenses have been allocated to 126 service providers in 75 countries, it is justifiable that the GSM system is often referred to as the Global System of Mobile communications.

The GSM specifications were released as 13 sets of recommendations [1], which are summarized in Table 27.1, covering various aspects of the system [3].

After a brief system overview in Section 27.2 and the introduction of physical and logical channels in Section 27.3 we embark upon describing aspects of mapping logical channels onto physical resources for speech and control channels in Sections 27.4 and 27.5, respectively. These details can be found in recommendations R.05.02 and R.05.03. These recommendations and all subsequently enumerated ones are to be found in [1]. Synchronization issues are considered in Section 27.6. Modulation (R.05.04), transmission via the standardized wideband GSM channel models (R.05.05), as well as adaptive radio link control (R.05.06 and R.05.08), discontinuous transmission (**DTX**) (R.06.31), and voice activity detection (**VAD**) (R.06.32) are highlighted in Sections 27.7–27.10, whereas a summary of the fundamental GSM features is offered in Section 27.11.

0-8493-8597-0/99/$0.00+$.50

© 1999 by CRC Press LLC

TABLE 27.1 GSM Recommendations [R.01.01]

R.00	*Preamble* to the GSM recommendations
R.01	*General structure* of the recommendations, description of a GSM network, associated recommendations, vocabulary, etc.
R.02	*Service aspects:* bearer-, tele- and supplementary services, use of services, types and features of mobile stations (MS), licensing and subscription, as well as transferred and international accounting, etc.
R.03	*Network aspects,* including network functions and architecture, call routing to the MS, technical performance, availability and reliability objectives, handover and location registration procedures, as well as discontinuous reception and cryptological algorithms, etc.
R.04	*Mobile/base station (BS) interface and protocols,* including specifications for layer 1 and 3 aspects of the open systems interconnection (OSI) seven-layer structure.
R.05	*Physical layer on the radio path,* incorporating issues of multiplexing and multiple access, channel coding and modulation, transmission and reception, power control, frequency allocation and synchronization aspects, etc.
R.06	*Speech coding specifications,* such as functional, computational and verification procedures for the speech codec and its associated voice activity detector (VAD) and other optional features.
R.07	*Terminal adaptors for MSs,* including circuit and packet mode as well as voiceband data services.
R.08	*Base station and mobile switching center* (MSC) *interface,* and transcoder functions.
R.09	*Network interworking* with the public switched telephone network (PSTN), integrated services digital network (ISDN) and, packet data networks.
R.10	*Service interworking, short message service.*
R.11	*Equipment specification and type approval specification* as regards to MSs, BSs, MSCs, home (HLR) and visited location register (VLR), as well as system simulator.
R.12	*Operation and maintenance,* including subscriber, routing tariff and traffic administration, as well as BS, MSC, HLR and VLR maintenance issues.

27.2 Overview

The system elements of a GSM public land mobile network (**PLMN**) are portrayed in Fig. 27.1, where their interconnections via the standardized interfaces A and Um are indicated as well. The mobile station (**MS**) communicates with the serving and adjacent base stations (**BS**) via the radio interface Um, whereas the BSs are connected to the mobile switching center (**MSC**) through the network interface A. As seen in Fig. 27.1, the MS includes a mobile termination (MT) and a terminal equipment (TE). The TE may be constituted, for example, by a telephone set and fax machine. The MT performs functions needed to support the physical channel between the MS and the base station, such as radio transmissions, radio channel management, channel coding/decoding, speech encoding/decoding, and so forth.

The BS is divided functionally into a number of base transceiver stations (BTS) and a base station controller (BSC). The BS is responsible for channel allocation (R.05.09), link quality and power budget control (R.05.06 and R.05.08), signalling and broadcast traffic control, frequency hopping (**FH**) (R.05.02), handover (**HO**) initiation (R.03.09 and R.05.08), etc. The MSC represents the gateway to other networks, such as the public switched telephone network (**PSTN**), integrated services digital network (**ISDN**) and packet data networks using the interworking functions standardized in recommendation R.09. The MSC's further functions include paging, MS location updating (R.03.12), HO control (R.03.09), etc. The MS's mobility management is assisted by the home location register (**HLR**) (R.03.12), storing part of the MS's location information and routing incoming calls to the visitor location register (**VLR**) (R.03.12) in charge of the area, where the paged MS roams. Location update is asked for by the MS, whenever it detects from the received and decoded broadcast

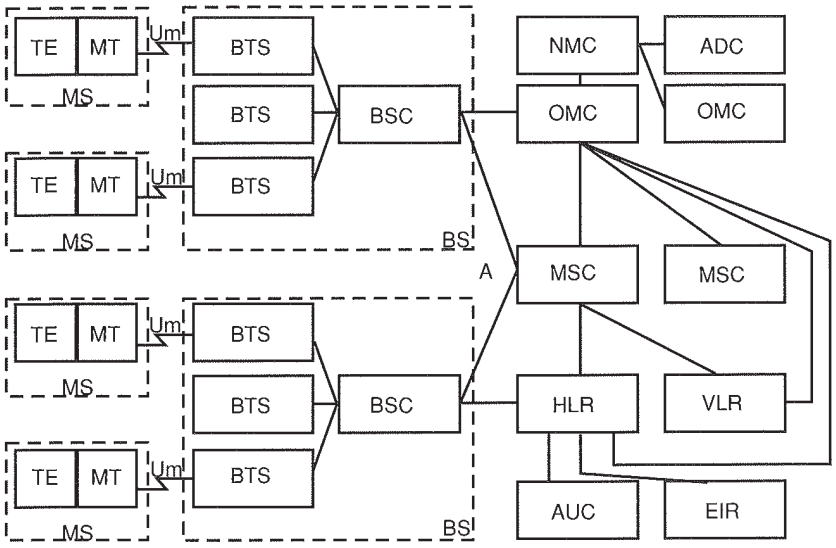


FIGURE 27.1 Simplified structure of GSM PLMN © ETT [4].

control channel (**BCCH**) messages that it entered a new location area. The HLR contains, amongst a number of other parameters, the international mobile subscriber identity (**IMSI**), which is used for the authentication (R.03.20) of the subscriber by his authentication center (**AUC**). This enables the system to confirm that the subscriber is allowed to access it. Every subscriber belongs to a home network and the specific services that the subscriber is allowed to use are entered into his HLR. The equipment identity register (**EIR**) allows for stolen, fraudulent, or faulty mobile stations to be identified by the network operators. The VLR is the functional unit that attends to a MS operating outside the area of its HLR. The visiting MS is automatically registered at the nearest MSC, and the VLR is informed of the MSs arrival. A roaming number is then assigned to the MS, and this enables calls to be routed to it. The operations and maintenance center (**OMC**), network management center (**NMC**) and administration center (**ADC**) are the functional entities through which the system is monitored, controlled, maintained and managed (R.12).

The MS initiates a call by searching for a BS with a sufficiently high received signal level on the BCCH carrier; it will await and recognize a frequency correction burst and synchronize to it (R.05.08). Now the BS allocates a bidirectional signalling channel and also sets up a link with the MSC via the network. How the control frame structure assists in this process will be highlighted in Section 27.5. The MSC uses the IMSI received from the MS to interrogate its HLR and sends the data obtained to the serving VLR. After authentication (R.03.20) the MS provides the destination number, the BS allocates a traffic channel, and the MSC routes the call to its destination. If the MS moves to another cell, it is reassigned to another BS, and a handover occurs. If both BSs in the handover process are controlled by the same BSC, the handover takes place under the control of the BSC, otherwise it is performed by the MSC. In case of incoming calls the MS must be paged by the BSC. A paging signal is transmitted on a paging channel (**PCH**) monitored continuously by all MSs, and which covers the location area in which the MS roams. In response to the paging signal, the MS performs an access procedure identical to that employed when the MS initiates a call.

27.3 Logical and Physical Channels

The GSM logical traffic and control channels are standardized in recommendation R.05.02, whereas their mapping onto physical channels is the subject of recommendations R.05.02 and R.05.03. The GSM system's prime objective is to transmit the logical traffic channel's (**TCH**) speech or data information. Their transmission via the network requires a variety of logical control channels. The set of logical traffic and control channels defined in the GSM system is summarized in Table 27.2. There are two general forms of speech and data traffic channels: the full-rate traffic channels (**TCH/F**), which carry information at a gross rate of 22.8 kb/s, and the half-rate traffic channels (**TCH/H**), which communicate at a gross rate of 11.4 kb/s. A physical channel carries either a full-rate traffic channel, or two half-rate traffic channels. In the former, the traffic channel occupies one timeslot, whereas in the latter the two half-rate traffic channels are mapped onto the same timeslot, but in alternate frames.

TABLE 27.2 GSM Logical Channels ©ETT [4]

Logical Channels					
Duplex BS ↔ MS Traffic Channels: TCH		Control Channels: CCH			
FEC-coded Speech	FEC-coded Data	Broadcast CCH BCCH BS → MS	Common CCH CCCH	Stand-alone Dedicated CCH SDCCH BS ↔ MS	Associated CCH ACCH BS ↔ MS
TCH/F 22.8 kb/s	TCH/F9.6 TCH/F4.8 TCH/F2.4 22.8 kb/s	Freq. Corr. Ch: FCCH	Paging Ch: PCH BS → MS	SDCCH/4	Fast ACCH: FACCH/F FACCH/H
TCH/H 11.4 kb/s	TCH/H4.8 TCH/H2.4 11.4 kb/s	Synchron. Ch: SCH	Random Access Ch: RACH MS → BS	SDCCH/8	Slow ACCH: SACCH/TF SACCH/TH SACCH/C4 SACCH/C8
		General Inf.	Access Grant Ch: AGCH BS → MS		

For a summary of the logical control channels carrying signalling or synchronisation data, see Table 27.2. There are four categories of logical control channels, known as the BCCH, the common control channel (**CCCH**), the stand-alone dedicated control channel (**SDCCH**), and the associated control channel (**ACCH**). The purpose and way of deployment of the logical traffic and control channels will be explained by highlighting how they are mapped onto physical channels in assisting high-integrity communications.

A physical channel in a time division multiple access (**TDMA**) system is defined as a timeslot with a timeslot number (TN) in a sequence of TDMA frames. The GSM system, however, deploys TDMA combined with frequency hopping (FH) and, hence, the physical channel is partitioned in both time and frequency. Frequency hopping (R.05.02) combined with interleaving is known to be very efficient in combatting channel fading, and it results in near-Gaussian performance even over hostile Rayleigh-fading channels. The principle of FH is that each TDMA burst is transmitted via a different RF channel (**RFCH**). If the present TDMA burst happened to be in a deep fade, then the next burst most probably will not be. Consequently, the physical channel is defined as a sequence of radio frequency

channels and timeslots. Each carrier frequency supports eight physical channels mapped onto eight timeslots within a TDMA frame. A given physical channel always uses the same TN in every TDMA frame. Therefore, a timeslot sequence is defined by a TN and a TDMA frame number FN sequence.

27.4 Speech and Data Transmission

The speech coding standard is recommendation R.06.10, whereas issues of mapping the logical speech traffic channel's information onto the physical channel constituted by a timeslot of a certain carrier are specified in recommendation R.05.02. Since the error correction coding represents part of this mapping process, recommendation R.05.03 is also relevant to these discussions. The example of the full-rate speech traffic channel (**TCH/FS**) is used here to highlight how this logical channel is mapped onto the physical channel constituted by a so-called normal burst (**NB**) of the TDMA frame structure. This mapping is explained by referring to Figs. 27.2 and 27.3. Then this example will be extended to other physical bursts such as the frequency correction (**FCB**), synchronization (**SB**), access (**AB**), and dummy burst (**DB**) carrying logical control channels, as well as to their TDMA frame structures, as seen in Figs. 27.2 and 27.6.

The regular pulse excited (**RPE**) speech encoder is fully characterized in the following references: [3, 5, 7]. Because of its complexity, its description is beyond the scope of this chapter. Suffice to say that, as it can be seen in Fig. 27.3, it delivers 260 b/20 ms at a bit rate of 13 kb/s, which are divided into three significance classes: class 1a (50 b), class 1b (132 b) and class 2 (78 b). The class-1a bits are encoded by a systematic (53, 50) cyclic error detection code by adding three parity bits. Then the bits are reordered and four zero tailing bits are added to periodically reset the memory of the subsequent half-rate, constraint length five convolutional codec (**CC**) CC$(2, 1, 5)$, as portrayed in Fig. 27.3. Now the unprotected 78 class-2 bits are concatenated to yield a block of 456 b/20 ms, which implies an encoded bit rate of 22.8 kb/s. This frame is partitioned into eight 57-b subblocks that are block diagonally interleaved before undergoing intraburst interleaving. At this stage each 57-b subblock is combined with a similar subblock of the previous 456-b frame to construct a 116-b burst, where the flag bits *hl* and *hu* are included to classify whether the current burst is really a TCH/FS burst or it has been stolen by an urgent fast associated (**FACCH**) control channel message. Now the bits are encrypted and positioned in a NB, as depicted at the bottom of Fig. 27.2, where three tailing bits (**TB**) are added at both ends of the burst to reset the memory of the Viterbi channel equalizer (**VE**), which is responsible for removing both the channel-induced and the intentional controlled intersymbol interference [6].

The 8.25-b interval duration guard period (GP) at the bottom of Fig. 27.2 is provided to prevent burst overlapping due to propagation delay fluctuations. Finally, a 26-b equalizer training segment is included in the center of the normal traffic burst. This segment is constructed by a 16-b Viterbi channel equalizer training pattern surrounded by five quasiperiodically repeated bits on both sides. Since the MS has to be informed about which BS it communicates with, for neighboring BSs one of eight different training patterns is used, associated with the so-called BS color codes, which assist in identifying the BSs.

This 156.25-b duration TCH/FS NB constitutes the basic timeslot of the TDMA frame structure, which is input to the Gaussian minimum shift keying (**GMSK**) modulator to be highlighted in Section 27.7, at a bit rate of approximately 271 kb/s. Since the bit interval is $1/(271 \text{ kb/s}) = 3.69\ \mu\text{s}$, the timeslot duration is $156.25 \cdot 3.69 \approx 0.577$ ms. Eight such normal bursts of eight appropriately staggered TDMA users are multiplexed onto one (**RF**) carrier giving, a TDMA frame of $8 \cdot 0.577 \approx 4.615$-ms duration, as shown in Fig. 27.2. The

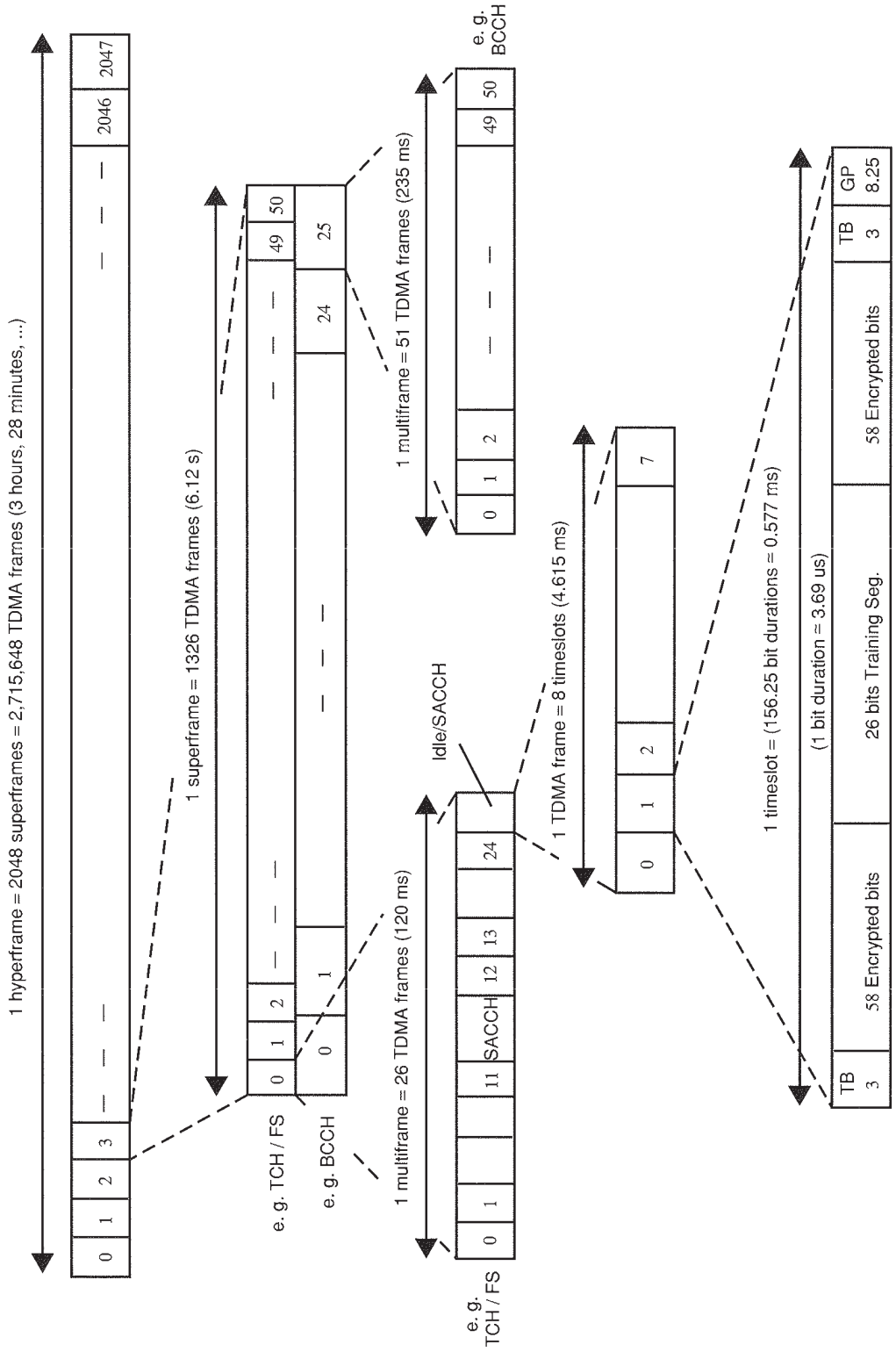


FIGURE 27.2 The GSM TDMA frame structure ©ETT [4].

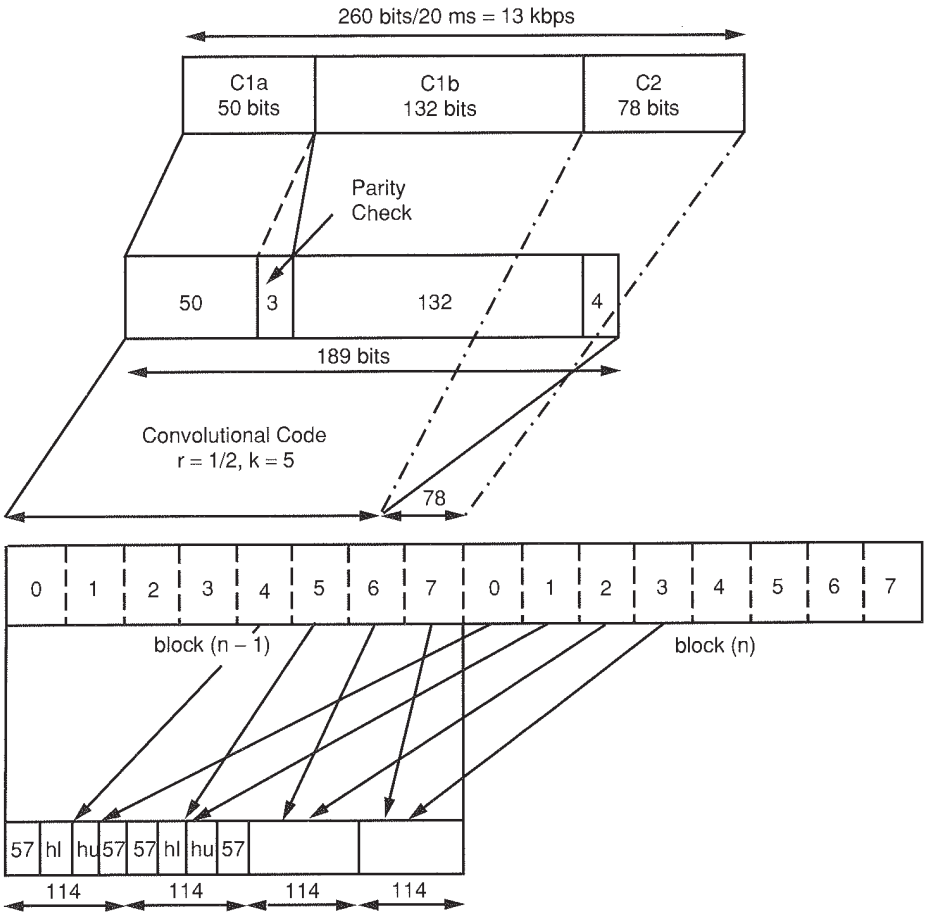


FIGURE 27.3 Mapping the TCH/FS logical channel onto a physical channel, ©ETT [4].

physical channel as characterized earlier provides a physical timeslot with a throughput of 114 b/4.615 ms = 24.7 kb/s, which is sufficiently high to transmit the 22.8 kb/s TCH/FS information. It even has a reserved capacity of 24.7 − 22.8 = 1.9 kb/s, which can be exploited to transmit slow control information associated with this specific traffic channel, i.e., to construct a so-called slow associated control channel (**SACCH**), constituted by the SACCH TDMA frames, interspersed with traffic frames at multiframe level of the hierarchy, as seen in Fig. 27.2.

Mapping logical data traffic channels onto a physical channel is essentially carried out by the channel codecs [8], as specified in recommendation R.05.03. The full- and half-rate data traffic channels standardized in the GSM system are: **TCH/F9.6, TCH/F4.8, TCH/F2.4,** as well as **TCH/H4.8, TCH/H2.4,** as was shown earlier in Table 27.2. Note that the numbers in these acronyms represent the data transmission rate in kilobits per second. Without considering the details of these mapping processes we now focus our attention on control signal transmission issues.

27.5 Transmission of Control Signals

The exact derivation, forward error correcting (**FEC**) coding and mapping of logical control channel information is beyond the scope of this chapter, and the interested reader is referred to ETSI, 1988 (R.05.02 and R.05.03) and Hanzo and Stefanov, 1992, for a detailed discussion. As an example, the mapping of the 184-b SACCH, FACCH, BCCH, SDCCH, PCH, and access grant control channel (**AGCH**) messages onto a 456-b block, i.e., onto four 114-b bursts is demonstrated in Fig. 27.4. A double-layer concatenated FIRE-code/convolutional code scheme generates 456 bits, using an overall coding rate of $R = 184/456$, which gives a stronger protection for control channels than the error protection of traffic channels.

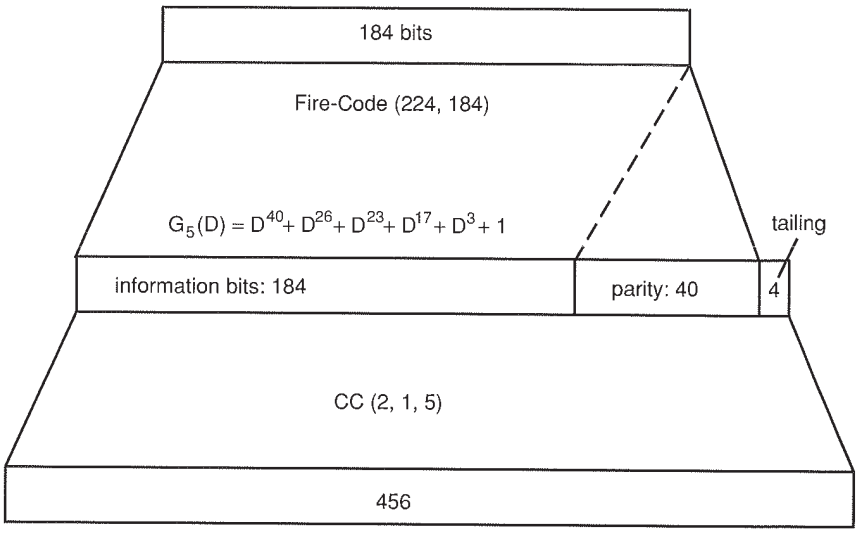


FIGURE 27.4 FEC in SACCH, FACCH, BCCH, SDCCH, PCH and AGCH, ©ETT [4].

Returning to Fig. 27.2 we will now show how the SACCH is accommodated by the TDMA frame structure. The TCH/FS TDMA frames of the eight users are multiplexed into multiframes of 24 TDMA frames, but the 13th frame will carry a SACCH message, rather than the 13th TCH/FS frame, whereas the 26th frame will be an idle or dummy frame, as seen at the left-hand side of Fig. 27.2 at the multiframe level of the traffic channel hierarchy. The general control channel frame structure shown at the right of Fig. 27.2 is discussed later. This way 24-TCH/FS frames are sent in a 26-frame multiframe during $26 \cdot 4.615 = 120$ ms. This reduces the traffic throughput to $(24/26) \cdot 24.7 = 22.8$ kb/s required by TCH/FS, allocates $(1/26) \cdot 24.7 = 950$ b/s to the SACCH and wastes 950 b/s in the idle frame. Observe that the SACCH frame has eight timeslots to transmit the eight 950-b/s SACCHs of the eight users on the same carrier. The 950-b/s idle capacity will be used in case of half-rate channels, where 16 users will be multiplexed onto alternate frames of the TDMA structure to increase system capacity. Then 16, 11.4-kb/s encoded half-rate speech TCHs will be transmitted in a 120-ms multiframe, where also 16 SACCHs are available.

The FACCH messages are transmitted via the physical channels provided by bits stolen from their own host traffic channels. The construction of the FACCH bursts from 184 control bits is identical to that of the SACCH, as also shown in Fig. 27.4 but its 456-b

frame is mapped onto eight consecutive 114-b TDMA traffic bursts, exactly as specified for TCH/FS. This is carried out by stealing the even bits of the first four and the odd bits of the last four bursts, which is signalled by setting $hu = 1, hl = 0$ and $hu = 0, hl = 1$ in the first and last bursts, respectively. The unprotected FACCH information rate is 184 b/20 ms = 9.2 kb/s, which is transmitted after concatenated error protection at a rate of 22.8 kb/s. The repetition delay is 20 ms, and the interleaving delay is $8 \cdot 4.615 = 37$ ms, resulting in a total of 57-ms delay.

In Fig. 27.2 at the next hierarchical level, 51-TCH/FS multiframes are multiplexed into one superframe lasting 51·120 ms = 6.12 s, which contains 26·51 = 1326-TDMA frames. In the case of 1326-TDMA frames, however, the frame number would be limited to $0 \leq \boldsymbol{FN} \leq 1326$ and the encryption rule relying on such a limited range of FN values would not be sufficiently secure. Then 2048 superframes were amalgamated to form a hyperframe of $1326 \cdot 2048 = 2{,}715{,}648$-TDMA frames lasting $2048 \cdot 6.12$ s $\approx$ 3 h 28 min, allowing a sufficiently high FN value to be used in the encryption algorithm. The uplink and downlink traffic-frame structures are identical with a shift of three timeslots between them, which relieves the MS from having to transmit and receive simultaneously, preventing high-level transmitted power leakage back to the sensitive receiver. The received power of adjacent BSs can be monitored during unallocated timeslots.

In contrast to duplex traffic and associated control channels, the simplex BCCH and CCCH logical channels of all MSs roaming in a specific cell share the physical channel provided by timeslot zero of the so-called BCCH carriers available in the cell. Furthermore, as demonstrated by the right-hand side section of Fig. 27.2, 51 BCCH and CCCH TDMA frames are mapped onto a $51 \cdot 4.615 = 235$-ms duration multiframe, rather than on a 26-frame, 120-ms duration multiframe. In order to compensate for the extended multiframe length of 235 ms, 26 multiframes constitute a 1326-frame superframe of 6.12-s duration. Note in Fig. 27.5, that the allocation of the uplink and downlink frames is different, since these control channels exist only in one direction.

Specifically, the random access channel (**RACH**) is only used by the MSs in the uplink direction if they request, for example, a bidirectional SDCCH to be mapped onto an RF channel to register with the network and set up a call. The uplink RACH has a low capacity, carrying messages of 8-b/235-ms multiframe, which is equivalent to an unprotected control information rate of 34 b/s. These messages are concatenated FEC coded to a rate of 36 b/235 ms = 153 b/s. They are not transmitted by the NB derived for TCH/FS, SACCH, or FACCH logical channels, but by the AB, depicted in Fig. 27.6 in comparison to a NB and other types of bursts to be described later. The FEC coded, encrypted 36-b AB messages of Fig. 27.6 contain among other parameters, the encoded 6-b BS identifier code (**BSIC**) constituted by the 3-b PLMN color code and 3-b BS color code for unique BS identification. These 36 b are positioned after the 41-b synchronization sequence, which has a high wordlength in order to ensure reliable access burst recognition and a low probability of being emulated by interfering stray data. These messages have no interleaving delay, while they are transmitted with a repetition delay of one control multiframe length, i.e., 235 ms.

Adaptive time frame alignment is a technique designed to equalize propagation delay differences between MSs at different distances. The GSM system is designed to allow for cell sizes up to 35 km radius. The time a radio signal takes to travel the 70 km from the base station to the mobile station and back again is 233.3 μs. As signals from all the mobiles in the cell must reach the base station without overlapping each other, a long guard period of 68.25 b (252 μs) is provided in the access burst, which exceeds the maximum possible propagation delay of 233.3 μs. This long guard period in the access burst is needed when the mobile station attempts its first access to the base station or after a handover has occurred.

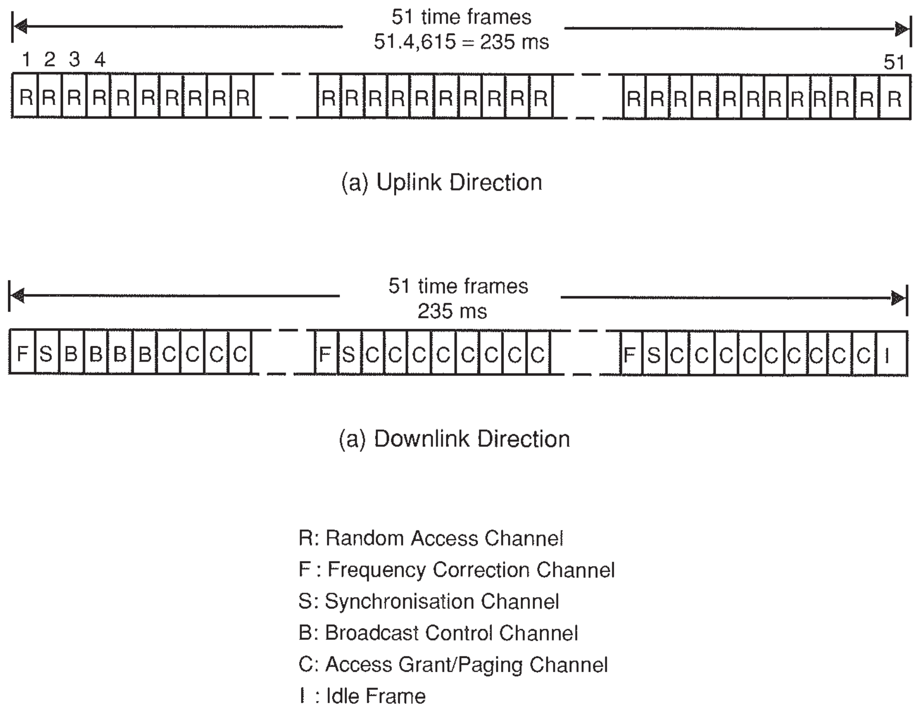


FIGURE 27.5 The control multiframe, ©ETT [4].

When the base station detects a 41-b random access synchronization sequence with a long guard period, it measures the received signal delay relative to the expected signal from a mobile station of zero range. This delay, called the timing advance, is signalled using a 6-b number to the mobile station, which advances its timebase over the range of 0–63 b, i.e., in units of 3.69 μs. By this process the TDMA bursts arrive at the BS in their correct timeslots and do not overlap with adjacent ones. This process allows the guard period in all other bursts to be reduced to $8.25 \cdot 3.69\ \mu\text{s} \approx 30.46\ \mu\text{s}$ (8.25 b) only. During normal operation, the BS continuously monitors the signal delay from the MS and, if necessary, it will instruct the MS to update its time advance parameter. In very large traffic cells there is an option to actively utilize every second timeslot only to cope with higher propagation delays, which is spectrally inefficient, but in these large, low-traffic rural cells it is admissible.

As demonstrated by Fig. 27.2, the downlink multiframe transmitted by the BS is shared amongst a number of BCCH and CCCH logical channels. In particular, the last frame is an idle frame (I), whereas the remaining 50 frames are divided in five blocks of ten frames, where each block starts with a frequency correction channel (**FCCH**) followed by a synchronization channel (**SCH**). In the first block of ten frames the FCCH and SCH frames are followed by four BCCH frames and by either four AGCH or four PCH. In the remaining four blocks of ten frames, the last eight frames are devoted to either PCHs or AGCHs, which are mutually exclusive for a specific MS being either paged or granted a control channel.

The FCCH, SCH, and RACH require special transmission bursts, tailored to their missions, as depicted in Fig. 27.6. The FCCH uses frequency correction bursts (FCB) hosting a specific 142-b pattern. In partial response GMSK it is possible to design a modulating data sequence, which results in a near-sinusoidal modulated signal imitating an unmodulated carrier exhibiting a fixed frequency offset from the RF carrier utilized. The synchronization

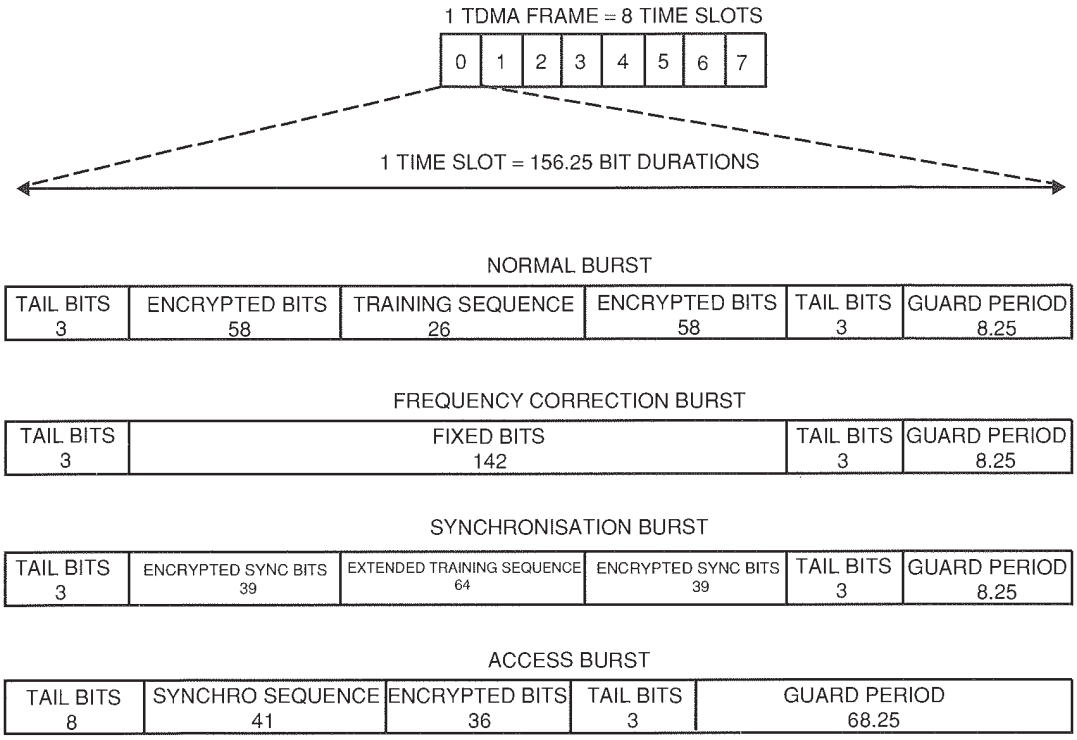


FIGURE 27.6 GSM burst structures, © ETT [4].

channel transmits SB hosting a $16 \cdot 4 = 64$-b extended sequence exhibiting a high-correlation peak in order to allow frame alignment with a quarter-bit accuracy. Furthermore, the SB contains $2 \cdot 39 = 78$ encrypted FEC-coded synchronization bits, hosting the BS and PLMN color codes, each representing one of eight legitimate identifiers. Lastly, the AB contain an extended 41-b synchronization sequence, and they are invoked to facilitate initial access to the system. Their long guard space of 68.25-b duration prevents frame overlap, before the MS's distance, i.e., the propagation delay becomes known to the BS and could be compensated for by adjusting the MS's timing advance.

27.6 Synchronization Issues

Although some synchronization issues are standardized in recommendations R.05.02 and R.05.03, the GSM recommendations do not specify the exact BS-MS synchronization algorithms to be used, these are left to the equipment manufacturers. A unique set of timebase counters, however, is defined in order to ensure perfect BS-MS synchronism. The BS sends FCB and SB on specific timeslots of the BCCH carrier to the MS to ensure that the MS's frequency standard is perfectly aligned with that of the BS, as well as to inform the MS about the required initial state of its internal counters. The MS transmits its uniquely numbered traffic and control bursts staggered by three timeslots with respect to those of the BS to prevent simultaneous MS transmission and reception, and also takes into account the required timing advance (**TA**) to cater for different BS-MS-BS round-trip delays.

The timebase counters used to uniquely describe the internal timing states of BSs and MSs are the quarter-bit number ($\boldsymbol{QN} = 0$–624) counting the quarter-bit intervals in bursts,

bit number ($\boldsymbol{BN}$ = 0–156), timeslot number ($\boldsymbol{TN}$ = 0–7) and TDMA Frame Number ($FN = 0\text{–}26 \cdot 51 \cdot 2048$), given in the order of increasing interval duration. The MS sets up its timebase counters after receiving a SB by determining QN from the 64-b extended training sequence in the center of the SB, setting $TN = 0$ and decoding the 78-encrypted, protected bits carrying the 25-SCH control bits.

The SCH carries frame synchronization information as well as BS identification information to the MS, as seen in Fig. 27.7, and it is provided solely to support the operation of the radio subsystem. The first 6 b of the 25-b segment consist of three PLMN color code bits

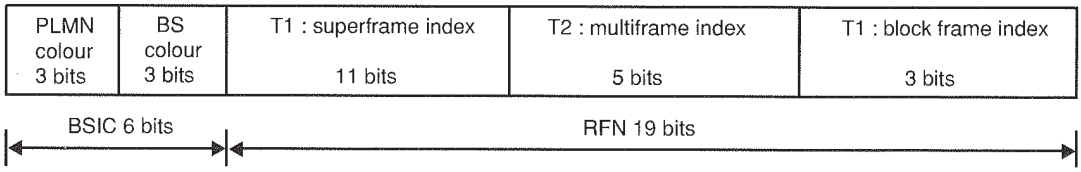


FIGURE 27.7 Synchronization channel (SCH) message format, ©ETT [4].

and three BS color code bits supplying a unique BS identifier code (BSIC) to inform the MS which BS it is communicating with. The second 19-bit segment is the so-called reduced TDMA frame number **RFN** derived from the full TDMA frame number FN, constrained to the range of $[0\text{–}(26 \cdot 51 \cdot 2048) - 1] = (0\text{–}2{,}715{,}647)$ in terms of three subsegments $T1$, $T2$, and $T3$. These subsegments are computed as follows: $T1(11\,\text{b}) = [FN \text{ div } (26 \cdot 51)]$, $T2(5\,\text{b}) = (FN \bmod 26)$ and $T3'(3\text{b}) = [(T3 - 1) \text{ div } 10]$, where $T3 = (FN \bmod 5)$, whereas div and mod represent the integer division and modulo operations, respectively. Explicitly, in Fig. 27.7 $T1$ determines the superframe index in a hyperframe, $T2$ the multiframe index in a superframe, $T3$ the frame index in a multiframe, whereas $T3'$ is the so-called signalling block index [1–5] of a frame in a specific 51-frame control multiframe, and their roles are best understood by referring to Fig. 27.2. Once the MS has received the SB, it readily computes the FN required in various control algorithms, such as encryption, handover, etc., as

$$FN = 51\,[(T3 - T2) \bmod 26] + T3 + 51 \cdot 26 \cdot T1, \qquad \text{where } T3 = 10 \cdot T3' + 1$$

27.7 Gaussian Minimum Shift Keying Modulation

The GSM system uses constant envelope partial response GMSK modulation [6] specified in recommendation R.05.04. Constant envelope, continuous-phase modulation schemes are robust against signal fading as well as interference and have good spectral efficiency. The slower and smoother are the phase changes, the better is the spectral efficiency, since the signal is allowed to change less abruptly, requiring lower frequency components. The effect of an input bit, however, is spread over several bit periods, leading to a so-called partial response system, which requires a channel equalizer in order to remove this controlled, intentional intersymbol interference (ISI) even in the absence of uncontrolled channel dispersion.

The widely employed partial response GMSK scheme is derived from the full response minimum shift keying (MSK) scheme. In MSK the phase changes between adjacent bit periods are piecewise linear, which results in discontinuous-phase derivative, i.e., instantaneous frequency at the signalling instants, and hence widens the spectrum. Smoothing these phase changes, however, by a filter having a Gaussian impulse response [6], which

is known to have the lowest possible bandwidth, this problem is circumvented using the schematic of Fig. 27.8, where the GMSK signal is generated by modulating and adding two quadrature carriers. The key parameter of GMSK in controlling both bandwidth and interference resistance is the 3-dB down filter-bandwidth $\times$ bit interval product $(B \cdot T)$, referred to as normalized bandwidth. It was found that as the $B \cdot T$ product is increased from 0.2 to 0.5, the interference resistance is improved by approximately 2 dB at the cost of increased bandwidth occupancy, and best compromise was achieved for $B \cdot T = 0.3$. This corresponds to spreading the effect of 1 b over approximately 3-b intervals. The spectral efficiency gain due to higher interference tolerance and, hence, more dense frequency reuse was found to be more significant than the spectral loss caused by wider GMSK spectral lobes.

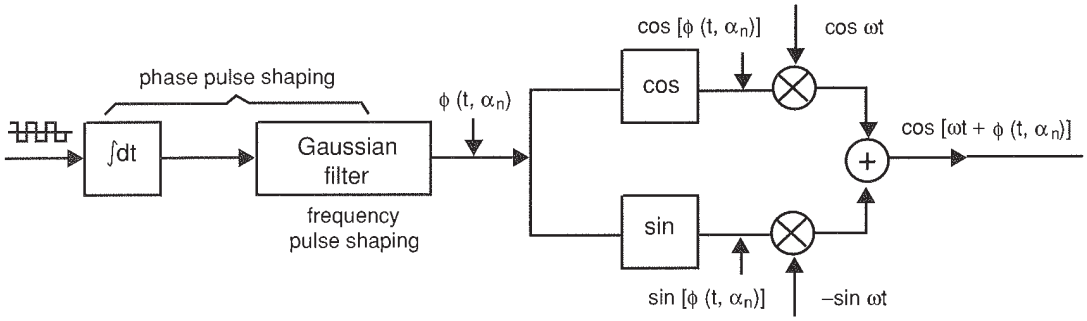


FIGURE 27.8 GMSK modulator schematic diagram, ©ETT [4].

The channel separation at the TDMA burst rate of 271 kb/s is 200 kHz, and the modulated spectrum must be 40 dB down at both adjacent carrier frequencies. When TDMA bursts are transmitted in an on-off keyed mode, further spectral spillage arises, which is mitigated by a smooth power ramp up and down envelope at the leading and trailing edges of the transmission bursts, attenuating the signal by 70 dB during a 28- and 18-μs interval, respectively.

27.8 Wideband Channel Models

The set of 6-tap GSM impulse responses [2] specified in recommendation R.05.05 is depicted in Fig. 27.9, where the individual propagation paths are independent Rayleigh fading paths, weighted by the appropriate coefficients h_i corresponding to their relative powers portrayed in the figure. In simple terms the wideband channel's impulse response is measured by transmitting an impulse and detecting the received echoes at the channel's output in every D-spaced so-called delay bin. In some bins no delayed and attenuated multipath component is received, whereas in others significant energy is detected, depending on the typical reflecting objects and their distance from the receiver. The path delay can be easily related to the distance of the reflecting objects, since radio waves are travelling at the speed of light. For example, at a speed of 300,000 km/s, a reflecting object situated at a distance of 0.15 km yields a multipath component at a round-trip delay of 1 μs.

The typical urban (**TU**) impulse response spreads over a delay interval of 5 μs, which is almost two 3.69-μs bit-intervals duration and, therefore, results in serious ISI. In simple terms, it can be treated as a two-path model, where the reflected path has a length of 0.75 km, corresponding to a reflector located at a distance of about 375 m. The hilly

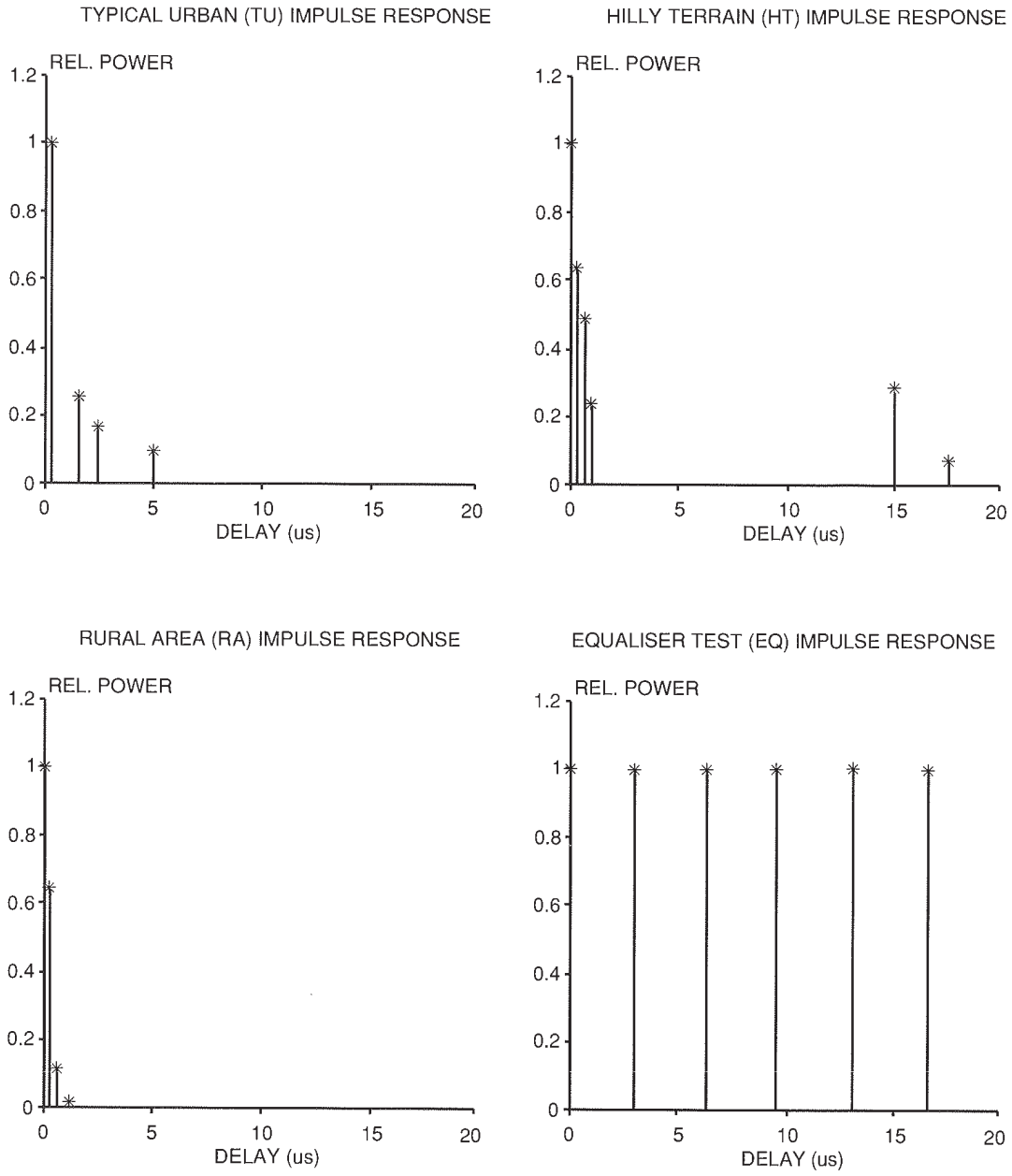


FIGURE 27.9 Typical GSM channel impulse responses, ©ETT [4].

terrain (HT) model has a sharply decaying short-delay section due to local reflections and a long-delay path around 15 μs due to distant reflections. Therefore, in practical terms it can be considered a two- or three-path model having reflections from a distance of about 2 km. The rural area (**RA**) response seems the least hostile amongst all standardized responses, decaying rapidly inside 1-b interval and, therefore, is expected to be easily combated by the channel equalizer. Although the type of the equalizer is not standardized, partial response systems typically use VEs. Since the RA channel effectively behaves as a single-path nondispersive channel, it would not require an equalizer. The fourth standardized

impulse response is artificially contrived in order to test the equalizer's performance and is constituted by six equidistant unit-amplitude impulses representing six equal-powered independent Rayleigh-fading paths with a delay spread over 16 μs. With these impulse responses in mind, the required channel is simulated by summing the appropriately delayed and weighted received signal components. In all but one case the individual components are assumed to have Rayleigh amplitude distribution, whereas in the RA model the main tap at zero delay is supposed to have a Rician distribution with the presence of a dominant line-of-sight path.

27.9 Adaptive Link Control

The adaptive link control algorithm portrayed in Fig. 27.10 and specified in recommendation R.05.08 allows for the MS to favor that specific traffic cell which provides the highest probability of reliable communications associated with the lowest possible path loss. It also decreases interference with other cochannel users and, through dense frequency reuse, improves spectral efficiency, whilst maintaining an adequate communications quality, and facilitates a reduction in power consumption, which is particularly important in hand-held MSs. The handover process maintains a call in progress as the MS moves between cells, or when there is an unacceptable transmission quality degradation caused by interference, in which case an intracell handover to another carrier in the same cell is performed. A radio-link failure occurs when a call with an unacceptable voice or data quality cannot be improved either by RF power control or by handover. The reasons for the link failure may be loss of radio coverage or very high-interference levels. The link control procedures rely on measurements of the received RF signal strength (**RXLEV**), the received signal quality (**RXQUAL**), and the absolute distance between base and mobile stations (DISTANCE).

RXLEV is evaluated by measuring the received level of the BCCH carrier which is continuously transmitted by the BS on all time slots of the B frames in Fig. 27.5 and without variations of the RF level. A MS measures the received signal level from the serving cell and from the BSs in all adjacent cells by tuning and listening to their BCCH carriers. The root mean squared level of the received signal is measured over a dynamic range from -103 to -41 dBm for intervals of one SACCH multiframe (480 ms). The received signal level is averaged over at least 32 SACCH frames ($\approx$15 s) and mapped to give RXLEV values between 0 and 63 to cover the range from -103 to -41 dBm in steps of 1 dB. The RXLEV parameters are then coded into 6-b words for transmission to the serving BS via the SACCH.

RXQUAL is estimated by measuring the bit error ratio (**BER**) before channel decoding, using the Viterbi channel equalizer's metrics [6] and/or those of the Viterbi convolutional decoder [8]. Eight values of RXQUAL span the logarithmically scaled BER range of 0.2–12.8% before channel decoding.

The absolute DISTANCE between base and mobile stations is measured using the timing advance parameter. The timing advance is coded as a 6-b number corresponding to a propagation delay from 0 to $63 \cdot 3.69$ μs $= 232.6$ μs, characteristic of a cell radius of 35 km.

While roaming, the MS needs to identify which potential target BS it is measuring, and the BCCH carrier frequency may not be sufficient for this purpose, since in small cluster sizes the same BCCH frequency may be used in more than one surrounding cell. To avoid ambiguity a 6-b BSIC is transmitted on each BCCH carrier in the SB of Fig. 27.6. Two other parameters transmitted in the BCCH data provide additional information about the BS. The binary flag called **PLMN_PERMITTED** indicates whether the measured BCCH carrier belongs to a PLMN that the MS is permitted to access. The second Boolean flag, **CELL_BAR_ACCESS,** indicates whether the cell is barred for access by the MS, although

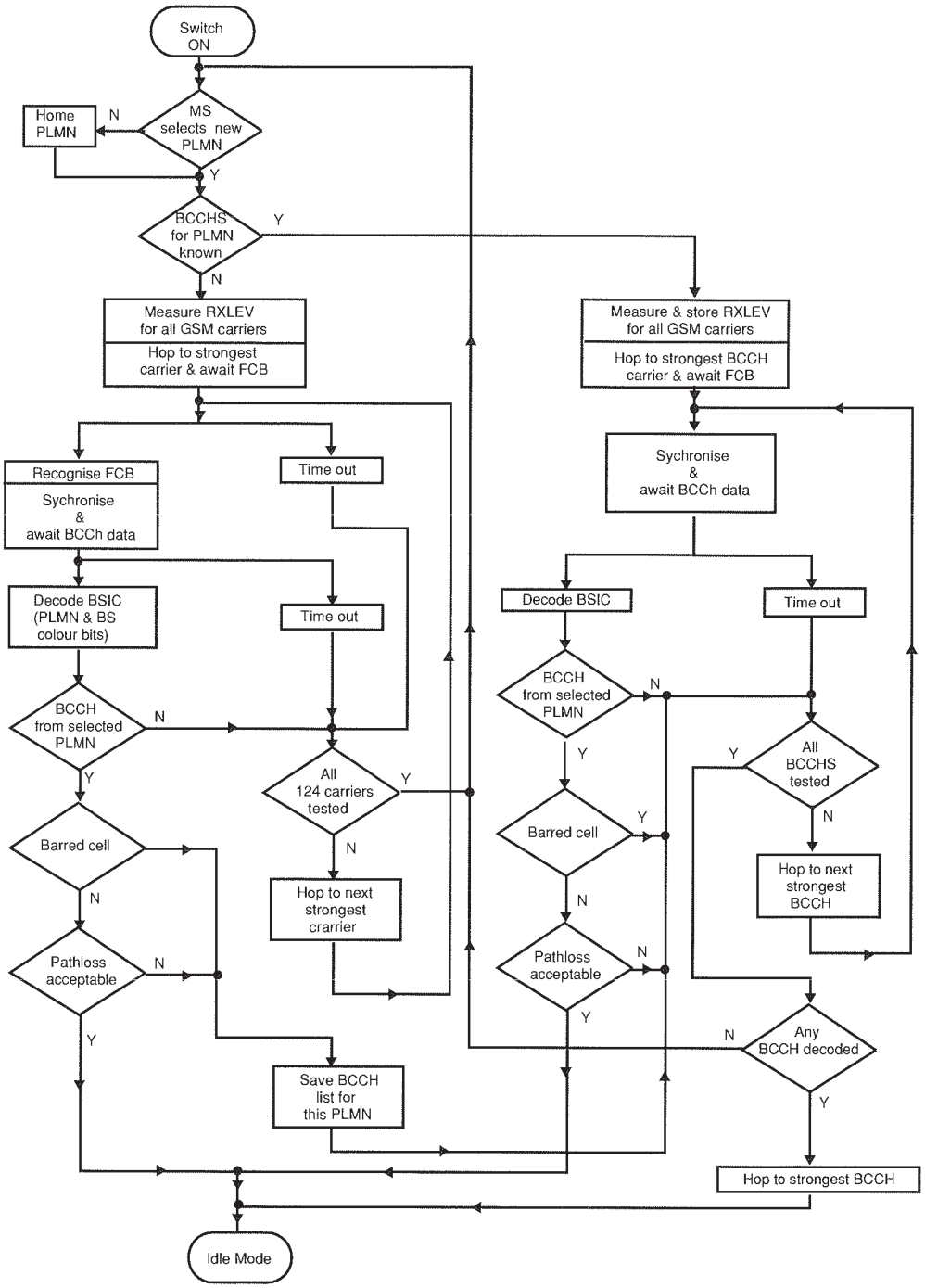


FIGURE 27.10 Initial cell selection by the MS, ©ETT [4].

it belongs to a permitted PLMN. A MS in idle mode, i.e., after it has just been switched on or after it has lost contact with the network, searches all 125 RF channels and takes readings of RXLEV on each of them. Then it tunes to the carrier with the highest RXLEV and searches for FCB in order to determine whether or not the carrier is a BCCH carrier. If it is not, then the MS tunes to the next highest carrier, and so on, until it finds a BCCH carrier, synchronizes to it and decodes the parameters BSIC, PLMN_PERMITTED and CELL_BAR_ ACCESS in order to decide whether to continue the search. The MS may store the BCCH carrier frequencies used in the network accessed, in which case the search time would be reduced. Again, the process described is summarized in the flowchart of Fig. 27.10.

