Introduction to wireless communications

Pierre A. Humblet

Institut Eurecom
2229 route des Cretes, B.P. 193
06904 Sophia-Antipolis, France

Paper presented at ECCTD'93, Davos, Switzerland, August 1993.

1 Introduction

The tremendous development of communication systems is one of the main features of the last decades. These systems have become ubiquitous and reach virtually every home in large parts of the world. The resulting increase in traffic has been made possible by, or has necessitated, major improvements in the backbone of the networks, for example digital transmission and switching, and the use of optical fibers.

However the development has not led to dramatic qualitative improvements for most, even in the developed countries. Now, as 40 years ago, telephony remains the main electronic communication vehicle, and the changes there are not breathtaking. Today, one can dial directly, instead of through an operator, to any part of the world, and be almost assured of an excellent quality call. Some countries have introduced a simple videotext service. In the mass media, television has changed from black and white to color.

This is in contrast to the professional area, where the coupling of communications and computers has led to an interconnected universe. Vast amounts of data are exchanged between machines located in the same building, and across continents. This phenomenon has been made possible by the digitalization of, and the introduction of computers in, the telecommunication network. Progress in this area is not slowing down, new advances in optical and switching technologies promise continual improvements.

Those advances have led to changes in person to person communications in the context of their work. For example researchers worldwide, and workers inside many companies, have come to rely on electronic mail. The desire to transmit data to interconnect computers has led to progress in signal processing and telephone modem technology, which in turn have made the fax machine a ubiquitous and indispensable instrument.

As we approach the end of the millennium we witness a major change in mass market applications as the development of wireless mobile technology will make it possible to reach anyone, anywhere, at any time (for the better or worse), instead of having to track a person by probing his presence near fixed telephone stations.

Wireless communications is not a new technology. Fixed installations have existed for most of the century. The development of vacuum tubes, and the invention of FM modulation, have led to significant deployment of mobile systems during world war II.

Starting in the late 1940's, the development of mathematical communication theory, of Shannon's information theory, and of coding theory, have led to an excellent understanding of the field of digital communications. This quickly gave rise to the development of sophisticated but expensive systems that were mostly limited to space and military uses.

On the other hand, in the private civilian world, simple mobile systems were deployed. They evolved over time from manual to automatic operation, but had usually a single central site covering an entire region. This severely limited the number of calls that could be handled, as these systems received relatively small frequency allocations.

It was not a problem at first, because the bulkiness and cost of the mobile equipment limited its appeal. The invention and mass fabrication of the transistor, and later of the integrated circuit, changed the situation. By the 1970's it was widely recognized that the demand outstripped the offer. With hindsight, how much it did was vastly underestimated.

1.1 Car telephony

In the 1970's, the Bell Laboratories introduced a system based on the cellular concept, which leads to much greater frequency efficiency as we shall see, and in 1979 the first large scale public system, the Advanced Mobile Phone Service (AMPS), was launched in the USA. It was and continues to be a success, fueled by the miniaturization of electronics that by 1990 had largely changed an automobile based product into one relying on hand-held sets. The transmission technology used in the system was not innovative, and it did not rely on advances in digital communication theory and practice.

In 1981 Europe was led by the Nordic countries into introducing its own systems, and by the mid 80's all major countries had equipped themselves with a home brewed version. The American and European systems, as well as those in Japan, have many common features. They use analog FM modulation, with channel bandwidths varying between 12.5 and 30 kHz. Different channels use separate frequency bands. This is called Frequency Division Multiple Access (FDMA). The two directions of traffic, to and from the mobile, are carried in different bands, widely separated from each other. This is called Frequency Division Duplexing (FDD).

The lack of uniformity between the systems was partially justified by the different frequency allocations that existed in different countries. However, as soon as 1978 a common frequency band at 900MHz had been reserved for mobile communications in Europe, and by 1982 the Conférence Européenne des Postes et Télécommunications (CEPT), including

the public administrations of more than 20 countries, had created a body to recommend a new mobile communications standard for Europe. The GSM (Groupe Special Mobile) system, which is currently being deployed, is the result of that effort. The meaning of the acronym has evolved to stand for "Global System for Mobile communications".

GSM is the first major digital civilian land mobile system. Advances in VLSI technology have now made it possible to implement in a small hand-held device many of the sophisticated modulation and coding techniques that were previously reserved for military and spatial uses. The GSM system still relies on FDMA, with frequencies spaced 200 kHz apart. Each frequency band carries digital modulation at 271kb/s, using a modulation format that one can view as digital frequency modulation. Each 271 kb/s bit stream is shared, by Time Division Multiple Access (TDMA), between many channels, 8 in the current version of GSM (16 in future systems). Effectively the system carries 8 calls using 200 kHz of bandwidth, consuming about 25 kHz per call. This does not appear to be an improvement over older analog systems, but we will see later that this simple reasoning is inadequate.

Meanwhile in the USA the AMPS system was reaching saturation. The number of subscribers in 1992 was about 8.8 millions, which compares with 4.4 millions in 1990 and 90,000 in 1984. A more efficient digital system had to be introduced. However no new frequency bands were available, and the new system had to share the available frequencies with the old one. This severely limited the design options.

Initially a consensus was reached on a "second generation" digital system using both FDMA and TDMA. It is known as IS-54. The frequency spacing is kept at 30 kHz. Each frequency band carries digital modulation at 48.6 kb/s, using a linear modulation format. Each digital carrier is shared by 3 channels (6 in future systems) in TDMA mode, thus multiplying the spectral efficiency by a factor of 3 (or 6) over AMPS. Network operators would thus convert channels one by one from the old format to the new one. By doing so they would eventually carry 1 phone call in 10 kHz, and later in 5 kHz.

Before this system could be deployed Qualcomm corporation proposed another design, that used the available technology much more aggressively. It is based on a channel sharing technique called "spread spectrum" or "code division multiple access" (CDMA) and it promised much higher spectral efficiencies than IS-54.

Thus North America and Europe seem to be moving in opposite directions. North America, forfeiting the continent wide compatibility of the equipment, will see the deployment of a dual mode technology, where all handsets support the old analog standard but use different advanced digital technologies. In Europe on the other hand, the old disparate standards are progressively abandoned in favor of a new common system.

There is more to a continent wide network than just having a common standard. It must also be possible for users registered in one area to roam, i.e. originate or receive calls in other networks. This capability was not present in the original AMPS system, but a number of proprietary communication links between mobile networks have been established. In second generation systems like GSM, great emphasis is placed on roaming, and to the solution to the attendant billing, security, and operational issues.

The GSM system has also been expanded to provide services in the 1800MHz band. It is then known as DCS1800.

1.2 Cordless systems

Another class of mobile system exists, cordless telephones. In the first generation, they are stand-alone consumer products that do not require any interoperability specifications. Each cordless phone comes with its own base station, and in fact should only operate with that base station for privacy and billing reasons.

This first generation of products is limited by small coverage and sensitivity to interference. The better handsets have access to many channels and automatically choose the one with least interference. This process is completely distributed, there is no central control.

Second generation cordless telephones are designed not only for residential use, but they are now seen as having four applications. The residential environment can be extended to the business environment, i.e. access to a private switch (known as PABX). Yet another extension is for public pay service (telepoint). They also offer a network of base stations, with no or limited ability to move between base stations while a call is in progress. Finally the same technology could also connect homes to the fixed telephone network, replacing the local loop. This essentially amounts to placing the base station on a telephone pole outside the home. The future liberalisation of local telephone service could greatly benefit from this possibility.

The first second generation system is logically known as Cordless Telephone 2 (CT2). It originated in Great Britain, where originally four operators offered similar but incompatible telepoint services. Eventually a standard evolved, it is also a FDMA system. The frequency separation is 100kHz, and the bit rate is 72 kb/s. The modulation format is Frequency Shift Keying. Contrary to systems for car telephony, both directions of a call alternate on the same frequency, with a period of 2 ms. The emphasis of this system is on simplicity and low weight, not on very high performance.

As mentioned earlier, CT2 was originally introduced in Great Britain as telepoint system (outgoing calls only). It initially floundered as a commercial operation. Now the system seems to be expanding in Great Britain after having met much success in the Far East. It has also started another life across the Channel. After a trial operation in Strasbourg, France Telecom is offering it in Paris under the name of Bi-Bop. At the outset the operator will install about 4000 base stations all over Paris, each usable within a radius of about 200m.

Another cordless system, the Digital European Cordless Telephone (DECT) has channels spaced 1.728MHz apart, modulated at 1.152 Mb/s, using almost the same modulation format as GSM. Each frequency carries 12 channels, in both directions, with a period of 10 ms. It can be seen as an advanced version of CT2. Its main application appears to be as an access to PABX.

1.3 Future systems

When one thinks of wireless communication, the first picture that comes to mind is that of talkie-walkies, or citizen band sets, where mobile users communicate directly with each other. Such systems were very popular for a relatively brief period. They are very different from car radio systems and cordless phones that we have described, where the aim is to provide wireless access to the traditional telephone network. Accordingly a significant part of the cost lies in the fixed equipment, such as the base stations, the switches, and the cables interconnecting the fixed equipment.

Support of mobile users also necessitates major upgrades to the networking software of telephone networks, and many see systems such a GSM as a part of the implementation of a future "Intelligent Network", providing novel and useful services on top of the current telephone network infrastructure.

This leads us to a vision of a third generation, of a universal mobile network combining the best features of the existing systems. It is called "Personal Communication Systems". It would be completely integrated in the intelligent network, and would offer new functionalities, particularly in the area of data transmission. It would be coupled with satellite systems, which are also being developed. They provide access to parts of the globe with low population density, where the development of a fixed infrastructure is impractical or not justified economically.

The goal is to provide communication from everyone, with everyone, or with sources of data or knowledge, anywhere and at any time. How to get there is not clear.

For example, it is still debated whether this goal requires the development of a single handset, capable of both "mobile" and "cordless" operation, or if there should be different, but tightly linked, services . Despite the similarity of their goal, there are major differences between car radio and cordless phone applications. One is that of user expectations.

Users recognize that being able to place calls from a car is a significant improvement over the statu quo ante. Consequently they are willing to pay more, to tolerate occasional quality or accessibility deficiencies, to access systems that are congested at peak times, to suffer from marginal voice quality and delay, to operate on a medium of dubious privacy, and to watch carefully the charge of their batteries (which can get recharged while the handset is in the car).

On the other hand, cordless systems are a replacement for the ubiquitous fixed telephone set. Thus they expect excellent voice quality and low delay, low cost, permanent availability, no delay to get a dial tone, and no hassle with the batteries.

The operating conditions are also very different. Those for use in car must tolerate higher speed, which poses problems both for the receiver and for the system: the cells must be larger to limit the number of hand-offs. Not only do pedestrians travel more slowly, they also seem more likely to move a few feet toward a window, or to a street corner, to get good quality communications. This cannot be expected from the driver of a car.

This brief section suffices to indicate that future systems will be determined not only

by technology, but by a number of behavioral and economic factors whose prediction is risky!

1.4 Plan of the paper

This introductory section gave a brief history of wireless communications in a mobile environment, and introduced the major systems used in the world today. The next section describes the system part of a typical mobile communication system, including the fixed infrastructure, and the mobility management and security issues. It is followed by section 3 which discusses the statistical properties of the mobile communication channel, introducing both mathematical and empirical models of propagation. Section 4 introduces more carefully the concept of cells, and of frequency reuse. They are of fundamental economic importance to the operation of mobile telephone systems. It is followed by a review of communication theory applied to the models of section 3. Throughout the paper we provide illustrations based mainly on the European GSM. The final section is devoted to a brief review the system proposed by Qualcomm as it has a number of interesting features.

2 Mobile Communication Systems

There is much more to a mobile cellular network than just transmission on a radio channel. It is in fact quite a complicated system that must interact with the users at the mobile terminals, with the fixed telephone network, and with the operation and maintenance services of the operating company. It must manage the movement of the users, as well as authenticate them, and bill them for the services they use.

One should also keep in mind that a modern telephone network is really made of two distinct networks: a "lower level" network capable of establishing circuit switched calls (essentially bit pipes) between two points, and a "higher level" network used by the switching machines to exchange messages with each others. That higher level network can be thought of as a packet switched computer network. It makes use of the same transmission facilities as the lower level network. The dichotomy has been inherited by most mobile networks.

This section gives a high level view of the functions listed above. Many of the names and implementation details come from the GSM environment, but they are present in similar fashion in most systems.

2.1 Physical structure

In addition to the mobile stations, each mobile network in turn is composed of 3 key elements, interconnected by land lines. They are the Mobile Switching Center, the Base

Station Controller, and the Base Station Transmitters. The simplified structure of a GSM network appears in figure 1.

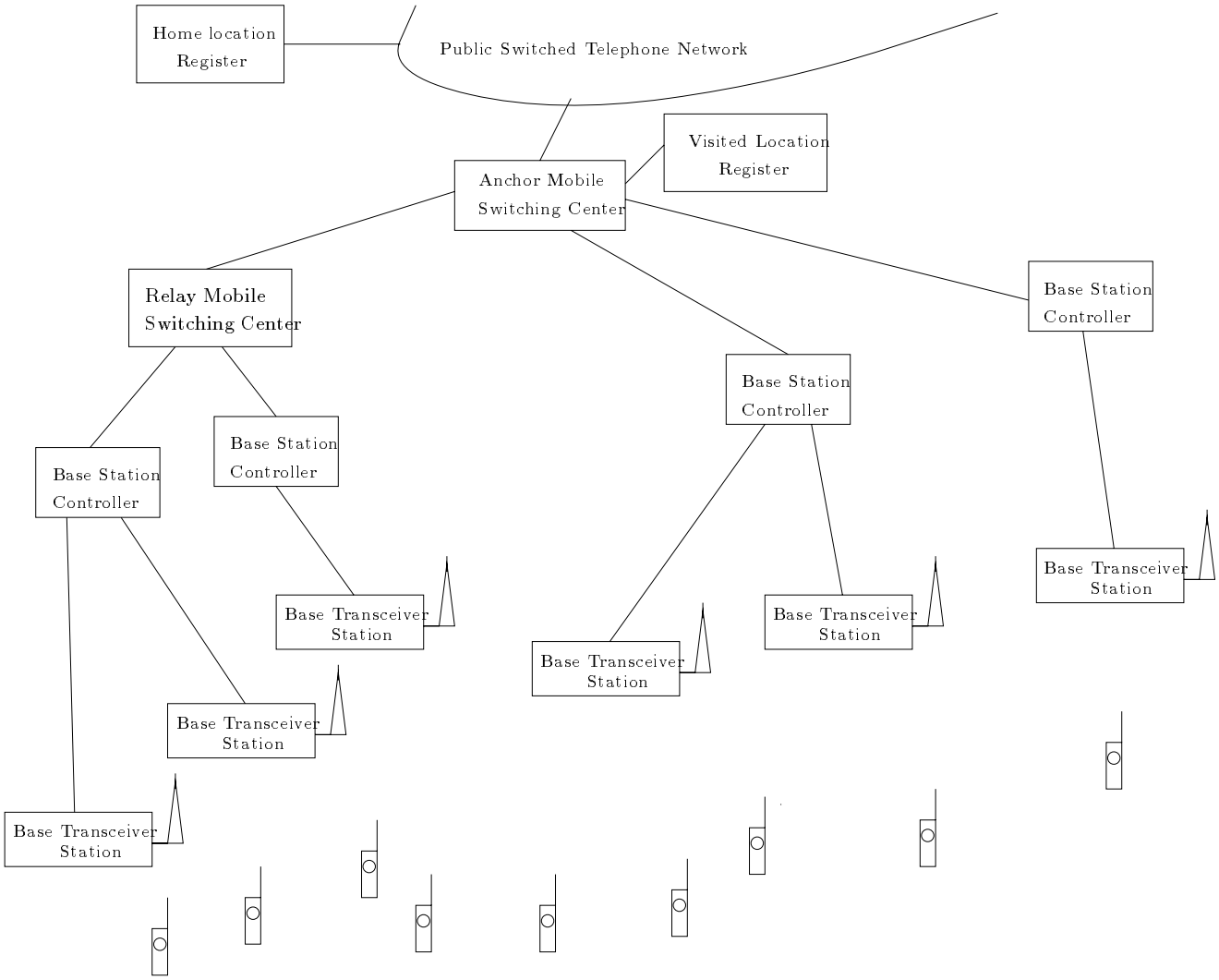


Figure 1: Structure of a GSM mobile network

The heart of a mobile cellular network is actually a switch, very much like as a telephone switch. It is called a Mobile Switching Center (MSC). Larger cellular networks actually have many MSC's. An MSC is not only responsible to setup and disconnect circuits, but also to manage the movements of the users. It should really be thought of as the combination of a switch and of a general purpose computer. Of course, modern telephone switches already have the functionality of a general purpose computer.

