

CHAPTER 2

CODING, MODULATION, AND MULTIPLE-ACCESS

Wayne Stark

OVERVIEW

The goal in wireless communication systems is to provide mobile, fully integrated, low cost, reliable, seamless systems with the capability of providing voice, high speed data, and video with minimum latency and long battery life. However, there are many obstacles and bottlenecks to achieving this in practice. These bottlenecks include the following:

- limited spectrum
- limited energy
- multipath fading and propagation loss
- cost
- standardization processes

This chapter of the report focuses on coding, modulation, and multiple-access techniques for wireless communications. To put the discussion in perspective, a short background on the development of wireless communication systems, focusing mainly on cellular systems is provided. The following sections discuss coding and modulation and various modulation techniques; present a comparative analysis of research and development in Europe, Japan, and the United States; and, finally, make recommendations for further research in wireless communications.

BACKGROUND

Cellular Systems

Wireless communications became a commercial success in the early 1980s with the introduction of cellular communication systems in the United States, Japan, and Europe. All of these systems were based on frequency division multiple-access whereby a user during a call was assigned a given frequency for transmission to a base station (uplink) and a given frequency for reception from the base station (downlink). The modulation technique adopted was frequency modulation (FM) for the voice signal. The bandwidth assigned to each user was around 30 kHz (depending on the country). Frequency division multiple access (FDMA) and FM modulation were well known techniques/technology available to system designers at the time. It is interesting to note that a university professor in the early 1980s actually proposed code division multiple-access (CDMA) as a technique for cellular communications, but AT&T/Bell Labs (pre-divestiture) discredited it. At the time digital technology was not ready for a modulation technique requiring

sophisticated processing algorithms. Nevertheless, the basic research component of this effort affected future wireless communication systems.

The designers of cellular systems faced many challenges that were unique to mobile environments. These challenges included time-varying multipath fading and interference from other users. The problem of multipath fading was handled by simply increasing the amount of transmitted power. Most of the cell phones were mounted in cars, handheld units were not available, and battery power was not the initial driving force. Interference was minimized by not reusing a given frequency in an adjacent cell. Finally, the application was purely voice communications.

In Europe, many different first generation systems operated in different frequency bands, and thus roaming between countries was not possible with a single cell phone. This lack of interoperability in Europe was one of the key drivers towards developing a second generation system in Europe. In the United States, the increased use of cell phones in the mid-1980s led to a shortage of capacity in major markets (Los Angeles, New York, Chicago). In order to achieve higher capacity (users/cell/Hz) second generation systems began development in the mid-late 1980s. Second generation systems (in the United States, Europe, and Japan) are all based on digital transmission techniques whereby the voice signal is encoded into a sequence of bits. The voice-coded data is then encoded for error correction and modulated using digital modulation techniques. Going to a digital modulation technique has several advantages, including the spectral efficiency of digital modulation, the capability of combining speech and data services, and the improved security of digital techniques. The first proposal for second generation systems in the United States was based on time-division multiple-access (TDMA) whereby each 30 kHz of bandwidth was slotted in time so that three users could use the same spectrum. This was possible because of the compression algorithms applied to speech waveforms. In these systems error correction was introduced to mitigate the effect of multipath fading. Interference is not a predominant issue in this design as different users use either different frequency or time slots or nonadjacent cells. Like first generation systems, essentially the only application for second generation systems (both cellular and personal communication systems (PCS) that use the 1900 MHz frequency band) is voice, although some low speed data communication capabilities are also possible.

It is interesting to note that the modulation technique chosen in Europe (Gaussian minimum shift-keying) for second generation systems is a constant envelope technique while the modulation technique chosen in the United States and Japan is $\pi/4$ differential phase shift keying (DPSK) with raised cosine filtering, which is a nonconstant envelope technique. Constant envelope techniques ensure that the envelope of the transmitted signal is a constant. This fact allows the power amplifier used by the mobile system to operate near saturation without distorting the signal. The most energy efficient operation of an amplifier is near saturation as this is when the power added efficiency is largest. The disadvantage of constant envelope modulation techniques is that their bandwidth efficiency is small relative to modulation techniques that have fluctuating envelopes. On the other hand nonconstant envelope modulation techniques need very linear amplification in order to preserve the signal shape (no distortion). Nonlinearities applied to a nonconstant envelope technique create both adjacent channel power and in-band intermodulation distortion. The adjacent channel power essentially widens the bandwidth occupancy of the signal while the in-band intermodulation distortion acts like additional noise in the system. No distortion is incurred when linear amplification is used with nonconstant envelope modulation techniques. However, linear amplification is possible mainly by operating the amplifier with a small input signal (large backoff) where the energy efficiency of the amplifier is smallest. This characteristic of nonlinear amplifiers makes large power efficiency and bandwidth efficiency hard to achieve simultaneously.