The adaptive power control is based on RXLEV measurements. In every SACCH multiframe the BS compares the RXLEV readings reported by the MS or obtained by the base station with a set of thresholds. The exact strategy for RF power control is determined by the network operator with the aim of providing an adequate quality of service for speech and data transmissions while keeping interferences low. Clearly, adequate quality must be achieved at the lowest possible transmitted power to keep cochannel interferences low, which implies contradictory requirements in terms of transmitted power. The criteria for reporting radio link failure are based on the measurements of RXLEV and RXQUAL performed by both the mobile and base stations, and the procedures for handling link failures result in the re-establishment or the release of the call, depending on the network operator's strategy.

The handover process involves the most complex set of procedures in the radio-link control. Handover decisions are based on results of measurements performed both by the base and mobile stations. The base station measures RXLEV, RXQUAL, DISTANCE, and also the interference level in unallocated time slots, whereas the MS measures and reports to the BS the values of RXLEV and RXQUAL for the serving cell and RXLEV for the adjacent cells. When the MS moves away from the BS, the RXLEV and RXQUAL parameters for the serving station become lower, whereas RXLEV for one of the adjacent cells increases.

27.10 Discontinuous Transmission

Discontinuous transmission (DTX) issues are standardized in recommendation R.06.31, whereas the associated problems of voice activity detection VAD are specified by R.06.32. Assuming an average speech activity of 50% and a high number of interferers combined with frequency hopping to randomize the interference load, significant spectral efficiency gains can be achieved when deploying discontinuous transmissions due to decreasing interferences, while reducing power dissipation as well. Because of the reduction in power consumption, full DTX operation is mandatory for MSs, but in BSs, only receiver DTX functions are compulsory.

The fundamental problem in voice activity detection is how to differentiate between speech and noise, while keeping false noise triggering and speech spurt clipping as low as possible. In vehicle-mounted MSs the severity of the speech/noise recognition problem is aggravated by the excessive vehicle background noise. This problem is resolved by deploying a combination of threshold comparisons and spectral domain techniques [1, 3]. Another important associated problem is the introduction of noiseless inactive segments, which is mitigated by comfort noise insertion (**CNI**) in these segments at the receiver.

27.11 Summary

Following the standardization and launch of the GSM system its salient features were summarized in this brief review. Time division multiple access (TDMA) with eight users per carrier is used at a multiuser rate of 271 kb/s, demanding a channel equalizer to combat dispersion in large cell environments. The error protected chip rate of the full-rate traffic channels is 22.8 kb/s, whereas in half-rate channels it is 11.4 kb/s. Apart from the full- and half-rate speech traffic channels, there are 5 different rate data traffic channels and 14 various control and signalling channels to support the system's operation. A moderately complex, 13 kb/s regular pulse excited speech codec with long term predictor (**LTP**) is used, combined with an embedded three-class error correction codec and multilayer interleaving to provide sensitivity-matched unequal error protection for the speech bits. An overall speech delay of 57.5 ms is maintained. Slow frequency hopping at 217 hops/s yields substantial performance gains for slowly moving pedestrians.

TABLE 27.3 Summary of GSM Features

System feature	Specification
Up-link bandwidth, MHz	890–915 = 25
Down-link bandwidth, MHz	935–960 = 25
Total GSM bandwidth, MHz	50
Carrier spacing, KHz	200
No. of RF carriers	125
Multiple access	TDMA
No. of users/carrier	8
Total No. of channels	1000
TDMA burst rate, kb/s	271
Modulation	GMSK with BT = 0.3
Bandwidth efficiency, b/s/Hz	1.35
Channel equalizer	yes
Speech coding rate, kb/s	13
FEC coded speech rate, kb/s	22.8
FEC coding	Embedded block/ convolutional
Frequency hopping, hop/s	217
DTX and VAD	yes
Maximum cell radius, km	35

Constant envelope partial response GMSK with a channel spacing of 200 kHz is deployed to support 125 duplex channels in the 890–915-MHz up-link and 935–960-MHz down-link bands, respectively. At a transmission rate of 271 kb/s a spectral efficiency of 1.35-bit/s/Hz is achieved. The controlled GMSK-induced and uncontrolled channel-induced intersymbol interferences are removed by the channel equalizer. The set of standardized wideband GSM channels was introduced in order to provide bench markers for performance comparisons. Efficient power budgeting and minimum cochannel interferences are ensured by the combination of adaptive power and handover control based on weighted averaging of up to eight up-link and down-link system parameters. Discontinuous transmissions assisted by reliable spectral-domain voice activity detection and comfort-noise insertion further reduce interferences and power consumption. Because of ciphering, no unprotected information is sent via

the radio link. As a result, spectrally efficient, high-quality mobile communications with a variety of services and international roaming is possible in cells of up to 35 km radius for signal-to-noise and interference ratios in excess of 10–12 dBs. The key system features are summarized in Table 27.3.

Defining Terms

A3: Authentication algorithm
A5: Cyphering algorithm
A8: Confidential algorithm to compute the cyphering key
AB: Access burst
ACCH: Associated control channel
ADC: Administration center
AGCH: Access grant control channel
AUC: Authentication center
AWGN: Additive Gaussian noise
BCCH: Broadcast control channel
BER: Bit error ratio
BFI: Bad frame indicator flag
***BN*:** Bit number
BS: Base station
BS-PBGT: BS powerbudget: to be evaluated for power budget motivated handovers
BSIC: Base station identifier code
CC: Convolutional codec
CCCH: Common control channel
CELL_BAR_ACCESS: Boolean flag to indicate, whether the MS is permitted to access the specific traffic cell
CNC: Comfort noise computation
CNI: Comfort noise insertion
CNU: Comfort noise update state in the DTX handler
DB: Dummy burst
DL: Down link
DSI: Digital speech interpolation to improve link efficiency
DTX: Discontinuous transmission for power consumption and interference reduction
EIR: Equipment identity register
EOS: End of speech flag in the DTX handler
FACCH: Fast associated control channel
FCB: Frequency correction burst
FCCH: Frequency correction channel
FEC: Forward error correction
FH: Frequency hopping
***FN*:** TDMA frame number
GMSK: Gaussian minimum shift keying

GP: Guard space

HGO: Handover in the VAD

HLR: Home location register

HO: Handover

HOCT: Handover counter in the VAD

HO_MARGIN: Handover margin to facilitate hysteresis

HSN: Hopping sequence number: frequency hopping algorithm's input variable

IMSI: International mobile subscriber identity

ISDN: Integrated services digital network

LAI: Location area identifier

LAR: Logarithmic area ratio

LTP: Long term predictor

MA: Mobile allocation: set of legitimate RF channels, input variable in the frequency hopping algorithm

MAI: Mobile allocation index: output variable of the FH algorithm

MAIO: Mobile allocation index offset: initial RF channel offset, input variable of the FH algorithm

MS: Mobile station

MSC: Mobile switching center

MSRN: Mobile station roaming number

MS_TXPWR_MAX: Maximum permitted MS transmitted power on a specific traffic channel in a specific traffic cell

MS_TXPWR_MAX(n): Maximum permitted MS transmitted power on a specific traffic channel in the nth adjacent traffic cell

NB: Normal burst

NMC: Network management center

NUFR: Receiver noise update flag

NUFT: Noise update flag to ask for SID frame transmission

OMC: Operation and maintenance center

PARCOR: Partial correlation

PCH: Paging channel

PCM: Pulse code modulation

PIN: Personal identity number for MSs

PLMN: Public land mobile network

PLMN_ PERMITTED: Boolean flag to indicate whether the MS is permitted to access the specific PLMN

PSTN: Public switched telephone network

***QN*:** Quarter bit number

R: Random number in the authentication process

RA: Rural area channel impulse response

RACH: Random access channel

RF: Radio frequency

RFCH: Radio frequency channel

RFN: Reduced TDMA frame number: equivalent representation of the TDMA frame number that is used in the synchronization channel
RNTABLE: Random number table utilized in the frequency hopping algorithm
RPE: Regular pulse excited
RPE-LTP: Regular pulse excited codec with long term predictor
RS-232: Serial data transmission standard equivalent to CCITT V24. interface
RXLEV: Received signal level: parameter used in handovers
RXQUAL: Received signal quality: parameter used in handovers
S: Signed response in the authentication process
SACCH: Slow associated control channel
SB: Synchronization burst
SCH: Synchronization channel
SCPC: Single channel per carrier
SDCCH: Stand-alone dedicated control channel
SE: Speech extrapolation
SID: Silence identifier
SIM: Subscriber identity module in MSs
SPRX: Speech received flag
SPTX: Speech transmit flag in the DTX handler
STP: Short term predictor
TA: Timing advance
TB: Tailing bits
TCH: Traffic channel
TCH/F: Full-rate traffic channel
TCH/F2.4: Full-rate 2.4-kb/s data traffic channel
TCH/F4.8: Full-rate 4.8-kb/s data traffic channel
TCH/F9.6: Full-rate 9.6-kb/s data traffic channel
TCH/FS: Full-rate speech traffic channel
TCH/H: Half-rate traffic channel
TCH/H2.4: Half-rate 2.4-kb/s data traffic channel
TCH/H4.8: Half-rate 4.8-kb/s data traffic channel
TDMA: Time division multiple access
TMSI: Temporary mobile subscriber identifier
***TN*:** Time slot number
TU: Typical urban channel impulse response
TXFL: Transmit flag in the DTX handler
UL: Up link
VAD: Voice activity detection
VE: Viterbi equalizer
VLR: Visiting location register

References

[1] European Telecommunications Standardization Institute. Group Speciale Mobile or Global System of Mobile Communication (GSM) Recommendation, ETSI Secretariat, Sophia Antipolis Cedex, France, 1988.

[2] Greenwood, D. and Hanzo, L., Characterisation of mobile radio channels, In *Mobile Radio Communications.* Steele, R., Ed., Chap. 2, 92–185. IEEE Press–Pentech Press, London, 1992.

[3] Hanzo, L. and Stefanov, J., The Pan-European digital cellular mobile radio system—known as GSM. In *Mobile Radio Communications,* Steele, R., Ed., Chap. 8, 677–773, IEEE Press–Pentech Press, London, 1992.

[4] Hanzo, L. and Steele, R., The Pan-European mobile radio system, Pts. 1 and 2, *European Trans. on Telecomm.,* 5(2), 245–276, 1994.

[5] Salami, R.A., Hanzo, L., et al., Speech coding. In *Mobile Radio Communications,* Steele, R., Ed., Chap. 3, 186–346. IEEE Press–Pentech Press, London, 1992.

[6] Steele, R. Ed., *Mobile Radio Communications,* IEEE Press–Pentech Press, London, 1992.

[7] Vary, P. and Sluyter, R.J., MATS-D speech codec: Regular-pulse excitation LPC, *Proceedings of Nordic Conference on Mobile Radio Communications.* 257–261, 1986.

[8] Wong, K.H.H. and Hanzo, L., Channel coding. In *Mobile Radio Communications.* Steele, R., Ed., Chap. 4, 347–488. IEEE Press–Pentech Press, London, 1992.

28

Speech and Channel Coding for North American TDMA Cellular Systems

28.1 Introduction 28-1
28.2 Modulation of Digital Voice and Data Signals 28-2
28.3 Speech Coding Fundamentals 28-2
28.4 Channel Coding Considerations 28-4
28.5 VSELP Encoder 28-4
28.6 Linear Prediction Analysis and Quantization 28-5
28.7 Bandwidth Expansion 28-8
28.8 Quantizing and Encoding the Reflection Coefficients 28-8
28.9 VSELP Codebook Search 28-8
28.10 Long-Term Filter Search 28-9
28.11 Orthogonalization of the Codebooks 28-9
28.12 Quantizing the Excitation and Signal Gains 28-9
28.13 Channel Coding and Interleaving 28-10
28.14 Bad Frame Masking 28-11
28.15 ACELP Encoder 28-11
28.16 Algebraic Codebook Structure and Search 28-11
28.17 Quantization of the Gains for ACELP Encoding 28-12
28.18 Channel Coding for ACELP Encoding 28-12
28.19 Conclusions 28-12
Defining Terms 28-13
References 28-13
Further Information 28-14

Paul Mermelstein
INRS-Télécommunications
University of Québec

28.1 Introduction

The goals of this chapter are to give the reader a tutorial introduction and high-level understanding of the techniques employed for speech transmission by the IS-54 digital cellular standard. It builds on the information provided in the standards document but is not meant to be a replacement for it. Separate standards cover the control channel used for the setup of calls and their handoff to neighboring cells, as well as the encoding of data signals for transmission. For detailed implementation information the reader should consult the most recent standards document [9].

0-8493-8597-0/99/$0.00+$.50

© 1999 by CRC Press LLC

IS-54 provides for encoding bidirectional speech signals digitally and transmitting them over cellular and microcellular mobile radio systems. It retains the 30-kHz channel spacing of the earlier advanced mobile telephone service (AMPS), which uses analog frequency modulation for speech transmission and frequency shift keying for signalling. The two directions of transmission use frequencies some 45 MHz apart in the band between 824 and 894 MHz. AMPS employs one channel per conversation in each direction, a technique known as frequency division multiple access (FDMA). IS-54 employs time division multiple access (TDMA) by allowing three, and in the future six, simultaneous transmissions to share each frequency band. Because the overall 30-kHz channelization of the allocated 25 MHz of spectrum in each direction is retained, it is also known as a FDMA-TDMA system. In contrast, the later IS-95 standard employs code division multiple access (CDMA) over bands of 1.23 MHz by combining several 30-kHz frequency channels.

Each frequency channel provides for transmission at a digital bit rate of 48.6 kb/s through use of differential quadrature-phase shift key (DQPSK) modulation at a 24.3-kBd channel rate. The channel is divided into six time slots every 40 ms. The full-rate voice coder employs every third time slot and utilizes 13 kb/s for combined speech and channel coding. The six slots provide for an eventual half-rate channel occupying one slot per 40 ms frame and utilizing only about 6.5 kb/s for each call. Thus, the simultaneous call carrying capacity with IS-54 is increased by a factor 3(factor 6 in the future) above that of AMPS. All digital transmission is expected to result in a reduction in transmitted power. The resulting reduction in intercell interference may allow more frequent reuse of the same frequency channels than the reuse pattern of seven cells for AMPS. Additional increases in erlang capacity (the total call-carrying capacity at a given blocking rate) may be available from the increased trunking efficiency achieved by the larger number of simultaneously available channels. The first systems employing dual-mode AMPS and TDMA service were put into operation in 1993.

In 1996 the TIA introduced the IS-641 enhanced full rate codec. This codec consists of 7.4 kb/s speech coding following the algebraic code-excited linear prediction (ACELP) technique [7], and 5.6 kb/s channel coding. The 13 kb/s coded information replaces the combined 13 kb/s for speech and channel coding introduced by the IS-54 standard. The new codec provides significant enhancements in terms of speech quality and robustness to transmission errors. The quality enhancement for clear channels results from the improved modeling of the stochastic excitation by means of an algebraic **codebook** instead of the two trained VSELP codebooks. Improved robustness to transmission errors is achieved by employing predictive quantization techniques for the linear-prediction filter and gain parameters, and increasing the number of bits protected by forward error correction.

28.2 Modulation of Digital Voice and Data Signals

The modulation method used in IS-54 is $\pi/4$ shifted differentially encoded quadrature phase-shift keying (DPSK). Symbols are transmitted as changes in phase rather than their absolute values. The binary data stream is converted to two binary streams X_k and Y_k formed from the odd- and even-numbered bits, respectively. The quadrature streams I_k and Q_k are formed according to

$$\begin{aligned} I_k &= I_{k-1}\cos\left[\Delta\phi\left(X_k, Y_k\right)\right] - Q_{k-1}\sin\left[\Delta\phi\left(X_k, Y_k\right)\right] \\ Q_k &= I_{k-1}\sin\left[\Delta\phi\left(X_k, Y_k\right)\right] + Q_{k-1}\cos\left[\Delta\phi\left(X_k, Y_k\right)\right] \end{aligned}$$

where I_{k-1} and Q_{k-1} are the amplitudes at the previous pulse time. The phase change $\Delta\phi$ takes the values $\pi/4, 3\pi/4, -\pi/4$, and $-3\pi/4$ for the dibit (X_k, Y_k) symbols (0,0), (0,1),

(1,0) and (1,1), respectively. This results in a rotation by $\pi/4$ between the constellations for odd and even symbols. The differential encoding avoids the problem of 180° phase ambiguity that may otherwise result in estimation of the carrier phase.

The signals I_k and Q_k at the output of the differential phase encoder can take one of five values, $0, \pm 1, \pm 1/\sqrt{2}$ as indicated in the constellation of Fig. 28.1. The corresponding impulses are applied to the inputs of the I and Q baseband filters, which have linear phase and square root raised cosine frequency responses. The generic modulator circuit is shown in Fig. 28.2. The rolloff factor α determines the width of the transition band and its value is 0.35,

$$|H(f)| = \begin{cases} 1, & 0 \le f \le (1-\alpha)/2T \\ \sqrt{1/2\{1 - \sin[\pi(2fT-1)/2\alpha]\}}, & (1-\alpha)/2T \le f \le (1+\alpha)/2T \\ 0, & f > (1+\alpha)/2T \end{cases}$$

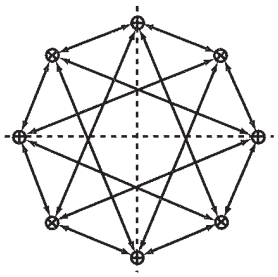


FIGURE 28.1 Constellation for $\pi/4$ shifted QPSK modulation. *Source*: TIA, 1992. Cellular System Dual-mode Mobile Station–Base Station Compatibility Standard TIA/EIA IS-54. With permission.

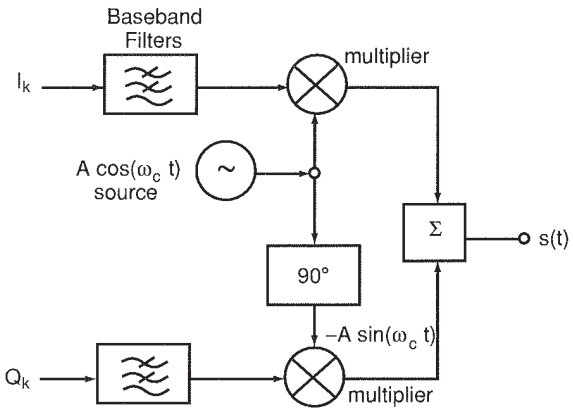


FIGURE 28.2 Generic modulation circuit for digital voice and data signals. *Source*: TIA, 1992. Cellular System Dual-mode Mobile Station–Base Station Compatibility Standard TIA/EIA IS-54.

28.3 Speech Coding Fundamentals

The IS-54 standard employs a vector-sum excited linear prediction (VSELP) coding technique. It represents a specific formulation of the much larger class of code-excited linear prediction (CELP) coders [2] that have proved effective in recent years for the coding of speech at moderate rates in the range 4–16 kb/s. VSELP provides reconstructed speech with a quality that is comparable to that available with frequency modulation and analog transmission over the AMPS system. The coding rate employed is 7.95 kb/s. Each of the six slots per frame carry 260 b of speech and channel coding information for a gross information rate of 13 kb/s. The 260 b correspond to 20 ms of real time speech, transmitted as a single burst.

For an excellent recent review of speech coding techniques for transmission, the reader is referred to Gersho, 1994 [3]. Most modern speech coders use a form of analysis by synthesis coding where the encoder determines the coded signal one segment at a time by feeding candidate excitation segments into a replica of a synthesis filter and selecting the segment that minimizes the distortion between the original and reproduced signals. Linear prediction coding (LPC) techniques [1] encode the speech signal by first finding an optimum linear filter to remove the short-time correlation, passing the signal through that LPC filter to obtain a residual signal, and encoding this residual using much fewer bits than would have been required to code the original signal with the same fidelity. In most cases the coding of the residual is divided into two steps. First, the long-time correlation due to the periodic pitch excitation is removed by means of an optimum one-tap filter with adjustable gain and lag. Next, the remaining residual signal, which now closely resembles a white-noise signal, is encoded. Code-excited linear predictors use one or more codebooks from which they select replicas of the residual of the input signal by means of a closed-loop error-minimization technique. The index of the codebook entry as well as the parameters of all the filters are transmitted to allow the speech signal to be reconstructed at the receiver. Most code-excited coders use trained codebooks. Starting with a codebook containing Guassian signal segments, entries that are found to be used rarely in coding a large body of speech data are iteratively eliminated to result in a smaller codebook that is considered more effective.

The speech signal can be considered quasistationary or stationary for the duration of the speech frame, of the order of 20 ms. The parameters of the short-term filter, the LPC coefficients, are determined by analysis of the autocorrelation function of a suitably windowed segment of the input signal. To allow accurate determination of the time-varying pitch lag as well as simplify the computations, each speech frame is divided into four 5-ms subframes. Independent pitch filter computations and residual coding operations are carried out for each subframe.

The speech decoder attempts to reconstruct the speech signal from the received information as best possible. It employs a codebook identical to that of the encoder for excitation generation and, in the absence of transmission errors, would produce an exact replica of the signal that produced the minimized error at the encoder. Transmission errors do occur, however, due, to signal fading and excessive interference. Since any attempt at retransmission would incur unacceptable signal delays, sufficient error protection is provided to allow correction of most transmission errors.

28.4 Channel Coding Considerations

The sharp limitations on available bandwidth for error protection argue for careful consideration of the sensitivity of the speech coding parameters to transmission errors. Pairwise

interleaving of coded blocks and convolutional coding of a subset of the parameters permit correction of a limited number of transmission errors. In addition, a cyclic redundancy check (CRC) is used to determine whether the error correction was successful. The coded information is divided into three blocks of varying sensitivity to errors. Group 1 contains the most sensitive bits, mainly the parameters of the LPC filter and frame energy, and is protected by both error detection and correction bits. Group 2 is provided with error correction only. The third group, comprising mostly the fixed codebook indices, is not protected at all.

The speech signal contains significant temporal redundancy. Thus, speech frames within which errors have been detected may be reconstructed with the aid of previously correctly received information. A bad-frame masking procedure attempts to hide the effects of short fades by extrapolating the previously received parameters. Of course, if the errors persist, the decoded signal must be muted while an attempt is made to hand off the connection to a base station to/from which the mobile may experience better reception.

28.5 VSELP Encoder

A block diagram of the VSELP speech encoder [4] is shown in Fig. 28.3. The excitation signal is generated from three components, the output of a long term or pitch filter, as well as entries from two codebooks. A weighted synthesis filter generates a synthesized approximation to the frequency-weighted input signal. The weighted mean square error between these two signals is used to drive the error minimization process. This weighted error is considered to be a better approximation to the perceptually important noise components than the unweighted mean square error. The total weighted square error is minimized by adjusting the pitch lag and the codebook indices as well as their gains. The decoder follows the encoder closely and generates the excitation signal identically to the encoder but uses an unweighted linear-prediction synthesis filter to generate the decoded signal. A spectral postfilter is added after the synthesis filter to enhance the quality of the reconstructed speech.

The precise data rate of the speech coder is 7950 b/s or 159 b per time slot, each corresponding to 20 ms of signal in real time. These 159 b are allocated as follows: 1) short-term filter coefficients, 38 bits; 2) frame energy, 5 bits; 3) pitch lag, 28 bits; 4) codewords, 56 bits; and 5) gain values, 32 bits.

28.6 Linear Prediction Analysis and Quantization

The purpose of the LPC analysis filter is to whiten the spectrum of the input signal so that it can be better matched by the codebook outputs. The corresponding LPC synthesis filter $A(z)$ restores the short-time speech spectrum characteristics to the output signal. The transfer function of the tenth-order synthesis filter is given by

$$A(z) = \frac{1}{1 - \sum_{i=1}^{N_p} \alpha_i z^{-i}}$$

The filter predictor parameters $\alpha_1, \ldots, \alpha_{N_p}$ are not transmitted directly. Instead, a set of **reflection coefficients** $r_1, \ldots, r_{N_p}$ are computed and quantized. The predictor parameters are determined from the reflection coefficients using a well-known backward recursion algorithm [6].

A variety of algorithms are known that determine a set of reflection coefficients from a windowed input signal. One such algorithm is the fixed point **covariance lattice**, FLAT,

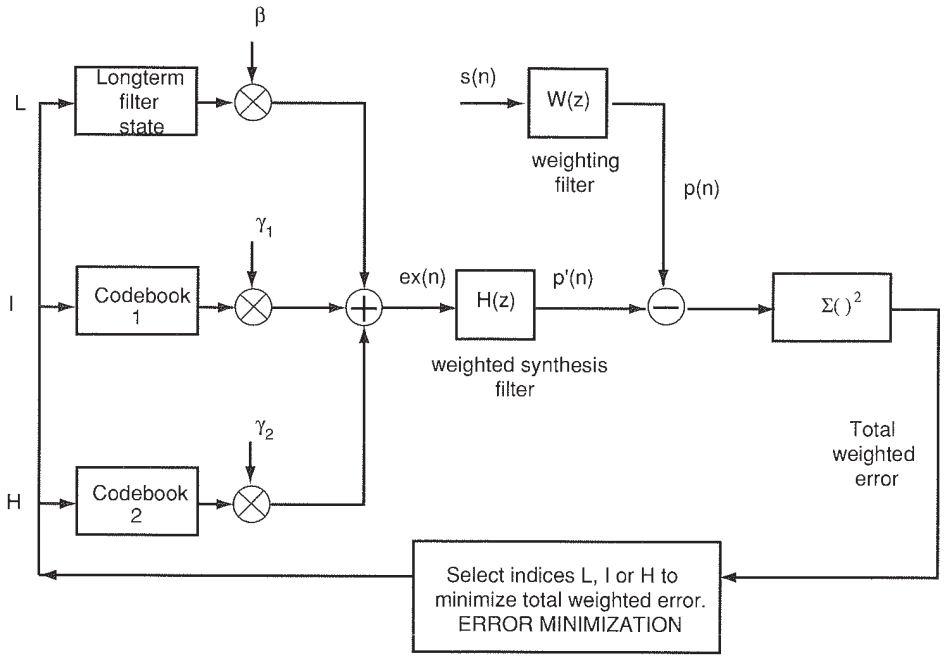


FIGURE 28.3 Black diagram of the speech encoder in VSELP. TIA. 1992. Cellular system Dual-mode Mobile Station–Base Station Compatibility Standard. TIA/EIA IS-54.

which builds an optimum inverse lattice stage by stage. At each stage j, the sum of the mean-squared forward and backward residuals is minimized by selection of the best reflection coefficient r_j. The analysis window used is 170 samples long, centered with respect to the middle of the fourth 5-ms subframe of the 20-ms frame. Since this centerpoint is 20 samples from the end of the frame, 65 samples from the next frame to be coded are used in computing the reflection coefficient of the current frame. This introduces a lookahead delay of 8.125 ms.

The FLAT algorithm first computes the covariance matrix of the input speech for $N_A = 170$ and $N_p = 10$,

$$\phi(i,k) = \sum_{n=N_p}^{N_A-1} s(n-i)s(n-k), \qquad 0 \le i, \quad k \le N_p \ ,$$

Define the forward residual out of stage j as $f_j(n)$ and the backward residual as $b_j(n)$. Then the autocorrelation of the initial forward residual $F_0(i,k)$ is given by $\phi(i,k)$. The autocorrelation of the initial backward residual $B_0(i,k)$ is given by $\phi(i+1,k+1)$ and the initial cross correlation of the two residuals is given by $C_0(i,k) = \phi(i,k+1)$ for $0 \le i,k \le N_{p-1}$. Initially j is set to 1. The **reflection coefficient** at each stage is determined as the ratio of the cross correlation to the mean of the autocorrelations. A block diagram of the computations is shown in Fig. 28.4. By quantizing the reflection coefficients within the computation loops, reflection coefficients at subsequent stages are computed taking into account the quantization errors of the previous stages. Specifically,

$$\begin{aligned} C'_{j-1} &= C_{j-1}(0,0) + C_{j-1}(N_p - j, N_p - j) \\ F'_{j-1} &= F_{j-1}(0,0) + F_{j-1}(N_p - j, N_p - j) \\ B'_{j-1} &= B_{j-1}(0,0) + B_{j-1}(N_p - j, N_p - j) \end{aligned}$$

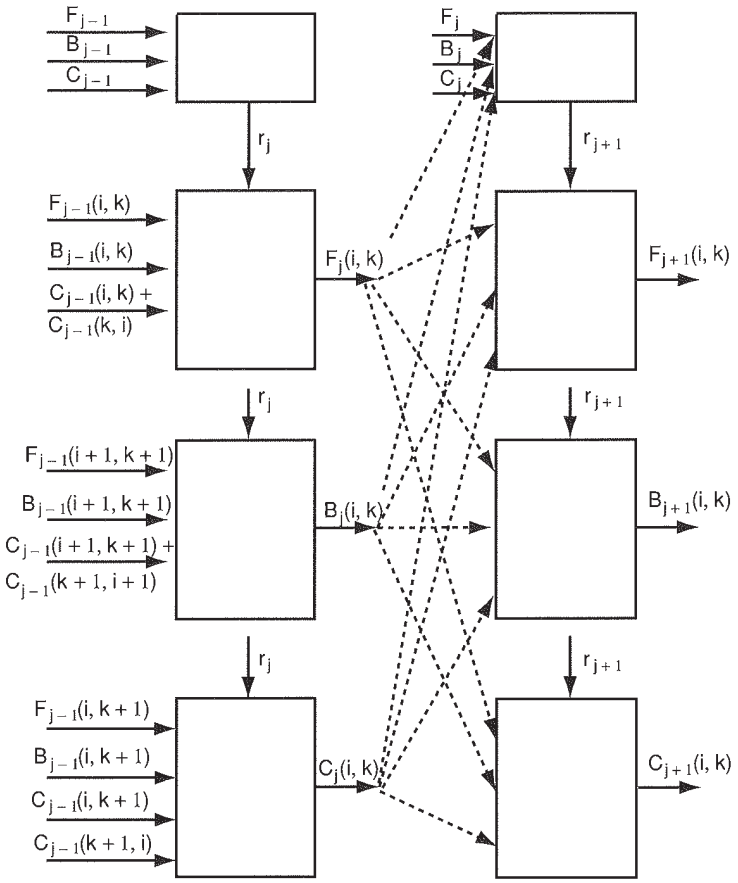


FIGURE 28.4 Block diagram for lattice covariance computations.

and

$$r_j = \frac{-2C'_{j-1}}{F'_{j-1} + B'_{j-1}}$$

Use of two sets of correlation values separated by $N_p - j$ samples provides additional stability to the computed reflection coefficients in case the input signal changes form rapidly.

Once a quantized reflection coefficient r_j has been determined, the resulting auto- and cross correlations can be determined iteratively as

$$\begin{aligned} F_j(i,k) &= F_{j-1}(i,k) + r_j[C_{j-1}(i,k) + C_{j-1}(k,i)] + r_j^2 B_{j-1}(i,k) \\ B_j(i,k) &= B_{j-1}(i+1,k+1) + r_j[C_{j-1}(i+1,k+1) + C_{j-1}(k+1,i+1)] \\ &\quad + r_j^2 F_{j-1}(i+1,k+1) \end{aligned}$$

and

$$\begin{aligned} C_j(i,k) &= C_{j-1}(i,k+1) + r_j[B_{j-1}(i,k+1) + F_{j-1}(i,k+1)] \\ &\quad + r_j^2 C_{j-1}(k+1,i) \end{aligned}$$

These computations are carried out iteratively for $r_j, j = 1, \ldots, N_p$.

28.7 Bandwidth Expansion

Poles with very narrow bandwidths may introduce undesirable distortions into the synthesized signal. Use of a binomial window with effective bandwidth of 80 Hz suffices to limit the ringing of the LPC filter and reduce the effect of the LPC filter selected for one frame on the signal reconstructed for subsequent frames. To achieve this, prior to searching for the reflection coefficients, the $\phi(i,k)$ is modified by use of a window function $w(j), j = 1, \ldots, 10$, as follows:

$$\phi'(i,k) = \phi(i,k)w(|i-k|)$$

28.8 Quantizing and Encoding the Reflection Coefficients

The distortion introduced into the overall spectrum by quantizing the reflection coefficients diminishes as we move to higher orders in the reflection coefficients. Accordingly, more bits are assigned to the lower order coefficients. Specifically, 6, 5, 5, 4, 4, 3, 3, 3, 3, and 2 b are assigned to $r_1, \ldots, r_{10}$, respectively. Scalar quantization of the reflection coefficients is used in IS-54 because it is particularly simple. **Vector quantization** achieves additional quantizing efficiencies at the cost of significant added complexity.

It is important to preserve the smooth time evolution of the linear prediction filter. Both the encoder and decoder linearly interpolate the coefficients α_i for the first, second and third subframes of each frame using the coefficients determined for the previous and current frames. The fourth subframe uses the values computed for that frame.

28.9 VSELP Codebook Search

The codebook search operation selects indices for the long-term filter (pitch lag L) and the two codebooks I and H so as to minimize the total weighted error. This closed-loop search is the most computationally complex part of the encoding operation, and significant effort has been invested to minimize the complexity of these operations without degrading performance. To reduce complexity, simultaneous optimization of the codebook selections is replaced by a sequential optimization procedure, which considers the long-term filter search as the most significant and therefore executes it first. The two vector-sum codebooks are considered to contribute less and less to the minimization of the error, and their search follows in sequence. Subdivision of the total codebook into two vector sums simplifies the processing and makes the result less sensitive to errors in decoding the individual bits arising from transmission errors.

Entries from each of the two vector-sum codebooks can be expressed as the sum of basis vectors. By orthogonalizing these basis vectors to the previously selected codebook component(s), one ensures that the newly introduced components reduce the remaining errors. The subframes over which the codebook search is carried out are 5 ms or 40 samples long. An optimal search would need exploration of a 40-dimensional space. The vector-sum approximation limits the search to 14 dimensions after the optimal pitch lag has been selected. The search is further divided into two stages of 7 dimensions each. The two codebooks are specified in terms of the fourteen, 40-dimensional basis vectors stored at the encoder and decoder. The two 7-b indices indicate the required weights on the basic vectors to arrive at the two optimum codewords.

The codebook search can be viewed as selecting the three best directions in 40-dimensional space, which when summed result in the best approximation to the weighted input signal.

The gains of the three components are determined through a separate error minimization process.

28.10 Long-Term Filter Search

The long-term filter is optimized by selection of a lag value that minimizes the error between the weighted input signal $p(n)$ and the past excitation signal filtered by the current weighted synthesis filter $H(z)$. There are 127 possible coded lag values provided corresponding to lags of 20–146 samples. One value is reserved for the case when all correlations between the input and the lagged residuals are negative and use of no long term filter output would be best. To simplify the convolution operation between the impulse response of the weighted synthesis filter and the past excitation, the impulse response is truncated to 21 samples or 2.5 ms. Once the lag is determined, the untruncated impulse response is used to compute the weighted long-term lag vector.

28.11 Orthogonalization of the Codebooks

Prior to the search of the first codebook, each filtered basis vector may be made orthogonal to the long-term filter output, the zero-state response of the weighted synthesis filter $H(z)$ to the long-term prediction vector. Each orthogonalized filtered basis vector is computed by subtracting its projection onto the long-term filter output from itself.

Similarly, the basis vectors of the second codebook can be orthogonalized with respect to both the long-term filter output and the first codebook output, the zero-state response of $H(z)$ to the previously selected summation of first-codebook basis vectors. In each case the codebook excitation can be reconstituted as

$$u_{k,i}(n) = \sum_{m=1}^{M} \theta_{im} v_{k,m}(n)$$

where $k = 1, 2$ for the two codebooks, $i = I$ or H the 7-b code vector received, $v_{k,m}$ are the two sets of basis vectors, and $\theta_{im} = +1$ if bit m of codeword $i = 1$ and -1 if bit m of codeword $i = 0$. Orthogonalization is not required at the decoder since the gains of the codebooks outputs are determined with respect to the weighted nonorthogonalized code vectors.

28.12 Quantizing the Excitation and Signal Gains

The three codebook gain values β, γ_1, and γ_2 are transformed to three new parameters $GS, P0$ and $P1$ for quantization purposes. GS is an energy offset parameter that equalizes the input and output signal energies. It adjusts the energy of the output of the LPC synthesis filter to equal the energy computed for the same subframe at the encoder input. $P0$ is the energy contribution of the long-term prediction vector as a fraction of the total excitation energy within the subframe. Similarly, $P1$ is the energy contribution of the code vector selected from the first codebook as a fraction of the total excitation energy of the subframe. The transformation reduces the dynamic range of the parameters to be encoded. An 8-b **vector quantizer** efficiently encodes the appropriate $(GS, P0, P1)$ vectors by selecting the vector which minimizes the weighted error. The received and decoded values β, γ_1, and γ_2 are computed from the received $(GS, P0, P1)$ vector and applied to reconstitute the decoded signal.

28.13 Channel Coding and Interleaving

The goals of channel coding are to reduce the impairments in the reconstructed speech due to transmission errors. The 159 b characterizing each 20-ms block of speech are divided into two classes, 77 in class 1 and 82 in class 2. Class 1 includes the bits in which errors result in a more significant impairment, whereas the speech quality is considered less sensitive to the class- 2 bits. Class 1 generally includes the gain, pitch lag, and more significant reflection coefficient bits. In addition, a 7-b cyclic redundancy check is applied to the 12 most perceptually significant bits of class 1 to indicate whether the error correction was successful. Failure of the CRC check at the receiver suggests that the received information is so erroneous that it would be better to discard it than use it. The error correction coding is illustrated in Fig. 28.5.

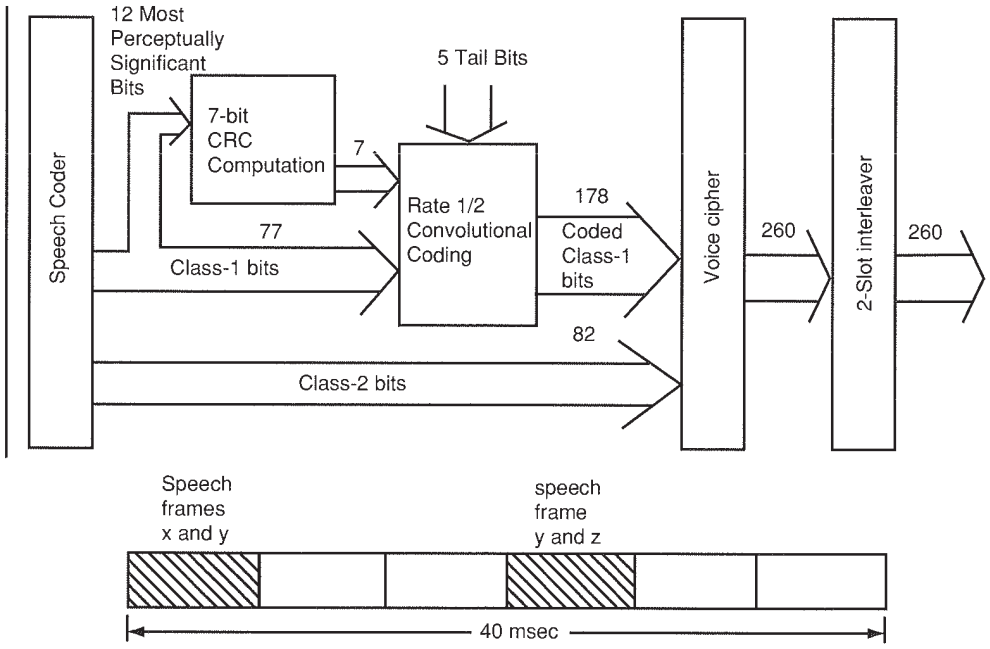


FIGURE 28.5 Error correction insertion for speech coder. Source TIA, 1992. Cellular Systems Dual-Mode Mobile Station–Base Station Compatibility Standards. TIA/EIA IS-54. With permission.

The error correction technique used is rate 1/2 convolutional coding with a constraint length of 5 [5]. A tail of 5 b is appended to the 84 b to be convolutionally encoded to result in a 178-b output. Inclusion of the tail bits ensures independent decoding of successive time slots and no propagation of errors between slots.

Interleaving the bits to be transmitted over two time slots is introduced to diminish the effects of short deep fades and to improve the error-correction capabilities of the channel coding technique. Two speech frames, the previous and the present, are interleaved so that the bits from each speech block span two transmission time slots separated by 20 ms. The interleaving attempts to separate the convolutionally coded class-1 bits from one frame as much as possible in time by inserting noncoded class-2 bits between them.

28.14 Bad Frame Masking

A CRC failure indicates that the received data is unusable, either due to transmission errors resulting from a fade, or from pre-emption of the time slot by a control message (fast associated control channel, FACCH). To mask the effects that may result from leaving a gap in the speech signal, a masking operation based on the temporal redundancy between adjacent speech blocks has been proposed. Such masking can at best bridge over short gaps but cannot recover loss of signal of longer duration. The bad frame masking operation may follow a finite state machine where each state indicates an operation appropriate to the elapsed duration of the fade to which it corresponds. The masking operation consists of copying the previous LPC information and attenuating the gain of the signal. State 6 corresponds to error sequences exceeding 100 ms, for which the output signal is muted. The result of such a masking operation is generation of an extrapolation in the gap to the previously received signal, significantly reducing the perceptual effects of short fades. No additional delay is introduced in the reconstructed signal. At the same time, the receiver will report a high frequency of bad frames leading the system to explore handoff possibilities immediately. A quick successful handoff will result in rapid signal recovery.

28.15 ACELP Encoder

The ACELP encoder employs linear prediction analysis and quantization techniques similar to those used in VSELP and discussed in Section 28.6. The frame structure of 20 ms frames and 5 ms subframes is preserved. Linear prediction analysis is carried out for every frame. The ACELP encoder uses a long-term filter similar to the one discussed in Section 28.10 and represented as an adaptive codebook. The nonpredictable part of the LPC residual is represented in terms of ACELP codebooks, which replace the two VSELP codebooks shown in Fig. 28.3.

Instead of encoding the reflection coefficients as in VSELP, the information is transformed into line-spectral frequency pairs (LSP) [8]. The LSPs can be derived from linear prediction coefficients, a 10th order analysis generating 10 line-spectral frequencies (LSF), 5 poles, and 5 zeroes. The LSFs can be vector quantized and the LPC coefficients recalculated from the quantized LSFs. As long as the interleaved order of the poles and zeroes is preserved, quantization of the LSPs preserves the stability of the LPC synthesis filters. The LSPs of any frame can be better predicted from the values calculated and transmitted corresponding to previous frames, resulting in additional advantages. The long-term means of the LSPs are calculated for a large body of speech data and stored at both the encoder and decoder. First-order moving-average prediction is then used for the mean-removed LSPs. The time-prediction technique also permits use of predicted values for the LSPs in case uncorrectable transmissions errors are encountered, resulting in reduced speech degradation. To simplify the vector quantization operations, each LSP vector is split into 3 subvectors of dimensions 3, 3, and 4. The three subvectors are quantized with 8, 9, and 9 bits respectively, corresponding to a total bit assignment of 26 bits per frame for LPC information.

28.16 Algebraic Codebook Structure and Search

Algebraic codebooks contain relatively few pulses having nonzero values leading to rapid search of the possible innovation vectors, the vectors which together with the ACB output form the excitation of the LPC filter for the current subframe. In this implementation the 40-position innovation vector contains only four nonzero pulses and each can take on only

values +1 and −1. The 40 positions are divided into four tracks and one pulse is selected from each track. The tracks are generally equally spaced but differ in their starting value, thus the first pulse can take on positions 0, 5, 10, 15, 20, 25, 30, or 35 and the second has possible positions 1, 6, 11, 16, 21, 26, 31, or 36. The first three pulse positions are coded with 3 bits and the fourth pulse position (starting positions 3 or 4) with 4 bits, resulting in a 17-bit sequence for the algebraic code of each subframe.

The algebraic codebook is searched by minimizing the mean square error between the weighted input speech and the weighted synthesized speech over the time span of each subframe. In each case the weighting is that produced by a perceptual weighting filter that has the effect of shaping the spectrum of the synthesis error signal so that it is better masked by spectrum of the current speech signal.

28.17 Quantization of the Gains for ACELP Encoding

The adaptive codebook gain and the fixed (algebraic) codebook gains are vector quantized using a 7-bit codebook. The gain codebook search is performed by minimizing the mean-square of the weighted error between the original and the reconstructed speech, expressed as a function of the adaptive codebook gain and a fixed codebook correction factor. This correction factor represents the log energy difference between a predicted gain and an estimated gain. The predicted gain is computed using fourth-order moving-average prediction with fixed coefficients on the innovation energy of each subframe. The result is a smoothed energy profile even in the presence of modest quantization errors. As discussed above in case of the LSP quantization, the moving-average prediction serves to provide predicted values even when the current frame information is lost due to transmission errors. Degradations resulting from loss of one or two frames of information are thereby mitigated.

28.18 Channel Coding for ACELP Encoding

The channel coding and interleaving operations for ACELP speech coding are similar to those discussed in Section 28.13 for VSELP coding. The number of bits protected by both error-detection (parity) and error-correction convolutional coding is increased to 48 from 12. Rate 1/2 convolutional coding is used on the 108 more significant bits, 96 class-1 bits, 7 CRC bits and the 5 tail bits of the convolutional coder, resulting in 216 coded class-1 bits. Eight of the 216 bits are dropped by puncturing, yielding 208 coded class-1 bits which are then combined with 52 nonprotected class-2 bits. As compared to the channel coding of the VSELP encoder, the numbers of protected bits is increased and the number of unprotected bits is reduced while keeping the overall coding structure unchanged.

28.19 Conclusions

The IS-54 digital cellular standard specifies modulation and speech coding techniques for mobile cellular systems that allow the interoperation of terminals built by a variety of manufacturers and systems operated across the country by a number of different service providers. It permits speech communication with good quality in a transmission environment characterized by frequent multipath fading and significant intercell interference. Generally, the quality of the IS-54 decoded speech is better at the edges of a cell than the corresponding AMPS transmission due to the error mitigation resulting from channel coding. Near a base station or in the absence of significant fading and interference, the IS-54 speech quality is

reported to be somewhat worse than AMPS due to the inherent limitations of the analysis–synthesis model in reconstructing arbitrary speech signals with limited bits. The IS-641 standard coder achieves higher speech quality, particularly at the edges of heavily occupied cells where transmission errors may be more numerous. At this time no new systems following the IS-54 standard are being introduced. Most base-stations have been converted to transmit and receive on the IS-641 standard as well and use of IS-54 transmissions is dropping rapidly. At the time of its introduction in 1996 the IS-641 coder represented the state of the art in terms of toll quality speech coding near 8 kb/s, a significant improvement over the IS-54 coder introduced in 1990. These standards represent reasonable engineering compromises between high performance and complexity sufficiently low to permit single-chip implementations in mobile terminals.

Both IS-54 and IS-641 are considered second generation cellular standards. Third generation cellular systems promise higher call capacities through better exploitation of the time-varying transmission requirements of speech conversations, as well as improved modulation and coding in wider spectrum bandwidths that achieve similar bit-error ratios but reduce the required transmitted power. Until such systems are introduced, the second generation TDMA systems can be expected to provide many years of successful cellular and personal communications services.

Defining Terms

Codebook: A set of signal vectors available to both the encoder and decoder.

Covariance lattice algorithm: An algorithm for reduction of the covariance matrix of the signal consisting of several lattice stages, each stage implementing an optimal first-order filter with a single coefficient.

Reflection coefficient: A parameter of each stage of the lattice linear prediction filter that determines 1) a forward residual signal at the output of the filter-stage by subtracting from the forward residual at the input a linear function of the backward residual, also 2) a backward residual at the output of the filter stage by subtracting a linear function of the forward residual from the backward residual at the input.

Vector quantizer: A quantizer that assigns quantized vectors to a vector of parameters based on their current values by minimizing some error criterion.

References

[1] Atal, B.S. and Hanauer, S.L., Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.*, 50, 637–655, 1971.

[2] Atal, B.S. and Schroeder, M., Stochastic coding of speech signals at very low bit rates. *Proc. Int. Conf. Comm.*, 1610–1613, 1984.

[3] Gersho, A., Advances in speech and audio compression. *Proc. IEEE,* 82, 900–918, 1994.

[4] Gerson, I.A. and Jasiuk, M.A., Vector sum excited linear prediction (VSELP) speech coding at 8 kbps. *Int. Conf. Acoust. Speech and Sig. Proc.*, ICASSP90, 461–464, 1990.

[5] Lin S. and Costello, D., *Error Control Coding: Fundamentals and Application*, Prentice Hall, Englewood Cliffs, NJ, 1983.

[6] Makhoul, J., Linear prediction, a tutorial review. *Proc. IEEE,* 63, 561–580, 1975.

[7] Salami, R., Laflamme, C., Adoul, J.P., and Massaloux, D., A toll quality 8 kb/s speech codec for the personal communication system (PCS). *IEEE Trans. Vehic. Tech.*, 43, 808–816, 1994.

[8] Soong, F.K. and Juang, B.H., Line spectrum pair (LSP) and speech data compression. *Proc. ICASSP'84*, 1.10.1–1.10.4, 1984.

[9] Telecommunications Industry Association, EIA/TIA Interim Standard, Cellular System Dual-mode Mobile Station–Base Station Compatibility Standard IS-54B, TIA/EIA, Washington, D.C., 1992.

Further Information

For a general treatment of speech coding for telecommunications, see N.S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice Hall, Englewood, NJ, 1984. For a more detailed treatment of linear prediction techniques, see J. Markel and A. Gray, *Linear Prediction of Speech*, Springer–Verlag, NY, 1976.

29

The British Cordless Telephone Standard: CT-2

29.1 History and Background 29-1
29.2 The CT-2 Standard 29-2
29.3 The Radio Interface 29-2
Transmission Issues • Multiple Access and Burst Structure • Power Ramping, Guard Period, and Propagation Delay • Power Control
29.4 Burst Formats 29-5
29.5 Signalling Layer Two (L2) 29-7
General Message Format • Fixed Format Packet
29.6 CPP-Initiated Link Setup Procedures 29-10
29.7 CFP-Initiated Link Setup Procedures 29-10
29.8 Handshaking 29-13
29.9 Main Features of the CT-2 System 29-16
Defining Terms 29-16
References 29-17

Lajos Hanzo
University of Southampton

29.1 History and Background

Following a decade of world-wide research and development (R&D), cordless telephones (**CT**) are now becoming widespread consumer products, and they are paving the way towards ubiquitous, low-cost personal communications networks (**PCN**) [7, 8]. The two most well-known European representatives of CTs are the digital European cordless telecommunications (**DECT**) system [1, 5] and the CT-2 system [2, 6]. Three potential application areas have been identified, namely, domestic, business, and public access, which is also often referred to as telepoint (**TP**).

In addition to conventional voice communications, CTs have been conceived with additional data services and local area network (**LAN**) applications in mind. The fundamental difference between conventional mobile radio systems and CT systems is that CTs have been designed for small to very small cells, where typically benign low-dispersion, dominant line-of-sight (**LOS**) propagation conditions prevail. Therefore, CTs can usually dispense with channel equalizers and complex low-rate speech codecs, since the low-signal dispersion allows for the employment of higher bit rates before the effect of channel dispersion becomes a limiting factor. On the same note, the LOS propagation scenario is associated with mild fading or near-constant received signal level, and when combined with appropriate small-cell power-budget design, it ensures a high average signal-to-noise ratio (**SNR**).

0-8493-8597-0/99/$0.00+$.50

© 1999 by CRC Press LLC

These prerequisites facilitate the employment of high-rate, low-complexity speech codecs, which maintain a low battery drain. Furthermore, the deployment of forward error correction codecs can often also be avoided, which reduces both the bandwidth requirement and the power consumption of the portable station (PS).

A further difference between public land mobile radio (**PLMR**) systems [3] and CTs is that whereas the former endeavor to standardize virtually all system features, the latter seek to offer a so-called access technology, specifying the common air interface (**CAI**), access and signalling protocols, and some network architecture features, but leaving many other characteristics unspecified. By the same token, whereas PLMR systems typically have a rigid frequency allocation scheme and fixed cell structure, CTs use dynamic channel allocation (**DCA**) [4]. The DCA principle allows for a more intelligent and judicious channel assignment, where the base station (BS) and PS select an appropriate traffic channel on the basis of the prevailing traffic and channel quality conditions, thus minimizing, for example, the effect of cochannel interference or channel blocking probability.

In contrast to PLMR schemes, such as the Pan-European global system of mobile communications (GSM) system [3], CT systems typically dispense with sophisticated mobility management, which accounts for the bulk of the cost of PLMR call charges, although they may facilitate limited hand-over capabilities. Whereas in residential applications CTs are the extension of the public switched telephone network (PSTN), the concept of omitting mobility management functions, such as location update, etc., leads to telepoint CT applications where users are able to initiate but not to receive calls. This fact drastically reduces the network operating costs and, ultimately, the call charge at a concomittant reduction of the services rendered.

Having considered some of the fundamental differences between PLMR and CT systems let us now review the basic features of the CT-2 system.

29.2 The CT-2 Standard

The European CT-2 recommendation has evolved from the British standard **MPT-1375** with the aim of ensuring the compatibility of various manufacturers' systems as well as setting performance requirements, which would encourage the development of cost-efficient implementations. Further standardization objectives were to enable future evolution of the system, for example, by reserving signalling messages for future applications and to maintain a low PS complexity even at the expense of higher BS costs. The CT-2 or MPT 1375 CAI recommendation is constituted by the four following parts.

1. *Radio interface:* Standardizes the radio frequency (RF) parameters, such as legitimate channel frequencies, the modulation method, the transmitter power control, and the required receiver sensitivity as well as the carrier-to-interference ratio (CIR) and the time division duplex (**TDD**) multiple access scheme. Furthermore, the transmission burst and master/slave timing structures to be used are also laid down, along with the scrambling procedures to be applied.
2. *Signalling layers one and two:* Defines how the bandwidth is divided among signalling, traffic data, and synchronization information. The description of the first signalling layer includes the dynamic channel allocation strategy, calling channel detection, as well as link setup and establishment algorithms. The second layer is concerned with issues of various signalling message formats, as well as link establishment and re-establishment procedures.

3. *Signalling layer three:* The third signalling layer description includes a range of message sequence diagrams as regards to call setup to telepoint BSs, private BSs, as well as the call clear down procedures.
4. *Speech coding and transmission:* The last part of the standard is concerned with the algorithmic and performance features of the audio path, including frequency responses, clipping, distortion, noise, and delay characteristics.

Having briefly reviewed the structure of the CT-2 recommendations let us now turn our attention to its main constituent parts and consider specific issues of the system's operation.

29.3 The Radio Interface

29.3.1 Transmission Issues

In our description of the system we will adopt the terminology used in the recommendation, where the PS is called cordless portable part (**CPP**), whereas the BS is referred to as cordless fixed part (**CFP**). The channel bandwidth and the channel spacing are 100 kHz, and the allocated system bandwidth is 40 MHz, which is hosted in the range of 864.15–868.15 MHz. Accordingly, a total of 40 RF channels can be utilized by the system.

The accuracy of the radio frequency must be maintained within ±10 kHz of its nominal value for both the CFP and CPP over the entire specified supply voltage and ambient temperature range. To counteract the maximum possible frequency drift of 20 kHz, automatic frequency correction (**AFC**) may be used in both the CFP and CPP receivers. The AFC may be allowed to control the transmission frequency of only the CPP, however, in order to prevent the misalignment of both transmission frequencies.

Binary frequency shift keying (FSK) is proposed, and the signal must be shaped by an approximately Gaussian filter in order to maintain the lowest possible frequency occupancy. The resulting scheme is referred to as Gaussian frequency shift keying (**GFSK**), which is closely related to Gaussian minimum shift keying (GMSK) [7] used in the DECT [1] and GSM [3] systems.

Suffice to say that in M-arry FSK modems the carrier's frequency is modulated in accordance with the information to be transmitted, where the modulated signal is given by

$$S_i(t) = \sqrt{\frac{2E}{T}} \cos\left[\omega_i t + \Phi\right] \qquad i = 1, \ldots, M$$

and E represents the bit energy, T the signalling interval length, ω_i has M discrete values, whereas the phase Φ is constant.

29.3.2 Multiple Access and Burst Structure

The so-called TDD multiple access scheme is used, which is demonstrated in Fig. 29.1. The simple principle is to use the same radio frequency for both uplink and downlink transmissions between the CPP and the CFP, respectively, but with a certain staggering in time. This figure reveals further details of the burst structure, indicating that 66 or 68 b per TDD frame are transmitted in both directions.

There is a 3.5- or 5.5-b duration guard period (**GP**) between the uplink and downlink transmissions, and half of the time the CPP (the other half of the time the CFP) is transmitting with the other part listening, accordingly. Although the guard period wastes some channel capacity, it allows a finite time for both the CPP and CFP for switching from

transmission to reception and vice versa. The burst structure of Fig. 29.1 is used during normal operation across an established link for the transmission of adaptive differential pulse code modulated (ADPCM) speech at 32 kb/s according to the CCITT G721 standard in a so-called B channel or bearer channel. The D channel, or signalling channel, is used for the transmission of link control signals. This specific burst structure is referred to as a multiplex one (**M1**) frame.