A network must keep a database with information about its customers. It is called the Home Location Register (HLR). Its functionality will be described later. The HLR can be accessed by a local MSC when a user operates in his home network, or by a remote MSC when the user roams in another network. To minimize the amount of control traffic,

a copy of the relevant information is kept in a local database. It is called the Visiting Location Register (VLR), and it is usually associated with an MSC.

An MSC is connected by regular land lines both to the fixed telephone network, and to Base Station Controllers (BSC). The protocol used on those line is called "System 7". It was developed for the exchange of messages between telephone switches.

A BSC is basically a small switch with a substantial computation capability. It controls a few dozens of Base Station Transmitters (BTS) to which it is attached by land lines. It is responsible mainly for the allocation of radio channels and the hand-offs. The protocol between a BSC and a BTS does not follow System 7, is closer to that of the data communications world.

Each BTS is located near an antenna. It consists mostly of transmitter and receiver equipment. Its main functionality is to interface the fixed network to the radio channel, thus insuring both modulation/demodulation and speech coding/decoding. The architecture has evolved, and the speech coding/decoding is often done in equipment that is sited away from the BTS, actually close to an MSC. A BTS has very small management responsibilities. It normally carries traffic for a few dozens of calls. The signal transmitted by a BTS can be heard in the adjacent area, called a cell.

The Mobile Stations (MS) are equipped with speech processing and radio communication equipment, of course, as well as equipment to measure the radio environment. They interface with the user via the keyboard, microphone and loudspeaker, and possibly with data interfaces to connect computers and fax machines. They also contain a slot for a smart card, i.e. a credit size plastic card in which are embedded a processor and a memory. Its role will be described in the next section.

2.2 User, security, call and mobility management

Each user subscribes to the services of a home network. At subscription time he receives a smart card with a unique identity, the International Mobile Subscriber Identity (IMSI). That smart card can be placed in any GSM mobile station, for example one located in a rented car, and it can be used in any GSM system, as determined by the type of subscription. The separation of the card from the mobile station offers increased flexibility and security compared to previous systems.

The information contained in the smart card is also maintained in a Home Location Register (and possibly in a distinct Authentication Center which we will not consider as separate in this discussion), which is essentially a database maintained by the home network provider. That database also contains a description of the service features subscribed to, and of the last known location of the subscriber.

We now examine how the system operates, examining first how the user and the network establish initial contacts.

2.2.1 Location update and authentication

BTS's repeatedly broadcast their identity and relevant control information on a control channel. In particular, BTS's are grouped in Location Areas (LA), a concept we will examine more below, and they broadcast the LA identity. When a mobile station is turned on, it scans all the control channels and determines a good BTS to communicate with. The choice depends on signal levels, as well as on the identity of the BTS. For example, an MS will normally attempt to communicate with a BTS belonging to the network that has issued the smart card.

If that BTS is not in the LA last visited by the mobile, the MS opens a communication channel with the BTS, and eventually with the MSC, for the purpose of updating its location. If the IMSI does not already appear in the VLR, the MSC contacts the HLR, and let it know where the customer is currently located. If the location is new, the HLR records the information and it notifies the previous VLR that the customer has left its area. In parallel it also conveys security and subscription information to the MSC. The exchange is illustrated in figure 2.

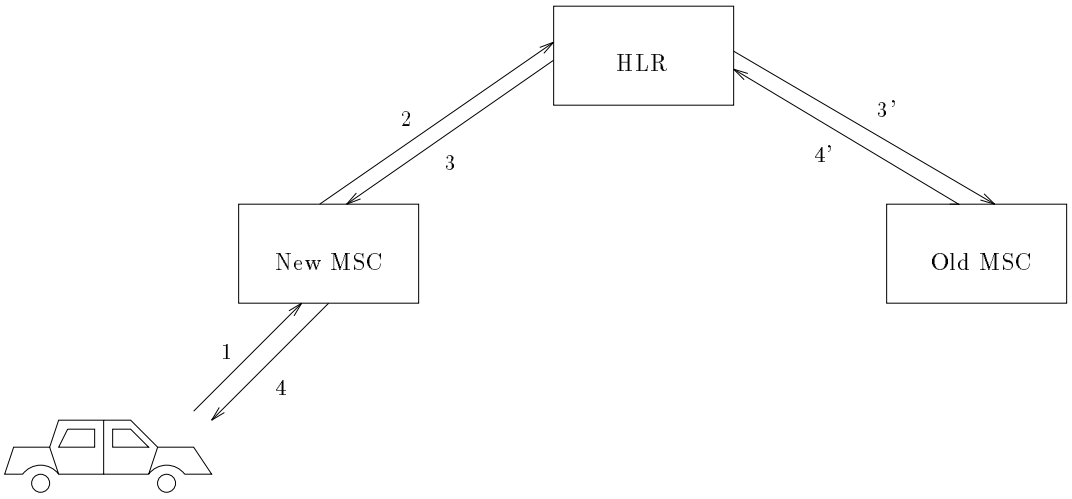


Figure 2: Sequence of interactions for location updates

Assuming the customer's subscription allows him to communicate in the visited network, the HLR will provide the VLR with subscription information, and with means of authenticating the MS. This consists of a set of triples {RAND, SRES, Kc}. RAND is a random number. SRES (for Signed RESult) and Kc (Key) result from applying RAND to a secret function that is encoded in the smart card and in the HLR, and that is unique to the IMSI. It is unknown to the customer and to all the other elements of the networks, except the smart card and the authentication part of the HLR.

To verify the identity of the IMSI, the visited VLR sends RAND to the MS, via the MSC, the BSC and the BTS. The MS passes RAND to its smart card, which in turns computes SRES and Kc. SRES is passed back to the VLR over the radio channel. If it matches the SRES computed at the HLR, the identity is verified. The MSC then has the

option to have the data transmission on the radio link be encrypted and decrypted with the key Kc, which was never transmitted over the air! The system should be designed so that knowledge of RAND and SRES does not give too much information about Kc.

Encryption takes place by "exclusive-oring" blocks of transmitted bits with a sequence S that is a time varying function of the key Kc. The decryption procedure is identical. Reportedly the customer is unaware whether encryption is actually used, and the algorithm that produces the sequence S from the key Kc can be country dependent. Thus nations can choose (or be dictated by others) what level of privacy to provide to their citizens.

As described, the IMSI is still transmitted over the air, which would permit an eavesdropper to detect what IMSI is making calls. Actually, when a mobile first contacts the MSC, it is issued an alias, the Temporary Mobile Subscriber Indentity (TIMSI) which is used preferably to the IMSI. The main reason to use the TIMSI is that it is shorter, and thus reduces the length of some messages. It has the side effect of protecting the caller's identity from all but eavesdroppers who have followed the exchange of messages since the very beginning.

Once it has registered with the MSC, a mobile station breaks the connection and goes in idle mode, where it just keeps monitoring adjacent base stations and where it listens to a special channel, the paging channel. As a Mobile Station moves about, it may switch from one BTS to another, not informing anybody as long as it remains in the same Location Area. When this occurs, it must open a connection with the new BTS and eventually with an MSC to signal its new location. Authentication takes place again, a new key Kc is computed, and a new TIMSI is issued.

2.2.2 Call processing and billing

Let us examine now how calls are handled. If the MS initiates a call, it opens a connection with the BTS, BSC and MSC, and gives the called number. The MSC checks the VLR to see if the call is authorized (there may be restrictions on allowed calls, say to international numbers). This information was downloaded from the HLR at the time of the location update. If all goes well, the MSC reaches the called number through the fixed telephone network (or another mobile network) and it connects the parties. Some time after the call is over, the MSC sends a billing record to the home network.

Assume now that a call is made from a fixed telephone to a mobile. The phone number of a mobile is really a pointer to an entry in the user's HLR. There could be many phone numbers associated with a single HLR entry, say one for voice calls and one for fax. The fixed network connects the source telephone to a switch, called a GMSC (Gateway MSC) normally located near the HLR. It interrogates the HLR to learn of the current location of the customer. The HLR then verifies if the call is authorized. If so, it contacts the visited MS, passes information about the type of call (voice or fax) and obtains what amounts to a temporary telephone number to reach the MS through the "lower layer" of the telephone system. That number is communicated back through the HLR to the GMSC, which establishes a call to the visited MSC. The sequence of messages is illustrated

in figure 3. That MSC consults its VLR to learn the Location Area of the mobile, and

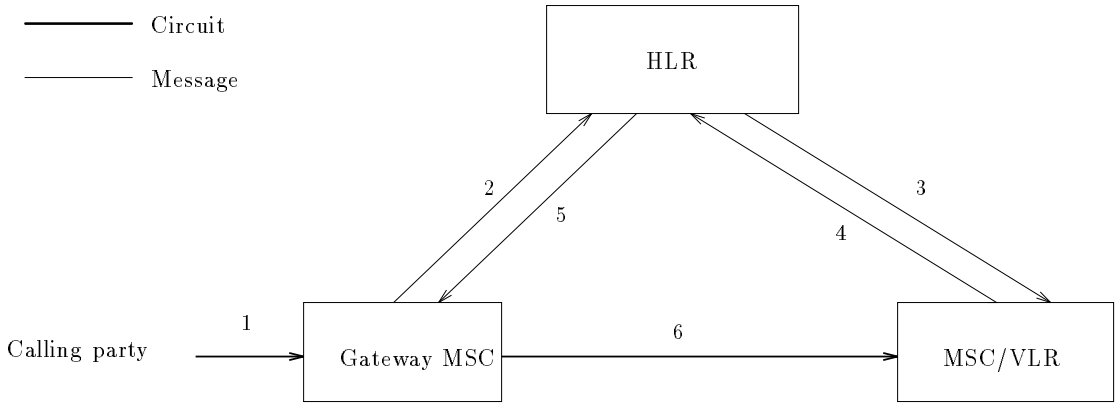


Figure 3: Sequence of interactions for incoming calls

issues a paging message to the mobile through all BTS's sited in the LA. One sees that if an LA is large, fewer location updates are required, but more paging messages are needed. It is to allow the tradeoff that LA's are used.

Hopefully the MS will hear the paging message and it will open a call through a BTS and a BSC to the MSC. The parties can then be connected.

It is interesting to consider who will pay for the call. A general principle is that the calling party should pay a fixed amount, independent of the location of the mobile. Thus the leg from the calling telephone to the GMSC is normally paid by the calling party. If the mobile station is in its home network, it may not be charged at all. However if it is roaming in another network, it will be charged for the leg of the call from the GMSC to the mobile. In that case the GMSC must issue a billing record to the mobile, and it is for that reason that it is located in the home network.

This has some interesting consequences. Two British families traveling together by car to Italy may decide to keep in touch by mobile radio while driving through France, say to debate on the next rest area. Even though they may only be some meters apart, the call will go though a GMSC located in Great Britain. They will each be charged for an international call: the calling party for a call from France to Great Britain, the other for a call from Great Britain to France!

There is another aspect to cost and billing. While those British cars travel in France, their mobile radios are updating their locations with French MSC's, keeping French computers busy and French lines humming. If they reach the Italian border without having made any call, the French operator will not have collected a centime for rendering the location tracking service.

2.2.3 Hand-Off procedures

We have considered above how a mobile keeps the system updated about its location, and how calls are made. We have not examined yet how calls are maintained while a mobile

station is moving, potentially changing cells.

In this area there has been a definite evolution over time, from giving the entire responsibility to the fixed network in AMPS, to involving the mobile station in GSM, to giving the entire responsibility to the mobile station in some cordless systems.

Let us consider why an hand-off might be needed. The first and most obvious reason is that the mobile has moved out of range of a base station, and needs to communicate with another one to maintain the call. However, it may be beneficial to change base station before communication becomes impossible, because doing so would allow reduced power levels and thus abate the interference caused in other calls. This forms the second reason. There is a third reason that is linked to traffic conditions. When a cell is heavily loaded with traffic, it may be beneficial to hand-off some calls to other cells, even if higher power levels are required.

From this discussion, one sees that hand-off decisions involve both radio and traffic considerations, and it is this dual aspect that drives decision on where to perform the computation.

The traffic conditions are perfectly known to the MSC, and pose no particular measurement issues, although communicating the information can demand much capacity. Conditions on the radio link are another issue. In older systems like AMPS, all the base stations in an area monitor the power levels of transmissions, even transmissions from mobiles handled by other base stations. The measurements are reported to the MSC where hand-off decisions are made.

In newer systems like GSM, the BTS still measures the quality of the communications with the mobiles it handles, but the mobiles themselves are responsible for monitoring power levels and delays (indicating distances) with neighboring BTS's. This is called Mobile Assisted Hand-off. Having a Mobile Station scan all frequencies to find the BTS's would be time consuming. To make the monitoring more fruitful, the BSC communicates to the MS a list of up to 6 frequencies to be monitored. About once or twice per second, the measurement results are communicated back by the mobile on a signaling channel that is associated with the channel carrying the voice or the data.

In GSM the decision about hand-offs is made by the BSC in charge of the call, based on the information available there. The choice of the BSC as the center of decision for hand-offs is motivated by the desire to spread the processing load. To take global traffic conditions into account, the BSC may suggest ranked alternatives to the MSC.

Once a decision is made, based on all available measurements, new paths must be established. One can view figure 1 as a tree, in which there is an old path from a BTS to the anchor MSC, and a new path from the new BTS to the anchor MSC. The anchor MSC is the MSC that carries the call to the fixed network. Those two paths will merge at some some point, called the switching point. It can be the BTS itself (the old and new BTS may coincide if the hand-off simply involves a frequency change), the BSC, a relay MSC, or the anchor MSC. In any case, messages go up the hierarchy to the switching point, and down to the new BSC. Terrestrial circuits facilities are established as needed from the switching point to the BTS.

Once the preparations are complete, the new BSC notifies the old one, which notifies the MS through the old BTS. The MS switches frequency, contacts the new BTS and establishes the connection. If everything is successful, the new BTS notifies the switching point and the terrestrial circuits to the old BTS are released.

The hand-off procedure must be executed quickly. To this effect, the messages it entails on the radio link are not transmitted on normal signaling channels, but actually preempt the transmission of speech or data. The interruptions are sufficiently brief and infrequent to be hardly noticeable.

This rapid survey of some system aspects should have given some appreciation of the vast and complicated distributed software structure that supports mobile networks. We will now change topic and review how waves propagate in a radio network.

3 Statistical Description of Multipath channels

Radio channels are well modeled by linear systems, as Maxwell's equation are essentially linear in the medium and at the power levels of interest. Non linearities are weak and they can be neglected. They occur for example because of rust and ice in antennas systems. However radio channels often vary with time, and they must be modeled as randomly time-variant linear systems.

The source of the time variation depends on the system. In fixed shortwave ionospheric radio communications in the HF band and tropospheric scatter (beyond the horizon) communications above 300MHz, the time variant impulse response of the channel is a consequence of the movements of the layers of the earth atmosphere that reflect the waves. In mobile communication channels that are of interest here, the variation with time arises both from the movements of the mobile (in a car, or carried by a moving pedestrian), and from the movements of neighboring objects that act as reflectors, such as passing cars and trucks, or doors and windows.

Communication theorists model the channel as a random process, and characterize it statistically. Results on this topic go back 40 years, we review them here.

