Contents

Preface

1. Introduction

1.1. Basic Terms
1.2. Social Significance of Telecommunication
1.3. History of Telecommunication
1.4. Structure of the Book

2. Signals

2.1. Classification of Signals
2.1.1. Deterministic Signals
2.1.2. Spectral Decomposition of Signals
2.2 Stochastic Processes
2.2.1. Superficial Characterisation of Stochastic Processes
2.2.2. Stationary and Ergodic Processes
2.2.3. Linear Transforms of Stationary Processes
2.2.4. Physical Meaning of Spectral Density Function

3. Sampling. Discrete Source Coding

3.1. Relation Between Analog Signals and Series of Samples
3.1.1. The Spectrum of Sampled Series
3.1.2. Signal Reconstruction from Equidistant Samples
3.1.3. Spectral Density of Random Series
3.1.4. Quantization Noise. Nonlinear Quantization
3.2. Symbol Series as Information
3.2.1. Coding Density
3.2.2. Coding of Symbol Series

4. Fundamentals of Acoustics

4.1. Physical Aspects of Sound
4.2. Physiological Characteristic of the Human Hearing
4.3. Production of Artificial Sound Field
4.4. Sound Transmission. Quality Requirements
4.5. Electroacoustic Transducers
4.6. Microphones
4.7. Loudspeakers
4.8. Headphones

5. Fundamentals of Photometry and Colorimetry

5.1. Perceptional, Physical and Psychophysical Features
5.2. The Human Eye and Vision
5.3. Thermal Radiation. Photometry
5.4. Colorimetry
5.5. The Picture
5.5.1. Transmission of Monochrome Moving Pictures
5.5.2. Transmission of Colour Moving Pictures

6. Communication Channels

6.1. Analog Channels
6.1.1. Time-Invariant Linear Distortion
6.1.2. Echo and Reverberation
6.1.3. Time-Variant Linear Distortion
6.1.4. Nonlinear Distortion
6.1.5. Additive Noise
6.2. Digital Channels
6.2.1. Memoryless Channels
6.2.2. Principles of Error-Detection
6.2.3. Capacity of Binary Symmetric Channels

7. Error Control Coding

7.1. Basic Terms of Coding
7.2 Linear Codes
7.3. Simple Linear Codes
7.3.1. Repeat Code
7.3.2. Single Parity-Check Code
7.3.3. Two Dimensional Parity-Check Code
7.3.4. Hamming Code

8. Guided Wave Connections

8.1 Introduction
8.2. Specification of the Signal Transmission
8.3. Transmission by TEM Waveguides
8.3.1. Duplex Telephone Connection
8.3.2. Structure of the TEM Waveguides
8.3.3. Transmission Model of TEM Waveguides
8.3.4. Amplification of Audio Signals

8.4. Transmission by Dielectric Waveguides
- 8.4.1. Principle of Dielectric Waveguides
- 8.4.2. Optical Transmitters and Receivers
- 8.4.3. Transmission Model of Dielectric Waveguides
- 8.4.4. Structure of the Dielectric Waveguides

9. Radio Channels

9.1. Electromagnetic Spectrum, Frequency Bands

9.2. The Radio Channel

9.3. Antennas
- 9.3.1. Antenna as a Spatial Filter
- 9.3.2. Directional Characteristics of the Antennas
- 9.3.3. Unidirectional Effect and the Gain
- 9.3.4. Effective Area of the Receiver Antenna

9.4. Wave Propagation Modes
- 9.4.1. Direct Wave and Free Field Attenuation
- 9.4.2. Refraction
- 9.4.3. Ground Back-Scatter
- 9.4.4. Ground Wave Multipath Propagation
- 9.4.5. Diffraction
- 9.4.6. Surface Wave Propagation
- 9.4.7. Tropospherical Scatter
- 9.4.8. Ionospherical Propagation

10. Noise

10.1. Thermal noise

10.2. Noise Characteristics of a Transmission System

10.3. Noise Figure of Multistage Systems

10.4. Effective Noise Temperature of Composite Sources

10.5. Signal to Noise Balance of Radio Communications

10.6. Quantization Noise

11. Analog Modulation Schemes

11.1. Amplitude Modulation (AM)
- 11.1.1. Sinewave Modulated AM Signals
- 11.1.2. AM Signal Demodulation

11.2 Angle Modulation
- 11.2.1. FM Signal Demodulation

12. Digital Modulation Schemes

12.1. Baseband Modulation
    12.1.1. Baseband Shaping of Impulses
    12.1.2. The Error Ratio
12.2. Carrier Modulations
    12.2.1. Structure of the Binary Modulation Systems

13. Channel Allocation

13.1. Frequency and Time Division Multiplexing
13.2. Code Division
13.3. Dynamic Channel Allocation. Polling Methods.
13.4. Random Access Time Division Procedures
    13.4.1. Aloha-Type Random Access Procedures
    13.4.2. Carrier Sense Procedures
    13.4.3. Conflict Resolving Algorithms

14. Public Switched Telephone Network

14.1. Structure of the Telephone Network
14.2 Telephone Sets
14.3. Automatic Telephone Exchanges
14.4. Stored Program Control Exchanges
14.5. Digital Exchanges

15. Traffic Theory

15.1. Terms and Definitions
    15.1.1. Mathematical Model of Telephone Traffic
15.2. Loss Systems
    15.2.1. Full-Availability Group
    15.2.2. Link Systems
15.3. Delay Systems

16. Point-to-Point Transmission

16.1 Quality Requirements
16.2 General Block Diagrams of Transmission Systems
16.3. Microwave Transmission Systems
    16.3.1. General Features
    16.3.2. Digital Radio Transmission

16.3.3. Digital Radio Relays
16.3.4. Digital Satellite Communication
16.3.5. Analog Radio-Relay Systems
16.4. Optical Transmission Systems
16.4.1. Intensity Modulated Optical Transmission
16.4.2. Coherent Optical Transmission

17. Mobile Communications

17.1. Main Features
17.1.1. Regulations of the Mobile Services
17.1.2. Structure and Operation of Terrestrial Stations
17.1.3. Mobile Radio Channels
17.2. Private Systems
17.3. Public Mobile Telecommunications
17.3.1. Specific Features and Devices
17.3.2. Pan-European Mobile Telephone System
17.4. Possible Trends of the Development

18. Integrated Services Data Network

18.1. Concept of the ISDN
18.2. Basic Rate Access
18.2.1. Access Delay
18.3. Primary Rate Access
18.4. Clock Synchronization, PDH and SDH Systems

19. Radio Broadcasting

19.1. AM Broadcasting
19.1.1. Basic Parameters
19.1.2. AM Transmitters
19.1.3. Basic Aspects of the Signal Reception
19.1.4. AM Receivers
19.2. FM Broadcasting
19.2.1. Basic Parameters
19.2.2. FM Transmitters
19.2.3. FM Receivers
19.3. Stereo Broadcasting
19.3.1 FM Stereo Broadcasting
19.4. Digital Broadcasting

20. Television Broadcasting

20.1. Structure of TV Pictures
20.2. Monochrome Video Signal
20.3. Colour TV Pictures
20.4. Proper Reproduction of Colours
20.5. Analog Color-TV Systems
20.5.1. The NTSC System
20.5.2. The SECAM System
20.5.3. The PAL System
20.5.4. The MAC System
20.6. Digital TV Systems

21. Data Broadcasting. Distribution of Programs

21.1. Definition of Data Broadcasting
21.2. Data Broadcasting Networks
21.3. Data Broadcasting on Program Broadcasting Networks
21.3.1. Data Broadcast in AM Program Broadcasting
21.3.2. Data Broadcast in UHF-FM Program Broadcasting
21.3.3. Data Broadcast in TV Broadcasting. Teletext
21.3.4. Single Frequency Networks
21.3.5. Satellite Systems
21.4. Distribution of Programs
21.4.1. Direct Broadcasting Satellites
21.4.2. The AM-micro Broadcasting System
21.5. Cable TV Networks
21.5.1. Structure of the CATV Systems
21.5.2. Signals and System Parameters
21.5.3. Development Trends

22. Radars. Active Microwave Remote Sensing

22.1. Effective Area
22.2. Measurement of Target Parameters
22.3. Modulation Procedures
22.4. Detection
22.5. Accuracy and Resolution
22.6. Microwave Active Remote Sensing
22.6.1. Scattering Meter
22.6.2. Image Processing Radars

23. Radio Astronomy, Microwave Passive Sensing, Radio Navigation

- 23.1. Radio Astronomy
  - 23.1.1. Radio Sources
  - 23.1.2. Radiometer
  - 23.1.3. Radio Telescopes
- 23.2. Microwave Passive Remote Sensing
  - 23.2.1. Principle of the Measurement
  - 23.2.2. Measurement of the Humidity of the Soil
- 23.3. Radio Navigation
  - 23.3.1. Classical Navigation Systems
  - 23.3.2. Global Positioning System

24. Audio and Video Recording

- 24.1. Mechanical Sound Recording
- 24.2. Magnetic Audio Recording
- 24.3. Magnetic Video Recording
- 24.4. Optical Recording
- 24.5. Digital Audio Recording
- 24.6. Home Cameras

Appendix

- Telecommunication Law

Index

1. INTRODUCTION

More and more electrical engineers throughout the world are dealing with telecommunication technology. In this introductory chapter we discuss the basic terms of telecommunication and its significance for the human society. Main events of the telecommunication history will also be reviewed. The following part deals with the role of telecommunication and its location within the studies of electrical engineering. Finally, the structure of the book is presented.

1.1. Basic Terms

The term telecommunication denotes the entirety of technical solutions developed for the transmission of the information between arbitrary two points, to any distance. Distortions and errors caused by the transmission have to be kept low at levels within reasonable expenses. To understand this general definition better, let us complete it by some remarks :

- Communication is a transactional process of sharing meaning. People had communicated with each other even before telecommunication existed. Telecommunication however significantly contributes to cover their needs.
- There are several kinds of information to be transmitted, such as speech, music, text, picture, data etc.
- Essentially, telecommunication utilizes electromagnetic phenomena. One of the fundamental problems of telecommunication is to maintain the information fidelity despite of the unavoidable interfering effects.
- Telecommunication is a service. The user expects a constant service at any circumstances. The reliability of telecommunication is of great importance.
- Interconnection of two users for the time of transmission is the task of switching technique. Transmission of information between two users is the task of the transmission technique. Public exchange network is a typical example for the first task while the second task is solved by the wirebound telecommunication.
- The telecommunication engineer has to consider economical aspects, as well. Taking into consideration the restrictive conditions, he has to find the balance between cost and quality.

Telecommunication services can be classified by several points of view, such as:

- kind of the information to be transmitted,
- number of users, taking part in the communication,
- role (possibilities) of the users, i.e. whether the communication is unidirectional, bidirectional (dialogue) or multidirectional (conference).

If there are more than two participants using the same telecommunication service then the units interconnecting them form a network. In the case of broadcasting, the information is unidirectional transmitted from one source to several sinks. In the case of data collection, the information is sent from several sources to one common sink.

Communication link between the users is said to be switched if it is set up just for the time interval required by the users. In such a case the network contains switching units (telephone exchange) as well. To make the telecommunication network more efficient, the transmission paths and the switching networks are allocated for multiple use. The channel allocation techniques challenge further fundamental problems in telecommunication technology. To give an economically reasonable solution for satisfying the users' demands, problems of collision and queuing have to be examined.

The above classification of telecommunication services is listed in detail in Table 1.1. There is one more important factor: whether the communication is taking place between fixed or between mobile points. Recently, the mobile telecommunication is coming into the limelight.

Table 1.1. Overview of Telecommunication services

Kind of Information	Direction of Communication[1]				Type of Network[2]				Service
	UD	BD	MD	BC	DA	PP	SW	DS	
Speech		+					+	+	Telephone
			+				+	+	Remote conference
	+						+	+	Exact time
	+			+					Radio
Music	+			+					Radio
	+			+					Cable radio
Text	+					+	(+)	(+)	Telegram
		+	(+)				+	+	Telex, E-mail
Stationary picture		+					+	+	Fax
Moving picture	+			+					TV
Data	+				+				Remote measurement
	+				+				Remote supervision
	+					+			Remote control
			+				+	+	Computer Network

Note 1: UD: unidirectional link, BD: bidirectional link, MD: multidirectional link.
Note 2: BC: broadcasting, DA: data acquisition; PP: point to point communication, SW: switched link, DS: distributed network.

1.2. Social Significance of Telecommunication

The participants of the telecommunication services can be classified as follows:

- users (participants, subscribers, consumers), employing the services but knowing nothing about the technical details,
- servicing staff, satisfying the user demands by planning, building and operating the network,
- manufacturers, developing, manufacturing and selling the equipment needed for the service,
- authority, regulating the technical and economical conditions of the telecommunication service.

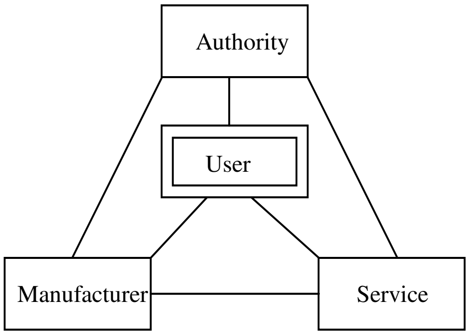


Figure 1.1. Participants of the Telecommunications

As it is shown in Fig. 1.1., it is the user who is in the focus of the telecommunication services. In Table 1.2., domestic telephone and television service is characterized by the number of telephone trunk lines and TV subscriptions, per 100 inhabitants. Although these parameters do not completely show the quality of the service but they are important data which characterize the state of the art and the trend of these services. The number of TV subscribers out of 100 inhabitants can be qualified as good, however the number of trunk lines per 100 inhabitants is enormously low.

Table 1.2. Telephone and TV Services in Hungary

	1970	1980	1990	1992
Number of telephone trunk lines (per 100 inhabitants)	3.86	5.76	8.76	11.3
Number of TV subscribers (per 100 inhabitants)	17.1	25.8	27.9	

The consequence of this deficiency is illustrated in Fig. 1.2. where the relation between the GDP and the number of trunk lines is shown. It turns out from the figure that there is a strong correlation between the GDP and the number of trunk lines normalized to the unity of population.

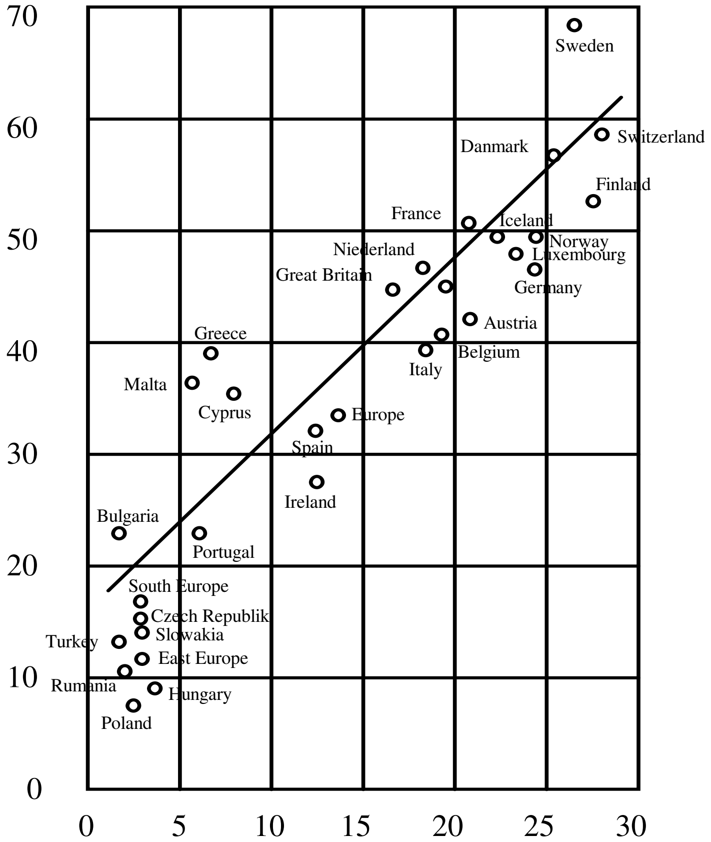


Figure 1.2. Relation Between the GDP and the Number of Telephone Lines

Since there are several other characteristics related to the GDP, it reflects the general state of development in the country. As it also turns out of the figure, the data of the East European countries are significantly under the regressive line.

It is interesting to estimate how much investment would be needed to increase the domestic trunk line density to 40 percent. In 1992, there were 1.2 million trunk lines in Hungary. Since there are 10.6 million inhabitants in Hungary, 4.24 million trunk lines would be needed on the whole for achieving the 40 percent density, i.e. 3.04 million new trunk lines should be installed. Taking $1.000 as the unity price of one trunk line, the investment would require about 3 billion dollars.

Further mapping of the social significance of telecommunication is left to the Reader. As a directive, let us point to the military demands, the applications in traffic, in catastrophe prevention, in technological telecommunication (e.g. electrical energy system), etc.

1.3. History of Telecommunication

The start of information transmission by means of an electrical device can be dated back to 1837, when the telegraph was invented by Morse. Morse encoded the letters of the alphabet in simple codes and can thus be considered the forerunner of the information and coding theory. Bell's invention of speech transmission (1876) initiated the development of the telephone. The start and progress in radio transmission can be connected with Marconi's activity (about 1900).

The progress was crucially influenced by the invention of electrical signal amplification, first by vacuum tube (1907) and later by semiconductors (1948). Semiconductor technology was significantly extended by the development of microelectronics and photonics. Since the first computer appeared, the computer and the telecommunication techniques support each other's development. As an important consequence of this support communication software has appeared, and its influence is still growing. Recently speech, picture and data transmissions are undergoing a revolutionary change into one integrated service.

The following three tables are characteristic for the development of telecommunication. Table 1.3., published by the Newsweek in October 1987, shows five milestones of the progress. Table 1.4. resumes the development of the theoretical background. The dates given here indicates the first release of fundamental papers or patents. Results presented there were further developed by several authors thus creating series of significant disciplines. Famous Hungarians, having contributed to this progress are listed in Table 1.5.

Table 1.3. Ten Milestones of the Progress in Telecommunication

Telephone (1876)	A. G. Bell
Radio Waves (1887-1907)	H. Hertz, A. Popov, G. Marconi
Television (1936)	British Broadcasting Corp.
Radio Telephone (1946)	Cellular System (1981)
Computer (1946)	Electronic Numeric Integrator and Computer (ENIAC), University of Pennsylvania
Satellite Transceiver (1962)	Telstar, Bell Laboratories
Microprocessor (1971)	INTEL Corporation
Digital Exchange (1976)	No. 4. ESS, Bell Laboratories
Optical Cable (1977)	Corning Glass Works
Local Area Network (1979)	Ethernet, Xerox-Intel-DEC

Table 1.4. Main Theoretical Fundamentals of the Progress

Network Theory	Ohm 1827, Kirchoff 1847 Heaviside 1900, Bode 1945
Electromagnetic Field Theory	Maxwell 1873
Traffic Theory	Erlang 1917
Signal Transmission and Modulation	Nyquist, Hartley 1920-28 Armstrong (FM) 1936 Reeves (PCM) 1937
Network Synthesis	Foster 1924, Cauer 1926-44 Brune 1931, Darlington 1939
Statistical Communication Theory	Rice, Wiener, Kotelnikov 1944-47
Information Theory and Coding	Shannon, Hamming 1948-50
Signal Processing	Cooley, Tukey 1965

Table 1.5. Outstanding Hungarian Contributions

Puskás Tivadar (1844-1893)	Telephone exchange, telephone courier, 1893
Pollák Antal (1865-1943)	High-speed telegraph, 1898
Virág József (1870-1901)	High-speed telegraph, 1898
Békésy György (1899-1972)	Research of the hearing mechanism, Nobel Price in 1961
Neumann János (1903-1957)	Principles of electronic computers
Bay Zoltán (1900-1992)	Reflection of radar signals from the Moon (1946)
Gábor Dénes (1900-1979)	Invention of holography, Nobel price in 1971
Kozma László (1902-1983)	Design of telephone exchange, Construction of computer, Kossuth price in 1948
Rényi Alfréd (1921-1970)	Information theory, Kossuth price in 1949 and 1954

1.4. Structure of the Book

This book takes aim of college students taking part in courses of electrical engineering education. The structure and content of the book presume preliminary training and preliminary studies with respect to the parallel subjects of the similar aim and prepare the students for the module choice where they will continue their studies in branches focused on special topics. The subject Telecommunication Systems is delivered four hours weekly during a period of 14-15 weeks. This condition as well as the price of the book limit the structure of the book.

Students of the subject have already a fair knowledge of computer science (digital circuits, programming and information theory) so that the current subject does not include these topics in spite of their importance (convergence of computer and telecommunication technologies, decisive role of communication software, computer networks). The same applies to electronics and microelectronics.

Albeit telecommunication systems are based on microelectronics, circuit implementations are discussed in other subjects. The book builds upon the knowledge of probability theory and there are several joining points to the preceding subject - Networks and Systems.

The book is divided into 24 chapters discussing the fundamental methods and services of telecommunication systems.

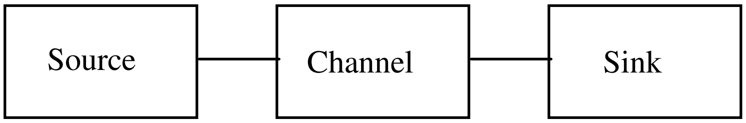


Figure 1.3 The Simplest Model of the Information Transmission

The fundamental task of telecommunication is the transmission of information from the source to the sink so that its model consists of the source, the sink and the transmission channel as it is shown in Fig. 1.3. The channel properties are determined by the transmission medium and by the characteristics of the circuits interfacing the source and the sink to the transmission channel (see Fig. 1.4.).

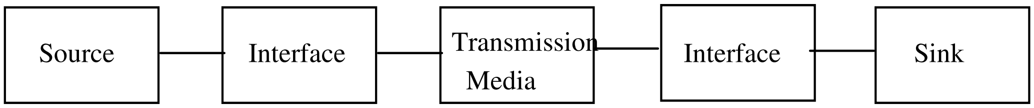


Figure 1.4. Model of the Information Transmission

Properties and mathematical description of the information sources and sinks are given in chapters 2., 3., 4. and 5. Chapters 2. and 3. present the general mathematical apparatus for signal description, Chapter 4. gives insight to the properties of sound and Chapter 5. to that of the picture. General properties of the telecommunication channel are described in chapter 6. Chapters 7. and 8. present the wire-bound and wireless transmission media. Chapter 9. is devoted to noise effects. Source - channel interfacing techniques discussed in Chapters 10., 11. and 12. consider analog and digital modulation procedures as well as the signal encoding and decoding techniques.

Chapter 13. opens the second part of the book which deals with telecommunication networks and services. As one of the basic procedures, channel multiplexing is treated in Chapter 13. One of the most widely-spread services is the public telephone network described in Chapter 14. This topic leads to the mass servicing theory and to the traffic theory which is described in Chapter 15. As an important implementation of these theories, mobile telecommunication is described in Chapter 17.

Chapter 16 discusses the terrestrial and satellite microwave transmission systems. Integrated service telecommunication networks are presented in Chapter 18. Chapters 19., 20. and 21. deal with the questions of audio, video and data broadcasting. Chapters 22. and 23. dealing with navigation, radio location and remote sensing are also of special importance. Chapter 24. is a review of telecommunication services in consumer electronics. The book includes an Appendix listing the laws about broadcasting and frequency regulations to illustrate the relations between the telecommunication and the society.

The chapter sequence is ruled by the logical order given above. The chapters, however, are self-explanatory sections which can also be studied independently.

References

[1] P.G. Fontolliet: Telecommunication Systems. Artech House, Dedham. MA, 1986.

[2] A.B. Carlson: Communication Systems. McGraw-Hill, New York, 1986.

[3] A Nobel-díjasok kislexikona. Gondolat Kiadó, Budapest, 1974.