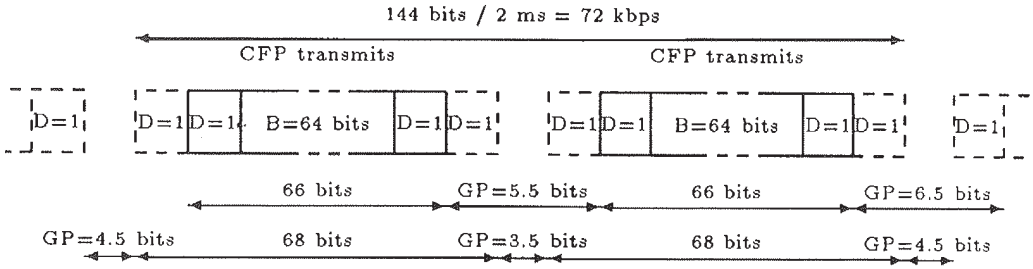


FIGURE 29.1 M1 burst and TDD frame structure.

Since the speech signal is encoded according to the CCITT G721 recommendation at 32 kb/s the TDD bit rate must be in excess of 64 kb/s in order to be able to provide the idle guard space of 3.5- or 5.5-b interval duration plus some signalling capacity. This is how channel capacity is sacrificed to provide the GP. Therefore, the transmission bit rate is stipulated to be 72 kb/s and the transmission burst length is 2 ms, during which 144-b intervals can be accommodated. As it was demonstrated in Fig. 29.1, 66 or 68 b are transmitted in both the uplink and downlink burst, and taking into account the guard spaces, the total transmission frame is constituted by $(2{\cdot}68) + 3.5 + 4.5 = 144$ b or equivalently, by $(2{\cdot}66) + 5.5 + 4.5 = 144$ b. The 66-b transmission format is compulsory, whereas the 68-b format is optional. In the 66-b burst there is one D bit dedicated to signalling at both ends of the burst, whereas in the 68-b burst the two additional bits are also assigned to signalling. Accordingly, the signalling rate becomes 2 b/2 ms or 4 b/2 ms, corresponding to 1 kb/s or 2 kb/s signalling rates.

29.3.3 Power Ramping, Guard Period, and Propagation Delay

As mentioned before and suggested Fig. 29.1, there is a 3.5- or 5.5-b interval duration GP between transmitted and received bursts. Since the signalling rate is 72 kb/s, the bit interval becomes about $1/(72\text{ kb/s}) \approx 13.9\ \mu$s and, hence, the GP duration is about 49 μs or 76 μs. This GP serves a number of purposes. Primarily, the GP allows the transmitter to ramp up and ramp down the transmitted signal level smoothly over a finite time interval at the beginning and end of the transmitted burst. This is necessary, because if the transmitted signal is toggled instantaneously, that is equivalent to multiplying the transmitted signal by a rectangular time-domain window function, which corresponds in the frequency domain to convolving the transmitted spectrum with a sinc function. This convolution would result in spectral side-lobes over a very wide frequency range, which would interfere with adjacent channels. Furthermore, due to the introduction of the guard period, both the CFP and CPP can tolerate a limited propagation delay, but the entire transmitted burst must arrive within the receivers' window, otherwise the last transmitted bits cannot be decoded.

29.3.4 Power Control

In order to minimize the battery drain and the cochannel interference load imposed upon cochannel users, the CT-2 system provides a power control option. The CPPs must be able to transmit at two different power levels, namely, either between 1 and 10 mW or at a level between 12 and 20 dB lower. The mechanism for invoking the lower CPP transmission level is based on the received signal level at the CFP. If the CFP detects a received signal strength more than 90 dB relative to 1 μV/m, it may instruct the CPP to drop its transmitted level by the specified 12–20 dB. Since the 90-dB gain factor corresponds to about a ratio of 31,623, this received signal strength would be equivalent for a 10-cm antenna length to an antenna output voltage of about 3.16 mV. A further beneficial ramification of using power control is that by powering down CPPs that are in the vicinity of a telepoint-type multiple-transceiver CFP, the CFP's receiver will not be so prone to being desensitised by the high-powered close-in CPPs, which would severely degrade the reception quality of more distant CPPs.

29.4 Burst Formats

As already mentioned in the previous section on the radio interface, there are three different subchannels assisting the operation of the CT-2 system, namely, the *voice/data channel* or *B channel,* the *signalling channel* or *D channel,* and the *burst synchronization channel* or **SYN** *channel.* According to the momentary system requirements, a variable fraction of the overall channel capacity or, equivalently, a variable fraction of the bandwidth can be allocated to any of these channels. Each different channel capacity or bandwidth allocation mode is associated with a different burst structure and accordingly bears a different name. The corresponding burst structures are termed as multiplex one (M1), multiplex two (**M2**), and multiplex three (**M3**), of which multiplex one used during the normal operation of established links has already been described in the previous section. Multiplex two and three will be extensively used during link setup and establishment in subsequent sections, as further details of the system's operation are unravelled.

Signalling layer one (L1) defines the burst formats multiplex one–three just mentioned, outlines the calling channel detection procedures, as well as link setup and establishment techniques. *Layer two* (**L2**) deals with issues of acknowledged and unacknowledged information transfer over the radio link, error detection and correction by retransmission, correct ordering of messages, and link maintenance aspects.

The burst structure multiplex two is shown in Fig. 29.2. It is constituted by two 16-b D-channel segments at both sides of the 10-b *preamble (P)* and the 24-b frame synchronization pattern (SYN), and its signalling capacity is 32 b/2 ms = 16 kb/s. Note that the M2 burst does not carry any B-channel information, it is dedicated to synchronization purposes. The 32-b D-channel message is split in two 16-b segments in order to prevent that any 24-b fraction of the 32-b word emulates the 24-b SYN segment, which would result in synchronization misalignment.

Since the CFP plays the role of the master in a telepoint scenario communicating with many CPPs, all of the CPP's actions must be synchronized to those of the CFP. Therefore, if the CPP attempts to initiate a call, the CFP will reinitiate it using the M2 burst, while imposing its own timing structure. The 10-b preamble consists of an alternate zero/one sequence and assists in the operation of the clock recovery circuitry, which has to be able to recover the clock frequency before the arrival of the SYN sequence, in order to be able to detect it. The SYN sequence is a unique word determined by computer search, which has a sharp autocorrelation peak, and its function is discussed later. The way the M2

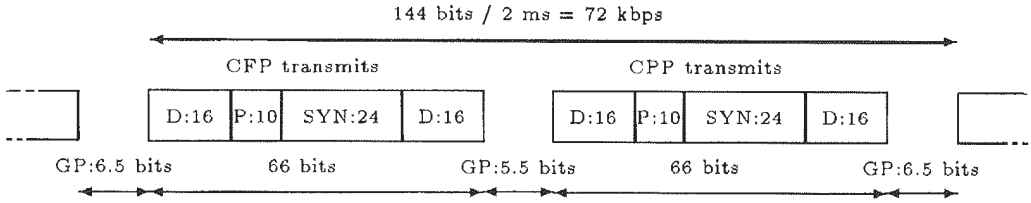


FIGURE 29.2 CT2 multiplex two burst structure.

and M3 burst formats are used for signalling purposes will be made explicit in our further discussions when considering the link setup procedures.

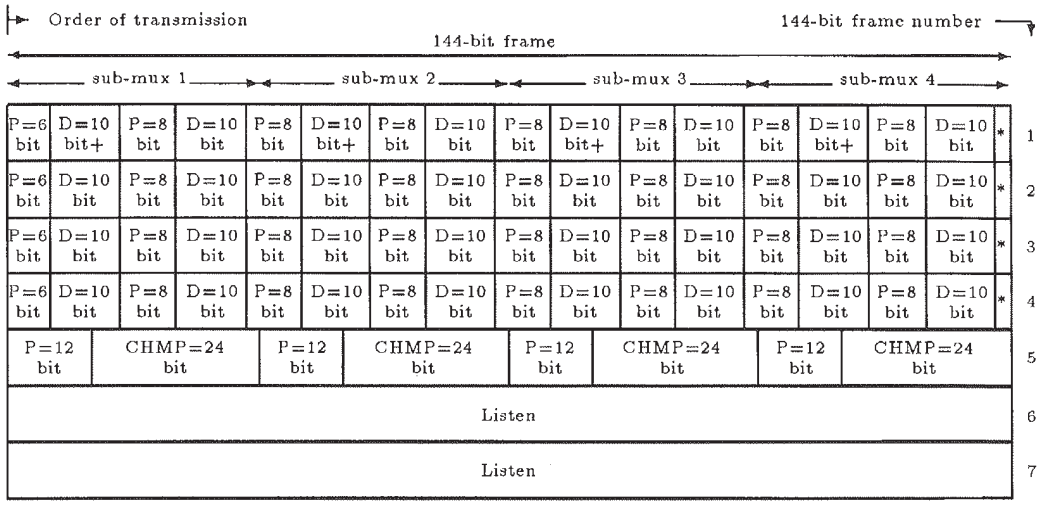
The specific SYN sequences used by the CFP and the CPP are shown in Table 29.1 along with the so-called *channel marker* (**CHM**) sequences used for synchronization purposes by the M3 burst format. Their differences will be made explicit during our further discourse. Observe from the table that the sequences used by the CFP and CPP, namely, SYNF, **CHMF** and SYNP, **CHMP,** respectively, are each other's bit-wise inverses. This was introduced in order to prevent CPPs and CFPs from calling each other directly. The CHM sequences are used, for instance, in residential applications, where the CFP can issue an M2 burst containing a 24-b CHMF sequence and a so-called poll message mapped on to the D-channel bits in order to wake up the specific CPP called. When the called CPP responds, the CFP changes the CHMF to SYNF in order to prevent waking up further CPPs unnecessarily.

TABLE 29.1 CT-2 Synchronization Patterns

	MSB (sent last)					LSB (sent first)
CHMF	1011	1110	0100	1110	0101	0000
CHMP	0100	0001	1011	0001	1010	1111
SYNCF	1110	1011	0001	1011	0000	0101
SYNCP	0001	0100	1110	0100	1111	1010

Since the CT-2 system does not entail mobility functions, such as registration of visiting CPPs in other than their own home cells, in telepoint applications all calls must be initiated by the CPPs. Hence, in this scenario when the CPP attempts to set up a link, it uses the so-called multiplex three burst format displayed in Fig. 29.3. The design of the M3 burst reflects that the CPP initiating the call is oblivious of the timing structure of the potentially suitable target CFP, which can detect access attempts only during its receive window, but not while the CFP is transmitting. Therefore, the M3 format is rather complex at first sight, but it is well structured, as we will show in our further discussions. Observe in the figure that in the M3 format there are five consecutive 2-ms long 144-b transmitted bursts, followed by two idle frames, during which the CPP listens in order to determine whether its 24-b CHMP sequence has been detected and acknowledged by the CFP. This process can be followed by consulting Fig. 29.6, which will be described in depth after considering the detailed construction of the M3 burst.

The first four of the five 2-ms bursts are identical D-channel bursts, whereas the fifth one serves as a synchronization message and has a different construction. Observe, furthermore, that both the first four 144-b bursts as well as the fifth one contain four so-called submultiplex segments, each of which hosts a total of $(6 + 10 + 8 + 10 + 2) = 36$ b. In



Notes: * 2 bit P

1/ Transmission is continuous for five bursts or 10 ms, then off for two burst periods or 4 ms

2/ The 20 bits of D chan are repeated in each of the 4 sub-mux's before D changes

3/ The D chan sync. word (SYNCD) always begins at the start of the slots marked +

FIGURE 29.3 CT2 multiplex three burst structure.

the first four 144-b bursts there are $(6 + 8 + 2) = 16$ one/zero clock-synchronizing P bits and $(10 + 10) = 20$ D bits or signalling bits. Since the D-channel message is constituted by two 10-b half-messages, the first half of the D-message is marked by the + sign in the figure. As mentioned in the context of M2, the D-channel bits are split in two halves and interspersed with the preamble segments in order to ensure that these bits do not emulate valid CHM sequences. Without splitting the D bits this could happen upon concatenating the one/zero P bits with the D bits, since the tail of the SYNF and SYNP sequences is also a one/zero segment. In the fifth 144-b M3 burst, each of the four submultiplex segments is constituted by 12 preamble bits and 24 CPP channel marker (CHMP) bits.

The four-fold submultiplex M3 structure ensures that irrespective of how the CFP's receive window is aligned with the CPP's transmission window, the CFP will be able to capture one of the four submultiplex segments of the fifth M3 burst, establish clock synchronization during the preamble, and lock on to the CHMP sequence. Once the CFP has successfully locked on to one of the CHMP words, the corresponding D-channel messages comprising the CPP identifier can be decoded. If the CPP identifier has been recognized, the CFP can attempt to reinitialize the link using its own master synchronization.

29.5 Signalling Layer Two (L2)

29.5.1 General Message Format

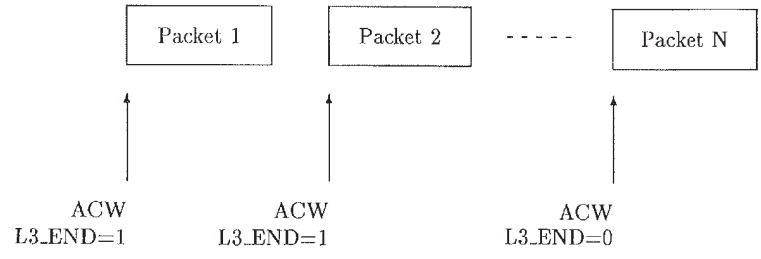
The signalling L2 is responsible for acknowledged and un-acknowledged information transfer over the air interface, error detection and correction by retransmission, as well as for the correct ordering of messages in the acknowledged mode. Its further functions are the link end-point identification and link maintenance for both CPP and CFP, as well as the definition of the L2 and **L3** interface.

Compliance with the L2 specifications will ensure the adequate transport of messages between the terminals of an established link. The L2 recommendations, however, do not define the meaning of messages, this is specified by L3 messages, albeit some of the messages are undefined in order to accommodate future system improvements.

The L3 messages are broken down to a number of standard packets, each constituted by one or more codewords (CW), as shown in Fig. 29.4. The codewords have a standard length of eight octets, and each packet contains up to six codewords. The first codeword in a packet is the so-called address codeword (ACW) and the subsequent ones, if present, are data codewords (**DCW**). The first octet of the ACW of each packet contains a variety of parameters, of which the binary flag **L3_END** is indicated in Fig. 29.4, and it is set to zero in the last packet. If the L3 message transmitted is mapped onto more than one packet, the packets must be numbered up to N. The address codeword is always preceded by a 16-b D-channel frame synchronization word **SYNCD.** Furthermore, each eight-octet CW is protected by a 16-b parity-check word occupying its last two octets. The binary Bose–Chaudhuri–Hocquenghem BCH(63,48) code is used to encode the first six octets or 48 b by adding 15 parity b to yield 63 b. Then bit 7 of octet 8 is inverted and bit 8 of octet 8 added such that the 64-b codeword has an even parity. If there are no D-channel packets to send, a 3-octet idle message **IDLE_D** constituted by zero/one reversals is transmitted. The 8-octet format of the ACWs and DCWs is made explicit in Fig. 29.5, where the two parity check octets occupy octets 7 and 8. The first octet hosts a number of control bits. Specifically, bit 1 is set to logical one for an ACW and to zero for a DCW, whereas bit 2 represents the so-called format type **FT** bit. $FT = 1$ indicates that variable length packet format is used for the transfer of L3 messages, whereas $FT = 0$ implies that a fixed length link setup is used for link end point addressing end service requests. FT is only relevant to ACWs, and in DCWs it has to be set to one.



a/ L2 packet: Contains at least an ACW and 0...5 DCWs



b/ L3 message: N packets

FIGURE 29.4 General L2 and L3 message format.

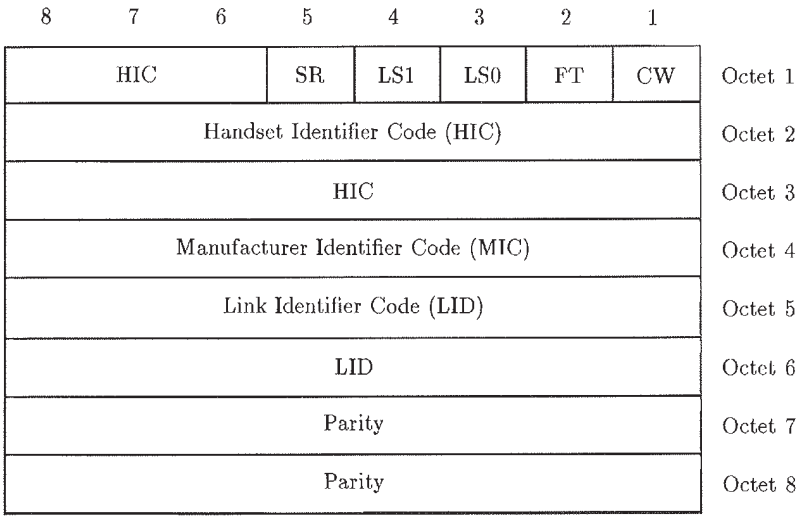


FIGURE 29.5 Fixed format packets mapped on M1, M2, and M3 during link initialization and on M1 and M2 during handshake.

29.5.2 Fixed Format Packet

As an example, let us focus our attention on the fixed format scenario associated with $FT = 0$. The corresponding codeword format defined for use in M1, M2, and M3 for link initiation and in M1 and M2 for handshaking is displayed in Fig. 29.5. Bits 1 and 2 have already been discussed, whereas the 2-bit link status (**LS**) field is used during call setup and handshaking. The encoding of the four possible LS messages is given in Table 29.2. The aim of these LS messages will become more explicit during our further discussions with reference to Fig. 29.6 and Fig. 29.7. Specifically, **link_request** is transmitted from the CPP to the CFP either in an M3 burst as the first packet during CPP-initiated call setup and link re-establishment, or returned as a poll response in an M2 burst from the CPP to the CFP, when the CPP is responding to a call. **Link_grant** is sent by the CFP in response to a link_request originating from the CPP. In octets 5 and 6 it hosts the so-called link identification (**LID**) code, which is used by the CPP, for example, to address a specific CFP or a requested service. The LID is also used to maintain link reference during handshake exchanges and link re-establishment. The two remaining link status handshake messages, namely, ID_OK and ID_lost, are used to report to the far end whether a positive confirmation of adequate link quality has been received within the required time-out period. These issues will be revisited during our further elaborations. Returning to Fig. 29.5, we note that the fixed packet format ($FT = 0$) also contains a 19-b handset identification code (**HIC**) and an 8-b manufacturer identification code (**MIC**). The concatenated HIC and MIC fields jointly from the unique 27-b portable identity code (**PIC**), serving as a link end-point identifier. Lastly, we have to note that bit 5 of octet 1 represents the signalling rate (**SR**) request/response bit, which is used by the calling party to specify the choice of the 66- or 68-b M1 format. Specifically, $SR = 1$ represents the four bit/burst M1 signalling format. The first 6 octets are then protected by the parity check information contained in octets 7 and 8.

TABLE 29.2 Encoding of Link Status Messages

LS1	LS0	Message
0	0	Link_request
0	1	Link_grant
1	0	ID_OK
1	1	ID_lost

29.6 CPP-Initiated Link Setup Procedures

Calls can be initiated at both the CPP and CFP, and the call initiation and detection procedures invoked depend on which party initiated the call. Let us first consider calling channel detection at the CFP, which ensues as follows. Under the instruction of the CFP control scheme, the RF synthesizer tunes to a legitimate RF channel and after a certain settling time commences reception. Upon receiving the M3 bursts from the CPP, the automatic gain control (AGC) circuitry adjusts its gain factor, and during the 12-b preamble in the fifth M3 burst, bit synchronization is established. This specific 144-b M3 burst, is transmitted every 14 ms, corresponding to every seventh 144-b burst. Now the CFP is ready to bit-synchronously correlate the received sequences with its locally stored CHMP word in order to identify any CHMP word arriving from the CPP. If no valid CHMP word is detected, the CFP may retune itself to the next legitimate RF channel, etc.

As mentioned, the call identification and link initialization process is shown in the flowchart of Fig. 29.6. If a valid 24-b CHMP word is identified, D-channel frame synchronization can take place using the 16-b SYNCD sequence and the next 8-octet L2 D-channel message delivering the link_request handshake portrayed earlier in Fig. 29.5 and Table 29.2 is decoded by the CFP. The required $16+64 = 80$ D bits are accommodated in this scenario by the $4 \cdot 20 = 80$ D bits of the next four 144-b bursts of the M3 structure, where the 20 D bits of the four submultiplex segments are transmitted four times within the same burst before the D message changes. If the decoded LID code of Fig. 29.5 is recognized by the CFP, the link may be reinitialized based on the master's timing information using the M2 burst associated with SYNF and containing the link_grant message addressed to the specific CPP identified by its PID.

Otherwise the CFP returns to its scanning mode and attempts to detect the next CHMP message. The reception of the CFP's 24-b SYNF segment embedded in the M2 message shown previously in Fig. 29.2 allows the CPP to identify the position of the CFP's transmit and receive windows and, hence, the CPP now can respond with another M2 burst within the receive window of the CFP. Following a number of M2 message exchanges, the CFP then sends a L3 message to instruct the CPP to switch to M1 bursts, which marks the commencement of normal voice communications and the end of the link setup session.

29.7 CFP-Initiated Link Setup Procedures

Similar procedures are followed when the CPP is being polled. The CFP transmits the 24-b CHMF words hosted by the 24-b SYN segment of the M2 burst shown in Fig. 29.2 in order to indicate that one or more CPPs are being paged. This process is displayed in the flowchart of Fig. 29.7, as well as in the timing diagram displayed in Fig. 29.8. The M2 D-channel messages convey the identifiers of the polled CPPs.

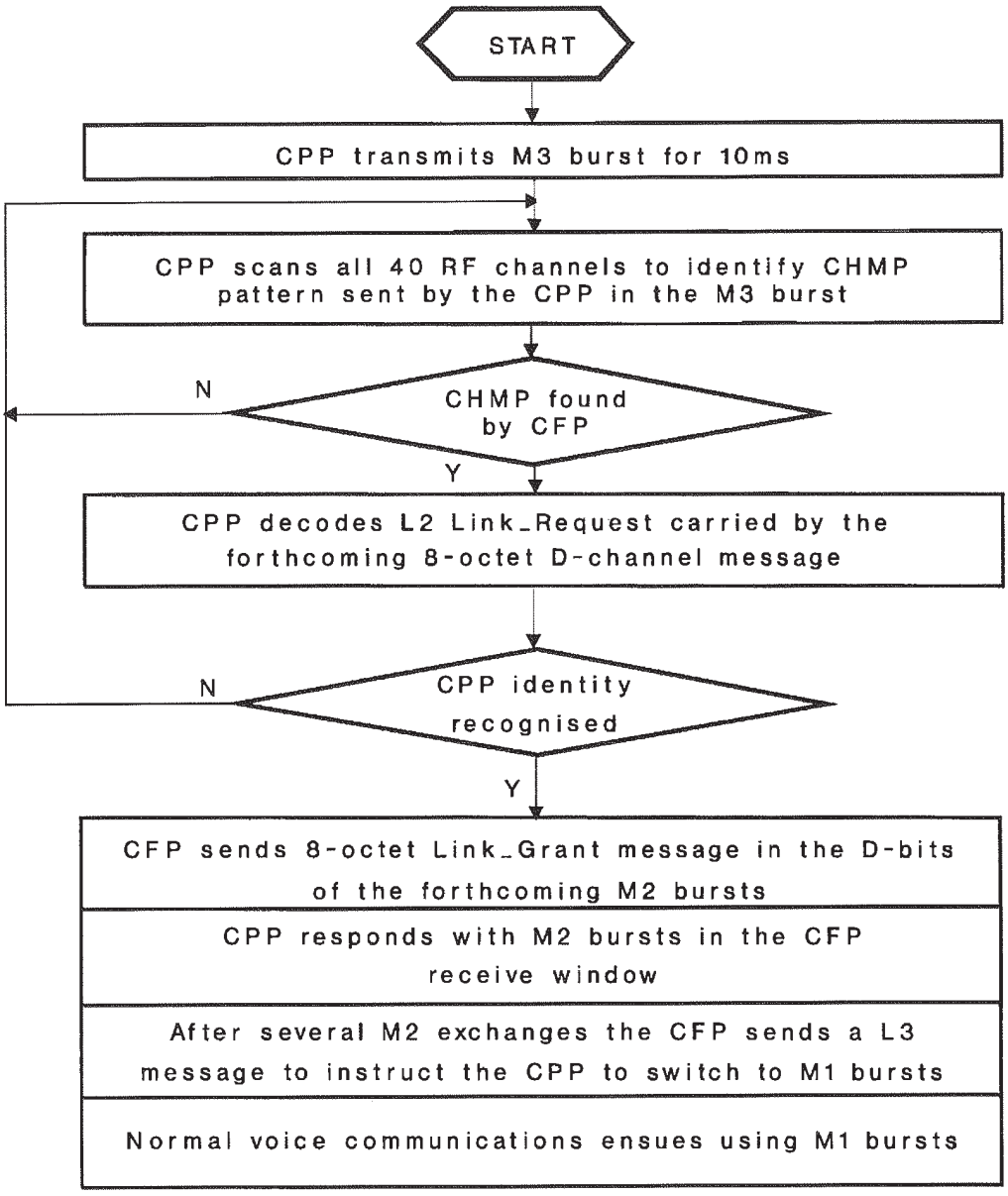


FIGURE 29.6 Flowchart of the CT-2 link initialization by the CPP.

The CPPs keep scanning all 40 legitimate RF channels in order to pinpoint any 24-b CHMF words. Explicitly, the CPP control scheme notifies the RF synthesizer to retune to the next legitimate RF channel if no CHMF words have been found on the current one. The synthesizer needs a finite time to settle on the new center frequency and then starts receiving again. Observe in Fig. 29.8 that at this stage only the CFP is transmitting the M2 bursts; hence, the uplink-half of the 2-ms TDD frame is unused.

Since the M2 burst commences with the D-channel bits arriving from the CFP, the CPP receiver's AGC will have to settle during this 16-b interval, which corresponds to about $16 \cdot 1/[72 \text{ kb/s}] \approx 0.22$ ms. Upon the arrival of the 10 alternating one–zero preamble bits,

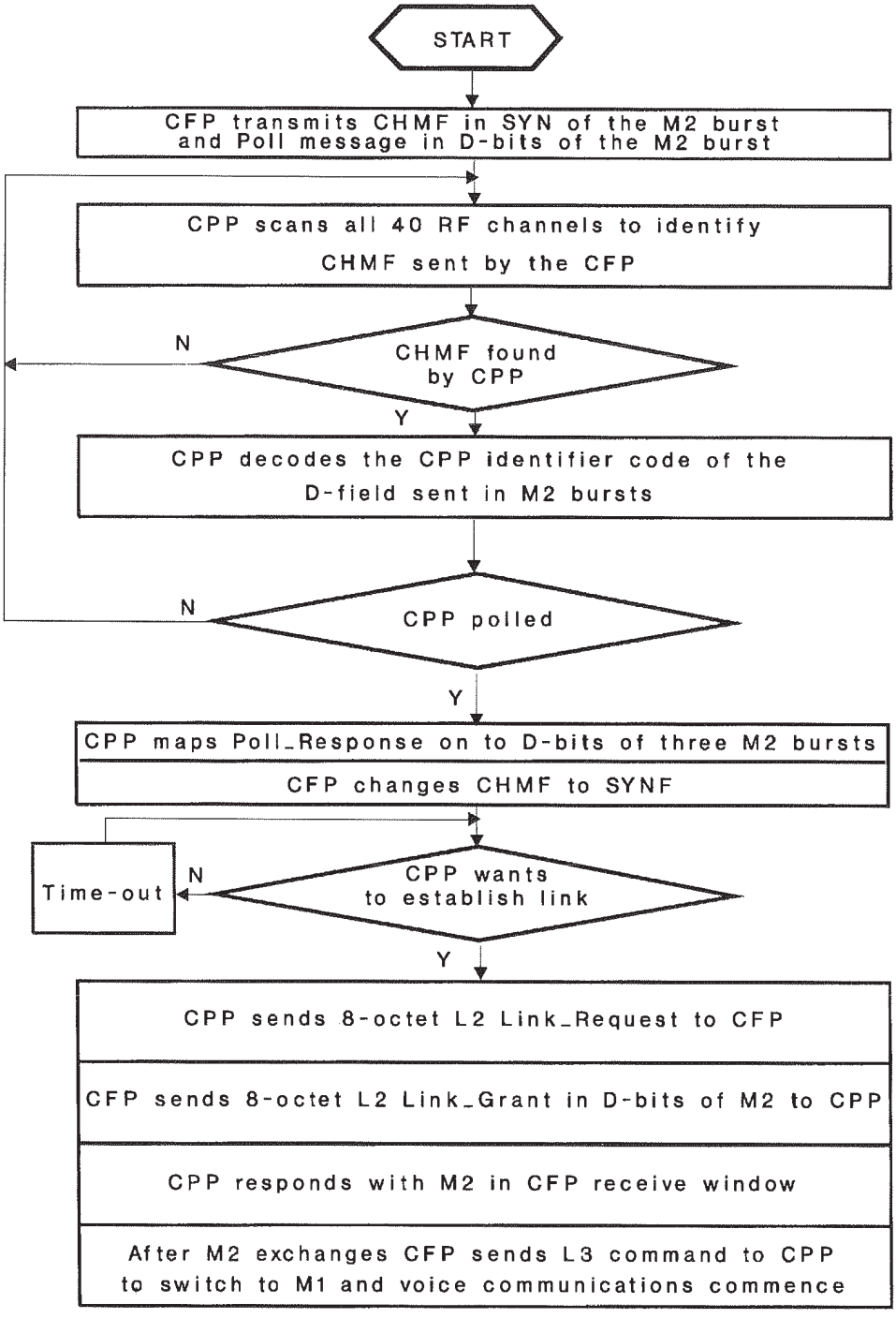


FIGURE 29.7 Flowchart of the CT-2 link initialization by the CFP.

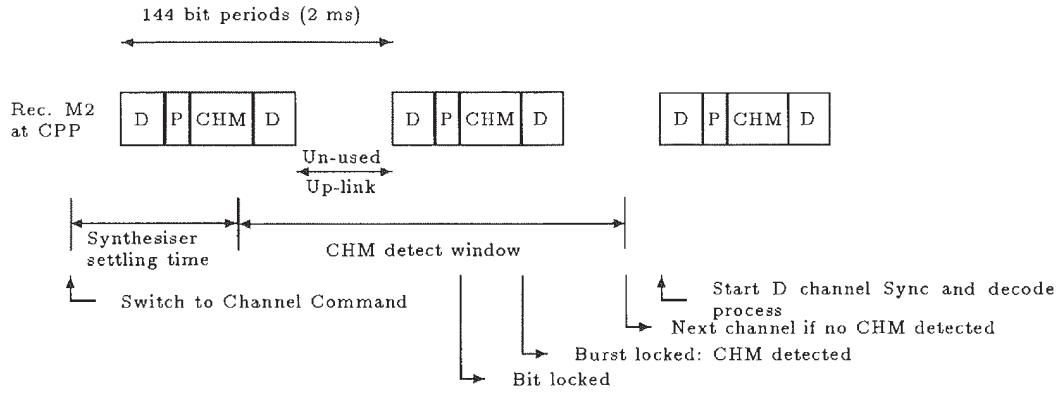


FIGURE 29.8 CT-2 call detection by the CPP.

bit synchronization is established. Now the CPP is ready to detect the CHMF word using a simple correlator circuitry, which establishes the appropriate frame synchronization. If, however, no CHMF word is detected within the receive window, the synthesizer will be retuned to the next RF channel, and the same procedure is repeated, until a CHMF word is detected.

When a CHMF word is correctly decoded by the CPP, the CPP is now capable of frame and bit synchronously decoding the D-channel bits. Upon decoding the D-channel message of the M2 burst, the CPP identifier (**ID**) constituted by the LID and PID segments of Fig. 29.5 is detected and compared to the CPP's own ID in order to decide as to whether the call is for this specific CPP. If so, the CPP ID is reflected back to the CFP along with a SYNP word, which is included in the SYN segment of an uplink M2 burst. This channel scanning and retuning process continues until a legitimate incoming call is detected or the CPP intends to initiate a call.

More precisely, if the specific CPP in question is polled and its own ID is recognized, the CPP sends its poll_response message in three consecutive M2 bursts, since the capacity of a single M2 burst is 32 D bits only, while the handshake messages of Fig. 29.5 and Table 29.2 require 8 octets preceded by a 16-b SYNCD segment. If by this time all paged CPPs have responded, the CFP changes the CHMF word to a SYNF word, in order to prevent activating dormant CPPs who are not being paged. If any of the paged CPPs intends to set up the link, then it will change its poll_response to a L2 link_request message, in response to which the CFP will issue an M2 link_grant message, as seen in Fig. 29.7, and from now on the procedure is identical to that of the CPP-initiated link setup portrayed in Fig. 29.6.

29.8 Handshaking

Having established the link, voice communications is maintained using M1 bursts, and the link quality is monitored by sending handshaking (**HS**) signalling messages using the D-channel bits. The required frequency of the handshaking messages must be between once every 400 ms and 1000 ms. The CT-2 codewords ID_OK, ID_lost, link_request and link_grant of Table 29.2 all represent valid handshakes. When using M1 bursts, however, the transmission of these 8-octet messages using the 2- or 4-b/2ms D-channel segment must be spread over 16 or 32 M1 bursts, corresponding to 32 or 64 ms.

Let us now focus our attention on the *handshake protocol* shown in Fig. 29.9. Suppose that the CPP's handshake interval of Thtx_p = 0.4 s since the start of the last transmitted handshake has expired, and hence the CPP prepares to send a handshake message HS_p.

If the CPP has received a valid HS_f message from the CFP within the last Thrx_p = 1s, the CPP sends an HS_p = ID_OK message to the CFP, otherwise an ID_Lost HS_p. Furthermore, if the valid handshake was HS_f = ID_OK, the CPP will reset its HS_f lost timer Thlost_p to 10 s. The CFP will maintain a 1-s timer referred to as Thrx_f, which is reset to its initial value upon the reception of a valid HS_p from the CPP.

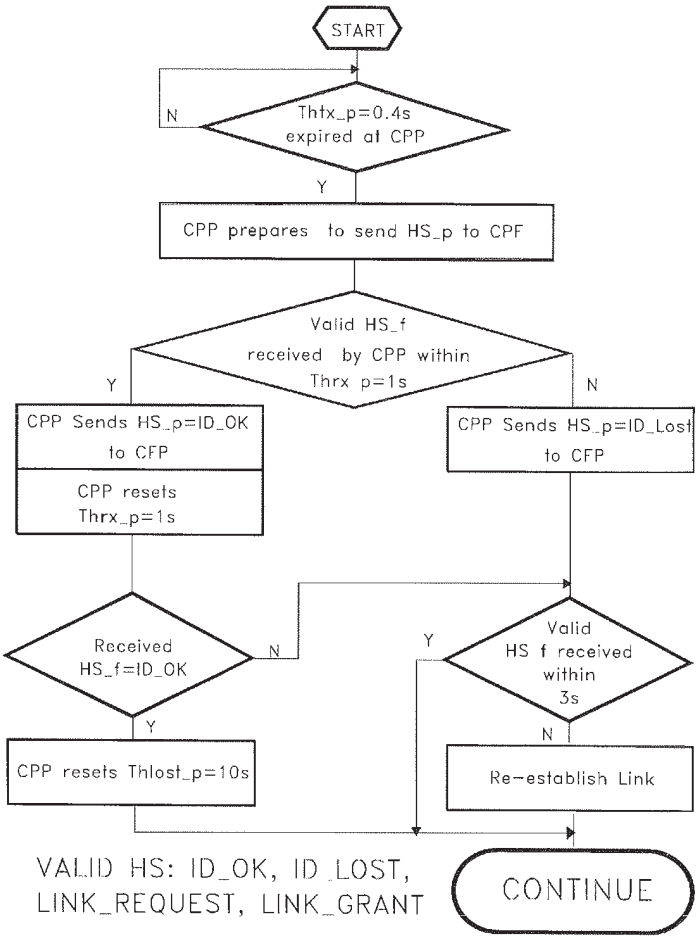


FIGURE 29.9 CT-2 handshake algorithms.

The CFP's actions also follow the structure of Fig. 29.9 upon simply interchanging CPP with CFP and the descriptor _p with _f. If the Thrx_f = 1 s timer expires without the reception of a valid HS_p from the CPP, then the CFP will send its ID_Lost HS_f message to the CPP instead of the ID_OK message and will not reset the Thlost_f = 10 s timer. If, however, the CFP happens to detect a valid HS_p, which can be any of the ID_OK, ID_Lost, link_request and link_grant messages of Table 29.2, arriving from the CPP, the CFP will reset its Thrx_f = 1 s timer and resumes transmitting the ID_OK HS_f message instead of the ID_Lost. Should any of the HS messages go astray for more than 3 s, the CPP or the CFP may try and re-establish the link on the current or another RF channel.

Again, although any of the ID_OK, ID_Lost, link_request and link_grant represent valid handshakes, only the reception of the ID_OK HS message is allowed to reset the Thlost = 10 s timer at both the CPP and CFP. If this timer expires, the link will be relinquished and the call dropped.

The handshake mechanism is further augmented by referring to Fig. 29.10, where two different scenarios are examplified, portraying the situation when the HS message sent by the CPP to the CFP is lost or, conversely, that transmitted by the CFP is corrupted.

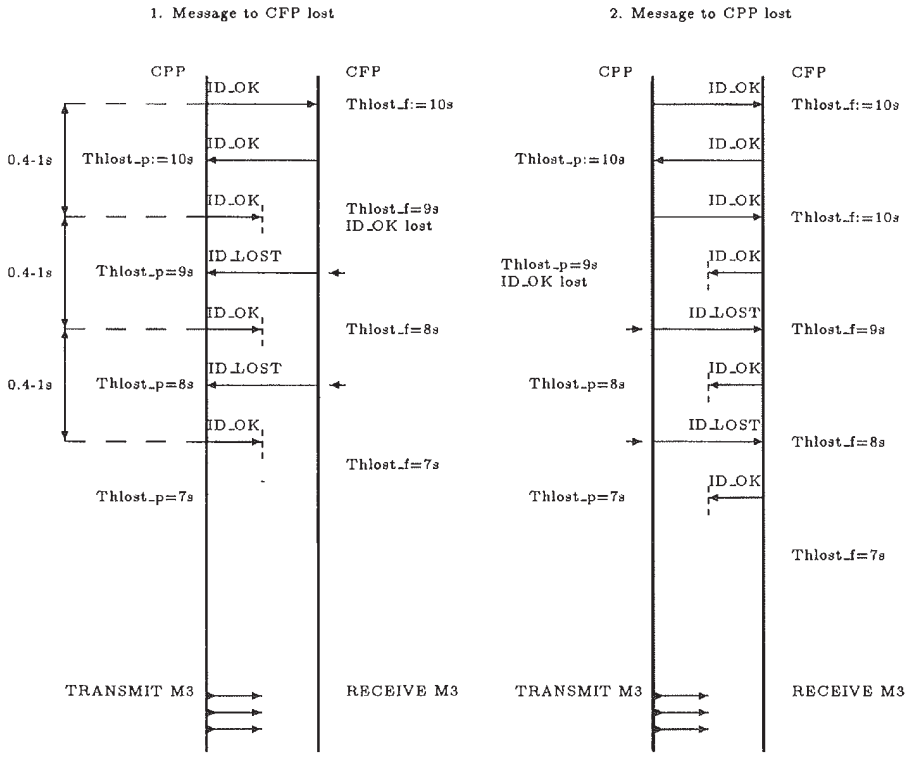


FIGURE 29.10 Handshake loss scenarios.

Considering the first scenario, during error-free communications the CPP sends HS_p = ID_OK, and upon receiving it the CFP resets its Thlost_f timer to 10 s. In due course it sends an HS_f = ID_OK acknowledgement, which also arrives free from errors. The CPP resets the Thlost_f timer to 10 s and, after the elapse of the 0.4–1 s handshake interval, issues an HS_p = ID_OK message, which does not reach the CFP. Hence, the Thlost_f timer is now reduced to 9 s and an HS_f = ID_Lost message is sent to the CPP. Upon reception of this, the CPP now cannot reset its Thlost_p timer to 10 s but can respond with an HS_p = ID_OK message, which again goes astray, forcing the CFP to further reduce its Thlost_f timer to 8 s. The CFP issues the valid handshake HS_f = ID_Lost, which arrives at the CPP, where the lack of HS_f = ID_OK reduces Thlost_p to 8 s. Now the corruption of the issued HS_p = ID_OK reduces Thlost_f to 7 s, in which event the link may be reinitialized using the M3 burst. The portrayed second example of Fig. 29.10 can be easily followed in case of the scenario when the HS_f message is corrupted.

29.9 Main Features of the CT-2 System

In our previous discourse we have given an insight in the algorithmic procedures of the CT-2 MPT 1375 recommendation. We have briefly highlighted the four-part structure of the standard dealing with the radio interface, signalling layers 1 and 2, signalling layer 3, and the speech coding issues, respectively. There are forty 100-kHz wide RF channels in the band 864.15–868.15 MHz, and the 72 kb/s bit stream modulates a Gaussian filtered FSK modem. The multiple access technique is TDD, transmitting 2-ms duration, 144-b M1 bursts during normal voice communications, which deliver the 32-kb/s ADPCM-coded speech signal. During link establishment the M2 and M3 bursts are used, which were also portrayed in this treatise, along with a range of handshaking messages and scenarios.

Defining Terms

AFC: Automatic frequency correction
CAI: Common air interface
CFP: Cordless fixed part
CHM: Channel marker sequence
CHMF: CFP channel marker
CHMP: CPP channel marker
CPP: Cordless portable part
CT: Cordless telephone
DCA: Dynamic channel allocation
DCW: Data code word
DECT: Digital European cordless telecommunications system
FT: Frame format type bit
GFSK: Gaussian frequency shift keying
GP: Guard period
HIC: Handset identification code
HS: Handshaking
ID: Identifier
L2: Signalling layer 2
L3: Signalling layer 3
LAN: Local area network
LID: Link identification
LOS: Line of sight
LS: Link status
M1: Multiplex one burst format
M2: Multiplex two burst format
M3: Multiplex three burst format
MIC: Manufacturer identification code
MPT-1375: British CT2 standard
PCN: Personal communications network
PIC: Portable identification code

PLMR: Public land mobile radio
SNR: Signal-to-noise ratio
SR: Signalling rate bit
SYN: Synchronization sequence
SYNCD: 16-b D-channel frame synchronization word
TDD: Time division duplex multiple access scheme
TP: Telepoint

References

[1] Asghar, S., Digital European cordless telephone (DECT), In *The Mobile Communications Handbook,* Chap. 30, CRC Press, Boca Raton, FL, 1995.

[2] Gardiner, J.G., Second generation cordless (CT-2) telephony in the UK: telepoint services and the common air-interface, *Elec. & Comm. Eng. J.,* 71–78, Apr. 1990.

[3] Hanzo, L., The Pan-European mobile radio system, In *The Mobile Communications Handbook,* Chap. 25, CRC Press, Boca Raton, FL, 1995.

[4] Jabbari, B., Dynamic channel assignment, In *The Mobile Communications Handbook,* Chap. 21, CRC Press, Boca Raton, FL, 1995.

[5] Ochsner, H., The digital European cordless telecommunications specification DECT. In *Cordless telecommunication in Europe.* Tuttlebee, W.H.M., Ed., 273–285. Springer-Verlag, 1990.

[6] Steedman, R.A.J., The Common Air Interface MPT 1375. In *Cordless Telecommunication in Europe.* Tuttlebee, W.H.W. Ed., 261–272, Springer-Verlag, 1990.

[7] Steele, R., Ed., *Mobile Radio Communications,* Pentech Press, London, 1992.

[8] Tuttlebee, W.H.W., Ed., *Cordless Telecommunication in Europe,* Springer-Verlag, 1990.

30

Half-Rate Standards

Wai-Yip Chan
Illinois Institute of Technology

Ira Gerson
Motorola Corporate Systems Research Laboratories

Toshio Miki
NTT Mobile Communication Network, Inc.

30.1 Introduction 30-1
30.2 Speech Coding for Cellular Mobile Radio Communications 30-2
30.3 Codec Selection and Performance Requirements 30-3
30.4 Speech Coding Techniques in the Half-Rate Standards 30-4
30.5 Channel Coding Techniques in the Half-Rate Standards 30-6
30.6 The Japanese Half-Rate Standard 30-8
30.7 The European GSM Half-Rate Standard 30-11
30.8 Conclusions 30-15
Defining Terms 30-15
Acknowledgment 30-15
References 30-16
Further Information 30-16

30.1 Introduction

A half-rate speech coding standard specifies a procedure for digital transmission of speech signals in a digital cellular radio system. The speech processing functions that are specified by a half-rate standard are depicted in Fig. 30.1. An input speech signal is processed by a

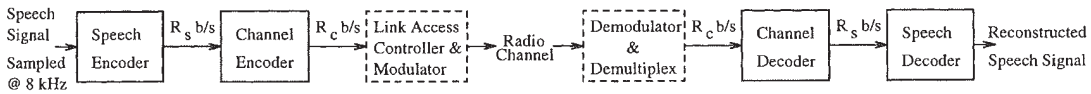


FIGURE 30.1 Digital speech transmission for digital cellular radio. Boxes with solid outlines represent processing modules that are specified by the half-rate standards.

speech encoder to generate a digital representation at a *net bit rate* of R_s bits per second. The encoded bit stream representing the input speech signal is processed by a *channel encoder* to generate another bit stream at a *gross bit rate* of R_c bits per second, where $R_c > R_s$. The channel encoded bit stream is organized into data frames, and each frame is transmitted as payload data by a radio-link access controller and modulator. The net bit rate R_s counts the number of bits used to describe the speech signal, and the difference between the gross and net bit rates $(R_c - R_s)$ counts the number of error protection bits needed by the *channel decoder* to correct and detect transmission errors. The output of

0-8493-8597-0/99/$0.00+$.50
© 1999 by CRC Press LLC

the channel decoder is given to the *speech decoder* to generate a *quantized* version of the speech encoder's input signal. In current digital cellular radio systems that use time-division multiple access (TDMA), a voice connection is allocated a fixed transmission rate (i.e., R_c is a constant). The operations performed by the speech and channel encoders and decoders and their input and output data formats are governed by the half-rate standards.

Globally, three major TDMA cellular radio systems have been developed and deployed. The initial digital speech services offered by these cellular systems were governed by *full-rate standards.* Because of the rapid growth in demand for cellular services, the available transmission capacity in some areas is frequently saturated, eroding customer satisfaction. By providing essentially the same voice quality but at half the gross bit rates of the full-rate standards, half-rate standards can readily double the number of callers that can be serviced by the cellular systems. The gross bit rates of the full-rate and half-rate standards for the European Groupe Speciale Mobile (GSM), Japanese Personal Digital Cellular[1] (PDC), and North American cellular (IS-54) systems are listed in Table 30.1. The three systems were developed and deployed under different time tables. Their disparate full- and half-bit rates partly reflect this difference. At the time of writing (January, 1995), the European and the Japanese systems have each selected an algorithm for their respective half-rate **codec.** Standardization of the North American half-rate codec has not reached a conclusion as none of the candidate algorithms has fully satisfied the standard's requirements. Thus, we focus here on the Japanese and European half-rate standards and will only touch upon the requirements of the North American standard.

TABLE 30.1 Gross Bit Rates Used for Digital Speech Transmission in Three TDMA Cellular Radio Systems

	Gross Bit Rate, b/s	
Standard Organization and Digital Cellular System	Full Rate	Half Rate
European Telecommunications Standards Institute (ETSI), GSM	22,800	11,400
Research & Development Center for Radio Systems (RCR), PDC	11,200	5,600
Telecommunication Industries Association (TIA), IS-54	13,000	6,500

30.2 Speech Coding for Cellular Mobile Radio Communications

Unlike the relatively benign transmission media commonly used in the public-switched telephone network (PSTN) for analog and digital transmission of speech signals, mobile radio channels are impaired by various forms of fading and interference effects. Whereas proper engineering of the radio link elements (modulation, power control, diversity, equalization, frequency allocation, etc.) ameliorates fading effects, burst and isolated bit errors still occur frequently. The net effect is such that speech communication may be required to be operational even for bit-error rates greater than 1%. In order to furnish reliable voice communication, typically half of the transmitted payload bits are devoted to error correction and detection.

[1]Personal Digital Cellular was formerly Japanese Digital Cellular (JDC).

It is common for low-bit-rate speech codecs to process samples of the input speech signal one frame at a time, e.g., 160 samples processed once every 20 ms. Thus, a certain amount of time is required to gather a block of speech samples, encode them, perform channel encoding, transport the encoded data over the radio channel, and perform channel decoding and speech synthesis. These processing steps of the speech codec add to the overall end-to-end transmission delay. Long transmission delay hampers conversational interaction. Moreover, if the cellular system is interconnected with the PSTN and a four-wire to two-wire (analog) circuit conversion is performed in the network, feedbacks called *echoes* may be generated across the conversion circuit. The echoes can be heard by the originating talker as a delayed and distorted version of his/her speech and can be quite annoying. The annoyance level increases with the transmission delay and may necessitate (at additional costs) the deployment of **echo cancellers**.

A consequence of user mobility is that the level and other characteristics of the acoustic background noise can be highly variable. Though acoustic noise can be minimized through suitable acoustic transduction design and the use of adaptive filtering/cancellation techniques [9, 13, 15], the speech encoding algorithm still needs to be robust against background noise of various levels and kinds (e.g., babble, music, noise bursts, and colored noise).

Processing complexity directly impacts the viability of achieving a circuit realization that is compact and has low-power consumption, two key enabling factors of equipment portability for the end user. Factors that tend to result in low complexity are fixed-point instead of floating-point computation, lack of complicated arithmetic operations (division, square roots, transcendental functions), regular algorithm structure, small data memory, and small program memory. Since, in general, better speech quality can be achieved with increasing speech and channel coding delay and complexity, the digital cellular mobile-radio environment imposes conflicting and challenging requirements on the speech codec.

30.3 Codec Selection and Performance Requirements

The half-rate speech coding standards are drawn up through competitive testing and selection. From a set of candidate codec algorithms submitted by contending organizations, the one algorithm that meets basic selection criteria and offers the best performance is selected to form the standard. The codec performance measures and codec testing and selection procedures are set out in a test plan under the auspices of the organization (Table 30.1) responsible for the standardization process (see, e.g., [16]). Major codec characteristics evaluated are speech quality, delay, and complexity. The full-rate codec is also evaluated as a *reference codec,* and its evaluation scores form part of the selection criteria for the codec candidates.

The speech quality of each candidate codec is evaluated through listening tests. To conduct the tests, each candidate codec is required to process speech signals and/or encoded bit streams that have been preprocessed to simulate a range of operating conditions: variations in speaker voice and level, acoustic background noise type and level, channel error rate, and stages of **tandem coding**. During the tests, subjects listen to processed speech signals and judge their quality levels or annoyance levels on a five-point opinion scale. The opinion scores collected from the tests are suitably averaged over all trials and subjects for each test condition (see [11], for mean opinion score (MOS) and degradation mean opinion score). The categorical opinion scales of the subjects are also calibrated using *modulated noise reference units* (*MNRUs*) [3]. Modulated noise better resembles the distortions created by speech codecs than noise that is uncorrelated with the speech signal. Modulated noise is generated by multiplying the speech signal with a noise signal. The resultant modulated

noise is scaled to a desired power level and then added to the uncoded (clean) speech signal. The ratio between the power level of the speech signal and that of the modulated noise is expressed in decibels and given the notation *dBQ*. Under each test condition, subjects are presented with speech signals processed by the codecs as well as speech signals corrupted by modulated noise. Through presenting a range of modulated-noise levels, the subjects' opinions are calibrated on the dBQ scale. Thereafter, the mean opinion scores obtained for the codecs can also be expressed on that scale.

For each codec candidate, a profile of scores is compiled, consisting of speech quality scores, delay measurements, and complexity estimates. Each candidate's score profile is compared with that of the reference codec, ensuring that basic requirements are satisfied (see, e.g., [12]). An overall figure of merit for each candidate is also computed from the profile. The candidates, if any, that meet the basic requirements then compete on the basis of maximizing the figure of merit.

Basic performance requirements for each of the three half-rate standards are summarized in Table 30.2. In terms of speech quality, the GSM and PDC half-rate codecs

are permitted to underperform their respective full-rate codecs by no more than 1 dBQ averaging over all test conditions and no more than 3 dBQ within each test condition. More stringently, the North American half-rate codec is required to furnish a speech-quality profile that is statistically equivalent to that of the North American full-rate codec as determined by a specific statistical procedure for multiple comparisons [16]. Since various requirements on the half-rate standards are set relative to their full-rate counterparts, an indication of the *relative* speech quality between the three half-rate standards can be deduced from the test results of De Martino [2] comparing the three full-rate codecs. The maximum delays in Table 30.2 apply to the total of the delays through the speech and channel encoders and decoders (Fig. 30.1). Codec complexity is computed using a formula that counts the computational operations and memory usage of the codec algorithm. The complexity of the half-rate codecs is limited to 3 or 4 times that of their full-rate counterparts.

TABLE 30.2 Basic Performance Requirements for the Three Half-Rate Standards

	Basic performance requirements		
Digital Cellular Systems	Min. Speech Quality, dBQ Rel. to Full Rate	Max. Delay, ms	Max. Complexity Rel. to Full Rate
Japanese (PDC)	−1 average, −3 maximum	94.8	3×
European (GSM)	−1 average, −3 maximum	90	4×
North American (IS-54)	Statistically equivalent	100	4×

30.4 Speech Coding Techniques in the Half-Rate Standards

Existing half-rate and full-rate standard coders can be characterized as *linear-prediction based analysis-by-synthesis* (LPAS) speech coders [4]. LPAS coding entails using a time-varying all-pole filter in the decoder to synthesize the quantized speech signal. A short segment of the signal is synthesized by driving the filter with an *excitation* signal that is either *quasiperiodic* (for *voiced* speech) or *random* (for *unvoiced* speech). In either case, the excitation signal has a *spectral envelope* that is relatively flat. The synthesis filter serves to shape the spectrum of the excitation input so that the spectral envelope of the synthesized output resembles the filter's magnitude frequency response. The magnitude response often has prominent peaks; they render the *formants* that give a speech signal its

phonetic character. The synthesis filter has to be adapted to the current frame of input speech signal. This is accomplished with the encoder performing a linear prediction (LP) analysis of the frame: the inverse of the all-pole synthesis filter is applied as an LP *error filter* to the frame, and the values of the filter parameters are computed to minimize the energy of the filter's output error signal. The resultant filter parameters are quantized and conveyed to the decoder for it to update the synthesis filter.

Having executed an LP analysis and quantized the synthesis filter parameters, the LPAS encoder performs analysis-by-synthesis (ABS) on the input signal to find a suitable excitation signal. An ABS encoder maintains a *copy* of the decoder. The encoder examines the possible outputs that can be produced by the decoder copy in order to determine how best to instruct (using transmitted information) the actual decoder so that it would output (synthesize) a good approximation of the input speech signal. The decoder copy tracks the state of the actual decoder, since the latter evolves (under ideal channel conditions) according to information received from the encoder. The details of the ABS procedure vary with the particular excitation model employed in a specific coding scheme. One of the earliest seminal LPAS schemes is *code excited linear prediction* (*CELP*) [4]. In CELP, the excitation signal is obtained from a **codebook** of *code vectors,* each of which is a candidate for the excitation signal. The encoder searches the codebook to find the one code vector that would result in a best match between the resultant synthesis output signal and the encoder's input speech signal. The matching is considered best when the energy of the difference between the two signals being matched is minimized. A *perceptual weighting filter* is usually applied to the difference signal (prior to energy integration) to make the minimization more relevant to human perception of speech fidelity. Regions in the frequency spectrum where human listeners are more sensitive to distortions are given relatively stronger weighting by the filter and vice versa. For instance, the concentration of spectral energy around the formant frequencies gives rise to stronger *masking* of coder noise (i.e., rendering the noise less audible) and, therefore, weaker weighting can be applied to the formant frequency regions. For masking to be effective, the weighting filter has to be adapted to the time-varying speech spectrum. Adaptation is achieved usually by basing the weighting filter parameters on the synthesis filter parameters.

The CELP framework has evolved to form the basis of a great variety of speech coding algorithms, including all existing full- and half-rate standard algorithms for digital cellular systems. We outline next the basic CELP encoder-processing steps, in a form suited to our subsequent detailed descriptions of the PDC and GSM half-rate coders. These steps have accounted for various computational efficiency considerations and may, therefore, deviate from a conceptual functional description of the encoder constituents.