3.1 The channel impulse response

Assume we transmit $x'(t)$, a passband signal around some carrier frequency f_c. It can represent a component of either a magnetic or an electric field. It admits the representation $x'(t) = \Re(x(t)e^{j2\pi f_c t})$ where $x(t)$ is a low pass signal called the complex envelope of $x'(t)$.

If we model the channel as a time variant linear system, the received signal will be

$$y'(t) = \int d\tau \; c'(\tau, t) x'(t - \tau) \tag{1}$$

where $c'(\tau, t)$ is the time variant impulse response of the channel, specifying the output

at time t due to an input at $t - \tau$. If the channel was time invariant, we would simply write $c'(\tau)$.

The complex envelope of $y'(t)$, $y(t)$ is then given by

$$y(t) = \int d\tau \, c(\tau, t) x(t - \tau) \tag{2}$$

where

$$c(\tau, t) = c'(\tau, t) e^{-j2\pi f_c \tau} \tag{3}$$

For example a channel may consists of a series of discrete echos that originate from scattering, diffraction or reflection on buildings, as illustrated in figure 4. Letting the echo amplitudes and delays be $a_i(t)$ and $d_i(t)$ we obtain

$$c(\tau, t) = \sum_i a_i(t) e^{-j2\pi f_c d_i(t)} \delta(\tau - d_i(t)). \tag{4}$$

The coefficients a_i can be complex if a phase shift is associated with a reflection.

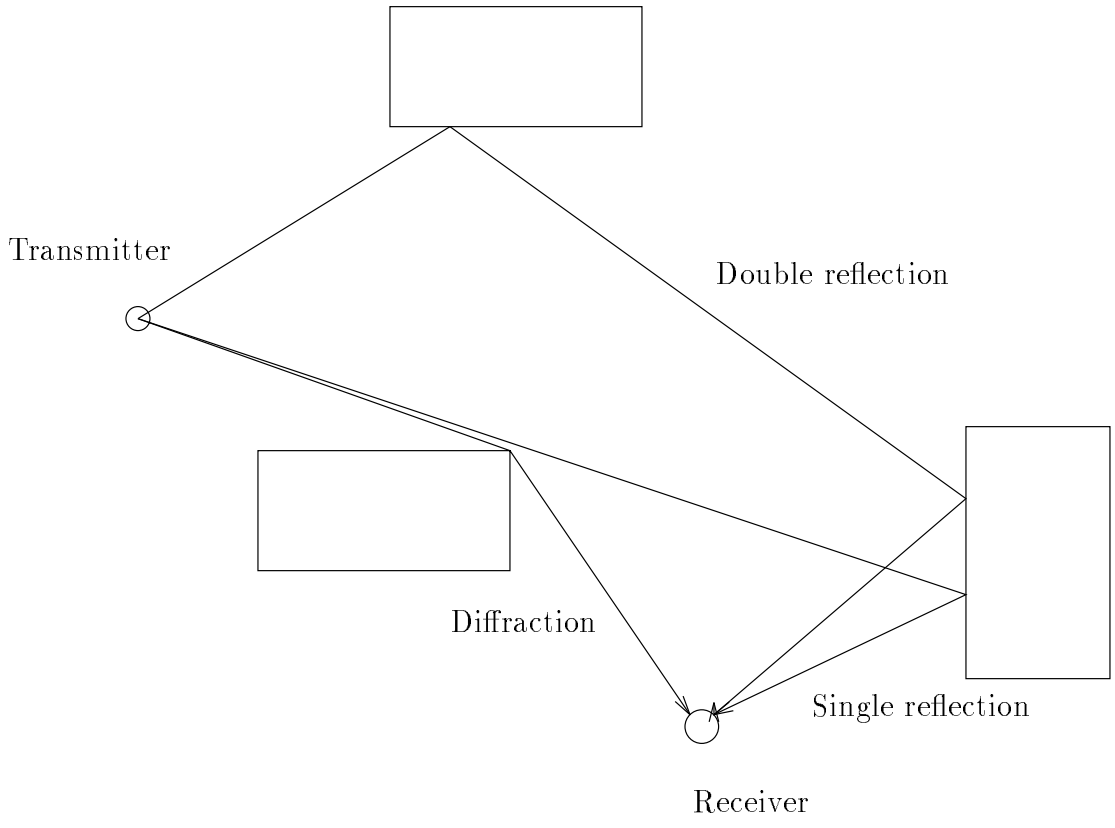


Figure 4: Multipath propagation

In the simplest case there are only two paths with constant amplitudes and and constant delays

$$c(\tau, t) = a_1 e^{-j2\pi f_c d_1} \delta(\tau - d_1) + a_2 e^{-j2\pi f_c d_2} \delta(\tau - d_2)) \tag{5}$$

The channel is now stationary and its frequency response is

$$\begin{aligned} C(f) &= a_1 e^{-j2\pi(f_c+f)d_1} + a_2 e^{-j2\pi(f_c+f)d_2} \\ |C(f)|^2 &= |a_1|^2 + |a_2|^2 + 2\Re(a_1 a_2^* e^{-j2\pi(f_c+f)(d_1-d_2)}) \\ &= |a_1|^2 + |a_2|^2 + 2|a_1||a_2|\cos(\phi + 2\pi(f_c+f)(d_1-d_2)) \\ &= (|a_1| - |a_2|)^2 + 4|a_1||a_2|\cos^2(\frac{\phi}{2} + \pi(f_c+f)(d_1-d_2)) \end{aligned}$$

where ϕ is the phase of $a_1 a_2^*$. One sees that the phase and magnitude of $C(f)$ vary wildly with frequency, with a period $1/(d_1-d_2)$. The power transfer oscillates around $|a_1|^2+|a_2|^2$, and its minimum is $(|a_1| - |a_2|)^2$. This can be very close to a null when $|a_1| \approx |a_2|$.

The fact that nulls are present when paths have comparable attenuations is a very important characteristic of radio channels, it remains true in more complicated situations. The behavior is exhibited in figure 5 for $a_1 = 1$ and $a_2 = .9$.

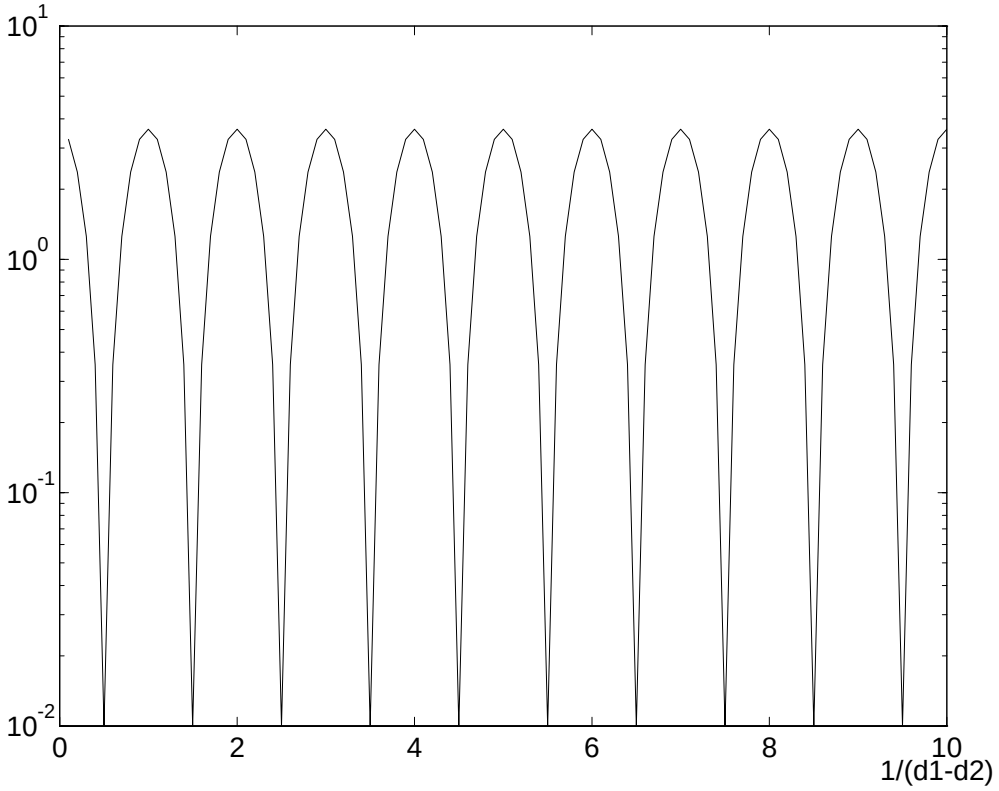


Figure 5: Magnitude of the frequency response of a two path channel with $a_1 = 1$ and $a_2 = .9$

There a number of methods to measure the impulse response $c(\tau, t)$.

It can be measured directly by sending an impulse at the carrier frequency through the channel.

Another method is to send a wideband signal $x(t)$, to measure the received signal $y(t) = c(\tau, t) * x(t)$ and to perform a deconvolution to recover $c(\tau, t)$. This is only possible if $x(t)$ is invertible, and if the channel does not change during the transmission of the signal.

A third method is to send a signal with a complex envelope $x(t)$ that has a sharp autocorrelation function. One way to construct such an $x(t)$ is to have

$$x(t) = \sum_{k=1}^{n} s_k p(t/T - k) \tag{6}$$

where $p(t)$ is 1 on $[0, 1]$ and 0 elsewhere, and the s_k are from a binary maximal length shift register sequence, with the 0 and 1's mapped into +1 and -1. This $x(t)$ has duration nT and an autocorrelation

$$R_x(t) = \int d\tau \, x(t + \tau) x(\tau) \tag{7}$$

that is equal to the signal energy nT when $t = 0$, but that for $|t| > T$ has sidelobes that are much smaller, typically by a factor of n. Those sequences are used in *spread spectrum* systems, we will meet them later.

The receiver demodulates the signal and passes it through a matched filter, i.e. a filter with impulse response $x(-t)$, suitably delayed to make it causal. The signal observed will be

$$x(\tau) * c(\tau, t) * x(-\tau) = R_x(\tau) * c(\tau, t) \tag{8}$$

where we have assumed that $c(\tau, t)$ did not change during the transmission of the signal. If the differential delays between the paths are larger than T, the individual paths can be resolved by the central component of $R_x(t)$, as long as their a_i's are large enough to be distinguished from a strong path exciting a sidelobe of the correlation function.

The discrete path model applies to the propagation of waves in a wide frequency band. If one includes in the description of the channel the response of the filters in the transmitter and receivers, then each path is spread (convolved with) the response of the filters. In particular if two paths are separated in time by less than the response time of the filter, they will not be distinguishable.

Waves take about 3.3 μs to travel 1 km, thus delays in urban areas are of the order of a few μs. In hilly terrain they can increase to a few tens of μs. The frequency response is thus flat over bandwidths smaller than hundreds or tens of kHz. In micro and picocells environment one should recall that light travels one foot in 1 ns, and the delays are of the order hundreds of ns. The frequency response is flat over bandwidths measured in MHz.

3.2 Effect of the time variations

Consider now what happens if we allow changes to occur in the impulse response. In this case it is helpful to work in the frequency domain and to send a signal $x'(t) =$

$2\cos(2\pi(f_c+f)t)$. The received low pass complex signal will then be

$$\begin{aligned} y(t) &= \int d\tau\, c(\tau,t) e^{j2\pi f(t-\tau)} \\ &= C(f,t) e^{j2\pi f t} \end{aligned} \tag{9}$$

$C(f,t)$ is simply the Fourier transform (over τ) of $c(\tau,t)$. It is an instantaneous Fourier transform.

To gain some insight into the situation, we examine a model with only one path and perform a first order approximation for $a_1(t)$ and $d_1(t)$ around t_0, using ' to denote derivatives. Thus

$$\begin{aligned} a_1(t) &\approx a_1(t_0) + a_1'(t_0)(t-t_0) \\ d_1(t) &\approx d_1(t_0) + d_1'(t_0)(t-t_0) \\ C(f,t) &\approx (a_1(t_0) + a_1'(t_0)(t-t_0)) e^{-j2\pi(f_c+f)(d_1(t_0)+d_1'(t_0)(t-t_0))} \\ &\approx a_1(t_0) e^{-j2\pi(f_c+f)d_1(t_0)} e^{-j2\pi d_1'(t_0)(t-t_0)} \\ y(t) &\approx C(f,t_0) e^{-j2\pi(f_c+f)d_1'(t_0)(t-t_0)} e^{j2\pi f t} \end{aligned}$$

where we have kept only the most significant terms, assuming that the amplitude of the path varies insignificantly over time periods of the order of a cycle of the carrier signal. We see that the frequency response is itself a sinusoid in time, and that the received signal will not be at the same frequency as the transmitted signal, but at a frequency shifted by $(f_c+f)d_1'(t_0) \approx f_c d_1'(t_0)$. This is due to the *Doppler effect*. It changes the input frequency by an offset $f_c d'(t_0)$ called the *Doppler frequency*.

If the difference in delay is due to a relative movement at speed v and angle α between the mobile and the incident wave (figure 6), the rate of change of the path length is $-v\cos(\alpha)$ and the rate of change of delay is $-v\cos(\alpha)/c$, where c is the speed of light. The Doppler frequency is then $-vf_c\cos(\alpha)/c = -v\cos(\alpha)/\lambda$. The maximum doppler frequency is equal to v/λ, where λ is the wavelength. At $f_c = 900$ MHz the maximum Doppler frequency is about 1 Hz when the mobile speed is 1 km/h.

Note than in the figure the angle α was measured in elevation in a vertical plane. In reality it must be measured in the 3 dimensional space.

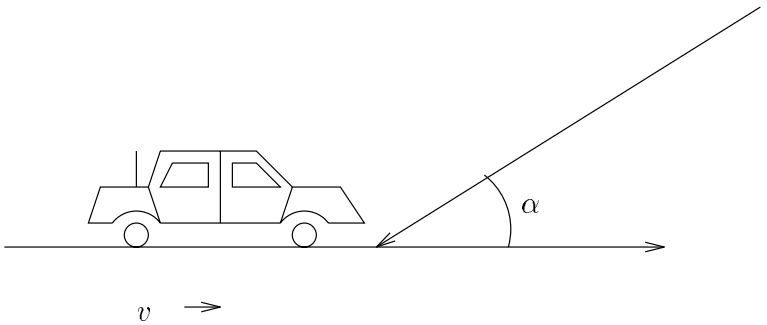


Figure 6: Geometry for the Doppler frequency

If there are many paths

$$C(f,t) = \sum_i a_i(t_0) e^{-j2\pi(f_c+f)d_i(t_0)} e^{-j2\pi(f_c+f)d_1'(t_0)(t-t_0)} \tag{10}$$

Here we see two phenomena. If all paths have the same d_i', then the received signal simply has a frequency offset, as for the single path case. However if the various paths have different delay variations, the frequency response of the channel will change radically over times $1/f_c\Delta(d')$. This is called fading.

$\Delta(d')$ is an ill-defined concept representing the typical difference of the delay variations. In urban areas the reflections come from all around the mobile, thus the d_i' can be negative as well as positive and $f_c\Delta(d')$ has the same order of magnitude as the maximum Doppler frequency.

3.3 Local statistical description of the channel

Up to this point we have described the channel by the deterministic functions $c(\tau,t)$ and $C(f,t)$. We have seen that these functions are highly variable, and they are best treated as two dimensional stochastic processes. We will proceed to do so, defining a number of interesting correlation functions and power spectral densities.

In many radio transmission systems in which there is no line of sight path between the transmitter and the receiver, each $a_i(t)$ is modeled as a complex random variable with independent identically distributed real and imaginary parts, with Gaussian statistics. Thus the magnitude square of $a_i(t)$ has a centered chi-square distribution with two degrees of freedom, and its phase is uniformly distributed on $[0, 2\pi]$. The distribution of the magnitude is Rayleigh, and the name of the model is "Rayleigh fading channel".