[4] Magyar Életrajzi Lexikon. Akadémiai Kiadó, Budapest, 1982.

[5] Magyar Statisztikai Zsebkönyv, 1989. Statisztikai Kiadó, Budapest, 1990.

[6] Magyarok a természettudomány és technika történetében. OMIKK, Budapest, 1992.

2. SIGNALS

Introduction

In telecommunication systems, signals are time- and/or place-dependent physical quantities (or their mathematical representations) which have some kind of meaningful content. In this sense, the time-dependent output of an electroacoustical transducer (e.g. a microphone), or the time- and space-dependent sound pressure in a certain point of the field, or even the blackness of a photograph as the function of plane co-ordinates, can be regarded as signal.

Functions are obvious mathematical models of the signals. In the simplest but typical case these functions are scalar (often complex) depending on one variable only (which is usually time). Methods discussed in this chapter refer strictly to these functions although it has to be mentioned that the same methods are used when more general signals (vectorial, multivariable) are described.

2.1. Classification of Signals

Signals frequently used in practice can be classified according to the "richness" of their domain and range. Back to the classical example, the output signal of a microphone is continuous both in its domain and in its range. Such a signal is called an *analog*. Another group of signals exists with instantaneous values making up a finite set of numbers. In this case we speak about discrete range or *discrete amplitude* signals. Another case is when the instantaneous values of the signal are important only in discrete points of time (usually at $t = kT+ t_o$, $k = 0, \pm 1,...$). This signal is said to be *discrete in time*. In fact, this is not a real signal but a series of numbers. In today's telecommunication, signals discrete both in time and amplitude are of huge importance. These signals are called *digital* and their significance lies in the fact that only digital signals can be evaluated by computers.

Another aspect of signal classification may be the purpose the signal is analyzed for. It is a typical task to compare and qualify similarity of two signals, for instance the time-dependent sound pressure and the electrical output signal of a microphone. In such a case, it is obviously impossible to characterize the quality of the microphone by comparing just a single continuous sound, say the vowel "a", it is necessary to examine several different functions (or sets of functions). On top of that, occurrence and significance of the examined functions are not necessarily the same so that signal analysis as the analysis of sets of functions is related also to the terms used in probability calculus. Because theory and terms of stochastic processes give proper frames to such an analysis, the class of *stochastic signals* has been introduced.

On the other hand, if -by experience- the behaviour of a system can be well judged by the response to a single previously defined function, the analysis is said to be done with a *deterministic* signal. The deterministic signal is a very "pleasant" analyzing tool if it can be defined by a simple equation; however the sense of such conclusions is, however, more limited (E.g., perfect transmission of the vowel "a" does not conclude the same for the vowel "i"). In short, deterministic signals are concrete functions while stochastic signal can be interpreted as set of functions which have some similar characteristic as well.

2.1.1. Deterministic Signals

To classify and to characterize deterministic signals used in telecommunication practice, some useful categories and terms have to be introduced. Some of these are reviewed in the following.

1.) *Finite time (finite hold)* signal: the $x(t)$, $t\in(-\infty,\infty)$ signal is said to be finite time if a pair of $t_1 > -\infty$, and $t_2 < +\infty$, exists, so that $x(t) = 0$ for $\forall t < t_1$ and $\forall t > t_2$.

2.) *Absolutely integrable* signal: the $x(t)$, $t\in(-\infty,\infty)$ signal is said to be absolutely integrable if $\int_{-\infty}^{+\infty} |x(t)|\,dt < +\infty$.

3.) *Energy* signal: the $x(t)$, $t\in(-\infty,\infty)$ signal is said to have finite energy if

$$E_X = \int_{-\infty}^{+\infty} |x(t)|^2\,dt < +\infty$$

4.) *Limited* signal: the $x(t)$, $t\in(-\infty,\infty)$ signal is said to be limited if a $K<+\infty$, exists, so that $|x(t)| < K$ for $\forall t\in(-\infty,\infty)$.

Note: if there is such time t exists where $|x(t)| = K$ then K is said to be the (absolute) peak value of the signal.

5.) *Finite average* signal: the $x(t)$, $t\in(-\infty,\infty)$ signal is said to have finite average if

$$A_x = \lim_{\substack{t_1\to-\infty \\ t_2\to+\infty}} \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} x(t)\,dt \text{ exists and is finite.}$$

Note: A_X is called as the DC average of the signal.

6.) *Power* signal: the $x(t)$, $t\in(-\infty,\infty)$ signal is said to have a finite average power if

$$P_X = \lim_{\substack{t_1\to-\infty \\ t_2\to+\infty}} \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} |x(t)|^2\,dt \text{ exists and is finite.}$$

7.) *Periodic* signal: the $x(t)$, $t\in(-\infty,\infty)$ signal is said to be periodic in T if $x(t+T) = x(t)$ for $\forall$t. T is said to be the fundamental period if there no such a $T_o < T$ exists, for which $x(t+T_o) = x(t)$ is valid for $\forall$t.

8.) *Harmonic* (sinusoidal) signal: the $x(t)$, $t \in (-\infty, \infty)$ signal is said to be harmonic if for some A,Ω and Φ values $x(t) = A \cos(\Omega t + \Phi)$ for $\forall$t.
A is the amplitude, Ω is the angular frequency and Φ is the phase of the signal.

Note: In technical literature, harmonic signals are widely expressed in complex form: $x(t) = A\, e^{j(\Omega t+\Phi)}$.

9.) *Quasi-periodic* signal: the $x(t) = \sum_i A_i \cos(\omega_i t + \Phi_i)$ is said to be quasi-periodic if the ratios of angular frequencies are irrational numbers.

It can be seen from the list above that essentially two groups of deterministic signals important for the practice exist: One of them (1-3) consists of the impulse-like signals (bursts), and steady signals form the other group (4-9).

2.1.2. Spectral Decomposition of Signals

For the analysis of linear time-invariant systems, it is preferable to handle the input (or output) signal as the sum of harmonic signals since it is the case when the influence of the system on the signal can easily be estimated. Therefore, it is an important question what are the conditions for a signal to be composed of such harmonic components or of any other composite form. This signal composition is denoted as spectral. Let us recall two characteristic examples of spectral composition.

Fourier-Series of a Periodic Signal

If $x(t)$ is a continuous signal periodic in T *it* can be expressed as

$$x(t) = \sum_i X_i \exp(j2\pi i t/T) \qquad (2.1)$$

The series of the above function is uniformly convergent and its coefficients are computed as follows

$$X_i = \frac{1}{T}\int_0^T x(t)\exp(-j2\pi i t/T)\, \mathrm{d}t \qquad (2.2)$$

Fourier Transform of Absolutely Integrable Signals

Let x(t), $t \in (-\infty, \infty)$ be an absolutely integrable function. In this case, $x(t)$ can be expressed as

$$x(t) = \int_{-\infty}^{+\infty} X(f)\, e^{j2\pi f t}\, \mathrm{d}f \qquad (2.3)$$

where

$$X(f) = \int_{-\infty}^{+\infty} x(t)\, e^{-j2\pi f t}\, dt \qquad (2.4)$$

The function *X*(.) is called the Fourier transform of *x*(.). It can be seen that in this case an integral is used for composition instead of a sum. Although this is a great difference from the mathematical point of view, it is more important that from the technical point of view the two compositions seem to be similar.

Spectral decomposition of signals introduces further practical aspects for signal classification and some specific terms related to the frequency domain:

1. *Band limited* signal: *x*(.) is said to be band limited within the $f_1<f_2$ range if it has no components outside the ranges $[f_1,f_2]$ and $[-f_2,-f_1]$.

2. *Narrowband* signal: *x*(.) is said to be narrowband if $(f_2-f_1)/f_2<<1$.

3. *Effective* frequency: that $B > 0$, for which $$B^2 = \frac{\int_{-\infty}^{+\infty} |x'(t)|^2\, dt}{\int_{-\infty}^{+\infty} |x(t)|^2\, dt}$$

(2.5)

Notice that the numerator of the eq. (2.5) is the energy of the derived signal and the denominator is the energy of the signal itself. Obviously, the above equation can be extended to the periodic and quasi-periodic signals as well.

2.2. Stochastic Processes

Let us consider the following examples:

Example 1. Suppose the signal appearing on the secondary coil of a line transformer is given by

$$\xi_t = A \cos(\Omega t+\Phi)$$

where the value of the phase shift $\Phi \in [0,2\pi]$ depends on the moment of switching the power on related to the first negative-to-positive transition of the line voltage time function following the switching. An obvious model can be set up by saying that Φ is a probability variable uniformly distributed over $[0,2\pi]$.

Example 2. One possible way of transmission of integer numbers α_i, (i=0, ± 1,...) is to compose a sum of so called elementary signals *x*(.) limited to duration *T* in the following form

$$\xi_t = \sum_i \alpha_i\, x(t-i\, T)$$

This signal will differ for various $\alpha = \{\alpha_i, i = 0, \pm1,..\}$. The analysis of such a signal on the base of all possible combinations of α may cause difficulties if specific statistical relations between the elements of the α series are not known. It is much more reasonable to make use of such an information and to say (if possible) that the numbers α_i, i=0,±1,... are independent, equally distributed probability variables.

The common feature of the above examples is that the signal (a time-dependent function) appeared as an object depending (also) on probability variables. A given probability variable (example 1) or a given series of probability variable (example 2) results in one concrete function. Such a function is called the *realization* of the random signal (a stochastic process) so that the random signal can be considered as a set of concrete functions. It is a certain deficiency of such a view that it hardly reflects what the characteristic and what the specific elements of the set are.

On the other hand, in both examples the observed quantity (the signal) is a probability variable at any moment so that it can be even considered as a set of probability variables with continuous parameter *t*. It is obvious that for those *t*-s which are "close" to each other, probability variables with "similar" values belong. (In example 1, too, "similar" values of probability variables belong to those parameters which are about one period of time apart from each other).

Anyway, the above way of thinking is useful even in cases when (unlike in our examples) the signal values cannot be expressed as probability variables. As we shall see, the knowledge of the adequate statistical properties of signal values is enough to answer several important practical questions.

2.2.1. Superficial Characterization of Stochastic Processes

The most important parameter of the random signal ξ are the values of its realization at an arbitrary time *t*. As this value, ξ_t, is also a probability variable, its behaviour is characterized by probabilities

$$F_\xi(x,t) = P(\xi_t \leq x) \tag{2.6}$$

where F_ξ is a two variable function called as one dimensional distribution (or amplitude distribution) of the process ξ. The range of the F_ξ is the [0,1] interval and as a function of the first variable it is monotonously increasing and continuous from the right side. If ξ_t is a probability variable with continuous range (which is typical for analog signals) then

$$f_\xi(x,t) = \frac{\partial}{\partial_x} F_\xi(x,t) \tag{2.7}$$

is the so called *one dimensional probability density function* which gives a detailed characterization of the signal behaviour, similar to the distribution function.

Another important question is the probability that an instantaneous signal value gets out of the limits of a given interval, e.g. $[-A,A]$. Knowing F_ξ and f_ξ, the answer is obvious.

The expected value of probability variables ξ_t gives a superficial but often useful characterization of the signal behaviour:

$$m_\xi = M(\xi_t) = \int_{-\infty}^{+\infty} x\, f_\xi(x,t)\ \mathrm{d}x, \tag{2.8}$$

Similarly, another useful parameter is the expected value of the signal power in time t:

$$M(\xi_t^2) = \int_{-\infty}^{+\infty} x^2\, f_\xi(x,t)\ \mathrm{d}x \tag{2.9}$$

The function $m_\xi(t)$, $t \in (-\infty,\infty)$ is called the time-dependent expected value of the process ξ.

In certain cases the knowledge of the one-dimensional probability distribution function is not sufficient. For instance, such is the case when knowing ξ_{t1}, the value of ξ_{t2} ($t_2 \neq t_1$) has to be determined. Of course, ξ_{t2} may be estimated by $m_\xi(t_2)$ but (especially if t_2 is close to t_1) we obviously do not use reasonable the knowledge of ξ_{t2}. The estimation can be more precise if we know the two-dimensional distribution function characterizing the common behaviour of ξ_{t1} and ξ_{t2} This function is the vectorial probability distribution

$$F_\xi(x_1,x_2,t_1,t_2) = P(\xi_{t1} < x_1 \text{ and } \xi_{t2} < x_2) \tag{2.10}$$

The corresponding two dimensional density function (if it exists) can also be defined:

$$f_\xi(x_1,x_2,t_1,t_2) = \frac{\partial^2}{\partial_{x_1}\partial_{x_2}} F_\xi(x_1,x_2,t_1,t_2) \tag{2.11}$$

In some practical cases even the two-dimensional distribution is unknown. The solution of the estimation problem may be obtained, however, if L_ξ *autocorrelation function* is known

$$L_\xi(t_1,t_2) = M(\xi_{t1}, \xi_{t2}), \qquad t_1,t_2 \in (-\infty,\infty) \tag{2.12}$$

It is easy to find such problems where the distribution of the probability variable $\xi = (\xi_{t1},\xi_{t2},\ldots\xi_{tn})$ has to be known for solving the problem in full depth. On the other hand, there are also several signals whose probability distribution function completely determined just by two simple functions, namely by the time-dependent expected value and by the autocorrelation function. That is why many important problems can be solved even if not too much is known about them.

2.2.2. Stationary and Ergodic Processes

In the strict sense of the word, processes characterized by distribution functions insensitive to time-shift are called (strongly) *stationary*. More precisely, the signal is said to be stationary if it is true for all n>0, all series of $t_1,t_2,...t_n$, and for all τ, that

$$F_\xi^{(n)}(x_1,x_2,..x_n,\ t_1+\tau,t_2+\tau,..t_n+\tau) = F_\xi^{(n)}(x_1,x_2,..x_n,\ t_1,t_2..t_n)$$

This invariance to time-shift shows the invariability, i.e. homogeneity of the signal in time. Obviously, the parameters are simpler in this case, e.g. the one-dimensional distribution function is reduced to a single variable function

$$F_\xi^{(1)}(x_1,t_1) = F_\xi^{(1)}(x_1,\ 0) = F_\xi(x)$$

while the two-dimensional distribution is in fact a function of three variables only:

$$F_\xi^{(2)}(x_1,x_2,t_1,t_2) = F_\xi^{(2)}(x_1,x_2,\ 0,\ t_2\text{-}t_1)$$

The expected value is independent of time: $M(\xi_t) = m_\xi(t) = m_\xi,$ (2.13)

while the autocorrelation function depends only on the interval between t_1 and t_2:

$$M(\xi_{t1},\xi_{t2}) = L(t_1,t_2) = R(t_2\text{-}t_1) \tag{2.14}$$

Moreover, since L_ξ is a symmetrical function of t_1 and t_2, R_ξ is an even function.

It may also frequently happen that the distribution functions of a signal are not known but the signal fulfils the conditions given by the equations (2.13) and (2.14). Even in this case it is possible to solve some important practical problems. Processes exhibiting such a character are therefore classified individually and called *weekly stationary* signals.

It is an everyday experience that there are such signals or phenomena, realizations of which are quite different but their character, their long-time average or the characteristic rhythm of their fluctuations are the same or much alike. This feature is so much the more important as there is a chance just for such signals to give a good picture about the behaviour of other realizations by examining one of them. More precisely, a process is called *ergodic* if almost any of its realizations is suitable to deduce any of its distribution functions. It can be proved that a strongly stationary process is ergodic or can be composed as a mixture of ergodic processes. To estimate the parameters of such processes, time averaging is used. It can be shown for ergodic processes, for instance that the average

$A(\xi) = \lim_{T\to\infty}\frac{1}{T}\int_{t_0}^{t_0+T}\xi_t\,dt$ is equal to the expected value $m(\xi)$ of the process with a probability equal to one.

2.2.3. Linear Transforms of Stationary Processes

Suppose that ξ is an ergodic process with an expected value m_ξ. In accordance with the conclusion which was made at the end of the previous chapter, this expected value can be considered a DC-component of the process.

Let us determine the expected value of the signal appearing at a filter output if the input signal of the filter is the process ξ. The expected value of the output signal is expected to be dependent on m_ξ and on the DC-gain of the filter.

Suppose $h(.)$ is the impulse response of the filter so that the output signal at time t is

$$\eta_t = \int_{-\infty}^{+\infty} h(\tau)\,\xi_{t-\tau}\,d\tau$$

and its expected value is

$$M(\eta_t) = M\left(\int_{-\infty}^{+\infty} h(\tau)\,\xi_{t-\tau}\,d\tau\right)$$

If the impulse response satisfies certain conditions, then the computation of the expected value and the integration can be swapped:

$$M(\eta_t) = \int_{-\infty}^{+\infty} h(\tau)\,M(\xi_{t-\tau})\,d\tau = \int_{-\infty}^{+\infty} h(\tau)\,m_\xi\,d\tau$$

As m_ξ does not depend on the integration, it can be put in front of the sign and the remaining term can be extended by $e^{-j0\tau}$ (=1)

$$m_\xi(t) = m_\xi \int_{-\infty}^{+\infty} h(\tau)\,d\tau = m_\xi \int_{-\infty}^{+\infty} h(\tau)\,e^{-j0\tau}\,d\tau = m_\xi\,H(0)$$

which is exactly the result what was expected. $H(0)$ is the frequency response of the filter at zero frequency (the Fourier transform of the filter impulse response).

A further interesting question is whether the output autocorrelation function of the filter can be determined if the autocorrelation function of the stationary input signal is known. Will the signal remain stationary at all ? Without losing generality, we restrict to zero-mean processes. Due to the limited space, instead of a detailed discussion, only the result is presented. With appropriate restrictions the output signal η remains stationary and if the Fourier transform

$$s_\xi(f) = \int_{-\infty}^{+\infty} R_\xi(\tau)\,e^{-j2\pi f\tau}\,d\tau \qquad (2.15)$$

exists then the Fourier transform of the autocorrelation function of the output signal is

$$s_\eta(f) = s_\xi(f)\,|H(f)|^2, \quad \text{if} \quad f \in (-\infty, \infty) \tag{2.16}$$

The Fourier transform of the autocorrelation function is called the *spectral density* function of the stationary process. This is an even function as R_ξ is also even. As it will be shown in the following chapter, the function is also non negative, too.

2.2.4. Physical Meaning of Spectral Density Function

As we have seen previously, deterministic signals can be synthesised from a sum from an integral of harmonic components. It is an interesting question whether the same could be done with stationary random signals. The answer is positive but results in the definition of a not too expressive new integral term. Instead, we rest contented with an approach which is satisfactory in the majority of practical cases although it is not an alternative to a more thorough survey of spectral decomposition.

Supposing that a stationary process can also be composed of a sum of harmonic signals, it is quite acceptable that the output signal of a narrowband bandpass filter may contain only spectral components falling into the passband of the filter. Let us examine the power of the filtered signal. Using the notation of Fig. 2.1.

$$M(\eta_t^2) = R(0) = \int_{-\infty}^{+\infty} s_\eta(f)\, e^{j2\pi f 0}\, df = \int_{-\infty}^{+\infty} s_\xi(f)\, |H(f)|^2\, df$$

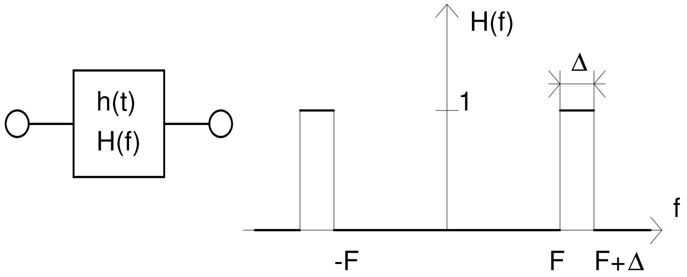


Fig. 2.1 Amplitude Response of an Ideal Bandpass Filter

(Here we exploited that R_η is the inverse Fourier transform of s_η.) Taking also into account that s_ξ is an even function and using the H shown in the figure

$$M(\eta_t^2) = 2 \int_{F}^{F+\Delta} s_\xi(f)\, df$$

If the bandwidth Δ is small then according to the theorem of integration calculus

$$M(\eta_t^2) \cong 2\, s_\xi(F)\, \Delta \tag{2.17}$$

Since only components having frequency near to F appear at the filter output, the resulting power is given by $M(\eta_t^2)$. However this quantity depends on the value of $s_\xi(F)$ representing a kind of spectral intensity distribution.

Control Questions

1. When is it reasonable to analyze a system by means of a stochastic signal?
2. How can a stationary signal be characterized?
3. What is the definition of the ergodic process?
4. When is the knowledge of the autocorrelation function essential?
5. What is the definition of the spectral density function?

Exercises

1. A stationary process having the expected value zero and constant spectral density at every frequency $f < B$ is called band-limited white noise. Compute the autocorrelation function of such a process.

2. Compute the spectral density function of the process characterized by the following autocorrelation function:

$$R_\xi(\tau) = R_o \exp(-|\tau|/T)$$

References

[1] Csibi S. et al.: Információ közlése és feldolgozása. Tankönyvkiadó, 1986.

[2] Gihman I.I.-Szkorohod A.V.: Bevezetés a stochasztikus folyamatok elméletébe. Mûszaki könyvkiadó, 1975.

[3] Papoulis A.: Probability, Random Variables and Stochastic Processes. McGraw Hill, New York, 1965.

3. SAMPLING. DISCRETE SOURCE CODING

It is a well-known fact that today's semiconductor technology is capable to produce digital devices having extremely high complexity at an affordable price, volume and weight. These features can also be utilized for solving tasks related to analog signals if conversion of the analog signal to series of symbols discrete both in time and in amplitude is possible. More precisely, the question is what are the pros and contras of mapping an analog signal into a series of symbols discrete in time and amplitude. In fact, investigations of these problems are covered by the term *sampling*.

Digital devices with computer-like architecture reduce eventually all tasks to operations with binary (two-state) symbols. Any series of symbols (e.g. sampled analog signal or a written text) can be converted into a series of binary symbols. The length of the converted (encoded) series is, however, not indifferent. Possibilities and limits of unambiguous coding will be discussed in the second section of this chapter.

3.1. Relation Between Analog Signals and Series of Samples

3.1.1. The Spectrum of Sampled Series

A widely used method of signal generation is that numbers stored in a computer-like device are periodically converted to electrical quantities, e.g. voltage. Such a device is called digital-to-analog converter (DAC) and it is usually integrated into one circuit. The output signal of the DAC is smoothed by a filter to eliminate unwanted components. It is an interesting question how the number stored have to be chosen to generate signals with specified shape.

To specify the task more precisely, let us suppose that a DAC and a smoothing filter are used to generate an absolutely integrable signal with the spectrum $X_e(.)$ by means of so far unknown series of numbers x_i, (i=0, ±1,..) entering the DAC input (see Fig. 3.1.). Suppose that the operation of the DAC is periodic in time (T). It is also important whether the converter is operated by narrow or by wide impulses.

The latter method is rather practical while the previous one serves as the computational model. Notice that the true digital-to-analog converter can be constructed by an ideal DAC and an ideal smoothing filter with impulse response $m(.)$. Using the notations of Fig. 3.1., the impulse response of the smoothing filter is

$$h(t)=m(t)*g(t), \qquad t\in(-\infty,\infty)$$

It also follows from the model that there are impulses of the magnitude x_iT/Δ and width Δ at the output of an ideal DAC. The filters respond to such excitation by the impulse response, so that

$$x(t) = T\sum_i x_i h(t-iT) \tag{3.1}$$

For sake of simplicity, let us assume that the output sample x_i is generated at time iT. The output signals of the ideal and the real DAC are also shown in Fig. 3.1.