In the late 1980s Qualcomm made a proposal to use direct-sequence spread-spectrum multiple-access (also called code division multiple-access) for cellular systems. Qualcomm's initial claims of significant (more than 10 times) increase of capacity of cellular systems captured the attention of service providers. This effort developed into a later second generation standard known as IS-95. The increase in capacity, it was claimed, was due to exploitation of the voice activity factor, Rake reception, which allowed exploitation of the multipath fading and frequency reuse of 1 (every cell uses all frequencies), which allowed more efficient use

of the frequency spectrum over a geographical area. In addition, multiple users spreading their signals over a wide bandwidth with user unique codes allowed for multiple users using the same spectrum at the same time. The IS-95 system uses 1.25 MHz of spectrum. The IS-95 system first became available commercially around 1995.

A significant event for wireless communication systems occurred in the mid-1990s when the Federal Communication Commission (FCC) decided to auction off spectra in the 1.9 GHz band. This opened up a pair 60 MHz frequency bands between the 1850-1910 and 1930-1990 MHz spectra for use by those with winning bids in the auction. This band was called the personal communication system (PCS) band, and systems operating in this band are called PCS systems. The modulation and coding techniques that were chosen for these systems were virtually all-cellular type systems except they were shifted up in frequency. The operating characteristics (coding, modulation, and multiple-access) were identical. Nevertheless this provided additional capacity for wireless communications and allowed for increased competition.

The overall goal of first and second generation systems is primarily voice communications. The innovations between first and second generations were mainly in going to digital modulation techniques and the use of error control coding. Third generation systems are now being designed and implemented to handle not only voice but high speed (from 384 kbps to 2 Mbps) data, although no one really knows what the market for these services will bear.

Other Wireless Systems

Cellular communication systems are clearly the most widely used wireless communication systems. However, other systems have different characteristics that deserve mentioning. To begin with, many military communication systems need to operate in a very different type of environment. First, in a military communication system, there may not be the possibility of dividing a region into cells and deploying a base station in each cell. Military systems tend to be highly mobile and may possibly operate in unfriendly environments where the cellular type of deployment is not possible. Military systems must also face the possibility of hostile interference (jamming). One of the consequences of not having base stations is that relaying messages is required. In other words, multihop communications over multiple wireless links is necessary. This creates a whole new class of communication problems and constraints.

Another wireless communication system of interest is a wireless local area network (WLAN), whereby a wireless link replaces the wired LAN. Currently such systems tend to operate in the Industrial-Scientific-Medical (ISM) band (902-928 MHz, 2400-2483 MHz, 5725-5780 MHz). This band is available for unlicensed use (in the United States) provided either the power levels are small enough or spread-spectrum modulation techniques are employed with larger transmitted power. The 2.4 GHz band is available worldwide while the 900 MHz band is not available in some parts of the world (e.g., Europe).

Another wireless system gaining significant attention is a cable replacement wireless system called Bluetooth. This is being developed by a consortium of communications and computer companies (e.g., Ericsson, Nokia, IBM, Intel, Motorola). The objective of a Bluetooth system is to replace the cables connecting different components in a computer system with wireless links. Because it is a cable replacement the objective is for a very low cost, low range system. It is also viewed as a means for connecting a mobile computer with a cell phone in a wireless manner such that communication between the mobile computer and a hidden (e.g., in the briefcase) cell phone can be used to connect to the Internet. It can also connect a headset to a cell phone without wires. The range is about 10 meters with 0 dBm (1 milliwatt) transmitted power, but can be increased to 100 meters with larger power. Bluetooth uses the 2.4 GHz ISM band with frequency-hopped spread-spectrum. The maximum data rate is 721 kbps. It handles up to 8 devices in a so-called "pico-net" and can have up to 10 pico-nets operating in a coverage area. The networks created with Bluetooth are ad-hoc networks implying multiple-hops per end-to-end connection.

A competing system is HomeRF. HomeRF is geared more towards higher data rates and higher transmitted power. As with Bluetooth, HomeRF is a frequency-hopping system operating in the 2.4 GHz band.