1. LP analysis on the current frame of input speech to determine the coefficients of the all-pole synthesis filter;
2. quantization of the LP filter parameters;
3. determination of the open-loop **pitch period** or lag;
4. adapting the perceptual weighting filter to the current LP information (and also pitch information when appropriate) and applying the adapted filter to the input speech signal;
5. formation of a filter cascade (which we shall refer to as *perceptually weighted synthesis filter*) consisting of the LP synthesis filter, as specified by the quantized parameters in step 2, followed by the perceptual weighting filter;
6. subtraction of the *zero-input response* of the perceptually weighted synthesis filter (the filter's decaying response due to past input) from the perceptually weighted

input speech signal obtained in step 4;

7. an *adaptive codebook* is searched to find the most suitable periodic excitation, i.e., when the perceptually weighted synthesis filter is driven by the best code vector from the adaptive codebook, the output of the filter cascade should best match the difference signal obtained in step 6;
8. one or more nonadaptive excitation codebooks are searched to find the most suitable random excitation vectors that, when added to the best periodic excitation as determined in step 7 and with the resultant sum signal driving the filter cascade, would result in an output signal best matching the difference signal obtained in step 6.

Steps 1–6 are executed once per frame. Steps 7 and 8 are executed once for each of the *subframes* that together constitute a frame. Step 7 may be skipped depending on the pitch information from step 3, or if step 7 were always executed, a *nonperiodic excitation* decision would be one of the possible outcomes of the search process in step 7. Integral to steps 7 and 8 is the determination of gain (scaling) parameters for the excitation vectors. For each frame of input speech, the filter and excitation and gain parameters determined as outlined are conveyed as encoded bits to the speech decoder.

In a properly designed system, the data conveyed by the channel decoder to the speech decoder should be free of errors most of the time, and the speech signal synthesized by the speech decoder would be identical to that as determined in the speech encoder's ABS operation. It is common to enhance the quality of the synthesized speech by using an adaptive *postfilter* to attenuate coder noise in the perceptually sensitive regions of the spectrum. The postfilter of the decoder and the perceptual weighting filter of the encoder may seem to be functionally identical. The weighting filter, however, influences the selection of the best excitation among available choices, whereas the postfilter actually shapes the spectrum of the synthesized signal. Since postfiltering introduces its own distortion, its advantage may be diminished if tandem coding occurs along the end-to-end communication path. Nevertheless, proper design can ensure that the net effect of postfiltering is a reduction in the amount of audible codec noise [1]. Excepting postfiltering, all other speech synthesis operations of an LPAS decoder are (effectively) duplicated in the encoder (though the converse is not true). Using this fact, we shall illustrate each coder in the sequel by exhibiting only a block diagram of its encoder or decoder but not both.

30.5 Channel Coding Techniques in the Half-Rate Standards

Crucial to the maintenance of quality speech communication is the ability to transport coded speech data across the radio channel with minimal errors. Low-bit-rate LPAS coders are particularly sensitive to channel errors; errors in the bits representing the LP parameters in one frame, for instance, could result in the synthesis of nonsensical sounds for longer than a frame duration. The error rate of a digital cellular radio channel with no channel coding can be catastrophically high for LPAS coders. The amount of tolerable transmission delay is limited by the requirement of interactive communication and, consequently, *forward error control* is used to remedy transmission errors. "Forward" means that channel errors are remedied in the receiver, with no additional information from the transmitter and, hence, no additional transmission delay. To enable the channel decoder to correct channel errors, the channel encoder conveys more bits than the amount generated by the speech encoder. The additional bits are for error *protection*, as errors may or may not occur in any

particular transmission epoch. The ratio of the number of encoder input (information) bits to the number of encoder output (code) bits is called the (channel) *coding rate*. This is a number no more than one and generally decreases as the error protection power increases. Though a lower channel coding rate gives more error protection, fewer bits will be available for speech coding. When the channel is in good condition and, hence, less error protection is needed, the received speech quality could be better if bits devoted to channel coding were used for speech coding. On the other hand, if a high channel coding rate were used, there would be uncorrected errors under poor channel conditions and speech quality would suffer. Thus, when nonadaptive forward error protection is used over channels with nonstationary statistics, there is an inevitable tradeoff between quality degradation due to uncorrected errors and that due to expending bits on error protection (instead of on speech encoding).

Both the GSM and PDC half-rate coders use *convolutional coding* [14] for error correction. Convolutional codes are sliding or sequential codes. The encoder of a rate $m/n, m < n$ convolutional code can be realized using m shift registers. For every m information bits input to the encoder (one bit to each of the m shift registers), n code bits are output to the channel. Each code bit is computed as a modulo-2 sum of a subset of the bits in the shift registers. Error protection overhead can be reduced by exploiting the unequal sensitivity of speech quality to errors in different positions of the encoded bit stream. A family of *rate-compatible punctured convolutional codes* (RCPCCs) [10] is a collection of related convolutional codes; all of the codes in the collection except the one with the lowest rate are derived by *puncturing* (dropping) code bits from the convolutional code with the lowest rate. With an RCPCC, the channel coding rate can be varied on the fly (i.e., variable-rate coding) while a sequence of information bits is being encoded through the shift registers, thereby imparting on different segments in the sequence different degrees of error protection.

For decoding a convolutional coded bit stream, the *Viterbi algorithm* [14] is a computationally efficient procedure. Given the output of the demodulator, the algorithm determines the most likely sequence of data bits sent by the channel encoder. To fully utilize the error correction power of the convolutional code, the amplitude of the demodulated *channel symbol* can be quantized to more bits than the minimum number required, i.e., for subsequent *soft decision decoding*. The minimum number of bits is given by the number of channel-coded bits mapped by the modulator onto each channel symbol; decoding based on the minimum-rate bit stream is called *hard decision* decoding. Although soft decoding gives better error protection, decoding complexity is also increased.

Whereas convolutional codes are most effective against randomly scattered bit errors, errors on cellular radio channels often occur in bursts of bits. These bursts can be broken up if the bits put into the channel are rearranged after demodulation. Thus, in *block interleaving*, encoded bits are read into a matrix by row and then read out of the matrix by column (or vice versa) and then passed on to the modulator; the reverse operation is performed by a *deinterleaver* following demodulation. Interleaving increases the transmission delay to the extent that enough bits need to be collected in order to fill up the matrix.

Owing to the severe nature of the cellular radio channel and limited available transmission capacity, uncorrected errors often remain in the decoded data. A common countermeasure is to append an error detection code to the speech data stream prior to channel coding. When residual channel errors are detected, the speech decoder can take various remedial measures to minimize the negative impact on speech quality. Common measures are repetition of speech parameters from the most recent good frames and gradual muting of the possibly corrupted synthesized speech.

The PDC and GSM half-rate standard algorithms together embody some of the latest advances in speech coding techniques, including: *multimodal coding* where the coder

configuration and bit allocation change with the type of speech input; *vector quantization* (*VQ*) [5] of the LP filter parameters; higher precision and improved coding efficiency for pitch-periodic excitation; and postfiltering with improved tandeming performance. We next explore the more distinctive features of the PDC and GSM speech coders.

30.6 The Japanese Half-Rate Standard

An algorithm was selected for the Japanese half-rate standard in April 1993, following the evaluation of 12 submissions in a first round, and four final candidates in a second round [12]. The selected algorithm, called pitch synchronous innovation CELP[2] (PSI-CELP), met all of the basic selection criteria and scored the highest among all candidates evaluated. A block diagram of the PSI-CELP encoder is shown in Fig. 30.2, and bit allocations are summarized in Table 30.3. The complexity of the coder is estimated to be approximately 2.4 times that of the PDC full-rate coder. The frame size of the coder is 40 ms, and its subframe size is 10 ms. These sizes are longer than those used in most existing CELP-type standard coders. However, LP analysis is performed twice per frame in the PSI-CELP coder.

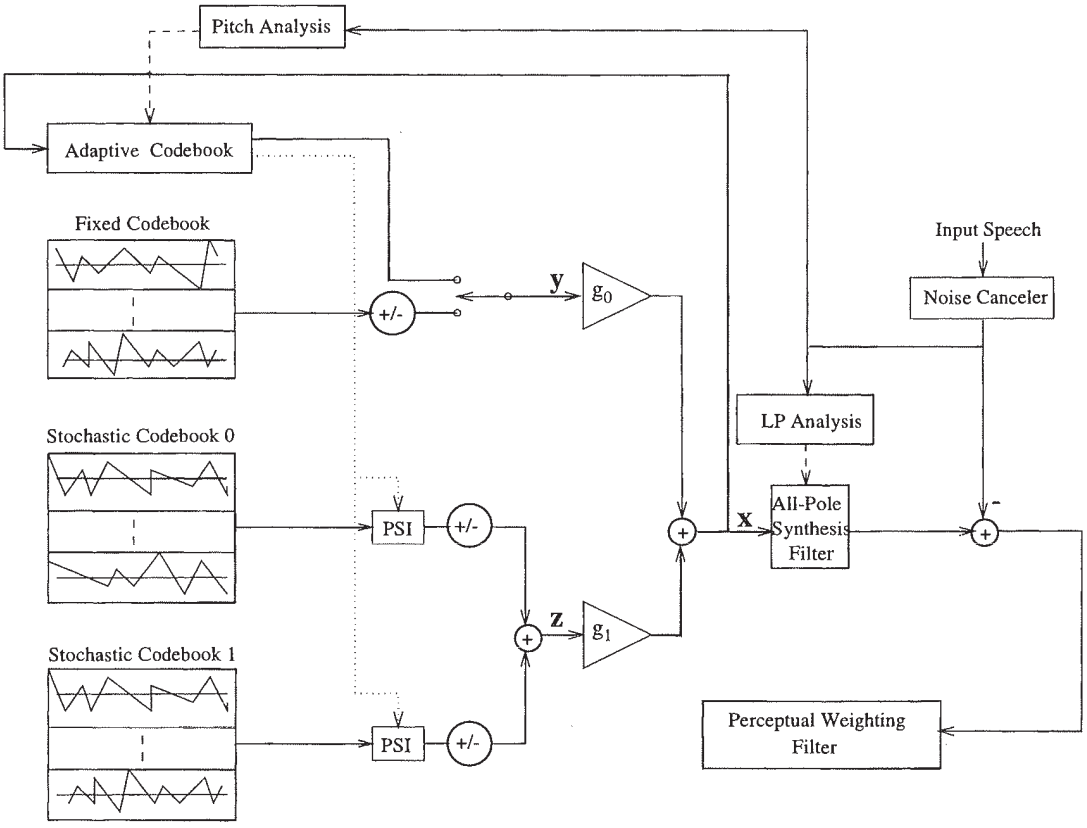


FIGURE 30.2 Basic structure of the PSI-CELP encoder.

[2]There were two candidate algorithms named PSI-CELP in the PDC half-rate competition. The algorithm described here was contributed by NTT Mobile Communications Network, Inc. (NTT DoCoMo).

TABLE 30.3 Bit Allocations for the PSI-CELP Half- Rate PDC Speech Coder

Parameter	Bits	Error Protected Bits
LP synthesis filter	31	15
Frame energy	7	7
Periodic excitation	8 × 4	8 × 4
Stochastic excitation	10 × 4	0
Gain	7 × 4	3 × 4
Total	138	66

A distinctive feature of the PSI-CELP coder is the use of an adaptive noise canceller [13, 15] to suppress noise in the input signal prior to coding. The input signal is classified into various modes, depending on the presence or absence of background noise and speech and their relative power levels. The current active mode determines whether *Kalman filtering* [9] is applied to the input signal and whether the parameters of the Kalman filter are adapted. Kalman filtering is applied when a significant amount of background noise is present or when both background noise and speech are strongly present. The filter parameters are adapted to the statistics of the speech and noise signals in accordance with whether they are both present or only noise is present.

The LP filter parameters in the PSI-CELP coder are encoded using VQ. A tenth-order LP analysis is performed every 20 ms. The resultant filter parameters are converted to 10 *line spectral frequencies* (LSFs).[3] The LSF parameters have a naturally increasing order, and together are treated as the ordered components of a vector. Since the speech spectral envelope tends to evolve slowly with time, there is intervector dependency between adjacent LSF vectors that can be exploited. Thus, the two LSF vectors for each 40-ms frame are paired together and jointly encoded. Each LSF vector in the pair is split into three subvectors. The pair of subvectors that cover the same vector component indexes are combined into one composite vector and vector quantized. Altogether, 31 b are used to encode a pair of LSF vectors. This three-way *split VQ*[4] scheme embodies a compromise between the prohibitively high complexity of using a large vector dimension and the performance gain from exploiting intra- and intervector dependency.

The PSI-CELP encoder uses a perceptual weighting filter consisting of a cascade of two filter sections. The sections exploit the pitch-harmonic structure and the LP spectral-envelope structure of the speech signal, respectively. The pitch-harmonic section has four parameters, a pitch lag and three coefficients, whose values are determined from an analysis of the periodic structure of the input speech signal. Pitch-harmonic weighting reduces the amount of noise in between the pitch harmonics by aggregating coder noise to be closer to the harmonic frequencies of the speech signal. In high-pitched voice, the harmonics are spaced relatively farther apart, and pitch-harmonic weighting becomes correspondingly more important.

The excitation vector $\boldsymbol{x}$ (Fig. 30.2) is updated once every subframe interval (10 ms) and is constructed as a *linear combination* of two vectors

$$\boldsymbol{x} = g_0\boldsymbol{y} + g_1\boldsymbol{z} \tag{30.1}$$

where g_0 and g_1 are scalar gains, $\boldsymbol{y}$ is labeled as the *periodic* component of the excitation and $\boldsymbol{z}$ as the *stochastic* or *random* component. When the input speech is voiced, the ABS op-

[3] Also known as line spectrum pairs (LSPs).
[4] Matrix quantization is another possible description.

eration would find a value for $\boldsymbol{y}$ from the *adaptive codebook* (Fig. 30.2). The codebook is constructed out of past samples of the excitation signal $\boldsymbol{x}$; hence, there is a feedback path into the adaptive codebook in Fig. 30.2. Each code vector in the adaptive codebook corresponds to one of the 192 possible pitch lag L values available for encoding; the code vector is populated with samples of $\boldsymbol{x}$ beginning with the Lth sample backward in time. L is not restricted to be an integer, i.e., *fractional pitch period* is permitted. Successive values of L are more closely spaced for smaller values of L; short, medium, and long lags are quantized to one-quarter, one-half, and one sampling-period resolution, respectively. As a result, the *relative* quantization error in the encoded pitch frequency (which is the reciprocal of the encoded pitch lag) remains roughly constant with increasing pitch frequency. When the input speech is unvoiced, $\boldsymbol{y}$ would be obtained from the fixed codebook (Fig. 30.2). To find the best value for $\boldsymbol{y}$, the encoder searches through the aggregate of 256 code vectors from both the adaptive and fixed codebooks. The code vector that results in a synthesis output most resembling the input speech is selected. The best code vector thus chosen also implicitly determines the voicing condition (voiced/unvoiced) and the pitch lag value L^* most appropriate to the current subframe of input speech. These parameters are said to be determined in a *closed-loop* search.

The stochastic excitation $\boldsymbol{z}$ is formed as a sum of two code vectors, each selected from a *conjugate codebook* (Fig. 30.2) [13]. Using a pair of conjugate codebooks each of size 16 code vectors (4 b) has been found to improve robustness against channel errors, in comparison with using one single codebook of size 256 code vectors (8 b). The synthesis output due to $\boldsymbol{z}$ can be decomposed into a sum of two orthogonal components, one of which points in the same direction as the synthesis output due to the periodic excitation $\boldsymbol{y}$ and the other component points in a direction orthogonal to the synthesis output due to $\boldsymbol{y}$. The latter synthesis output component of $\boldsymbol{z}$ is kept, whereas the former component is discarded. Such decomposition enables the two gain factors g_0 and g_1 to be separately quantized. For voiced speech, the conjugate code vectors are preprocessed to produce a set of *pitch synchronous innovation* (PSI) vectors. The first L^* samples of each code vector are treated as a fundamental period of samples. The fundamental period is replicated until there are enough samples to populate a subframe. If L^* is not an integer, interpolated samples of the code vectors are used (upsampled versions of the code vectors can be precomputed). PSI has been found to reinforce the periodicity and substantially improve the quality of synthesized voiced speech.

The postfilter in the PSI-CELP decoder has three sections, for enhancing the formants, the pitch harmonics, and the high frequencies of the synthesized speech, respectively. Pitch-harmonic enhancement is applied only when the adaptive codebook has been used. Formant enhancement makes use of the decoded LP synthesis filter parameters, whereas a refined pitch analysis is performed on the synthesized speech to obtain the values for the parameters of the pitch-harmonic section of the postfilter. A first-order high-pass filter section compensates for the low-pass spectral tilt [1] of the formant enhancement section.

Of the 138 speech data bits generated by the speech encoder every 40-ms frame, 66 b (Table 30.3) receive error protection and the remaining 72 speech data bits of the frame are not error protected. An error detection code of 9 *cyclic redundancy check* (*CRC*) bits is appended to the 66 b and then submitted to a rate 1/2, punctured convolutional encoder to generate a sequence of 152 channel coded bits. Of the unprotected 72 b, the 40 b that index the excitation codebooks (Table 30.3) are remapped or *pseudo-Gray coded* [17] so as to equalize their channel error sensitivity. As a result, a bit error occurring in an index word is likely to cause about the same amount of degradation regardless of the bit error position in the index word. For each speech frame, the channel encoder emits 224 b of payload data.

The payload data from two adjacent frames are interleaved before transmission over the radio link.

Uncorrected errors in the most critical 66 b are detected with high probability as a CRC error. A finite state machine keeps track of the recent history of CRC errors. When a sequence of CRC errors is encountered, the power level of the synthesized speech is progressively suppressed, so that muting is reached after four consecutive CRC errors. Conversely, following the cessation of a sequence of CRC errors, the power level of the synthesized speech is ramped up gradually.

30.7 The European GSM Half-Rate Standard

A *vector sum excited linear prediction* (VSELP) coder, contributed by Motorola, Inc., was selected in January 1994 by the main GSM technical committee as a basis for the GSM half-rate standard. The standard was finally approved in January 1995. VSELP is a generic name for a family of algorithms from Motorola; the North American full-rate and the Japanese full-rate standards are also based on VSELP. All VSELP coders make use of the basic idea of representing the excitation signal by a linear combination of *basis vectors* [6]. This representation renders the excitation codebook search procedure very computationally efficient. A block diagram of the GSM half-rate decoder is depicted in Fig. 30.3 and bit allocations are tabulated in Table 30.4. The coder's frame size is 20 ms, and each frame comprises four subframes of 5 ms each. The coder has been optimized for execution on a processor with 16-b word length and 32-b accumulator. The GSM standard is a *bit exact* specification: in addition to specifying the codec's processing steps, the numerical formats and precisions of the codec's variables are also specified.

TABLE 30.4 Bit Allocations for the VSELP Half-Rate GSM Coder

Parameter	Bits/subframe	Bits/frame
LP synthesis filter		28
Soft interpolation		1
Frame energy		5
Mode selection		2
Mode 0		
Excitation code I	7	28
Excitation code H	7	28
Gain code G_s, P_0	5	20
Mode 1, 2, and 3		
Pitch lag L (first subframe)		8
Difference lag (subframes 2, 3, 4)	4	12
Excitation code J	9	36
Gain code G_s, P_0	5	20
Total		112

The synthesis filter coefficients in GSM VSELP are encoded using the *fixed point lattice technique* (FLAT) [8] and vector quantization. FLAT is based on the *lattice filter* representation of the linear prediction error filter. The tenth-order lattice filter has 10 stages, with the ith stage, $i \in \{1, \ldots, 10\}$, containing a *reflection coefficient* parameter r_i. The lattice filter has an *order-recursion* property such that the best prediction error filters of all orders less than ten are all embedded in the best tenth-order lattice filter. This means that once the values of the lower order reflection coefficients have been optimized, they do not have to be reoptimized when a higher order predictor is desired; in other words, the coefficients can be optimized sequentially from low to high orders. On the other hand, if the lower order coefficients were suboptimal (as in the case when the coefficients are quantized), the higher

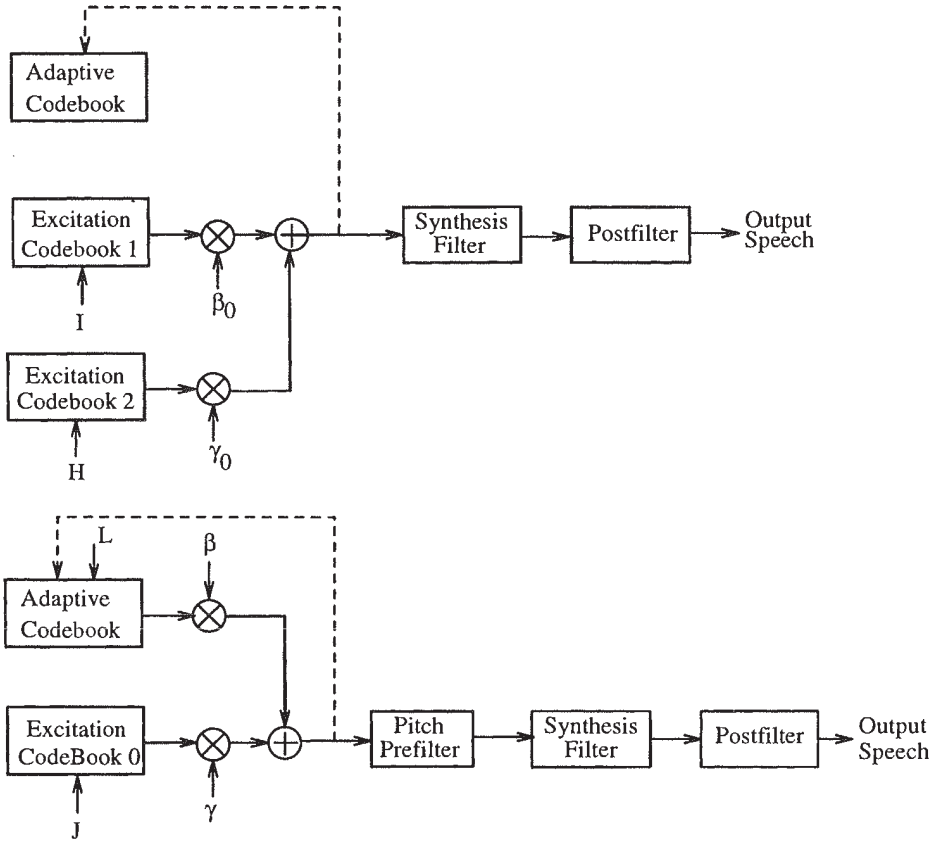


FIGURE 30.3 Basic structure of the GSM VSELP decoder. Top is for mode 0 and bottom is for modes 1, 2, and 3.

order coefficients could still be selected to minimize the prediction *residual* (or error) energy at the output of the higher order stages; in effect, the higher order stages can compensate for the suboptimality of lower order stages.

In the GSM VSELP coder, the ten reflection coefficients $\{r_1, \ldots, r_{10}\}$ that have to be encoded for each frame are grouped into three coefficient vectors $\boldsymbol{v}_1 = [r_1 r_2 r_3], \boldsymbol{v}_2 = [r_4 r_5 r_6], \boldsymbol{v}_3 = [r_7 r_8 r_9 r_{10}]$. The vectors are quantized sequentially, from $\boldsymbol{v}_1$ to $\boldsymbol{v}_3$, using a b_i-bit VQ codebook C_i for $\boldsymbol{v}_i$, where b_i, $i = 1, 2, 3$ are 11, 9, and 8 b, respectively. The vector $\boldsymbol{v}_i$ is quantized to minimize the prediction error at the energy output of the jth stage of the lattice filter where r_j is the highest order coefficient in the vector $\boldsymbol{v}_i$. The computational complexity associated with quantizing $\boldsymbol{v}_i$ is reduced by searching only a small subset of the code vectors in C_i. The subset is determined by first searching a *prequantizer* codebook of size c_i bits, where c_i, $i = 1, \ldots, 3$ are 6, 5, and 4 b, respectively. Each code vector in the prequantizer codebook is associated with $2^{b_i - c_i}$ code vectors in the target codebook. The subset is obtained by pooling together all of the code vectors in C_i that are associated with the top few best matching prequantizer code vectors. In this way, a factor of reduction in computational complexity of nearly $2^{b_i - c_i}$ is obtained for the quantization of $\boldsymbol{v}_i$.

The half-rate GSM coder changes its configuration of excitation generation (Fig. 30.3) in accordance with a *voicing mode* [7]. For each frame, the coder selects one of four possible voicing modes depending on the values of the *open-loop* pitch-prediction gains computed

for the frame and its four subframes. Open loop refers to determining the pitch lag and the pitch-predictor coefficient(s) via a direct analysis of the input speech signal or, in the case of the half-rate GSM coder, the perceptually weighted (LP-weighting only) input signal. Open-loop analysis can be regarded as the opposite of closed-loop analysis, which in our context is synonymous with ABS. When the pitch-prediction gain for the frame is weak, the input speech signal is deemed to be unvoiced and mode 0 is used. In this mode, two 7-b *trained* codebooks (excitation codebooks 1 and 2 in Fig. 30.3) are used, and the excitation signal for each subframe is formed as a linear combination of two code vectors, one from each of the codebooks. A trained codebook is one designed by applying the coder to a representative set of speech signals while optimizing the codebook to suit the set. Mode 1, 2, or 3 is chosen depending on the strength of the pitch-prediction gains for the frame and its subframes. In these modes, the excitation signal is formed as a linear combination of a code vector from an 8-b adaptive codebook and a code vector from a 9-b trained codebook (Fig. 30.3). The code vectors that are summed together to form the excitation signal for a subframe are each scaled by a gain factor (β and γ in Fig. 30.3). Each mode uses a gain VQ codebook specific to that mode.

As depicted in Fig. 30.3, the decoder contains an adaptive pitch prefilter for the voiced modes and an adaptive postfilter for all modes. The filters enhance the perceptual quality of the decoded speech and are not present in the encoder. It is more conventional to locate the pitch prefilter as a section of the postfilter; the distinctive placement of the pitch prefilter in VSELP was chosen to reduce artifacts caused by the time-varying nature of the filter. In mode 0, the encoder uses an LP spectral weighting filter in its ABS search of the two excitation codebooks. In the other modes, the encoder uses a pitch-harmonic weighting filter in cascade with an LP spectral weighting filter for searching excitation codebook 0, whereas only LP spectral weighting is used for searching the adaptive codebook. The pitch-harmonic weighting filter has two parameters, a pitch lag and a coefficient, whose values are determined in the aforementioned open-loop pitch analysis.

A code vector in the 8-b adaptive codebook has a dimension of 40 (the duration of a subframe) and is populated with past samples of the excitation signal beginning with the Lth sample back from the present time. L can take on one of 256 different integer and fractional values. The best adaptive code vector for each subframe can be selected via a complete ABS; the required exhaustive search of the adaptive codebook is, however, computationally expensive. To reduce computation, the GSM VSELP coder makes use of the aforementioned open-loop pitch analysis to produce a list of *candidate lag values*. The open-loop pitch-prediction gains are ranked in decreasing order, and only the lags corresponding to top-ranked gains are kept as candidates. The final decisions for the four L values of the four subframes in a frame are made jointly. By assuming that the four L values can not vary over the entire range of all possible 256 values in the short duration of a frame, the L of the first subframe is coded using 8 b, and the L of each of the other three subframes is coded *differentially* using 4 b. The 4 b represent 16 possible values of deviation relative to the lag of the previous subframe. The four lags in a frame trace out a *trajectory* where the change from one time point to the next is restricted; consequently, only 20 b are needed instead of 32 b for encoding the four lags. Candidate trajectories are constructed by linking top ranked lags that are commensurate with differential encoding. The best trajectory among the candidates is then selected via ABS.

The trained excitation codebooks of VSELP have a special vector sum structure that facilitates fast searching [6]. Each of the 2^b code vectors in a b-bit trained codebook is formed as a linear combination of b *basis vectors*. Each of the b scalar weights in the linear combination is restricted to have a binary value of either 1 or -1. The 2^b code vectors in the codebook are obtained by taking all 2^b possible combinations of values of the weights. A

substantial storage saving is incurred by storing only b basis vectors instead of 2^b code vectors. Computational saving is another advantage of the vector-sum structure. Since filtering is a linear operation, the synthesis output due to each code vector is a linear combination of the synthesis outputs due to the individual basis vectors, where the same weight values are used in the output linear combination as in forming the code vector. A vector sum codebook can be searched by first performing synthesis filtering on its b basis vectors. If, for the present subframe, another trained codebook (mode 0) or an adaptive codebook (mode 1, 2, 3) had been searched, the filtered basis vectors are further orthogonalized with respect to the signal synthesized from that codebook, i.e., each filtered basis vector is replaced by its own component that is orthogonal to the synthesized signal. Further complexity reduction is obtained by examining the code vectors in a sequence such that two successive code vectors differ in only one of the b scalar weight values; that is, the entire set of 2^b code vectors is searched in a *Gray coded* sequence. With successive code vectors differing in only one term in the linear combination, it is only necessary in the codebook search computation to progressively track the difference [6].

The total energy of a speech frame is encoded with 5 b (Table 30.4). The two gain factors (β and γ in Fig. 30.3) for each subframe are computed after the excitation codebooks have been searched and are then transformed to parameters G_s and P_0 to be vector quantized. Each mode has its own 5-b gain VQ codebook. G_s represents the energy of the subframe relative to the total frame energy, and P_0 represents the fraction of the subframe energy due to the first excitation source (excitation codebook 1 in mode 0, or the adaptive codebook in the other modes).

An *interpolation bit* (Table 30.4) transmitted for each frame specifies to the decoder whether the LP synthesis filter parameters for each subframe should be obtained from interpolating between the decoded filter parameters for the current and the previous frames. The encoder determines the value of this bit according to whether interpolation or no interpolation results in a lower prediction residual energy for the frame. The postfilter in the decoder operates in concordance with the actual LP parameters used for synthesis.

The speech encoder generates 112 b of encoded data (Table 30.4) for every 20-ms frame of the speech signal. These bits are processed by the channel encoder to improve, after channel decoding at the receiver, the uncoded bit-error rate and the detectability of uncorrected errors. Error detection coding in the form of 3 CRC bits is applied to the most critical 22 data bits. The combined 25 b plus an additional 73 speech data bits and 6 *tail bits* are input to an RCPCC encoder (the tail bits serve to bring the channel encoder and decoder to a fixed terminal state at the end of the payload data stream). The 3 CRC bits are encoded at rate 1/3 and the other 101 b are encoded at rate 1/2, generating a total of 211 channel coded bits. These are finally combined with the remaining 17 (uncoded) speech data bits to form a total of 228 b for the payload data of a speech frame. The payload data from two speech frames are interleaved for transmission over four timeslots of the GSM TDMA channel.

With the Viterbi algorithm, the channel decoder performs soft decision decoding on the demodulated and deinterleaved channel data. Uncorrected channel errors may still be present in the decoded speech data after Viterbi decoding. Thus, the channel decoder classifies each frame into three integrity categories: bad, unreliable, and reliable, in order to assist the speech decoder in undertaking error concealment measures. A frame is considered bad if the CRC check fails or if the received channel data is close to more than one candidate sequence. The latter evaluation is based on applying an adaptive threshold to the metric values produced by the Viterbi algorithm over the course of decoding the most critical 22 speech data bits and their 3 CRC bits. Frames that are not bad may be classified

as unreliable, depending on the metric values produced by the Viterbi algorithm and on channel reliability information supplied by the demodulator.

Depending on the recent history of decoded data integrity, the speech decoder can take various error concealment measures. The onset of bad frames is concealed by repetition of parameters from previous reliable frames, whereas the persistence of bad frames results in power attenuation and ultimately muting of the synthesized speech. Unreliable frames are decoded with normality constraints applied to the energy of the synthesized speech.

30.8 Conclusions

The half-rate standards employ some of the latest techniques in speech and channel coding to meet the challenges posed by the severe transmission environment of digital cellular radio systems. By halving the bit rate, the voice transmission capacity of existing full-rate digital cellular systems can be doubled. Although advances are still being made that can address the needs of quarter-rate speech transmission, much effort is currently devoted to enhancing the speech quality and robustness of full-rate (GSM and IS-54) systems, aiming to be closer to *toll quality*. On the other hand, the imminent introduction of competing wireless systems that use different modulation schemes [e.g., coded division multiple access (CDMA)] and/or different radio frequencies [e.g., personal communications systems (PCS)] is poised to alleviate congestion in high-user-density areas.

Defining Terms

Codebook: An ordered collection of all possible values that can be assigned to a scalar or vector variable. Each unique scalar or vector value in a codebook is called a *codeword*, or *code vector* where appropriate.

Codec: A contraction of *(en)coder–decoder*, used synonymously with the word *coder*. The encoder and decoder are often designed and deployed as a pair. A half-rate standard codec performs speech as well as channel coding.

Echo canceller: A signal processing device that, given the source signal causing the echo signal, generates an estimate of the echo signal and subtracts the estimate from the signal being interfered with by the echo signal. The device is usually based on a discrete-time adaptive filter.

Pitch period: The fundamental period of a voiced speech waveform that can be regarded as periodic over a short-time interval (quasiperiodic). The reciprocal of pitch period is *pitch frequency* or simply, *pitch*.

Tandem coding: Having more than one encoder–decoder pair in an end-to-end transmission path. In cellular radio communications, having a radio link at each end of the communication path could subject the speech signal to two passes of speech encoding–decoding. In general, repeated encoding and decoding increases the distortion.

Acknowledgment

The authors would like to thank Erdal Paksoy and Mark A. Jasiuk for their valuable comments.

References

[1] Chen, J.-H. and Gersho, A., Adaptive postfiltering for quality enhancement of coded speech. *IEEE Trans. Speech & Audio Proc.*, 3(1), 59–71, 1995.

[2] De Martino, E., Speech quality evaluation of the European, North-American and Japanese speech codec standards for digital cellular systems. In *Speech and Audio Coding for Wireless and Network Applications,* Atal, B.S., Cuperman, V., and Gersho, A., Eds., 55–58, Kluwer Academic Publishers, Norwell, MA, 1993.

[3] Dimolitsas, S., Corcoran, F.L., and Baraniecki, M.R., Transmission quality of North American cellular, personal communications, and public switched telephone networks. *IEEE Trans. Veh. Tech.*, 43(2), 245–251, 1994.

[4] Gersho, A., Advances in speech and audio compression. *Proc. IEEE,* 82(6), 900–918, 1994.

[5] Gersho, A. and Gray, R.M., *Vector Quantization and Signal Compression,* Kluwer Academic Publishers, Norwell, MA, 1991.

[6] Gerson, I.A. and Jasiuk, M.A., Vector sum excited linear prediction (VSELP) speech coding at 8 kbps. In *Proceedings, IEEE Intl. Conf. Acoustics, Speech, & Sig. Proc.*, 461–464, April, 1990.

[7] Gerson, I.A. and Jasiuk, M.A., Techniques for improving the performance of CELP—type speech coders. *IEEE J. Sel. Areas Comm.*, 10(5), 858–865, 1992.

[8] Gerson, I.A., Jasiuk, M.A., Nowack, J.M., Winter, E.H., and Müller, J.-M., Speech and channel coding for the half-rate GSM channel. In *Proceedings, ITG-Report 130 on Source and Channel Coding,* 225–232. Munich, Germany, Oct., 1994.

[9] Gibson, J.D., Koo, B., and Gray, S.D., Filtering of colored noise for speech enhancement and coding. *IEEE Trans. Sig. Proc.*, 39(8), 1732–1742, 1991.

[10] Hagenauer, J., Rate-compatible punctured convolutional codes (RCPC codes) and their applications. *IEEE Trans. Comm.*, 36(4), 389–400, 1988.

[11] Jayant, N.S. and Noll, P., *Digital Coding of Waveforms,* Prentice-Hall, Englewood Cliffs, NJ, 1984.

[12] Masui, F. and Oguchi, M.,. Activity of the half rate speech codec algorithm selection for the personal digital cellular system. *Tech. Rept. of IEICE,* RCS93-77(11), 55–62 (in Japanese), 1993.

[13] Ohya, T., Suda, H., and Miki, T., 5.6 kbits/s PSI-CELP of the half-rate PDC speech coding standard. In *Proceedings, IEEE Veh. Tech. Conf.*, 1680–1684, June, 1994.

[14] Proakis, J.G., *Digital Communications,* 3rd ed., McGraw-Hill, New York, 1995.

[15] Suda, H., Ikeda, K., and Ikedo, J., Error protection and speech enhancement schemes of PSI-CELP, *NTT R & D.* (Special issue on PSI-CELP speech coding system for mobile communications), 43(4), 373–380, (in Japanese), 1994.

[16] Telecommunication Industries Association (TIA). Half-rate speech codec test plan V6.0. TR45.3.5/93.05.19.01, 1993.

[17] Zeger, K. and Gersho, A., Pseudo-Gray coding. *IEEE Trans. Comm.*, 38(12), 2147–2158, 1990.

Further Information

Additional technical information on speech coding can be found in the books, periodicals, and conference proceedings that appear in the list of references. Other relevant publications not represented in the list are *Speech Communication,* Elsevier Science Publishers; *Advances in Speech Coding,* B. S. Atal, V. Cuperman, and A, Gersho, eds., Kluwer Academic Publishers; and *Proceedings of the IEEE Workshop on Speech Coding.*

31

Wireless Video Communications

31.1 Introduction 31-1
31.2 Wireless Video Communications 31-2
Recommendation H.223
31.3 Error Resilient Video Coding 31-5
A Standard Video Coder • Error Resilient Video Decoding • Classification of Error-Resilience Techniques
31.4 MPEG-4 Error Resilience Tools 31-9
Resynchronization • Data Partitioning • Reversible Variable Length Codes (RVLCs) • Header Extension Code (HEC) • Adaptive Intra Refresh (AIR)
31.5 H.263 Error Resilience Tools 31-14
Slice Structure Mode (Annex K) • Independent Segment Decoding (ISD) (Annex R) • Error Tracking (Appendix I) • Reference Picture Selection (Annex N)
31.6 Discussion 31-16
Defining Terms 31-17
References 31-17
Further Information 31-18

Madhukar Budagavi
Texas Instruments

Raj Talluri
Texas Instruments

31.1 Introduction

Recent advances in technology have resulted in a rapid growth in mobile communications. With this explosive growth, the need for reliable transmission of mixed media information—audio, video, text, graphics, and speech data—over wireless links is becoming an increasingly important application requirement. The bandwidth requirements of raw video data are very high (a 176 × 144 pixels, color video sequence requires over 8 Mb/s). Since the amount of bandwidth available on current wireless channels is limited, the video data has to be compressed before it can be transmitted on the wireless channel. The techniques used for video compression typically utilize predictive coding schemes to remove redundancy in the video signal. They also employ variable length coding schemes, such as Huffman codes, to achieve further compression.

The wireless channel is a noisy fading channel characterized by long bursts of errors [8]. When compressed video data is transmitted over wireless channels, the effect of channel errors on the video can be severe. The variable length coding schemes make the compressed bitstream sensitive to channel errors. As a result, the video decoder that is decoding the corrupted video bitstream can easily lose synchronization with the encoder. Predictive coding techniques, such as **block motion compensation,** which are used in current video compression standards, make the matter worse by quickly propagating the effects of channel

0-8493-8597-0/99/$0.00+$.50

© 1999 by CRC Press LLC

errors across the video sequence and rapidly degrading the video quality. This may render the video sequence totally unusable.

Error control coding [5], in the form of **Forward Error Correction (FEC)** and/or **Automatic Repeat reQuest (ARQ),** is usually employed on wireless channels to improve the channel conditions. FEC techniques prove to be quite effective against random bit errors, but their performance is usually not adequate against longer duration burst errors. FEC techniques also come with an increased overhead in terms of the overall bitstream size; hence, some of the coding efficiency gains achieved by video compression are lost. ARQ techniques typically increase the delay and, therefore, might not be suitable for real-time videoconferencing. Thus, in practical video communication schemes, error control coding is typically used only to provide a certain level of error protection to the compressed video bitstream, and it becomes necessary for the video coder to accept some level of errors in the video bitstream. Error-resilience tools are introduced in the video codec to handle these residual errors that remain after error correction.

The emphasis in this chapter is on discussing relevant international standards that are making wireless video communications possible. We will concentrate on both the error control and source coding aspects of the problem. In the next section, we give an overview of a wireless video communication system that is a part of a complete wireless multimedia communication system. The International Telecommunication Union—Telecommunications Standardization Sector (ITU-T) H.223 [1] standard that describes a method of providing error protection to the video data before it is transmitted is also described. It should be noted that the main function of H.223 is to multiplex/demultiplex the audio, video, text, graphics, etc., which are typically communicated together in a videoconferencing application—error protection of the transmitted data becomes a requirement to support this functionality on error-prone channels. In Section 31.3, an overview of error-resilient video coding is given. The specific tools adopted into the International Standards Organization (ISO)/International Electrotechnical Commission (IEC) Motion Picture Experts Group (MPEG) v.4 (i.e., MPEG-4) [7] and the ITU-T H.263 [3] video coding standards to improve the error robustness of the video coder are described in Sections 31.4 and 31.5, respectively.

Table 31.1 provides a listing of some of the standards that are described or referred to in this chapter.

31.2 Wireless Video Communications

Figure 31.1 shows the basic block diagram of a wireless video communication system [10]. Input video is compressed by the video encoder to generate a compressed bitstream. The transport coder converts the compressed video bitstream into data units suitable for transmission over wireless channels. Typical operations carried out in the transport coder include channel coding, framing of data, modulation, and control operations required for accessing the wireless channel. At the receiver side, the inverse operations are performed to reconstruct the video signal for display.

In practice, the video communication system is part of a complete multimedia communication system and needs to interact with other system components to achieve the desired functionality. Hence, it becomes necessary to understand the other components of a multimedia communication system in order to design a good video communication system. Figure 31.2 shows the block diagram of a wireless multimedia terminal based on the ITU-T H.324 set of standards [4]. We use the H.324 standard as an example because it is the first videoconferencing standard for which mobile extensions were added to facilitate use on

TABLE 31.1 List of Relevant Standards

ISO/IEC 14496-2 (MPEG-4)	Information Technology—Coding of Audio-Visual Objects: Visual
H.263 (Version 1 and Version 2)	Video coding for low bitrate communication
H.261	Video codec for audiovisual services at p X 64 kbit/s
H.223	Multiplexing protocol for low bitrate multimedia communication
H.324	Terminal for low bitrate multimedia communication
H.245	Control protocol for multimedia communication
G.723.1	Dual rate speech coder for multimedia communication transmitting at 5.3 and 6.3 kbit/s

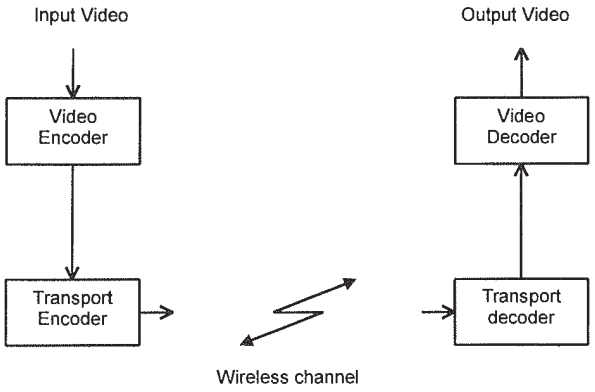


FIGURE 31.1 A wireless video communication system.

wireless channels. The system components of a multimedia terminal can be grouped into three processing blocks: (1) audio, video, and data (the word *data* is used here to mean still images/slides, shared files, documents etc.), (2) control, and (3) multiplex-demultiplex blocks.

1. Audio, video, and data processing blocks—These blocks basically produce/consume the multimedia information that is communicated. The aggregate bitrate generated by these blocks is restricted due to limitations of the wireless channel and, therefore, the total rate allowed has to be judiciously allocated among these blocks. Typically, the video blocks use up the highest percentage of the aggregate rate, followed by audio and then data. H.324 specifies the use of H.261/H.263 for video coding and G.723.1 for audio coding.
2. Control block—This block has a wide variety of responsibilities all aimed at setting up and maintaining a multimedia call. The control block facilitates the set-up of compression methods and preferred bitrates for audio, video, and data to be used in the multimedia call. It is also responsible for end-to-network signalling for accessing the network and end-to-end signalling for reliable operation of the multimedia call. H.245 is the control protocol in the H.324 suite of standards that specifies the control messages to achieve the above functionality.
3. Multiplex-Demultiplex (MUX) block—This block multiplexes the resulting audio, video, data, and control signals into a single stream before transmission on the network. Similarly, the received bitstream is demultiplexed to obtain the audio, video, data, and control signals, which are then passed to their respec-

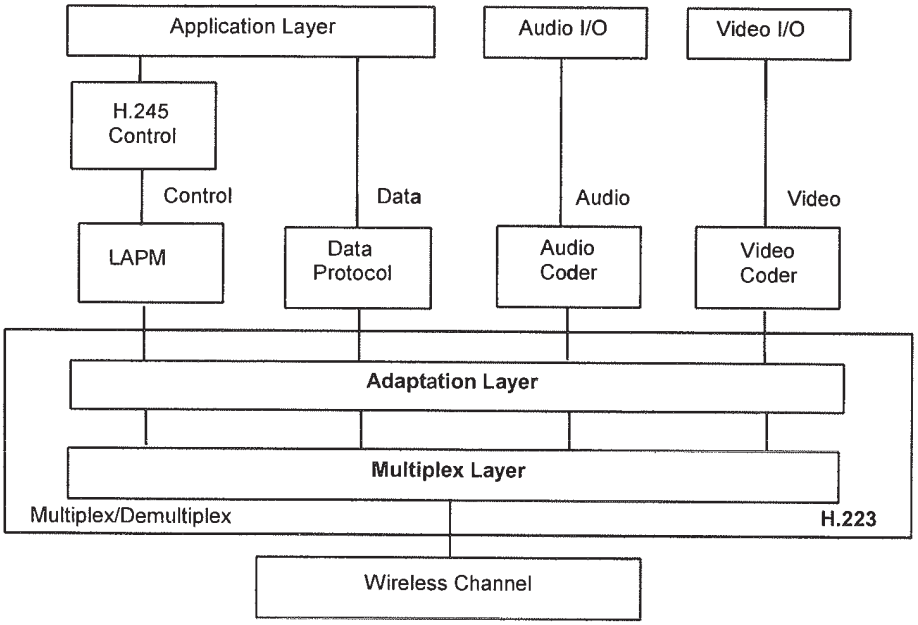


FIGURE 31.2 Configuration of a wireless multimedia terminal.

tive processing blocks. The MUX block accesses the network through a suitable network interface. The H.223 standard is the multiplexing scheme used in H.324.

Proper functioning of the MUX is crucial to the operation of the video communication system, as all the multimedia data/signals flow through it. On wireless channels, transmission errors can lead to a breakdown of the MUX resulting in, for example, nonvideo data being channeled to the video decoder or corrupted video data being passed on to the video decoder. Three annexes were specifically added to H.223 to enable its operation in error-prone environments. Below, we give a more detailed overview of H.223 and point out the levels of error protection provided by H.223 and its three annexes. It should also be noted that MPEG-4 does not specify a lower-level MUX like H.223, and thus H.223 can also be used to transmit MPEG-4 video data.

31.2.1 Recommendation H.223

Video, audio, data, and control information is transmitted in H.324 on distinct logical channels. H.223 determines the way in which the logical channels are mixed into a single bitstream before transmission over the physical channel (e.g., the wireless channel). The H.223 multiplex consists of two layers—the multiplex layer and the adaptation layer, as shown in Fig. 31.2. The multiplex layer is responsible for multiplexing the various logical channels. It transmits the multiplexed stream in the form of packets. The adaptation layer adapts the information stream provided by the applications above it to the multiplex layer below it by adding, where appropriate, additional octets for the purposes of error control and sequence numbering. The type of error control used depends on the type of information (audio/video/data/control) being conveyed in the stream. The adaptation layer provides error control support in the form of both FEC and ARQ.

H.223 was initially targeted for use on the benign general switched telephone network (GSTN). Later on, to enable its use on wireless channels, three annexes (referred to as

Levels 1–3, respectively), were defined to provide improved levels of error protection. The initial specification of H.223 is referred to as Level 0. Together, Levels 0–3 provide for a trade-off of error robustness against the overhead required, with Level 0 being the least robust and using the least amount of overhead and Level 3 being the most robust and also using the most amount of overhead.

1. H.223 Level 0—Default mode. In this mode the transmitted packet sizes are of variable length and are delimited by an 8-bit HDLC (High-level Data Link Control) flag (**01111110**). Each packet consists of a 1-octet header followed by the payload, which consists of a variable number of information octets. The header octet includes a Multiplex Code (MC) which specifies, by indexing to a multiplex table, the logical channels to which each octet in the information field belongs. To prevent emulation of the HDLC flag in the payload, bitstuffing is adopted.
2. H.223 Level 1 (Annex A)—Communication over low error-prone channels. The use of bitstuffing leads to poor performance in the presence of errors; therefore in Level 1, bitstuffing is not performed. The other improvement incorporated in Level 1 is the use of a longer 16-bit pseudo-noise synchronization flag to allow for more reliable detection of packet boundaries. The input bitstream is correlated with the synchronization flag and the output of the correlator is compared with a correlation threshold. Whenever the correlator output is equal to or greater than the threshold, a flag is detected. Since, bitstuffing is not performed, it is possible to have this flag emulated in the payload. However, the probability of such an emulation is low and is outweighed by the improvement gained by not using bitstuffing over error-prone channels.
3. H.223 Level 2 (Annex B)—Communication over moderately error-prone channels. When compared to the Level 1 operation, Level 2 increases the protection on the packet header. A Multiplex Payload Length (MPL) field, which gives the length of the payload in bytes, is introduced into the header to provide additional redundancy for detecting the length of the video packet. A (24,12,8) extended Golay code is used to protect the MC and the MPL fields. Use of error protection in the header enables robust delineation of packet boundaries. Note that the payload data is not protected in Level 2.
4. H.223 Level 3 (Annex C)—Communication over highly error-prone channels. Level 3 goes one step above Level 2 and provides for protection of the payload data. Rate Compatible Punctured Convolutional (RCPC) codes, various CRC polynomials, and ARQ techniques are used for protection of the payload data. Level 3 allows for the payload error protection overhead to vary depending on the channel conditions. RCPC codes are used for achieving this adaptive level of error protection because RCPC codes use the same channel decoder architecture for all the allowed levels of error protection, thereby reducing the complexity of the MUX.

31.3 Error Resilient Video Coding

Even after error control and correction, some amount of residual errors still exist in the compressed bitstream fed to the video decoder in the receiver. Therefore, the video decoder should be robust to these errors and should provide acceptable video quality even

in the presence of some residual errors. In this section, we first describe a standard video coder configuration that is the basis of many international standards and also highlight the potential problems that are encountered when compressed video from these systems is transmitted over wireless channels. We then give an overview of the strategies that can be adopted to overcome these problems. Most of these strategies are incorporated in the MPEG-4 video coding standard and the H.263 (Version 2) video coding standard [3]. The original H.263 standard [2] which was standardized in 1996 for use in H.324 terminals connected to GSTN is referred to as Version 1. Version 2 of the H.263 standard provides additional improvements and functionalities (which include error-resilience tools) over the Version 1 standard. We will use H.263 to refer to both Version 1 and Version 2 standards and a distinction will be made only when required.

31.3.1 A Standard Video Coder

Redundancy exists in video signals in both spatial and temporal dimensions. Video coding techniques exploit this redundancy to achieve compression. A plethora of video compression techniques have been proposed in the literature, but a hybrid coding technique consisting of block motion compensation (BMC) and discrete cosine transforms (DCT) has been found to be very effective in practice. In fact, most of the current video coding standards such as H.263 and MPEG-4, which provide state-of-the-art compression performance, are all based on this hybrid coding technique. In this hybrid BMC/DCT coding technique, BMC is used to exploit temporal redundancy and the DCT is used to reduce spatial redundancy.

Figure 31.3 illustrates a standard hybrid BMC/DCT video coder configuration. Pictures are coded in either of two modes—interframe (INTER) or intraframe (INTRA) mode. In intraframe coding, the video image is encoded without any relation to the previous image, whereas in interframe coding, the current image is predicted from the previous image using BMC, and the difference between the current image and the predicted image, called the residual image, is encoded. The basic unit of data which is operated on is called a macroblock (MB) and is the data (both **luminance and chrominance** components) corresponding to a block of 16 $\times$ 16 pixels. The input image is split into disjoint macroblocks and the processing is done on a macroblock basis. Motion information, in the form of **motion vectors,** is calculated for each macroblock. The motion compensated prediction residual error is then obtained by subtracting each pixel in the macroblock with its motion shifted counterpart in the previous frame. Depending on the mode of coding used for the macroblock, either the image macroblock or the corresponding residual image macroblock is split into blocks of size 8 $\times$ 8 and an 8 $\times$ 8 DCT is applied to each of these 8 $\times$ 8 blocks. The resulting DCT coefficients are then quantized. Depending on the quantization step-size, this will result in a significant number of zero-valued coefficients. To efficiently encode the DCT coefficients that remain nonzero after quantization, the DCT coefficients are zig-zag scanned, and run-length encoded and the run-lengths are variable length encoded before transmission. Since a significant amount of correlation exists between the neighboring macroblocks' motion vectors, the motion vectors are themselves predicted from already transmitted motion vectors and the motion vector prediction error is encoded. The motion vector prediction error and the mode information are also variable length coded before transmission to achieve efficient compression.

The decoder uses a reverse process to reconstruct the macroblock at the receiver. The variable length codewords present in the received video bitstream are decoded first. For INTER macroblocks, the pixel values of the prediction error are reconstructed by inverse quantization and inverse DCT and are then added to the motion compensated pixels from the previous frame to reconstruct the transmitted macroblock. For INTRA macroblocks,

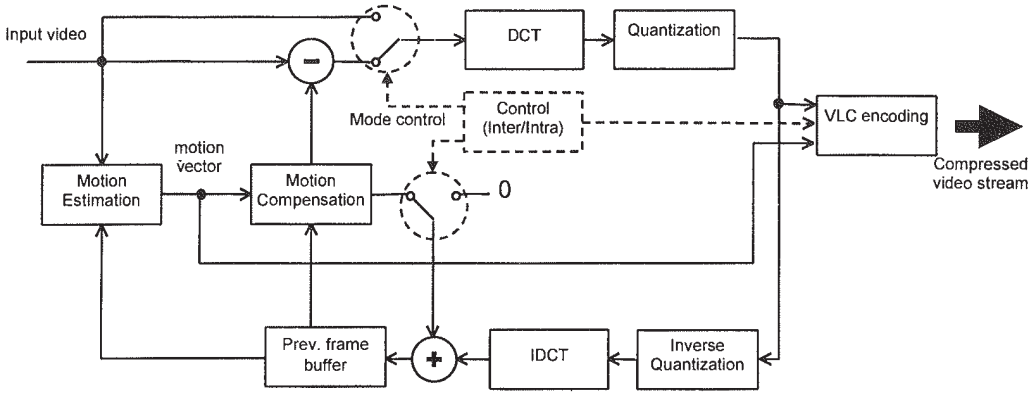


FIGURE 31.3 A standard video coder.

inverse quantization and inverse DCT directly result in the transmitted macroblock. All macroblocks of a given picture are decoded to reconstruct the whole picture.

31.3.2 Error Resilient Video Decoding

The use of predictive coding and variable length coding (VLC), though very effective from a compression point of view, makes the video decoding process susceptible to transmission errors. In VLC, the boundary between codewords is implicit. The compressed bitstream has to be read until a full codeword is encountered; the codeword is then decoded to obtain the information encoded in the codeword. When there are transmission errors, the implicit nature of the boundary between codewords typically leads to an incorrect number of bits being used in VLC decoding and, thus, subsequently results in a loss of synchronization with the encoder. In addition, the use of predictive coding leads to the propagation of these transmission errors to neighboring spatial blocks and to subsequently decoded frames, which leads to a rapid degradation in the reconstructed video quality.

To minimize the disastrous impact that transmission errors can have on the video decoding process, the following stages are incorporated in the video decoder to make it more robust:

- Error detection and localization
- Resynchronization
- Data recovery
- Error concealment

Figure 31.4 shows an error resilient video decoder configuration. The first step involved in robust video coding is the detection of errors in the bitstream. The presence of errors in the bitstream can be signaled by the FEC used in the multiplex layer. The video coder can also detect errors whenever illegal VLC codewords are encountered in the bitstream or when the decoding of VLC codewords leads to an illegal value of the decoded information (e.g., occurrence of more than 64 DCT coefficients for an 8×8 DCT block). Accurate detection of errors in the bitstream is a very important step, since most of the other error resilience techniques can only be invoked if an error is detected.