A justification of this model is that the reflection with delay $d_i(t)$ is really made up of very many waves that have been scattered by the reflector, and that arrive at the receiver at about the same time (with respect to the bandwidth of the receiver) but with different amplitudes a_{ki}' and phases ϕ_{ki}. The in-phase and quadrature components of the ith path are given respectively by

$$\begin{aligned} a_{i1} &= \sum_k a_{ki}' \cos(\phi_{ki}) \\ a_{i2} &= \sum_k a_{ki}' \sin(\phi_{ki}) \end{aligned}$$

If the ϕ_{ki} modulo π have a distribution that is uniform on $[0,\pi]$ then a_{i1} and a_{i2} are uncorrelated. If there are many scattered waves of comparable magnitude corresponding to each reflection, the Central Limit Theorem can be invoked to justify the Gaussian statistics of a_i referred to above. The mean square of $|a_i|^2$ is given by $\sum_k |a_{ki}|^2$, thus it depends on the local propagation conditions.

When there is a direct path, or many paths well separated in time and generated by near perfect reflectors, the Central Limit Theorem cannot be invoked. The magnitude

of a wave that is the sum of a deterministic component and a Gaussian component has a Rician distribution. The probability of having a signal with magnitude less than a small threshold is considerably higher for the Rayleigh than for the Rician case. This is illustrated in figure 7. The Rayleigh fading model is thus usually considered to be a worst

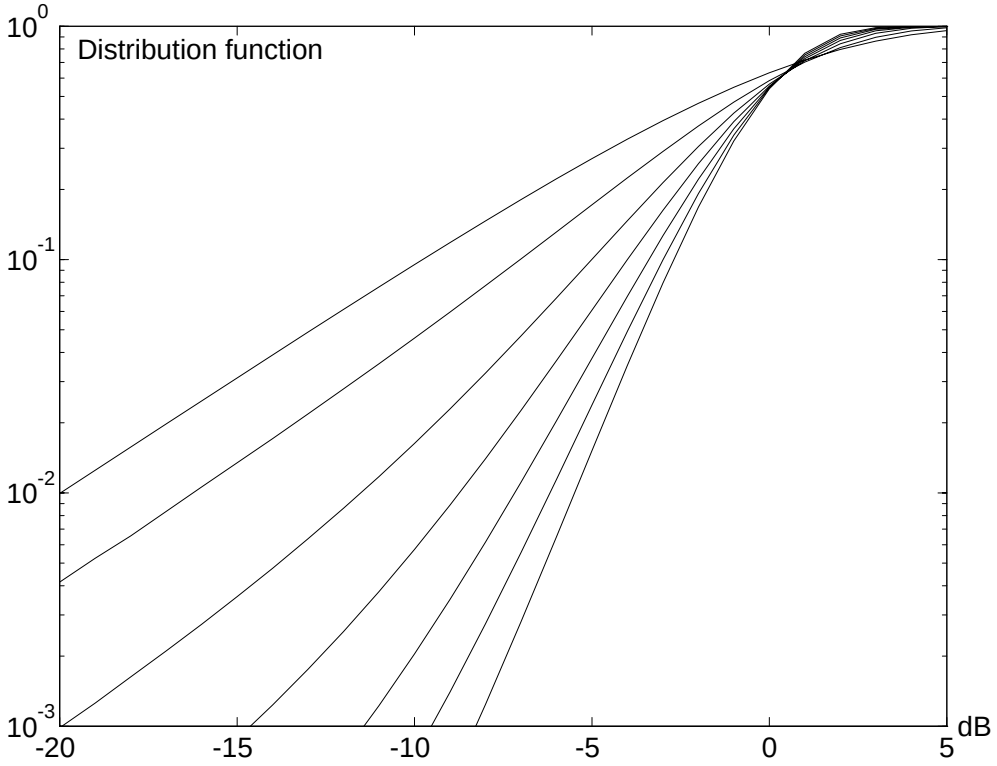


Figure 7: Probability distribution functions of Rayleigh (top curve) and Rician random variables. The horizontal units are in dB relative to RMS value. The Rician curves are for ratios of direct to scattered powers equal to 2, 4, 6, 8, 10 and 12.

case situation, and it is widely used for that reason. Due to its simplicity, we will use it in this tutorial, but one should recall that it can be unduly pessimistic when a direct path is present.

The channel impulse response is thus taken to be a two dimensional 0 mean Gaussian random process. $c(\tau, t)$ is actually a function of the positions of the transmitters and receivers, and this complicates the matter. Assume we have a fixed transmitter. By making use again of the Central Limit Limit, one can argue that $c(\tau, t)$ is jointly Gaussian over time and space. However the impulse response seen at a "random point in space" is a linear combination of Gaussian distributions, which will not be Gaussian except if the variance is constant over space. This will turn out to be important later.

The random process $c(\tau, t)$ is characterized by the correlation function

$$R_c(\tau_1, \tau_2, t_1, t_2) = E(c^*(\tau_1, t_1)c(\tau_2, t_2)) \tag{11}$$

If one assumes that the channel is stationary, (or at least Wide Sense Stationary) the correlation function depends only on $\Delta t = t_2 - t_1$.

This is easy to justify for the time scales of interest here if the mobile is stationary and the variations are due to movement of the ionosphere or to the traffic in the vicinity. However stationarity in presence of movements of the mobile requires more assumptions, e.g. on the spatial stationarity of $c(\tau, t)$ and on the way the mobile moves.

To make progress one also assumes that reflections occurring with different delays are uncorrelated. This is called *Uncorrelated Scattering*. It is harder to justify, as different paths may have some segments in common. Mathematically this results in

$$R_c(\tau_1, \tau_2, t_1, t_2) = \phi(\tau_1, t_2 - t_1)\delta(\tau_1 - \tau_2) \tag{12}$$

and the channel is completely characterized by $\phi(\tau, \Delta t)$.

In particular $\phi(\tau, 0)$ is the average power response of the channel as a function of the time delay τ. It is called the "delay power spectrum", which is unfortunate because it connotes the notion of frequency, and also the "multipath intensity profile", which describes the function much better. In can be estimated by transmitting very wide band signals and cross-correlating two delayed versions of the received signal, or by observing the output of a matched filter, as described above.

Measurements reveal that $\phi(\tau, 0)$ is only significant over a finite range of values, called the "multipath spread" of the channel, denoted by T_s.

T_s is directly related to the maximum difference between the propagation times of the significant echos, as already discussed above.

Earlier we have introduced the instantaneous frequency response $C(f, t)$. Looking at it from a statistical point of view, the frequency response is itself a Gaussian random process in f and t, as it is the linear transformation of a Gaussian process. Its correlation function is given by

$$\begin{aligned}
R_C(f_1, f_2, t_1, t_2) &= EC^*(f_1, t_1)C(f_2, t_2) \\
&= \frac{1}{2}\int_{-\infty}^{\infty} d\tau_1 \int_{-\infty}^{\infty} d\tau_2\, R_c(\tau_1, \tau_2, t_1, t_2)e^{j2\pi(f_1\tau_1 - f_2\tau_2)} \\
&= \int_{-\infty}^{\infty} d\tau_1 \int_{-\infty}^{\infty} d\tau_2\, \phi(\tau_1, t_2 - t_1)\delta(\tau_1 - \tau_2)e^{j2\pi(f_1\tau_1 - f_2\tau_2)} \\
&= \int_{-\infty}^{\infty} d\tau_1\, \phi(\tau_1, t_2 - t_1)e^{j2\pi(f_1 - f_2)\tau_1} \\
&= \Phi(f_2 - f_1, t_2 - t_1)
\end{aligned}$$

where $\Phi(f, \Delta t)$ is the Fourier transform of $\phi(\tau, \Delta t)$. The stationarity in f is a result of the assumption of uncorrelated scattering. $\Phi(\Delta f, \Delta t)$ is called the *spaced-frequency spaced-time* correlation function of the channel. It can be measured directly by transmitting two

tones separated by Δf and cross-correlating the two separately received signals with a relative delay Δt.

In particular $\Phi(\Delta f, 0)$ gives an indication of the coherence of the channel in the frequency domain. It is the Fourier transform of $\phi(\tau, 0)$, which is time limited to $\approx T_s$, thus it is frequency limited to $\approx 1/T_s = F_c$. F_c is called the *coherence bandwidth* of the channel.

We have discussed the τ parameter in $\phi(\tau, \Delta t)$, we now have to consider the Δt, i.e. how time invariant the channel response is. Intuitively, one should be able to define a "coherence time" T_c over which the channel doesn't change. For $\Delta t < T_c$ one expects $\phi(\tau, \Delta t)$ to be close to $\phi(\tau, 0)$.

As we have seen before, variations of the channel response will result in a broadening or a frequency shift (due to the Doppler effect) of the transmitted signal spectrum. It is thus meaningful to measure the expected power density received at frequency f_2 when a tone at frequency f_1 is transmitted.

If a tone $e^{j2\pi f_1 t}$ is transmitted, the received signal is from (9)

$$y(t) = C(f_1, t)e^{j2\pi f_1 t} \tag{13}$$

It is again a linear transformation of a Gaussian process, thus a Gaussian process itself. Its correlation function is given by

$$\begin{aligned} R_y(t_1, t_2) &= EC^*(f_1, t_1)C(f_1, t_2)e^{j2\pi f_1(t_1 - t_2)} \\ &= \Phi(0, t_2 - t_1)e^{j2\pi f_1(t_1 - t_2)} \end{aligned}$$

thus it is stationary and its spectral density at f_2 is given by

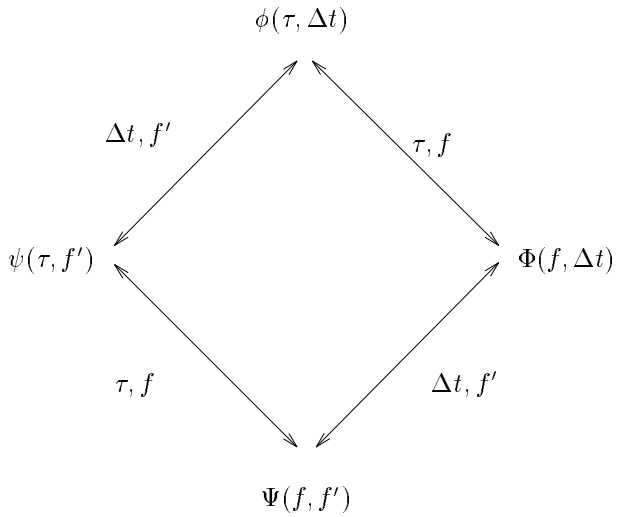
$$\begin{aligned} S_y(f_2) &= \int_{-\infty}^{\infty} dt\, R_y(t)e^{-j2\pi f_2 t} \\ &= \int_{-\infty}^{\infty} dt\, \Phi(0, t)e^{-j2\pi(f_1 - f_2)t} \\ &= \Psi(0, f_2 - f_1) \end{aligned}$$

where $\Psi(f, f')$ is the Fourier transform (over Δt) of $\Phi(f, \Delta t)$.

This time we see that the power density at f_2 when a tone at f_1 is transmitted depends only on $f_2 - f_1$. It can be measured by frequency domain wideband sounders. $\Phi(0, f')$ is called the *Doppler spectrum of the channel*. A time invariant channel has a Doppler spectrum concentrated at 0. If a vehicle is moving at a steady pace and receives a single signal, the Doppler spectrum will be concentrated at the Doppler frequency. Usually $\Phi(0, f')$ is only significant for f' in a small range called the Doppler spread $F_s \approx 1/T_c$.

There remains one Fourier transform to define, $\psi(\tau, f')$, which is both the Fourier transform of $\phi(\tau, \Delta t)$ with respect to Δt, and is also the inverse Fourier transform of $\Psi(f, f')$. This function is called the "scattering function" of the channel and it measures the channel response vs. Doppler frequency and time spread .

The relationship between the transforms appear in figure 8.



$\phi(\tau, 0)$ = Multipath intensity profile
(delay power spectrum)

$\Phi(f, \Delta t)$ = Spaced-frequency spaced-time correlation function

$\Phi(0, \Delta t)$ = Spaced-time correlation function

$\Psi(0, f')$ = Doppler power spectrum

$\psi(\tau, f')$ = Scattering function

Figure 8: Relationship between the correlation functions

In simulations a simplified model is often adopted. When a mobile moves at velocity v in a direction α with respect to an incident wave, the doppler shift is $-f_c v \cos(\alpha)/c$. For a pure carrier at frequency f_c, the complex signal envelope is then

$$x(t) = Ae^{-j2\pi f_c \frac{v}{c} t \cos(\alpha)}. \tag{14}$$

Reflections of the signal occur near the receiver, and the angle of arrival can be considered as uniformly distributed on $[0, 2\pi]$. The correlation of the received signal is

$$\begin{aligned} R_x(t_1, t_2) &= E|A|^2 e^{j2\pi f_c \frac{v}{c}(t_1 - t_2)\cos(\alpha)} \\ &= |a|^2 J_0(2\pi f_d (t_1 - t_2)) \end{aligned}$$

where J_0 is a Bessel function and $f_d = f_c v/c = v/\lambda$.

We see that $x(t)$ is a WSS process, and its spectral density is given by

$$S_x(f) = \begin{cases} \frac{|A|^2}{2\pi f_d} \frac{1}{\sqrt{1 - \left(\frac{f}{f_d}\right)^2}} & |f| \leq f_d \\ 0 & \text{else} \end{cases} \tag{15}$$

A normalized plot of $S_x(f)$ appears in figure 9.

The previous model is used in computer or hardware simulations. There the signal is broken in a number of branches that are delayed and multiplied by complex Gaussian random processes with the spectral density (15), with amplitudes dependent of the branch. The GSM standard defines several such channels, representing different propagation conditions (Urban, suburban, ...). A block diagram of such a model with 3 paths appears in figure 10.

3.4 Channel classification

One of the key consideration when using a fading channel is whether its response can be measured and used in adaptive receivers to provide good noise immunity. For this to be possible, the spread of the response (T_s) must be much smaller than the coherence time of the channel (T_c). A channel that can be measured is said to be *underspread*. If it cannot it is said to be *overspread*. The parameters of the urban mobile channel used as example above make it underspread ($T_s < T_c$, or $T_s F_s < 1$). Most radio channels also fall into that category, with the exception of long distance communication with fast flying airplanes.

As discussed before, the channel response can be measured by sending tones and estimating the frequency response (with a resolution inversely proportional to the duration of the signal, thus of $1/T_c$), or by sending a wide band signal of bandwidth W measuring the time response directly with a resolution $1/W$. We call a "wideband" a signal with a bandwidth much bigger than the inverse of its time duration. We will study them more in the context of Spread Spectrum systems. The correlation and power spectra can be obtained by averaging, and they can be computed from each other by the relations given above.

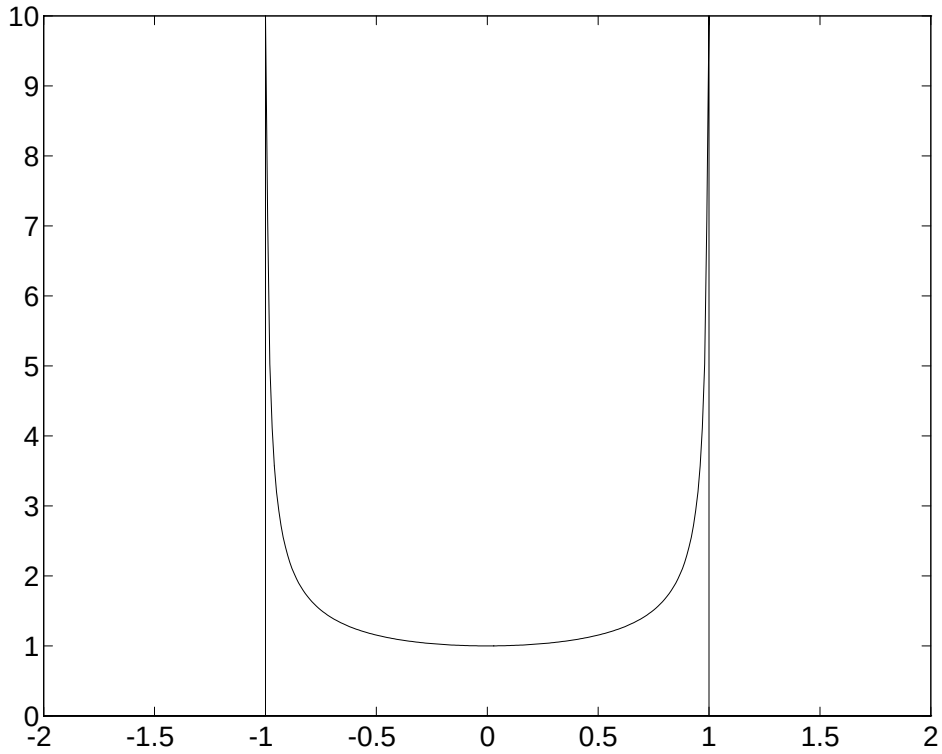


Figure 9: Spectral density $S_x(f/f_d)$

The parameters T_s and T_c (or F_c and F_s) are also important in relation with the signaling format. If the basic pulses carrying the information last for a time T less than T_s they will be subject to Intersymbol Interference (ISI), as reception from neighboring pulses will overlap. Its bandwidth will be more than $F_c \approx 1/T_s$, and the signal will see a frequency response that is non constant and may have deep nulls. One then says that the channel is "frequency-selective". This qualifier is somewhat misleading as it applies to the channel with respect to a specific signal.