Products with data rates in the range of 2 Mbps with operating distances on the order of 150 feet are possible.

The development of a variety of communication systems is shown in Fig. 2.1 as a function of data rates and user mobility/cell sizes. This figure illustrates the fact that higher data rates are possible at lower mobilities or decreased cell size. This is due to two considerations. The first (and main) consideration is that at smaller distances the propagation loss is less and thus for a given power level, higher values of E_b/N_0 , the received signal-to-noise ratio, are possible. Another effect is that at high mobilities the channel is harder to estimate and thus proper demodulation/decoding becomes more difficult. This is compensated for, to a certain extent, by the fact that generally error control coding works better (for a fixed block length) when the channel is memory-less (independent fading on different symbols).

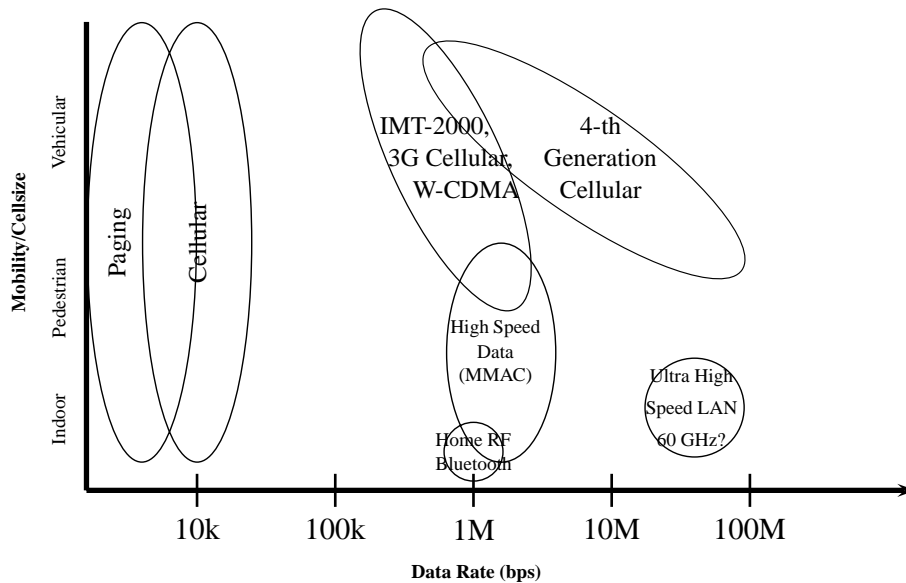


Fig. 2.1. The development of a variety of communication systems is shown as a function of data rates and user mobility/cell sizes.

CODING AND MODULATION

Coding and modulation provide the means of mapping information into waveforms such that the receiver (with an appropriate demodulator and decoder) can recover the information in a reliable manner. The simplest model for a communication system is that of an additive white Gaussian noise (AWGN) system. In this model a user transmits information by sending one of M possible waveforms in a given time, period T , with a given amount of energy. The rate of communication, R , in bits per second is $\log_2(M)/T$. The signal occupies a given bandwidth W Hz. The normalized rate of communications is R/W measured in bits/second/Hz. The received signal is the sum of the transmitted signal and white Gaussian noise (noise occupying all frequencies). The optimum receiver for deciding which of the M signals was transmitted filters the received waveform to remove as much noise as possible while retaining as much signal as possible. For a fixed amount of energy, the more waveforms (the larger M) the harder it is for the receiver to distinguish which waveform was transmitted. There is a fundamental tradeoff between the energy efficiency of a communication system and the bandwidth efficiency. This fundamental tradeoff is shown in Fig. 2.2. In this figure the possible normalized rate of transmission (measured in bits per second per Hz) is shown as a function of the received signal-to-noise ratio E_b/N_0 for arbitrarily reliable communication. Here, E_b is the amount of energy received per information bit while N_0 is the power spectral density of the noise. The curves labeled AWGN place no restrictions on the type of transmitted waveform except that the average energy must be constrained so that the received signal energy per bit is E_b . The curve labeled BPSK restricts

the modulation (but not the coding) to binary phase shift keying. The curve labeled QPSK is for quaternary phase shift keying and the 8-PSK curve is for 8-ary phase shift keying. Clearly at low rates and low E_b/N_0 there is virtually no loss in using QPSK modulation with the best coding compared to the best modulation and coding. Also shown in the figure is what can be achieved with certain coding schemes. While these curves show the best possible transmission rate for a given energy, no restrictions are placed on the amount of delay incurred and on the complexity of implementation. It has been the goal of communication researchers and engineers to achieve performance close to the fundamental limits with small complexity and delay.