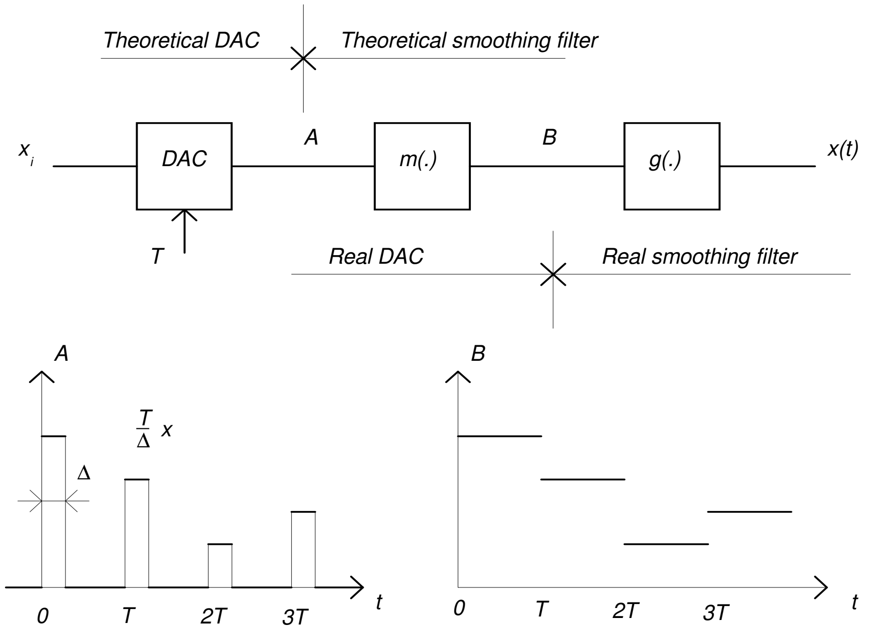


Fig. 3.1. Reconstruction of Analog Signal from Stored Samples

Fourier transform of $x(.)$ obviously exists if h is absolutely integrable and the sum of x_i (i=0,±1,...) exists. In this case

$$x(f)=\int_{-\infty}^{\infty}x(t)\,e^{-j2\pi ft}dt=T\sum_i x_i\;H(f)\,e^{-j2\pi fiT}$$

that is

$$X(f)\text{€}=\text{€}H(f)\text{€}T\text{€}\sum_i\;x_\mathrm{i}\text{€}\mathrm{e}^{-\mathrm{j}2\pi\mathrm{fiT}} \qquad (3.2)$$

where $H(.)$ is the transfer function of the smoothing filter.

From eq. (3.2) it follows that

$$X_\mathrm{m}(f)\text{€}=\text{€}T\text{€}\sum_i\;x_\mathrm{i}\text{€}\mathrm{e}^{-\mathrm{j}2\pi\mathrm{fiT}} \qquad (3.3)$$

behaves in the same way as if it was the Fourier transform of an absolutely integrable signal. This behaviour establishes the terminology for X_m as being the spectrum of the series $x_\mathrm{i,}$ i=0,±1,...

In fact, X_m exhibits the usual symmetry properties of the spectra of the real signals:

$$X_\mathrm{m}(-f)\text{€}=\text{€}X_\mathrm{m}{}^*(f), \qquad \forall f\text{€}\in\text{€}(-\infty,\infty)$$

moreover it is periodic in $1/T$:

$$X_\mathrm{m}(f+\frac{1}{T})\text{€}=\text{€}X_\mathrm{m}(f),\quad \forall f\text{€}\in\text{€}(-\infty,\infty)$$

Periodicity also means that (3.3) is the Fourier series of $X_m(f)$, i.e. a given X_m can be realized by Fourier decomposition resulting in series of x_i, i=0,±1,... .

So it can be registered that, on the base of stored samples, it is possible to generate a signal with spectrum X_e by means of a DAC and a filter:

$$X_e(f) = H(f)X_m(f), \quad \forall f \in (-\infty,\infty)$$

where X_m is a spectrum periodic in 1/T.

In the above example , X_e was given and H and X_m had to be chosen. In different practical cases the choice depends on the actual situation. The solution is relatively universal if the specification is band limited, e.g.

$$X_e(f) = 0, \text{ whenever } \quad |f| > B < 1/(2T).$$

In this case, $H(f)$ may be a lowpass filter (with the cutoff frequency at B) and

$$X_m(f) = \sum_k X_e(f+k/T), \quad \forall f \in (-\infty, \infty).$$

Fig. 3.2. illustrates the relation between X_m and X_e. Furthermore, it shows the passband and stopband of the smoothing filter. It also can be seen that in the case of $B < 1/(2T)$ an unspecified region between the passband and the stopband exists which is needed for the realization of the filter. Samples to be stored can be generated by the Fourier decomposition of the X_m, $i = 0,\pm1$, ... :

$$x_i = \int_{-B}^{B} X_m(f) e^{j2\pi f i T} df = \int_{-\infty}^{\infty} X_e(f) e^{j2\pi f i T} df \qquad (3.5)$$

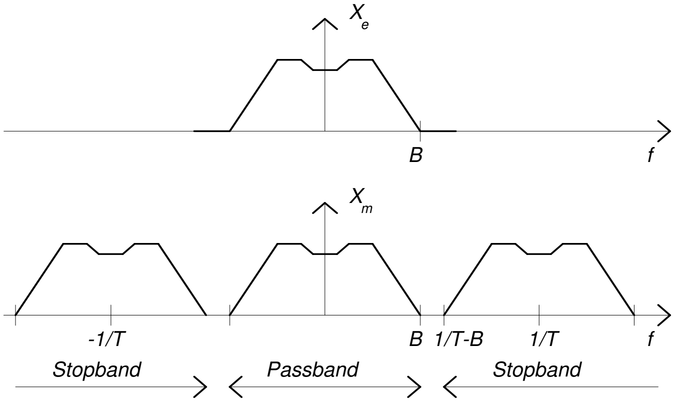


Fig. 3.2. Restoration of the Spectrum X_e

3.1.2. Signal Reconstruction from Equidistant Samples

Suppose that samples of the signal $x(t)$ were taken so that the time interval between the samples is identical, i.e.

$$x_i = x(iT), \quad i=0,\pm1,...$$

In this case the spectrum of the sampled signal (if x is absolutely integrable, then it certainly exists) is:

$$X(f)\text{€}=\text{€}T\text{€}\sum_{i} x(iT)\text{€}e^{-j2\pi fiT}$$

It may be expected that the spectrum of the sampled series and that of the signal x are in a simple relation with each other. Indeed, it is true that

$$X_{\mathrm{m}}(f)\text{€}=\text{€}\sum_{k} X(f+k/T), \quad \forall f\text{€}\in\text{€}(-\infty,\infty) \tag{3.6}$$

As the right side of the eq. (3.6) is the function of f periodic in $1/T$, it can be expressed in the form of Fourier series. Computing the coefficients, we obtain exactly $x(i\tau)$, $i=0,\pm 1,...$ (Q.e.d.).

There is an especially interesting practical case, in which -€using some suitable sampling frequency $f_s=1/T$€- it is true for all frequencies that only one non-zero element is in eq. (3.6) (the cumulative spectrum is not aliased). That means that the analog signal can be reconstructed from the samples of x, equidistant in T by means of the system shown in Fig. 3.1., provided that the transfer function matches the spectrum of the signal x, i.e.:

$$H(f)=\begin{cases} 1, \text{if } x(f)\neq 0 \\ 0, \text{if} x_m(f)\neq 0 \text{ and } x(f)=0 \\ \text{arbitrary, otherwise} \end{cases}$$

The straightforward consequence of the above statement is the Shannon's sampling theorem: An absolutely integrable signal can by reconstructed from its samples equidistant in T by means of an ideal lowpass filter with the cutoff frequency B€<€$1/(2T)$.

Shannon's theorem is valid not only for absolutely integrable signals but for harmonic signals and for stationary stochastic processes, too.

3.1.3. Spectral Density of Random Series

Let us examine the properties of a random signal reconstructed by means of a DAC and a smoothing filter from the series of random numbers. Suppose that ξ_i are probability variables with uniform distribution and zero-mean, moreover that they are uncorrelated, i.e.

$$\mathrm{M}(\zeta_i\zeta_j)=\begin{cases} \sigma^2, \text{if } j=i; \\ 0, \text{if } j\neq i. \end{cases}$$

The latter method is rather practical while the previous one serves as the computational model. Notice that the true digital-to-analog converter can be constructed by an ideal DAC and an ideal smoothing filter with impulse response $m(.)$. Using the notations of Fig. 3.1., the impulse response of the smoothing filter is

$$h(t)=m(t)*g(t), \qquad t\in(-\infty,\infty)$$

Suppose the smoothing filter of the signal generator is ideal and its cut-off frequency is 1/(2*T*) so that its frequency response is

$$h(t) = \frac{1}{T}\frac{\sin(\pi t/T)}{\pi t/T}, \quad \forall t \in (-\infty,\infty)$$

Under such a set of conditions it can be proved that the generated signal will be a stationary, zero-mean signal with the spectral density

$$s_\eta(f) = \begin{cases} \sigma^2 T, \text{ if } |f| < 1/(2T) \\ 0, \text{otherwise} \end{cases} \qquad (3.7)$$

3.1.4. Quantization Noise. Nonlinear Quantization

When an analog signal is converted to digital form, the analog samples are replaced by code word belonging to a finite set *N* of code words. This means that there are only *N* different samples which can be exactly represented. It may be a natural (but not necessary) requirement that multiples of a basic unit are assigned to the samples. Generally (but again, not necessarily) *n* bit binary code words are assigned to the samples accordingly to a simple rule. For instance, two-s complement code is such a widely used representation.

Using *n* bit code words, it is possible to distinguish $N=2^n$ sample values. The analog-to-digital converter (ADC) generates the following exact sample values:

$$x = \Delta i, \quad i = -N/2,...0,1,...N/2-1 \qquad (3.8)$$

Conversion range is an important parameter of such an ADC. If this range is defined by the interval $(-C,C)$ then the magnitude of the quantization steps will be $\Delta = 2C/N$.

Of course, the sample is usually not equal to any of the precise values given by (3.8). In fact, the converter substitutes a sample by the code word representing the value which is closest to the actual value of the sample. Thus the precise value *x* differs from the actual value *x*:

$$\hat{x} = x + \varepsilon$$

where ε can vary between -Δ/2 and Δ/2.

The difference ε is called the *quantization noise*. In simple models, the quantization noise is modelled by a uniformly distributed probability variable. Furthermore, it is also assumed that the instantaneous values of the quantization noise added to different samples are not correlated, i.e.

$$M(\varepsilon_1 \cdot \varepsilon_2) = 0$$

This model of the quantization noise is useful when the bandwidth of the stationary stochastic signal is relatively great in comparison to the sampling frequency ($B \sim f_s/2$).

In Chapter 3.1.3. we have seen that the noise process reconstructed from uncorrelated noise samples has a constant spectral density within the range of $|f|<f_s/2$ and its power is

equal to that of the samples. Since the signal reconstructing system is linear, the signal appearing at the output is

$$\hat{x}(t)\text{€}=\text{€}x(t)\text{€}+\text{€}\varepsilon(t)$$

where x(t) is the original input signal. The power of the process $\varepsilon(t)$ can be computed from the distribution of the samples:

$$P_{\varepsilon}\text{€}=\text{€}M(\varepsilon^2(t))\text{€}=\text{€}M(\varepsilon^2)\text{€}=\text{€}\int_{-\Delta/2}^{\Delta/2} x^2\ f_{\varepsilon}(x)dx=\ldots=\frac{\Delta^2}{12}.$$

The subjective measure of the effect caused by the quantization noise can be well defined by the ratio of the reconstructed signal power and the quantization noise power, which is called the *signal-to-noise ratio*. The maximum amplitude of the sinusoidal signal is C so that the maximum power is

$$P_{\mathrm{x}}\text{€}=\text{€}C^2/2,$$

The signal-to-noise ratio is then

$$\mathrm{SNR}\text{€}=\frac{P_x}{P_{\varepsilon}}=\frac{C^2}{2}:\frac{\Delta^2}{12}=6(C/\Delta)^2$$

Knowing that $\mathrm{C}/\Delta=N/2=2^{n-1}$, the signal-to-noise ratio for the maximum amplitude sinewave is

$$\mathrm{SNR}\text{€}=\frac{3}{2}2^{2n}, i.e.\,\mathrm{SNR}[\mathrm{dB}]=1.74+n6.02 \qquad (3.9)$$

Of course, the signal-to-noise ratio is significantly smaller if the power of the converted signal is well bellow the permissible limit.

In a significant part of telecommunication applications, the average power of the sampled and A/D converted signals is within a range of about 36 dB. More precisely, the signal at the ADC input may have the maximum amplitude C but might also have just 1/4000th power (C/64 amplitude, i.e. -36 dB) of the previous one. Should it be required to have 36 dB signal-to-noise ratio even for such a low-level signal, it would result in an unnecessarily great signal-to-noise ratio for high level signals, e.g. for the maximum signal it would be 36+36=72 dB, which could be satisfied by n=12.

The representation can be made more dense if the precise sample values are not chosen as equidistant. In the range of $|x|<C/64=C_{\mathrm{o}}$, let us have the distance $\Delta_{\mathrm{o}}=C/32$ so that 64 divisions are in this range. This is just enough to satisfy the 36 dB SNR for the low-level signals. In the next range where $C_{\mathrm{o}}<|x|<2C_{\mathrm{o}}$ the distance is doubled to $2\Delta_{\mathrm{o}}$ so that for the signals with the amplitude $2C_{\mathrm{o}}$ the SNR remains the same but the number of samples is only 32 in this range. This procedure can be continued until the entire range $|x|<C$ is covered. It is easy to count that only 256 samples shall be precisely represented using this

procedure so that the n=8 bit code word length meets the above SNR requirement. The price we have to pay for this kind of logarithmic conversion is that the relation between the analog samples and the code words assigned to them is not as easily seen as it was for the linear code. In practice, the logarithmic compression of 8 bit code words is performed by 13 bit ADCs and an appropriate postprocessing of the obtained samples.

3.2. Symbol Series as Information

It is a frequent task to replace elements of a finite set of symbols by another set of symbols, e.g. to convert a text into a series of 0's and 1's. In the model of such a task, the elements of the set to be converted are called *source symbols* and the conversion procedure is called *coding*. The properties, possibilities and limits of coding are examined by the information theory, specifically by the source coding theory. This theory is motivated by the fact that the extent of the coded text, the *coding density*, is by far not indifferent for the user. To be able to define this question we have to characterize the source of the coded symbols.

The essential parameter of the source is the set of its symbols, the source alphabet. The source is well defined by listing all its symbols, e.g. $a_1,a_2,...$€a_n. The source message is understood as a series of symbols to be encoded. These may be so long that they even cannot be taken into consideration in source modelling; in this case we are speaking about infinite series of symbols. It is not a too good description of a real source but it is a simple and well handled model if the source symbols listed in message are supposed to be independent and to have equal random distribution. In this case the source is called *stationary* and *memoryless* and it is fully characterized by the source symbol distribution, i.e. the system of probabilities

$$p_k€=€P(y_i=a_k), \quad i=0,\pm 1,... \; ; \; k=1,2,...n$$

3.2.1. Coding Density

Let us examine the binary coding of a source of n elements with the distribution P€=€$(p_1,...p_n)$. Obviously, the source symbols can be represented by the series of 0's and 1's having the same length k if it is true that

$$2^k€\geq€n$$

so that the code can be unambiguously decoded. Greater coding density can be achieved if code words of different lengths may be used. It is relatively easy to decode such codes for which one can decide after reading a certain number of code bits whether they form a valid code or not. This type of code is called the prefix code. It is easy to see that as far as coding density is concerned, the prefix codes do not restrict the coding possibilities.

Obviously, to be able to decode a code unambiguously, the code lengths l_i (i=1,2,...n) must be longer than a certain minimum length. This relation is given by the Kraft inequality stating that a code can be unambiguously decoded if it is true for the code lengths that

$$\sum_{i=1}^{n} 2^{-l_i} \le 1. \qquad (3.10)$$

From the point of view of the total length of the message, the expected value of the code word length is of importance. If the probability of an l_i long code word is p_i then the expected value of the code word length is

$$\lambda\text{€}=\text{€}\sum_{i=1}^{n}\text{€}l_\mathrm{i}\text{€}p_\mathrm{i}. \qquad (3.11)$$

Obviously, it is reasonable to assign short code lengths to symbols with great probability and vice versa. It can be proved that for any (unambiguous) code

$$\lambda\text{€}\ge\text{€}\sum_{i=1}^{n} p_\mathrm{i}\text{€}\mathrm{ld}(1/p_\mathrm{i}) \qquad (3.12)$$

It is also true that choosing

$$\mathrm{ld}(1/p_\mathrm{i})\text{€}\le\text{€}l_\mathrm{i}\text{€}<\text{€}\mathrm{ld}(1/p_\mathrm{i})\text{€}+\text{€}1,$$

(3.10) is satisfied so that a code exists for which

$$\sum_{i=1}^{n}\text{€}p_\mathrm{i}\text{€}\mathrm{ld}(1/p_\mathrm{i})\text{€}\le\text{€}\lambda\text{€}<\text{€}1\text{€}+\text{€}\sum_{i=1}^{n} p_\mathrm{i}\text{€}\mathrm{ld}(1/p_\mathrm{i})$$

There is a particular function of p_i probabilities, a feature of source distribution playing role in the limits obtained for average code length. This feature is called *source entropy*:

$$H(P)\text{€}=\text{€}\sum_{i=1}^{n} p_\mathrm{i}\text{€}\mathrm{ld}(1/p_\mathrm{i}) \qquad (3.13)$$

3.2.2. Coding of Symbol Series

If the source entropy is not too great (≅€1€bit/symbol) then the limit (3.12) given for the average code length is not too strict. For the practical point of view it is not indifferent whether the average length of a code is closer to the shortest or to the longest code length.

The limit given by eq. (3.12) ensures the existence of highly efficient code in the case of high entropy sources. It is possible to use a source equivalent to the original one but having high entropy. As the source symbols let us take the K symbol messages of the original source. This extended source will have n^{K} symbols and its entropy will be $K\text{€}H(P)$ since the source is memoryless. So that for the average length of the best code representing the original K set of symbols, the following inequality can be given:

$$K\text{€}H(P)\text{€}\le\text{€}\lambda^{(\mathrm{K})}\text{€}<\text{€}1\text{€}+\text{€}K\text{€}H(P)$$

The average code length assigned to a single symbol of the original source can thus arbitrarily approximate the entropy of the original source.

The source entropy thus fully determines the possible coding density of the source and in this sense it is characteristic for the information content of the source messages. The length of the bit series of the code matching the source best may serve as a measure of the information content of the source messages.

Control questions

1. What conditions are needed to define the spectrum of series of numbers?
2. What are the characteristics of a spectrum of a continuous equidistantly sampled signal?
3. Under what conditions is there no correlation between the samples of quantization noise?
4. What is coding density limited by?
5. Why is it more advantageous to encode series of symbols instead of encoding single symbols?

Exercises

1. What is the spectrum of the series x_i€=€2^{-i}, i€=0,1,... ?
2. What is the entropy of the probability distribution p_i€=€2^{-i}, i€=€1,2,...?

References

[1] Csibi S. et al.: Információ közlése és feldolgozása. Tankönyvkiadó, 1986.
[2] Tusnády G.-Ziermann M.: Idõsorok analízise. Mûszaki Könyvkiadó, 1986.
[3] Lindner T.-Lugosi G.: Bevezetés az információelméletbe. Tankönyvkiadó, 1990.

4. FUNDAMENTALS OF ACOUSTICS. SOUND. ELECTROACOUSTICAL TRANSDUCERS.

A significant part of the incoming information is perceived by our ears and then processed by hearing. Sound processing and transmission is, therefore, one of the important areas of telecommunication systems. To be able to design equipment used in this field, one must become well acquainted with the physical properties of sound and with the psychophysical characteristics of the ear. This chapter describes physical parameters of the sound, quantities which handle the physiological characteristics of hearing, artificial sound field and the sound perception and reproduction techniques.

4.1. Physical Aspects of Sound

Sound is the mechanical vibration of an elastic medium. Influenced by an outer force, the particles of the elastic material are displaced and because of the elastic force and the inertness they begin to vibrate. Vibration propagates in solids, liquids and gaseous materials, as well. As the human ear perceives generally airborne sounds, the generation, propagation and perception of airborne sound is of essential importance.

Airborne sound appears as the fluctuation of air pressure. Sound pressure can therefore be regarded as an alternating component superposed to constant (or very slowly varying) atmospheric pressure (see Fig. 4.1.). Air pressure *P(t)* in a given point of space can be expressed as the sum of the atmospheric pressure P_0 and of the sound pressure *p(t).*

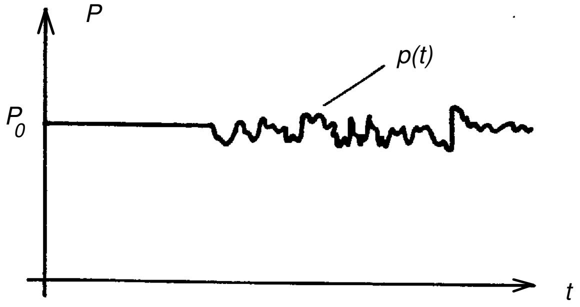


Figure 4.1. The Air Pressure

$$P(t) = P_0 + p(t) \tag{4.1}$$

In the following discussion we will examine only sound pressure. To characterize the magnitude of sound, the effective value of sound pressure is used. The standardized unit of sound pressure is the Pascal (1 Pa = 1 N/m^2). (Atmospheric pressure is about 100.000 Pa).

Sound pressure is measured by microphones. Instead of its actual value, sound pressure is usually given in dB being compared to certain reference pressure. 20 μPa is used as reference, since this is the level of the 1 kHz sinusoidal sound just yet audible by the average human ear. The sound pressure level can therefore be given as SPL = 20 lg p/p$_o$

$$L_p = 20\lg\frac{p}{p_0} \qquad (4.2)$$

Pressure difference generated in one point of the space tends to equalize towards the adjacent parts. During this process, particles of the air are displaced generating thus a new pressure difference in the neighbouring space.

So the sound pressure variation propagates in the form of sound waves. The distance between two points having the same phase of the sound wave is called the *wavelength*. The product of the frequency and the wavelength is equal to the *propagation velocity* of the sound wave:

$$c = f \cdot \lambda \qquad (4.3)$$

The propagation velocity of the sound is 340 m/sec.

If the sound source is concentrated in a single point and if there is no obstacle in the surrounding field then the sound waves are spherical. Far enough from the source, the curvature of the sphere can be neglected and the wave is supposed to be plain (Fig. 4.2.). For plain waves the ratio of sound pressure and particle velocity is constant:

$$\frac{p}{\nu} = \rho_0 c = 410 \;\; \frac{\mathrm{kg}}{\mathrm{m}^2 \cdot \mathrm{s}}, \qquad (4.4)$$

where ρ_o is the air density.

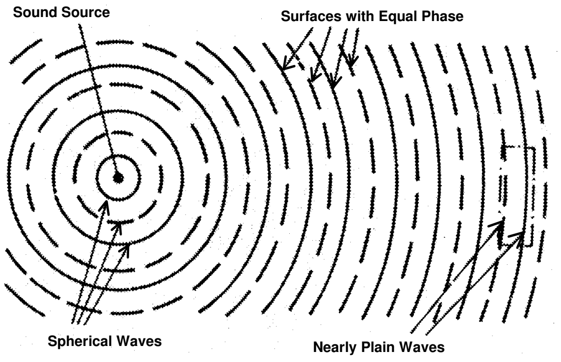


Figure 4.2. Spherical and Plane Waves

Sound can also be characterized by the acoustic power which comes through a unit of area. This quantity is called *sound intensity* and its value can be expressed as the product of sound pressure and particle velocity:

I€=€p.v€=€p^2/(ρ_o€c) (4.5)

Intensity is usually given in related form in dB, too. The reference is the same as for sound pressure. It can be seen easily that the reference value is I_o= 1 pW/m^2 which is the intensity of a just yet audible 1€kHz tone. Intensity can, therefore, be expressed as

$$L_I = 10\lg\frac{I}{I_0} \tag{4.6}$$

4.2. Physiological Characteristic of the Human Hearing

Human hearing is limited both in frequency and amplitude. According to the test measurements carried out on a very large amount of people, levels of just audible sound pressures were determined. The average of the measured values is called the *threshold of audibility*. The threshold of audibility is strongly frequency dependent. The ear is most sensitive in the range of some few kHz, below and beyond this range the sensitivity is smaller (see Fig. 4.3.) It can be seen that the range of the audible signals is between 20 Hz and 20 kHz. Too loud sounds cause pain, the lowest of such a sound pressure is called the *threshold of pain*. Musical sounds and the speech are within these limits. It can also be seen that the frequency range and the amplitude range of music are remarkably greater than that of speech.