If the signal lasts much longer than T_s then ISI will not be a major concern. If that signal is narrowband, its bandwidth will be narrower than the coherence bandwidth F_c and it will see a flat channel response. One says that the channel is "frequency-nonselective". Not having ISI is good of course, but having a "frequency-nonselective" channel is like putting all of one's eggs in one basket. It is an all or nothing proposition. If the signal is in the fade, reception will be very bad. One can remedy the situation by coding.

Implicit in the previous paragraph was the fact that the signal had to last less than "T_c", the coherence time of the channel, otherwise it does not make sense to talk about a frequency response. The conditions $T \gg T_s$ and $T \ll T_c$ are only possible together when the channel is underspread. When they hold one says that the channel is a "slowly fading channel". Again, this is with respect to a specific signal.

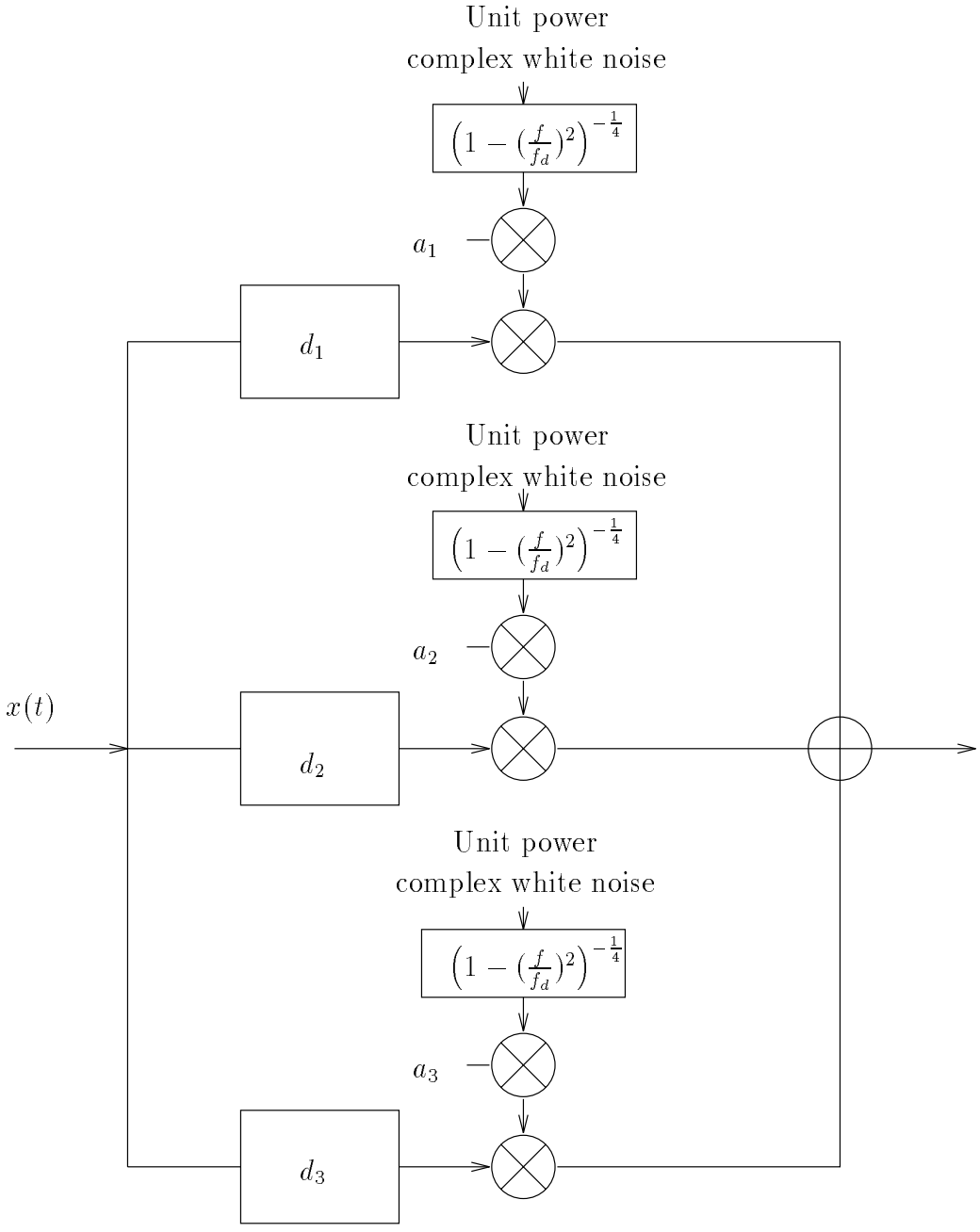


Figure 10: Block diagram of a mobile radio channel simulator

3.5 Wide area statistical description of the radio channel

In this section we consider the statistical characterization of a channel over a wide area. In line with the discussion in the previous section, the description is decomposed in two parts, one about "small areas" and one on "large areas".

The small area has a radius of a few tens of wavelength, and one admits that the statistics of the received signal is position invariant in that area. The various correlation functions can then be estimated by sounding the channel repeatedly within the area. Measurements often support the Rayleigh model when there is no direct path. We have seen that the mean power transfer depends on the paths that are locally significant.

Large areas usually cover regions having consistent physical characteristics, for example a downtown area, or a suburb. One can then take the average of the local correlation functions over the large area, and define concepts such as average multipath intensity profile, coherence bandwidth, etc...

The mean power transfer of the local areas is handled differently, due to its importance and to the fact that it exhibits wide fluctuations. Its value depends on local conditions, particularly mask effects of buildings, hills, etc.., that are known as *shadowing*.

One first attempts to take distance into account by calculating the average received power versus distance. This is often fitted to curves that run the gamut between simple relationships like $d^{-\alpha}$, where α often turns out to be a number between 3 and 4, and the sophisticated model due to Hata, based on measurements made by Okumara in the Tokyo area. The Hata-Okumara formulas predict the attenuation based on the distance, the heights of the antennas, the frequency of operation, and the type of terrain.

Distance alone does not suffice to predict the small area average received power, and the remaining discrepancy is characterized by an empirical probability distribution. It is often reported to be approximately log-normal, i.e. it is Gaussian when expressed in dB. This can be somewhat justified by the central limit theorem, by arguing that the reception conditions depend on the product of the attenuations in the areas traversed by the signal.

The standard deviation σ of the underlying Gaussian distribution is expressed in dB, typical values range between 5 and 10 dB. It is an important number, as it determines *coverage*, i.e. the percentage of an area that does not suffer attenuation above some threshold. Let the expected attenuation at distance d be $A(d)$. Half of the area located at distance d actually sees a larger attenuation. 16% of the area suffers an attenuation greater than $A(d) + \sigma$, 1% and attenuation greater than $A(d) + 2.3\sigma$.

Now that we understand better the propagation condition, we will focus on the cell layout. The concept of cell is central to modern mobile systems and determines both the economics and traffic carrying capabilities of the network.

4 The cellular concept

Frequency reuse is a central part of the design and operation of modern mobile networks. It is that feature that distinguishes them from older systems where a central transmitter serves a very large geographical area. In cellular systems, a large area is divided into "cells". One talks of "macrocells" in the case of mobile telephony systems designed for car use, such as AMPS and GSM. There the cell radius varies between, say, 500 m and 35 km. Systems designed mostly for outdoor use by pedestrians have "microcells". Their radius is of the order of tens of, or a few hundreds of, meters. Finally "nanocells" are designed for indoor use, and are of the size of buildings. Contrary to their larger siblings, nanocells are 3 dimensional. Their extend is limited vertically by the building floors.

A number of frequencies are used within a cell. As radio waves attenuates with distance, frequencies can be reused in distant areas. For discussion purposes, it is useful to conceptually partition a plane terrain in hexagonal regions, as sketched in figure 11, and to take their radius as the unit of length. The choice of the hexagon comes purely from the fact that it is the geometric shape closest to a circle that paves the plane.

The set of available frequencies can also be partitioned in subsets, numbered from 1 to K, and allocated to the cells. Sometimes it is necessary to add extra constraints. For example in some systems the emission spectrum of the signals overlaps neighboring channels, thus adjacent frequencies cannot share the same, or even adjacent cells. We do not consider the extra constraints here, but we focus on allocating the frequencies to maximize the minimum distance between cells using the same set of frequencies.

In a regular hexagonal pattern, the vector joining the center of regions sharing the same frequency has the form $iu_v + ju_d$, where i and j are integer, u_v is a vertical vector of length $\sqrt{3}$ and u_d is a diagonal vector with a slope of 60° with the same length. It follows that the minimum distance between matching centers has the form $d = \sqrt{3(i^2 + ij + j^2)}$. The set of points in the plane closest to the center of one of these hexagons is itself an hexagon (indicated with dashed lines in the bottom parts of figure 11) of radius $\sqrt{i^2 + ij + j^2}$. The number of sets is thus given by

$$K = i^2 + ij + j^2 \tag{16}$$

and one has the relation

$$d = \sqrt{3K}. \tag{17}$$

Figure 11 displays reuse patterns with $K = 3, 4, 7, 9$. If a system can support F frequencies, and each frequency can support T simultaneous calls, each cell can support FT/K simultaneous users. From economic arguments, it is clear that K should be minimized.

How small can K be made? It is essentially limited by interference from neighboring cells. Consider a user at the periphery (distance 1) of its cell receiving from a base station at the center. It also receives weaker signals from the 6 closest cells sending at the same frequency, and from others beyond. The 6 closest cells are about at distance d. If one

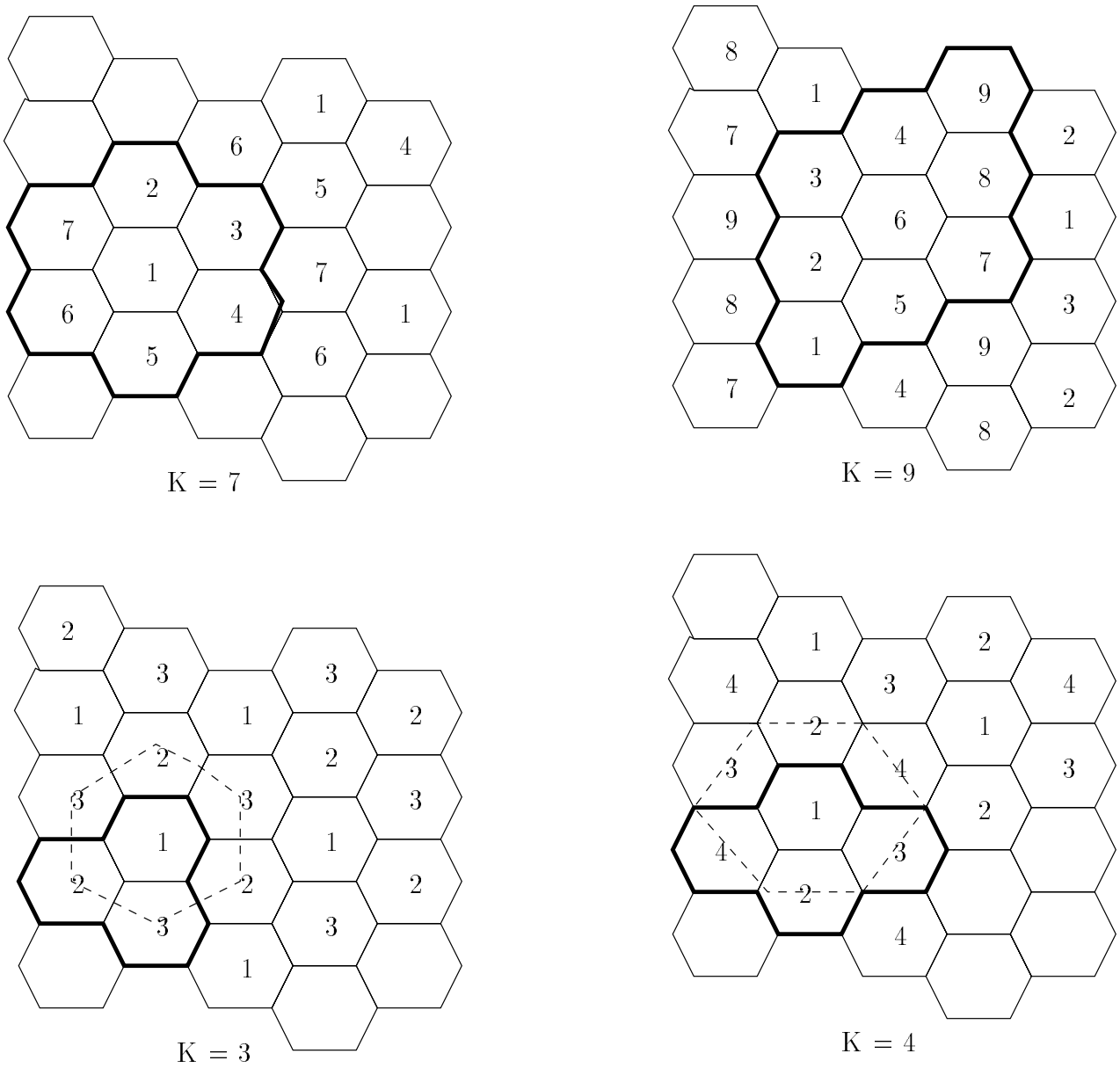


Figure 11: Hexagonal reuse patterns

admits that power decays with $d^{-\alpha}$ [1], the interference power to carrier power ratio will be

$$I/C = 6d^{-\alpha} = 6(3K)^{\alpha/2}, \tag{18}$$

and

$$K = \frac{1}{3}\left(\frac{6C}{I}\right)^{\frac{2}{\alpha}} \tag{19}$$

α is usually taken as a number between 3.5 and 4 in urban areas. One conclusion is clear: the larger the propagation α, the smaller K.

How large K can actually be still depends on C/I. The value of $C/I = 18$ dB is usually quoted for analog FM systems. Designers of digital systems hope to operate at much smaller values, possibly 7 to 10 dB. The reader will recall from the introduction that the efficiency of digital systems, measured in kHz of bandwidth per voice channel, was not that different from the efficiency of analog systems. The discussion in this section introduces the concept of frequency reuse K, which increases the appeal of digital systems.

The previous approximation has ignored a number of factors. Geometric effects, such as the neglect of the distant interferers, and the inequality of the distances from a point to the neighboring cells, are immediately noticeable. However they do not matter very much except for small K. The main factor that we have ignored is the effect of fading.

In absence of dynamic power control tracking the slow lognormal fading due to shadowing, the carrier level must be increased by a level between 1 and 2 times the standard deviation (in dB), depending of what fraction of the cells must have satisfactory service. This adds about 7 to 10 dB to the required C/I. Of course the level of I also varies due to interference, but the impact is reduced by the law of large numbers (in this case, large is only 6!).