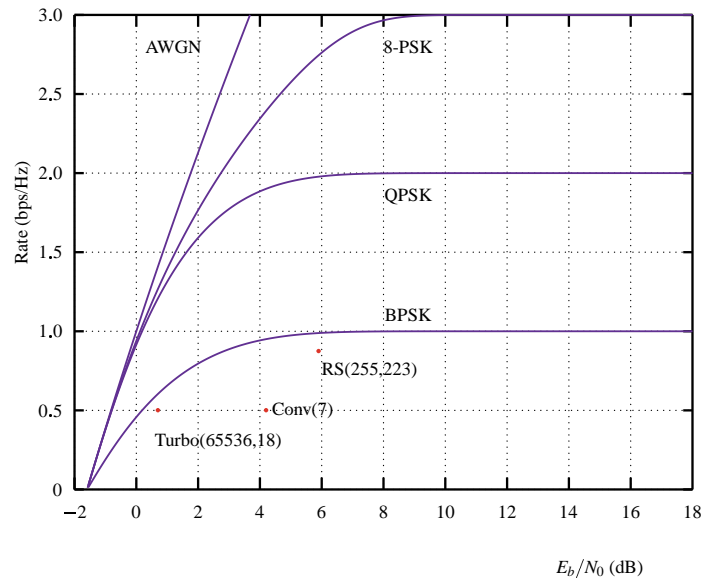


Fig. 2.2. Possible transmission rates versus signal-to-noise ratios for an additive white Gaussian noise channel.

For a wireless communications system, the AWGN model is much too simplistic. In a wireless communication system the transmitted signal typically propagates over several distinct paths before reaching the receiving antenna. Depending on the relative phases of the received signal the multiple signals could interfere in a destructive manner or in a constructive manner. The result of the multiple paths is that the received signal amplitude is sometimes attenuated severely when the signals from different paths cancel destructively, while sometimes the signal amplitude becomes relatively large because of constructive interference. The nature of the interference is, in general, time varying and frequency dependent. This is generally called time and frequency selective fading. A typical time response for a multipath fading channel is shown in Fig. 2.3.

The received signal varies more quickly as the vehicle speed increases. In the original analog cellular systems in order to compensate for the multipath fading, the transmitter increased or decreased the amount of transmitted power. As with the additive white Gaussian noise channel, there are fundamental limits on the rate of transmission for a given average received energy-to-noise ratio ($\overline{E_b/N_0}$). In the simplest model the received signal energy is modeled as a Rayleigh distributed random variable, independent from symbol to symbol. With this assumption the transmissions rates possible, as a function of the average received signal-to-noise ratio, are shown in Fig. 2.4. The gray curves represent the performance possible in an additive white Gaussian noise channel while the dark curves represent the performance with Rayleigh fading. The assumption in this figure is that the channel bandwidth is very narrow and so the result of fading is to only change the amplitude of the signal and not distort the signal in any other way. This is clearly not valid for many communication systems (especially wide bandwidth systems like direct-sequence CDMA).

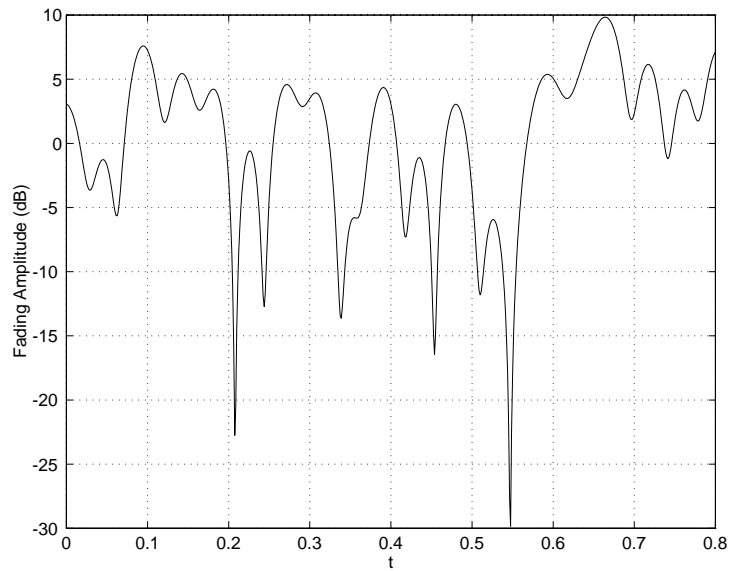


Fig. 2.3. Received signal strength as a function of time for vehicle velocity 10 mph.