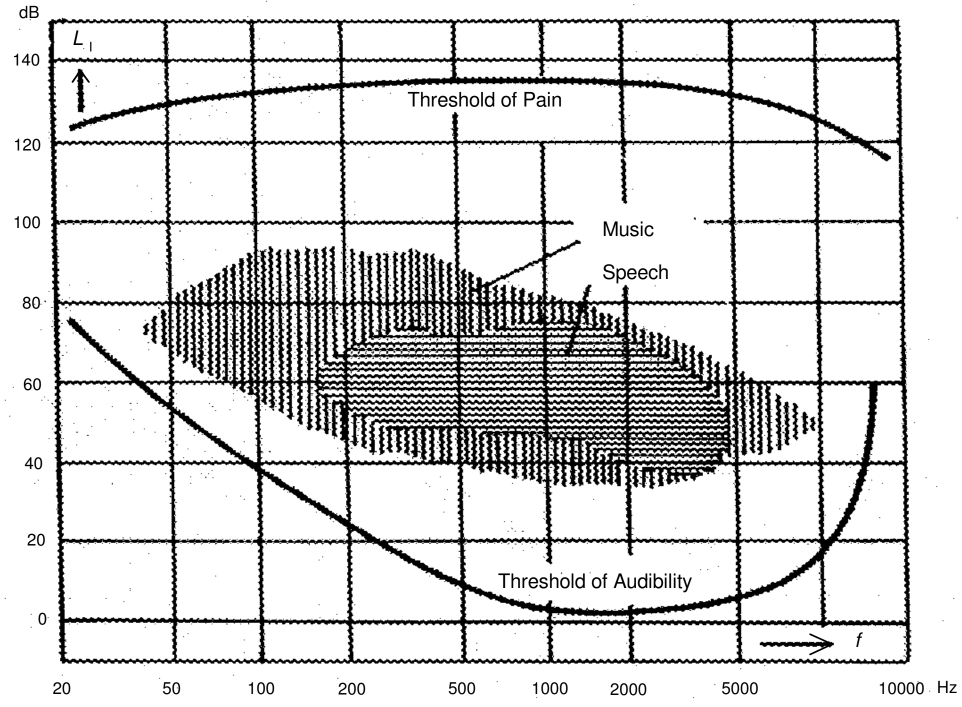


Figure 4.3. Limits of Human Hearing

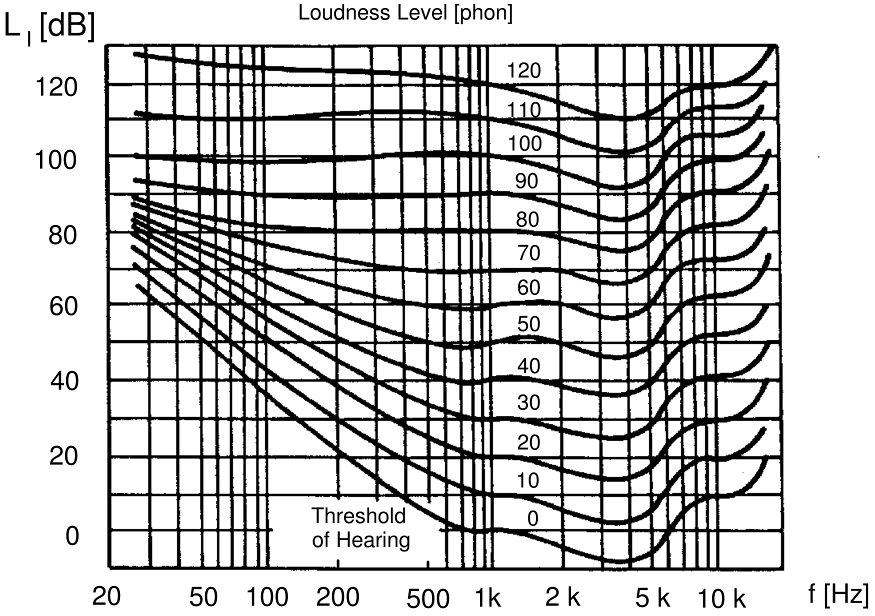


Figure 4.4. Equal Loudness Level Contours

For the subjective judgement of the sound level, the term of *loudness level* was introduced. Loudness level of an arbitrary sound is as many *phon*-s as many dB-s is the sound pressure level of the 1 kHz tone of the same loudness. (In the loudness evaluation test, the listener alternately listens to and compares the measured sound with the reference sound which can be varied by himself until the two sounds are judged to be equal.)

Connecting the points with the same loudness along the frequency axis, we obtain the so called Fletcher-Munson curves (see Fig. 4.4.). Loudness level of a sound at a given frequency and pressure level can be read as the value of L_N belonging to a certain curve on the diagram. Loudness is thus suitable for the comparison of sounds with different frequencies.

To evaluate the resulting level of simultaneous sounds, the term *loudness* has been introduced. Loudness is denoted as N and its unit is called *son*. If the loudness level is greater than 40 phon then the loudness can be computed as

$$N = 2^{\frac{L_N - 40}{10}}$$

Masking is another term related to the simultaneous presence of two different sounds. Masking means covering the weaker sound with a stronger sound when each has a different frequency. Masking has been examined for sinusoidal sounds and for narrow- and broadband noises, respectively. Fig. 4.5. shows the increase of the audibility threshold caused by 1 kHz narrow-band masking noises. As it can be seen in the figure, high frequency sounds are easier to mask.

Spatial parameters of sound are also very important. First of all, *direction of the sound* has to be mentioned. In the horizontal plane, the sound is localized by the difference of the sound pressures at our ears. At low frequencies the phase difference is detected while on higher frequencies a difference in intensity arrises due to the shadowing caused by the head. To perceive direction in the vertical plane, the head has to be moved up and down.

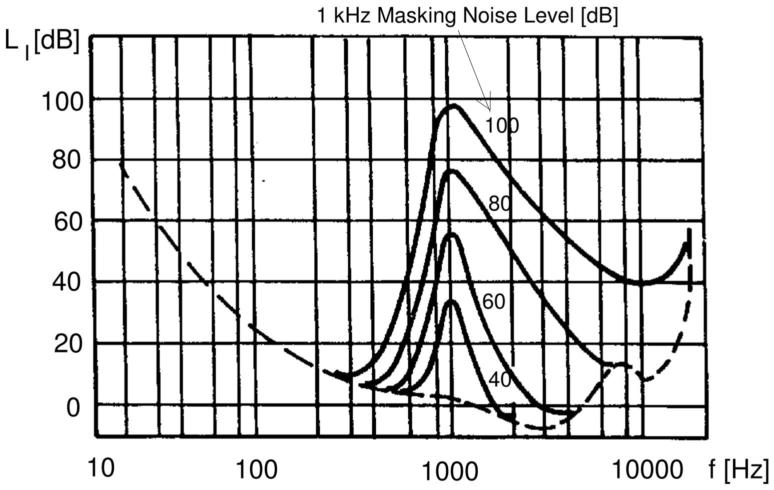


Figure 4.5. Masking

4.3. Production of Artificial Sound Field

It has been a long-existing need to produce such an sound field which contains all the information important for the human ear. The extent of satisfaction of such a need is different for the various areas of communication systems as it is not always necessary to consider the entire frequency band or the 120 dB dynamic range and/or the binaural effect of the audible sounds.

Let us consider the steps of the sound field producting process. First, the sound is perceived by a microphone in the original field. To get more spatial information, several microphones are sometimes used. Microphones convert the sound into electrical signals proportional to sound pressure. The signals are then processed, e.g. signals of the microphones are individually amplified, filtered, reverberated, added, etc. The processed signal is then passed through the communication channel which can be either wired or wireless.

In simple cases, e.g. in direct telephone communication there is no need for signal processing. Signals received by the *receiver* are then converted so that they become suitable for feeding the *loudspeaker* or the *headphone*. These transducers convert electrical signals into sound. The process described above is performed in real-time, i.e. the artificial sound field is follows (almost) simultaneously the original sound field.

In other cases the processed signal is stored by a *sound recorder*. In these cases the sound-record (e.g. a disc, a cassette, etc.) is spread to the customers who can listen to it whenever they want.

If the sound field is recorded by just one microphone or the signals of more microphones are added, the sound is transmitted or recorded as *monophonic*. Of course, in this case directional reception is not possible, all the sound sources of the original field will be heard from one direction in the restored field.

To produce *binaural effect*, at least two properly chosen signals have to be recorded and then separately transmitted. This principle was realized by the *stereophonic* system which was introduced in the '60-s and worked out for AM and FM broadcast as for several audio recording systems, as well. For the best stereo effect, it is recommended to arrange the two monitoring loudspeakers and the position of the listener so that they form an equilateral triangle.

In the '70-s, the number of independent sound channels was extended to four thus establishing the so called *quadrophonic* system. In this system, the best effect can be achieved by a quadratic arrangement of the loudspeakers and placing the listener right in the middle of the square. Quadrophony, however, has not been widely spread mainly because of financial reasons. All the three methods of sound field reproduction are presented in Fig. 4.6.

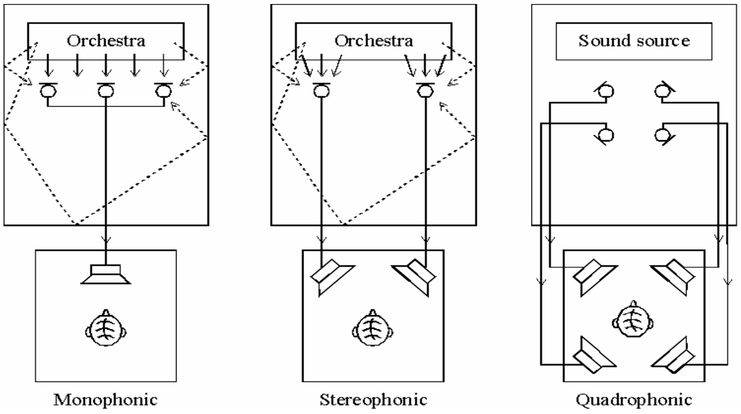


Figure 4.6. Sound Field Reproduction Techniques

4.4. Sound Transmission. Quality Requirements.

Besides the number of independent channels, there are also great differences in the bandwidth, dynamic range, signal to noise ratio and distortion of the transmitted signals.

The system *bandwidth* is usually given by a lower and an upper limit which are defined by the frequencies on which the signal level decreases 3 dB-s, compared to the middle frequency level. *Signal to noise ratio* is given by the effective value of the signal and that of the noise, expressed in dB. *Dynamic range* is the ratio of the strongest and the weakest portions of the signal, given in dB. It follows from the last two definitions that the dynamic range cannot be greater than the signal to noise ratio. *Harmonic distortion* is given by the ratio of the effective value of the harmonic signals and the effective value of the fundamental frequency signal, expressed in %.

Requirements for the quantities listed above strongly depend on whether speech or music has to be transmitted and reproduced. Requirements for the speech transmission are less than those for the so-called Hi-Fi quality which is suitable for high fidelity reproduction of the music. These quality requirements were first declared by the German standard DIN 45500. The highest technical level is represented by the so-called studio quality. At first, this quality could be achieved only by expensive radio and TV studio equipment, nowadays several digital audio products of consumer electronics meet the standard.

For telephone systems, economical speech transmission and retention of speech intelligibility is important. As it can be seen from the Fig. 4.3., speech signals components fall into the region of some hundreds of Hz to 4-5 kHz. According to listening tests, speech intelligibility will be maintained even if the upper frequency limit is set to 3000 Hz. To identify a speaker, a somewhat wider range is needed so that frequency range from 300 Hz to 3400 Hz has been standardized for telephone systems. Other requirements are also very modest, the signal to noise ratio is not higher than 20-25 dB and the distortion can be as high as 5-10%.

Somewhat better parameters are achieved in short and medium wave *amplitude modulated* (AM) radio broadcast where music programs are also transmitted. Here the upper frequency limit is 4.5 kHz and the signal to noise ratio can be as high as 40 dB.

Even better are the parameters of *frequency modulated* (FM) radio broadcast which is less sensitive to noise. FM broadcasting stations operate in the VHF (very high frequency) range and the frequency distance between two neighbouring stations is 300 kHz. The transmitted frequency range extends from 50 Hz to 15 kHz, the signal to noise ratio is higher than 60 dB and the distortion can be kept under 1%. Sound parameters of the terrestrial television broadcast are similar to these.

Since the end of the '60-s, FM stereo broadcast has also been widely used. Two independent signals of the same quality as for the FM mono broadcast are transmitted. *Crosstalk* between the two channels is about -40 dB which is low enough for good stereo effect. The only drawback is a worse noise immunity. In the '70-s, some experimental quadrophonic broadcasts in the VHF range were realized. According to the listening tests, the localization of the sources was very good in all directions.

Since the end of the '80-s, *satellite TV broadcasts* are spreading more and more. These programs are usually transmitted with multiple audio channels for different purposes. For instance, they can be used for multilingual or stereophonic accompanying sound or even for independent transmission of additional radio broadcasting programs.

Nowadays *Digital Satellite Radio* (DSR) broadcast is coming up. The special feature of this system is that besides the digital sound, additional data (subcodes) are transmitted enabling listeners to group and select different sorts of programs (e.g. news, pop music, etc.).

Let us now make an overview of the different sound recording techniques. The oldest of them is the *mechanical recording* which went through great changes until it appeared in present shape (record player). Frequency range recorded on a present-day disc is between 40 Hz and 16 kHz, the signal to noise ratio reaches 50-60 dB and the harmonic distortion is about 1%. Two (stereo) channels are recorded, the crosstalk is about -30...-40 dB.

The *analog magnetic recorders*, better known as tape recorders form the greatest group of recording devices. These recorders are designed for a great variety of applications starting with dictaphones suitable for speech-recording only, and ending with high-end multi-channel studio equipment. Currently, the parameters of a good-quality cassette tape recorder (*deck*) are nearly as good as those of a record player. The accompanying sound of a normal video player has a somewhat worse quality then a cassette tape recorder. The upper frequency limit is about 8 kHz as its tape speed is slower. The Hi-Fi video recorders record two sound channels with the bandwidth from 40 Hz to 16 kHz and with 70 dB signal to noise ratio.

Digital *compact disc* (CD) uses 16-bit resolution and 44.1 kHz sampling frequency to replay two independent sound channels in 10 Hz to 20 kHz frequency range, with 96 dB signal to noise ratio and harmonic distortion less than 0.005%, thus producing a stereo sound field of excellent quality.

The *Rotary Head Digital Audio Tape Recorder* (R-DAT) works with 12 and 16 bit resolution and with sampling frequencies 32, 44.1 and 48 kHz. Using sampling frequencies above 40 kHz, the same quality as for the CD is reached.

The overview given above has to be regarded as the present "state of the art", as newer and newer equipment are appearing day by day, e.g. Philips Digital Compact Cassette (DCC), Sony Minidisc (MD) and others.

4.5. Electroacoustic Transducers

Electroacoustic transducers are devices transforming electrical energy into acoustic energy or vice versa. Transformation is carried out in two steps. First, electric energy is converted into mechanical energy by means of an electromechanical transducer. The important part of such a transducer is a mechanical vibrating system which is rigidly attached to a diaphragm. The mechanical vibration forces the air particles adjacent to the diaphragm to move so that mechanical energy is converted into the acoustic energy of the propagating sound waves. In the *inverse* effect, incoming sound waves bring the diaphragm and the mechanical system into motion so that a signal proportional to the motion is generated at the electrical output. In certain transducers the diaphragm and the mechanical part cannot be clearly distinguished from each other.

The so-called *controlled* transducers use the input signal to control output energy of an external source. This principle has the advantage of being able to control much greater energies than the input energy. That is why these transducers are used to be called *active*, as well. Carbon microphones for telephone sets are a good example for such kind of transducers. In the following we shall discuss the most frequently used electromechanical transducers.

The *electromagnetic transducer* (Fig. 4.7.) consists of a permanent magnet, soft magnetic pole-shoe, an anchor and a spring. At the standstill position the gap size is *s*/2, attracting force of the magnet is in equilibrium with the force of the spring. In the presence of a current flowing through the coil, attracting force gets stronger and the air gap becomes smaller. On the contrary, a current flowing in the opposite direction weakens the attracting force thus the anchor moves off. In the inverse operation, force generated by the sound pressure moves the anchor. In accordance with the direction of the movement, a flux change will induce voltage in the coil.

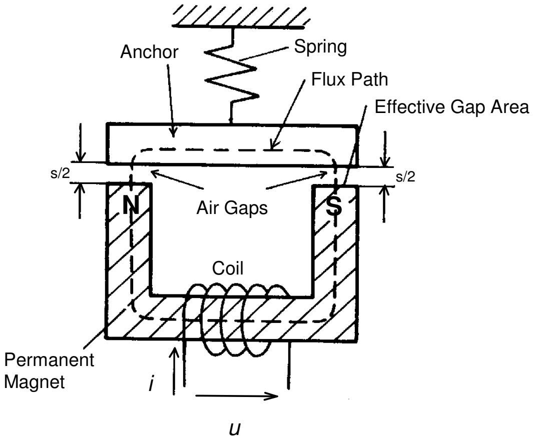


Figure 4.7. Electromagnetic Transducer

The *electrodynamic transducer* (Fig. 4.8.) has a fixed air gap. Inside the air gap, there is a magnetic field in which a conductor is moving. At the ends of the conductor, a voltage is induced which is proportional to the flux density in the gap, to the length of the conductor and to the velocity of the moving conductor. Thus the motional energy is transformed into electrical energy. When the conductor is driven by a current, a force is generated which is also proportional to the flux density, to the conductor length and to the magnitude of the current. To increase the efficiency of the transducer, a moving coil is used instead of a single piece of a straight conductor, since the entire wire length is of importance.

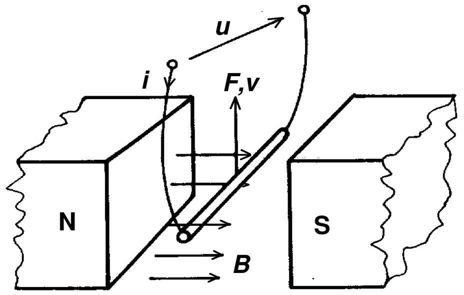


Figure 4.8. Electrodynamic Transducer

The *electrostatic transducer* (Fig 4.9.) is essentially a capacitor consisting of one fixed and one moving electrode. The moving electrode is made of a thin metal foil acting also as the microphone diaphragm. The fixed electrode is called back plate and is made of a thick piece of metal. To ensure linear operation of the transducer, a DC voltage source is connected to the microphone through a high value resistor. The electrostatic force attracts the thin diaphragm towards the back plate. As the edges of the diaphragm are fixed, the foil becomes deformed. This deformation is strengthened or weakened by an additional AC voltage, depending on its sign with respect to the DC voltage so that the diaphragm starts to move. If sound pressure is acting on diaphragm, it is again more or less deformed and due to this, capacity of the condenser formed by the two electrodes will change. As the charge of the capacitor is unable change during rapid changes of the capacitance, the voltage changes and the change can be monitored on the resistance R.

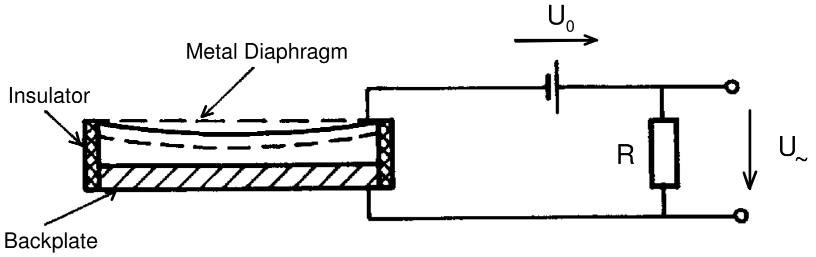


Figure 4.9. Electrostatic Transducer

The *piezoelectric transducer* makes use of the ability of certain materials to produce unbalanced charge distribution on their surface when they are deformed. In the inverse effect, deformation appears when electrical field is applied.

4.6. Microphones

Because of the great variety of requirements for the transmitted and recorded sound, several types of microphones are used for sound reception. A microphone can be characterized by its sensitivity, frequency response of the sensitivity and by the directional characteristic. *Sensitivity* is defined as the output voltage related to the unity of pressure. *Frequency response* is the sensitivity given as the function of frequency. *Directional characteristic* expresses the dependence of the sensitivity on the direction of the incoming waves (see Fig. 4.10.).

Directional characteristic depends on the microphone housing. In the case of closed housing, the microphone is *omnidirectional* (sensitivity is the same in all directions). If the housing is opened (both sides of the diaphragm can be accessed by the sound pressure) the microphone is *bilateral*, i.e. it is entirely insensitive to side sound-waves and most sensitive to sounds coming in from the directions perpendicular to the diaphragm. The third important type is the so called *cardioid* microphone which has great sensitivity to one direction only and there is a total suppression in the backward direction.

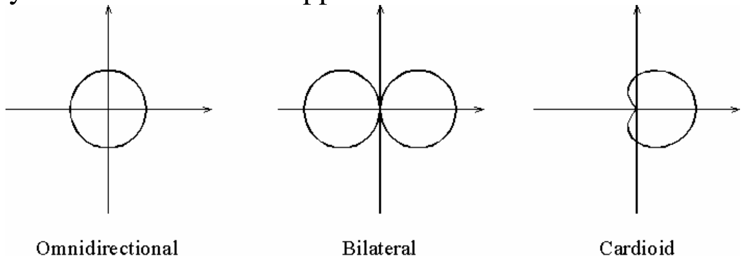


Fig. 4.10. Microphone Directivity Patterns

In telephone sets, mass produced cheap *carbon microphones* are used. Such a microphone acts as a resistor varying its resistance according to the change of the sound pressure on the diaphragm. This is due to the change of the contact resistance of carbon powder placed between two gold plated contacts. The bottom electrode is fixed and isolated from the housing while the upper electrode moves together with the diaphragm. Since the contact resistance is a nonlinear function of the diaphragm displacement, the distortion of the microphone is rather high. The wide-spread use of this type of microphones is due to its great output signal. Currently, they are loosing importance as fully electronic telephone sets enter the market.

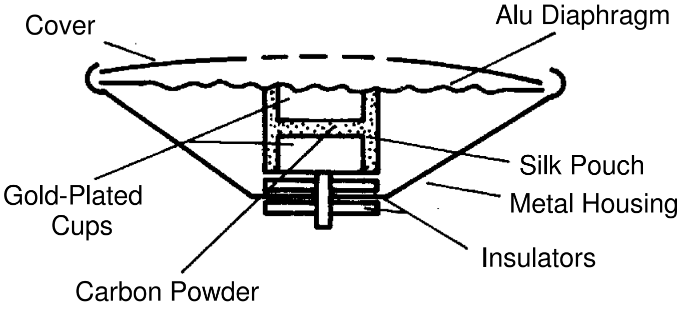


Figure 4.11. Carbon Microphone

Both in studio and consumer applications, *dynamic microphones* are frequently used (see Fig. 4.12.). Induced output voltage appears at the ends of the moving coil, inserted into the air gap of a permanent magnet. The coil moves together with the diaphragm and the motion is proportional to the incoming sound pressure. The magnet and the diaphragm are placed into a housing closed from the front side by a protecting grid. Opening the housing, choosing proper grid parameters and placing additional acoustic elements, broadband frequency response and arbitrary directional characteristic can be realized.