With power control, C is much more stable, but the dynamics of I is increased, as it depends now on two attenuations: on the desired propagation path of the interfering signals, and on the undesired interference path. Overall the effect is beneficial.

Other effects should also be taken into account.

Hand-offs should have a large impact as they tend to keep a mobile transmitting to the "best" base station, usually in the sense of minimizing the required power. This reduces I. Frequency hopping on a burst basis (frequency hopping will be further detailed below), together with coding, also have a beneficial effect. In heavy traffic, when all frequencies and time slots are in use, frequency hopping reduces the fluctuations of I, but not its average level. Essentially it increases the number of interferers but decrease their individual effects, so that the law of large numbers helps the situation. How much it does depends on the amount of interleaving and coding.

The previous calculations can best be qualified of speculations! The models are crude and many of the key parameters are unknown. They can be manipulated at will by

[1] 60 GHz is seriously considered as an operating frequency for picocells because it has a very strong absorption line of oxygen, about 14 dB/km. Attenuation increases exponentially with distance, and the $d^{-\alpha}$ model does not apply. Excellent frequency reuse is essential in picocells.

proponents or opponents of various systems. Analog FM systems are reportedly used with a K of about 21. GSM systems hope to operate with K between 4 and 12.

The previous discussion on increasing K has focused on maximizing the number of frequencies available per site, which is a key economic consideration for the operators. The initial promise of the cellular concept was different. It was to allow arbitrary high traffic in a given area by using more cells, each with a smaller area. Thus this would make best use of the frequency spectrum, viewed as a public commodity. Unfortunately, that promise has not been fully realized.

Dividing cells has proven to be more difficult that anticipated. Firstly, as a cell becomes smaller, the base station must be located at the center with a relative precision that remains constant, thus with an absolute precision that increases. It is not always easy to find to a convenient and affordable site at arbitrary spots in the center of a city.

Secondly, the attenuation parameter α is actually dependent on distance, and its gets closer to 2 (from 4) when the cell radius is diminished. In the case of microcell, the exponent even appears to decrease to 0, as the street can form a waveguide. This reduces the benefits of the smaller cells.

Thirdly, as cells get smaller, hand-offs become more frequent, loading the signaling system and increasing its cost. This is due both to the smaller size of the cell, and to the fact that their shape are farther from circular, being more determined by obstacles than by path length loss. Thus a larger fraction of the area cell is close to its periphery.

These reasons have made it uneconomical (or unfeasible) for operators to reduce cell diameters below about 500m.

We can illustrate the consequences of these numbers by a numerical example. The area of Paris within the circumferential highway has a radius of about 5 km. It could be covered by about 100 cells with a 500 m radius. In GSM at 900 MHz, each cell would offer at most 124 frequencies, each supporting at most 8 (later 16) calls. With 8 slots and $K = 12$, the system can support $100 \times 124 \times 8/12 = 8266$ simultaneous calls. If each subscriber makes a 2 minute call during the busy hour (and is willing to wait for a free channel), the number of subscribers will be at most $30 \times 8266 \approx 250000$.

When 16 slots are available, and if K can be decreased to 4, the number of calls increases to about 100000 and the number of subscribers to 3 millions. On the other hand, if the cell radius is increased to 1 km, all the previous numbers must be divided by 4, reducing the number of subscribers to 62500 or 750000.

The significance of these numbers can be debated, but they make clear that within engineering factors that do not seem to be well understood, the system varies from giving service to a few percents of the population, to being almost universal. Whatever the capability of the current system, the next one is likely to provide universal service.

5 Communication Theory

We have seen a previous section that the mobile radio channel is subject to fading. In this section we explore the consequences of fading on the communication performances, and we examine the measures one can take to reduce the degradation caused by fading. We use the tools and results from communication theory, which is concerned with the modulation and detection of digital signals transmitted over noisy channels. We will first review classical communication theory on the additive white Gaussian noise channel, then extend the results to fading channels.

5.1 Modulation methods and signal space

The view we take is that we transmit a message i, possibly representing many bits, by sending a sequence of symbols, b_k^i. The value of a discrete symbols b are communicated by sending a waveform $h(b,t)$. The first argument indicates that the waveform depends on the value of the symbol, the second indicates the dependency on time. The waveform corresponding to the ith message is

$$\begin{aligned} x^i(t) &= \sum_k h(b_k^i, t - kT) \\ x'^i(t) &= \Re(\sum_k h(b_k^i, t - kT) e^{j2\pi f_c t}) \end{aligned}$$

The first form corresponds to the complex envelope, the second to the modulated signal. One sees that the waveform corresponding to the kth symbol is delayed by kT. $1/T$ is known as the *baud rate*.

In simple cases $h(b,t)$ takes the simple form

$$h(b,t) = bh(t) \tag{20}$$

i.e. the modulation is linear and the system is known as *Pulse Amplitude Modulation*. $h(t)$ might be a simple pulse, like

$$h(t) = \begin{cases} 1 & t \in [0,T] \\ 0 & \text{else} \end{cases} \tag{21}$$

This $h(t)$ has a spectrum that is not well confined. Other pulse shapes, such as *root raised cosine*, have better spectral properties and are often used when the channel is bandlimited, or when there are constraints on out-of-channel emissions.

Because $h()$ is part of a complex envelope, it can be complex and so can the b_k^i in the linear case. The name of the modulation is then *Quadrature Amplitude Modulation*. For example, binary phase modulation (or antipodal signaling) corresponds to having $b = \pm 1$, and quaternary phase modulation corresponds to $b = \pm 1$ or $\pm\sqrt{-1}$.

Non linear modulations are also used. One of the most popular is *binary orthogonal Frequency Shift Keying* where the symbols b are binary, say 0 and 1, and

$$\begin{aligned} h(0,t) &= 1,\ t \in [0,T] \\ h(1,t) &= e^{j2\pi nt/T},\ t \in [0,T] \end{aligned}$$

Those waveforms have a frequency difference n/T, where n is the modulation index. One verifies that $h(0,t)$ and $h(1,t)$ are orthogonal when n is an integer multiple of $1/2$.

The binary orthogonal case can be generalized to M-ary by introducing more waveforms, and it can be generalized to the bi-orthogonal as follows in the two dimensional case:

$$\begin{aligned} h(0,t) &= 1,\ t \in [0,T] \\ h(1,t) &= -1,\ t \in [0,T] \\ h(2,t) &= e^{j2\pi nt/T},\ t \in [0,T] \\ h(3,t) &= -e^{j2\pi nt/T},\ t \in [0,T] \end{aligned}$$

The terms "linear" and "non linear" modulations are traditional and come from the world of analog systems. There is no fundamental difference between the two in the case of digital transmission. The function space spanned by the $h(a,t)$ simply has dimension one or two for PAM and QAM, dimension two for binary orthogonal and bi-orthogonal systems, and M for M-ary orthogonal.

5.2 The additive white Gaussian noise channel

One of the simplest and most useful channel models is that of the additive white Gaussian noise channel. There the received waveform is equal to the transmitted waveform (possibly with some attenuation) to which white Gaussian noise with spectral density $N_0/2$ has been added, i.e.

$$y(t) = Ax(t) + n(t) \tag{22}$$

Detection theory teaches that if all messages are equally likely (we will always make this assumption) then the receiver minimizing the message error probability is a correlation receiver. In principle it generates locally all the possible waveforms $x^i(t)$ and finds the one that minimizes

$$\int dt\, |x^i(t) - y(t)|^2 \tag{23}$$

We will denote this integral by $\|x^i - y\|^2$, it is induced by the innerproduct

$$\langle x(t), y(t) \rangle = \int dt\, x^*(t) y(t) \tag{24}$$

Instead of finding the message i that maximizes (23) one can as well maximize

$$\Re(\langle y(t), x^i(t) \rangle) - \frac{\|x^i(t)\|^2}{2} \tag{25}$$

The probability that a message i will be incorrectly decoded into another message j is then upperbounded by

$$Q(\sqrt{\frac{\|x^i - x^j\|^2}{2N_0}}) < e^{-\frac{\|x^i - x^j\|^2}{4N_0}} \tag{26}$$

where $Q(z)$ is the complementary distribution function of a zero mean, unit variance Gaussian random variable

$$\begin{aligned} Q(z) &= \int_z^\infty du \, \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \\ &\leq \frac{1}{2} e^{-\frac{z^2}{2}} \end{aligned}$$

The upperbound exhibits the correct exponential behavior.

This form of the receiver is impractical because there are too many possible messages over which to minimize (23) or to maximize (25). A slight conceptual simplification occurs if one observes that (25) can be computed from

$$y_{b,k} = \langle y(t), h(b, t - kT) \rangle, \ \forall b, k. \tag{27}$$

A long correlation has been broken into many shorter ones.

Great simplifications occur if $h(b,t)$ is orthogonal to $h(b', t - kT)$, $k \neq 0$, i.e. if there is no Intersymbol Interference, and if the b_k^i are statistically independent. In that case it is optimal to do symbol by symbol decoding, i.e. to make independent decisions about each of the symbol b_k^i. To that effect the receiver minimizes $\|y(t)h(b_k^i, t - kT)\|^2$ or equivalently maximizes over b_k^i

$$\Re(\langle y(t), h(b_k^i, t - kT) \rangle) - \left\| h(b_k^i, t) \right\|^2 / 2 \tag{28}$$

The symbol error probability is then given by

$$P_e = \alpha Q(A \min_{b \neq b'} \sqrt{\frac{\|h(b,t) - h(b',t)\|^2}{2N_0}}) \tag{29}$$

where α is a number between $1/M$ and 1, where M denotes the number of possible values for b_k^i.

In particular, for antipodal modulation where $\mathcal{E}$ denotes the signal energy at the receiver, $\mathcal{E} = A^2 \|h(t)\|^2$

$$P_e = Q(\sqrt{\frac{2\mathcal{E}}{N_0}}) \tag{30}$$

while for binary orthogonal modulation

$$P_e = Q(\sqrt{\frac{\mathcal{E}}{N_0}}) \tag{31}$$

One sees that binary orthogonal modulation is not as energy efficient.

5.3 Additive noise

The previous section has focused on the importance of the signal to noise ratio $\mathcal{E}/N_0$. In final analysis, it is the additive noise and not the fading that ultimately limits the performances of mobile communications systems. If there was no additive noise, fading would not matter as arbitrarily small signal could be detected accurately. In this section we review the origin of the Gaussian noise, and give typical values.

There are two main sources of additive noise. One is interference from other users in the same system, and that one is managed through frequency reuse schemes and power control. The other is background radiation noise. Studies have shown that the background radiation noise is close to the basic limits predicted by physics, i.e. it is a white and Gaussian noise with one sided spectral density kT, where k is Boltzmann's constant, $1.3810^{-23} J/K$ and T is the temperature is Kelvin.

Mobile communications antennas are directed at the terrestrial environment, and not at the sky (which is cooler), thus T can be taken as about $300K$ and $kT \approx 4.1410^{-21} J = -203.8$ dB with respect to a Joule. In addition to the background radiation noise, the input stages of the receiver also generate noise. The quality of the receiver is measured by the Noise Figure which gives the ratio of the total noise to the nominal background noise discussed earlier. Good receivers have noise figures of a few dB, such as 6. Thus the one sided noise spectral density level N_0 is about -197.8 dB relative to a Joule.

For reliable communication in absence of fading, and without coding, the energy per bit must be about $\mathcal{E} = 10N_0$. Useful bit rates are typically about 10^4 bits per second. In that case the received power must be -147.8 dBW, where a dBW is a dB relative to a Watt. However systems using TDMA must be capable of receiving at a higher bit rate, 277 kB/s (uncoded) in the case of GSM. To receive that bit stream reliably requires -133.4 dBW, or -103.4 dBm. The specifications actually require -104 dBm, in excellent agreement with the previous derivation.

The amount of attenuation that can be tolerated depends on the transmitted power of course, but one sees that it can be considerable. In our scenario a 1 mW transmitter can tolerate about 117.8 dB of attenuation for a 10 kb/s bit rate.

This attenuation ultimately limits the useful range of the transmitter, although propagation delays can set lower limits in TDMA systems. Receivers located well within the range see signals that are much stronger than required, and they would enjoy almost perfect reception, if it was not for interference and fading.

5.4 Effect of channel dispersion. Rake receiver

In the ideal case considered before, the channel only adds noise. We already know that a radio channel is characterized by a time variant impulse response $c(\tau, t)$. For the moment we will neglect the variation with t and we assume that $c(\tau)$ has been measured. From the point of view of the receiver, the only difference is that the received signals have the

form

$$h'(b,t) = c(t) * c(t) \tag{32}$$

and thus it should compute

$$\begin{aligned}
\langle y(t), h(b,t)\rangle &= \int dt\, y^*(t) h'(b,t) \\
&= \int dt\, y^*(t) (\int d\tau'\, c(t-\tau) h(b,\tau)) \\
&= \int d\tau\, (\int dt\, y(t) c(-(\tau - t)))^* h(b,\tau) \\
&= \langle y(t) * c^*(-t), h(b,t)\rangle
\end{aligned}$$

This last equation shows that one can as well pass the received signal through a filter with response $c^*(-t)$ and then process is as if there was no channel. The filter with response $c^*(-t)$ is called a *channel matched filter*.

Symbol by symbol detection is still optimal if $h'(b,t)$ is orthogonal to $h(b', t-kT)$, $k \neq 0$, i.e. if the channel has not introduced intersymbol interference. There is a subtle point here: there may not be significant intersymbol interference even if $c(t)$ extends over an interval greater than T. This is the case if $h(t)$ is a spread spectrum waveform of the type given in (6) (except that the duration T here is T/n there), and if $c(t)$ consists of discrete paths with delay separation different from a multiple of T. For example if

$$c(t) = a_1 \delta(t - d_1) + a_2 \delta(t - d_2) \tag{33}$$

then

$$\begin{aligned}
\langle h'(t), h'(t-kT)\rangle &= (|a_1|^2 + |a_2|^2) R_h(kT) + a_1^* a_2 R_h(kT - d_1 + d_2) + a_1 a_2^* R_h(-kT + d_1 - d_2) \\
&\approx 0 \text{ if } |kT - d_1 + d_2| > T/n
\end{aligned}$$

Even the restriction on path separations can be removed if the spreading pattern is changed for each baud.

We apply this theory to mobile radio, using the popular multipath model. Each path has an attenuation a and a delay d, thus

$$c(t) = \sum_j a_j \delta(t - d_j)$$

In that case the channel matched filter is itself given by

$$c(-t - d) = \sum_j b_j^* \delta(t + d_j - d) \tag{34}$$

where the delay d is chosen to be $\max(b_j)$ to make the response causal, and $R_c(t)$ is given by (neglecting the delay d)

$$R_c(t) = \sum_j |a_j|^2 \delta(t) + \sum_{i,j,\, i \neq j} a_i a_j^* \delta(t - \tau_i + \tau_j) \tag{35}$$

with a main impulse at $t = 0$ and many smaller impulses scattered on either side. This type of receiver is called a Rake receiver, it was first designed for military applications in the 1950's. Its anme evokes either the action of gathering all the paths into a single main impulse a 0 (which is not quite accurate as we will see), or the shape of its impulse response.

In the frequency domain, the concatenated channel has response $|C(f)|^2$, thus the main effect of the Rake receiver is to make the phase constant. In presence of fading, for example as displayed in figure 5 for a two path channel, all the peaks and valleys are still present, they are even emphasized. This seems unexpected, the large coefficient of the $\delta(t)$ term suggests a flat concatenated response in the frequency domain.

The implementation of a Rake receiver seems to require the capability to measure the impulse response of the channel, and thus a wideband signal. However if the transmitted signals $h(b,t)$ are narrowband, there is no advantage per se in measuring $c(t)$ and resolving all the paths. All that is required is to measure $C(f)$ over the bandwidth of the $h(b,t)$, and this can be done by using training sequences having the same bandwidth as the signal. There is an advantage to using wideband signals, they can lead to better performances as they provide natural "diversity" as we will see below.