Due to the use of VLC, the location in the bitstream where the decoder detects an error is not the same location where the error has actually occurred but some undetermined distance away from it. This is shown in Fig. 31.5. Once an error is detected, it also implies that the

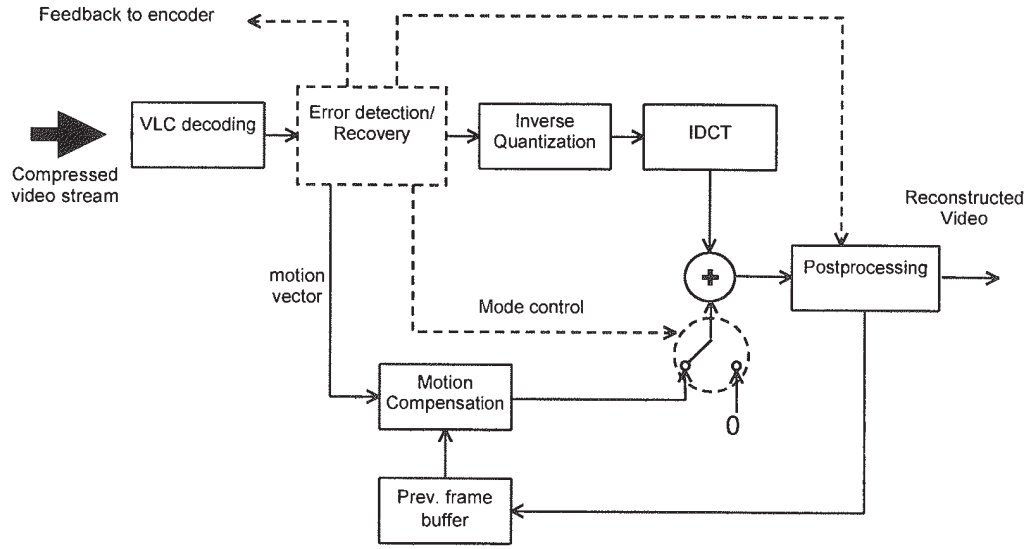


FIGURE 31.4 Error resilient video decoder.

decoder is not in synchronization with the encoder. Resynchronization schemes are then employed for the decoder to fall back into lock step with the encoder. While constructing the bitstream, the encoder inserts unique resynchronization words into the bitstream at approximately equally spaced intervals. These resynchronization words are chosen such that they are unique from the valid video bitstream. That is, no valid combination of the video algorithm's VLC tables can produce these words. The decoder, upon detection of an error, seeks forward in the bitstream looking for this known resynchronization word. Once this word is found, the decoder then falls back in synchronization with the encoder. At this point, the decoder has detected an error, regained synchronization with the encoder, and isolated the error to be between the two resynchronization points. Since the decoder can only isolate the error to be somewhere between the resynchronization points but not pinpoint its exact location, all of the data that corresponds to the macroblocks between these two resynchronization points needs to be discarded. Otherwise, the effects of displaying an image reconstructed from erroneous data can cause highly annoying visual artifacts.

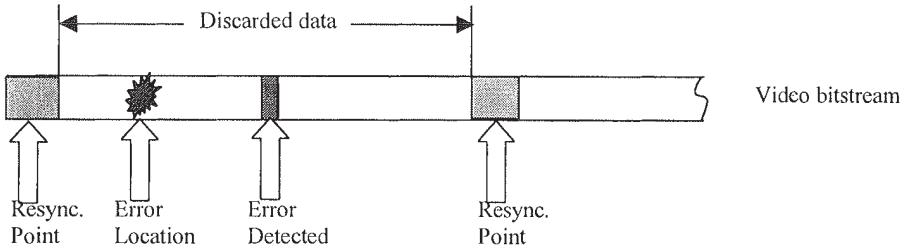


FIGURE 31.5 At the decoder, it is usually not possible to detect the error at the actual error occurrence location; hence, all the data between the two resynchronization points may need to be discarded.

Some data recovery techniques, such as "reversible decoding," enable the decoder to salvage some of the data between the two resynchronization points. These techniques advocate

the use of a special kind of VLC table at the encoder in coding the DCTs and motion vector information. These special VLCs have the property that they can be decoded both in the forward and reverse directions. By comparing the forward and reverse decoded data, the exact location of the error in the bit stream can be localized more precisely and some of the data between the two resynchronization points can be salvaged. The use of these reversible VLCs (RVLCs) is part of the MPEG-4 standard and will be described in greater detail in the following sections.

After data recovery, the impact of the data that is deemed to be in error needs to be minimized. This is the error concealment stage. One simple error concealment strategy is to simply replace the luminance and chrominance components of the erroneous macroblocks with the luminance and chrominance of the corresponding macroblocks in the previous frame of the video sequence. While this technique works fairly well and is simple to implement, more complex techniques use some type of estimation strategies to exploit the local correlation that exists within a frame of video data to come up with a better estimate of the missing or erroneous data. These error concealment strategies are essentially postprocessing algorithms and are not mandated by the video coding standards. Different implementations of the wireless video systems utilize different kinds of error concealment strategies based on the available computational power and the quality of the channel.

If there is support for a decoder feedback path to the encoder as shown in Fig. 31.3, this path can be used to signal detected errors. The feedback information from the decoder can be used to retransmit data or to influence future encoder action so as to stop the propagation of detected errors in the decoder. Note that for the feedback to take place, the network must support a back channel.

31.3.3 Classification of Error-Resilience Techniques

In general, techniques to improve the robustness of the video coder can be classified into three categories based on whether the encoder or the decoder plays a primary part in improving the error robustness [10]. *Forward error resilience* techniques refer to those techniques where the encoder plays the primary part in improving the error robustness, typically by introducing redundancy in the transmitted information. In *postprocessing* techniques, the decoder plays the primary part and does concealment of errors by estimation and interpolation (e.g., spatial-temporal filtering) using information it has already received. In *interactive error* resilience techniques, the decoder and the encoder interact to improve the error resilience of the video coder. Techniques that use decoder feedback come under this category.

31.4 MPEG-4 Error Resilience Tools

MPEG-4 is an ISO/IEC standard being developed by the Motion Pictures Expert Group. Initially MPEG was aimed primarily at low-bit-rate communications; however, its scope was later expanded to be much more of a multimedia coding standard [7]. The MPEG-4 video coding standard is the first video coding standard to address the problem of efficient representation of visual objects of arbitrary shape. MPEG-4 was also designed to provide "universal accessibility," i.e., the ability to access audio-visual information over a wide range of storage and transmission media. In particular, because of the proliferation of wireless communications, this implied development of specific tools to enable error-resilient transmission of compressed data over noisy communication channels.

A number of tools have been incorporated into the MPEG-4 video coder to make it more error resilient. All these tools are basically forward error resilience tools. We describe below each of these tools and its advantages.

31.4.1 Resynchronization

As mentioned earlier, a video decoder that is decoding a corrupted bitstream may lose synchronization with the encoder (i.e., it is unable to identify the precise location in the image where the current data belongs). If remedial measures are not taken, the quality of the decoded video rapidly degrades and becomes unusable. One approach is for the encoder to introduce resynchronization markers in the bitstream at various locations. When the decoder detects an error, it can then look for this resynchronization marker and regain synchronization.

Previous video coding standards such as H.261 and H.263 (Version 1) logically partition each of the images to be encoded into rows of macroblocks called Group Of Blocks (GOBs). These GOBs correspond to a horizontal row of macroblocks for **QCIF** images. Figure 31.6 shows the GOB numbering scheme for H.263 (Version 1) for QCIF resolution. For error resilience purposes, H.263 (Version 1) provides the encoder an option of inserting resynchronization markers at the beginning of each of the GOBs. Hence, for QCIF images these resynchronization markers are allowed to occur only at the left edge of the images. The smallest region that the error can be isolated to and concealed in this case is thus one row of macroblocks.

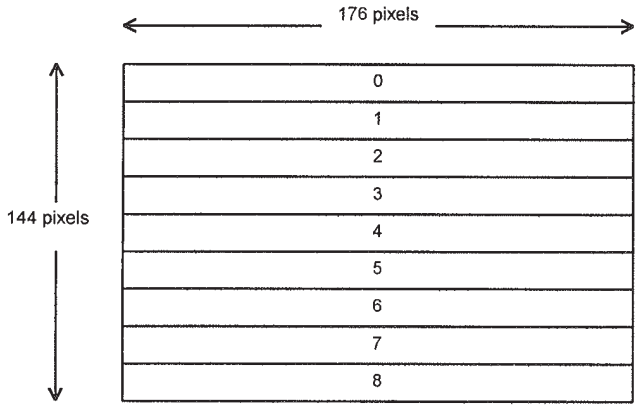


FIGURE 31.6 H.263 GOB numbering for a QCIF image.

In contrast, the MPEG-4 encoder is not restricted to inserting the resynchronization markers only at the beginning of each row of macroblocks. The encoder has the option of dividing the image into video packets. Each video packet is made up of an integer number of consecutive macroblocks in raster scan order. These macroblocks can span several rows of macroblocks in the image and can even include partial rows of macroblocks. One suggested mode of operation for the MPEG-4 encoder is for it to insert a resynchronization marker periodically at approximately every K bits. Note that resynchronization markers can only be placed at a macroblock boundary and, hence, the video packet length cannot be constrained to be exactly equal to K bits. When there is a significant activity in one part of the image, the macroblocks corresponding to these areas generate more bits than other parts

of the image. If the MPEG-4 encoder inserts the resynchronization markers at uniformly spaced bit intervals, the macroblock interval between the resynchronization markers is a lot closer in the high activity areas and a lot farther apart in the low activity areas. Thus, in the presence of a short burst of errors, the decoder can quickly localize the error to within a few macroblocks in the important high activity areas of the image and preserve the image quality in these important areas. In the case of H.263 (Version 1), where the resynchronization markers are restricted to be at the beginning of the GOBs, it is only possible for the decoder to isolate the errors to a fixed GOB independent of the image content. Hence, effective coverage of the resynchronization marker is reduced when compared to the MPEG-4 scheme. The recommended spacing of the resynchronization markers in MPEG-4 is based on the bitrates. For 24 Kb/s, it is recommended to insert them at intervals of 480 bits and for bitrates between 25 Kb/s to 48 Kb/s, it is recommended to place them at every 736 bits. Figures 31.7(a) and (b) illustrate the placement of resynchronization markers for H.263 (Version 1) and MPEG-4.

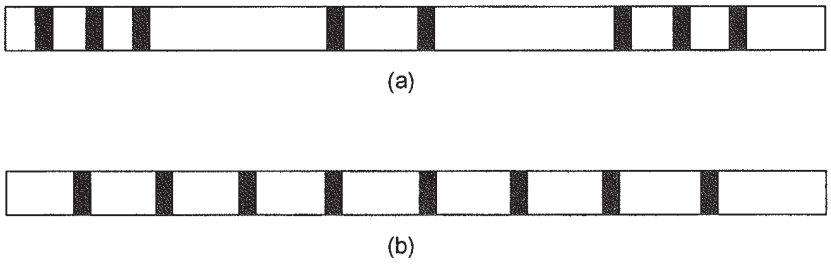


FIGURE 31.7 Position of resynchronization markers in the bitstream for (a) H.263 (Version 1) encoder with GOB headers and for (b) an MPEG-4 encoder with video packets.

Note that in addition to inserting the resynchronization markers at the beginning of each video packet, the encoder also needs to remove all data dependencies that exist between the data belonging to two different video packets within the same image. This is required so that even if one of the video packets in the current image is corrupted due to errors, the other packets can be decoded and utilized by the decoder. In order to remove these data dependencies, the encoder inserts two additional fields in addition to the resynchronization marker at the beginning of each video packet, as shown in Fig. 31.8. These are, (1) the absolute macroblock number of the first macroblock in the video packet, *Mb. No.*, (which indicates the spatial location of the macroblock in the current image), (2) the quantization parameter, QP, which denotes the initial quantization parameter used to quantize the DCT coefficients in the video packet. The encoder also modifies the predictive encoding method used for coding the motion vectors such that there are no predictions across the video packet boundaries. Also shown in Fig. 31.8 is a third field, labeled *HEC*. Its use is discussed in a later section.



FIGURE 31.8 An MPEG-4 video packet.

31.4.2 Data Partitioning

Data partitioning in MPEG-4 provides enhanced error localization and error concealment capabilities. The data partitioning mode partitions the data within a video packet into a motion part and a texture part (DCT coefficients) separated by a unique Motion Marker (MM), as shown in Fig. 31.9. All the syntactic elements of the video packet that have motion-related information are placed in the motion partition and all the remaining syntactic elements that relate to the DCT data are placed in the texture partition. If the texture information is lost, data partitioning enables the salvation of motion information, which can then be used to conceal the errors in a more effective manner.



FIGURE 31.9 A data partitioned MPEG-4 video packet.

The motion marker is computed from the motion VLC tables using a search program such that it is Hamming distance 1 from any possible valid combination of the motion VLC tables [9]. The motion marker is uniquely decodable from the motion VLC tables, and it indicates to the decoder the end of the motion information and the beginning of the DCT information. The number of macroblocks in the video packet is implicitly known after encountering the motion marker. Note that the motion marker is only computed once based on the VLC tables and is fixed in the standard. Based on the VLC tables in MEPG-4, the motion marker is a 17-bit word whose value is **1 1111 0000 0000 0001.**

31.4.3 Reversible Variable Length Codes (RVLCs)

As was shown in Fig. 31.5, if the decoder detects an error during the decoding of VLC codewords, it loses synchronization and hence typically has to discard all the data up to the next resynchronization point. RVLCs are designed such that they can be instantaneously decoded both in the forward and the backward direction. When the decoder detects an error while decoding the bitstream in the forward direction, it jumps to the next resynchronization marker and decodes the bitstream in the backward direction until it encounters an error. Based on the two error locations, the decoder can recover some of the data that would have otherwise been discarded. This is shown in Fig. 31.10, which shows only the texture part of the video packet—only data in the shaded area is discarded. Note that if RVLCs were not used, all the data in the texture part of the video packet would have to be discarded. RVLCs thus enable the decoder to better isolate the error location in the bitstream.

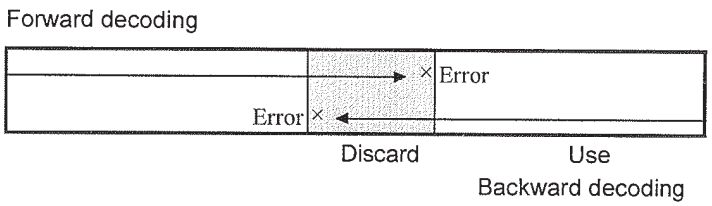


FIGURE 31.10 Use of reversible variable length codes.

Figure 31.11 shows the comparison of performance of resynchronization, data partitioning, and RVLC techniques for 24 Kb/s QCIF video data. The experiments involved transmission of three video sequences, each of duration 10s, over a bursty channel simulated by a 2-state Gilbert model [6]. The burst duration on the channel is 1 ms and the burst occurrence probability is 10^{-2}. Figure 31.11, which plots the average peak signal-to-noise ratios of the received video frames, shows that data partitioning and RVLC provide improved performance when compared to using only resynchronization markers.

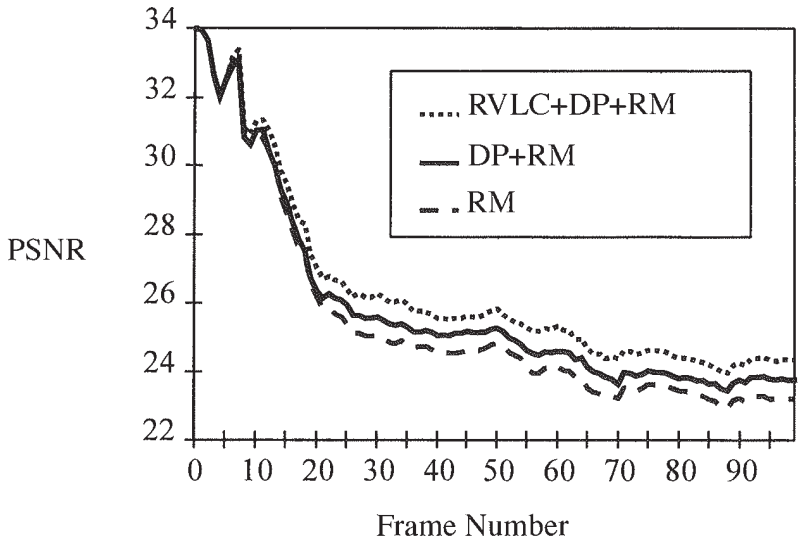


FIGURE 31.11 Performance comparison of resynchronization, data partitioning, and RVLC over a bursty channel simulated by a 2-state Gilbert model. Burst durations are 1ms long and the probability of occurrence of a burst is 10^{-2}. Legend: RM—resynchronization marker; DP—data partitioning; RVLC—reversible variable length codes.

31.4.4 Header Extension Code (HEC)

Some of the most important information that the decoder needs in order to decode the video bitstream is in the video frame header data. This data includes information about the spatial dimensions of the video data, the time stamps associated with the decoding and the presentation of this video data, and the type of the current frame (INTER/INTRA). If some of this information is corrupted due to channel errors, the decoder has no other recourse but to discard all the information belonging to the current video frame. In order to reduce the sensitivity of this data, a technique called Header Extension Code (HEC) was introduced into the MPEG-4 standard. In each video packet, a 1-bit field called HEC is present. The location of HEC in the video packet is shown in Fig. 31.8. For each video packet, when HEC is set, the important header information that describes the video frame is repeated in the bits following the HEC. This information can be used to verify and correct the header information of the video frame. The use of HEC significantly reduces the number of discarded video frames and helps achieve a higher overall decoded video quality.

31.4.5 Adaptive Intra Refresh (AIR)

Whenever an INTRA macroblock is received, it basically stops the temporal propagation of errors at its corresponding spatial location. The procedure of forcefully encoding some macroblocks in a frame in INTRA mode to flush out possible errors is called INTRA refreshing. INTRA refresh is very effective in stopping the propagation of errors, but it comes at the cost of a large overhead. Coding a macroblock in INTRA mode typically requires many more bits when compared to coding the macroblock in INTER mode. Hence, the INTRA refresh technique has to be used judiciously.

For areas with low motion, simple error concealment by just copying the previous frame's macroblocks works quite effectively. For macroblocks with high motion, error concealment becomes very difficult. Since the high motion areas are perceptually the most significant, any persistent error in the high motion area becomes very noticeable. The AIR technique of MPEG-4 makes use of the above facts and INTRA refreshes the motion areas more frequently, thereby allowing the corrupted high motion areas to recover quickly from errors.

Depending on the bitrate, the AIR approach only encodes a fixed and predetermined number of macroblocks in a frame in INTRA mode (the exact number is not standardized by MPEG-4). This fixed number might not be enough to cover all the macroblocks in the motion area; hence, the AIR technique keeps track of the macroblocks that have been refreshed (using a "refresh map") and in subsequent frames refreshes any macroblocks in the motion areas that might have been left out.

31.5 H.263 Error Resilience Tools

In this section, we discuss four error resilience techniques which are part of the H.263 standard—*slice structure mode* and *independent segment decoding,* which are forward error resilience features, and *error tracking* and *reference picture selection,* which are interactive error resilience techniques. Error tracking was introduced in H.263 (Version 1) as an appendix, whereas the remaining three techniques were introduced in H.263 (Version 2) as annexes.

31.5.1 Slice Structure Mode (Annex K)

The slice structured mode of H.263 is similar to the video packet approach of MPEG-4 with a slice denoting a video packet. The basic functionality of a slice is the same as that of a video packet—providing periodic resynchronization points throughout the bistream. The structure of a slice is shown in Fig. 31.12. Like an MPEG-4 video packet, the slice consists of a header followed by the macroblock data. The SSC is the slice start code and is identical to the resynchronization marker of MPEG-4. The MBA field, which denotes the starting macroblock number in the slice, and the SQUANT field, which is the quantizer scale coded nonpredictively, allow for the slice to be coded independently.

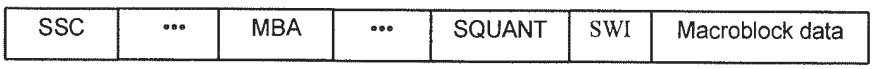


FIGURE 31.12 Structure of a slice in H.263/Annex K.

The slice structured mode also contains two submodes which can be used to provide additional functionality. The submodes are

- Rectangular slice submode (RSS)—This allows for rectangular shaped slices. The rectangular region contained in the slice is specified by SWI+1 (See Fig. 31.12 for the location of the SWI field in the slice header), which gives the width of the rectangular region, and MBA, which specifies the upper left macroblock of the slice. Note that the height of the rectangular region gets specified by the number of macroblocks contained in the slice. This mode can be used, for example, to subdivide images into rectangular regions of interest for region-based coding.
- Arbitrary slice submode (ASO)—The default order of transmission of slices is such that the MBA field is strictly increasing from one slice to the next transmitted slice. When ASO is used, the slices may appear in any order within the bitstream. This mode is useful when the wireless network supports prioritization of slices which might result in out-of-order arrival of video slices at the decoder.

31.5.2 Independent Segment Decoding (ISD) (Annex R)

Even though the slice structured mode eliminates decoding dependency between neighboring slices, errors in slices can spatially propagate to neighboring slices in subsequent frames due to motion compensation. This happens because motion vectors in a slice can point to macroblocks of neighboring slices in the reference picture. Independent segment decoding eliminates this from happening by restricting the motion vectors within a predefined segment of the picture from pointing to other segments in the picture, thereby helping to contain the error to be within the erroneous segment. This improvement in the localization of errors, however, comes at a cost of a loss of coding efficiency. Because of this restriction on the motion vectors, the motion compensation is not as effective, and the residual error images use more bits.

For ease of implementation, the ISD mode puts restrictions on segment shapes and on the changes of segment shapes from picture to picture. The ISD mode cannot be used with the slice structured mode (Annex K) unless the rectangular slice submode of Annex K is active. This prevents the need for treating awkward shapes of slices that can otherwise arise when Annex K is not used with rectangular slice submode. The segment shapes are not allowed to change from picture to picture unless an INTRA frame is being coded.

31.5.3 Error Tracking (Appendix I)

The error tracking approach is an INTRA refresh technique but uses decoder feedback of errors to decide which macroblocks in the current image to code in INTRA mode to prevent the propagation of these errors. When there are no errors on the channel, normal coding (which usually results in the bit-efficient INTER mode being selected most of the time) is used. The use of decoder feedback allows the system to adapt to varying channel conditions and minimizes the use of forced INTRA updates to situations when there are channel errors.

Because of the time delay involved in the decoder feedback, the encoder has to track the propagation of an error from its original occurrence to the current frame to decide which macroblocks should be INTRA coded in the current frame. A low complexity algorithm was proposed in Appendix I of H.263 to track the propagation of errors. However, it should be noted that the use of this technique is not mandated by H.263. Also, H.263 itself does not standardize the mechanism by which the decoder feedback of error can be sent. Typically, H.245 control messages are used to signal the decoder feedback for error tracking purposes.

31.5.4 Reference Picture Selection (Annex N)

The Reference Picture Selection (RPS) mode of H.263 also relies on decoder feedback to efficiently stop the propagation of errors. The back channel used in RPS mode can be a separate logical channel (e.g., by using H.245), or if two-way communication is taking place, the back channel messages can be sent multiplexed with the encoded video data. In the presence of errors, the RPS mode allows the encoder to be instructed to select one of the several previously correctly received and decoded frames as the reference picture for motion compensation of the current frame being encoded. This effectively stops the propagation of error. Note that the use of RPS requires the use of multiple frame buffers at both the encoder and the decoder to store previously decoded frames. Hence, the improvement in performance in the RPS mode has come at the cost of increased memory requirements.

31.6 Discussion

In this chapter we presented a broad overview of the various techniques that enable wireless video transmission. Due to the enormous amount of bandwidth required, video data is typically compressed before being transmitted, but the errors introduced by the wireless channels have a severe impact on the compressed video information. Hence, special techniques need to be employed to enable robust video transmission. International standards play a very important role in communications applications. The two current standards that are most relevant to video applications are ISO MPEG-4 and ITU H.263. In this chapter, we detailed these two standards and explained the error resilient tools that are part of these standards to enable robust video communication over wireless channels. A tutorial overview of these tools has been presented and the performance of these tools has been described.

There are, however, a number of other methods that further improve the performance of a wireless video codec that the standards do not specify. If the encoder and decoder are aware of the limitations imposed by the communication channel, they can further improve the video quality by using these methods. These methods include encoding techniques such as rate control to optimize the allocation of the effective channel bit rate between various parts of video to be transmitted and intelligent decisions on when and where to place INTRA refresh macroblocks to limit the error propagation. Decoding methods such as superior error concealment strategies that further conceal the effects of erroneous macroblocks by estimating them from correctly decoded macroblocks in the spatiotemporal neighborhood can also significantly improve the effective video quality.

This chapter has mainly focused on the error resilience aspects of the video layer. There are a number of error detection and correction strategies, such as Forward Error Correction (FEC), that can further improve the reliability of the transmitted video data. These FEC codes are typically provided in the systems layer and the underlying network layer. If the video transmission system has the ability to monitor the dynamic error characteristics of the communication channel, joint source-channel coding techniques can also be effectively employed. These techniques enable the wireless communication system to perform optimal trade-offs in allocating the available bits between the source coder (video) and the channel coder (FEC) to achieve superior performance.

Current video compression standards also support *layered* coding methods. In this approach, the compressed video information can be separated into multiple layers. The *base* layer, when decoded, provides a certain degree of video quality and the *enhancement* layer, when received and decoded, then adds to the base layer to further improve the video quality. In wireless channels, these base and enhancement layers give a natural method of partitioning the video data into more important and less important layers. The base layer can be

protected by a stronger level of error protection (higher overhead channel coder) and the enhancement layer by a lesser strength coder. Using this Unequal Error Protection (UEP) scheme, the communication system is assured of a certain degree of performance most of the time through the base layer, and when the channel is not as error prone and the decoder receives the enhancement layer, this scheme provides improved quality.

Given all these advances in video coding technology, coupled with the technological advances in processor technology, memory devices, and communication systems, wireless video communications is fast becoming a very compelling application. With the advent of higher bandwidth third generation wireless communication systems, it will be possible to transmit compressed video in many wireless applications, including mobile videophones, videoconferencing systems, PDAs, security and surveillance applications, mobile Internet terminals, and other multimedia devices.

Defining Terms

Automatic Repeat reQuest (ARQ): An error control system in which notification of erroneously received messages is sent to the transmitter which then simply retransmits the message. The use of ARQ requires a feedback channel and the receiver must perform error detection on received messages. Redundancy is added to the message before transmission to enable error detection at the receiver.

Block motion compensation (BMC): Motion compensated prediction that is done on a block basis; that is, blocks of pixels are assumed to be displaced spatially in a uniform manner from one frame to another.

Forward Error Correction: Introduction of redundancy in data to allow for correction of errors without retransmission.

Luminance and chrominance: Luminance is the brightness information in a video image, whereas chrominance is the corresponding color information.

Motion vectors: Specifies the spatial displacement of a block of pixels from one frame to another.

QCIF: Quarter Common Intermediate Format (QCIF) is a standard picture format that defines the image dimensions to be 176 $\times$ 144 (pixels per line $\times$ lines per picture) for luminance and 88 $\times$ 72 for chrominance.

References

[1] International Telecommunications Union—Telecommunications Standardization Sector, Recommendation H.223: Multiplexing protocol for low bitrate multimedia communications, Geneva, 1996.

[2] International Telecommunications Union—Telecommunications Standardization Sector, Recommendation H.263: Video coding for low bitrate communication, Geneva, 1996.

[3] International Telecommunications Union—Telecommunications Standardization Sector, Draft Recommendation H.263 (Version 2): video coding for low bitrate communication, Geneva, 1998.

[4] International Telecommunications Union—Telecommunications Standardization Sector, Recommendation H.324: terminal for low bit rate multimedia communications, Geneva, 1996.

[5] Lin, S. and Costello, D.J., Jr., *Error Control Coding: Fundamentals and Applications,* Prentice-Hall, Englewood Cliffs, NJ, 1983.

[6] Miki, T., et al., Revised error pattern generation programs for core experiments on error resilience, *ISO/IEC JTC1/SC29/WG11 MPEG96/1492,* Maceio, Brazil, Nov. 1996.

[7] International Organization for Standardization, Committee draft of Tokyo (N2202): information technology—coding of audio-visual objects: visual, *ISO/IEC 14496-2,* Mar. 1998.

[8] Sklar, B., Rayleigh fading channels in mobile digital communication systems, Pt. I: Characterization, *IEEE Commun. Mag.,* 35, 90–100, 1997.

[9] Talluri, R., et al., Error concealment by data partitioning, to appear in *Signal Processing: Image Commun.,* 1998.

[10] Wang, Y. and Zhu, Q., Error control and concealment for video communication: a review, *IEEE Trans. Circuits Syst. Video Technol.,* 86(5), 974–997, May 1998.

Further Information

A broader overview of wireless video can be found in the special issue of *IEEE Communications Magazine,* June 1998. Wang and Zhu [10] provide an exhaustive review of error concealment techniques for video communications. More details on MPEG-4 and ongoing Version 2 activities in MPEG-4 can be found on the web page `http://drogo.cselt.it/mpeg/standards/mpeg-4/mpeg-4.htm`. H.263 (Version 2) activities are tracked on the web page `http://www.ece.ubc.ca/spmg/research/motion/h263plus/`. Most of the ITU-T recommendations can be obtained from the web site `http://www.itu.org`. The special issue of *IEEE Communications Magazine,* December 1996, includes articles on H.324 and H.263.

Current research relevant to wireless video communications is reported in a number of journals including *IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Image Processing, IEEE Transactions on Vehicular Technology, Signal Processing: Image Communication.* The *IEEE Communications Magazine* regularly reports review articles relevant to wireless video communications. Conferences of interest include the IEEE International Conference on Image Processing (ICIP), IEEE Vehicular Technology Conference (VTC), and IEEE International Conference on Communications (ICC).

32

Wireless LANs

32.1 Introduction 32-1
32.2 Physical Layer Design 32-2
32.3 MAC Layer Protocols 32-4
Reservation-TDMA (R-TDMA) • Distributed Foundation Wireless MAC (DFWMAC) • Randomly Addressed Polling (RAP)
32.4 Network Layer Issues 32-8
Alternative View of Mobile Networks • A Proposed Architecture • Networking Issues
32.5 Transport Layer Design 32-11
32.6 Conclusions 32-13
Defining Terms 32-13
References 32-13
Further Information 32-14

Suresh Singh
Oregon State University

32.1 Introduction

A proliferation of high-performance portable computers combined with end-user need for communication is fueling a dramatic growth in wireless **local area network** (LAN) technology. Users expect to have the ability to operate their portable computer globally while remaining connected to communications networks and service providers. Wireless LANs and cellular networks, connected to high-speed networks, are being developed to provide this functionality.

Before delving deeper into issues relating to the design of wireless LANs, it is instructive to consider some scenarios of user mobility.

1. A simple model of user mobility is one where a computer is physically moved while retaining network connectivity at either end. For example, a move from one room to another as in a hospital where the computer is a hand-held device displaying patient charts and the nurse using the computer moves between wards or floors while accessing patient information.
2. Another model situation is where a group of people (at a conference, for instance) set up an ad-hoc LAN to share information as in Fig. 32.1.
3. A more complex model is one where several computers in constant communication are in motion and continue to be networked. For example, consider the problem of having robots in space collaborating to retrieve a satellite.

0-8493-8597-0/99/$0.00+$.50

© 1999 by CRC Press LLC

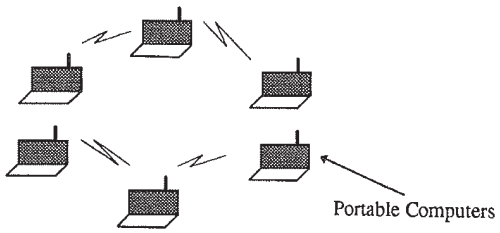


FIGURE 32.1 Ad-hoc wireless LAN.

A great deal of research has focused on dealing with physical layer and **medium access control (MAC)** layer protocols. In this chapter we first summarize standardization efforts in these areas. The remainder of the chapter is then devoted to a discussion of networking issues involved in wireless LAN design. Some of the issues discussed include routing in wireless LANs (i.e., how does data find its destination when the destination is mobile?) and the problem of providing service guarantees to end users (e.g., error-free data transmission or bounded delay and bounded bandwidth service, etc.).

32.2 Physical Layer Design

Two media are used for transmission over wireless LANs, infrared and radio frequency. RF LANs are typically implemented in the industrial, scientific, and medical (ISM) frequency bands 902–928 MHz, 2400–2483.5 MHz and 5725–5850 MHz. These frequencies do not require a license allowing the LAN product to be portable, i.e., a LAN can be moved without having to worry about licensing.

IR and RF technologies have different design constraints. IR receiver design is simple (and thus inexpensive) in comparison to RF receiver design because IR receivers only detect the amplitude of the signal not the frequency or phase. Thus, a minimal of filtering is required to reject interference. Unfortunately, however, IR shares the electromagnetic spectrum with the sun and incandescent or fluorescent light. These sources of modulated infrared energy reduce the signal-to-noise ratio of IR signals and, if present in extreme intensity, can make the IR LANs inoperable. There are two approaches to building IR LANs.

1. The transmitted signal can be focused and aimed. In this case the IR system can be used outdoors and has an area of coverage of a few kilometers.
2. The transmitted signal can be bounced off the ceiling or radiated omnidirectionally. In either case, the range of the IR source is 10–20 m (i.e., the size of one medium-sized room).

RF systems face harsher design constraints in comparison to IR systems for several reasons. The increased demand for RF products has resulted in tight regulatory constraints on the allocation and use of allocated bands. In the U.S., for example, it is necessary to implement spectrum spreading for operation in the ISM bands. Another design constraint is the requirement to confine the emitted spectrum to a band, necessitating amplification at higher carrier frequencies, frequency conversion using precision local oscillators, and selective components. RF systems must also cope with environmental noise that is either naturally occurring, for example, atmospheric noise or man made, for example, microwave ovens, copiers, laser printers, or other heavy electrical machinery. RF LANs operating in the ISM frequency ranges also suffer interference from amateur radio operators.

Operating LANs indoors introduces additional problems caused by multipath propagation, Rayleigh fading, and absorption. Many materials used in building construction are

opaque to IR radiation resulting in incomplete coverage within rooms (the coverage depends on obstacles within the room that block IR) and almost no coverage outside closed rooms. Some materials, such as white plasterboard, can also cause reflection of IR signals. RF is relatively immune to absorption and reflection problems. Multipath propagation affects both IR and RF signals. The technique to alleviate the effects of multipath propagation in both types of systems is the same use of aimed (directional) systems for transmission enabling the receiver to reject signals based on their angle of incidence. Another technique that may be used in RF systems is to use multiple antennas. The phase difference between different paths can be used to discriminate between them.

Rayleigh fading is a problem in RF systems. Recall that Rayleigh fading occurs when the difference in path length of the same signal arriving along different paths is a multiple of half a wavelength. This causes the signal to be almost completely canceled out at the receiver. Because the wavelengths used in IR are so small, the effect of Rayleigh fading is not noticeable in those systems. RF systems, on the other hand, use wavelengths of the order of the dimension of a laptop. Thus, moving the computer a small distance could increase/decrease the fade significantly.

Spread spectrum transmission technology is used for RF-based LANs and it comes in two varieties: direct-sequence spread spectrum (DSSS) and frequency-hopping spread spectrum (FHSS). In a FHSS system, the available band is split into several channels. The transmitter transmits on one channel for a fixed time and then hops to another channel. The receiver is synchronized with the transmitter and hops in the same sequence; see Fig. 32.2(a). In DSSS systems, a random binary string is used to modulate the transmitted signal. The relative rate between this sequence and user data is typically between 10 and 100; see Fig. 32.2(b).

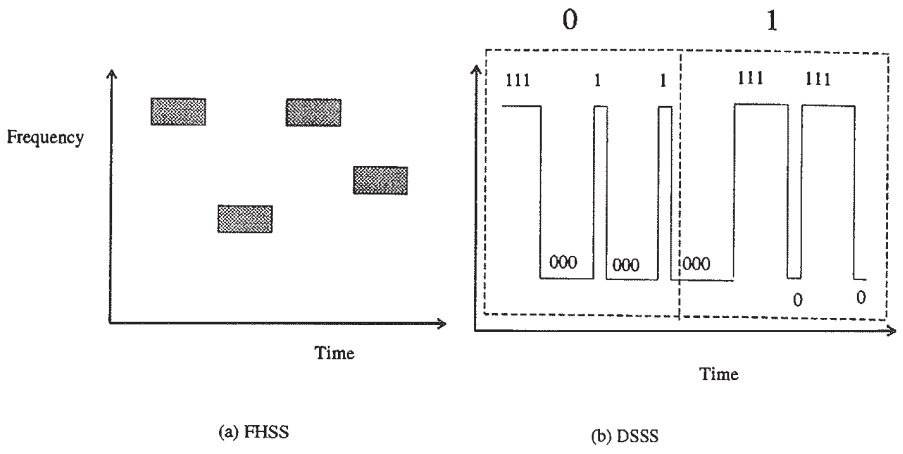


FIGURE 32.2 Spread spectrum.

The key requirements of any transmission technology is its robustness to noise. In this respect DSSS and FHSS show some differences. There are two possible sources of interference for wireless LANs: the presence of other wireless LANs in the same geographical area (i.e., in the same building, etc.) and interference due to other users of the ISM frequencies. In the latter case, FHSS systems have a greater ability to avoid interference because the hopping sequence could be designed to prevent potential interference. DSSS systems, on the other hand, do exhibit an ability to recover from interference because of the use of the spreading factor [Fig. 32.2(b)].

It is likely that in many situations several wireless LANs may be collocated. Since all wireless LANs use the same ISM frequencies, there is a potential for a great deal of interference. To avoid interference in FHSS systems, it is necessary to ensure that the hopping sequences are orthogonal. To avoid interference in DSSS systems, on the other hand, it is necessary to allocate different channels to each wireless LAN. The ability to avoid interference in DSSS systems is, thus, more limited in comparison to FHSS systems because FHSS systems use very narrow subchannels (1 MHz) in comparison to DSSS systems that use wider subchannels (for example, 25 MHz), thus, limiting the number of wireless LANs that can be collocated. A summary of design issues can be found in [1].

32.3 MAC Layer Protocols

MAC protocol design for wireless LANs poses new challenges because of the in-building operating environment for these systems. Unlike wired LANs (such as the ethernet or token ring), wireless LANs operate in strong multipath fading channels where channel characteristics can change in very short distances resulting in unreliable communication and unfair channel access due to capture. Another feature of the wireless LAN environment is that carrier sensing takes a long time in comparison to wired LANs; it typically takes between 30 and 50 μs (see [4]), which is a significant portion of the packet transmission time. This results in inefficiencies if the CSMA family of protocols is used without any modifications.

Other differences arise because of the mobility of users in wireless LAN environments. To provide a building (or any other region) with wireless LAN coverage, the region to be covered is divided into cells as shown in Fig. 32.3. Each cell is one wireless LAN, and adjacent cells use different frequencies to minimize interference. Within each cell there is an access point called a **mobile support station** (MSS) or base station that is connected to some wired network. The mobile users are called **mobile hosts (MH)**. The MSS performs the functions of channel allocation and providing connectivity to existing wired networks; see Fig. 32.4. Two problems arise in this type of an architecture that are not present in wired LANs.

1. The number of nodes within a cell changes dynamically as users move between cells. How can the channel access protocol dynamically adapt to such changes efficiently?
2. When a user moves between cells, the user has to make its presence known to the other nodes in the cell. How can this be done without using up too much bandwidth? The protocol used to solve this problem is called a handoff protocol and works along the following lines: A switching station (or the MSS nodes working together, in concert) collects signal strength information for each mobile host within each cell. Note that if a mobile host is near a cell boundary, the MSS node in its current cell as well as in the neighboring cell can hear its transmissions and determine signal strengths. If the mobile host is currently under the coverage of MSS M1 but its signal strength at MSS M2 becomes larger, the switching station initiates a handoff whereby the MH is considered as part of M2's cell (or network).

The mode of communication in wireless LANs can be broken in two: communication from the mobile to the MSS (called *uplink* communication) and communication in the reverse direction (called *downlink* communication). It is estimated that downlink communication accounts for about 70–80% of the total consumed bandwidth. This is easy to see because most of the time users request files or data in other forms (image data, etc.) that consume

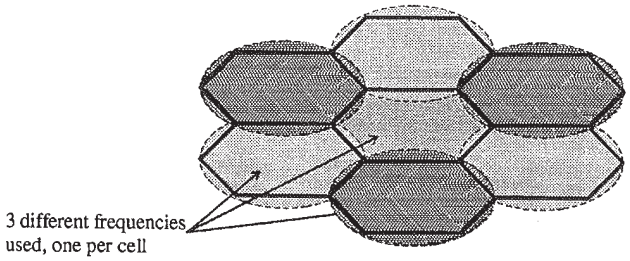


FIGURE 32.3 Cellular structure for wireless LANs (note frequency reuse).

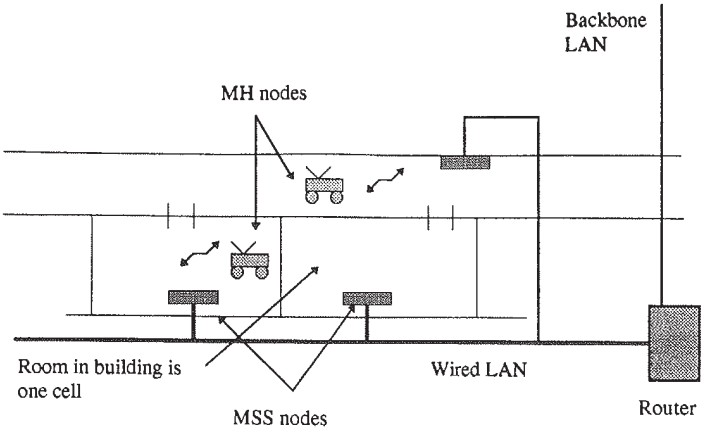


FIGURE 32.4 In-building LAN (made up of several wireless LANs).

much more transmission bandwidth than the requests themselves. In order to make efficient use of bandwidth (and, in addition, guarantee service requirements for real-time data), most researchers have proposed that the downlink channel be controlled entirely by the MSS nodes. These nodes allocate the channel to different mobile users based on their current requirements using a protocol such as **time division multiple access** (TDMA). What about uplink traffic? This is a more complicated problem because the set of users within a cell is dynamic, thus making it infeasible to have a static channel allocation for the uplink. This problem is the main focus of MAC protocol design.

What are some of the design requirements of an appropriate MAC protocol? The IEEE 802.11 recommended standard for wireless LANs has identified almost 20 such requirements, some of which are discussed here (the reader is referred to [3], for further details). Clearly any protocol must maximize throughput while minimizing delays and providing fair access to all users. In addition to these requirements, however, mobility introduces several new requirements.

1. The MAC protocol must be independent of the underlying physical layer transmission technology adopted (be it DSSS, FHSS or IR).
2. The maximum number of users can be as high as a few hundred in a wireless LAN. The MAC protocol must be able to handle many users without exhibiting catastrophic degradation of service.
3. The MAC protocols must provide secure transmissions because the wireless medium is easy to tap.
4. The MAC protocol needs to work correctly in the presence of collocated networks.
5. It must have the ability to support ad-hoc networking (as in Fig. 32.1).

6. Other requirements include the need to support priority traffic, preservation of packet order, and an ability to support multicast.

Several contention-based protocols currently exist that could be adapted for use in wireless LANs. The protocols currently being looked by IEEE 802.11 include protocols based on **carrier sense multiple access (CSMA)**, polling, and TDMA. Protocols based on **code division multiple access** (CDMA) and **frequency division multiple access** (FDMA) are not considered because the processing gains obtained using these protocols are minimal while, simultaneously, resulting in a loss of flexibility for wireless LANs.

It is important to highlight an important difference between networking requirements of ad-hoc networks (as in Fig. 32.1) and networks based on cellular structure. In cellular networks, all communication occurs between the mobile hosts and the MSS (or base station) within that cell. Thus, the MSS can allocate channel bandwidth according to requirements of different nodes, i.e., we can use centralized channel scheduling for efficient use of bandwidth. In ad-hoc networks there is no such central scheduler available. Thus, any multiaccess protocol will be contention based with little explicit scheduling. In the remainder of this section we focus on protocols for cell-based wireless LANs only.

All multiaccess protocols for cell-based wireless LANs have a similar structure; see [3].

1. The MSS announces (explicitly or implicitly) that nodes with data to send may contend for the channel.
2. Nodes interested in sending data contend for the channel using protocols such as CSMA.
3. The MSS allocates the channel to successful nodes.
4. Nodes transmit packets (contention-free transmission).
5. MSS sends an explicit acknowledgment (ACK) for packets received.

Based on this model we present three MAC protocols.

32.3.1 Reservation-TDMA (R-TDMA)

This approach is a combination of TDMA and some contention protocol (see PRMA in [7]). The MSS divides the channel into slots (as in TDMA), which are grouped into frames. When a node wants to transmit it needs to reserve a slot that it can use in every consecutive frame as long as it has data to transmit. When it has completed transmission, other nodes with data to transmit may contend for that free slot. There are four steps to the functioning of this protocol.

a. At the end of each frame the MSS transmits a feedback packet that informs nodes of the current reservation of slots (and also which slots are free). This corresponds to steps 1 and 3 from the preceding list.
b. During a frame, all nodes wishing to acquire a slot transmit with a probability ρ during a free slot. If a node is successful it is so informed by the next feedback packet. If more than one node transmits during a free slot, there is a collision and the nodes try again during the next frame. This corresponds to step 2.
c. A node with a reserved slot transmits data during its slot. This is the contention-free transmission (step 4).
d. The MSS sends ACKs for all data packets received correctly. This is step 5.

The R-TDMA protocol exhibits several nice properties. First and foremost, it makes very efficient use of the bandwidth, and average latency is half the frame size. Another big

benefit is the ability to implement power conserving measures in the portable computer. Since each node knows when to transmit (nodes transmit during their reserved slot only) it can move into a power-saving mode for a fixed amount of time, thus increasing battery life. This feature is generally not available in CSMA-based protocols. Furthermore, it is easy to implement priorities because of the centralized control of scheduling. One significant drawback of this protocol is that it is expensive to implement (see [2]).

32.3.2 Distributed Foundation Wireless MAC (DFWMAC)

The CSMA/CD protocol has been used with great success in the ethernet. Unfortunately, the same protocol is not very efficient in a wireless domain because of the problems associated with cell interference (i.e., interference from neighboring cells), the relatively large amount of time taken to sense the channel (see [6]) and the hidden terminal problem (see [12, 13]). The current proposal is based on a CSMA/collision avoidance (CA) protocol with a four-way handshake; see Fig. 32.5.

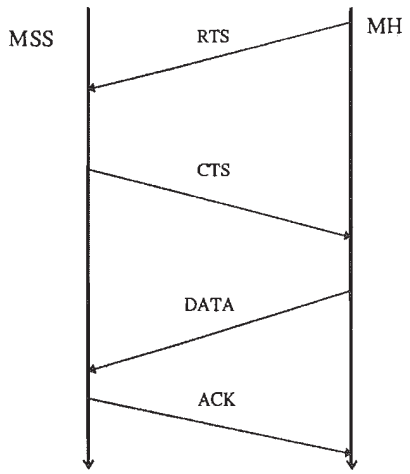


FIGURE 32.5 CSMA/CA and four-way handshaking protocol.

The basic operation of the protocol is simple. All MH nodes that have packets to transmit compete for the channel by sending ready to transmit (RTS) messages using nonpersistent CSMA. After a station succeeds in transmitting a RTS, the MSS sends a clear to transmit (CTS) to the MH. The MH transmits its data and then receives an ACK. The only possibility of collision that exists is in the RTS phase of the protocol and inefficiencies occur in the protocol, because of the RTS and CTS stages. Note that unlike R-TDMA it is harder to implement power saving functions. Furthermore, latency is dependent on system load making it harder to implement real-time guarantees. Priorities are also not implemented. On the positive side, the hardware for this protocol is very inexpensive.

32.3.3 Randomly Addressed Polling (RAP)

In this scheme, when a MSS is ready to collect uplink packets it transmits a READY message. At this point all nodes with packets to send attempt to grab the channel as follows.

a. Each MH with a packet to transmit generates a random number between 0 and P.

b. All active MH nodes simultaneously and orthogonally transmit their random numbers (using CDMA or FDMA). We assume that all of these numbers are received correctly by the MSS. Remember that more than one MH node may have selected the same random number.

c. Steps a and b are repeated L times.

d. At the end of L stages, the MSS determines a stage (say, k) where the total number of distinct random numbers was the largest. The MSS polls each distinct each random number in this stage in increasing order. All nodes that had generated the polled random number transmit packets to the MSS.

e. Since more than one node may have generated the same random number, collisions are possible. The MSS sends a ACK or NACK after each such transmission. Unsuccessful nodes try again during the next iteration of the protocol.

The protocol is discussed in detail in [4] and a modified protocol called GRAP (for group RAP) is discussed in [3]. The authors propose that GRAP can also be used in the contention stage (step 2) for TDMA- and CSMA-based protocols.

32.4 Network Layer Issues

An important goal of wireless LANs is to allow users to move about freely while still maintaining all of their connections (network resources permitting). This means that the network must route all packets destined for the mobile user to the MSS of its current cell in a transparent manner. Two issues need to be addressed in this context.

- How can users be addressed?
- How can active connections for these mobile users be maintained?

Ioanidis, Duchamp, and Maguire [8] propose a solution called the IPIP (IP-within-IP) protocol. Here each MH has a unique **internet protocol** (IP) address called its home address. To deliver a packet to a remote MH, the source MSS first broadcasts an address resolution protocol (ARP) request to all other MSS nodes to locate the MH. Eventually some MSS responds. The source MSS then encapsulates each packet from the source MH within another packet containing the IP address of the MSS in whose cell the MH is located. The destination MSS extracts the packet and delivers it to the MH. If the MH has moved away in the interim, the new MSS locates the new location of the MH and performs the same operation. This approach suffers from several problems as discussed in [11]. Specifically, the method is not scaleable to a network spanning areas larger than a campus for the following reasons.

1. IP addresses have a prefix identifying the campus subnetwork where the node lives; when the MH moves out of the campus, its IP address no longer represents this information.
2. The MSS nodes serve the function of routers in the mobile network and, therefore, have the responsibility of tracking all of the MH nodes globally causing a lot of overhead in terms of message passing and packet forwarding; see [5].

Teraoka and Tokoro [11], have proposed a much more flexible solution to the problem called virtual IP (VIP). Here every mobile host has a virtual IP address that is unchanging regardless of the location of the MH. In addition, hosts have physical network addresses

(traditional IP addresses) that may change as the host moves about. At the transport layer, the target node is always specified by its VIP address only. The address resolution from the VIP address to the current IP address takes place either at the network layer of the same machine or at a gateway. Both the host machines and the gateways maintain a cache of VIP to IP mappings with associated timestamps. This information is in the form of a table called *address mapping table* (AMT). Every MH has an associated *home gateway*. When a MH moves into a new subnetwork, it is assigned a new IP address. It sends this new IP address and its VIP address to its home gateway via a *VipConn* control message. All intermediate gateways that relay this message update their AMT tables as well. During this process of updating the AMT tables, all packets destined to the MH continue to be sent to the old location. These packets are returned to the sender, who then sends them to the home gateway of the MH. It is easy to see that this approach is easily scaleable to large networks, unlike the IPIP approach.

32.4.1 Alternative View of Mobile Networks

The approaches just described are based on the belief that mobile networks are merely an extension of wired networks. Other authors [10] disagree with this assumption because there are fundamental differences between the mobile domain and the fixed wired network domain. Two examples follow.

1. The available bandwidth at the wireless link is small; thus, end-to-end packet retransmission for transmission control protocol (TCP)-like protocols (implemented over datagram networks) is a bad idea. This leads to the conclusion that transmission within the mobile network must be connection oriented. Such a solution, using virtual circuits (VC), is proposed in [5].
2. The bandwidth available for a MH with open connections changes dynamically since the number of other users present in each cell varies randomly. This is a feature not present in fixed high-speed networks where, once a connection is set up, its bandwidth does not vary much. Since bandwidth changes are an artifact of mobility and are dynamic, it is necessary to deal with the consequences (e.g., buffer overflow, large delays, etc.) locally to both, i.e., shield fixed network hosts from the idiosyncrasies of mobility as well as to respond to changing bandwidth quickly (without having to rely on end-to-end control). Some other differences are discussed in [10].

32.4.2 A Proposed Architecture

Keeping these issues in mind, a more appropriate architecture has been proposed in Ghai and Singh [5], and Singh [10]. Mobile networks are considered to be different and separate from wired networks. Within a mobile network is a three-layer hierarchy; see Fig. 32.6. At the bottom layer are the MHs. At the next level are the MSS nodes (one per cell). Finally, several MSS nodes are controlled by a **supervisor host (SH)** node (there may be one SH node per small building). The SH nodes are responsible for flow control for all MH connections within their domain; they are also responsible for tracking MH nodes and forwarding packets as MH nodes roam. In addition, the SH nodes serve as a *gateway* to the wired networks. Thus, any connection setup from a MH to a fixed host is broken in two, one from the MH to the SH and another from the SH to the fixed host. The MSS nodes in this design are simply connection endpoints for MH nodes. Thus, they are simple devices that implement the MAC protocols and little else. Some of the benefits of this design are

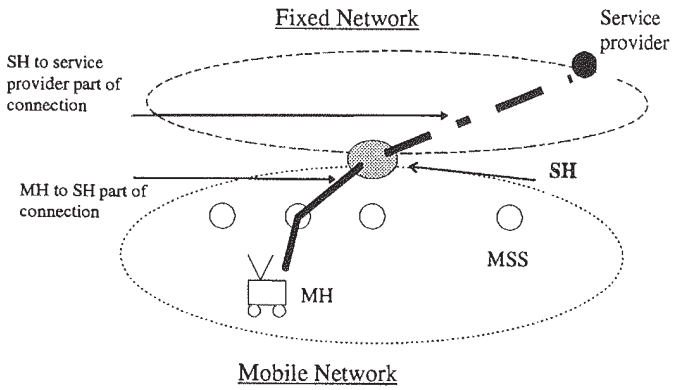


FIGURE 32.6 Proposed architecture for wireless networks.

as follows.

1. Because of the large coverage of the SH (i.e., a SH controls many cells) the MH remains in the domain of one SH much longer. This makes it easy to handle the consequences of dynamic bandwidth changes locally. For instance, when a MH moves into a crowded cell, the bandwidth available to it is reduced. If it had an open ftp connection, the SH simply buffers undelivered packets until they can be delivered. There is no need to inform the other endpoint of this connection of the reduced bandwidth.
2. When a MH node sets up a connection with a service provider in the fixed network, it negotiates some quality of service (QOS) parameters such as bandwidth, delay bounds, etc. When the MH roams into a crowded cell, these QOS parameters can no longer be met because the available bandwidth is smaller. If the traditional view is adopted (i.e., the mobile networks are extensions of fixed networks) then these QOS parameters will have to be renegotiated each time the bandwidth changes (due to roaming). This is a very expensive proposition because of the large number of control messages that will have to be exchanged. In the approach of Singh [10], the service provider will never know about the bandwidth changes since it deals only with the SH that is accessed via the wired network. The SH bears the responsibility of handling bandwidth changes by either buffering packets until the bandwidth available to the MH increases (as in the case of the ftp example) or it could discard a fraction of real-time packets (e.g., a voice connection) to ensure delivery of most of the packets within their deadlines. The SH could also instruct the MSS to allocate a larger amount of bandwidth to the MH when the number of buffered packets becomes large. Thus, the service provider in the fixed network is shielded from the mobility of the user.

32.4.3 Networking Issues

It is important for the network to provide connection-oriented service in the mobile environment (as opposed to connectionless service as in the internet) because bandwidth is at a premium in wireless networks, and it is, therefore, inadvisable to have end-to-end retransmission of packets (as in TCP). The proposed architecture is well suited to providing connection-oriented service by using VCs.

In the remainder of this section we look at how virtual circuits are used within the mobile network and how routing is performed for connections to mobile hosts. Every connection set up with one or more MH nodes as a connection endpoint is routed through the SH nodes and each connection is given a unique VC number. The SH node keeps track of all MH nodes that lie within its domain. When a packet needs to be delivered to a MH node, the SH first buffers the packet and then sends it to the MSS at the current location of the MH or to the predicted location if the MH is currently between cells. The MSS buffers all of these packets for the MH and transmits them to the MH if it is in its cell. The MSS discards packets after transmission or if the SH asks it to discard the packets. Packets are delivered in the correct order to the MH (without duplicates) by having the MH transmit the expected sequence number (for each VC) during the initial handshake (i.e., when the MH first enters the cell). The MH sends ACKs to the SH for packets received. The SH discards all packets that have been acknowledged. When a MH moves from the domain of SH1 into the domain of SH2 while having open connections, SH1 continues to forward packets to SH2 until either the connections are closed or until SH2 sets up its own connections with the other endpoints for each of MH's open connections (it also gives new identifiers to all these open connections). The detailed protocol is presented in [5].

The SH nodes are all connected over the fixed (wired) network. Therefore, it is necessary to route packets between SH nodes using the protocol provided over the fixed networks. The VIP protocol appears to be best suited to this purpose. Let us assume that every MH has a globally unique VIP address. The SHs have both a VIP as well as a fixed IP address. When a MH moves into the domain of a SH, the IP address affixed to this MH is the IP address of the SH. This ensures that all packets sent to the MH are routed through the correct SH node. The SH keeps a list of all VIP addresses of MH nodes within its domain and a list of open VCs for each MH. It uses this information to route the arriving packets along the appropriate VC to the MH.