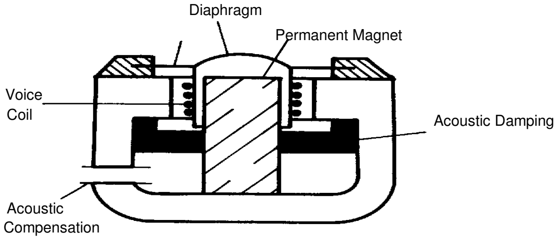


Figure 4.12. Dynamic Microphone

The *condenser microphone* (Fig. 4.13.) is a device suitable for both studio applications and for measurements. There is a disk-shaped back-plate, isolated from the cylindrical metal housing at the end of which a metal diaphragm is stretched. The distance between the two electrodes is about 0.01 mm. Alternating voltage, which appears on the resistance is amplified by a low-noise high input impedance preamplifier. The sensitivity of such a microphone can be calibrated and remains stable during a long period of time because of the precise construction.

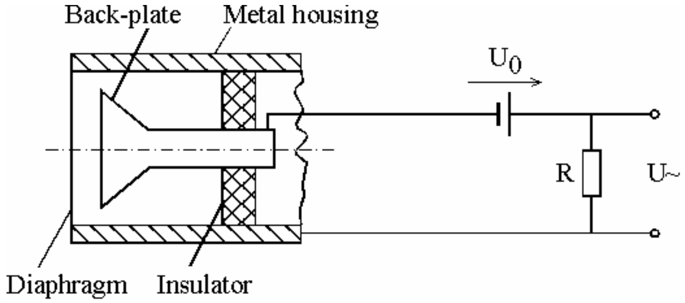


Figure 4.13. Condenser Microphone

For consumer applications, *piezoelectric microphones* are used (Fig. 4.14.). A piece of a dual crystal (bimorph) is fixed at one end to the housing while its other end is attached to the diaphragm. The incoming pressure bends the bimorph to produce two voltages of opposite sign which are then simply added. As the output impedance of the device is high, high input impedance amplifier has to be used for further signal amplification.

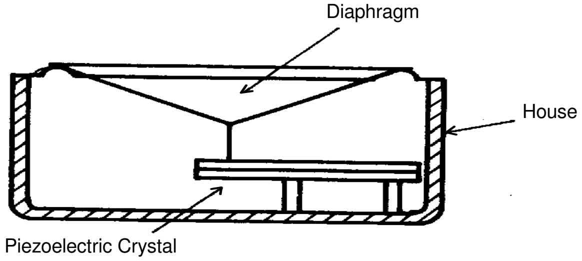


Figure 4.14. Piezoelectric Microphone

4.7. Loudspeakers

The last step of artificial sound field production is the transformation of electric energy into acoustic energy. This transformation is performed by loudspeakers. Like with microphones, different types of loudspeakers have been developed.

The *dynamic loudspeaker* (Fig. 4.15.) is the most widely used type. The moving coil of the loudspeaker is placed into the air gap of a magnetic circuit. The moving coil is attached to a conical diaphragm which is supported by an inside spider and by an outside rim to keep the motion axial. Frame attached to the magnet is supporting the rim and the leads of the moving coil end also on the frame. By the mutual action of the moving coil current and the magnetic field, an axial force is generated. This force brings the diapraghm into motion producing thus sound waves. For modest quality requirements (e.g. for AM radio receivers) one loudspeaker is satisfactory. The full audio range can be covered by two or three loudspeakers, each designed for a different frequency range.

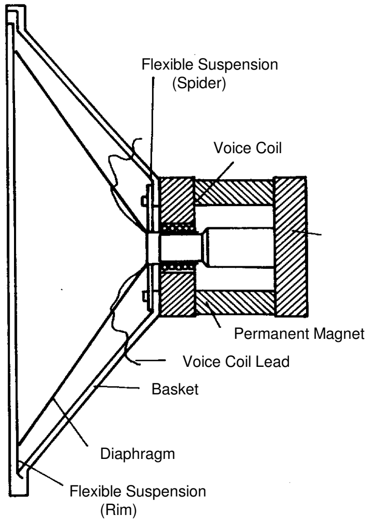


Figure 4.15. Dynamic Loudspeaker

Efficiency of sound emission can be improved by better acoustic matching. Such an improvement can be achieved by a *horn* with exponentially growing cross sectional area. The transducer is placed into the throat of the horn. The only drawback of the horn is that a very great horn is needed for a good low frequency response, so that horn is used only for sound reinforcement requiring modest sound quality (where the lack of the basses can be tolerated).

As a curiosity, the *condenser loudspeaker* shall be mentioned. To achieve good low frequency response, the diaphragm area has to be extremely great and a special high voltage power supply and matching transformer are also needed to provide proper DC bias and impedance matching. Distortion of such a device is small, frequency response is flat but the price is rather high.

4.8. Headphones

Headphones produce a sound field restricted only to the ear cavity. *Magnetic* headphone is one of the most important type as this is used as telephone receiver. A simplified sketch of such a headphone is shown on Fig. 4.16. The air gap of the fixed part is chosen great so that the flux lines close through the moving anchor. The anchor is attached to a flexible diaphragm. Magnetic field generated by the coil strengthens or weakens the static field so that the anchor vibrates around its default position. This vibration produces a sound pressure in the closed cavity of the ear. The crucial point of the construction is the size of the air gap. For consumer applications, dynamic headphones are most widely used. In fact, these are small dynamic loudspeakers and because of their size, broadband frequency response can more easily be achieved than with conventional loudspeakers.

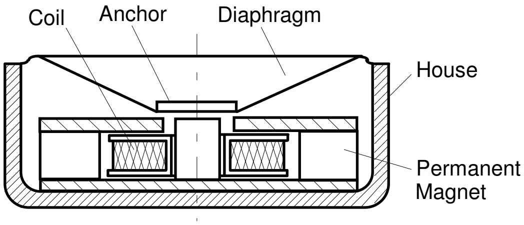


Figure 4.16. Magnetic Headphone

Control questions

1. How can sound be characterized physically?
2. What are the limits of hearing?
3. What are the psychophysical characteristic of the sound?
4. What kind of artificial sound field are reproduced?
5. What are the most important microphone types?
6. What are the most important devices of reproduction?

Exercises

1. How many dB is the sound pressure level of 1 Pa?
2. What are the loudness values of sinusoidal sounds of frequencies 110, 200 and 2000 Hz, if their intensity is 60 dB?
3. Determine the entire loudness of the previous three sounds!

References

[1] Valkó I. Péter: Az elektroakusztika alapjai. Akadémiai kiadó. 1963.
[2] Ivar Veit: Mûszaki akusztika. Mûszaki Könyvkiadó. 1977.
[3] Tarnóczy Tamás: Teremakusztika I. Akadémiai Kiadó. 1986.
[4] Ferenczy Pál: Hírközléselmélet

5. FUNDAMENTALS OF PHOTOMETRY AND COLORIMETRY. THE PICTURE.

5.1. Perceptional, Physical and Psychophysical Features

The quantities normally encountered in communication engineering can be expressed in objective terms such as power, voltage, etc., using units that are independent of the observer. The perception of light and colour are partially subjective phenomena, however, akin to the appreciation of music by the ear and the brain.

Quantities used in photometry and colorimetry can be presented in three different ways as shown in Fig. 5.1. The most upper row presents the obvious description, i.e. how light is actually perceived by the eye and then processed by the brain. Of course, this presentation is purely subjective so that the corresponding quantities cannot be numerically expressed as they are just perceptional equivalents of the terms used in photometry.

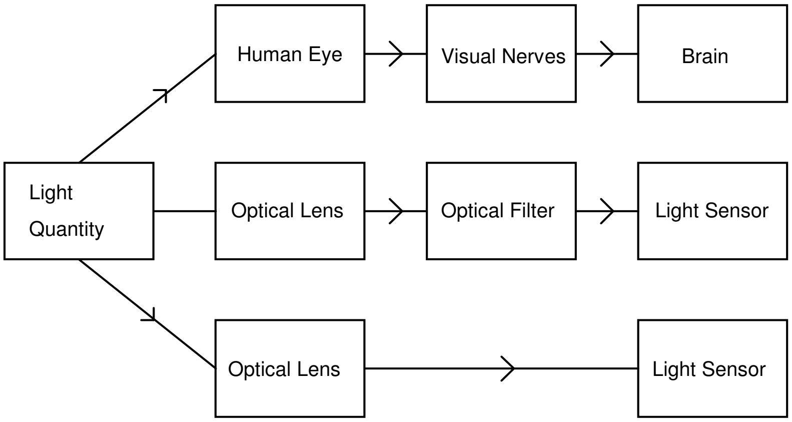


Figure 5.1 Perceptional, Psychophysical and Physical Representations of Light

In the middle row of the Fig. 1. the so called *psychophysical* representation can be seen. Here the light is passed to a sensor through a lens and an optical filter which -together with the characteristic of the sensor- satisfies the internationally accepted standard for the human organ of vision. This standard has been released by the International Lighting Committee, abbreviated as *CIE* (Commission International de l'Ènclairage).The representation is called psychophysical because on one hand it takes into account the properties and limits of the human vision but on the other hand it is an objective measure which can be expressed numerically as well.

Finally, the third of the possible representations is given in the bottom row of the figure. This is a pure physical representation where the lighting quantities are measured just like any other physical quantity without subjective limits. Obviously, this is also an objective representation.

5.2. The Human Eye and Vision

Vision is understood as the perception of visible radiation by the human eye. The organ of vision is a collective term including the eye, the visual nerves and certain areas of the brain. These parts transform the light stimulus to optical excitation resulting in the sense of vision. Picture of the outside world is optically transformed by the *retina,* located on the rear side of the eye. The retina is a light sensitive thin layer containing the terminations of the *visual nerves* (rods and cones), the nerve cells and the supporting tissue. The two different kinds of visual nerve terminations play an important role in vision. It is highly probable that the *rods* are the elements responsible for the perception of strong light and the *cones* enable the sensation under poor light conditions.

Let us see some further terms for describing vision. Lightness, hue and colourfulness are three other sensational parameters playing important roles in vision. *Lightness* is a property characterizing the amount of light that a certain surface is emitting. As it can be seen, this definition fully corresponds to that used in everyday life. *Hue* is another parameter of vision resulting in naming the colours as blue, green, yellow red, purple etc.

The third parameter, the *colourfulness* (or saturation) serves for the estimation where a perceived colour can be located between white and the pure (spectral) colour, provided this later has the same lightness and hue as the colour examined. The saturation grade is usually given by the adjectives attached to the collars, e.g. light green, pastel blue, dark red, faint yellow, etc. All these three sensational parameters have their objective (psychophysical) equivalents which will be discussed in the following chapters.

5.3. Thermal Radiation. Photometry.

The fundamental input to any television system is radiant energy in the form of light. *Photometry*, the measurement of light, is therefore of great importance to television engineers.

As a standard of light a well defined source is needed, parameters of which can be exactly reproduced at any time. As it is known from physics, a perfectly black body (i.e. one that absorbs all radiant energy falling upon it) is suitable for this purpose. When a black body is heated, the power radiated from a given area of its surface has a magnitude and spectral distribution determined solely by its temperature. The energy distribution of such a radiator is expressed by the Planck's law, which gives the spectral concentration of radiant exitance *M* as follows:

$$M_{e\lambda} = C_1 \lambda^{-5} (e^{C_2/\lambda T} - 1)^{-1} \qquad (5.1)$$

where C_1 and C_2 are radiation constants, λ is the wavelength and *T* is the temperature.

To have also a well defined photometer, an ’artificial eye’ has been constructed which simulates the light sensitivity of the human eye. The relative response of the normal human eye to monochromatic light at the different spectral frequencies has been determined experimentally by the CIE and standardized in 1924. This is known as the *photophic spectral luminous efficiency function* and is illustrated in Fig. 5.2. The symbol of this function is $V(\lambda)$ and it is usually expressed as a function of the wavelength of light (in air).

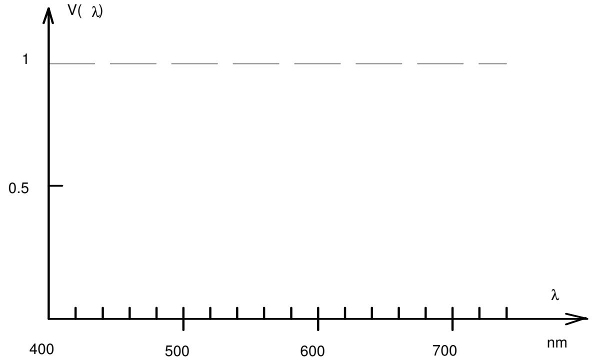


Figure 5.2 The Photophic Spectral Luminous Efficiency Function

To determine the photophic luminous efficiency function, the following procedure was carried out: First, light of constant intensity was emitted and its frequency was varied until the lightness perceived by the observer was found to be maximum. This occurred at a frequency about 5.410^{14}€Hz, corresponding to wavelength λ_m€=€555€nm. Then the wavelength was set to another λ and the power was adjusted again until the lightness was judged to be the same as at λ_m. $V(\lambda)$ could be computed then as the ratio of the radiated power at λ_m and λ, respectively. Of course, this experiment has been carried out by many observers and the resulting average was used to define the so called *CIE standard eye* which is an optical sensor with sensitivity corresponding to the function $V(\lambda)$.

The photophic luminous efficiency function serves as a link between the subjective response of the human eye and normal physical measurement techniques. It thus provide the basis for a group a photometric units. As one of the most important of them, *luminance* is defined as the measure of luminous flux (radiated power) per unit solid angle and per unit projected area.

5.4. Colorimetry

Colorimetry is based on the fact that observers can match colours with additive mixtures of three reference stimuli in amounts known as *tristimulus values*. Using reference stimuli at specified wavelengths, CIE has defined a standard set of tristimulus values to match each different wavelength of the spectrum.

Let us see, how these values have been determined. The sketch of the experimental set up is shown in Fig. 5.3. There is a white, non mirroring wedge illuminated by an unknown source of light from the right and by known sources from the left side. In front of the wedge, there is an observer seeing both sides of the wedge. The observer acts as the 'sensor' of this subjective colorimeter. It is up to him to adjust the intensity of the three known sources to achieve matching between the two sides. After that, the unknown colour can be objectively characterized by the readout of the radiation strength of the three known sources.

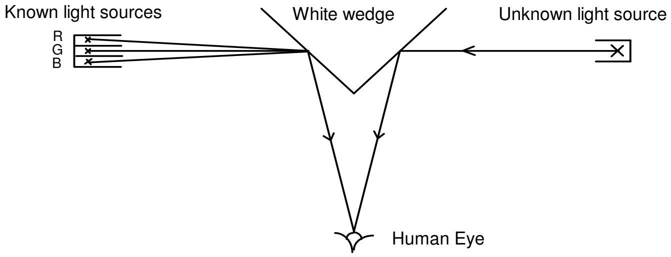


Figure 5.3. The Comparative Colorimeter

Experience has shown, however, that the spectral distribution of two compared colours might differ even when they match. What is more, two collars can perfectly match even if their radiation have different spectral distributions. This fact is of essential importance in typography as this is the only way to reproduce colourful pictures by using just three conveniently chosen colours. Different colour stimuli perceived as being the same are called *izochromatic*.

To be able to repeat the above procedure, it was necessary to standardize various elements of the system. The CIE has therefore defined a standard set of reference colour stimuli, and a standard set of tristimulus values for them; these data constitute the CIE 1931 standard colorimetric observer. The reference-colour stimuli are radiations of wavelength 700 nm for the red stimulus (R), 546.1 nm for the green stimulus (G) and 435.8 nm for the blue stimulus (B). The tristimulus values were chosen to match the typical white colour. There is a great imbalance in the three amounts (the amount of green being the greatest, and the amount of blue being much smaller). As white is a colour that is not biased towards either red, green or blue, new relative units of R and B were chosen so that the amounts be equal to the amount of green.

Series of measurements have been carried out with the standard colorimetric observer to find the different tristimulus values for different colours. To make use of the resulting huge data file, CIE has worked up a specific 'map' of colours. As three stimuli are assigned to each colour, a three-dimensional co-ordinate system would have been needed to plot the actual co-ordinates. To simplify this representation (at the expense of loosing the lightness information), co-ordinate transformation and some other calculations have been done resulting in a two-dimensional chart called *chromacity diagram*. In spite of this, the suitability of the diagram for all colorimetric measurements without the need of the related mathematical apparatus, gives chromacity diagram an outstanding importance.

The CIE chromacity diagram is shown in Fig. 5.4. The *x*, *y* co-ordinates are called *chromacity co-ordinates* and are calculated from the original tristimulus values *X*, *Y* and *Z* as

$$x = \frac{X}{X+Y+Z}, \quad y = \frac{Y}{X+Y+Z}. \tag{5.2}$$

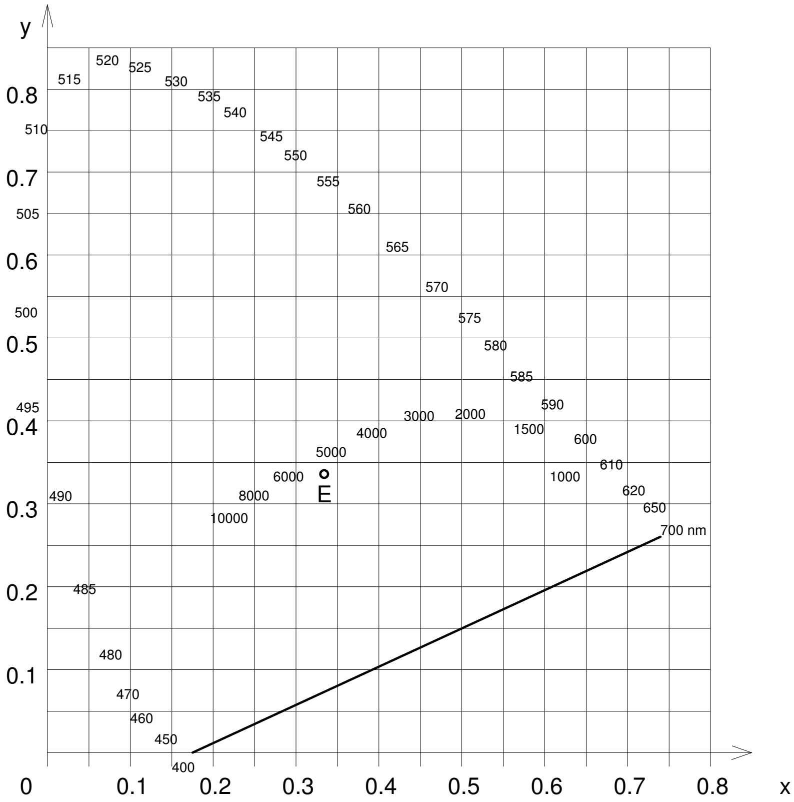


Figure 5.4. CIE Chromacity Diagram

The positions of spectral colours (the spectrum locus) are shown by the curved line and are given by the corresponding wavelengths in nm. The points representing non- spectral (pale) colours are inside the curved line. The straight line at the bottom of the chart connects the red and the blue spectral colours, so that non spectral colours mixed of red and blue (e.g. purple, violet, etc.) are located along this line. The illuminant (point representing the normal white colour) is located in the middle and is denoted as *E*. Colours radiated by the black body in the temperature range 1000-10000 K are also presented in the figure.

It follows from the previous explanation that any colour assigned to a point inside the curved line and lying on the straight line going through point *E* can be mixed by the addition of white *E* and the spectral colour given by the intersection of the straight and curved lines, respectively. Of course, if the point of intersection lies on the bottom line ('purple line') then the weighted mixture of red and blue should be taken instead of spectral colours.

The chromacity diagram only shows the proportions of tristimulus values; hence bright and dim colours having the same proportions belong to the same point. For this reason, the illuminant point also represents grey colours; and orange and brown colours, for example, tend to plot at similar positions to each other. That is why luminance has to be given as an additional information for the unambiguous definition of a certain colour.

There are some other important terms related to the previous notation which will be discussed on the basis of Fig. 5.5. Let us choose an arbitrary point denoted as C representing a colour which can be composed of the white colour E and the spectral colour C_d. Since the latter plays the dominant role in the hue, the wavelength of the spectral colour C_d is called the *dominant wavelength* and is denoted as λ_d.

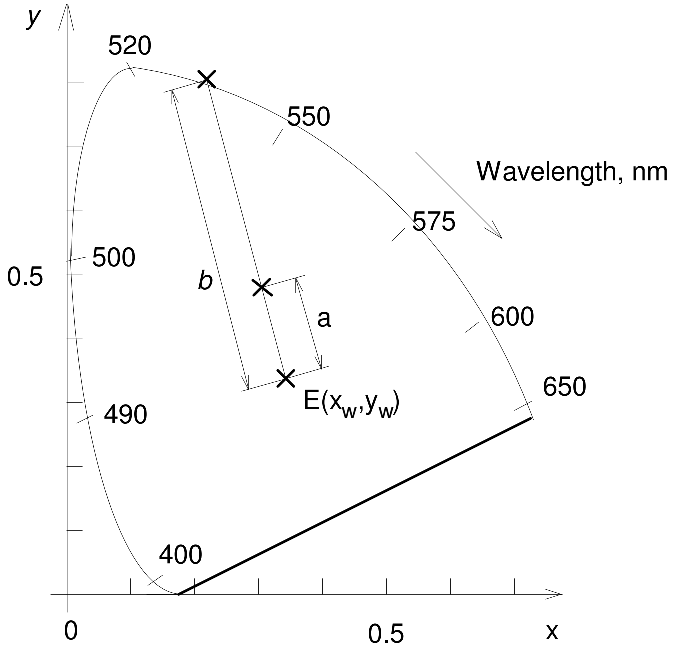


Figure 5.5. Dominant Wavelength and Excitation Purity

Since there is no difference in the hue wherever the point C is located on the line connecting the points E and C_d, the dominant wavelength can be considered the psychophysical equivalent of the hue. Similarly, the *excitation purity* p_e can be derived as the psychophysical equivalent of saturation:

$$p_e = a / b \qquad (5.3.)$$

where a and b are the distances between C and E and between C and C_d, respectively.

5.5. The Picture

To transmit a picture or a series of pictures through a communication channel, it has first to be converted to a time function. Suppose we have a monochromatic picture composed of a great number of light and dark elements. Although this picture is continuous in fact, it can be divided into small picture elements called *pixels* and characterized by their luminance, provided the elements are so small that their luminance can be taken as constant.

The procedure of converting the picture into a time function is as follows: First, the picture is divided into numbered pixels and the luminance of the individual pixels is determined and converted into the corresponding analog values. These values are then put in the same order as the pixels are numbered. This process is performed by the camera which scans the picture pixel by pixel and generates electrical output voltage proportional to the luminance of the pixels. A similar process comes on at the receiver side where the cathode ray tube converts the changes of the signal to visible changes of the luminance.

It is also important to know how great the information content of a stationary and of moving pictures is or should be. To determine these values, first we have to decide into how many pixels a picture has to be divided. Since the ’final receiver’ of all pictures is the human eye, we have to decide about its resolution. For monochromatic pictures the resolution is about 2', i.e. if the visual angle between two pixels is less then 2' the eye is not able to differentiate between them. As it was also determined experimentally, the optimum visual angle for the whole picture is about 20° . The third important factor is the aspect ratio which was set to 4:3 which is the value commonly used in photography. From all these data summarized in Fig. 5.6., the number of pixels in the vertical and horizontal directions can be computed:

$$n_V = \frac{20°}{2} = \frac{20 \cdot 60}{2} = 600,\ n_H = \frac{4}{3} n_V = 800$$

which makes $n€=€n_1€n_2€=€4.8€10^5$ pixels for the whole picture. The last step is to decide how many tones a pixel shall have. Here we refer again to experimental results indicating that a picture seems to be natural if the number of tones chosen is about 100.