As we have seen, in absence of intersymbol interference, the probability of error is given by (30) and (31) for antipodal and orthogonal signals respectively. For antipodal signals the relevant signal energy $\mathcal{E}$ is

$$\mathcal{E} = \int df\, |C(f)|^2 |H(f)|^2 \tag{36}$$

Again the main function of the channel matched filter is to make the concatenated filter response be positive to insure that the integral will be large. If $H(f)$ is narrowband with respect to the channel coherence bandwidth, $\mathcal{E}$ might be small. However if it is wide, then from equation (35),

$$\mathcal{E} \approx \sum_i |a_i|^2 \int df\, |H(f)|^2 \tag{37}$$

5.5 Systems with memory

Up to this point we have been able to treat systems with symbol by symbol detection. This is optimal when there is no intersymbol interference, and when adjacent symbols are independent. Often neither of these conditions need to hold.

For example, the Minimum Shift Keying modulation is an orthogonal frequency modulation with frequency deviation $n = 1/2$ (the minimum possible to be orthogonal), and continuous phase. The phase continuity yields a narrower spectrum, but it poses a problem. Depending on what symbol has been transmitted, the phase of the two possible signals at the end of the baud will differ by π. The initial phase of the signal in a baud will thus depends on the symbol in the previous baud.

The memory in the system can be captured by the concept of "state" at the end of the baud. The situation can be nicely pictured in a *trellis diagram* such as the one depicted

in figure 12 for MSK signals. The horizontal axis corresponds to time, while the states are the circles displayed in the vertical direction. In the case of MSK, there are only two states, 0 (the top state) and π, the bottom state. The lines between the states correspond to transmitted symbols, 0 for dashed lines, 1 for solid lines. The equations, denoted $h(e,t)$ below, next to the edges correspond to the waveforms transmitted during a baud. They depend not only on the symbol transmitted, but also on the state.

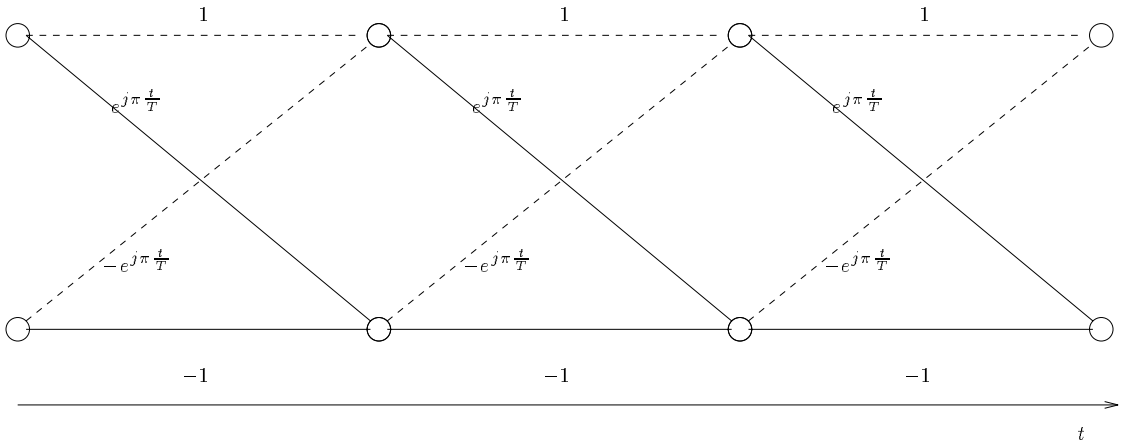


Figure 12: Trellis diagram for MSK signals

There are two important observations: firstly, to every message (i.e. sequence of symbols), corresponds one path through the trellis.

Secondly, the waveform $x^i(t)$ corresponding to the message can be "read" from the waveforms labeling the edges. For example in the case of figure 12 the message $i = (101)$ would have a waveform

$$\begin{aligned} x^i(t) &= e^{j\pi\frac{t}{T}},\ t \in [0,T] \\ &= -e^{j\pi\frac{t-T}{T}},\ t \in [T,2T] \\ &= 1,\ t \in [2T,3T] \end{aligned}$$

One sees that it is phase continuous as desired.

Thirdly, if the waveforms (properly shifted) in different sections of the trellis are orthogonal, then the message minimizing $\|x^i(t) - y(t)\|^2$ can be determined from the $\|h(e,t-kT) - y(t)\|^2$.

More precisely, one can label each edge with the weight $\|h(e,t-kT) - y(t)\|^2$ (using the corresponding e and k), and the desired path is the one with minimum weight sum. Finding the minimum is a shortest path problem in a loop free graph. It can be solved by Bellman's algorithm, which in this context is known as Viterbi's algorithm.

This method is applicable not only to memory introduced by the transmitter, as in the case of MSK, but also to memory introduced by the channel, i.e. intersymbol interference. The main limitation of trellis diagrams is that the number of states required is M^L, where

M is the number of possible symbols in each baud (2 in the case of MSK), and L is the memory of the channel (in bauds, only 1 in the case of MSK).

For example the GSM system was designed with intersymbol interference in mind. It operates with differential path delays of up to 15 μs, or about 5 symbols, thus $L = 5$ and 32 states may be required. It also uses a modulation system called GMSK (Gaussian MSK), a variant on the MSK modulation where the phase transitions are smoother but necessitates more memory. It appears that in some implementations the functions that "labels" the edges of the trellis are computed on the fly to take into account both the memory required by the modulation, and the measured channel impulse response. Of course signals are sampled, so that the continuous-time function is replaced by one or two samples.

As we have seen before, the probability of error between two messages i and j is determined by $\|x^i(t) - x^j(t)\|^2$. In the case of MSK in figure 12 one sees easily that this is $2T$, as long as all the sequences start in a given state, and end with a "flush symbol" to bring them in a given state. This $2T$ is to be compared to T for regular frequency shift keying (only one section of the trellis). Thus MSK makes up for the 3dB loss of frequency shift keying compared to phase modulation.

The MSK example illustrates also that performance can be enhanced by introducing dependencies between symbols. That is the object of coding theory. The dependency is expressed by the fact that fewer points in the signal space are used than are available. In traditional block and convolutional codes, this is achieved by increasing the dimension of the signal space, but keeping binary amplitudes in each dimension. In the case of Trellis Coded Modulation, the number of points is increased by adding extra levels. Thus MSK is a case of trellis code, where instead of using orthogonal signals during each baud, one uses bi-orthogonal signals.

We will say no more about coding, but will focus instead on the effect of fading.

5.6 Binary signaling over a slowly varying Rayleigh fading channel

This section studies the performance of binary orthogonal signals transmitted channel with additive white Gaussian noise and fading. The fading is taken as slow and frequency nonselective, i.e. the signal has a bandwidth less than the coherence bandwidth F_c of the channel. We let a denote the amplitude of the received signal.

Following equation (31), the probability of error conditioned on a is given by:

$$P_e = Q(\sqrt{\frac{a^2\mathcal{E}}{N_0}})$$

Up to now we have considered a as a constant. If there is slow Raleigh fading, a^2 has a chi-square distribution with two degrees of freedom, which turns out to be the same as

the exponential distribution. The previous expression for the probability of error can be considered as a conditional probability given a^2 and the unconditional error probability is then

$$\begin{aligned} P_e &= \int_{-\infty}^{\infty} du\, Q(\sqrt{\frac{u\mathcal{E}}{N_0}}) \frac{1}{\overline{a^2}} e^{-\frac{u}{\overline{a^2}}} \\ &= \frac{1}{2}\left(1 - \sqrt{\frac{\overline{a^2}\mathcal{E}}{2N_0 + \overline{a^2}\mathcal{E}}}\right) \end{aligned}$$

The quantity $\overline{a^2}\mathcal{E}/N_0$ is the average signal to noise ratio. When it is much greater than 1, as it must to get good performance, the error probability can be approximated by

$$P_e \approx \frac{N_0}{2\overline{a^2}\mathcal{E}} \tag{38}$$

thus it inversely proportional to the average signal to noise ratio, instead of decreasing exponentially with it. This is illustrated in figure 13. Thus if one desires an error prob-

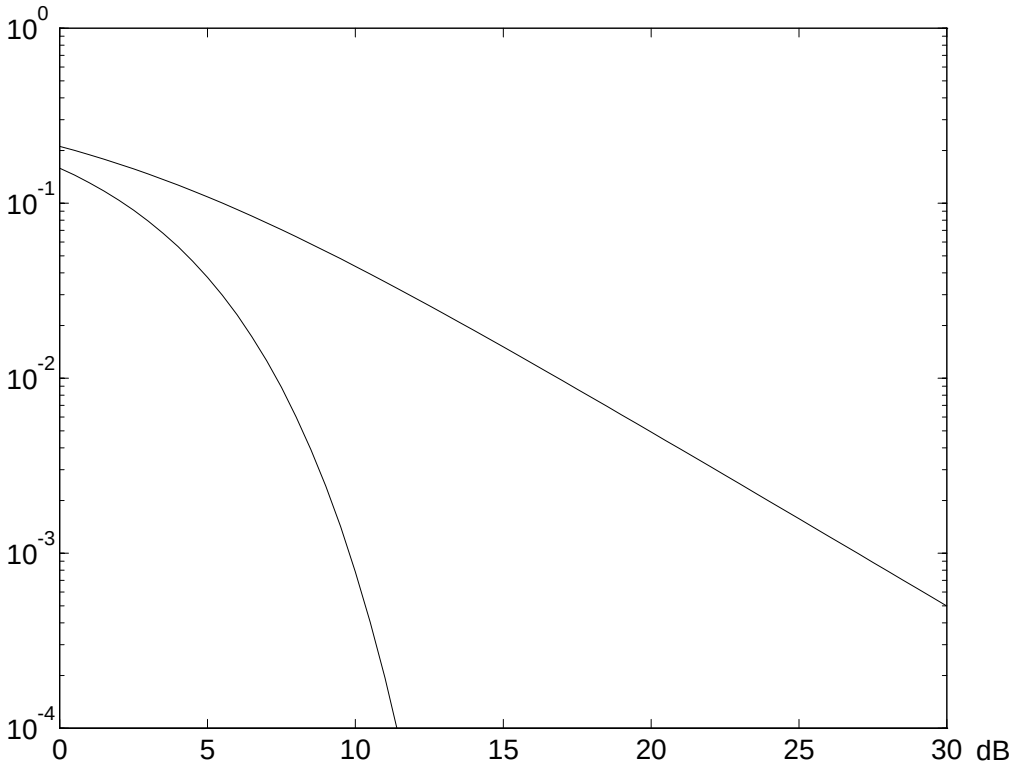


Figure 13: Probability of error of binary orthogonal signals with and without Rayleigh fading, versus the average signal to noise ratio

ability of 10^{-6}, common in land line communication systems, one needs a signal to noise ratio 13.5 dB (about 20) in absence of fading. In presence of Rayleigh fading, the same probability of error would require a signal to noise ratio of 5 10^5, or 57 dB, an increase by more than 4 orders of magnitude. More realistic error probabilities for fading channels are of the order of 10^{-3}. A signal to noise ratio of 10 dB suffices in absence of fading, 27 dB are necessary when fading is present.

Other modulation methods, such as phase modulation, require 3 dB less signal to noise ratio for the same error probability. Similarly, envelope detection methods, which do not rely on phase acquisition, only suffers a 3dB loss. Those small numbers pale in comparison with the influence of fading.

What is happening is clear. There are periods of time where the frequency nonselective channel is very close to a null, and power must be increased tremendously to lead to acceptable performances. The key to improve the situation is to reduce the likelihood of being in a deep fade. The main technique to this effect is called diversity transmission.

5.7 Diversity techniques

Diversity transmission techniques simply transmit the information on L independent fading channels. If the total power transmitted is constant, only a fraction $1/L$ can be used on any one channel. The likelihood that they will all be in a deep fade decreases geometrically fast with L, so that overall there is a significant gain.

A number of techniques are available to implement diversity. One possibility is to do it in the time domain, repeating the signal at intervals greater than T_c. This is known as *interleaving*. In a mobile environment T_c is often related to the speed of the mobile station, and having fixed users can cause systems based on time diversity to become ineffective.

Another possibility is to have the diversity in the frequency domain, transmitting the signal over many carriers spaced in frequency by more than the coherence bandwidth of the channel, $F_c \approx 1/T_s$. This is called *Frequency Hopping*. It is used in the GSM system.

A variation on this theme is to use a wideband signal (such as (6)) that spreads the information over a wide bandwidth. If a bandwidth W is used, one can expect to have a degree of diversity $L \approx W/F_c \approx WT_s$.

Yet another way is to use multiple antennas, separated enough to get independent replicas of the signal. A variation on this theme is to use co-located antennas sensitive to different components of the electric or magnetic fields. It is known that they can be chosen to fade independently.

Rather than simply using diversity transmission, coding should be used. In effect, each information bits is "spread" by coding over many signal dimensions, thus reducing the damage associated with each fade.

Finally, some or all the previous techniques can be combined. Virtually all second generation systems use time diversity under the form of interleaving, and GSM and the Qualcomm system use frequency diversity. In addition, base station are usually equiped

with two receive antennas. Currently only the stronger signal is processed, but more sophisticated algorithms are under study.

A unified analysis of diversity without coding is quite straightforward. Again one of two binary signals is sent. They can be quite a complicated signal, being composed of independent replicas, or being wideband. This time we will study the case of antipodal signal, or Binary Phase Shift Keying, instead of orthogonal signals. Thus we have a signal of the form $\pm h(t)$.

The received signal will be

$$y(t) = \pm \int d\tau \, c(\tau, t) h(t - \tau) + n(t) \tag{39}$$

where $n(t)$ is an additive white Gaussian noise process. Using the slow fading model, we assume that the receiver can measure $c(\tau, t)$, or more simply, can estimate the waveforms

$$h'(t) = \int d\tau \, c(\tau, t) h(t - \tau) \tag{40}$$

that would be received in absence of noise. In presence of additive white Gaussian noise, an optimal receiver will correlate $y(t)$ with $h'(t)$, obtaining

$$\begin{aligned} r &= \int dt \, h'(t) y(t) \\ &= \pm \int dt \, h'^2(t) + \int dt \, h(t) n(t) \\ &= \pm \mathcal{E}' + n \end{aligned}$$

where $\mathcal{E}'$ is the received signal energy and n is a zero mean Gaussian random variable with variance $\mathcal{E}' N_0/2$. An optimal receiver will decide 0 if $r > 0$, and 1 otherwise. If 0 is transmitted and error occurs if $n < -\mathcal{E}'$, thus conditioned on $\mathcal{E}'$ the proability of error is

$$P_e = Q(\sqrt{2\mathcal{E}'/N_0})$$

This formula displays the 3 dB advantage of antipodal signals over orthogonal signals. Up to this point we have essentially repeated the reasoning leading to (30).

We are now faced with the task of averaging P_e over the statistics of $\mathcal{E}'$. In the previous section, we had $\mathcal{E}' = a^2 \mathcal{E}$, where a^2 had a chi-square distribution. That results generalizes to the current situation. With the Wide Sense Stationary, Uncorrelated Scattering model of the channel outlined previously, the signal $h'(t)$ at the receiver is a Gaussian random process. Its correlation function $R_{h'}(t_1, t_2)$ can be expressed in terms of the correlation function of the channel response:

$$\begin{aligned} R_{h'}(t_1, t_2) &= E \int_{-\infty}^{\infty} d\tau_1 \, c^*(\tau_1, t_1) h^*(t_1 - \tau_1) \int_{-\infty}^{\infty} d\tau_2 \, c(\tau_2, t_2) h(t_2 - \tau_2) \\ &= \int_{-\infty}^{\infty} d\tau_1 \int_{-\infty}^{\infty} d\tau_2 \, \phi(\tau_1, t_1 - t_2) \delta(\tau_1 - \tau_2) h^*(t_1 - \tau_1) h(t_2 - \tau_2) \\ &= \int_{-\infty}^{\infty} d\tau_1 \, \phi(\tau_1, t_1 - t_2) h^*(t_1 - \tau_1) h(t_2 - \tau_1) \end{aligned}$$

$h'(t)$ possesses a Karuhnen-Loève expansion,

$$h'(t) = \sum_k h_i \zeta_i(t) \tag{41}$$

where the $\zeta_i(t)$'s are orthonormal over the observation interval and the h_i's are independent Gaussian random variables with variance λ_i's, the eigenvalues of the integral equation

$$\int dt_2 \, R_{h'}(t_1, t_2)\zeta_i(t_2) = \lambda_i \zeta_i(t_1) \tag{42}$$

and both the domain of integration and the region of equality are the observation interval.