32.5 Transport Layer Design

The transport layer provides services to higher layers (including the application layer), which include connectionless services like UDP or connection-oriented services like TCP. A wide variety of new services will be made available in the high-speed networks, such as continuous media service for real-time data applications such as voice and video. These services will provide bounds on delay and loss while guaranteeing some minimum bandwidth.

Recently variations of the TCP protocol have been proposed that work well in the wireless domain. These proposals are based on the traditional view that wireless networks are merely extensions of fixed networks. One such proposal is called I-TCP [2] for indirect TCP. The motivation behind this work stems from the following observation. In TCP the sender times out and begins retransmission after a timeout period of several hundred milliseconds. If the other endpoint of the connection is a mobile host, it is possible that the MH is disconnected for a period of several seconds (while it moves between cells and performs the initial greeting). This results in the TCP sender timing out and transmitting the same data several times over, causing the effective throughput of the connection to degrade rapidly. To alleviate this problem, the implementation of I-TCP separates a TCP connection into two pieces—one from the fixed host to another fixed host that is near the MH and another from this host to the MH (note the similarity of this approach with the approach in Fig. 32.6). The host closer to the MH is aware of mobility and has a larger timeout period. It serves as a type of gateway for the TCP connection because it sends ACKs back to the sender before receiving ACKs from the MH. The performance of I-TCP is far superior to traditional TCP for the mobile networks studied.

In the architecture proposed in Fig. 32.6, a TCP connection from a fixed host to a mobile host would terminate at the SH. The SH would set up another connection to the MH and would have the responsibility of transmitting all packets correctly. In a sense this is a similar idea to I-TCP except that in the wireless network VCs are used rather than datagrams. Therefore, the implementation of TCP service is made much easier.

A problem that is unique to the mobile domain occurs because of the unpredictable movement of MH nodes (i.e., a MH may roam between cells resulting in a large variation of available bandwidth in each cell). Consider the following example. Say nine MH nodes have opened 11-kb/s connections in a cell where the available bandwidth is 100 kb/s. Let us say that a tenth mobile host M10, also with an open 11-kb/s connection, wanders in. The total requested bandwidth is now 110 kb/s while the available bandwidth is only 100 kb/s. What is to be done? One approach would be to deny service to M10. However, this seems an unfair policy. A different approach is to penalize all connections equally so that each connection has 10-kb/s bandwidth allocated.

To reduce the bandwidth for each connection from 11 kb/s to 10 kb/s, two approaches may be adopted:

1. Throttle back the sender for each connection by sending control messages.
2. Discard 1-kb/s data for each connection at the SH. This approach is only feasible for applications that are tolerant of data loss (e.g., real-time video or audio).

The first approach encounters a high overhead in terms of control messages and requires the sender to be capable of changing the data rate dynamically. This may not always be possible; for instance, consider a teleconference consisting of several participants where each mobile participant is subject to dynamically changing bandwidth. In order to implement this approach, the data (video or audio or both) will have to be compressed at different ratios for each participant, and this compression ratio may have to be changed dynamically as each participant roams. This is clearly an unreasonable solution to the problem. The second approach requires the SH to discard 1-kb/s of data for each connection. The question is, how should this data be discarded? That is, should the 1 kb of discarded data be consecutive (or clustered) or uniformly spread out over the data stream every 1 s? The way in which the data is discarded has an effect on the final perception of the service by the mobile user. If the service is audio, for example, a random uniform loss is preferred to a clustered loss (where several consecutive words are lost). If the data is compressed video, the problem is even more serious because most random losses will cause the encoded stream to become unreadable resulting in almost a 100% loss of video at the user.

A solution to this problem is proposed in Seal and Singh [9], where a new sublayer is added to the transport layer called the *loss profile transport sublayer* (*LPTSL*). This layer determines how data is to be discarded based on special transport layer markers put by application calls at the sender and based on negotiated loss functions that are part of the QOS negotiations between the SH and service provider. Figure 32.7 illustrates the functioning of this layer at the service provider, the SH, and the MH. The original data stream is broken into *logical segments* that are separated by markers (or flags). When this stream arrives at the SH, the SH discards entire logical segments (in the case of compressed video, one logical segment may represent one frame) depending on the bandwidth available to the MH. The purpose of discarding entire logical segments is that discarding a part of such a segment of data makes the rest of the data within that segment useless—so we might as well discard the entire segment. Observe also that the flags (to identify logical segments) are inserted by the LPTSL via calls made by the application layer. Thus, the transport layer or the LPTSL does not need to know encoding details of the data stream. This scheme

is currently being implemented at the University of South Carolina by the author and his research group.

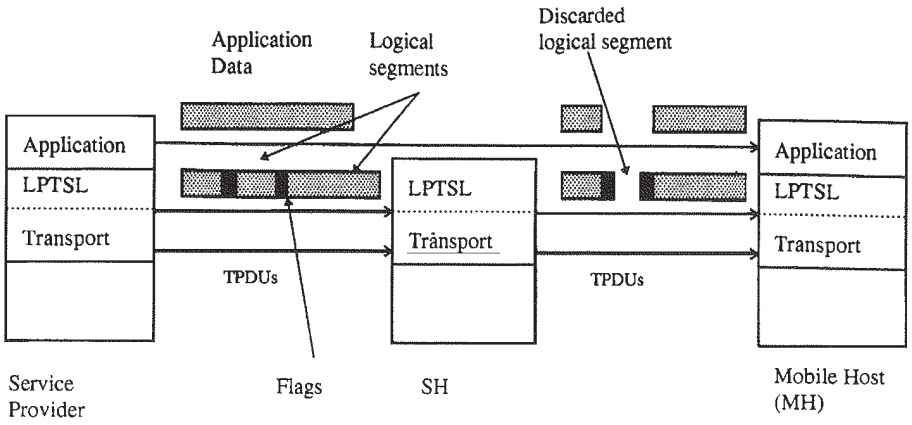


FIGURE 32.7 LPTSL, an approach to handle dynamic bandwidth variations.

32.6 Conclusions

The need for wireless LANs is driving rapid development in this area. The IEEE has proposed standards (802.11) for the physical layer and MAC layer protocols. A great deal of work, however, remains to be done at the network and transport layers. There does not appear to be a consensus regarding subnet design for wireless LANs. Our work has indicated a need for treating wireless LAN subnetworks as being fundamentally different from fixed networks, thus resulting in a different subnetwork and transport layer designs. Current efforts are underway to validate these claims.

Defining Terms

Carrier-sense multiple access (CSMA): Protocols such as those used over the ethernet.

Medium access control (MAC): Protocols arbitrate channel access between all nodes on a wireless LAN.

Mobile host (MH) nodes: The nodes of wireless LAN.

Supervisor host (SH): The node that takes care of flow-control and other protocol processing for all connections.

References

[1] Bantz, D.F. and Bauchot, F.J., Wireless LAN design alternatives. *IEEE Network*, 8(2), 43–53, 1994.

[2] Barke, A. and Badrinath, B.R., I-TCP: indirect TCP for mobile hosts. Tech. Rept. DCS-TR-314, Dept. Computer Science, Rutgers University, Piscataway, NJ, 1994.

[3] Chen, K.-C., Medium access control of wireless LANs for mobile computing. *IEEE Network*, 8(5), 50–63, 1994.

[4] Chen, K.-C. and Lee, C.H., RAP: a novel medium access control protocol for wireless data networks. *Proc. IEEE GLOBECOM'93*, IEEE Press, Piscataway, NJ, 08854. 1713–1717, 1993.

[5] Ghai, R. and Singh, S., An architecture and communication protocol for picocellular networks. *IEEE Personal Comm. Mag.*, 1(3), 36–46, 1994.

[6] Glisic, S.G., 1-Persistent carrier sense multiple access in radio channel with imperfect carrier sensing. *IEEE Trans. on Comm.*, 39(3), 458–464, 1991.

[7] Goodman, D.J., Cellular packet communications. *IEEE Trans. on Comm.*, 38(8), 1272–1280, 1990.

[8] Ioanidis, J., Duchamp, D., and Maguire, G.Q., IP-based protocols for mobile internetworking. *Proc. of ACM SIGCOMM'91*, ACM Press, New York, NY, 10036 (Sept.), 235–245, 1991.

[9] Seal, K. and Singh, S., Loss profiles: a quality of service measure in mobile computing. *J. Wireless Networks*, 2, 45–61, 1996.

[10] Singh, S., Quality of service guarantees in mobile computing. *J. of Computer Comm.*, 19, 359–371, 1996.

[11] Teraoka, F. and Tokoro, M., Host migration transparency in IP networks: the VIP approach. *Proc. of ACM SIGCOMM*, ACM Press, New York, NY, 10036 (Jan.), 45–65, 1993.

[12] Tobagi, F. and Kleinrock, L., Packet switching in radio channels: Part I carrier sense multiple access models and their throughput delay characteristic. *IEEE Trans. on Comm.*, 23(12), 1400–1416, 1975a.

[13] Tobagi, F. and Kleinrock, L., Packet switching in radio channels: Part II the hidden terminal problem in CSMA and busy-one solution. *IEEE Trans. on Comm.*, 23(12), 1417–1433, 1975b.

Further Information

A good introduction to physical layer issues is presented in Bantz [1] and MAC layer issues are discussed in Chen [3]. For a discussion of network and transport layer issues, see Singh [10] and Ghai and Singh [5].

33

Wireless Data

33.1 Introduction 33-1
33.2 Characteristics of Wireless Data Networks 33-2
Radio Propagation Characteristics
33.3 Market Issues 33-4
33.4 Modem Services Over Cellular Networks 33-4
33.5 Private Data Networks 33-6
ARDIS • MOBITEX • Ricochet Network
33.6 Cellular Data Networks and Services 33-10
Cellular Digital Packet Data (CDPD) • Digital Cellular Data Services
33.7 Other Developing Standards 33-14
TErrestrial Trunked Radio (TETRA)
33.8 Conclusions 33-15
References 33-15
Further Information 33-16

Allen H. Levesque
GTE Laboratories, Inc.

Kaveh Pahlavan
Worcester Polytechnic Institute

33.1 Introduction

Wireless data services and systems represent a steadily growing and increasingly important segment of the communications industry. While the wireless data industry is becoming increasingly diverse, one can identify two mainstreams that relate directly to users' requirement for data services. On one hand, there are requirements for relatively low-speed data services provided to mobile users over wide geographical areas, as provided by private mobile data networks and by data services implemented on common-carrier cellular telephone networks. On the other hand, there are requirements for high-speed data services in local areas, as provided by cordless private branch exchange (PBX) systems and wireless local area networks (LANs), as well as by the emerging personal communications services (PCS). Wireless LANs are treated in Chapter 32. In this chapter we mainly address wide-area wireless data systems, commonly called *mobile data systems*, and briefly touch upon data services to be incorporated into the emerging digital cellular systems.

Mobile data systems provide a wide variety of services for both business users and public safety organizations. Basic services supporting most businesses include electronic mail, enhanced paging, modem and facsimile transmission, remote access to host computers and office LANs, information broadcast services and, increasingly, Internet access. Public safety organizations, particularly law-enforcement agencies, are making increasing use of wireless data communications over traditional VHF and UHF radio dispatch networks, over commercial mobile data networks, and over public cellular telephone networks. In addition, there

0-8493-8597-0/99/$0.00+$.50

© 1999 by CRC Press LLC

are wireless services supporting vertical applications that are more or less tailored to the needs of specific companies or industries, such as transaction processing, computer-aided delivery dispatch, customer service, fleet management, and emergency medical services. Work currently in progress to develop the national Intelligent Transportation System (ITS) includes the definition of a wide array of new traveler services, many of which will be supported by standardized mobile data networks.

Much of the growth in use of wireless data services has been spurred by the rapid growth of the paging service industry and increasing customer demand for more advanced paging services, as well as the desire to increase work productivity by extending to the mobile environment the suite of digital communications services readily available in the office environment. There is also a desire to make more cost-efficient use of the mobile radio and cellular networks already in common use for mobile voice communications by incorporating efficient data transmission services into these networks. The services and networks that have evolved to date represent a variety of specialized solutions and, in general, they are not interoperable with each other. As the wireless data industry expands, there is an increasing demand for an array of attractively priced standardized services and equipment accessible to mobile users over wide geographic areas. Thus, we see the growth of nationwide privately operated service networks as well as new data services built upon the first and second generation cellular telephone networks. The implementation of PCS networks in the 2-GHz bands as well as the eventual implementation of third generation (3G) wireless networks will further extend this evolution.

In this chapter we describe the principal existing and evolving wireless data networks and the related standards activities now in progress. We begin with a discussion of the technical characteristics of wireless data networks.

33.2 Characteristics of Wireless Data Networks

From the perspective of the data user, the basic requirement for wireless data service is convenient, reliable, low-speed access to data services over a geographical area appropriate to the user's pattern of daily business operation. By low speed we mean data rates comparable to those provided by standard data modems operating over the public switched telephone network (PSTN). This form of service will support a wide variety of short-message applications, such as notice of electronic mail or voice mail, as well as short file transfers or even facsimile transmissions that are not overly lengthy. The user's requirements and expectations for these types of services are different in several ways from the requirements placed on voice communication over wireless networks. In a wireless voice service, the user usually understands the general characteristics and limitations of radio transmission and is tolerant of occasional *signal fades* and brief dropouts. An overall level of acceptable voice quality is what the user expects. In a data service, the user is instead concerned with the accuracy of delivered messages and data, the time-delay characteristics of the service network, the ability to maintain service while traveling about, and, of course, the cost of the service. All of these factors are dependent on the technical characteristics of wireless data networks, which we discuss next.

33.2.1 Radio Propagation Characteristics

The chief factor affecting the design and performance of wireless data networks is the nature of radio propagation over wide geographic areas. The most important mobile data systems operate in various land–mobile radio bands from roughly 100 to 200 MHz, the

specialized mobile radio (SMR) band around 800 MHz, and the cellular telephone bands at 824–894 MHz. In these frequency bands, radio transmission is characterized by distance-dependent field strength, as well as the well-known effects of *multipath fading*, signal shadowing, and signal blockage. The signal coverage provided by a radio transmitter, which in turn determines the area over which a mobile data receiving terminal can receive a usable signal, is governed primarily by the *power–distance relationship*, which gives signal power as a function of distance between transmitter and receiver. For the ideal case of single-path transmission in free space, the relationship between transmitted power P_t and received power P_r is given by

$$P_r/P_t = G_t G_r(\lambda/4\pi d)^2 \tag{33.1}$$

where G_t and G_r are the transmitter and receiver antenna gains, respectively, d is the distance between the transmitter and the receiver, and λ is the wavelength of the transmitted signal. In the mobile radio environment, the power-distance relationship is in general different from the free-space case just given. For propagation over an Earth plane at distances much greater than either the signal wavelength or the antenna heights, the relationship between P_t and P_r is given by

$$P_r/P_t = G_t G_r\left(h_1^2 h_2^2/d^4\right) \tag{33.2}$$

where h_1 and h_2 are the transmitting and receiving antenna heights. Note here that the received power decreases as the fourth power of the distance rather than the square of distance seen in the ideal free-space case. This relationship comes from a propagation model in which there is a single signal reflection with phase reversal at the Earth's surface, and the resulting received signal is the vector sum of the direct line-of-sight signal and the reflected signal. When user terminals are deployed in mobile situations, the received signal is generally characterized by rapid fading of the signal strength, caused by the vector summation of reflected signal components, the vector summation changing constantly as the mobile terminal moves from one place to another in the service area. Measurements made by many researchers show that when the fast fading is averaged out, the signal strength is described by a Rayleigh distribution having a log-normal mean. In general, the power-distance relationship for mobile radio systems is a more complicated relationship that depends on the nature of the terrain between transmitter and receiver.

Various propagation models are used in the mobile radio industry for network planning purposes, and a number of these models are described in [1]. Propagation models for mobile communications networks must take account of the terrain irregularities existing over the intended service area. Most of the models used in the industry have been developed from measurement data collected over various geographic areas. A very popular model is the *Longley–Rice model* [8, 14]. Many wireless networks are concentrated in urban areas. A widely used model for propagation prediction in urban areas is one usually referred to as the *Okumura–Hata model* [4, 9].

By using appropriate propagation prediction models, one can determine the range of signal coverage for a base station of given transmitted power. In a wireless data system, if one knows the level of received signal needed for satisfactory performance, the area of acceptable performance can, in turn, be determined. Cellular telephone networks utilize base stations that are typically spaced 1–5 mi apart, though in some mid-town areas, spacings of 1/2 mi or less are now being used. In packet-switched data networks, higher power transmitters are used, spaced about 5–15 mi apart.

An important additional factor that must be considered in planning a wireless data system is the in-building penetration of signals. Many applications for wireless data services involve the use of mobile data terminals inside buildings, for example, for trouble-shooting and

servicing computers on customers' premises. Another example is wireless communications into hospital buildings in support of emergency medical services. It is usually estimated that in-building signal penetration losses will be in the range of 15–30 dB. Thus, received signal strengths can be satisfactory in the outside areas around a building but totally unusable inside the building. This becomes an important issue when a service provider intends to support customers using mobile terminals inside buildings.

One important consequence of the rapid fading experienced on mobile channels is that errors tend to occur in bursts, causing the transmission to be very unreliable for short intervals of time. Another problem is signal dropouts that occur, for example, when a data call is handed over from one base station to another, or when the mobile user moves into a location where the signal is severely attenuated. Because of this, mobile data systems employ various error-correction and error-recovery techniques to insure accurate and reliable delivery of data messages.

33.3 Market Issues

There are two important trends that are tending to propel growth in the use of wireless data services. The first is the rapidly increasing use of portable devices such as laptop computers, pen-pads, notebook computers, and other similar devices. Increasingly, the laptop or notebook computer is becoming a standard item of equipment for traveling professional or business person, along with the cellular telephone and pager. This trend has been aided by the steady decrease in prices, increases in reliability, and improvements in capability and design for such devices. The second important trend tending to drive growth in wireless data services is the explosive growth in the use of the Internet. As organizations become increasingly reliant upon the Internet for their everyday operations, they will correspondingly want their employees to have convenient access to the Internet while travelling, just as they do in the office environment. Wireless data services can provide the traveler with the required network access in many situations where wired access to the public network is impractical or inconvenient. Mobile data communication services discussed here provide a solution for wireless access over wide areas. Recent estimates of traffic composition indicate that data traffic now accounts for less than 1% of the traffic on wireless networks, compared to 50% for wireline networks. Therefore, the potential for growth in the wireless data market is seen to be enormous.

33.4 Modem Services Over Cellular Networks

A simple form of wireless data communication now in common use is data transmission using modems or facsimile terminals over analog cellular telephone links. In this form of communication, the mobile user simply accesses a cellular channel just as he would in making a standard voice call over the cellular network. The user then operates the modem or facsimile terminal just as would be done from office to office over the PSTN. A typical connection is shown in Fig. 33.1, where the mobile user has a laptop computer and portable modem in the vehicle, communicating with another modem and computer in the office. Typical users of this mode of communication include service technicians, real estate agents, and traveling sales people. In this form of communication, the network is not actually providing a data service but simply a voice link over which the data modem or fax terminal can interoperate with a corresponding data modem or fax terminal in the office or service center. The connection from the mobile telephone switching office (MTSO) is a standard landline connection, exactly the same as is provided for an ordinary cellular telephone call.

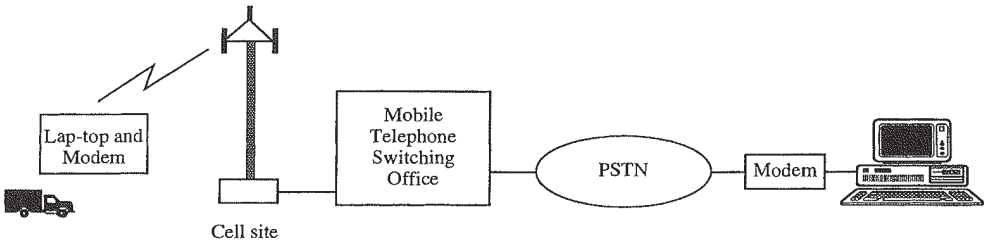


FIGURE 33.1 Modem operation over an analog cellular voice connection.

Many portable modems and fax devices are now available in the market and are sold as elements of the so-called "mobile office" for the traveling business person. Law enforcement personnel are also making increasing use of data communication over cellular telephone and dispatch radio networks to gain rapid access to databases for verification of automobile registrations and drivers' licenses. Portable devices are currently available that operate at transmission rates up to 9.6 or 14.4 kb/s. Error-correction modem protocols such as MNP-10, V.34, and V.42 are used to provide reliable delivery of data in the error-prone wireless transmission environment.

In another form of mobile data service, the mobile subscriber uses a portable modem or fax terminal as already described but now accesses a modem provided by the cellular service operator as part of a *modem pool*, which is connected to the MTSO. This form of service is shown in Fig. 33.2. The modem pool might provide the user with a choice of

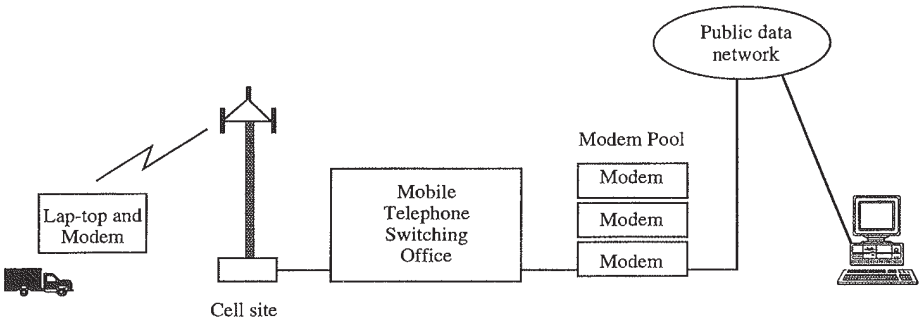


FIGURE 33.2 Cellular data service supported by modem pools in the network.

several standard modem types. The call connection from the modem pool to the office is a digital data connection, which might be supported by any of a number of public packet data networks, such as those providing X.25 service. Here, the cellular operator is providing a special service in the form of modem pool access, and this service in general carries a higher tariff than does standard cellular telephone service, due to the operator's added investment in the modem pools. In this form of service, however, the user in the office or service center does not require a modem but instead has a direct digital data connection to the desk-top or host computer.

Each of the types of wireless data transmission just described is in effect an appliqué onto an underlying cellular telephone service and, therefore, has limitations imposed by the characteristics of the underlying voice-circuit connection. That is, the cellular segment of the call connection is a circuit-mode service, which might be cost effective if the user needs to send long file transfers or fax transmissions but might be relatively costly if only

short messages are to be transmitted and received. This is because the subscriber is being charged for a circuit-mode connection, which stays in place throughout the duration of the communication session, even if only intermittent short message exchanges are needed. The need for systems capable of providing cost-effective communication of relatively short message exchanges led to the development of wireless packet data networks, which we describe next.

33.5 Private Data Networks

Here we describe three packet data networks that provide mobile data services to users in major metropolitan areas of the United States.

33.5.1 ARDIS

ARDIS is a two-way radio service developed as a joint venture between IBM and Motorola and first implemented in 1983. In mid-1994, IBM sold its interest in ARDIS to Motorola and early in 1998 ARDIS was acquired by the American Mobile Satellite Corporation. The ARDIS network consists of four network control centers with 32 network controllers distributed through 1250 base station in 400 cities in the U.S. The service is suitable for two-way transfers of data files of size less than 10 kilobytes, and much of its use is in support of computer-aided dispatching, such as is used by field service personnel, often while they are on customers' premises. Remote users access the system from laptop radio terminals, which communicate with the base stations. Each of the ARDIS base stations is tied to one of the 32 radio network controllers, as shown in Fig. 33.3. The backbone of the network is implemented with leased telephone lines. The four ARDIS hosts, located in Chicago, New York, Los Angeles, and Lexington, KY, serve as access points for a customer's mainframe computer, which can be linked to an ARDIS host using async, bisync, SNA, or X.25 dedicated circuits.

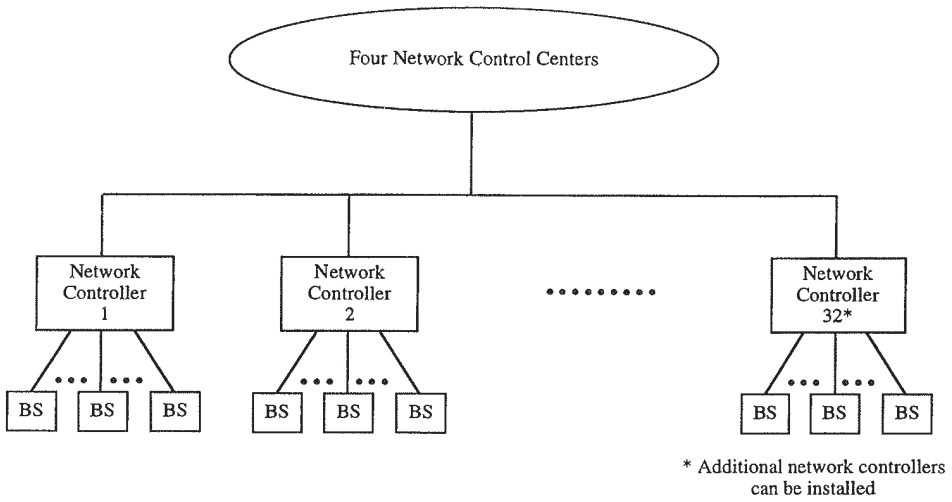


FIGURE 33.3 ARDIS network architecture.

The operating frequency band is 800 MHz, and the RF links use separate transmit and receive frequencies spaced by 45 MHz. The system was initially implemented with an RF channel data rate 4800 b/s per 25-kHz channel, using the MDC-4800 protocol. In 1993 the access data rate was upgraded to 19.2 kb/s, using the RD-LAP protocol, which provides a user data rate of about 8000 b/s. In the same year, ARDIS implemented a nationwide roaming capability, allowing users to travel between widely separated regions without having to preregister their portable terminals in each new region. The ARDIS system architecture is cellular, with cells overlapped to increase the probability that the signal transmission from a portable transmitter will reach at least one base station. The base station power is 40 W, which provides line-of-sight coverage up to a radius of 10–15 miles. The portable units operate with 4 W of radiated power. The overlapping coverage, combined with designed power levels, and error-correction coding in the transmission format, insures that the ARDIS can support portable communications from inside buildings, as well as on the street. This capability for in-building coverage is an important characteristic of the ARDIS service. The modulation technique is frequency-shift keying (FSK), the access method is frequency division multiple access (FDMA), and the transmission packet length is 256 bytes.

Although the use of overlapping coverage, almost always on the same frequency, provides reliable radio connectivity, it poses the problem of interference when signals are transmitted simultaneously from two adjacent base stations. The ARDIS network deals with this by turning off neighboring transmitters, for 0.5–1 s, when an outbound transmission occurs. This scheme has the effect of constraining overall network capacity.

The laptop portable terminals access the network using a random access method called data sense multiple access (DSMA) [11]. A remote terminal listens to the base station transmitter to determine if a "busy bit" is on or off. When the busy bit is off, the remote terminal is allowed to transmit. If two remote terminals begin to transmit at the same time, however, the signal packets may collide, and retransmission will be attempted, as in other contention-based multiple access protocols. The busy bit lets a remote user know when other terminals are transmitting and, thus, reduces the probability of packet collision.

33.5.2 MOBITEX

The MOBITEX system is a nationwide, interconnected trunked radio network developed by Ericsson and Swedish Telecom. The first MOBITEX network went into operation in Sweden in 1986, and networks have either been implemented or are being deployed in 22 countries. A MOBITEX operations association oversees the open technical specifications and coordinates software and hardware developments [6]. In the U.S., MOBITEX service was introduced by RAM Mobile Data in 1991. In 1992 Bell South Enterprises became a partner with RAM. Currently, RAM Mobile Data is a subsidiary of Bell South Corporation and operates under the name Bell South Wireless Data, LP. The Bell South MOBITEX service now covers over 90% of the U.S. urban business population with about 2000 base stations, and it provides automatic "roaming" across all service areas. By locating its base stations close to major business centers, the system provides a degree of in-building signal coverage. Although the MOBITEX system was designed to carry both voice and data service, the U.S. and Canadian networks are used to provide data service only. MOBITEX is an intelligent network with an open architecture that allows establishing virtual networks. This feature facilitates the mobility and expandability of the network [7, 12].

The MOBITEX network architecture is hierarchical, as shown in Fig. 33.4. At the top of the hierarchy is the network control center (NCC), from which the entire network is managed. The top level of switching is a national switch (MHX1) that routes traffic between service regions. The next level comprises regional switches (MHX2s), and below that are

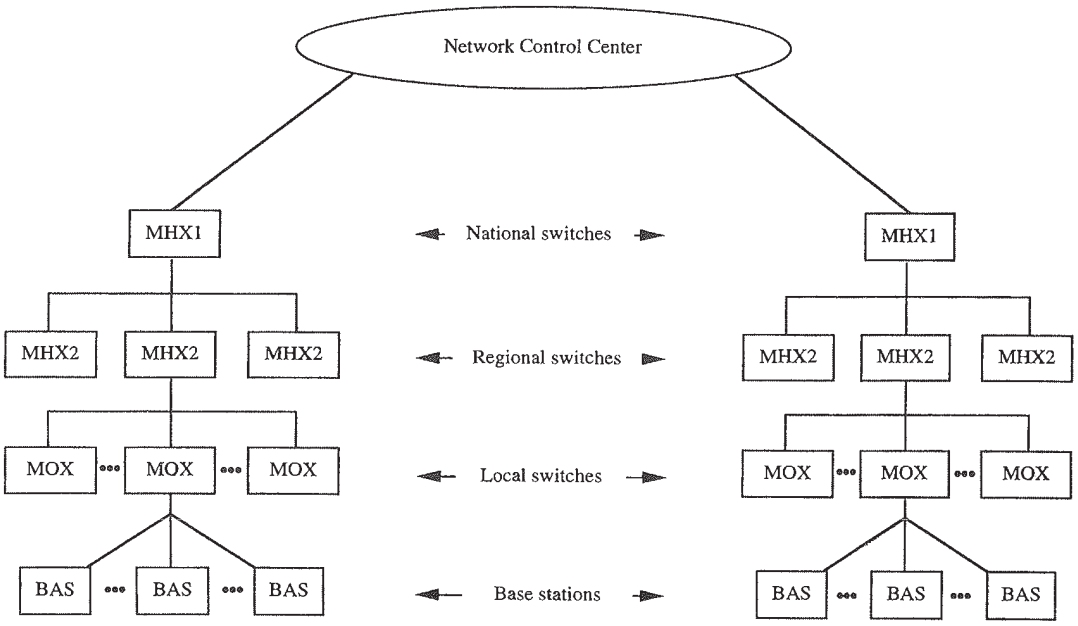


FIGURE 33.4 MOBITEX network architecture.

local switches (MOXs), each of which handles traffic within a given service area. At the lowest level in the network, multichannel trunked-radio base stations communicate with the mobile and portable data sets. MOBITEX uses packet-switching techniques, as does ARDIS, to allow multiple users to access the same channel at the same time. Message packets are switched at the lowest possible network level. If two mobile users in the same service area need to communicate with each other, their messages are relayed through the local base station, and only billing information is sent up to the network control center.

The base stations are laid out in a grid pattern using the same frequency reuse rules as are used for cellular telephone networks. In fact, the MOBITEX system operates in much the same way as a cellular telephone system, except that handoffs are not managed by the network. That is, when a radio connection is to be changed from one base station to another, the decision is made by the mobile terminal, not by a network computer as in cellular telephone systems.

To access the network, a mobile terminal finds the base station with the strongest signal and then registers with that base station. When the mobile terminal enters an adjacent service area, it automatically re-registers with a new base station, and the user's whereabouts are relayed to the higher level network nodes. This provides automatic routing of messages bound for the mobile user, a capability known as *roaming*. The MOBITEX network also has a store-and-forward capability.

The mobile units transmit at 896 to 901 MHz and the base stations at 935 to 940 MHz. The base stations use a trunked radio design employing 2 to 30 radio channels in each service area. The system uses dynamic power setting, in the range of 100 mW–10 W for mobile units and 100 mW–4 W for portable units. The Gaussian minimum shift keying (GMSK) modulation technique is used, with $BT = 0.3$ and noncoherent demodulation. The transmission rate is 8000 b/s half-duplex in 12.5-kHz channels, and the service is suitable for file transfers up to 20 kilobytes. The MOBITEX system uses a proprietary network-layer protocol called MPAK, which provides a maximum packet size of 512 bytes and a

24-b address field. Forward-error correction, as well as retransmissions, are used to ensure the bit-error-rate quality of delivered data packets. Fig. 33.5 shows the packet structure at various layers of the MOBITEX protocol stack. The system uses the reservation-slotted ALOHA (R-S-ALOHA) random access method.

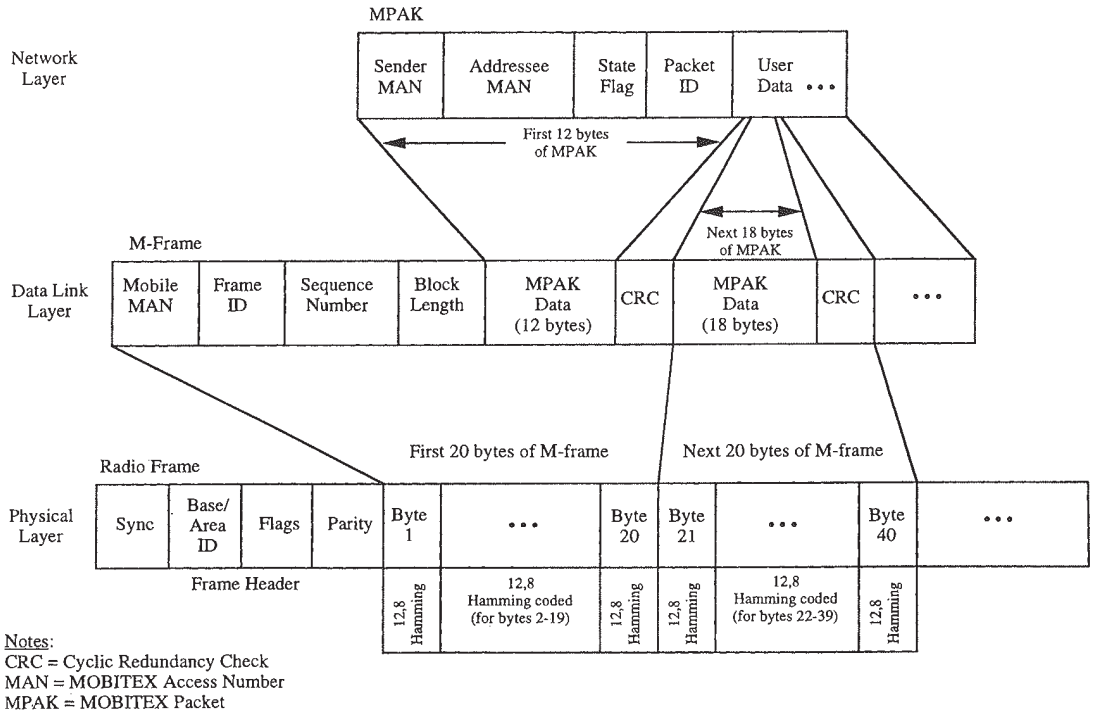


FIGURE 33.5 MOBITEX packet and frame structure at three layers of the protocol stack.

33.5.3 Ricochet Network

The Ricochet network is a wide-area wireless data network developed and operated by Metricom, Inc., headquartered in Los Gatos, California. Ricochet Network service is currently available in the greater San Francisco Bay Area, Seattle, and Washington, D.C. as well a number of university campuses, schools, and corporate campuses. Metricom plans to expand the service to other metropolitan areas as well. Ricochet is a spread-spectrum, frequency-hopping design, operating in an unlicensed band, the 902–928 MHz industrial, scientific, and medical (ISM) band. The Ricochet network employs shoebox-sized radio transceivers, called microcell radios, which are typically mounted on street light poles or utility poles. Each microcell unit accesses electric power from the street light itself, or from adjacent power lines, and is otherwise a self-contained unit. Each microcell radio employs 162 frequency-hopping channels and uses a randomly selected hopping sequence. This allows many subscribers to use the network simultaneously.

The microcell radios are typically installed every 1/4 to 1/2 mi in a regular geometric pattern. A mobile subscriber using a Ricochet modem accesses any of the microcell radios with data packets, and the packets are routed from one microcell radio to another by

means of an efficient routing protocol. Within a 20 sq mi radius containing approximately 100 microcell radios, the network has a wired access point (WAP) that collects the radio packets and formats them for transmission over a wired IP network backbone through a T1 frame relay connection. Each WAP and the microcell radios that communicate with it can support thousands of subscribers.

Data packets transmitted by a Ricochet modem can be routed to another Ricochet modem or to one of a number of gateways that allow subscribers to access other services. The current network design includes gateways to the Internet, to the PSTN, to an X.25 network, and to corporate Intranets and LANs. The current Ricochet network design provides user data rates up to about 128 kb/s, and higher data rates are being planned.

33.6 Cellular Data Networks and Services

33.6.1 Cellular Digital Packet Data (CDPD)

The cellular digital packet data (CDPD) system was designed to provide packet data services as an overlay onto the existing analog cellular telephone network, which is called advanced mobile phone service (AMPS). CDPD was developed by IBM in collaboration with the major cellular carriers. Any cellular carrier owning a license for AMPS service is free to offer its customers CDPD service without any need for further licensing. A basic concept of the CDPD system is to provide data services on a noninterfering basis with the existing analog cellular telephone services using the same 30-kHz channels. This is accomplished in either of two ways. First, one or a few AMPS channels in each cell site can be devoted to CDPD service. Second, CDPD is designed to make use of a cellular channel that is temporarily not being used for voice traffic and to move to another channel when the current channel is assigned to voice service. In most of the CDPD networks deployed to date, the fixed-channel implementation is being used.

The compatibility of CDPD with the existing cellular telephone system allows it to be installed in any AMPS cellular system in North America, providing data services that are not dependent on support of a digital cellular standard in the service area. The participating companies issued release 1.0 of the CDPD specification in July 1993, and release 1.1 was issued in late 1994 [2]. At this writing (mid-1998), CDPD service is implemented in many of the major market areas in the U.S. Typical applications for CDPD service include: electronic mail, field support servicing, package delivery tracking, inventory control, credit card verification, security reporting, vehicle theft recovery, traffic and weather advisory services, and a wide range of information retrieval services.

Although CDPD cannot increase the number of channels usable in a cell, it can provide an overall increase in user capacity if data customers use CDPD instead of voice channels. This capacity increase results from the inherently greater efficiency of a connectionless packet data service relative to a connection-oriented service, given bursty data traffic. That is, a packet data service does not require the overhead associated with setup of a voice traffic channel in order to send one or a few data packets. In the following paragraphs we briefly describe the CDPD network architecture and the principles of operation of the system. Our discussion follows [13], closely.

The basic structure of a CDPD network (Fig. 33.6) is similar to that of the cellular network with which it shares transmission channels. Each mobile end system (M-ES) communicates with a mobile data base station (MDBS) using the protocols defined by the air-interface specification, to be described subsequently. The MDBSs are typically collocated with the cell equipment providing cellular telephone service to facilitate the channel-sharing procedures. All of the MDBSs in a service area are linked to a mobile data intermediate system

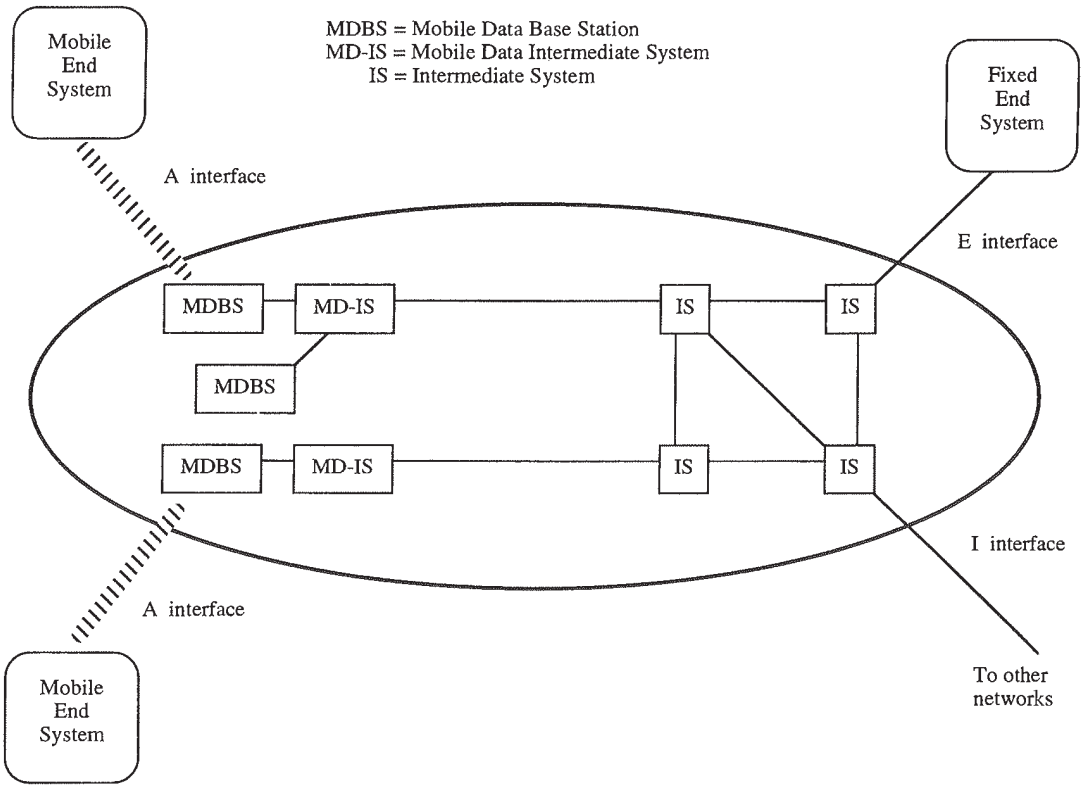


FIGURE 33.6 Cellular digital packet data network architecture.

(MD-IS) by microwave or wireline links. The MD-IS provides a function analogous to that of the mobile switching center (MSC) in a cellular telephone system. The MD-IS may be linked to other MD-ISs and to various services provided by end systems outside the CDPD network. The MD-IS also provides a connection to a network management system and supports protocols for network management access to the MDBSs and M-ESs in the network.

Service endpoints can be local to the MD-IS or remote, connected through external networks. A MD-IS can be connected to any external network supporting standard routing and data exchange protocols. A MD-IS can also provide connections to standard modems in the PSTN by way of appropriate modem interworking functions (modem emulators). Connections between MD-ISs allow routing of data to and from M-ESs that are roaming, that is, operating in areas outside their home service areas. These connections also allow MD-ISs to exchange information required for mobile terminal authentication, service authorization, and billing.

CDPD employs the same 30-kHz channelization as used in existing AMPS cellular systems throughout North America. Each 30-kHz CDPD channel supports channel transmission rates up to 19.2 kb/s. Degraded radio channel conditions, however, will limit the actual information payload throughput rate to lower levels, and will introduce additional time delay due to the error-detection and retransmission protocols.

The CDPD radio link physical layer uses GMSK modulation at the standard cellular carrier frequencies, on both forward (base-to-mobile) and reverse (mobile-to-base) links. The Gaussian pulse-shaping filter is specified to have bandwidth-time product $B_bT = 0.5$.

The specified B_bT product assures a transmitted waveform with bandwidth narrow enough to meet adjacent-channel interference requirements, while keeping the intersymbol interference small enough to allow simple demodulation techniques. The choice of 19.2 kb/s as the channel bit rate yields an average power spectrum that satisfies the emission requirements for analog cellular systems and for dual-mode digital cellular systems.

The forward channel carries data packets transmitted by the MDBS, whereas the reverse channel carries packets transmitted by the M-ESs. In the forward channel, the MDBS forms data frames by adding standard high level data link control (HDLC) terminating flags and inserted zero bits, and then segments each frame into blocks of 274 b. These 274 b, together with an 8-b *color code* for MDBS and MD-IS identification, are encoded into a 378-b coded block using a (63, 47) Reed–Solomon code over a 64-ary alphabet. A 6-b synchronization and flag word is inserted after every 9 code symbols. The flag words are used for reverse link access control. The forward link block structure is shown in Fig. 33.7.

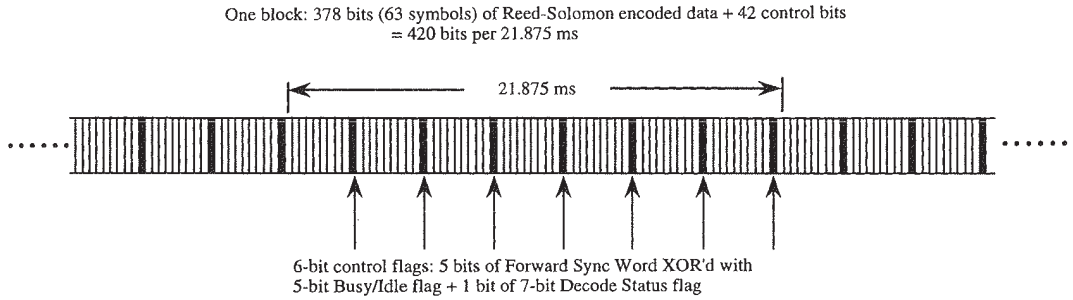


FIGURE 33.7 Cellular digital packet data forward link block structure.

In the reverse channel, when an M-ES has data frames to send, it formats the data with flags and inserted zeros in the same manner as in the forward link. That is, the reverse link frames are segmented and encoded into 378-b blocks using the same Reed–Solomon code as in the forward channel. The M-ES may form up to 64 encoded blocks for transmission in a single reverse channel transmission burst. During the transmission, a 7-b transmit continuity indicator is interleaved into each coded block and is set to all ones to indicate that more blocks follow, or all zeros to indicate that this is the last block of the burst. The reverse channel block structure is shown in Fig. 33.8.

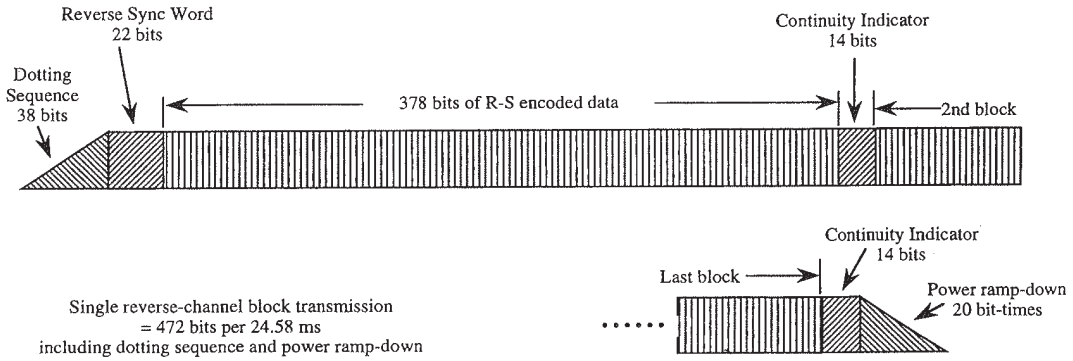


FIGURE 33.8 Cellular digital packet data reverse link block structure.

The media access control (MAC) layer in the forward channel is relatively simple. The receiving M-ES removes the inserted zeros and HDLC flags and reassembles data frames that were segmented into multiple blocks. Frames are discarded if any of their constituent blocks are received with uncorrectable errors.

On the reverse channel (M-ES to MDBS), access control is more complex, since several M-ESs must share the channel. CDPD uses a multiple access technique called digital sense multiple access (DSMA), which is closely related to the carrier sense multiple access/collision detection (CSMA/CD) access technique.

The network layer and higher layers of the CDPD protocol stack are based on standard ISO and Internet protocols. The CDPD specification stipulates that there be no changes to protocols above the network layer of the seven-layer ISO model, thus insuring the compatibility of applications software used by CDPD subscribers.

The selection of a channel for CDPD service is accomplished by the radio resource management entity in the MDBS. Through the network management system, the MDBS is informed of the channels in its cell or sector that are available either as dedicated data channels or as potential CDPD channels when they are not being used for analog cellular service, depending on which channel allocation method is implemented. For the implementation in which CDPD service is to use "channels of opportunity," there are two ways in which the MDBS can determine whether the channels are in use. If a communication link is provided between the analog system and the CDPD system, the analog system can inform the CDPD system directly about channel usage. If such a link is not available, the CDPD system can use a forward power monitor ("sniffer" antenna) to detect channel usage on the analog system. Circuitry to implement this function can be built into the cell sector interface.

Another version of CDPD called circuit-switched CDPD (C-SCDPD) is designed to provide service to subscribers traveling in areas where the local cellular service provider has not implemented the CDPD service. With C-SCDPD, the subscriber establishes a standard analog cellular circuit connection to a prescribed number, and then transmits and receives CDPD data packets over that circuit connection. The called number is a gateway that provides connection to the CDPD backbone packet network.

33.6.2 Digital Cellular Data Services

In response to the rapid growth in demand for cellular telephone service throughout the U.S and Canada, the Cellular Telecommunications Industry Association (CTIA) and the Telecommunications Industry Association (TIA) have been developing standards for new digital cellular systems to replace the existing analog cellular system (advanced mobile phone system or AMPS). Two air-interface standards have now been published. The IS-54 standard specifies a three-slot TDMA system, and the I-95 standard specifies a CDMA spread spectrum system. In both systems, a variety of data services are being planned.

Following the development of the IS-95 standard for CDMA voice service, the cellular industry has worked on defining various data services to operate in the same networks. The general approach taken in the definition of IS-95 data services has been to base the services on standard data protocols, to the greatest extent possible [17]. The previously-specified physical layer of the IS-95 protocol stack was adopted for the physical layer of the data services, with an appropriate radio link protocol (RLP) overlaid. The first CDMA data service to be defined was asynchronous ("start-stop" interface) data and Group-3 facsimile. This service provides for interoperability with many standard PSTN data modems as well as standard office fax machines. This service is in the category of *circuit-mode service,* since

a circuit connection is first established, just as with a voice call, and the circuit connection is maintained until the user disconnects.

Following the standardization of the asynchronous data service, the industry defined a service that carries packet-formatted data over a CDMA circuit connection. It is important to note that this is not a true packet-data service over the radio link, since the full circuit connection is maintained regardless of how little packet data is transmitted. One potential application for this type of service is to provide subscribers with CDPD access from a CDMA network.

It is recognized that in order to make use more efficient, it will be highly desirable to provide a contention-based packet data service in CDMA cellular networks. This is currently a subject of study in CDMA data services standardization groups.

In parallel with the CDMA data services efforts, another TIA task group, TR45.3.2.5, has defined standards for digital data services for the TDMA digital cellular standard IS-54 [15, 18]. As with the IS-95 data services effort, initial priority was given to standardizing circuit-mode asynchronous data and Group-3 facsimile services [16].

33.7 Other Developing Standards

33.7.1 TErrestrial Trunked Radio (TETRA)

As has been the case in North America, there is interest in Europe in establishing fixed wide-area standards for mobile data communications. Whereas the Pan-European standard for digital cellular, termed Global Systems for Mobile communications (GSM), will provide an array of data services, data will be handled as a circuit-switched service, consistent with the primary purpose of GSM as a voice service system. Therefore, the European Telecommunications Standards Institute (ETSI) began developing a public standard in 1988 for trunked radio and mobile data systems, and this standardization process continues today. The standards, which are now known generically as TErrestrial Trunked Radio (TETRA) (formerly Trans-European Trunked Radio), were made the responsibility of the ETSI RES 6 subtechnical committee [3]. In 1996, the TETRA standardization activity was elevated within RES-6 with the creation of project TETRA.

TETRA is being developed as a family of standards. One branch of the family is a set of radio and network interface standards for trunked voice (and data) services. The other branch is an air-interface standard optimized for wide-area packet data services for both fixed and mobile subscribers and supporting standard network access protocols. Both versions of the standard will use a common physical layer, based on $\pi/4$ differential quadrature phase shift keying ($\pi/4$-DQPSK) modulation operating at a channel rate of 36 kb/s in each 25-kHz channel. The composite data rate of 36 kb/s comprises four 9 kb/s user channels multiplexed in a TDMA format. The TETRA standard provides both connection-oriented and connectionless data services, as well as mixed voice and data services.

TETRA has been designed to operate in the frequency range from VHF (150 MHz) to UHF (900 MHz). The RF carrier spacing in TETRA is 25 kHz. In Europe, harmonized bands have been designated in the frequency range 380–400 MHz for public safety users. It is expected that commercial users will adopt the 410–430 MHz band. The Conference of European Posts and Telecommunications Administrations (CEPT) has made additional recommendations for use in the 450–470 MHz and 870–876/915–921 MHz frequency bands.

Table 33.1 compares the chief characteristics and parameters of the wireless data services described.

TABLE 33.1 Characteristics and Parameters of Five Mobile Data Services

System:	ARDIS	MOBITEX	CDPD	IS-95[b]	TETRA[b]
Frequency band Base to mobile, (MHz). Mobile to base, (MHz).	(800 band, 45-kHz sep.)	935–940[a] 896–901	869–894 824–849	869–894 824–849	(400 and 900 Bands)
RF channel spacing	25 kHz (U.S.)	12.5 kHz	30 kHz	1.25 MHz	25 kHz
Channel access/ multiuser access	FDMA/ DSMA	FDMA/ dynamic- R-S-ALOHA	FDMA/ DSMA	FDMA/ CDMA-SS	FDMA/ DSMA & SAPR[c]
Modulation method	FSK, 4-FSK	GMSK	GMSK	4-PSK/DSSS	$\pi/4$-QDPSK
Channel bit rate, kb/s	19.2	8.0	19.2	9.6	36
Packet length	Up to 256 bytes (HDLC)	Up to 512 bytes	24–928 b	(Packet service-TBD)	192 b (short) 384 b (long)
Open architecture	No	Yes	Yes	Yes	Yes
Private or Public Carrier	Private	Private	Public	Public	Public
Service Coverage	Major metro. areas in U.S.	Major metro. areas in U.S.	All AMPS areas	All CDMA cellular areas	European trunked radio
Type of coverage	In-building & mobile	In-building & mobile	Mobile	Mobile	Mobile

[a] Frequency allocation in the U.S. in the U.K., 380–450 MHz band is used.
[b] IS-95 and TETRA data services standardization in progress.
[c] Slotted-ALOHA packet reservation.

33.8 Conclusions

Mobile data radio systems have grown out of the success of the paging-service industry and the increasing customer demand for more advanced services. The growing use of portable, laptop, and palmtop computers and other data services will propel a steadily increasing demand for wireless data services. Today, mobile data services provide length-limited wireless connections with in-building penetration to portable users in metropolitan areas. The future direction is toward wider coverage, higher data rates, and capability for wireless Internet access.

References

[1] Bodson, D., McClure, G.F., and McConoughey, S.R., Eds., *Land-Mobile Communications Engineering*, Selected Reprint Ser., IEEE Press, New York, 1984.

[2] CDPD Industry Coordinator. Cellular Digital Packet Data Specification, Release 1.1, November 1994, Pub. CDPD, Kirkland, WA, 1994.

[3] Haine, J.L., Martin, P.M., and Goodings, R.L.A., A European standard for packet-mode mobile data, *Proceedings of Personal, Indoor, and Mobile Radio Conference (PIMRC'92)*, Boston, MA. Pub. IEEE, New York, 1992.

[4] Hata, M., Empirical formula for propagation loss in land-mobile radio services. *IEEE Trans. on Vehicular Tech.*, 29(3), 317–325, 1980.

[5] International Standards Organization (ISO), Protocol for providing the connectionless-mode network service. Pub. ISO 8473, 1987.

[6] Khan, M. and Kilpatrick, J., MOBITEX and mobile data standards. *IEEE Comm. Maga.*, 33(3), 96–101, 1995.

[7] Kilpatrick, J.A., Update of RAM Mobile Data's packet data radio service. *Proceedings of the 42nd IEEE Vehicular Technology Conference (VTC'92),* Denver, CO, 898–901, Pub. IEEE, New York, 1992.

[8] Longley, A.G. and Rice, P.L., Prediction of tropospheric radio transmission over irregular terrain. A computer method—1968, Environmental Sciences and Services Administration Tech. Rep. ERL 79-ITS 67, U.S. Government Printing Office, Washington, D.C., 1968.

[9] Okumura, Y., Ohmori, E., Kawano, T., and Fukuda, K., Field strength and its variability in VHF and UHF land-mobile service. *Review of the Electronic Communication Laboratory,* 16, 825–873, 1968.

[10] Pahlavan, K. and Levesque, A.H., Wireless data communications. *Proceedings of the IEEE,* 82(9), 1398–1430, 1994.

[11] Pahlavan, K. and Levesque, A.H., *Wireless Information Networks,* J. Wiley & Sons, New York, 1995.

[12] Parsa, K., The MOBITEX packet-switched radio data system. *Proceedings of the Personal, Indoor and Mobile Radio Conference, (PIMRC'92)*, Boston, MA, 534–538. Pub. IEEE, New York, 1992.