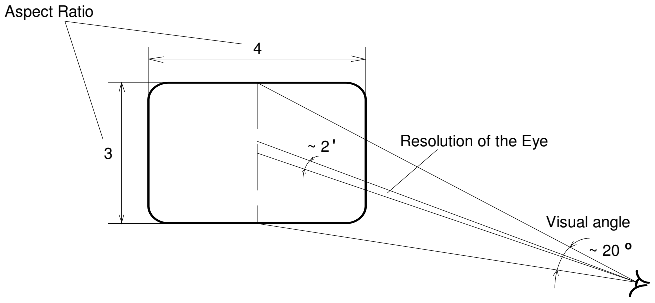


Figure 5.6 Parameters Determining the Number of Pixels

From the above data the information content of a monochromatic picture can be computed. Suppose the gradation number is *s*€=€100. The information content of one pixel is then

$$I_{\text{pixel}} = \log_2 s = \log_2 100 = 6.65 \text{ bit}$$

so that for the entire monochromatic picture

$$I_{\text{pic}}^{\text{mc}} = nI_{\text{pixel}} = 4.8 \cdot 10^5 \cdot 6.65 = 3.19 \cdot 10^6 \text{ bit} \qquad (5.4)$$

This means that to send one stationary picture which satisfies the above conditions, more than 3 million bits have to be transmitted through the channel.

5.5.1. Transmission of Moving Pictures

Similarly to the projection of films, we can make use of the inertia of the eye and transmit about 25-30 pictures in one second instead of a continuous transmission of ’all’, phases (which is impossible, anyway). The series of pictures transmitted with such a frequency will give the impression of a continuously varying picture. Of course, the individual pictures are divided into pixels in the same way as it is done when stationary picture is transmitted.

Suppose that we want to transmit 25 monochrome pictures every second. The information rate of such a number of pictures is

$$v_{\text{mpic}} = N \cdot I_{\text{pic}}^{\text{mc}} = 25 \cdot 3.19 \cdot 10^6 = 80 \text{ Mbit / s} \qquad (5.5)$$

so that to transmit monochrome TV pictures, a channel with the information rate of at least 80 Mbit/s is needed theoretically.

5.5.2. Transmission of Colour Pictures

Let us examine the information rate needed for the transmission of colour pictures. First, we have to make clear what the information surplus of a colour picture is in comparison to the monochrome one. We shall follow the procedure used for monochrome pictures, i.e. we compute first the information of one pixel and then multiply it by the pixel number to obtain the information capacity of the entire picture.

As we already know, one colour pixel is defined by three independent data. One of them is the luminance, and the other two contain the colour information. These data can be the chromacity co-ordinates used in colorimetry; red and blue colour-difference signals can also be used (detailed discussion of these terms is given in Chapter 20).

As for the gradation of monochrome pictures, it was also experienced that instead of the transmission of infinite number of colour tones it is sufficient to distinguish just 20 different colour gradations to reproduce a satisfying picture. That is, 20 different colour attributes can be assigned to each pixel.

Furthermore, we have to consider that the resolution of the human eye is poorer for colour pictures. According to experiments, the resolution partly depends on the pairing of the colours of adjacent pixels but it can be said that on the average the resolution is about one fifth of that for the monochrome picture, i.e. the minimum visual angle is about 10' compared to the 2' of the monochrome picture.

It can be concluded from the previous discussion that the size of a colour pixels may be five times greater than the size of the monochrome pixel, or in other words just one colour pixel is needed to cover the square grid of 5€x€5€=€25 monochrome pixels. So that human vision is pleased with a colour picture having the pixel structure 25 times more rough than a finely detailed monochrome picture.

At this point we can already compute the information surplus of the colour picture:

$$I_{\mathrm{pic}}^{\mathrm{c}} = 2n_1 n_2 \frac{1}{25} \log_2 20 = 166 \text{ kbit} \qquad (5.6.)$$

Let us compute now the resulting information of the colour picture:

$$I_{\mathrm{cpic}} = I_{\mathrm{pic}}^{\mathrm{mc}} + I_{\mathrm{pic}}^{\mathrm{c}} = 3.19 \cdot 10^6 + 0.166 \cdot 10^6 \text{ bit} = 3.36 \text{ Mbit}$$

Note that the colour information contributes very little to the information content of the picture.

Finally, the transfer rate for the colour picture is

$$v_{\mathrm{cpic}} = NI_{\mathrm{cpic}} = 25 \cdot 3.36 \cdot 10^6 \approx 84 \text{ Mbit / s}$$

which is the minimum channel capacity for the TV colour picture transmission. Practically, much greater capacity is needed, since neither the channel nor the time can be used up one hundred percent.

Control Questions:

1. How can the lighting quantities be classified? Name the features of these classes!
2. How has the photopic luminous efficiency function *V*(λ) been determined?
3. What are the definitions of the CIE psychophysical quantities? Interpret the meaning of these quantities!
4. What is the definition of a monochrome and a colour pixel?
5. How can be computed the information of a colour and a monochrome picture?

References

[1] Hunt R.W.G.-Darby P.J.: Light and Colour Principles. IBA Technical Review 22.

[2] Ferenczy P.: Video- és hangrendszerek. Mûszaki könyvkiadó, 1986.

6. COMMUNICATION CHANNELS

The goal of any kind of communication is to send or to copy an information to another place (or often to more places) with the help of appropriate devices and equipment. The information sender is often called the *source* while the receiver is called the *sink*. The path between the source and the sink is the communication channel.

If there are only two partners taking part in the communication, we speak about point-to-point communication. There are also multipoint systems with one source sending information to more sinks. In *simplex* systems, the information can be passed in one direction only. The system is said to be *half-duplex* if the communication is possible in both directions, but not at the same time. Finally, in *duplex* systems the communication may happen in both directions without any time restrictions. In this chapter the characteristics of *simplex*, point-to-point communication will be discussed and considered as the communication channel.

The simplest communication channels consist of a transfer medium and of transducers interfacing the information to this medium. The characteristics of such channels are influenced mainly by the transfer medium. In this sense wirebound and wireless channels are distinguished. From the user's point of view, however, it is more important how he can be connected to the channel, what expected quality of the channel can be and what kind of information can be sent.

A channel is said to be analog if analog signals are transmitted and received at its input and output. On the contrary, a digital channel transmits and receives digital signals or series of symbols between the input and the output points. These input and output points are called the *interfaces*.

Transmission between the source and the sink is often performed by cascaded channel sections. It may also happen that by means of convenient transducers, a digital channel is built upon an analog channel or, on the contrary, an analog channel is formed from a digital channel (see Fig. 6.1.).

When a channel is to be characterized, first of all, the essential differences between analog and digital channels shall be taken into account. In the following, the channel properties are discussed from this point of view.

6.1. Analog Channels

As we have seen in Fig. 6.1., an analog channel is a section of the communication channel receiving analog signal at the input interface and reproducing analog signal at the output interface. Such a channel can be characterized by the specification of the signals, for which the channel can provide a satisfying operation. More detailed characterization can be given if the effects produced by the channel are defined with the help of simple models. Sometimes it is difficult to characterize a parameter of the channel in the desired depth. In this case the channel is said to be uncertain or unspecified from the point of view of the given parameter. However, this uncertainty does not affect definitely the quality of transmission, it only sets some restrictions to the transmitted signals (for instance, telephone channels are not specified in the frequency range from 0 to 300 Hz but such a lack of specification is irrelevant since the transmitted signals do not contain such spectral components). There is another restriction of the input signal if some values of a certain signal parameter disturb the operation of the channel or have influence on other channels.

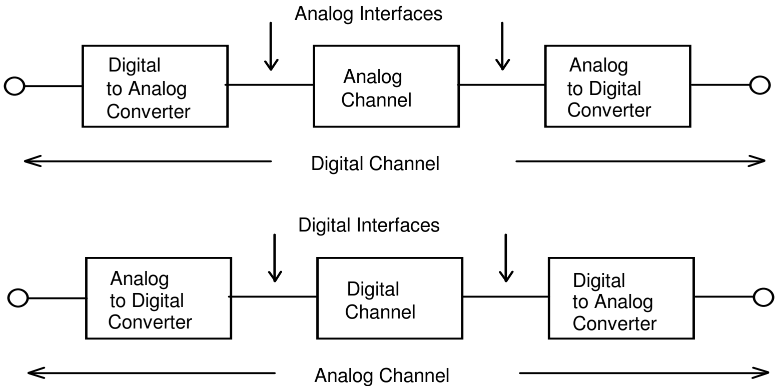

Analog Interfaces
Digital to Analog Converter
Analog Channel
Analog to Digital Converter
Digital Channel
Digital Interfaces
Analog to Digital Converter
Digital Channel
Digital to Analog Converter
Analog Channel

Fig. 6.1 Analog and Digital Channels Built Upon Each Other

Three effects are considered in commonly used channel models:

- linear distortion (which can be either time-invariant or time-variant),
- nonlinear distortion (which may be memoryless or looped),
- noises (noise is meant as an effect independent of the input signal).

6.1.1. Time-Invariant Linear Distortion

This is a kind of distortion typically caused by the attenuation and time delay the signal suffers when passing through the transmission medium and the interfacing devices. This distortion is present almost always and generally it is not too harmful as the time delay (if it is short) usually does not cause any problem and the channel attenuation can be compensated by appropriate amplification. The distortion of the channel is generally frequency-dependent; this dependence can be described by the channel frequency response $H_c(f)$, $f \in (-\infty,\infty)$. *Attenuation* and *phase* are the quantities derived from the frequency response and used for the practical characterization of the channel:

$$a(f) = -20\lg\left|H_c(f)\right| \qquad \text{and} \qquad \phi(f) = -\operatorname{arc} H_c(f) \quad (6.1)$$

It is easy to see that if the signal is attenuated by a_0 and suffers a delay T, the distortion can be modelled by the channel with

$$a(f) = a_0 \qquad \text{and} \qquad \phi(f) = 2\pi fT \qquad (6.2)$$

For those frequencies where the signal does not have any spectral components, the behaviour of the channel is indifferent.

It is an often case that different components of the signal have different attenuation and delay. This phenomenon is called *dispersion* and is described in detail by the frequency response H_c(.). For the superficial characterization of the dispersion, test impulses shown in Fig. 6.2. are used.

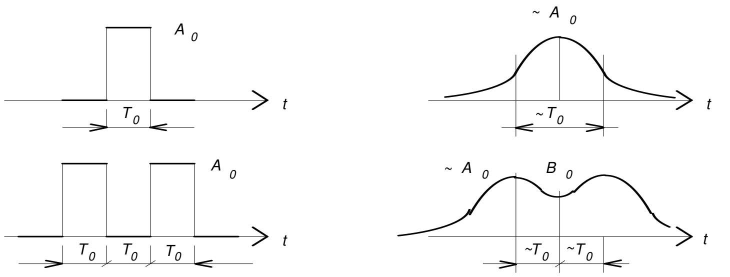


Fig. 2. Pulses Distorted by a Linear System

Sometimes the channel is used only in a narrow region near a frequency f_o where

$$\left|H_c(f)\right| = A_0 \quad \text{and } \phi(f) = \phi(f_0) + 2\pi(f - f_0)\tau_0 \qquad (6.3)$$

This is the case when the channel input signal is

$$x(t) = m(t)e^{j2\pi f_0 t} \qquad (6.4)$$

where *m*(.) is a slowly changing narrowband signal. Let *M* be the Fourier transform of *m* so that in the narrow band around f_o

$$X(f) = M(f - f_0)$$

and the signal at the channel output is then

$$Y(f) = X(f)H_c(f) = M(f - f_0)H_c(f),$$

i.e.

$$Y(f) = A_0 e^{-j\pi\phi(f_0)} M(f - f_0) e^{-j2\pi(f - f_0)\tau_0}$$

The inverse transform is then

$$y(t) = A_0\, m(t - \tau_0) e^{j(2\pi f_0 t - \phi(f_0))} \qquad (6.5)$$

Besides the amplification of the signal by A_o and the shift of its harmonic factor by $\Phi(f_o)$, it is important to notice that the envelope of the harmonic signal $m(.)$ remained essentially undistorted but suffered a delay τ_o. It follows from eq. (6.3) that

$$\tau_0 = \frac{1}{2\pi}\frac{\partial\phi(f)}{\partial f}\bigg|_{f-f_0} = \tau(f)\big|_{f=f_0} \qquad (6.6)$$

The function $\tau(f)$, $f\in(-\infty,\infty)$ which is the derivative of the phase characteristic is called *envelope delay* (or group delay). Generally, the envelope delay is a more illustrative term than the phase. If the envelope delay of the channel is frequency-dependent, the wideband signals may be significantly distorted. It is worth to remark that envelope delay of channels of a bandpass character has remarkable ripples at the edges of the passband.

6.1.2. Echo and Reverberation

Echo and reverberation are special kinds of linear time-invariant distortions which occur when the output signal is composed of several components of the input signal which have different delays and attenuations:

$$y(t) = \sum_i c_i x(t - T_i) \qquad (6.7)$$

This effect is often due to multipath propagation or caused by reflections from mismatched terminations. In the simplest case eq. (6.7) has only two members and usually a_o>>a_1. Suppose that a_o=1 so that

$$y(t) = x(t) + c_1 x(t - T_1) \qquad (6.8)$$

In the frequency domain, this distortion corresponds to the following frequency response:

$$A(f) = 1 + c_1 e^{-j2\pi f T_1} \qquad (6.9)$$

which is periodical in $1/T_1$ so that the spectrum is periodically deformed by ripples having amplitude a_1. In the case of speech signals, if T_1>>50 ms, the delayed sound is perceived by the ear separately as an echo. If the delay is smaller, the sound is perceived as being one but of a particularly hollow sounding. In the case of video signals, the echo blurs the picture contours or produces a ghost picture.

The other usual form of reverberation is when the output signal contains components generated by multiple reflections:

$$y(t)€=€\sum_i c^{i}\cdot x(t\text{-}iT), \qquad |c| < 1 \qquad (6.10)$$

Speech or music is echoing in such a case.

6.1.3. Time-Variant Linear Distortion

It is also a usual condition that the transfer function of a channel cannot be supposed to be constant even for a short period of time. The simplest form of this case is when the gain (or attenuation) of the channel fluctuates:

$$y(t) = A(t)\ x(t) \qquad (6.11)$$

Even such a simple model enables us to set up and answer several interesting questions. Important parameters of such a type of interference having a multiplicative character are the rate and the extent of the changes of $A(.)$. If these are slow compared to the changing of x and the amplitude varies just some few dB-s then this interference can be compensated relatively easily by automatic gain control (AGC). The real problem is caused by great ($\geq$10 dB) and fast changes of A which is typical for wireless communication.

A special type of the time-variant distortion is the so called *phase jitter* and *frequency shift*. This may happen typically when the output signal is a very particular transform of the input signal:

$$y(t) = x(t)\ \cos(\mu_t) - z(t)\ \sin(\mu_t) \qquad (6.12)$$

where $z(.)$ stands for the Hilbert transform of $x(.)$. (Hilbert transform is a linear distortion shifting the signal phase by $\pi/2$ rad at all frequencies.) Transformation given in (6.12) is time-variant because of the time dependence of μ_t. The effect can be well illustrated when x is a sinewave having frequency f_o. In this case

$$x(t) = \sin(2\pi f_0 t) \qquad \text{and} \qquad z(t) = -\cos(2\pi f_0 t)$$

so that

$$y(t) = \sin(2\pi f_0 t + \mu_t) \qquad (6.13)$$

Several μ functions, having sometimes quite an unusual character might come about in real applications. If μ is a stationary process in the usual sense of the word then this effect is called *phase jitter*. Other important case is when μ varies linearly in time:

$$\mu_t = \mu_0 + \Delta t \qquad (6.14)$$

In this case

$$y(t) = \sin\big(2\pi(f_0 + \Delta)t + \mu_0\big) \qquad (6.15)$$

Note that all sinusoidal components are shifted by the same frequency Δ. This shift essentially changes the signal shape, e.g. the transmitted data are so distorted that they even become unrecognizable. In musical signals, this leads to a particularly unpleasant sounding since the harmonic content characterizing musical sounds is significantly distorted even if Δ is small. In the speech signals the frequency shift is more acceptable since frequency offset of some few Hz does not degrade significantly the speech intelligibility.

6.1.4. Nonlinear Distortion

Modelling of the systems by linear transformations is simple but sometimes imperfect. In more precise models nonlinear effects should be taken also into consideration. The simplest nonlinear models are memoryless, i.e. the output signal at an arbitrary time *t* depends only on the input signal value in the same time:

$$y(t) = n(x(t)) \tag{6.16}$$

where *n*(.) is a single variable function, usually continuous. *Saturation* and *dead zone* effect (see Fig. 6.3.) can be presented as typical examples of memoryless nonlinearities.

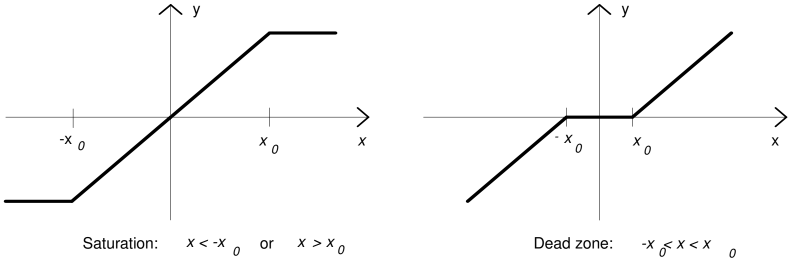


Fig. 6.3 Nonlinearity Caused by Saturation and Dead Zone

Function *n*(.) describing the nonlinear behaviour can often be decomposed into Taylor series, more precisely it can be substituted by the first few members of the Taylor series:

$$n(x) = b_0 + b_1 x + b_2 x^{2+} + b_3 x^3 \tag{6.17}$$

To determine the effect of such a nonlinearity, let us assume that the input signal is sinusoidal:

$$x(t) = U \cos(2\pi f_0 t)$$

The output signal will be then

$$y(t) = b_0 + b_1 U \cos(2\pi f_0 t) + b_2 U^2 \cos^2(2\pi f_0 t) + b_3 U^3 \cos^3(2\pi f_0 t)$$

or, using trigonometric identities:

$$y(t) = b_0 + \frac{1}{2} b_2 U^2 + (b_1 U + \frac{3}{4} b_3 U^3) \cos(2\pi f_0 t) +$$

$$+ \frac{3}{4} b_2 U^2 \cos(2\pi 2 f_0 t) + \frac{1}{4} b_3 U^3 \cos(2\pi 3 f_0 t)$$

This simple example may serve for making some more general conclusions. Namely, it can be stated that because of nonlinear distortion, new sinusoidal components are generated in the output signal which were not present in the input signal (in the example above, $2f_o$ and $3f_o$ are such a components).

The amplitude of the fundamental harmonic is a nonlinear function of the input signal amplitude and the amplitude of the harmonics are power functions of the input amplitude. These simple consequences enable us to characterize the nonlinearity by means of the 2nd, 3rd, etc. *harmonic distortion factor*, defined as the ratio of the corresponding harmonic amplitude to that of the fundamental component.

When examining the nonlinear behaviour of amplifiers, it is a common experience that increasing the input signal, the power of the components causing distortion starts to increase dramatically at a certain input level, thus indicating the start of the saturation. This effect can be used to define more precisely the overloading level of systems with higher complexity.

6.1.5. Additive Noise

Generally, different interferences (crosstalk, thermal noise, man-made noise, etc.) influencing the output signal of the transmission systems have to be considered as being looped and non-linear. However, it is also usual that these effects can be collected as one common factor ν_t which is independent of the signal itself and can be simply added to it:

$$y(t) = x(t) + \nu_t \qquad (6.18)$$

Such a type of noise is called *additive* and can obviously be modelled by a stationary stochastic process. Generally, it is not possible to characterize the process by its distributions.

Sometimes ν is composed of several independent noise sources of approximately the same magnitude. In this case ν can be well approximated by a Gaussian process and the primary parameters of the process can be determined by secondary parameters (expected value, autocorrelation function). As a typical model, the process with zero-mean and constant spectral density over a wide range is used.

Signal-to-noise ratio which is the ratio of the powers of x and ν, is usually a good parameter to characterize the influence of the additive noise from the point of view of the sink:

$$\frac{S}{N} = \frac{P_x}{M(\nu_t^2)} \qquad (6.19)$$

The value of the signal-to-noise ratio is usually given in the logarithmic scale in decibels as

$$SNR = 10\lg\left(\frac{S}{N}\right), \quad \text{dB} \qquad (6.20)$$

To characterize the signal and noise intensity, it is also convenient to introduce their power in logarithmic scale. The *absolute power level* of the signal S is defined as

$$s_{\text{signal}} = 10\lg\left(\frac{S}{S_0}\right) \qquad (6.21)$$

where S_o is a reference power (usually 1 mW). Signal-to-noise ratio is then given as the difference of the signal and noise power level in dB:

$$SNR = s_{\text{signal}} - s_{\text{noise}}, \quad \text{dB} \qquad (6.22)$$

6.2. Digital Channels

6.2.1. Memoryless Channels

Digital channels are understood as being systems accepting N_{in} kinds of input symbols (a_1,a_2,...) and capable to produce N_{out} kinds of output symbols (b_1,b_2,...). Such a channel can also be characterized by the rhythm the symbols are received and generated. This parameter is called the *symbol rate* (v_s). As N symbols can generate bit series in length ld(N), the so-called *data transfer rate* is

$$\nu_{\text{data}} = \nu_{\text{signal}} \, \text{ld}(N_{\text{in}}) \qquad (6.23.)$$

Generally, it is not warranted that the channel output symbols characterize unambiguously the input symbols. However it is often true that the actual output symbol η is determined solely by the actual input symbol ξ and the instantaneous "caprice" of the channel.

A channel is said to be *memoryless*, if its output is independent of previous symbols and of the response to those symbols. The behaviour of memoryless channel is characterized by conditional probabilities or the so called *system of transitional probabilities*:

$$p_{ij} = P(\eta = b_i | \xi = a_j), \quad i = 1,2,...,N_{\text{out}} \;\; j = 1,2,...,N_{\text{in}} \qquad (6.24)$$

A typical example of a memoryless channel is the binary symmetric channel (BSC), symbols of which can have two values, e.g. 0 and 1. Transitional probability is characterized by a single date, p:

$$p_{01} = p_{10} = p \qquad p_{11} = p_{00} = 1 - p$$

For channels which have the same set of input and output symbols, the transmission errors can be well evaluated by the so-called *probability of error*. This term, denoted as P_{e} means the probability of the event that the channel output signal is not equal to the input symbol:

$$P_e = P(\eta \neq \xi) \qquad (6.25)$$

Obviously, the probability of error may depend on the probabilities with which the source generates the individual input symbols. If p_{i} is the probability of sending the ith symbol then

$$P_e = \sum_{i=1}^{N} p_i \, P(\eta \neq a_i | \xi = a_i) = \sum_{i=1}^{N} p_i \, (1 - p_{ii}) \qquad (6.26)$$

It is interesting that for BSC the probability of error is independent of the source distribution: P_{e}€=€p.

There are several practical cases when the probability of error gives sufficient information about the usability of the channel. This is the case when the probability of error is small and the channel is used e.g. for transmission of coded speech. Occasional channel errors cause additional noise in the reconstructed speech signal but this may be tolerable and does not necessarily degrade the quality of the provided service.

The situation is quite different when the same channel is used e.g. for copying a computer program. If there is just one faulty bit in the copied program code, the program might become completely useless. In such a case it is necessary to recognize the errors caused by the channel and to correct the faulty symbols.

6.2.2. Principles of Error-Detection

Let us divide the source symbols into consecutive blocks each of *K* symbols and assign a supplement of *N*-*K* symbols to each block according to some appropriate rule. The blocks of *N* symbols are transmitted through the channel and checked in the receiver whether the relations between the first *K* symbols and the remaining *N*-*K* symbols match the defined rule. If the rule used for the supplement generation is well chosen, not only can we detect the errors but also deduce which symbols are defective. Of course, it is not indifferent how much the original message has to be lengthened to detect and correct the faulty symbols, since the transfer rate is decreased by the factor of *K*/*N*. The efficiency of error-correction is limited by the transition probabilities of the channel. This problem belongs to the information theory and will be discussed in Chapter 7.