We thus have

$$\begin{aligned} \mathcal{E}' &= \int dt \, |\sum_i h_i \zeta_i(t)|^2 \\ &= \sum_i |h_i|^2 \end{aligned}$$

Thus $\mathcal{E}'$ is a sum of squares of independent Gaussian variables. They actually occur in pairs with the same variance because they arise from a bandpass process. The ith pair has an exponential (i.e. chi-square with degree 2) distribution with mean λ_i. The simple fading situation in the previous section corresponds to the degenerate case where there is only a pair of eigenfunctions with a non-zero eigenvalue.

Having $\mathcal{E}'$ as a sum of independent exponential random variables, it is easy to write its characteristic function and to invert it to yield the probability density. If all the λ_i's are distinct its probability density is given by

$$p_{\mathcal{E}'}(x) = \sum_k \frac{\alpha_k}{\lambda_k} e^{-\frac{x}{\lambda_k}} \tag{43}$$

where

$$\alpha_k = \prod_{i,\, i \neq k} \frac{\lambda_k}{\lambda_k - \lambda_i} \tag{44}$$

and its expected value is

$$\mathcal{E} = E\mathcal{E}' = \sum_k \lambda_k \tag{45}$$

The conditional error probability can now be obtained by averaging P_e given in (41) over $\mathcal{E}'$, yielding

$$P_e = \frac{1}{2} \sum_k \alpha_k \left(1 - \sqrt{\frac{\lambda_k}{N_0 + \lambda_k}} \right) \tag{46}$$

In the case where the λ_k are either 0 or equal to a constant λ, $\mathcal{E}'$ has a chi-square distribution with $2L$ degrees of freedom, where L is the number of non zero λ_k. The probability of error is expressed as

$$P_e = \left(\frac{1-\mu}{2} \right)^L \sum_{k=0}^{L-1} \binom{L-1+k}{k} \left(\frac{1+\mu}{2} \right)^k \tag{47}$$

where $\mu = \sqrt{\lambda/(N_0+\lambda)}$. Also $\mathcal{E} = L\lambda$, which permits to write the error probability as a function of $\mathcal{E}/N_0$ and L.

In the case of high signal to noise ratio per dimension,

$$\mu \approx 1 - \frac{N_0}{2\lambda} \tag{48}$$

and

$$\begin{aligned}\left(\frac{1+\mu}{2}\right)^L &\approx \left(1-\frac{N_0}{4\lambda}\right)^L \\ &\approx e^{-\frac{L^2 N_0}{4\mathcal{E}}}\end{aligned}$$

For large enough $\mathcal{E}$ this can be approximated by 1. The sum in (47) then reduces to $\begin{pmatrix} 2L-1 \\ L \end{pmatrix}$ so that

$$P_e \approx \left(\frac{LN_0}{4\mathcal{E}}\right)^L \begin{pmatrix} 2L-1 \\ L \end{pmatrix} \tag{49}$$

It decreases as the signal to noise ratio raised to the Lth power. This is displayed in figure 14 where one sees that diversity can vastly reduce the negative effect of fading.

The derivation was quite general. It is straightforward to apply it when explicit diversity measures are taken, such as interleaving over times greater than T_c, or frequency hoping over frequencies greater than F_c. One can assume that all replicas have the same signal to noise ratios. The formula with distinct λ_k can be used for wideband signals that can resolve many paths. One can admit that each path corresponds to an eigenfunction. However it is not always easy to predict how many paths will be present, nor what their average attenuation is.

6 The Qualcomm CDMA system

In competition with the IS-54 system proposed to upgrade the US cellular system (AMPS), Qualcomm Inc. has developed a spread spectrum Code Division Multiple Access (CDMA) system. It promised much larger capacity increases (possibly 10 or 20) over AMPS than the IS-54 proposal (3, then 6). Spread spectrum and CDMA are well understood techniques that had their origin in military systems and have been applied in civilian satellite communication systems. As we will see, their application requires that the powers from various users be almost the same at the base station. This is hard to achieve in a mobile environment and it was of the major challenges solved by Qualcomm.

The system is compatible with AMPS in the sense that it uses the AMPS frequency bands for the up-link and the down-link, but a single CDMA channel is 1.23 MHz wide, which corresponds to 41 x 30kHz AMPS channels.

Both the up-link and the down-link rely on the 64 Walsh functions of order 64. Recall that the Walsh functions are continuous time functions that mimic the rows of a Hadamard

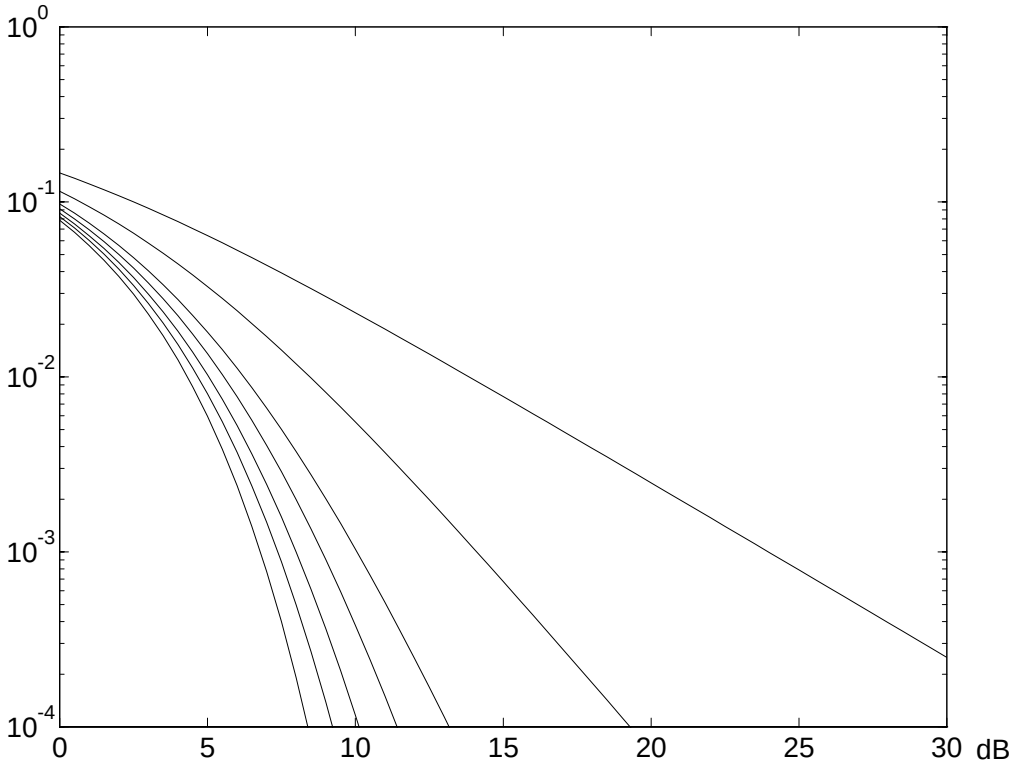


Figure 14: Probability of error for antipodal signals with fading and diversity $L = 1$, 2, 4, 6, 10, 20, and no fading, versus the average signal to noise ratio

matrix. Consequently they are orthogonal functions taking values of +1 and -1. The 8 Walsh functions (or the 8×8 Hadamard matrix) of order 8 are depicted below, the pattern extends to 64.

$$\mathcal{H}_8 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

The 8 Walsh functions of order 8.

Each symbol of the Walsh function lasts for $1/1.2288$ μs, so that each function lasts $1/19200$ s. The use of those functions is very different depending on the direction, we start with the direction to the mobile.

6.1 To the mobile station

In the direction to the mobile there are 64 channels, each modulating one of the Walsh functions. The outputs from the 64 channels are added and split into 2 copies. Each copy is multiplied by a pseudorandom (or PseudoNoise, PN) sequence at 1.2288 megachips/sec, repeating after 32768 chips, i.e. 37.5 times per second. A "chip" is the basic interval of a PN sequence. Those pseudorandom sequences are common to all base stations, but each base station uses a different offset, multiple of 64 chips (i.e. about 52 μs), with respect to absolute time. Thus there can be 512 offsets. In effect a base station is identified by its time shift. After multiplication, each copy modulates a branch of a quadrature modulator at the carrier frequency, which again is common to all base stations.

The pilot channel uses the first function, which is constant. Its signal level is constant and 4 to 6 dB higher than the other channels. A mobile station, which knows the common carrier frequency and PN sequences, can attempt to detect the pilot channel by a correlation operation. Each multipath component separated by more than about 1 μs will give rise to a spike at the output of the correlator, and the 3 strongest components (with about the same delay, thus from the same base station) are tracked and combined coherently. At this point a mobile station has effectively measured the channel from the "best" base station, and it is ready to synchronize. It also has an estimate of the loss between the base station and itself, and can adjust its transmit power accordingly. This is only a rough compensation because the up-link and down-link channel use different bands, thus suffer different attenuations. This measurement is useful to quickly reduce power, e.g. when coming out of a tunnel.

The synchronization channel uses the 32th Walsh function (4th in the table above, with alternating chips). Its information rate is 1200 b/s. Data is coded by a convolutional code to 2400 b/s, repeated twice (to 4800 b/s), interleaved over 1/37.5 s, and each resulting bit modulates 4 successive copies of the Walsh function. The synchronization messages contain time of day and other useful information, such as the identification of paging channels. Now the mobile is ready to process traffic. It listens for system information on an assigned paging channel, and switches to a data channel when an incoming call is signaled.

There can be up to 7 paging channels. Each transmits data at 2.4 kb/s, 4.8 kb/s or 9.6 kb/s with constant power, using a rate 1/2 convolutional code with constraint length 9. After coding (and possible repetition) the data rate is 19.2 kb/s. Bits are interleaved over 20 ms, and modulate one copy of the corresponding Walsh function. In addition to incoming call requests, a paging channel transmits system information and acknowledgements.

The data (or voice) channels are like the paging channels, with 3 main exceptions. First, the transmission power is variable. Second, some of the data bits are used to send power control signals to the mobile. Third, the data at 19.2 kb/s is scrambled by a PN sequence of length $2^{42} - 1$ ($\approx 4.39810^{12}$ chips or 7 years), with a time offset that is a function of the serial number of the corresponding mobile station. Of course the mobile station must apply the reverse operation. The necessary synchronization is achieved from

the synchronization channel. Speech is encoded by a variable rate vocoder at rates of 1.2 kb/s, 2.4 kb/s, 4.8 kb/s, or 9.6 kb/s depending on speaker activity, but the encoder output is duplicated as necessary to always be 9.6 kb/s. A convolutional code with rate 1/2 and constraint length 9 raises the bit rate to 19.2 kb/s. That signal is interleaved over 20 ms and transmitted. The transmission power is inversely proportional to the degree of duplication.

Thanks to the use of the orthogonal Walsh functions, there is no interference between the channels originating from the same base station, but there is interference from the other base stations. Their level depends on the fading characteristics both on the desired propagation path (because of power control), and on the interference path.

The mobile station measures the signal to interference ratio and request power level changes as required. The changes in base station power are in steps of .5 dB over a range of only 6dB.

An interesting peculiarity of the Qualcomm system is the *soft hand-off* feature. A mobile keeps tracking emissions from the base stations (identified by their time delay). When a hand-off occurs, the new base station starts transmitting together with the old one. The Rake receiver in the mobile station effectively combines both transmissions as if they were due to a multipath phenomenon.

A similar trick can be used to gain diversity in microcells environments. There the differential delay is too small for the Rake receiver to resolve the paths and to benefit from diversity. Diversity can be introduced "artificially" by installing a few antennas, each sending the same signal with an artificial delay of a few microseconds. All these emissions are "naturally" combined by the Rake receiver.

6.2 To the base station

In the direction to the mobile the same PN sequences are used. Speech is again encoded by a convolutional code. This time the code has a rate of 1/3 but still a constraint length of 9. Interleaving is again performed over 20 ms. Bits are grouped by 6, each of the 64 resulting symbols generates an appropriate Walsh function. This is in marked difference with the other direction, it corresponds to using 64-ary orthogonal signals. The Walsh function is combined with the long PN sequence, using the time offset corresponding to the mobile station, and the result is again modulated by the quadrature PN sequences and transmitted. No pilot tone is generated.

The receiver at the base station has a tracking receiver and four receivers that each lock on a significant path of the channel response. 64 correlators, one for each Walsh function, are associated with each receiver. Corresponding outputs are combined and the correlator with the largest output is identified, resulting in the recovery of a 6 bit symbol. Deinterleaving and decoding follow. All users from a given cell, and from neighboring cells, interfere at a base station. The receiver separates them by correlation with the correct spread spectrum patterns.

We have mentioned before that the mobile station changes its power level with the

intensity of the received pilot tone. This is not precise enough, and another power control loop has been developed. The base station measures the signal to interference ratio of each mobile station and sends one control bit every 1.25 ms (800 b/s). This control bit adjusts the power by .5 dB. The dynamic range is 85 dB.

6.3 Discussion

The Qualcomm system uses a bandwidth of 1.3 MHz (41 AMPS channels) for 62 channels, or about 20 kHz/channel. Its efficiency comes from maximal reuse, $K = 1$. This is made possible by the spread spectrum system, which spreads the 19.2 kb/s generated by a user over the full bandwidth, effectively reducing the interference level between users by a factor of about 60. If all 62 users were on all the time at the same level, the signal to interference ratio would be about 0 dB and the system would not work. Worse, if some users have a power that is too high by 1.5 dB, and others are too low by 1.5 dB, the signal to interference ratio would be -3 dB.

Thus the key question is how many users can be on simultaneously. Qualcomm relies on two factors: firstly, in voice conversations a speaker is only active 40% of the time. Nothing is transmitted during silences, and effectively the interference level is reduced to 40%. Secondly, each cell is divided into three sectors. As each sector covers only about a third of the users, interference is again reduced.

One might object that the same reductions could be achieved in TDMA or FDMA systems by having dynamic channel assignments on a demand basis. This is true, and experimental systems have proved it. The Qualcomm system has one statistical advantage: it is sensitive to the sum of many interferers. TDMA and FDMA systems are sensitive to having even a single interferer in the neighborhood. Ultimately the debate between the various approaches will be settled by field trials and in the marketplace.

7 Bibliography

Wireless mobile communication systems have been, and still are, the object of numerous studies. Rather than attempting to give a partial bibliography, we refer interested readers to some specialized books.

Proakis [5] gives a classical account of communication theory, and of its application to radio systems, while W. C. Y. Lee [2] is more focused on mobile systems. Parsons' recent book [4] treats the propagation aspect in depth. Calhoun [1] reviews the development of analog systems, and introduces the digital ones. He gives an interesting point of view on the history of the field, including the regulatory facets in the United States. Steele [6] has edited a recent volume that covers many aspects of propagation, modulation, demodulation, coding and speech processing, both at a theoretical level and as applied to existing and proposed systems. Mouly and Paulet [3] give a detailed account of the GSM system. It treats both the radio part and also the protocols that allow the interworking of a large system comprised of mobile stations and of an important fixed infrastructure.

References

[1] George Calhoun. *Digital Cellular Radio.* Artech House, Norwood, MA, 1988.

[2] William C. Y. Lee. *Mobile Communications Engineering.* McGraw-Hill, New York, 1982.

[3] Michel Mouly and Marie-Bernadette Pautet. *The GSM System for Mobile Communications.* Published by the authors, Paris, 1992.

[4] D. Parsons. *The Mobile Propagation Channel.* Pentech Press, London, 1992.

[5] J. G. Proakis. *Digital Communications.* McGraw-Hill, New York, 1989.

[6] Raymond Steele, editor. *Mobile Radio Communications.* Pentech Press, London, 1992.