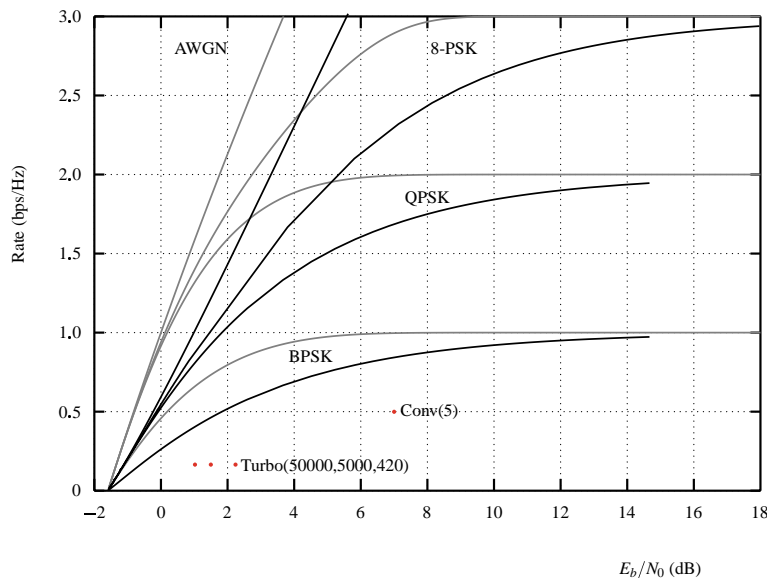


Fig. 2.4. Possible transmission rates versus signal-to-noise ratios for a Rayleigh fading channel.

A key observation from this figure is that there is not a significant loss in performance between what is possible in an additive white Gaussian noise channel and what is possible in a fading channel. For example for transmission rates less than 1/2 bps/Hz the loss in performance due to fading is less than 2 dB with the optimal coding and with BPSK modulation. However, for BPSK alone (without coding), the loss in performance compared to white Gaussian noise channels is on the order of 40 dB when the desired error probability is 10^{-5} . This is a huge loss and is due to the fluctuations of the signal amplitude. Basically the fading process sometimes attenuates the signal so that the conditional error probability is close to 1/2. Sometimes the fading accentuates the signal so that the conditional error probability is virtually zero. The average error probability then is dominated by the probability that the fading level is small. This can be overcome with one or a combination of several techniques. Antenna arrays whereby the received signal at different antennas fades independently is one such technique (discussed in Chapter 6). Another technique is time diversity through coding. In the simplest realization of this, information is transmitted multiple times

spaced far enough apart (in time) so that the fading is independent. At the receiver the signals are combined appropriately. In this manner the probability of error is dominated by the probability that the fading processes attenuate all the transmissions of a single bit. The probability of this event is much smaller than the probability that during a single time instant the fading process will cause significant attenuation. This is essentially a simple form of coding. Another simple form of coding is through frequency diversity. In this case the same information is transmitted over several different frequencies simultaneously. Because the channel is frequency selective not all the frequencies fade simultaneously. In this way diversity is achieved as long as the frequencies used are sufficiently separated (separation larger than the coherence bandwidth of the channel).

The conclusion from the previous discussion is that, for a fading channel, coding and/or diversity techniques are essential in providing performance close to optimal. As before, there are underlying assumptions that the delay is not a major constraint. In the time diversity system, identical data are transmitted but spread out in time. In order to achieve good performance the time separation needs to be sufficiently large so that the fading is nearly independent. In the frequency diversity case, a similar argument is made with the frequencies used. So a large time-bandwidth-space product is needed in order to achieve reasonable performance. In a wireless system, error control, coding and modulation are used to protect the data not only against the effects of fading but interference as well. Interference will be discussed in the multiple access section.