6.2.3. Capacity of Binary Symmetric Channels

Let us determine the maximum efficiency of error-free transmission through a binary symmetric channel whose probability of error is *p*. Suppose that the examined model is optimal, i.e. it consists of the above channel and an ideal backward channel informing about the received message (e.g. we are working on a**n** unreliable keyboard but we can see the output on the display).

As a first step, let us send a message of n_o bits so that $p{\cdot}n_o$ bits will be damaged. The error-correcting message thus should be a series of n_o bits, containing 0s with the probability 1-*p* (indicating that these bits were received correctly) and 1s with the probability *p* (indicating that the bit is faulty, the corresponding message bit has to be inverted).

So that the error-correcting message can be considered as a message having the distribution of symbols as follows:

$$P:\quad p_0 = 1 - p \qquad \text{and} \qquad p_1 = p\,.$$

Let us use source coding (see Chapter 3.2.) for the error correcting message. Using the optimum source coding, the length of the error-correcting message will be

$$n_1 = n_0\, H(P)$$

Of course, this message might also contain faulty bits in some positions but we can send similar error correcting messages until all errors are corrected.

For the error-free transmission of all *n* bits a total of

$$\sum_{i=0}^{\infty} n_i = n_0 + n_0 H^2(P) + \ldots = n_0 \frac{1}{1 - H(P)} \tag{6.27}$$

bits will be needed. The efficiency of the coding is

$$C = \frac{n_0}{\sum_{i=0}^{\infty} n_i} = 1 - H(P) \tag{6.28}$$

Obviously, efficiency will further be reduced if the error correcting messages are to be "built into" the original message in advance. However it can be proved that if *K*/*N*€<€*C* always such a code exists for which the probability of erroneous evaluation of blocks goes to zero if N → ∞.

The above example can be generalized for even more complex channels. With certain codes the probability of error can be made arbitrarily small provided the coding efficiency *K*/*N* is smaller than the capacity, a limit determined by the channel.

Control Questions

1. What is the difference between distortion and noise?
2. What are the conditions for a signal not to be distorted by a linear (time-invariant) distortion?
3 When is it reasonable to characterize the transmission quality by the signal-to-noise ratio?
4. What are binary symmetric channels?
5. What is channel capacity and in what sense does it limit transmission efficiency

Exercises

1. A lowpass filter having bandwidth of *B*[Hz] is tested by the double impulse given in Fig. 6.2. Estimate the value of *D* if *B*€=€1/*T*.
2. Determine the capacity of a binary erasure channel. The channel has three possible outputs: 0, 1 (representing the input symbols) and *x* which stands for a non readable symbol.

References

[1] Csibi S. et al: Információ közlése és feldolgozása. Tankönyvkiadó, 1986.

[2] Gyõrfi L.-Vajda I.: A hibajavító kódolás és a nyilvános kulcsú titkosítás elemei. TUB lecture notes, 1991.

[3] Gallager, R. G.: Information Theory and Reliable Communication. Wiley, New York, 1968.

7. ERROR CONTROL CODING

Let us consider a digital communication channel transmitting binary series (0s and 1s) entering the channel input. The channel consists of a modulator, the physical transmission medium and a demodulator. The modulator converts input 0s and 1s to pairs of signals suitable for transmission through the medium. During transmission, these signals are distorted and disturbed by noise. The infinite set of the received signals is then converted back to 0s and 1s by the demodulator using a decision rule. These decisions, however, are not free of errors. The probability of error would certainly be reduced if the transmitted power or the duration of the signals were increased. These methods are not used because neither poor efficiency nor low transfer rate is desirable. Fortunately, there is a procedure called *error control coding* to keep the probability of transmission errors at an acceptably low level.

There are two tasks in the error control coding: error *detection* and error *correction*. When error detection is used, receiver informs the transmitter on a backward channel that an error has been found and requests the transmitter to re-send the signal. In the case of error correction, the receiver is able to correct certain errors. Hybrid coding procedures are also used where the receiver first tries to correct the error and then checks the result by error detection.

7.1. Basic Terms of Coding

The basic communication structure is shown in Fig 7.1.

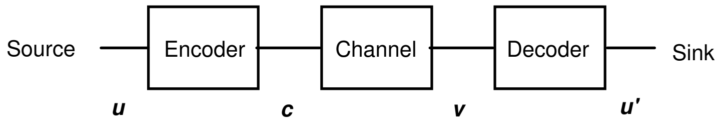


Figure. 7.1. Error Correction in the Communication Channel

The source transmits a k bit long binary *message* $\boldsymbol{u}$€=€$(u_1,u_2,...,u_k)$ through the communication channel towards the sink. The message is converted by the encoder to an n bit long binary *code word* $\boldsymbol{c} = (c_1,c_2,...,c_n)$. The *word received* at the channel output is $\boldsymbol{v} = (v_1,v_2,...,v_n)$ and is of the same length as c.

An error caused by the channel is said to occur at the m^{th} position if c_m€≠€v_m. Let t be the *number of all errors* occurring during the transfer of $\boldsymbol{c}$. Generally $d(\boldsymbol{c},\boldsymbol{v})$, the *Hamming distance* of arbitrary words c and v¯is defined as the number of those positions where these two words differ, i.e. $t = d(\boldsymbol{c},\boldsymbol{v})$.

Code is defined as a set C of 2^k binary vectors of length n, where k is the length of the binary message. This code is usually denoted as $C(n,k)$, where n and k are called the parameters of the code. The elements of the code are called *code words. Coding* is a reversible process which transforms messages into code words, i.e. different messages are transformed into different code words.

Decoding is performed in two consecutive steps: First, on the basis of a *decision rule*, the received word $\underline{v}$ is transformed into a decoded code word *c*' then according to the inverse of encoding, a *u*' message is assigned to the decoded code word $\underline{c}$. The most frequently used decision rule chooses a code word *c*' which has the shortest Hamming distance to the received word *v*, i.e.

$$\mathrm{d}(\mathbf{c'},\mathbf{v}) = \min \mathrm{d}(\mathbf{c},\mathbf{v}) \qquad \mathbf{c} \in \mathrm{C}$$

As it turns out from the previous discussion there are two main tasks of error correction coding. First, a code has to be set up which creates code words with Hamming distances as great as possible. After having such a code, a decision rule has to be constructed which finds the code word being of minimum distance from the received code without the need to look up all the code words. However if the code is short such a thorough examination is possible yet. If the code length, i.e. the parameter *n* is relatively small, (e.g. $n \leq 10$) then the so called *table lookup decoding* can be used.

Example 7.1.: Let us choose parameters *k*=2, *n*=5 and consider the following code *C*:

u	***c***
00	00000
01	01101
10	10110
11	11011

In this case, the first ten lines of the 32-row decoding table are as follows

v	***c'***	***u'***
00000	00000	00
10000	00000	00
01000	00000	00
11000	00000	00
00100	00000	00
10100	10110	10
01100	01101	01
11100	01101	01
00010	00000	00
10010	10110	10

Because the size of the memory is limited, table lookup decoding cannot be used generally. If the value of *k* is small (*k*€<<€*n*) then there are only few code words even in the case of long code lengths. This enables to decode the code word by computing the distances code word by code word.

For the typical cases however, neither of the two previous methods can be used. Just imagine the case when the code length *k* is 50 bit and there are $2^{50} \approx 10^{15}$ possible code words! To present a solution for such a case, further terms have to be introduced.

The minimum Hamming distance between the code words of a code is a very important parameter. It is called *code distance* and is denoted as $d_{\min}$. Thus, formally $d_{\min}$= min d(c,c'), $c \neq c'$, $c,c' \in C$. It is easy to check that the code distance in the example 7.1. is $d_{\min}$=€3.

The aim of *error detection* is to decide whether the received word is a code word or not. If the number of errors within one received code word is not more than t and $d_{\min}>t$ then it is certain that no combination of errors results in another valid code word. This is very important because it would not be possible to detect errors on the receiving side if the received code was a false code word.

Theorem 7.1.: Code C with the code distance $d_{\min}$ is able to detect maximum $d_{\min}$-1 errors.

A code is said to have error detection (or error correction) capability t if it is able to detect (or to correct) not less than t errors and there is at least one received word with t+1 errors where the errors cannot be detected. For instance, error detection capability of the code given in example 7.1. is 2, in accordance with the above theorem.

In the case of error correction, the question is what should be the condition for an unambiguous restoration of the transmitted code word c from the received word v. The formal condition is that for any other code word $\boldsymbol{c'}$

$$d(\mathbf{v}, \mathbf{c'}) > d(\mathbf{v}, \mathbf{c}) \qquad (7.1.)$$

i.e. $\underline{\mathbf{c}}$ has to be the closest to the received word. Since the Hamming distance is really a distance (non negative, symmetrical and fulfilling the triangle-inequalities) wherefore

$$d(\mathbf{v}, \mathbf{c'}) \geq d(\mathbf{c}, \mathbf{c'}) - d(\mathbf{v}, \mathbf{c}) \qquad (7.2.)$$

Thus the eq. (7.1). can be satisfied if the right side of the eq. (7.2.) is greater than $d(v,c)$, i.e. if

$$d(\boldsymbol{c},\boldsymbol{c'}) - d(\boldsymbol{v},\boldsymbol{c}) > d(\boldsymbol{v},\boldsymbol{c}) \qquad (7.3.)$$

which leads to $d(\boldsymbol{c},\boldsymbol{c'})$€>€2·$d(\boldsymbol{v},c)$. This condition is certainly satisfied (taking into account¯also that $d(\boldsymbol{c},\boldsymbol{c'})$€≥€$d_{\min}$, $\boldsymbol{c} \neq \boldsymbol{c'}$) if $d_{\min}/2 > d(\boldsymbol{v},\boldsymbol{c})$ is true.

Theorem 7.2.: Error correction ability of a code C with the code distance $d_{\min}$ is int[($d_{\min}$-1)/2].

In example 7.1., error correction ability is 1, i.e. the code is able to correct one error at any position of the faulty code word. It follows from theorem 7.2. that the code distance of a code being able to correct at most t errors is $d_{\min} \geq 2t$€+€1.

Now the following question can be put: How can a code be constructed with a sufficiently great code distance? To be able to answer this question, let us survey the terms of linear codes.

7.2. Linear Codes

The idea of linear coding can be demonstrated by a simple example.

Example 7.2.: Let us take the code used in example 7.1. It can be seen that the coding can be done by matrix multiplication $\boldsymbol{c} = \boldsymbol{u}\cdot\boldsymbol{G}$ where

$$\mathbf{G} = \begin{pmatrix} 10110 \\ 01101 \end{pmatrix}$$

G is a binary matrix of size 2x5. In this simple case the set of codes consists of the all-zero vector, the first and the second row the **G** matrix, and co-ordinate-by-co-ordinate modulo-2 (XOR) sum of the two rows. Thus the elements of the *C* code are generated as linear combinations of the rows of the matrix **G**. The code *C*, as a set of binary vectors, forms a linear space. Generalizing this remark we can understand the term of the linear code.

A binary code *C* is called *linear* if the set *C* is a *linear space*, i.e. if for all $\boldsymbol{c},\boldsymbol{c}'\in C$, $\boldsymbol{c} + \boldsymbol{c}'\in C$ is also true. Thus the all-zero code word (**0**) is also an element of the linear code, because $c + c = 0$ is true for arbitrary binary word **c**. Linear codes are significant since their code words are generated relatively simply and error detection and correction is also simpler than for nonlinear codes.

Terms commonly used for the space of real vectors remain valid in the space of binary vectors. Suppose that vectors $\mathbf{g}_1,\mathbf{g}_2,..,\mathbf{g}_k$ form a base of the linear space *C*, i.e. that with these vectors an arbitrary $\mathbf{c} \in C$ element can be generated as

$$\boldsymbol{c} = \sum_{i=1}^{k} u_i\boldsymbol{g}_i \qquad \mathrm{i} = 1,2,...,\mathrm{k}$$

Let us build a matrix **G** with size *k*x*n*, the rows of which are $\mathbf{g}_1,\mathbf{g}_2,...,\mathbf{g}_k$. Coding is done by $\mathbf{c} = \mathbf{u}\cdot\mathbf{G}$ and the matrix **G** is called obviously as *generator matrix.*

The code *C* can thus be succinctly given by an appropriate set of *k* code words instead of listing all of its 2^k code words. Furthermore, the coding is done by a simple rule. Notice that several generator matrices belong to the same code, i.e. as many as many different bases the actual linear space can have. On the other hand, there is only one generator matrix which encodes given messages into given code words.

Going back to the example 7.1., we can see that the coding was chosen so that the first two bits of the code word are identical with the corresponding two-bit message. This is advantageous because the second decoding step, i.e. asserting the proper message to a code word is trivial as decoding here simply means detachment of the first *k* bits of the code word. The principle of such coding can be generalized as follows:

An (*n*,*k*) code *C* is called *systematic* if the first *k* bits of its code word correspond to the message. Generator matrix of a systematic code is unambiguous and according to the rules of matrix multiplication it is obviously of the following form:

$$\mathbf{G} = (\mathbf{I}_k,\mathbf{B}), \qquad (7.4.)$$

where $\mathbf{I}_k$ is a unity matrix of size kxk, **B** is a matrix of size k€x€(n-k). Structure of the code word belonging to the message u is as follows

$$\mathbf{c} = (u_1,u_2 ,...,u_k,c_{k+1},c_{k+2} ,...,c_n)$$

The segment of the first k co-ordinates of the code word is called the *message segment* and the segment of the last n-k co-ordinates is called the *parity segment.*

To select code words of C out of a set of 2^n binary vectors n bit long, an H (n-k)€x€n binary matrix can be assigned to the linear code C. For this matrix

$$\mathbf{Hc}^T = 0 \qquad (7.5.)$$

is true if and only if **c**€€ €C. ($(.)^T$ stands for the transposed matrix.) A matrix with such a characteristics is called a *parity-check matrix*. If the code is systematic then

$$\mathbf{H} = (\mathbf{A},\mathbf{I}_{n-k}), \qquad (7.6.)$$

where

$$\mathbf{A} = -\mathbf{B}_T \qquad (7.7.)$$

and $\mathbf{I}_{n-k}$ is an (n-k)€x€(n-k) unity matrix. Equations (7.6.) and (7.7.) can easily be proved:

Starting from eq. (7.3.) and (7.5.), a chain of equations can be set up for an arbitrary pair of c and u belonging together:

$$\mathbf{Hc}^T = \mathbf{H}(\mathbf{uG})^T = \mathbf{HG}^T\mathbf{u}^T = \mathbf{0}$$

so that

$$\mathbf{HG}^T = \mathbf{0} \qquad (7.8.)$$

Substituting equations (7.4.) and (7.6.) into (7.8.)

$$\mathbf{HG}^T= (\mathbf{A},\mathbf{I}_{n-k})(\mathbf{I}_k,\mathbf{B})^T = \mathbf{A} + \mathbf{B}^T= \mathbf{0}$$

which validates equation (7.7.)

Example 7.3.: Suppose we have the code C of example 7.1. the generator matrix of which is given in example 7.2. Using the notation introduced above:

$\mathbf{B} = \begin{pmatrix} 110 \\ 101 \end{pmatrix}$, and taking into account that -1 = 1 (modulo 2) $\mathbf{A} = \begin{pmatrix} 11 \\ 10 \\ 01 \end{pmatrix}$

So the parity matrix **H** is $\mathbf{H} = \begin{pmatrix} 11100 \\ 10010 \\ 01001 \end{pmatrix}$

In the following, a procedure will be shown how to use the **H** matrix for decoding. Let **c** be the code word sent and **v** the word received. The difference of the two vectors is called the *error vector*:

$$\mathbf{e} = \mathbf{v} - \boldsymbol{c}$$

For instance, if **c** = (10110) and **v** = (11110) then **e** = (01000) showing that the 2nd co-ordinate was mistaken. Notice that using (7.5.)

$$\mathbf{Hv}^{T} = \mathbf{H}(\mathbf{c}+\mathbf{e})^{T} = \mathbf{Hc}^{T}+\mathbf{He}^{T} = \mathbf{He}^{T},$$

i.e. the value of $\mathbf{Hv}^T$ is dependent only on the error vector and is independent of the code word. The following quantity

$$\mathbf{s} = \mathbf{eH}^{T} \qquad (7.9.)$$

is called the *syndrome* of the error vector *e*. Syndromes of the code words are **0**. (Row vector $\mathbf{eH}^T$ corresponds to the column vector $\mathbf{He}^T$). Figure 7.2. visualizes this multiplication (7.9.)

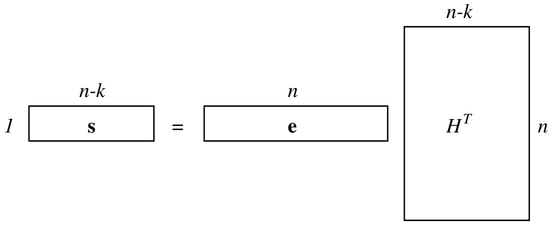


Fig. 7.2. Dimensional Presentation of the Syndrome Computation

Returning to example 7.3., syndrome of the error vector **e**=(01000) is **s**=(101). The length of the syndrome vector is *n-k*, that is 5-2=3.

As the syndromes of the error vectors are independent of codes, table lookup decoding of linear codes becomes possible. The structure of the table is as follows:

syndrome	error vector with min. errors
$\mathbf{s}_0=\mathbf{0}$	$\mathbf{e}_0=\mathbf{0}$
$\mathbf{s}_1$	$\mathbf{e}_1$
.	.
.	.
.	.
$\mathbf{s}_{2^{n-k}-1}$	$\mathbf{e}_{2^{n-k}-1}$

In the first row, there is the zero syndrome and the corresponding zero-error vector (error-free case). The length of the syndrome vectors is *n*-*k* so that the number of the syndromes is 2^{n-k}.

The steps of *syndrome decoding* are as follows:

1. Compute the syndrome **s** corresponding to the received word **v**.
2. Read out the predicted error vector corresponding to the computed *s* from the table.
3. Compute **c'= v - e**
4. Assign decoded message **u'** to the computed **c'**.

Let us illustrate the above procedure by an example!

Example 7.4.: Using the parity matrix **H** of the example 7.3., our (5,2) code *C* will result in the following table:

s	**e**
000	00000
100	00100
010	00010
110	10000
001	00001
101	01000
011	00011
111	01010

As it can be seen from the table, the code is able to correct all the five different single errors and also two double errors.

In the following let us overview the basic characteristics of some well-known and simple codes.

7.3. Simple Linear Codes

7.3.1. Repeat Code

The simplest error correcting code is the repeat code. In this case the length of the message is *k*€=€1, which is repeated *n*-times, so that the resulting code is *C*(*n*,1). According to only two possible messages (0 or 1), this code contains just two code words (000...0 and 111...1). Obviously, the code distance is *n* so that the error correction ability of the code is (*n*-1)/2 (preferably, *n* is an odd number).

Example 7.5.: The simplest repeat code being able to correct one error is the code made up of words (000) and (111). For instance, if the received word is (110) then (111) is decoded and the message is decided to be 1.

7.3.2. Single Parity-Check Code

The simplest error detection code is the single parity-check code. The code is generated in such a way that the message is extended by a single parity-check digit the value of which depends on the bits of the message. Parity-check can be chosen as even or odd, even-parity check is defined as modulo 2 sum of the elements while odd-parity check is its inverse, i.e. parity-check code is $C(n,n\text{-}1)$. It is easy to see that the code distance is 2. That is, if one bit of an arbitrary code is changed at arbitrary position, the parity is changed. Thus, by changing two bits, original parity (even or odd) is restored and another valid code is obtained which differs from the original code in two co-ordinates. Since the distance is 2, it follows from the theorem 7.1. that the single parity-check code is able to detect just one error.

7.3.3. Two Dimensional Parity-Check Code

Let us arrange the k bits of a message into a u pxq matrix**,** i.e. k=pq. Let us generate parity-check digits for each row and column and let us add the resulting digits at the end of each row and under each column respectively. Thus a c (p+1)x(q+1) matrix will be obtained if the bottom right element is also given somehow. Let us define this element as the "parity-check of the parities", i.e. parity-check digit for the bits standing in last row or the last column, respectively. It is easy to understand that this corner element is the same for the rows as for the columns.

The code distance is 4 which can be proved as follows: Suppose we have an even-parity check case and a u message matrix containing just one 1. In this case, the corresponding row and column parity bits will also be 1, same as the bottom right parity bit. That is, altogether four 1s will be in the code, so that the distance from the all-zero code word is 4 (Hamming distance is determined as the result of comparison of the corresponding matrix co-ordinates).

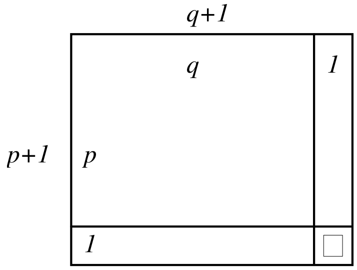


Figure 7.3 Two-Dimensional Parity Check Code

For any code, the Hamming distance of two arbitrary code words is equal to the number of non-zero elements found in their difference. Furthermore, for linear codes, the difference of two code words is also a code word. Consequently, minimum distance between code word pairs is equal to the number of 1s in the non-zero code word with the minimum number of 1s. As for the double parity-check code, there is no such a non-zero code word containing less than four 1-s, the code distance is really 4.

7.3.4. Hamming Code

Keeping in mind that the syndrome is computed as $\mathbf{s} = \mathbf{eH}^T$ (see also Fig. 7.2.), it can be seen that for vector $\mathbf{e}$ = (000...1...0) which contains just one 1 on the i.th coordinate, the i.th column of the matrix **H** can be assigned as its syndrome. Consequently, to have different syndromes for n different such error vectors, columns of the **H** have to be chosen to be different. This choice guarantees that the code corresponding to the matrix **H** is able to correct every single error. The code corresponding to the matrix **H** and the generator matrix **G** can be easily obtained by constructing the matrix **H** as systematic (7.6.) and then computing (7.7.).

Example 7.6.: For the (7,4) code C which is able to correct single errors, let us choose the following matrix **H** according to the construction rules described above:

$$\mathbf{H} = \begin{pmatrix} 1101100 \\ 1011010 \\ 0111001 \end{pmatrix}$$

The code length is $n = 2^3$-1 = 7 so that **H** has 7 different columns each containing $n - k = 7 - 4 = 3$ bit long non-zero binary vector. The generator matrix **G**, corresponding to this code is

$$\mathbf{G} = \begin{pmatrix} 1000110 \\ 0100101 \\ 0010011 \\ 0001111 \end{pmatrix}$$

The resulting code is a (7,4) Hamming code. Notice that this code is an optimum from a certain point of view. Namely, if the task is to construct a single-error correcting code with n = 7 then there is no such a code among them whose size is greater than 2^4 = 16. Of course, this construction can be generalized for parameters ($n = 2^r$-1, $k = 2^r$-1 - r) where r ≥ 3 is an integer number.

Exercises:

1. Let the generator matrix of a linear code be $\mathbf{G} = \begin{pmatrix} 010111 \\ 111010 \\ 110100 \end{pmatrix}$.

Determine:
a) the code words,
b) the systematic generator matrix,
c) the error correction ability,
d) the syndrome decoding table of the code.

2. Let us examine the (7,4) Hamming code. Is it true that if two errors have occurred at the transmission of a code word there is always a code word which differs from the received word in only one bit? Give an explanation!
3. Compute the probability of error for the decoding of (7,4) Hamming code if the probability of faulty bits at the channel input is *p* and the bit failures are independent.

Control Questions

1. How can the error ratio of the received bits be improved when transmitting information through a noisy communication channel?
2. How can error detection be used to improve quality of the transmission?
3. What is the algebraic structure of a linear code?
4. How many generator matrices can be assigned to a linear code?
5. List the methods used for decoding!
6. What would be the procedure of error correction for the two-dimensional parity code?
7. How can a code be constructed so that it is able to correct one error by constructing its parity matrix?