[13] Quick, R.R., Jr. and Balachandran, K., Overview of the cellular packet data (CDPD) system. *Proceedings of the Personal, Indoor and Mobile Radio Conference, (PIMRC'93)*, Yokohama, Japan, 338–343. Pub. IEEE, New York, 1993.

[14] Rice, P.L., Longley, A.G., Norton, K.A., and Barsis, A.P., Transmission loss predictions for tropospheric communication circuits. National Bureau of Standards, Tech. Note 101, Boulder, CO, 1967.

[15] Sacuta, A., Data standards for cellular telecommunications—a service-based approach. *Proceedings of the 42nd IEEE Vehicular Technology Conference,* Denver CO, 263–266. Pub. IEEE, New York, 1992.

[16] Telecommunications Industry Association. Async data and fax. Project No. PN-3123, and Radio link protocol 1. Project No. PN-3306, Nov. 14. Issued by TIA, Washington, D.C., 1994.

[17] Tiedemann, E., Data services for the IS-95 CDMA standard. Presented at Personal, Indoor and Mobile Radio Conf. PIMRC'93. Yokohama, Japan, 1993.

[18] Weissman, D., Levesque, A.H., and Dean, R.A., Interoperable wireless data. *IEEE Comm. Mag.,* 31(2), 68–77, 1993.

Further Information

Reference [11] provides a comprehensive survey of the wireless data field as of mid-1994. The monthly journals *IEEE Communications Magazine* and *IEEE Personal Communications Magazine,* and the bimonthly journal *IEEE Transactions on Vehicular Technology* report advances in many areas of mobile communications, including wireless data. For subscription information contact: IEEE Service Center, 445 Hoes Lane, P. O. Box 1331, Piscataway, NJ, 08855-1131. Phone (800)678-IEEE.

34

Wireless ATM: Interworking Aspects

Melbourne Barton
Bellcore

Matthew Cheng
Bellcore

Li Fung Chang
Bellcore

34.1 Introduction II-1
34.2 Background and Issues II-2
34.3 Wireless Interworking With Transit ATM Networks ... II-3
Integrated Wireless–Wireline ATM Network Architecture • Wireless Access Technology Options
34.4 The PCS-to-ATM Interworking Scenario II-10
Architecture and Reference Model • Signalling Link Evolution
34.5 QoS Support II-14
34.6 Conclusions II-18
Defining Terms II-19
References II-20
Further Information II-21

34.1 Introduction

The ATM Forum's **wireless asynchronous transfer mode (WATM)** Working Group (WG) is developing specifications intended to facilitate the use of ATM technology for a broad range of wireless network access and interworking scenarios, both public and private. These specifications are intended to cover the following two broad WATM application scenarios:

- *End-to-End WATM*—This provides seamless extension of ATM capabilities to mobile terminals, thus providing ATM virtual channel connections (VCCs) to wireless hosts. For this application, high data rates are envisaged, with limited coverage, and transmission of one or more ATM cells over the air.
- *WATM Interworking*—Here the fixed ATM network is used primarily for high-speed transport by adding mobility control in the ATM infrastructure network, without changing the non-ATM air interface protocol. This application will facilitate the use of ATM as an efficient and cost-effective infrastructure network for next generation non-ATM wireless access systems, while providing a smooth migration path to seamless end-to-end WATM.

This chapter focuses on the ATM interworking application scenario. It describes various interworking and non-ATM wireless access options and their requirements. A generic

0-8493-8597-0/99/$0.00+$.50
© 1999 by CRC Press LLC

personal communications services (PCS)[1]-to-ATM interworking scenario is described which enumerates the architectural features, protocol reference models, and signalling issues that are being addressed for mobility support in the ATM infrastructure network. Evolution strategies intended to eventually provide end-to-end WATM capabilities and a methodology to consistently support a range of quality of service (QoS) levels on the radio link are also described.

34.2 Background and Issues

ATM is the switching and multiplexing standard for **broadband integrated services digital network (BISDN),** which will ultimately be capable of supporting a broad range of applications over a set of high capacity multiservice interfaces. ATM holds out the promise of a single network platform that can simultaneously support multiple bandwidths and latency requirements for fixed access and wireless access services without being dedicated to any one of them. In today's wireline ATM network environment, the **user network interface (UNI)** is fixed and remains stationary throughout the connection lifetime of a call. The technology to provide fixed access to ATM networks has matured. Integration of fixed and wireless access to ATM will present a cost-effective and efficient way to provide future tetherless multimedia services, with common features and capabilities across both wireline and wireless network environments. Early technical results [19, 20, 25] have shown that standard ATM protocols can be used to support such integration and extend mobility control to the subscriber terminal by incorporating wireless specific layers into the ATM user and control planes.

Integration of wireless access features into wireline ATM networks will place additional demands on the fixed network infrastructure due primarily to the additional user data and signalling traffic that will be generated to meet future demands for wireless multimedia services. This additional traffic will allow for new signalling features including registration, call delivery, and handoff during the connection lifetime of a call. Registration keeps track of a wireless user's location, even though the user's communication link might not be active. Call delivery, establishes a connection link to/from a wireless user with the help of location information obtained from registration. The registration and call delivery functions are referred to as **location management.** Handoff is the process of switching (rerouting) the communication link from the old coverage area to the new coverage area when a wireless user moves during active communication. This function is also referred to as **mobility management.**

In June, 1996 the ATM Forum established a WATM WG to develop requirements and specifications for WATM. The WATM standards are to be compatible with ATM equipment adhering to the (then) current ATM Forum specifications. The technical scope of the WATM WG includes development of: (1) **radio access layer (RAL)** protocols for the physical (PHY), medium access control (MAC), and data link control (DLC) layers; (2) wireless control protocols for radio resource management; (3) mobile protocol extensions for ATM (mobile ATM) including handoff control, routing considerations, location management, traffic and QoS control, and wireless network management; and (4) wireless

[1]The term PCS is being used in a generic sense to mean emerging digital wireless systems, which support mobility in microcellular and other environments. It is currently defined in ANSI T1.702-1995 as "A set of capabilities that allows some combination of terminal mobility, personal mobility, and service profile management."

interworking functions for mapping between non-ATM wireless access and ATM signalling and control entities. Phase-1 WATM specifications are being developed for short-range, high-speed, end-to-end WATM devices using wireless terminals that operate in the 5 GHz frequency band. Operating speeds will be up to 25 Mb/s, with a range of 30 m–50 m indoor and 200 m–300 m outdoor. The European Telecommunications Standards Institute (ETSI) Broadband Radio Access Networks (BRAN) project is developing the RAL for Phase-1 WATM specifications using the **HIgh PERformance Radio LAN (HIPERLAN)** functional requirements. The ATM Forum plans to release the Phase-1 WATM specifications by the second quarter of 1999.

There are a number of emerging wireless access systems (including digital cellular, PCS, legacy LANs based on the IEEE 802.11 standards, satellite, and IMT-2000 systems), that could benefit from access to, and interworking with, the fixed ATM network. These wireless systems are based on different access technologies and require development of different **interworking functions (IWFs)** at the wireless access network and fixed ATM network boundaries to support WATM interworking. For example, a set of network interfaces have already been identified to support PCS access to the fixed ATM network infrastructure, without necessarily modifying the PCS air interface protocol to provide end-to-end ATM capabilities [4, 6, 28]. The WATM WG might consider forming sub-working groups which could work in parallel to identify other network interfaces and develop IWF specifications for each of (or each subset of) the wireless access options that are identified. These WATM interworking specifications would be available in the Phase-2 and later releases of the WATM standards.

Some service providers and network operators see end-to-end WATM as a somewhat limited service option at this time because it is being targeted to small enterprise networks requiring high-speed data applications, with limited coverage and low mobility. On the other hand, WATM interworking can potentially support a wider range of services and applications, including low-speed voice and data access, without mandatory requirements to provide over-the-air transmission of ATM cells. It will allow for wider coverage, possibly extending to macrocells, while supporting higher mobility. WATM interworking will provide potential business opportunities, especially for public switched telephone network (PSTN) operators and service providers, who are deploying emerging digital wireless technologies such as PCS. Existing wireless service providers (WSPs) with core network infrastructures in place can continue to use them while upgrading specific network elements to provide ATM transport. On the other hand, a new WSP entrant without such a network infrastructure can utilize the public (or private) ATM transport network to quickly deploy the WATM interworking service, and not be burdened with the cost of developing an overlay network. If the final goal is to provide end-to-end WATM services and applications, then WATM interworking can provide an incremental development path.

34.3 Wireless Interworking With Transit ATM Networks

Figure 34.1 shows one view of the architectural interworking that will be required between public/private wireless access networks and the fixed ATM network infrastructure. It identifies the network interfaces where modifications will be required to allow interworking between both systems. A desirable objective in formulating WATM specifications for this type of wireless access scenario should be to minimize modifications to the transit ATM network and existing/emerging wireless access system specifications. This objective can be largely met by limiting major modifications to the network interfaces between the boundaries of the transit ATM network and public/private wireless networks, and where possible,

adopting existing network standard processes (i.e., SS7, IS-41, MAP, AIN.3, Q.931, Q.932, Q.2931, etc.) to minimize development costs and upgrades to existing service providers' network infrastructure. Development of standard network interfaces that allow interworking of a reasonable subset of non-ATM digital wireless access systems with the fixed ATM network infrastructure insure that:

- Large-scale revisions and modifications are not necessary to comply with later versions of the WATM specifications to accommodate other emerging digital wireless access systems that do not require end-to-end ATM connectivity
- WATM systems are supported by open interfaces with a rich set of functionality to provide access to both ATM and non-ATM wireless access terminal devices
- WATM services can reach a much larger potential market including those markets providing traditional large-scale support for existing voice services and vertical voice features.

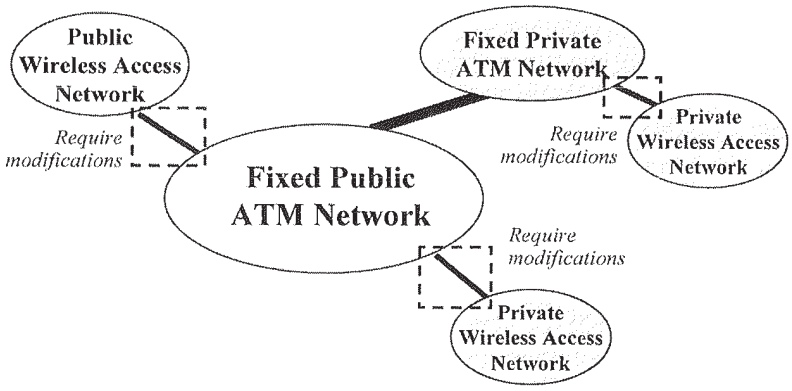


FIGURE 34.1 Wireless ATM interworking architecture.

34.3.1 Integrated Wireless–Wireline ATM Network Architecture

Figure 34.2 shows one example of a mature, multifunctional ATM transport network platform, which provides access to fixed and mobile terminals for wide-area coverage. Four distinct network interfaces are shown supporting: (1) fixed access with non-ATM terminal, (2) fixed access with ATM terminal, (3) wireless access with non-ATM terminal (WATM interworking), and (4) wireless access with ATM terminal (end-to-end WATM).

International Telecommunications Union (ITU) and ATM Forum standard interfaces either exist today or are being developed to support fixed access to ATM networks through various network interfaces. These include frame relay service (FRS) UNI, cell relay service (CRS) UNI, circuit emulation service (CES) UNI, and switched multimegabit data service (SMDS) subscriber NI (SNI). The **BISDN intercarrier interface (B-ICI)** specification [1] provides examples of wired IWFs that have been developed for implementation above the ATM layer to support intercarrier service-specific functions developed at the network nodes, and distributed in the public ATM/BISDN network. These distributed, service-specific functions are defined by B-ICI for FRS, CRS, CES, and SMDS. Examples of such functions include ATM cell conversion, clock recovery, loss of signal and alarm indi-

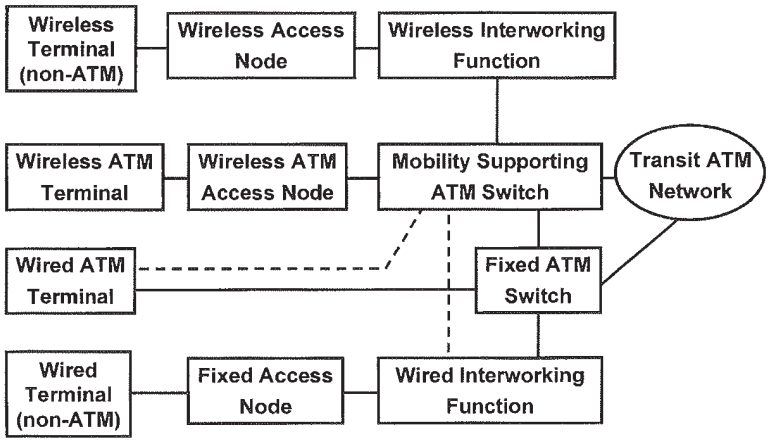


FIGURE 34.2 Wireless and wireline system integration with transit ATM Network.

cation detection, virtual channel identifier (VCI) mapping, access class selection, encapsulation/mapping, and QoS selection. In addition to the B-ICI, the **private network-network interface (PNNI)** specification [2] defines the basic call control signalling procedures (e.g., connection setup and release) in private ATM networks. It also has capabilities for autoconfiguration, scalable network hierarchy formation, topology information exchange, and dynamic routing.

On the wireless access side, existing ITU specifications provide for the transport of wireless services on public ATM networks (see, e.g., [10, 11]). For example, if the ATM UNI is at the mobile switching center (MSC), then message transfer part (MTP) 1 and 2 would be replaced by the PHY and ATM layers, respectively, the broadband ISDN user part (BISUP) replaced by MTP 3, and a common channel signalling (CCS) interface deployed in the ATM node. BISUP is used for ATM connection setup and any required feature control. If the ATM UNI is at the base station controller (BSC), then significant modifications are likely to be required. Equipment manufacturers have not implemented, to any large degree, the features that are available with the ITU specifications. In any case, these features are not sufficient to support the WATM scenarios postulated in this chapter.

The two sets of WATM scenarios postulated in this chapter are shown logically interfaced to the ATM network through a **mobility-enabled ATM (ME-ATM)** switch. This enhanced ATM access switch will have capabilities to support mobility management and location management. In addition to supporting handoff and rerouting of ATM connections, it will be capable of locating a mobile user anywhere in the network.

It might be desirable that functions related to mobility not be implemented in standard ATM switches so that network operators and service providers are not required to modify their ATM switches in order to accommodate WATM and related services. A feasible strategy that has been proposed is to implement mobility functions in servers (i.e., service control modules or SCMs) that are logically separated from the ATM switch. In these servers, all mobility features, service creation logic, and service management functions will be implemented to allow logical separation of ATM switching and service control from mobility support, service creation, and management. The open network interface between the ATM access switch and SCM will be standardized to enable multivendor operation. It would be left up to the switch manufacturer to physically integrate the SCM into a new ATM switching fabric, or implement the SCM as a separate entity.

34.3.2 Wireless Access Technology Options

Figure 34.3 identifies various digital wireless systems that could be deployed to connect mobile terminals to the transit ATM network through IWFs that have to be specified. The main emphasis is on developing mobility support in the transit ATM infrastructure network to support a range of wireless access technologies. Significant interests have been expressed, through technical contributions and related activities in the WATM WG, in developing specifications for the IWFs shown in Fig. 34.3 to allow access to the fixed ATM network through standard network interfaces. The main standardization issues that need to be addressed for each wireless access system are described below.

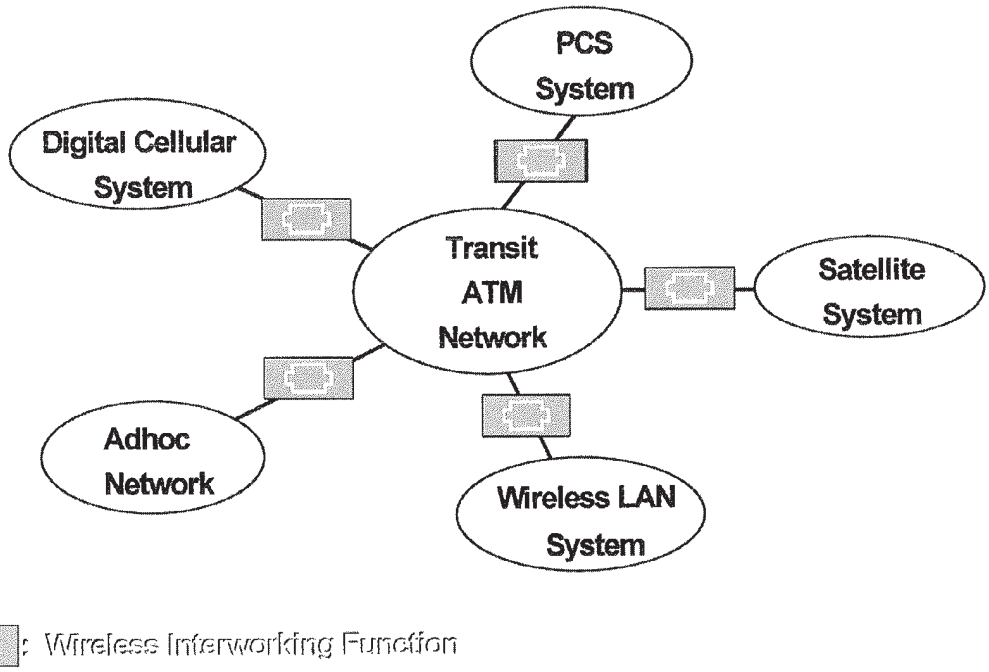


FIGURE 34.3 Various wireless access technologies supported by wireless ATM interworking.

PCS

This class of digital wireless access technologies include the low-tier PCS systems as described in Cox [8]. Digital cellular systems which are becoming known in the U.S. as high-tier PCS, especially when implemented in the 1.8-1.9 GHz PCS frequency bands, are addressed in the Digital Cellular section below. The PCS market has been projected to capture a significant share of the huge potential revenues to be generated by business and residential customers. In order to provide more flexible, widespread, tetherless portable communications than can be provided by today's limited portable communications approaches, low-power exchange access radio needs to be integrated with network intelligence provided by the wireline network.

Today, the network for supporting PCS is narrowband ISDN, along with network intelligence based on advanced intelligent network (AIN) concepts, and a signalling system for mobility/location management and call control based on the signalling system 7 (SS7)

network architecture. It is expected that the core network will evolve to BISDN/ATM over time with the capability to potentially integrate a wide range of network services, both wired and wireless, onto a single network platform. Furthermore, there will be no need for an overlay network for PCS. In anticipation of these developments, the WATM WG included the development of specifications and requirements for PCS access to, and interworking with, the ATM transit network in its charter and work plan. To date, several contributions have been presented at WATM WG meetings that have identified some of the key technical issues relating to PCS-to-ATM interworking. These include (1) architectures, (2) mobility and location management signalling, (3) network evolution strategies, and (4) PCS service scenarios.

Wireless LAN

Today, wireless LAN (WLAN) is a mature technology. WLAN products are frequently used as LAN extensions to access areas of buildings with wiring difficulties and for cross-building interconnect and nomadic access. Coverage ranges from tens to a few hundreds of meters, with data rates ranging from hundreds of kb/s to more than 10 Mb/s. Several products provide 1 or 2 Mb/s. ATM LAN products provide LAN emulation services over the connection-oriented (CO) ATM network using various architectural alternatives [24]. In this case, the ATM network provides services that permit reuse of existing LAN applications by stations attached directly to an ATM switch, and allow interworking with legacy LANs. Furthermore, the increasing importance of mobility in data access networks and the availability of more usable spectrum are expected to speed up the evolution and adoption of mobile access to WLANs. Hence, it is of interest to develop wireless LAN products that have LAN emulation capabilities similar to wireline ATM LANs.

The ETSI BRAN project is developing the HIPERLAN RAL technology for wireless ATM access and interconnection. It will provide short-range wireless access to ATM networks at approximately 25 Mb/s in the 5 GHz frequency band. HIPERLAN is an ATM-based wireless LAN technology that will have end-to-end ATM capabilities. It does not require the development of an IWF to provide access to ATM. A previous version of HIPERLAN (called HIPERLAN I), which has been standardized, supports data rates from 1–23 Mb/s in the 5 GHz band using a non-ATM RAL [27]. This (and other non-ATM HIPERLAN standards being developed) could benefit from interworking with the backbone ATM as a means of extending the marketability of these products in the public domain such as areas of mass transit and commuter terminals.

Several proposals have been submitted to IEEE 802.11 to provide higher speed extensions of current IEEE 802.11 systems operating in the 2.4 GHz region and the development of specifications for new systems operating in the 5 GHz frequency band. The proposed 2.4 GHz extensions support different modulation schemes, but are interoperable with the current IEEE 802.11 low rate PHY and are fully compliant with the IEEE 802.11 defined MAC. The 5 GHz proposals are not interoperable with the current 2.4 GHz IEEE 802.11 systems. One of the final three 5 GHz proposals being considered is based on orthogonal frequency division multiplexing (OFDM), or multicarrier modulation. The other two are single carrier proposals using offset QPSK (OQPSK)/offset QAM (OQAM), and differential pulse position modulation (DPPM). The OFDM proposal has been selected. With 16-QAM modulation on each subcarrier and rate-3/4 convolutional coding, the OFDM system has a peak data rate capability of 30 Mb/s.

It is clear that a whole range of WLAN systems either exist today, or are emerging, that are not based on ATM technology, and therefore cannot provide seamless access to the ATM infrastructure network. The development of IWF specifications that allow these

WLANs to provide such access through standard network interfaces will extend the range of applications and service features provided by WLANs. In Pahlavan [17], a number of architectural alternatives to interconnect WLAN and WATM to the ATM and/or LAN backbone are discussed, along with service scenarios and market and product issues.

Digital Cellular

Digital cellular mobile radio systems include the 1.8–1.9 GHz (high-tier PCS) and the 800–900 MHz systems that provide high-mobility, wide-area coverage over macrocells. Cellular radio systems at 800–900 MHz have evolved to digital in the form of Global System for Mobile Communications (GSM) in Europe, Personal Digital Cellular (PDC) in Japan, and IS-54 Time Division Multiple Access (TDMA) and IS-95 Code Division Multiple Access (CDMA) in the U.S. The capabilities in place today for roaming between cellular networks provide for even wider coverage. Cellular networks have become widespread, with coverage extending beyond some national boundaries. These systems integrate wireless access with large-scale mobile networks having sophisticated intelligence to manage mobility of users.

Cellular networks (e.g., GSM) and ATM networks are evolving somewhat independently. The development of IWFs that allow digital cellular systems to utilize ATM transport will help to bridge the gap. Cellular networks have sophisticated mechanisms for authentication and handoff, and support for rerouting through the home network. In order to facilitate the migration of cellular systems to ATM transport, one of the first issues that should be addressed is that of finding ways to enhance the basic mobility functions already performed by them for implementation in WATM. Contributions presented at WATM WG meetings have identified some basic mobility functions of cellular mobile networks (and cordless terminal mobility) which might be adopted and enhanced for WATM interworking. These basic mobility functions include rerouting scenarios (including path extension), location update, call control, and authentication.

Satellite

Satellite systems are among the primary means of establishing connectivity to untethered nodes for long-haul radio links. This class of applications has been recognized as an essential component of the National Information Infrastructure (NII) [16]. Several compelling reasons have been presented in the ATM Forum's WATM WG for developing standard network interfaces for **satellite ATM (SATATM)** networks. These include (1) ubiquitous wide area coverage, (2) topology flexibility, (3) inherent point-to-multipoint and broadcast capability, and (4) heavy reliance by the military on this mode of communications. Although the geostationary satellite link represents only a fraction of satellite systems today, WATM WG contributions that have addressed this interworking option have focused primarily on geostationary satellites. Some of these contributions have also proposed the development of WATM specifications for SATATM systems having end-to-end ATM capabilities.

Interoperability problems between satellite systems and ATM networks could manifest themselves in at least four ways.

1. Satellite links operate at much higher bit error rates (BERs) with variable error rates and bursty errors.
2. The approximately 540 ms round trip delay for geosynchronous satellite communications can potentially have adverse impacts on ATM traffic and congestion control procedures.

3. Satellite communications bandwidth is a limited resource, and might be incompatible with less bandwidth efficient ATM protocols.
4. The high availability rates (at required BERs) for delivery of ATM (e.g., 99.95%) is costly, hence the need to compromise between performance relating to availability levels and cost.

A number of experiments have been performed to gain insights into these challenges [21]. The results can be used to guide the development of WATM specifications for the satellite interworking scenario. Among the work items that have been proposed for SATATM access using geostationary satellites links are (1) identification of requirements for RAL and mobile ATM functions; (2) study of the impact of satellite delay on traffic management and congestion control procedures; (3) development of requirements and specifications for bandwidth efficient operation of ATM speech over satellite links; (4) investigation of various WATM access scenarios; and (5) investigation of frequency spectrum availability issues.

One interesting SATATM application scenario has been proposed to provide ATM services to multiuser airborne platforms via satellite links for military and commercial applications. In this scenario, a satellite constellation is assumed to provide contiguous overlapping coverage regions along the flight path of the airborne platforms. A set of interworked ground stations form the mobile enhanced ATM network that provides connectivity between airborne platforms and the fixed terrestrial ATM network via bent-pipe satellite links. Key WATM requirements for this scenario have been proposed, which envisage among other things modifications to existing PNNI signalling and routing mechanisms to allow for mobility of (ATM) switches.

In related work, the Telecommunications Industry Association (TIA) TR34.1 WG has also proposed to develop technical specifications for SATATM networks. Three ATM network architectures are proposed for bent-pipe satellites and three others for satellites with onboard ATM switches [23]. Among the technical issues that are likely to be addressed by TR34.1 are protocol reference models and architecture specifications, RAL specifications for SATATM, and support for routing, rerouting, and handoff of active connections. A liaison has been established between the ATM Forum and TR34.1, which is likely to lead to the WATM WG working closely with TR34.1 to develop certain aspects of the TR34.1 SATATM specifications.

Ad Hoc Networks

The term ad hoc network is used to characterize wireless networks that do not have dedicated terminals to perform traditional BS and/or wireless resource control functions. Instead, any mobile terminal (or a subset of mobile terminals) can be configured to perform these functions at any time. Ad hoc networking topologies have been investigated by wireless LAN designers, and are part of the HIPERLAN and IEEE 802.11 wireless LAN specifications [13]. As far as its application to WATM is concerned, low-cost, plug-and-play, and flexibility of system architecture are essential requirements. Potential application service categories include rapidly deployable networks for government use (e.g., military tactical networks, rescue missions in times of natural disasters, law enforcement operations, etc.), ad hoc business conferencing devoid of any dedicated coordinating device, and ad hoc residential network for transfer of information between compatible home appliances.

There are some unique interworking features inherent in ad hoc networks. For example, there is a need for location management functions not only to identify the location of terminals but also to identify the current mode of operation of such terminals. Hence, the WATM WG is considering proposals to develop separate requirements for ad hoc networks independent of the underlying wireless access technology. It is likely that requirements will

be developed for an ad hoc RAL, mobility management signalling functions, and location management functions for supporting interworking of ad hoc networks that provide access to ATM infrastructure networks.

The range of potential wireless access service features and wireless interworking scenarios presented in the five wireless access technologies discussed above is quite large. For example, unlicensed satellite systems could provide 32 kb/s voice, and perhaps up to 10 Mb/s for wireless data services. The main problem centers around the feasibility of developing specifications for IWFs to accommodate the range of applications and service features associated with the wireless access options shown in Fig. 34.3. The WATM WG might consider forming sub-working groups which would work in parallel to develop the network interface specifications for each (or a subset of each) of the above non-ATM wireless access options.

34.4 The PCS-to-ATM Interworking Scenario

This section presents a more detailed view of potential near-term and longer-term architectures and reference models for PCS-to-ATM interworking. A signalling link evolution strategy to support mobility is also described. Mobility and location management signalling starts with the current CCS network, which is based on the SS7 protocol and 56 kb/s signalling links, and eventually migrates to ATM signalling. The system level architectures, and the mobility and location management signalling issues addressed in this section serve to illustrate the range of technical issues that need to be addressed for the other WATM interworking options.

34.4.1 Architecture and Reference Model

The near-term approach for the PCS-to-ATM interworking scenario is shown in Fig. 34.4. The near-term approach targets existing PCS providers with network infrastructures in place, who wish to continue to use them while upgrading specific network elements (e.g., MSCs) to provide ATM transport for user data. The existing MSCs in the PCS network are upgraded to include fixed ATM interfaces. ATM is used for transport and switching of user data, while mobility/location management and call control signalling is carried by the SS7 network. No mobility support is required in the ATM network. Synchronization problems might develop because different traffic types relating to the same call may traverse the narrowband SS7 network and the broadband ATM network and arrive at the destination in an uncoordinated manner. Upgrading the SS7 network to broadband SS7 (e.g., T1 speeds or higher) should partially alleviate this potential problem.

A longer-term approach for PCS-to-ATM interworking has been proposed in some technical contributions to the WATM WG. This is illustrated in Fig. 34.5, together with the protocol stacks for both data and signalling. The ATM UNI is placed at the BSC, which acts as the PCS-to-ATM gateway. The ATM network carries both data and signalling. ATM cells are not transmitted over the PCS link. Communications between the BS and BSC are specific to the PCS access network, and could be a proprietary interface. Compared with existing/emerging BSCs, additional protocol layer functionality is required in the BSC to provide (1) transfer/translation and/or encapsulation of PCS **protocol data units (PDUs),** (2) ATM to wireless PDU conversion, and (3) a limited amount of ATM multiplexing/demultiplexing capabilities.

The BSC is connected to the ATM network through a ME-ATM access switch instead of a MSC. The ME-ATM switch provides switching and signalling protocol functions to support ATM connections together with mobility and location management. These functions could

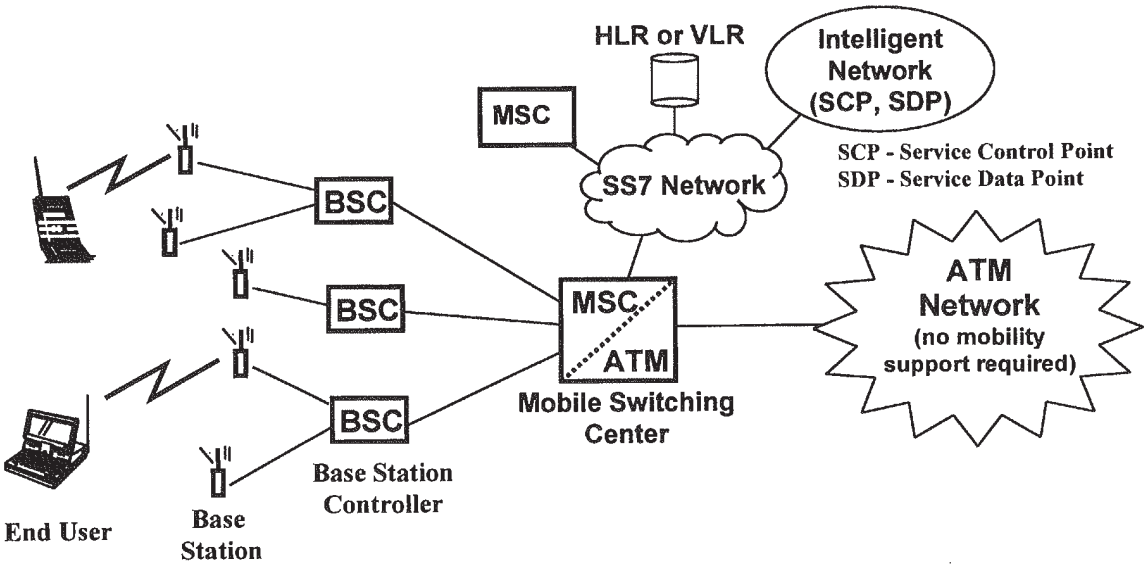


FIGURE 34.4 A near-term architecture for PCS-to-ATM interworking.

be implemented in servers that are physically separate from, but logically connected to, the ME-ATM switch. The WATM WG is expected to formulate requirements and specifications for these mobile-specific functions. On the other hand, an IWF can be introduced between the BSC and the ME-ATM switch shown in Fig. 34.5. In this case, the UNI is between the IWF and the ME-ATM switch, and another standard interface (not necessarily ATM) can be used to connect the BSC to the IWF. The BSC then requires no modification, but a new entity (i.e., the IWF) is required. The IWF will perform protocol conversion and it may serve multiple BSCs.

A unique feature of this architecture is that modifications to network entities to allow for interworking of PCS with the transit ATM network are only required at the edges of the ATM and wireless access networks, i.e., to the BSC and the ME-ATM access switch. In order to minimize the technical impact of mobility on existing/emerging transit ATM networks and PCS specifications, an initial interworking scenario is envisaged in which there are no (or only very limited) interactions between PCS and ATM signalling entities. PCS signalling would be maintained over the air interface, traditional ATM signalling would be carried in the control plane (C-Plane), and PCS signalling would be carried over the ATM network as user traffic in the user plane (U-Plane). In the long term, this architecture is expected to eventually evolve to an end-to-end ATM capability.

On the ATM network side, mobility and location management signalling is implemented in the user plane as the **mobile application part (MAP)** layer above the ATM adaptation layer (AAL), e.g., AAL5. The MAP can be based on the MAP defined in the existing PCS standards (e.g., IS41-MAP or GSM-MAP), or based on a new set of mobile ATM protocols. The U-plane is logically divided into two parts, one for handling mobility and location management signalling messages and the other for handling traditional user data. This obviates the need to make modifications to the ATM UNI and NNI signalling protocols, currently being standardized. The MAP layer also provides the necessary end-to-end reliability management for the signalling because it is implemented in the U-Plane, where reliable communication is not provided in the lower layers of the ATM protocol stack as is done in the signalling AAL (SAAL) layer. The MAP functionality can be distributed at

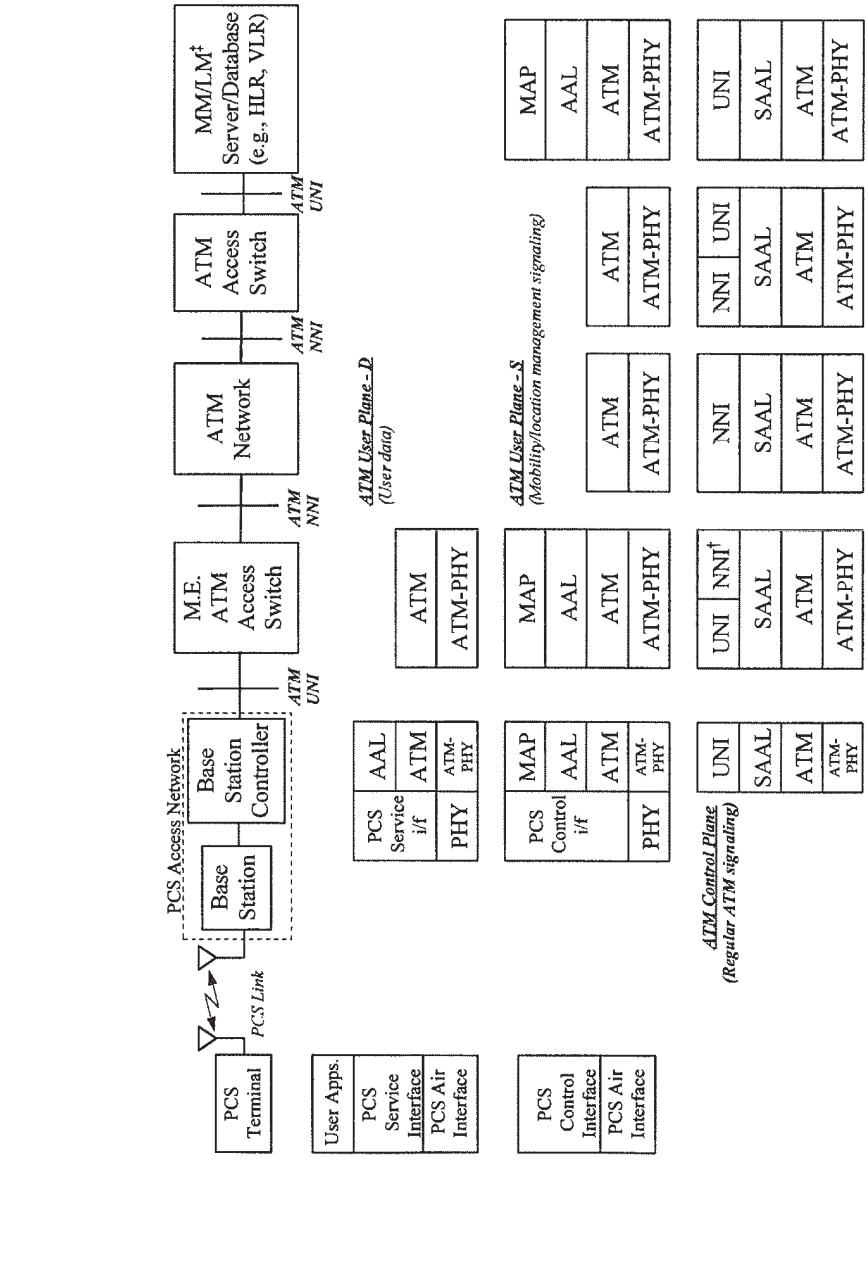


FIGURE 34.5 A longer-term architecture and reference model PCS-to-ATM interworking.
† NNI here is used in a generic sense, to refer to both public and private networks. In public networks, there will be an additional MTP3 layer between the NNI and SAAL layers.
‡ MM—Mobility Management, LM—Location Management.

individual BSCs or ME-ATM access switches or centralized in a separate server to further reduce modifications to existing implementations. The setup and release of ATM connections are still handled by the existing ATM signalling layer. This approach allows easy evolution to future wireless ATM architectures, which will eventually integrate the ATM and PCS network segments to form a homogeneous network to carry both mobility/location management and traditional ATM signalling in the C-Plane. Issues relating to mobility support for PCS interworking with ATM networks are addressed in more detail in Cheng [7].

34.4.2 Signalling Link Evolution

Signalling will play a crucial role in supporting end-to-end connections (without location restrictions) in an integrated WATM network infrastructure. The signalling and control message exchanges required to support mobility will occur more frequently than in wireline ATM networks. Today's CCS network, which is largely based on the SS7 protocol and 56 kb/s signalling links, will not be able to support the long-term stringent service and reliability requirements of WATM. Two broad deployment alternatives have been proposed for migrating the CCS/SS7 network (for wireline services) to using the broadband signalling platform [22].

- Migration to high-speed signalling links using the narrowband signalling platform supported by current digital transmission (e.g., DS1) facilities. The intent of this approach is to support migration to high-speed signalling links using the existing CCS infrastructure with modified signalling network elements. This would allow the introduction of high-speed (e.g., 1.5 Mb/s) links with possible minimal changes in the CCS network. One option calls for modifications of existing MTP2 procedures while maintaining the current protocol layer structure. Another option replaces MTP2 with some functionality of the ATM/SAAL link layer, while continuing to use the same transport infrastructure as DS1 and transport messages over the signalling links in variable length signal units delimited by flags.
- Alternative 2 supports migration of signalling links to a broadband/ATM signalling network architecture. Signalling links use both ATM cells and the ATM SAAL link layer, with signalling message transported over synchronous optical network (SONET) or existing DS1/DS3 facilities at rates of 1.5 Mb/s or higher. This alternative is intended primarily to upgrade the current CCS network elements to support an ATM-based interface, but could also allow for the inclusion of signalling transfer point (STP) functions in the ATM network elements to allow for internetworking between ATM and existing CCS networks.

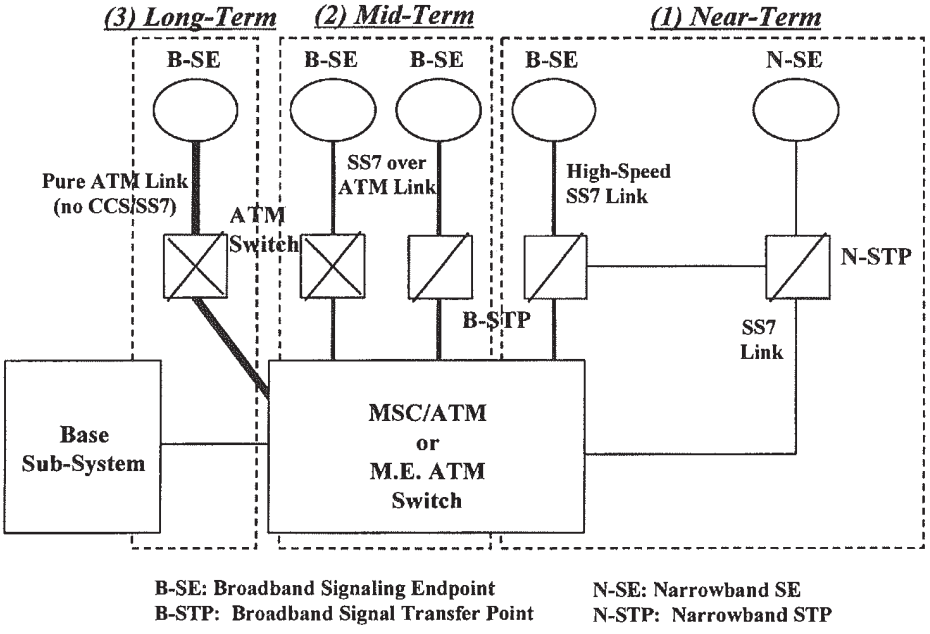
The second alternative provides a good vehicle for the long-term goal of providing high-speed signalling links on a broadband signalling platform supported by the ATM technology. One signalling network configuration and protocol option is the extended Q.93B signalling protocol over ATM in associated mode [22]. The PSTN's CCS networks are currently quasi-associated signalling networks. Q.93B is primarily intended for point-to-point bearer control in user-to-network access, but can also be extended for (link-to-link) switch-to-switch and some network-to-network applications. Standards activities to define high-speed signalling link characteristics in the SS7 protocol have been largely finalized. Standards to support SS7 over ATM are at various stages of completion. These activities provide a good basis for further evolving the SS7 protocol to provide the mobility and location management features that will be required to support PCS (and other wireless systems) access to the ATM network.

The functionality required to support mobility in cellular networks is currently defined as part of the MAP. Both IS-41 and GSM MAPs are being evolved to support PCS services with SS7 as the signalling transport protocol [14]. Quite similar to the two alternatives described above, three architectural alternatives have been proposed for evolving today's IS-41 MAP on SS7 to a future modified (or new) IS-41 on ATM signalling transport platform [28]. They are illustrated in Fig. 34.6. In the first (or near-term) approach, user data is carried over the ATM network, while signalling is carried over existing SS7 links. The existing SS7 network can also be upgraded to broadband SS7 network (e.g., using T1 links) to alleviate some of the capacity and delay constraints in the narrowband SS7 network. This signalling approach can support the near-term PCS-to-ATM interworking described in the previous subsection. In the second (or midterm) approach, a hybrid-mode operation is envisaged, with the introduction of broadband SS7 network elements into the ATM network. This results in an SS7-over-ATM signalling transport platform. Taking advantage of the ATM's switching and routing capabilities, the MTP3 layer could also be modified to utilize these capabilities and eliminate the B-STP functionality from the network. In the third phase (or long-term approach), ATM replaces the SS7 network with a unified network for both signalling and user data. No SS7 functionality exists in this approach. Here, routing of signalling messages is completely determined by the ATM layer, and the MAP may be implemented in a format other than the transaction capability application part (TCAP), so that the unique features of ATM are best utilized. The longer-term PCS-to-ATM interworking approach is best supported for this signalling approach. The performance of several signalling protocols for PCS mobility support for this long-term architecture is presented in Cheng [6, 7]. The above discussion mainly focuses on public networks. In private networks, there are no existing standard signalling networks. Therefore, the third approach can be deployed immediately in private networks.

There are several ways to achieve the third signalling approach in ATM networks. The first approach is to overlay another "network layer" protocol (e.g., IP) over the ATM network, but this requires the management of an extra network. The second approach is to enhance the current ATM network-network interface (e.g., B-ICI for public networks and PNNI for private networks) to handle the new mobility/location management signalling messages and information elements, but this requires modifications in the existing ATM network. The third approach is to use dedicated channels (PVCs or SVCs) between the mobility control signalling points. This does not require any modifications in existing ATM specifications. However, signalling latency may be high in the case of SVCs, and a full mesh of PVCs between all mobility control signalling points is difficult to manage. The fourth approach is to use the generic functional protocols (e.g., connection oriented-bearer independent (CO-BI) or connectionless-BI (CL-BI) transport mechanism) defined in ITU-T's Q.2931.2 [12]. This cannot be done in existing ATM networks without modifications, but these functions are being included in the next version of PNNI (PNNI 2.0) to provide a generic support for supplementary services. There are also technical contributions to the ATM Forum proposing "connectionless ATM" [26], which attempt to route ATM cells in a connectionless manner by using the routing information obtained through the PNNI protocol. However, the "connectionless ATM" concept is still being debated.

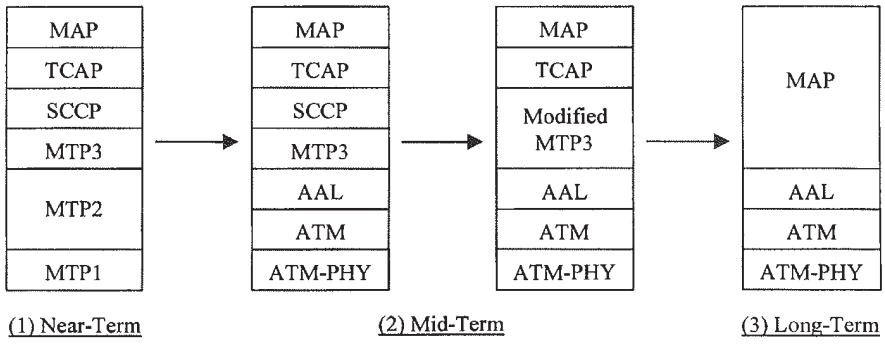
34.5 QoS Support

One immediate impact of adding mobility to an otherwise fixed ATM infrastructure network is the need to manage the changing QoS levels that are inherent in a mobile environment due to the vagaries of the wireless link, the need for rerouting of traffic due to handoff,



Note: A signaling endpoint is the mobility control at a mobile switch, server or database.

(a) Architecture



TCAP – Transaction Capabilities Application Part
SCCP – Signaling Connection Control Part

(b) Protocol stack

FIGURE 34.6 Signalling link evolution for mobility/location management over a public ATM network.

and available bandwidth, etc. Dynamic QoS negotiation and flow control will be required to flexibly support QoS guarantees for multimedia service applications that are likely to be encountered in this environment. QoS provisioning is based on the notion that the wireless channel is likely to demand more stringent measures than the fixed ATM network to support end-to-end QoS. QoS metrics include (1) throughput, (2) delay sensitivity, (3) loss sensitivity, and, (4) BER performance. Issues relating to end-to-end QoS provisioning in multimedia wireless networks are discussed in Naghshineh [15] and articles therein. Here, the focus is on BER maintenance in the context of PCS-to-ATM interworking using forward error correction (FEC) at the radio PHY layer. This is the first step towards developing a hybrid automatic repeat request (ARQ)/FEC protocol for error control of the wireless link, with FEC at the PHY supplemented by ARQ at the DLC layer. A comparison of commonly used FEC and ARQ techniques and their potential application to WATM is presented in Ayanoglu [3].

One adaptive FEC coding scheme that has been proposed for PCS-to-ATM interworking is based on the use of **rate-compatible punctured convolution (RCPC),** punctured Bose-Chaudhuri-Hocquenghem (BCH) or Reed-Solomon (RS) coding at the wireless PHY layer to provide unequal error protection of the wireless PDU [5]. These coding schemes can support a broad range of QoS levels consistent with the requirements of multimedia services, minimize the loss of information on the wireless access segment, and prevent misrouting of cells on the fixed ATM network. Code rate puncturing is a procedure used to periodically discard a set of predetermined coded bits from the sequence generated by an encoder for the purposes of constructing a higher rate code. With the rate-compatibility restriction, higher rate codes are embedded in the lower rate codes, allowing for continuous code rate variation within a data frame.

An example set of three wireless PDU formats that might be appropriate for a PCS system that provides access to the ATM network is shown in Table 34.1. It is desirable to establish a tight relationship between the wireless PDU and wireline ATM cell to minimize incompatibilities between them. This will reduce the complexity of the IWF at the PCS-to-ATM gateway by limiting the amount of processing required for protocol conversion. The wireless PDU can be tightly coupled to the wireline ATM cell in two ways.

1. Encapsulate the equivalent of a full 48-byte ATM information payload, along with wireless-specific overhead (and a full/compressed ATM header, if required by the mobile terminal), in the wireless PDU (e.g., PDU-3).
2. Transmit a submultiple of 48-byte ATM information payload, along with wireless-specific overhead (and a compressed ATM header, if required by the mobile terminal), in the wireless PDU (e.g., PDU-1 and PDU-2).

If the second option is used, then the network has to decide whether to send partially filled ATM cells over the ATM infrastructure network, or wait until enough wireless PDUs arrive to fill an ATM cell.

The performance of RCPC and punctured BCH (and RS) codes have been evaluated in terms of the decoded bit BER [5] on a Rayleigh fading channel. For RCPC coding, the Viterbi upper bound on the decoded BER is given by:

$$P_b \leq \frac{1}{P} \sum_{d=d_f}^{\infty} \beta_d P_d \tag{34.1}$$

where P is the puncturing period, β_d is the weight coefficient of paths having distance d, P_d is the probability of selecting the wrong path, and d_f is the free distance of the RCPC code.

TABLE 34.1 Examples of Wireless Protocol Data Unit (PDU) Formats

PDU type	PDU header (bits)	Information payload (bits)	PDU trailer (bits)	PDU size (bits)
PDU-1	24	128	—	152
PDU-2	40	256	8	304
PDU-3	56	384	16	456

Note: Information payloads are limited to submultiples of a 48-byte ATM cell information payload.

Closed-form expressions for P_d are presented in Hagenauer [9] for a flat Rayleigh fading channel, along with tables for determining β_d. Since the code weight structure of BCH codes is known only for a small subset of these codes, an upper BER performance bound is derived in the literature independent of the structure. Assume that for a t-error correcting BCH(n,k) code, a pattern of i channel errors ($i > t$) will cause the decoded word to differ from the correct word in $i + t$ bit positions. For flat Rayleigh fading channel conditions and differential quadrature phase-shift keying (DQPSK) modulation, the decoded BER for the BCH(n,k) code is [18]:

$$P_b = \frac{2}{3} \sum_{i=t+1}^{n} \frac{i+t}{n} (1 - P_s)^{n-1} P_s^i \tag{34.2}$$

where P_s is the raw symbol error rate (SER) on the channel. An upper bound on the decoded BER for a t-error correcting RS(n,k) code with DQPSK signalling on a flat Rayleigh fading channel is similar in form to (34.2). However, P_s should be replaced with the RS-coded digit error rate P_e, and the unit of measure for the data changed from symbols do digits. A transmission system that uses the RS code [from GF(2^m)][2] with M-ary signalling generates $r = m/\log_2 M$ symbols per m-bit digit. For statistically independent symbol errors, $P_e = 1 - (1 - P_s)^r$. For the binary BCH(n,k) code, $n = 2^m - 1$.

Table 34.2 shows performance results for RCPC and punctured BCH codes that are used to provide adaptive FEC on the PHY layer of a simulated TDMA-based PCS system that accesses the fixed ATM network. The FEC codes provide different levels of error protection for the header, information payload, and trailer of the wireless PDU. This is particularly useful when the wireless PDU header contains information required to route cells in the fixed ATM network, for example. Using a higher level of protection for the wireless PDU header increases the likelihood that the PDU will reach its destination, and not be misrouted in the ATM network.

The numerical results show achievable average code rates for the three example PDU formats in Table 34.1. The PCS system operates in a microcellular environment at 2 GHz, with a transmission bit rate of 384 kb/s, using DQPSK modulation. The channel is modeled as a time-correlated Rayleigh fading channel. The PCS transmission model assumes perfect symbol and frame synchronization, as well as perfect frequency tracking. Computed code rates are shown with and without the use of diversity combining. All overhead blocks in the wireless PDUs are assumed to require a target BER of 10^{-9}. On the other hand, the information payload has target BERs of 10^{-3} and 10^{-6}, which might be typical for voice

[2] GF(2^m) denotes the Galois Field of real numbers from which the RS code is constructed. Multiplication and addition of elements in this field are based on modulo-2 arithmetic, and each RS code word consists of m bits.

TABLE 34.2 Average Code Rates for RCPC Coding[a] and Punctured BCH Coding[b] for a 2-GHz TDMA-Based PCS System with Access to an ATM Infrastructure Network

Coding scheme	Average code rate (10^{-3} BER for information payload)			Average code rate (10^{-6} BER for information payload)		
	PDU-1	PDU-2	PDU-3	PDU-1	PDU-2	PDU-3
RCPC: no diversity	0.76	0.77	0.78	0.61	0.62	0.63
RCPC: 2-branch diversity	0.91	0.93	0.93	0.83	0.84	0.85
Punctured BCH: no diversity	0.42	0.44	0.51	0.34	0.39	0.46
Punctured BCH: 2-branch Diversity	0.83	0.81	0.87	0.75	0.76	0.82

[a] with soft decision decoding and no channel state information at the receiver.

[b] with and without 2-branch frequency diversity.

and data, respectively. Associated with this target BERs is a design goal of 20 dB for the SNR. The mother code rate for the RCPC code is $R = 1/3$, the puncturing period $P = 8$, and the memory length $M = 6$. For BCH coding, the parameter $m \geq 8$.

The numerical results in Table 34.2 show the utility of using code rate puncturing to improve the QoS performance of the wireless access segment. The results for punctured RS coding are quite similar to those for punctured BCH coding. Adaptive PHY layer FEC coding can be further enhanced by implementing an ARQ scheme at the DLC sublayer, which is combined with FEC to form a hybrid ARQ/FEC protocol [9] to supplement FEC at the PHY layer. This approach allows adaptive FEC to be distributed between the wireless PHY and DLC layers.

34.6 Conclusions

The ATM Forum is developing specifications intended to facilitate the use of ATM technology for a broad range of wireless network access and interworking scenarios, both public and private. These specifications are intended to cover requirements for seamless extension of ATM to mobile devices and mobility control in ATM infrastructure networks to allow interworking of non-ATM wireless terminals with the fixed ATM network. A mobility-enhanced ATM network that is developed from specifications for WATM interworking may be used in near-term cellular/PCS/satellite/wireless LAN deployments, while providing a smooth migration path to the longer-term end-to-end WATM application scenario. It is likely to be cost-competitive with other approaches that adopt non-ATM overlay transport network topologies.

This chapter describes various WATM interworking scenarios where the ATM infrastructure might be a public (or private) transit ATM network, designed primarily to support broadband wireline services. A detailed description of a generic PCS-to-ATM architectural interworking scenario is presented, along with an evolution strategy to eventually provide end-to-end WATM capabilities in the long term. One approach is described for providing QoS support using code rate puncturing at the wireless PHY layer, along with numerical results. The network architectures, protocol reference models, signalling protocols, and QoS management strategies described for PCS-to-ATM interworking can be applied, in varying degrees, to the other WATM interworking scenarios described in this chapter.

Defining Terms

Broadband integrated services digital network (BISDN): A cell-relay-based information transfer technology upon which the next-generation telecommunications infrastructure is to be based.

BISDN intercarrier interface (B-ICI): A carrier-to-carrier public interface that supports multiplexing of different services such as SMDS, frame relay, circuit emulation, and cell relay services.

High-Performance Radio LAN (HIPERLAN): Family of standards being developed by the European Telecommunications Standards Institute (ETSI) for high-speed wireless LANs, to provide short-range and remote wireless access to ATM networks and for wireless ATM interconnection.

Interworking functions (IWFs): A set of network functional entities that provide interaction between dissimilar subnetworks, end systems, or parts thereof, to support end-to-end communications.

Location management: A set of registration and call delivery functions.

Mobile application part (MAP): Application layer protocols and processes that are defined to support mobility services such as intersystem roaming and handoffs.

Mobility-enabled ATM (ME-ATM): An ATM switch with additional capabilities and features to support location and mobility management.

Mobility management: The handoff process associated with switching (rerouting) of the communication link from the old coverage area to the new coverage area when a wireless user moves during active communication.

Personal communications services (PCS): Emerging digital wireless systems which support mobility in microcellular and other environments, which have a set of capabilities that allow some combination of terminal mobility, personal mobility, and service profile management.

Protocol data units (PDUs): The physical layer message structure used to carry information across the communications link.

Private network–network interface (PNNI): The interface between two private networks.

User network interface (UNI): A standardized interface providing basic call control functions for subscriber access to the telecommunications network.

Radio access layer (RAL): A reference to the physical, medium access control, and data link control layers of the radio link.

Rate-compatible punctured convolution (RCPC): Periodic discarding of predetermined coded bits from the sequence generated by a convolutional encoder for the purposes of constructing a higher rate code. The rate-compatibility restriction insures that the higher rate codes are embedded in the lower rate codes, allowing for continuous code rate variation to change from low to high error protection within a data frame.