In 1993 a new coding technique (known as turbo codes) was shown to have exceptional performance in an additive white Gaussian noise environment, coming within 0.7 dB of the fundamental limit for a Gaussian channel with a code with block length on the order of 65,000 bits (Berrou et al. 1993). Since that discovery was made, considerable effort has begun on investigating these codes on other channels and with different block lengths. For an ideal Rayleigh fading channel (independent fades for each symbol) turbo codes with block length 50,000 approach within about 1.5 dB of the fundamental limit when the channel is known perfectly. For the white Gaussian noise channel, low density parity check codes are within 0.01 dB of the fundamental limit when the block length is very large. When the block length is shorter (as required by delay constraints in many applications) then the performance of turbo codes deteriorates to the point that traditional convolutional codes perform better. Third generation cellular systems will employ turbo codes for relatively long (e.g., larger than 300 bits) block length messages.

Many different modulation schemes are used in current wireless systems, among these binary phase shift keying (BPSK), Gaussian-filtered minimum shift keying (GMSK), $\pi/4$ DPSK, offset quadrature phase shift keying (OQPSK), and orthogonal frequency division multiplexing (OFDM) (multicarrier). There are a couple key issues when designing a modulation technique. One of these issues is whether the technique uses a constant envelope or a nonconstant envelope. Constant envelope modulation techniques can cope with amplifier nonlinearities but have larger bandwidth than nonconstant envelope modulation techniques. On the other hand, a power amplifier is most energy efficient when operating in the nonlinear region. Nonconstant envelope techniques have smaller bandwidth but need a very linear amplifier to avoid generating both in-band distortion and adjacent channel power. The goal is to have bandwidth efficiency and power efficiency simultaneously. However, with current amplifier designs there is a tradeoff between these two conflicting objectives.

Another key issue when dealing with modulation is intersymbol interference. A wireless channel generally has multipath fading, which causes intersymbol interference if the data symbol duration is the same magnitude or smaller than the delay spread of the channel. As the data rate increases, the amount of (number of symbols affected by) intersymbol interference increases. This generally increases the complexity of the receiver. One method to avoid this is to transmit information on many different carrier frequencies simultaneously. This makes the symbol duration on each carrier much longer (by a factor equal to the number of carriers) and thus decreases the amount of intersymbol interference. However, multicarrier modulation techniques have a particularly high fluctuation of the signal envelope; and thus to avoid

generating unwanted signals (in-band or adjacent channel) an amplifier with high backoff (low input drive level) is required, which means that the energy efficiency will be very small.

Another approach to dealing with multipath fading is to use wide bandwidth modulation techniques, generally referred to as spread-spectrum techniques. Because of the frequency and time selective nature of the wireless channel, a narrowband signal might experience a deep fade if the phases from multiple paths add up in a destructive manner at the receiver. These deep fades generally need extra protection to prevent errors by either increasing the power or adding additional redundancy for error control coding. On the other hand, if the signal has a wide bandwidth (relative to the inverse of the delay spread) then not all the frequencies in a given band will simultaneously be in a deep fade. As such the signal from the part of the spectrum that is not faded can still be recovered. One realization of this idea is that of a direct-sequence system that uses a Rake receiver to “collect” the energy from several paths (at different delays). The probability of all of the paths fading simultaneously becomes much smaller than the probability of one of the paths fading. Because of this the performance is significantly improved compared to a narrow-band system. However, the performance is limited by the bandwidth available.

MULTIPLE-ACCESS TECHNIQUES

Cellular systems divide a geographic region into cells where a mobile unit in each cell communicates with a base station. The goal in the design of cellular systems is to be able to handle as many calls as possible (this is called capacity in cellular terminology) in a given bandwidth with some reliability. There are several different ways to allow access to the channel. These include the following.

- frequency division multiple-access (FDMA)
- time division multiple-access (TDMA)
- time/frequency multiple-access
- random access
- code division multiple-access (CDMA)
 - frequency-hop CDMA
 - direct-sequence CDMA
 - multi-carrier CDMA (FH or DS)

As mentioned earlier, FDMA was the initial multiple-access technique for cellular systems. In this technique a user is assigned a pair of frequencies when placing or receiving a call. One frequency is used for downlink (base station to mobile) and one pair for uplink (mobile to base). This is called frequency division duplexing. That frequency pair is not used in the same cell or adjacent cells during the call. Even though the user may not be talking, the spectrum cannot be reassigned as long as a call is in place. Two second generation cellular systems (IS-54, GSM) use time/frequency multiple-access whereby the available spectrum is divided into frequency slots (e.g., 30 kHz bands) but then each frequency slot is divided into time slots. Each user is then given a pair of frequencies (uplink and downlink) and a time slot during a frame. Different users can use the same frequency in the same cell except that they must transmit at different times. This technique is also being used in third generation wireless systems (e.g., EDGE).