Satellite ATM (SATATM): A satellite network that provides ATM network access to fixed or mobile terminals, high-speed links to interconnect fixed or mobile ATM networks, or form an ATM network in the sky to provide user access and network interconnection services.

Wireless asynchronous transfer mode (WATM): An emerging wireless networking technology that extends ATM over the wireless access segment, and/or uses the ATM infrastructure as a transport network for a broad range of wireless network access scenarios, both public and private.

References

[1] ATM. B-ICI Specification, V 2.0. ATM Forum, 1995.

[2] ATM. Private Network-Network Interface (PNNI) Specification, Version 1.0. ATM Forum, 1996.

[3] Ayanoglu, E., et al., Wireless ATM: Limits, challenges, and protocols. *IEEE Personal Commun.*, 3(4), 18–34, Aug. 1996.

[4] Barton, M., Architecture for wireless ATM networks. *PIMRC'95.* 778–782, Sep. 1995.

[5] Barton, M., Unequal error protection for wireless ATM applications. *GLOBECOM'96,* 1911–1915, Nov. 1996.

[6] Cheng, M., Performance comparison of mobile assisted network controlled, and mobile-controlled hand-off signalling for TDMA PCS with an ATM backbone. *ICC'97.* Jun. 1977a.

[7] Cheng, M., Rajagopalan, S., Chang, L.F., Pollini, G.P., and Barton, M., PCS mobility support over fixed ATM networks. *IEEE Commun. Mag.*, 35(11), 82–92, Nov. 1997b.

[8] Cox, D.C., Wireless personal communications: what is it?. *IEEE Personal Commun. Mag.*, 2(2), 20–35, Apr. 1995.

[9] Hagenauer, J., Rate-compatible punctured convolution codes (RCPC codes) and their applications. *IEEE Trans. Commun.*, COM-36(4), 389–400, Apr. 1988.

[10] ITU. Message Transfer Part Level 3 Functions and Messages Using the Service of ITU-T Recommendations, Q.2140. International Telecommunications Union-Telecommunications Standardization Sector, Geneva, Switzerland. TD PL/11–97. 1995a.

[11] ITU. B-ISDN Adaptation Layer—Service Specific Coordination Function for Signaling at the Network Node Interface (SCCF) at NNI. ITU-T Q.2140. International Telecommunications Union, Telecommunications Standardization Sector, Geneva, Switzerland. 1995b.

[12] ITU. Digital Subscriber Signaling System No. 2—Generic Functional Protocol: Core Functions. ITU-T Recommendation Q.2931.2. International Telecommunications Union, Telecommunications Standardization Sector, Geneva, Switzerland. 1996.

[13] LaMaire, R.O., Krishna, A., Bhagwat, P., and Panian, J., Wireless LANs and mobile networking: standards and future directions. *IEEE Commun. Mag.*, 34(8), 86–94, Aug. 1996.

[14] Lin, Y.B. and Devries, S.K., PCS network signalling using SS7. *IEEE Commun. Mag.*, 2(3), 44–55, Jun. 1995.

[15] Naghshineh, M. and Willebeek-LeMair, M., End-to-end QoS provisioning in multimedia wireless/mobile networks using an adaptive framework. *IEEE Commun. Mag.*, 72–81, Nov. 1997.

[16] NSTC. *Strategic Planning Document—Information and Communications,* National Science and Technology Council. 10, Mar. 1995.

[17] Pahlavan, K., Zahedi, A., and Krishnamurthy, P., Wideband local wireless: wireless LAN and wireless ATM. *IEEE Commun. Mag.*, 34–40, Nov. 1997.

[18] Proakis, J.G., *Digital Communications.* McGraw-Hill, New York. 1989.

[19] Raychaudhuri, D. and Wilson, N.D., ATM-based transport architecture for multiservices wireless personal communication networks. *IEEE J. Select. Areas Commun.*, 12(8), 1401–1414, Oct. 1994.

[20] Raychaudhuri, D., Wireless ATM networks: architecture, system design, and prototyping. *IEEE Personal Commun.*, 3(4), 42–49, Aug. 1996.

[21] Schmidt, W.R., et al., Optimization of ATM and legacy LAN for high speed satellite communications. Transport Protocols for High-Speed Broadband Networks Workshop. *GLOBECOM'96.* Nov. 1996.

[22] SR. Alternatives for Signaling Link Evolution. Bellcore Special Report. *Bellcore SR-NWT-002897.* (1), Feb. 1994.

[23] TIA. TIA/EIA Telecommunications Systems Bulletin (TSB)—91. Satellite ATM Networks: Architectures and Guidelines. Telecommunications Industry Association. TIA/EIA/TSB-91. Apr. 1998.

[24] Truong, H.L. et al., LAN emulation on an ATM network. *IEEE Commun. Mag.,* 70–85, May 1995.

[25] Umehira, M. et al., An ATM wireless system for tetherless multimedia services. *ICUPC'95.* Nov. 1995.

[26] Veeraraghavan, M., Pancha, P., and Eng, K.Y., Connectionless Transport in ATM Networks. ATM Forum Contribution. ATMF/97-0141, 9–14, Feb. 1997.

[27] Wilkinson, T. et al., A report on HIPERLAN standardization. *Intl. J. Wireless Inform. Networks.* 2(2), 99–120, 1995.

[28] Wu, T.H. and Chang, L.F., Architecture for PCS mobility management on ATM transport networks. *ICUPC'95.* 763–768, Nov. 1995.

Further Information

Information supplementing the wireless ATM standards work may be found in the ATM Forum documents relating to the Wireless ATM Working Group's activities (web page `http://www.atmforum.com`). Special issues on wireless ATM have appeared in the August 1996 issue of *IEEE Personal Communications,* the January 1997 issue of the *IEEE Journal on Selected Areas in Communications,* and the November 1997 issue of *IEEE Communications Magazine.*

Reports on proposals for higher speed wireless LAN extensions in the 2.4 GHz and 5 GHz bands can be found at the IEEE 802.11 web site (`http://grouper.ieee.org/groups/802/11/Reports`). Additional information on HIPERLAN and related activities in ETSI BRAN can be obtained from their web site (`http://www.etsi.fr/bran`).

35

Wireless ATM: QoS and Mobility Management

Bala Rajagopalan
NEC C&C Research Laboratories

Daniel Reininger
NEC C&C Research Laboratories

35.1 Introduction 35-1
35.2 QoS in Wireless ATM 35-2
ATM QoS Model • QoS Approach in Wireless ATM • MAC Layer Functions • Network and Application Layer Functions
35.3 Mobility Management in Wireless ATM 35-18
Location Management in Wireless ATM • Network Entities Involved in LM • Location Management Functions and Control Flow • Connection Handover in Wireless ATM
35.4 Summary and Conclusions 35-29
References 35-30

35.1 Introduction

Wireless ATM (WATM) refers to the technology that enables ATM end-system mobility as well as tetherless access to ATM core networks. Wireless ATM has two distinct components: the radio access technology, and enhancements to existing ATM technology to support end-system mobility. The latter component is referred to as "MATM" (mobility-enhanced ATM) and it is independent of the radio access technology. The rationale for wireless ATM has been discussed at length elsewhere [1, 2]. In this chapter, we restrict our discussion to two challenging issues in wireless ATM: the provisioning of ATM quality of service (QoS) for connections that terminate on mobile end systems over a radio link, and the protocols for mobility management in the MATM infrastructure.

Figure 35.1 illustrates the WATM reference model considered in this chapter [3]. "W" UNI in this figure indicates the ATM user-network interface established over the wireless link. "M" NNI refers to the ATM network-node interface supporting mobility management protocols. The figure depicts the scenario where an MATM network has both end system mobility-supporting ATM switches (EMAS) and traditional ATM switches with no mobility support. Thus, one of the features of mobility management protocols in an MATM network is the ability to work transparently over switches that do not implement mobility support.

35.2 QoS in Wireless ATM

The type of QoS guarantees to be provided in wireless ATM systems is debatable [4]. On the one hand, the QoS model for traditional ATM networks is based on fixed terminals and

0-8493-8597-0/99/$0.00+$.50
© 1999 by CRC Press LLC

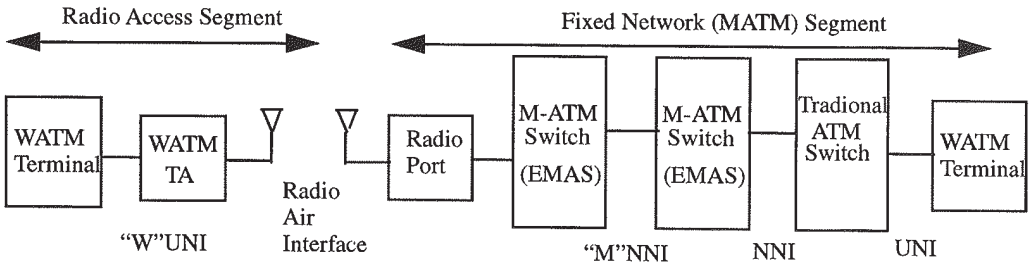


FIGURE 35.1 WATM reference model.

high quality links. Terminal mobility and error-prone wireless links introduce numerous problems [5]. On the other hand, maintaining the existing QoS model allows the transparent extension of fixed ATM applications into the domain of mobile networking. Existing prototype implementations have chosen the latter approach [6]–[8]. This is also the decision of the ATM Forum wireless ATM working group [9]. Our discussion, therefore, is oriented in the same direction, and to this end we first briefly summarize the existing ATM QoS model.

35.2.1 ATM QoS Model

Five service categories have been defined under ATM [10]. These categories are differentiated according to whether they support constant or variable rate traffic, and real-time or non-real-time constraints. The service parameters include a characterization of the traffic and a reservation specification in the form of QoS parameters. Also, traffic is policed to ensure that it conforms to the traffic characterization, and rules are specified for how to treat nonconforming traffic. ATM provides the ability to tag nonconforming cells and specify whether tagged cells are policed (and dropped) or provided with best-effort service.

Under UNI 4.0, the service categories are constant bit rate (CBR), real-time variable bit rate (rt-VBR), non-real-time variable bit rate (nrt-VBR), unspecified bit rate (UBR) and available bit rate (ABR). The definition of these services can be found in [10]. Table 35.1 summarizes the traffic descriptor parameters and QoS parameters relevant to each service category in ATM traffic management specifications version 4.0 [11]. Here, the traffic parameters are peak cell rate (PCR), cell delay variation tolerance (CDVT), sustainable cell rate (SCR), maximum burst size (MBS) and minimum cell rate (MCR). The QoS parameters are cell loss ratio (CLR), maximum cell transfer delay (max CTD) and cell delay variation (CDV). The explanation of these parameters can be found in [11].

Functions related to the implementation of QoS in ATM networks are usage parameter control (UPC) and connection admission control (CAC). In essence, the UPC function (implemented at the network edge) ensures that the traffic generated over a connection conforms to the declared traffic parameters. Excess traffic may be dropped or carried on a best-effort basis (i.e., QoS guarantees do not apply). The CAC function is implemented by each switch in an ATM network to determine whether the QoS requirements of a connection can be satisfied with the available resources. Finally, ATM connections can be either point-to-point or point-to-multipoint. In the former case, the connection is bidirectional, with separate traffic and QoS parameters for each direction, while in the latter case it is unidirectional. In this chapter, we consider only point-to-point connections for the sake of simplicity.

TABLE 35.1 ATM Traffic and QoS Parameters

	ATM Service Category				
Attribute	CBR	rt-VBR	nrt-VBR	UBR	ABR
Traffic Parameters					
PCR and CDVT	Yes	Yes	Yes	Yes	Yes
SCR and MBS	N/A	Yes	Yes	N/A	N/A
MCR	N/A	N/A	N/A	N/A	Yes
QoS Parameters					
CDV	Yes	Yes	No	No	No
Maximum CTD	Yes	Yes	No	No	No
CLR	Yes	Yes	Yes	No	No

35.2.2 QoS Approach in Wireless ATM

QoS in wireless ATM requires a combination of several mechanisms acting in concert. Figure 35.2 illustrates the various points in the system where QoS mechanisms are needed:

- *At the radio interface:* A QoS-capable medium access control (MAC) layer is required. The mechanisms here are resource reservation and allocation for ATM virtual circuits under various service categories, and scheduling to meet delay requirements. Furthermore, an error control function is needed to cope with radio link errors that can otherwise degrade the link quality. Finally, a CAC mechanism is required to limit access to the multiple access radio link in order to maintain QoS for existing connections.
- *In the network:* ATM QoS mechanisms are assumed in the network. In addition, a capability for QoS renegotiation will be useful. This allows the network or the mobile terminal (MT) to renegotiate the connection QoS when the existing connection QoS cannot be maintained during handover. Renegotiation may also be combined with *soft* QoS mechanisms, as described later. Finally, mobility management protocols must include mechanisms to maintain QoS of connections rerouted within the network during handover.
- *At the MT:* The MT implements the complementary functions related to QoS provisioning in the MAC and network layers. In addition, application layer functions may be implemented to deal with variations in the available QoS due to radio link degradation and/or terminal mobility. Similar functions may be implemented in fixed terminals communicating with MTs.

In the following, we consider the QoS mechanisms in some detail. We focus on the MAC layer and the new network layer functions such as QoS renegotiation and soft QoS. The implementation of existing ATM QoS mechanisms have been described in much detail by others (for example, see [12]).

35.2.3 MAC Layer Functions

The radio link in a wireless ATM system is typically a broadcast multiple access channel shared by a number of MTs. Different multiple access technologies are possible, for instance frequency, time, or code division multiple access (FDMA, TDMA, and CDMA, respectively). A combination of FDMA and *dynamic* TDMA is popular in wireless ATM implementations.

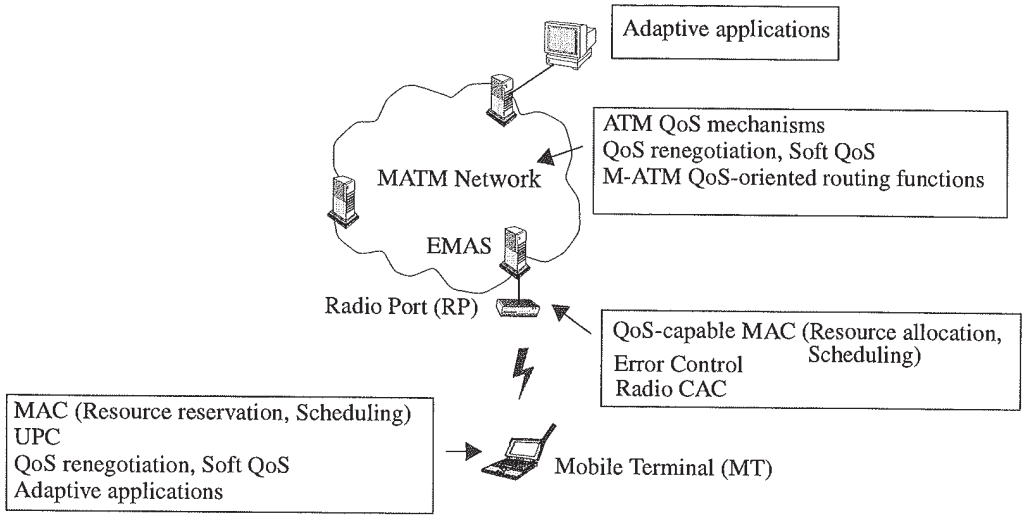


FIGURE 35.2 QoS mechanisms in wireless ATM.

That is, each radio port (RP) operates on a certain frequency band and this bandwidth is shared dynamically among ATM connections terminating on multiple MTs using a TDMA scheme [13]. ATM QoS is achieved under dynamic TDMA using a combination of a resource reservation/allocation scheme and a scheduling mechanism. This is further explained using the example of two wireless ATM implementations: NEC's WATMnet 2.0 prototype [13] and the European Union's Magic WAND (Wireless ATM Network Demonstrator) project [8].

Resource Reservation and Allocation Mechanisms (WATMnet 2.0)

WATMnet utilizes a TDMA/TDD (time division duplexing) scheme for medium access. The logical transmission frame structure under this scheme is shown in Fig. 35.3. As shown, this scheme allows the flexibility to partition the frame dynamically for downlink (from EMAS to MTs) and uplink (from MTs to EMAS) traffic, depending on the traffic load in each direction. Other notable features of this scheme are:

- A significant portion of each slot is used for forward error control (FEC)
- A separate contention region in the frame is used for MTs to communicate with the EMAS
- 8-byte control packets are used for bandwidth request and allocation announcements. An MT can tag request packets along with the WATM cells it sends or in the contention slots
- WATM cells are modified ATM cells with data link control (DLC) and cyclic redundancy check (CRC) information added

In the downlink direction, the WATM cells transported belong to various ATM connections terminating on different MTs. After such cells arrive at the EMAS from the fixed network, the allocation of TDMA slots for specific connections is done at the EMAS based on the connections' traffic and QoS parameters. This procedure is described in the next section. In the uplink direction, the allocation is based on requests from MTs. For bursty traffic, an MT makes a request only after each burst is generated. Once uplink slots are allocated to specific MTs, the transmission of cells from multiple active connections at an MT is again subject to the scheduling scheme. Both the request for uplink slot allocation

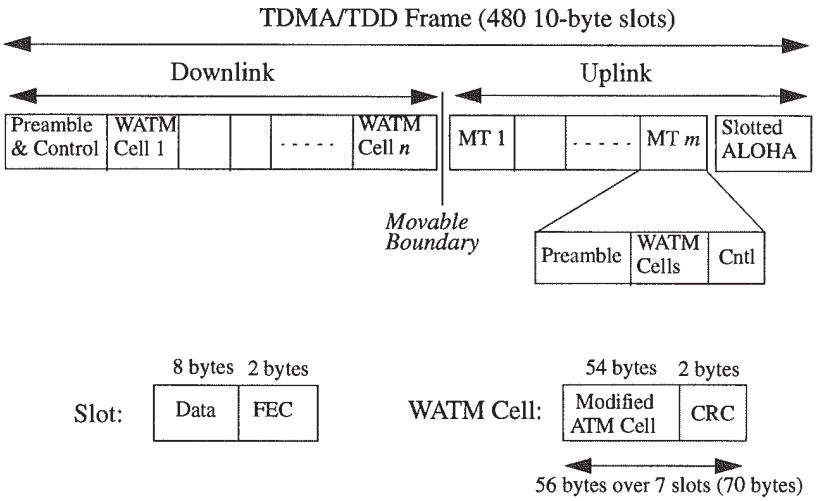


FIGURE 35.3 TDMA logical frame format.

from the MTs and the results from the EMAS are carried in control packets whose format is shown in Fig. 35.4. Here, the numbers indicate the bits allocated for various fields. The sequence number is used to recover from transmission losses. Request and allocation types indicate one of four types: CBR, VBR, ABR or UBR. The allocation packet has a start slot field which indicates where in the frame the MT should start transmission. The number of allocated slots is also indicated.

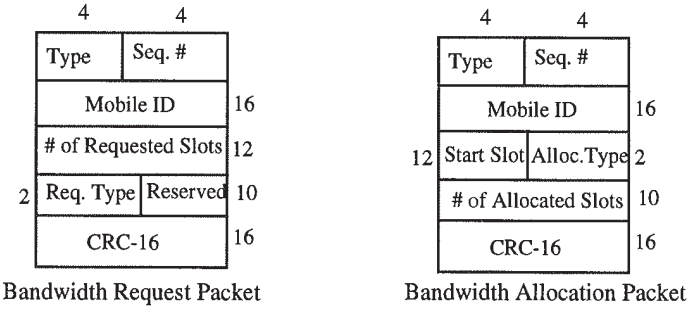


FIGURE 35.4 Bandwidth control packet formats.

The DLC layer implementation in WATMnet is used to reduce the impact of errors that cannot be corrected using the FEC information. The DLC layer is responsible for selective retransmission of cells with uncorrectable errors or lost cells. Furthermore, the DLC layer provides request/reply control interface to the ATM layer to manage the access to the wireless bandwidth, based on the instantaneous amount of traffic to be transmitted. The wireless ATM cell sent over the air interface is a modified version of the standard ATM cell with DLC and CRC information, as shown in Fig. 35.5. The same figure also shows the acknowledgment packet format for implementing selective retransmission. In this packet, the VCI field specifies the ATM connection for which the acknowledgment is being sent.

The sequence number field indicates the beginning sequence number from which the 16-bit acknowledgment bitmap indicates the cells correctly received (a "1" in the bit map indicates correct reception).

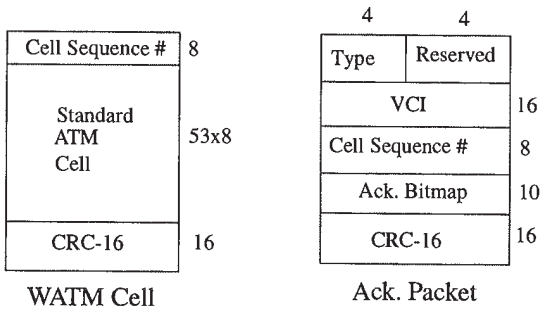


FIGURE 35.5 WATM formats at the DLC layer.

The TDMA/TDD scheme thus provides a mechanism for dynamic bandwidth allocation to multiple ATM connections. How active connections are serviced at the EMAS and the MTs to maintain their QoS needs is another matter. This is described next using Magic WAND as the example.

Scheduling (Magic WAND)

The Magic WAND system utilizes a TDMA scheme similar to that used by the WATM-net. This is shown in Fig. 35.6. Here, each MAC Protocol Data Unit (MPDU) consists of a header and a sequence of WATM cells from the same MT (or the EMAS) referred to as a *cell train*. In the Magic WAND system, the scheduling of both uplink and downlink transmissions is done at the EMAS. Furthermore, the scheduling is based on the fact that the frame length is variable. A simplified description of the scheduling scheme is presented below. More details can be found in [14, 15].

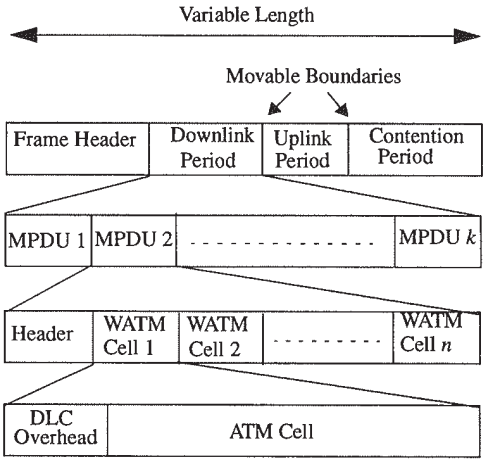


FIGURE 35.6 Magic WAND TDMA frame structure.

At the beginning of each TDMA frame, the scheduling function at the EMAS considers pending transmission requests, uplink and downlink, from active connections. The scheduler addresses two issues:

1. The determination of the number of cells to be transmitted from each connection in the frame, and
2. The transmission sequence of the selected cells within the frame

The objective of the scheduler is to regulate the traffic over the radio interface as per the declared ATM traffic parameters of various connections and to ensure that the delay constraints (if any) are met for these connections over this interface.

The selection of cells for transmission is done based on the service categories of active connections as well as their traffic characteristics. First, for each connection, a priority based on its service category is assigned. CBR connections are assigned the highest priority, followed by rt-VBR, nrt-VBR, ABR. In addition, for each active connection that is not of type UBR, a token pool is implemented. Tokens for a connection are generated at the declared SCR of the connection and tokens may be accumulated in the pool as long as their number does not exceed the declared MBS for the connection. The scheduler services active connections in two passes: in the first pass, connections are considered in priority order from CBR to ABR (UBR is omitted) and within each priority class only connections with a positive number of tokens in their pools are considered. Such connections are serviced in the decreasing order of the number of tokens in their pools. Whenever a cell belonging to a connection is selected for transmission, a token is removed from its pool. At the end of the first pass, either all the slots in the downlink portion of the frame are used up or there are still some slots available. In the latter case, the second pass is started. In this pass, the scheduler services remaining excess traffic in each of CBR, rt and nrt-VBR and ABR classes, and UBR traffic in the priority order. It is clear that in order to adequately service active connections, the mean bandwidth requirement of the connections cannot exceed the number of downlink slots available in each frame. The CAC function is used to block the setting up of new connections over a radio interface when there is a danger of overloading. Another factor that can result in overloading is the handover of connections. The CAC must have some knowledge of the expected load due to handovers so that it can limit new connection admissions. Preserving the QoS for handed over connections while not degrading existing connections at a radio interface requires good network engineering. In addition, mechanisms such as QoS renegotiation and soft QoS (Section 35.2.4) may be helpful.

Now, at the end of the selection phase, the scheduler has determined the number of cells to be transmitted from each active connection. Some of these cells are to be transmitted uplink while the others are to be transmitted downlink. The scheduler attempts to place a cell for transmission within the frame such that the cell falls in the appropriate portion of the frame (uplink or downlink, Fig. 35.6) and the delay constraint (CDT) of the corresponding connection is met. To do this, first the delay allowed over the radio segment is determined for each connection with a delay constraint (it is assumed that this value can be obtained during the connection routing phase by decomposing the path delay into delays for each hop). Then, for downlink cells, the arrival time for the cell from the fixed network is marked. For uplink cells, the arrival time is estimated from the time at which the request was received from the MT. The deadline for the transmission of a cell (uplink or downlink) is computed as the arrival time plus the delay allowed over the radio link.

The final placement of the cells in the frame is based on a three-step process, as illustrated by an example with six connections (Fig. 35.7). Here, Dn and Un indicate downlink and uplink cells with deadline = slot n, respectively, and D_j^i and U_j^i indicate downlink and

uplink cells of the ith connection with deadline = slot j, respectively. In the first step, the cells are ordered based on their deadlines [Fig. 35.7(a)]. Several cells belonging to the same connection may have been selected for transmission in a frame. When assigning a slot for the first cell of such a "cell train" the scheduler positions the cell such that its transmission will be before and as close to its deadline as possible. Some cells may have to be shifted from their previously allocated slots to make room for the new allocation. This is done only if the action does not violate the deadline of any cell. When assigning a slot for another cell in the train, the scheduler attempts to place it in the slot next to the one allocated to the previous cell. This may require shifting of existing allocations as before. This is shown in Figs. 35.7(b)–35.7(d). Here, the transmission frame is assumed to begin at slot 5.

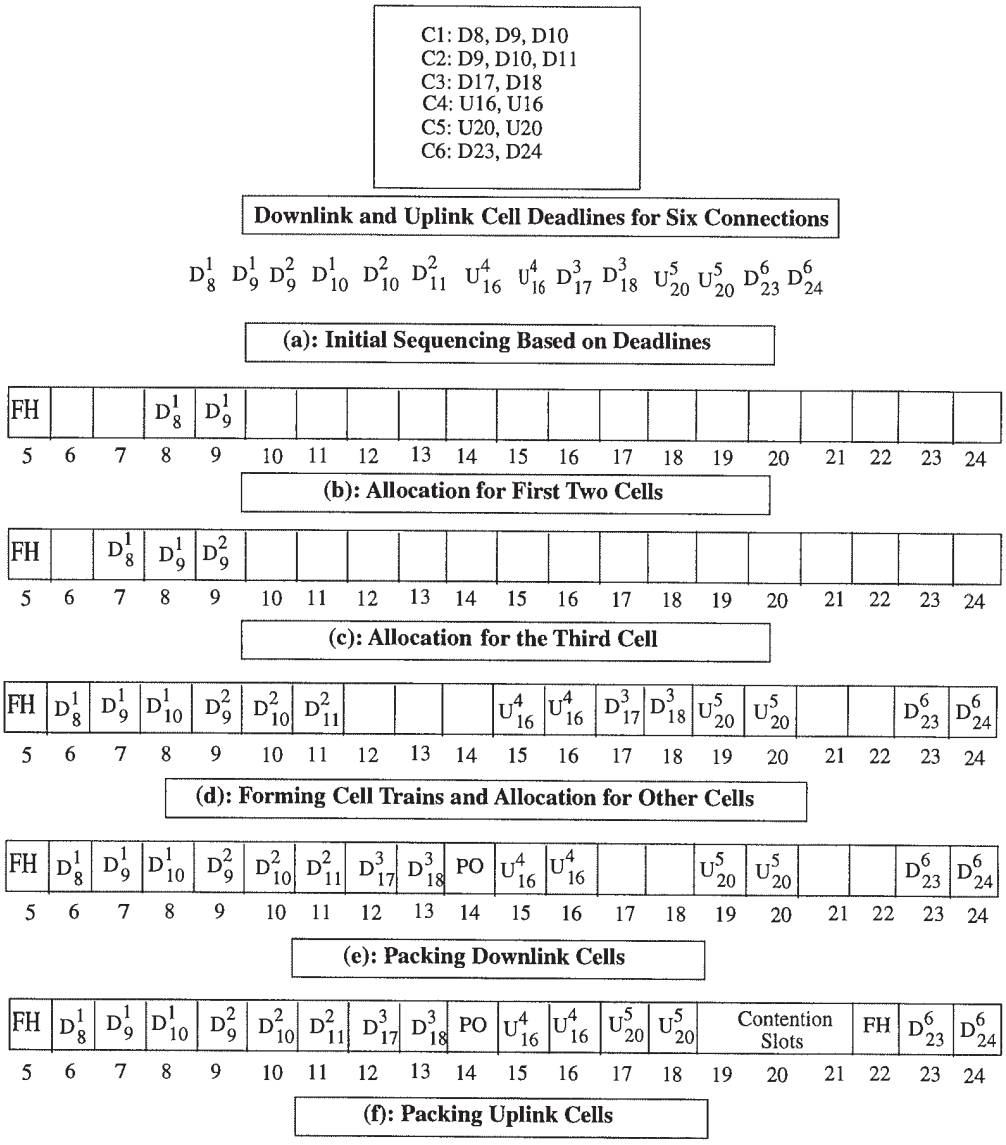


FIGURE 35.7 Scheduling example.

At the end of the first step, the cell sequencing may be such that uplink and downlink cells may be interleaved. The second step builds the downlink portion of the frame by first shifting all downlink cells occurring before the first uplink cell as close to the beginning of the frame as possible. In the space between the last such downlink cell and the first uplink cell, as many downlink cells as possible are packed. This is illustrated in Fig. 35.7(e). A slot for *period overhead* (PO) is added between the downlink and uplink portions. Finally, in the last step, the uplink cells are packed, by moving all uplink cells occurring before the next downlink cell as shown in Fig. 35.7(f). The contention slots are added after the last uplink cell and the remaining cells are left for the next frame.

Thus, scheduling can be a rather complicated function. The specific scheduling scheme used in the Magic WAND system relies on the fact that the frame length is variable. Scheduling schemes for other frame structures could be different.

35.2.4 Network and Application Layer Functions

Wireless broadband access is subject to sudden variations in bandwidth availability due to the dynamic nature of the service demand (e.g., terminals moving in and out of RPs coverage area, variable bit-rate interactive multimedia connections) and the natural constraints of the physical channel (e.g, fading and other propagation conditions). QoS control mechanisms should be able to handle efficiently both the mobility and the heterogeneous and dynamic bandwidth needs of multimedia applications. In addition, multimedia applications themselves should be able to adapt to terminal heterogeneity, computing limitations, and varying availability of network resources [16].

In this section, the network and application layer QoS control functions are examined in the context of NECs WATMnet system [6]. In this system, a concept called *soft-QoS* is used to effectively support terminal mobility, maintain acceptable application performance and high network capacity utilization. Soft-QoS relies on a QoS control framework that permits the allocation of network resources to dynamically match the varying demands of mobile multimedia applications. Network mechanisms under this framework for connection admission, QoS renegotiation, and handoff control are described. The soft-QoS concept and the realization of the soft-QoS controller based on this concept are described next. Finally, experimental results on the impact of network utilization and soft-QoS provisioning for video applications are discussed.

A Dynamic Framework for QoS Control

The bit-rates of multimedia applications vary significantly among sessions and within a session due to user interactivity and traffic characteristics. Contributing factors include the presence of heterogeneous media (e.g., video, audio, and images) compression schemes (e.g., MPEG, JPEG), presentation quality requirements (e.g., quantization, display size), and session interactivity (e.g., image scaling, VCR-like control). Consider, for example, a multimedia application using several media components or media objects (such as a multi-window multimedia user interface or future MPEG-4 encoded video) which allows users to vary the relative importance-of-presence (IoP) of a given media object to match the current viewing priorities. In this case there would be a strong dependency of user/application interaction on the bandwidth requirements of individual media components. Figure 35.8 shows the bit-rate when the user changes the video level-of-detail (LoD) during a session. A suitable network service for these applications should support bandwidth renegotiation to simultaneously achieve high network utilization and maintain acceptable performance. For this purpose, an efficient network service model should support traffic contract renegotia-

tion during a session. It has been experimentally verified that bandwidth renegotiation is key for efficient QoS support of network-aware adaptive multimedia applications, and that the added implementation complexity is reasonable [17, 18].

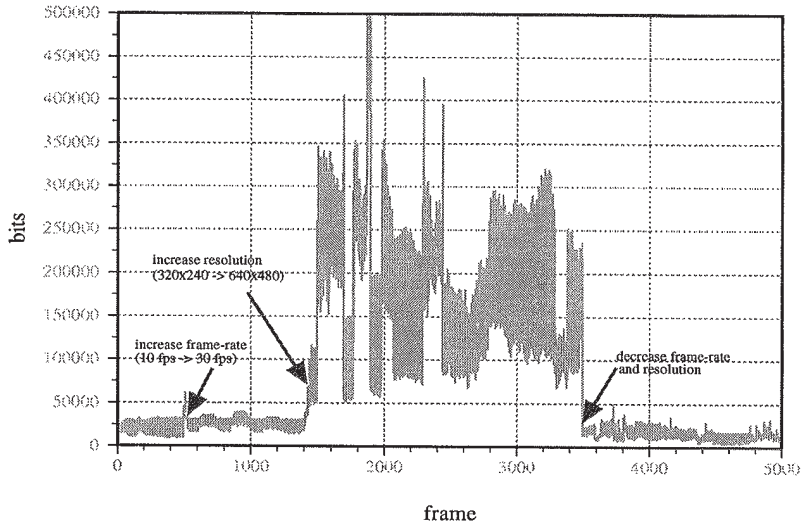


FIGURE 35.8 Video bit-rate changes on an interactive multimedia session.

In the mobile multimedia communication scenario, bandwidth renegotiation is particularly important. Conventional network services use static bandwidth allocation models that lack the flexibility needed to cope with multimedia interactivity and session mobility. These session properties enlarge the dynamic range of bandwidth requirements and make dynamic bandwidth management protocols a requirement for effective end-to-end QoS support. Renegotiation may be required during handover, as well as when resource allocation changes are warranted due to instantaneous application needs and sudden changes in network resource availability.

Figure 35.9 shows the system and API model for QoS control with bandwidth renegotiation. The application programming interface (API) between the adaptive application and the QoS control module is dynamic, i.e., its parameters can be modified during the session. For example, the Winsock 2 API under Microsoft Windows [19] allows the dynamic specification of QoS parameters suitable for the application. In addition, the API between the QoS controller and the network allows the traffic descriptor to be varied to track the bit-rate requirements of the bitstream. A new network service, called VBR$^+$ allows renegotiation of traffic descriptors between the network elements and the terminals [20]. VBR$^+$ allows multimedia applications to request "bandwidth-on-demand" suitable for their needs.

Soft-QoS Model

Although multimedia applications have a wide range of bandwidth requirements, most can gracefully adapt to sporadic network congestion while still providing acceptable performance. This graceful adaptation can be quantified by a softness profile [17]. Figure 35.10 shows the characteristics of a softness profile. The softness profile is a function defined on the scales of two parameters: satisfaction index and bandwidth ratio. The satisfaction index is based on the subjective mean-opinion-score (MOS), graded from 1 to 5; a minimum

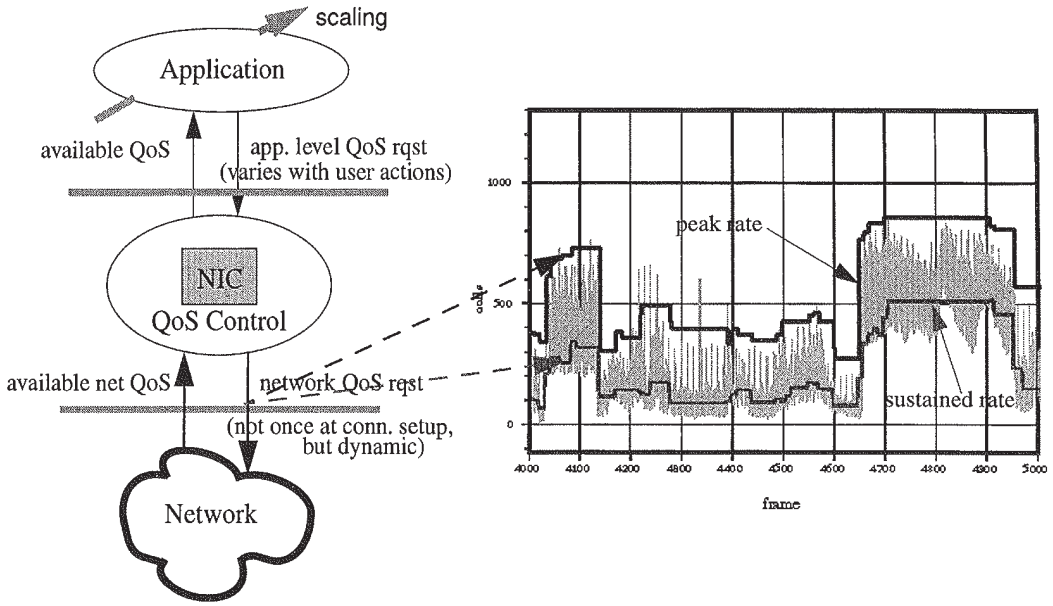
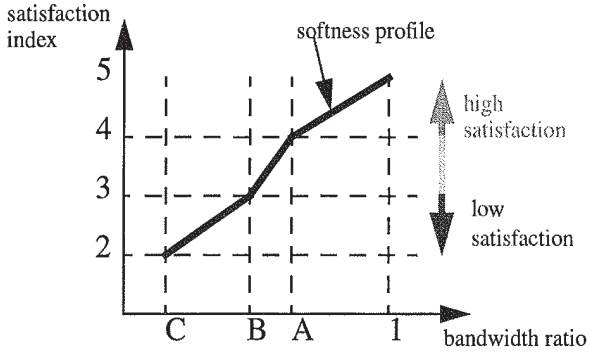


FIGURE 35.9 System and API model for QoS control with bandwidth renegotiation.

satisfaction divides the scale in two operational regions: the acceptable satisfaction region and the low satisfaction region. The bandwidth ratio is defined by dividing the current bandwidth allocated by the network to the bandwidth requested to maintain the desired application performance. Thus, the bandwidth ratio is graded from 0 to 1; a value of 1 means that the allocated bandwidth is sufficient to achieve the desired application performance. The point indicated as B is called the critical bandwidth ratio since it is the value that results in minimum acceptable satisfaction. As shown in Fig. 35.10, the softness profile is approximated by piecewise linear "S-shaped" function consisting of three linear segments. The slope of each linear segment represents the rate at which applications performance degrades (satisfaction index decreases) when the network allocates only a portion of the requested bandwidth: the steeper the slope is, the "harder" the corresponding profile is.



A ~ 1.1 B
B = "critical ratio"
C ~ 0.7 B

Application	B
teleconf.	0.2 - 0.6
VoD (action)	0.5 - 0.7
MoD (interactive)	0.3 - 0.8

FIGURE 35.10 Example softness profile.

The softness profile allows efficient match of application requirements to network resource availability. With the knowledge of the softness profile, network elements can perform soft-QoS control—QoS-fair allocation of resources among contending applications when congestion arises. Applications can define a softness profile that best represents their needs. For example, the softness profile for digital compressed video is based on the nonlinear relationship between coding bit-rate and quality, and the satisfaction index is correlated to the user perception of quality [21, 22]. While video-on-demand (VoD) applications may, in general, tolerate bit-rate regulations within a small dynamic range, applications such as surveillance or teleconference may have a larger dynamic range for bit-rate control. Other multimedia applications may allow a larger range of bit-rate control by resolution scaling [18]. In these examples, VoD applications are matched to a "harder" profile than the other, more adaptive multimedia applications. Users on wireless mobile terminals may select a "softer" profile for an application in order to reduce the connection's cost, while a "harder" profile may be selected when the application is used on wired desktop terminal. Thus, adaptive multimedia applications able to scale their video quality could specify their soft-QoS requirements dynamically to control the session's cost.

Figure 35.11 conceptually illustrates the role of application QoS/bandwidth renegotiation, service contract, and session cost in the service model. The soft-QoS service model is suitable for adaptive multimedia applications capable of gracefully adjusting their performance to variable network conditions. The service definition is needed to match the requirements of the application with the capabilities of the network. The service definition consists of two parts: a usage profile that specifies the target regime of operation and the service contract that statistically quantifies the soft-QoS service to be provided by the network. The usage profile, for example, can describe the media type (e.g., MPEG video), interactivity model (e.g., multimedia browsing, video conference), mobility model (indoors, urban semi-mobile, metropolitan coverage area), traffic, and softness profiles. The service contract quantifies soft-QoS in terms of the probability that the satisfaction of a connection will fall outside the acceptable range (given in its softness profile), the expected duration of "satisfaction outage," and the new connection blocking probability.

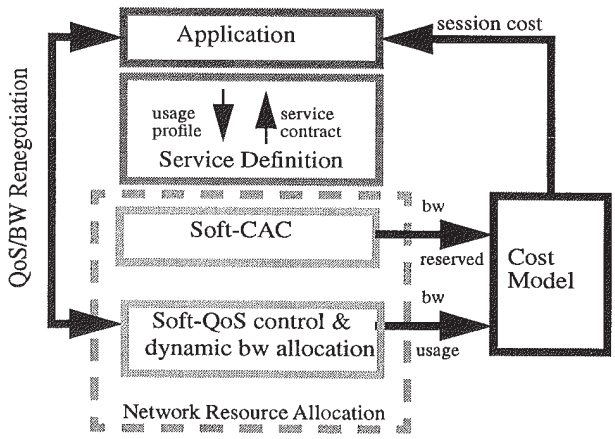


FIGURE 35.11 Service model for dynamic bandwidth allocation with soft-QoS.

Network resource allocation is done in two phases. First, a connection admission control procedure, called soft-CAC, checks the availability of resources on the terminals coverage

area at connection set-up time. The necessary resources are estimated based on the service definition. The new connection is accepted if sufficient resources are estimated to be available for the connection to operate within the service contract without affecting the service of other ongoing connections. Otherwise the connection is blocked. Second, while the connection is in progress, dynamic bandwidth allocation is performed to match the requirements of interactive variable bit-rate traffic. When congestion occurs, the soft-QoS control mechanism (re)-allocates bandwidth among connections to maintain the service of all ongoing connections within their service contracts. The resulting allocation improves the satisfaction of undersatisfied connections while maintaining the overall satisfaction of other connections as high as possible [17]. Under this model, connections compete for bandwidth in a "socially responsible" manner based on their softness profiles. Clearly, if a cost model is not in place, users would request the maximum QoS possible. The cost model provides feedback on session cost to the applications; the user can adjust the long-term QoS requirements to maintain the session cost within budget.

Soft-QoS Control in the WATMnet System

In the WATMnet system, soft-QoS control allows effective support of mobile multimedia applications with high network capacity utilization. When congestion occurs, the soft-QoS controller at the EMASs allocates bandwidth to connections based on their relative robustness to congestion given by the applications softness profiles. This allocation improves the satisfaction of undersatisfied connections while maintaining the overall satisfaction of other connections as high as possible. Within each EMAS, connections compete for bandwidth in a "socially responsible" manner based on their softness profiles.

ATM UNI signalling extensions are used in the WATMnet system to support dynamic bandwidth management. These extensions follow ITU-T recommendations for ATM traffic parameter modification while the connection is active [23]. Although these procedures are not finalized at the time of this writing, an overview of the current state of the recommendation is given next with an emphasis on its use to support soft-QoS in the mobile WATM scenario.

ITU-T Q.2963 allows all three ATM traffic parameters, PCR, SCR, and MBS, to be modified during a call. All traffic parameters must be increased or decreased; it is not possible to increase a subset of the parameters while decreasing the others. The user who initiates the modification request expects to receive from the network a new set of traffic parameters that are greater than or equal to (or less than) the existing traffic parameters if the modification request is an increase (or decrease). Traffic parameter modification is applicable only to point-to-point connections and may be requested only by the terminal that initiated the connection while in the active state.

The following messages are added to the UNI:

- MODIFY REQUEST message is sent by the connection owner to request modification of the traffic descriptor; its information element (IE) is the ATM traffic descriptor.
- MODIFY ACKNOWLEDGE message is sent by the called user or network to indicate that the modify request is accepted. The broadband report type IE is included in the message when the called user requires confirmation of the success of modification.
- CONNECTION AVAILABLE is an optional message issued by the connection owner to confirm the connection modification performed in the addressed user to requesting user direction. The need for explicit confirmation of modification is

indicated by the "modification confirmation" field in the MODIFY ACKNOWLEDGE broadband report IE.

- MODIFY REJECT message is sent by the called user or network to indicate that the modify connection request is rejected. The cause of the rejection is informed through the cause IE.

Figures 35.12, 35.13, and 35.14 show the use of these messages for successful, addressed user rejection and network rejection of modification request, respectively.

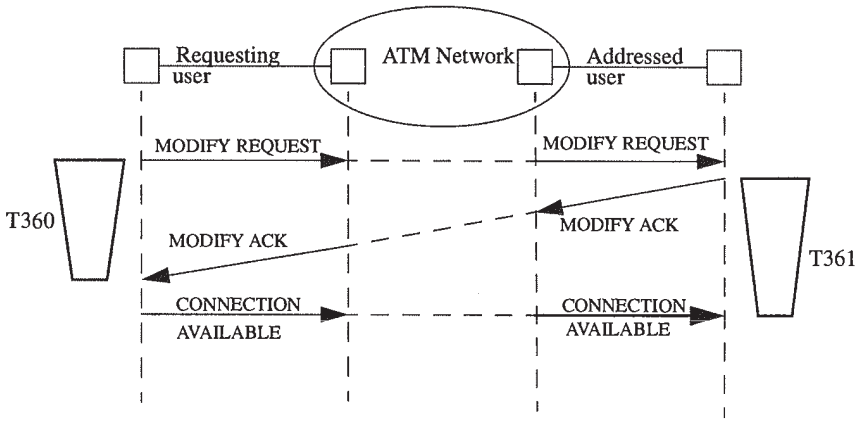


FIGURE 35.12 Successful Q2963 modification of ATM traffic parameters with (optional) confirmation.

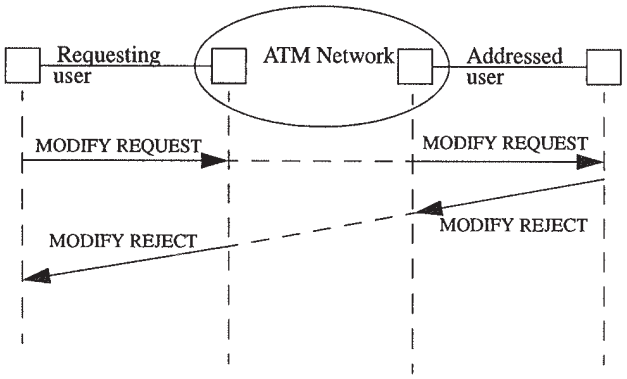


FIGURE 35.13 Addressed user rejection of modification.

Additionally, the soft-QoS control framework of the WATMnet system uses the following modifications to the Q2963 signalling mechanisms:

- BANDWIDTH CHANGE INDICATION (BCI) message supports network-initiated and called user-initiated modification. The message is issued by the network or called user to initiate a modification procedure. The traffic descriptor to be used by the connection owner when issuing the corresponding MODIFY REQUEST message is specified in the BCIs ATM traffic descriptor IE. Figure 35.15 illus-

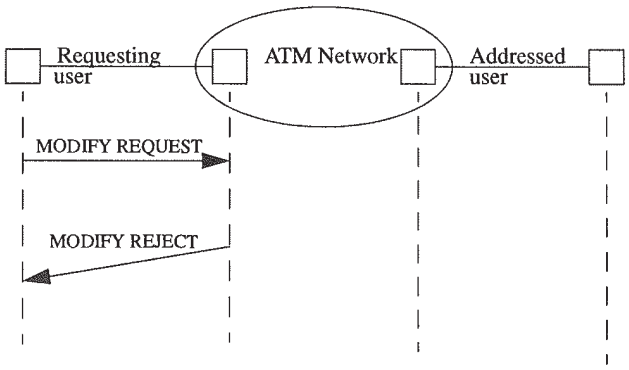


FIGURE 35.14 Network rejection of modification request.

trates the use of BCI for called user-initiated modification. Timer T362 is set when issuing the BCI message and cleared when the corresponding MODIFY REQUEST message is received; if T362 expires, the terminal and/or network element can modify the traffic policers to use the ATM traffic descriptor issued in the BCI message.

- Specification of softness profile and associated minimum acceptable satisfaction level (sat_{min}) in the MODIFY REQUEST message. The softness profile and sat_{min} are used for QoS-fair allocation within the soft-QoS control algorithm.
- Specification of available bandwidth fraction (ABF) for each ATM traffic descriptor parameter. ABF is defined as the ratio of the available to requested traffic descriptor parameter. This results in ABF-PCR, ABF-SCR, and ABF-MBS for the peak, sustained, and maximum burst size, respectively. These parameters are included in the MODIFY REJECT message. Using the ABF information, the connection owner may recompute the requested ATM traffic descriptor and reissue an appropriate MODIFY REQUEST message.

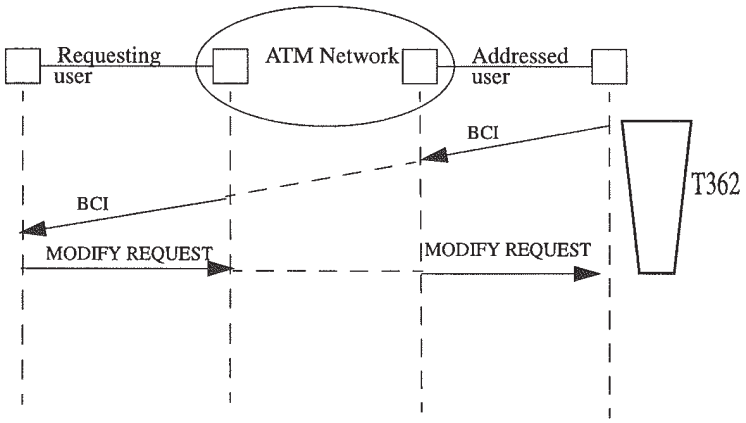


FIGURE 35.15 Use of BCI message for addressed user-initiated modification.

Two additional call states are defined to support modification. An entity enters the modify request state when it issues a MODIFY REQUEST of BCI message to the other

side of the interface. An entity enters the modify received state when it receives a MODIFY REQUEST of BCI message from the other side of the interface.

Soft-QoS control is particularly useful during the handover procedure as a new MT moves into a cell and places demands on resources presently allocated to connections from other MTs. In the present WATM baseline handover specification [9], an MT sends a prioritized list of VCs to the network during handover, but there is no specification as to what the MT or the network should do if not all these VCs can be accommodated at the new cell. A flexible way of prioritizing the bandwidth allocation to various session VCs is through their softness profiles. If a mobile terminal faces significant drop in bandwidth availability as it moves from one cell to another, rather than dropping the handover connections, the EMAS might be able to reallocate bandwidth among selected active connections in the new cell.

Within the soft-QoS framework, the soft-QoS controller selects a set of connections, called donors, and changes their bandwidth reservation so as to ensure satisfactory service for all [17]. This process is called network-initiated renegotiation. Network-initiated renegotiation improves the session handover success probability since multiple connections within and among sessions can share the available resources at the new EMAS, maintaining the satisfaction of individual connections above the minimum required. This mechanism allows multimedia sessions to transparently migrate the relative priority of connections as the MT moves across cells without a need to further specify details of the media session's content.

Figure 35.16 shows an MT moving into the coverage area of a new RP under a new EMAS and issuing a MODIFY REQUEST message (MRE). As a result, the EMAS might have

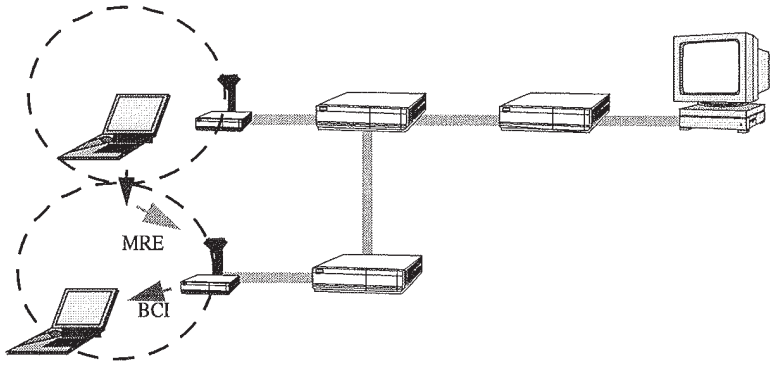


FIGURE 35.16 Handover procedure with soft-QoS in the WATMnet system.

to reallocate bandwidth of other connections under the RP to successfully complete the handover. This is accomplished by issuing BCI messages to a selected set of connections, called donors. At the time of receiving the first MRE message for a connection being handed over, no state exists for that connection within the new EMAS. This event differentiates the MRE messages from ongoing connections and connections being handed over. Different algorithmic provisions can be made to expedite bandwidth allocation to MRE messages of connections being handed over, reducing the probability of handover drop. For example, upon identifying a MRE from such connection, the soft QoS controller can use cached bandwidth reserves to maintain the satisfaction of the connection above the minimum. The size of the bandwidth cache is made adaptive to the ratio of handed-over to local bandwidth demand. The bandwidth cache for each RP can be replenished off-line using the network-initiated modification procedure. In this way, a handed-over connection need not wait for

the network-initiated modification procedure to end before being able to use the bandwidth. The outcome of the reallocation enables most connections to sustain a better than minimum application performance while resources become available. Short-term congestion may occur due to statistical multiplexing. If long-term congestion arises due to the creation of a hot spot, dynamic channel allocation (DCA) may be used to provide additional resources. It is also possible that if connection rerouting is required inside the network for handover, the required resources to support the original QoS request may not be available within the network along the new path. Renegotiation is a useful feature in this case also.

An important performance metric for the soft-QoS service model is the low satisfaction rate (LSR). LSR measures the probability of failing to obtain link capacity necessary to maintain acceptable application performance. Figure 35.17 compares the LSR with and without network-initiated modification over a wide range of link utilization for an MPEG-based interactive video application. The softness profiles for MPEG video were derived from

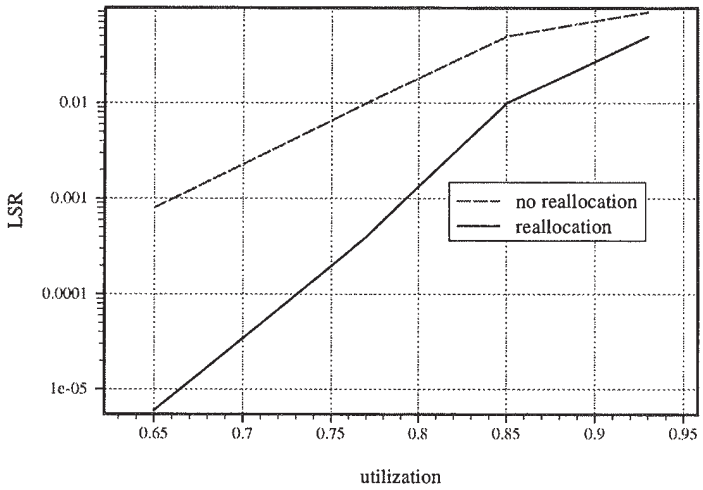


FIGURE 35.17 Effect of soft-QoS control with and without network-initiated modification.

empirical results reported in [21, 22]. The figure shows that network-initiated modification has an important contribution to soft-QoS control performance: robust operation (LSR $< 10^{-3}$) is achievable while maintaining 70 to 80% utilization. In the WATM scenario, the handoff success probability with soft-QoS control is related to the LSR by Prob(handoff success) $= \mathrm{P}(\mathrm{sat}^{+} > \mathrm{sat}_{\min}) > 1 - \mathrm{LSR} >> 1 - \mathrm{P}_b$, where sat^{+} represents the satisfaction after handover to the new AP completes. Since it is better to block a new connection than to drop an existing connection for lack of capacity, the condition $\mathrm{LSR} \ll \mathrm{P}_b$ is used, where P_b is the connection blocking probability. The operating goal for the system is to maintain utilization $> 70\%$, $\mathrm{LSR} \sim 10^{-5}$, $\mathrm{P}_b \sim 10^{-3}$.

Although the results presented are based on softness profiles for video, the definition of soft-QoS is appropriate for adaptive multimedia applications in general. Various representative softness profiles can be defined and refined as users' experience with distributed multimedia applications grows. New profiles can easily be incorporated within the framework as they become available.