Code division multiple-access techniques allow many users to simultaneously access a given frequency allocation. User separation at the receiver is possible because each user spreads the modulated waveform over a wide bandwidth using unique spreading codes. There are two basic types of CDMA. Direct-sequence CDMA (DS-SS) spreads the signal directly by multiplying the data waveform with a user-unique high bandwidth pseudo-noise binary sequence. The resulting signal is then mixed up to a carrier frequency and transmitted. The receiver mixes down to baseband and then re-multiplies with the binary $\{\pm 1\}$ pseudo-noise sequence. This effectively (assuming perfect synchronization) removes the pseudo-noise signal and what remains (of the desired signal) is just the transmitted data waveform. After removing the

pseudo-noise signal, a filter with bandwidth proportional to the data rate is applied to the signal. Because other users do not use completely orthogonal spreading codes, there is residual multiple-access interference present at the filter output.

This multiple-access interference can present a significant problem if the power level of the desired signal is significantly lower (due to distance) than the power level of the interfering user. This is called the near-far problem. Over the last 15 years there has been considerable theoretical research on solutions to the near-far problem beginning with the derivation of the optimal multiuser receiver and now with many companies (e.g., Fujitsu, NTT DoCoMo, NEC) building suboptimal reduced complexity multiuser receivers. The approach being considered by companies is either successive interference cancellation or parallel interference cancellation. One advantage of these techniques is that they generally do not require spreading codes with period equal to the bit duration. Another advantage is that they do not require significant complexity (compared to a minimum mean square error—MMSE—detector or a decorrelating detector). These interference cancellation detectors can also easily be improved by cascading several stages together.

As a typical example, Fujitsu has a multistage parallel interference canceler with full parallel structure that allows for short processing delay. Accurate channel estimation is possible using pilot and data symbols. Soft decision information is passed between stages, which improves the performance. Fujitsu's system uses 1-2 stages giving fairly low complexity. Fujitsu claims that the number of users per cell increases by about a factor of 2 (100%) compared to conventional receivers and 1.3 times if intercell interference is considered.

COMPARATIVE ANALYSIS

It is useful but difficult to compare the research being done in different countries in the area of wireless communications. First, the panel visited only a handful of companies in each region. Second, these were generally the larger companies with more visibility. Third, proprietary research was not included as part of any discussion. Nevertheless, based on what the members of the panel saw, the WTEC panel has attempted to compare the research activities in various regions, as summarized in Table 2.1 below.

Table 2.1
Comparison of Research Activities

Research Area	U.S.	Europe	Japan
Multiuser Detection Theory	*****	**	***
Multiuser Detection Implementation	***	***	*****
Coding Theory	*****	*****	***
Coding Practice	****	****	****
Multiple-Access	****	****	****

CONCLUSIONS

Throughout this study, and based partially on interactions with selected U.S., European, and Japanese companies, it is recognized that there is substantial need for systems research for future wireless applications. The following research areas are either emerging or evolving and are considered important for future health of wireless communication systems:

- new decoding algorithms for turbo codes for wireless channels
- new coding/modulation techniques for reducing the peak-to-mean envelope ratio, maximizing the data rate and providing large coding gain
- new approaches to jointly designing modulation techniques, and power amplifiers to simultaneously obtain high power added efficiency along with bandwidth efficiency

- new demodulation/decoding techniques to simultaneously combat the near-far problem and do channel decoding in multi-rate DS-CDMA systems
- communication problems unique to high frequency systems (e.g., channel estimation)
- joint channel estimation and decoding/demodulation algorithms
- multiple-access techniques for multi-rate systems with variable quality of service requirements
- space-time coding for systems with multiple antennas
- analog decoding techniques for high speed, low power systems
- ultra wideband systems and hardware design
- research in methodologies for an integrated approach to wireless communications (device layer: e.g., power and low noise amplifiers, mixers, filters; physical layer: coding, modulation; medium access layer: CDMA/FDMA/TDMA; data link layer: hybrid ARQ; network layer: routing protocols)

REFERENCES

Berrou, C., A. Glavieux, and P. Thitimajshima. 1993. "Near shannon limit error-correcting coding and decoding: turbo-codes." *Proceedings of the 1993 International Conference on Communications* 1064–1069.