References

[1] Gyõrfi L.-Vajda I.: A hibajavító kódolás és a nyilvános kulcsú titkosítás elemei. Jegyzet, 1991

[2] Vajda I.: Hibajavító kódolás mûszaki alkalmazásai. Jegyzet, Mérnöki Továbbképzõ Intézet, 1982

[3] Fritz J.-Csiszár I.: Információelmélet. Tankönyvkiadó, 1983

8. GUIDED WAVE CHANNELS

Since wireless transmissions allow to use the radio channels for telephone communications in reasonable cases only (mobile telephones are such an exception), the telephone communication is carried mostly by guided waves i.e. cable connections. The whole world is covered by telecommunication network. Technical product of such a great complexity can operate only under uniform technical specifications. Such a set of specifications is defined and regularly updated by the CCITT. The goal of the present chapter is to give insight to the wave propagation modes used for the wirebound connections and to present some of the requirements and procedures specified by the CCITT recommendations. For more detailed studies, the reader is referred to the information given in the References.

8.1. Specification of the Signal Transmission

The method used for the transmission depends to a great extent on the properties of the signal to be transmitted. In today's communication a great variety of signals (speech, music, steady picture, moving picture, text, data, communication software), signals necessary for the control of the telephone exchanges (e.g. dialling) and feeding current of remote power supplies (e.g. for microphones) are transmitted. At the beginning, separated telecommunication networks were developed for the transmission of the different signals. In the last decades, these services are undergoing integration on different levels.

Besides the above signals, transmission techniques includes also the transmission of multiplexed signals created for the more efficient (multiple) usage of the line. Multiplexing procedures are discussed in Chapter 13. The essential consequence of the multiplexing is the increase of the bandwidth of the transmitted signals, proportionally to the extent of the multiplexing.

The telephone service as the means of speech transmission is of greatest importance. When a speech transmission is specified, it has to be kept in mind that the communication is being performed between two human brains and that the microphone and the receiver are positioned a few centimeters apart from the mouse and from the ear. The essential aim of the speech transmission is to maintain speech intelligibility at a 95...97 per cent level while other signal parameters are of minor importance, e.g. the recognition of the speaker is not necessary. Requirements for the signal transmission have been derived from the results of the subjective tests carried out for this purpose (see Table 8.1.).

Table 8.1 Main Parameters of Speech Transmission

Equivalent acoustic attenuation	30-40 dB
Signal-to-noise ratio	10-20 dB
Signal-to-crosstalk or echo	25-35 dB
Frequency range	300-3400Hz
Attenuation ripples in the range	2-15dB
Delay (for duplex connection)	250ms
Time-dependent ripple of the delay	±30ms
Envelope delay as a function of frequency	±30-60 ms
Phase shift	indifferent

8.2. Transmission by TEM Waveguides

8.2.1. Duplex Telephone Connection

The block diagram of a bidirectional duplex audio frequency telephone connection is shown in Fig. 8.1. The connection is terminated by 2-wire sections with repeaters while the middle part consists of 4-wire sections with repeaters. The reason of this arrangement is a complex technical and economical problem which will be explained in this chapter.

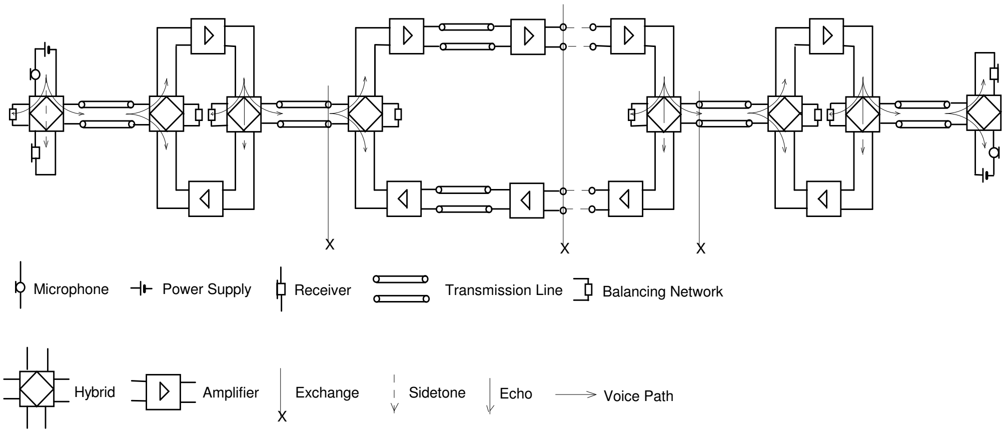


Figure 8.1. Duplex Audio Frequency Connection with Hybrids

The 2-wire sections are transformed to 4-wire sections by means of the so-called hybrid. A hybrid is a circuit with four ports. The input signal put at one of the ports is passed to the two neighbouring ports in such a way that the signal power is halved while there is no output at the port opposite to the input, provided the neighbour ports are terminated by the same impedance (see Fig. 8.2.a). Following each transformation, the signal is attenuated theoretically by 3 dB. Since the elements of the hybrid are not ideal, the total attenuation is about 3.5 dB as shown in Fig. 8.2.b.

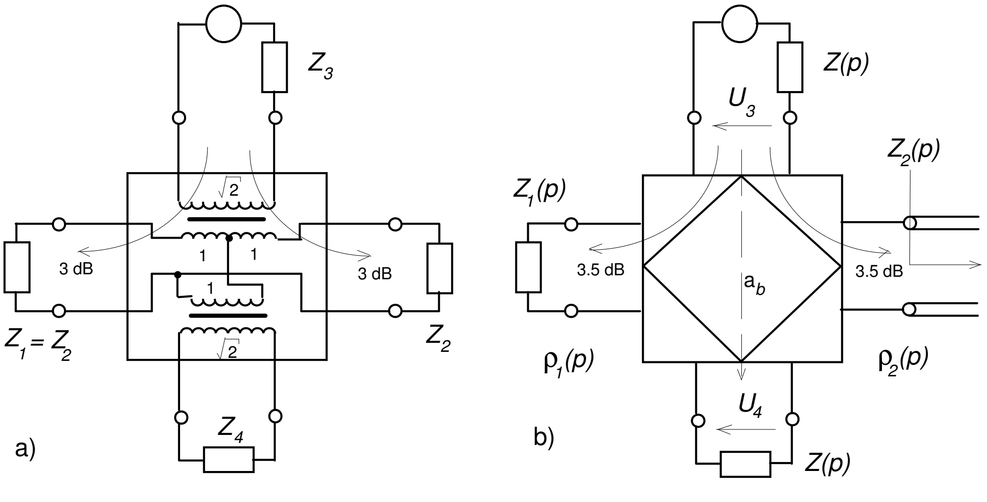


Figure 8.2 Power Distribution of an Ideal (a) and a Real (b) Hybrid

In the actual circuit the hybrid is terminated by a pair of amplifiers, input and output impedance of which being of the same value $Z(p)$. The transmission line has a frequency dependent wave impedance $Z_2(p)$ which is approximated by the balancing impedance $Z_1(p)$. As the consequence of the mismatch, the insertion loss undergoes a change but this is so small that we may still calculate with the 3.5 dB value. A certain amount of the signal, however, reaches also the opposite port producing thus an echo. The relative magnitude of this signal can be characterized by the backward attenuation a_z:

$$a_z(\omega)\text{€}=\text{€}20\lg|U_3(j\omega)/U_4(j\omega)|\text{€}=\text{€}7\mathrm{dB}-20\lg|\rho_2(j\omega)-\rho_1(j\omega)| \qquad (8.1)$$

where ρ_1 and ρ_2 are the reflection coefficients of the two-ports:

$$\rho_i(p)\text{€}=\text{€}[Z_i(p)-Z(p)][Z_i(p)+Z(p)]^{-1},\ i=1,2,\ |\rho_i(p)|\leq\text{€}1 \qquad (8.2)$$

The practical value of a_z is about 20-25 dB.

8.2.2. Structure of the TEM Waveguides

TEM waveguides consist of two metal conductors and a dielectric insulator between them. The distance of the conductors is small compared to the signal wavelength. In such waveguides only the TEM (transversal electromagnetic) base mode is present. Three different constructions are used as TEM waveguides:

1.) The so called *aerial line* is a pair of bronze wires mounted on telegraph poles. Due to the open structure, crosstalk with radio broadcasting limits the maximum usable frequency at 150 kHz (see Fig. 8.3.). The crosstalk between the aerial lines is reduced by regular swap of the two wires and by symmetrical line transformers.

2.) The symmetrical cable consists of two or four copper wires separated by an insulator. Flat cables, as well as aerial lines are used especially in the subscriber circuits. Twisted pair and star-quad cables are used in the 2-wire and 4-wire audio frequency circuits and in multiplex sections. The electromagnetic field of the symmetrical cable is more closed but crosstalk between such cables may occur by electromagnetic coupling if the cables are near to each other. This crosstalk can be compensated up to 600 kHz. In the case of a *single* symmetrical cable, the crosstalk is negligible and the maximum usable frequency is limited by the attenuation at 5 MHz (see Fig. 8.3.).

3.) The coaxial cable consists of two concentric conductors. The central wire is separated from the cylindrical outer shielding conductor by a cylindrical spacing layer of insulation. Coaxial cables cannot be used at frequencies lower than 60 kHz since the penetration depth of the electromagnetic waves propagating inside the cable reaches the thickness of the shielding tube which can be economically manufactured. The upper frequency is limited by the attenuation at 60 MHz (see Fig. 8.3.). Coaxial cables are used in multiplex connections.

In practice, common cables are made of the symmetrical and the coaxial cable, sheath of which consists of a metal-plastic-bitumen composition. The factory length of the air, terrestrial, river and undersea cables is about 1 km. The cables are wound on a drum.

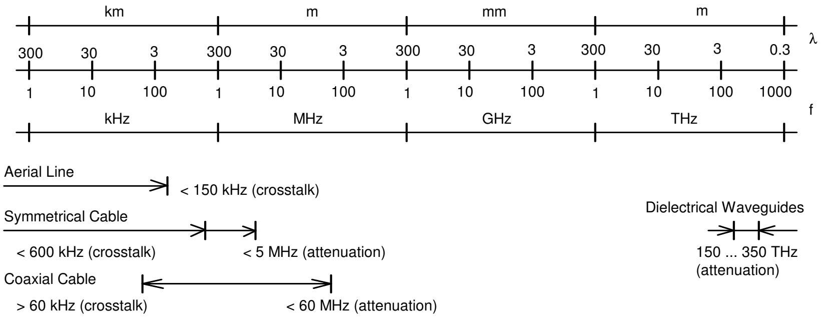


Figure 8.3. Frequency Range of the Waveguides Used in Telecommunication

8.2.3. Transmission Model of TEM Waveguides

TEM waveguides can be modelled by the transmission line. A section of such a line is characterized by the wave impedance $Z_0(p)$ and the propagation constant $\gamma(p)$. At the input the line is terminated by the generator impedance $Z_1(p)$ and at the output by the loading impedance $Z_2(p)$. Substituting $Z(p)$=$Z_0(p)$ into eq. (8.2), $\rho_i(p)$ can be computed. If $Z_i(p)$=$Z_0(p)$ then $\rho_i(p)$=0 so that voltages $U_i(p)$ are transmitted without reflection:

$$U_2(j\omega)€=€U_1(j\omega)€e^{-\gamma(j\omega)l}€=€U_1(j\omega)€e^{-[\alpha(\omega)+j\beta(\omega)]l} \qquad (8.3)$$

where $\alpha(\omega)$ is the wave attenuation constant and $\beta(\omega)$ is the phase shift constant:

$$a(\omega)€=€l\cdot\alpha(\omega) \;\; and \;\; b(\omega)€=€l\cdot\beta(\omega) \qquad (8.4)$$

Another essential transmission parameter is the delay. Suppose that a modulated signal (see chapter 11) is transmitted with carrier frequency ω_O so that $b(\omega)$ is constant within the passband range ω_1...ω_2. The modulation content (e.g. the envelope) will therefore be delayed by T_c, remains, however, undistorted while the carrier will be phase-shifted by T:

$$T_C€=€\tau_C(\omega_O), \;\; \tau_C(\omega)€=€db(\omega)/d\omega€=€l€d\beta(\omega)/d\omega€=€l/v_C(\omega), \qquad (8.5)$$

$$T_f€=€\tau_f(\omega_O), \;\; \tau_f(\omega)€=€db(\omega)/d\omega€=€l€d\beta(\omega)/d\omega€=€l/v_f(\omega), \qquad (8.6)$$

where $\tau_c(\omega)$ and $\tau_f(\omega)$ are the group and the phase delays, respectively, while $v_c(\omega)$ and $v_j(\omega)$ are the group and the phase velocities. It follows from Table 8.1. that the above $\beta(\omega)$ is suitable for speech transmission if ω_1≤€300€Hz and ω_2≥€3400€Hz.

$Z_0(p)$ and $\gamma(p)$ can be derived from R[Ω/km], G[S/km], L[H/km] and C[F/km] parameters of the TEM waveguides which are given for the unity of the length. At extreme frequencies the following approximate relations are valid:

$$\alpha(\omega)\big|_{\omega\to 0} \approx \sqrt{RG}, \qquad \alpha(\omega)\big|_{\omega\to\infty} \approx Q\sqrt{RG}, \qquad (8.7)$$

$$\beta(\omega)\big|_{\omega\to 0} \approx \omega Q\sqrt{LC}, \quad \beta(\omega)\big|_{\omega\to\infty} \approx \omega\sqrt{LC}, \qquad (8.8)$$

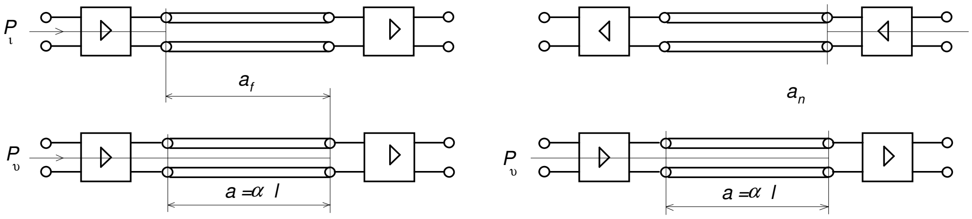
$$Q = \frac{1}{2}\left(\sqrt{\frac{RC}{GL}} + \sqrt{\frac{GL}{RC}}\right) \geq 1, \quad Z_0(\omega)\Big|_{\omega\to\infty} \approx \sqrt{\frac{L}{C}} \qquad (8.9)$$

In the whole frequency range (0...∞) the attenuation is directly proportional while the delay is inversely proportional to the value of Q, typical value of which is about 8 to 30 for TEM waveguides. In accordance with Table 8.1., this is favourable since the specification for the delay should be fulfilled for long-haul connections as well, while the attenuation can be compensated by amplifiers. For the different TEM waveguides the value of α is in the range from 0.03 to 4 dB/km.

8.2.4. Amplification of Audio Signals

The distance which has to be bridged over by wire connection is divided into amplified sections. The length of such sections is limited essentially by the near-end crosstalk. The near and the far-end crosstalks are defined by the direction of the signals (same or opposite) propagating on the two neighbouring wires as shown in Fig. 8.4. The useful power is denoted here as P_u and the power of the interfering signal as P_i, a_n and a_f are the near and the far-end crosstalk attenuations, K denotes the crosstalk protection factor. If a_k, α and K are given, l can be calculated.

Far End Crosstalk: Near End Crosstalk:



$K = a_f$ -αl, dB $K = a_n$ -αl, dB

Figure 8.4. Crosstalk Between Amplified Sections

Attenuations a_1 and a_2 of the 4-wire circuits shown in Fig. 8.1. are measured from left to right and from right to left, respectively, at the 2-wire terminals of the two hybrids. The attenuation of two 4-wire circuits is then a_1-7 and a_2-7 dB. If, however, during the operation of the exchanges the hybrids are temporarily terminated by open circuits then their attenuation will be 7 dB in accordance with eq. 8.1. and 8.2. Let a_l denote the loop attenuation of the 4-wire circuit terminated by open circuits. The sufficient condition of stability is a_l€>€0€dB. This condition can be interpreted also as a_l/2€>€0€dB resulting in value being close to attenuations a_1 and a_2:

$$\frac{a_l}{2} = \frac{a_1-7+7+a_2-7+7}{2} = \frac{a_1+a_2}{2} \quad \left[\text{dB}\right] \qquad (8.10)$$

i.e. if a_1€=€a_2€=€a then a_l/2€=€a. New circuits are designed with a_l/2€=€7 dB. The attenuation of the branches are 0 dB in this case, just compensating the attenuation of the cables thus arbitrary long distance can be bridged by the 4-wire amplified circuit.

Although 4-wire circuits are more expensive than the 2-wire ones, they are reasonable anyway since being used as trunk lines between two exchanges they serve many users. This feature is less effective at the ends of the connection therefore amplified 2-wire circuits are used there. The distance achievable by 2-wire circuits is strongly limited, however, because 2-wire amplifiers introduce additional loops into the system which can interact and thus degrade the stability of the system. Therefore the number of the loops should not be greater than three in the whole system.

Besides feedback, the finite attenuation of the hybrid causes echo, as well. The greater the delay, the more disturbing the echo is. The echoes of the 4-wire circuits are important since they do not have attenuation even for long lengths. For this reason, certain attenuation has to be allowed in long-haul communications or another solution has to be found to reduce the echo.

8.3. Transmission by Dielectric Waveguides

8.3.1. Principle of Dielectric Waveguides

Observing the light propagation in a glass fibre with high optical purity, α€=€0.2 dB/km minimum attenuation can be found in the frequency range between the infrared and the visible light. There is a 54 THz wide range between 1.3 and 1.7 μm where the attenuation does not exceed 0.5 dB. Outside this range, the light interacts with the electrons of the glass which increases the attenuation. In the case of small contamination caused by moisture, attenuation peaks appear in the above range so that newer minima of attenuation are present at 0.85 and 1.3 μm with attenuation values 2.5 and 0.6 dB/km, respectively.

Optical fibres are of huge importance for today's telecommunication since the resources of the Earth could not cover the demand for copper in the case of further usage of TEM waveguides.

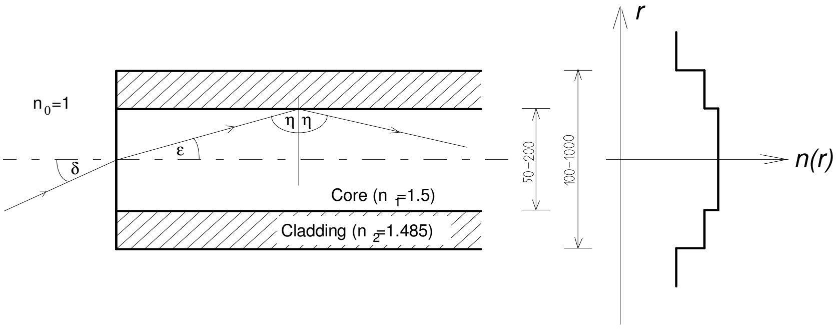


Figure 8.5. Step Index (SI) Optical Fibre

Glass fibre with homogeneous refraction index cannot be used as waveguide because of the dispersion. Surrounding, however, the glass core by a cladding with smaller refraction index as shown in Fig. 8.5., total reflection of the incoming rays can be achieved for sin $\eta \leq n_2/n_1$. *Numerical aperture* (*NA*) is the term used to define the maximum input angle $\delta_{\max}$ for which the beam stays within the core:

$$NA\text{€}=\text{€}\sin\text{€}\delta_{\max}\text{€}=\text{€}n_1\cdot\sin\text{€}\varepsilon_{\max}\text{€}=\text{€}n_1\cdot\cos\text{€}\eta_{\min}=\text{€}\sqrt{n_1^2-n_2^2}$$

(8.11)

8.3.2. Optical Transmitters and Receivers

The operation of light transmitters and detectors is based on the interaction of photons and electrons. For instance, semiconductor diodes are the devices operating under conditions providing such interaction. The main parameters of these devices are summarized in Tables 8.2 and 8.3. LD stands for Laser Diode, LED for Light Emitting Diode and APD for Avalanche Photo Diode.

Table 8.2 Main Parameters of Light Transmitters

Parameter	LED	LD
Emitted power	2 - 5 mW	15 - 20 mW
Incident loss	15 - 25 dB	3 dB
Middle of the band	0.8, 1.3, 1.55 μm	1.3, 1.55 μm
	350, 230, 200 THz	230, 200 THz
Unmodulated bandwidth	30 - 100 nm	0.5 - 5 nm
	10 - 30 THz	0.15 - 1.5 THz
Max. modulation freq.	50 - 100 MHz	3 - 20 GHz
Life expectancy	10^8 h	10^6 h

Table 8.3. Main Parameters of Light Detectors

Parameter	Si APD	Ge APD	In, GaAs	APD
Middle of the band $[\mu m]$	0.85	1.3	1.3	1.5
Max. modulation freq.[GHz]	0.15-1.5	0.5-2.5	0.5-4	05-2

Knowing the system specification, the proper device can be chosen. The light transmitters are strongly nonlinear, therefore impulse optical transmission (TDM) is mainly used for telecommunication.

8.3.3. Transmission Model of Dielectric Waveguides

The solution of the characteristic equation for electromagnetic field propagating in the SI optical fibre is shown in Fig. 8.6. Here β is the well known wave phase-shift constant, *c* is the free-space velocity of light, and the refraction indices are independent of frequency. There are only hybrid modes propagating along dielectric waveguides. The fundamental mode is propagating at all frequencies, the other modes are propagating above their own limit frequencies. If the free-space wave length is λ then the number of modes having their limit frequency lower than $f\text{€}=\text{€}c/\lambda$ is:

$$M\text{€}\cong\text{€}(\pi\cdot NA\ d/\lambda)^2\ , \qquad \text{if} \quad M \text{»} 1. \qquad (8.12)$$

There are about 1000 modes propagating in SI fibres. The group delay of the individual modes at a working frequency is different (see Fig.€8.5.) which - according to eq. (8.5.) results in mode dispersion of the received signal with group-delay difference $\Delta\tau_m\text{€}=\text{€€}(1/v_{cmin}\text{-}1/v_{cmax})$. Supposing the received signal to be a $\Delta\tau$ wide Gaussian impulse, at least $\Delta\tau_m$ time should be left between two transmitted impulses. The corresponding bandwidth can be expressed as

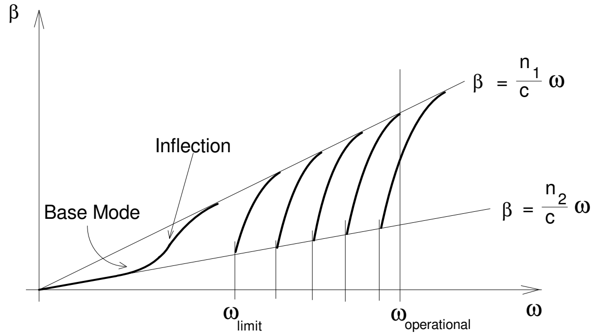

β
$\beta = \frac{n_1}{c}\,\omega$
Inflection
Base Mode
$\beta = \frac{n_2}{c}\,\omega$
ω_{limit}
$\omega_{\text{operational}}$
ω

Figure 8.6. Characteristics of Propagating Modes

$$B\text{€}=\text{€}0.44/\Delta\tau$$

(8.13)

so that the mode dispersion can be characterized by the product $B_m \cdot l$:

$$B_m \cdot l\text{€}=\text{€}0.44/(1/v_{\text{cmin}}\text{€}-\text{€}1/v_{\text{cmax}})\text{€}=\text{€}K,$$

(8.14)

where K is a constant characteristic for the fibre. Inhomogeneities of the fibre cause mode-coupling , however, which results in an experimental relation:

$$B_m \cdot l^{\zeta}\text{€}=\text{€}K \tag{8.15}$$

where $\zeta\text{€}=\text{€}0.5...1$. The mode dispersion can be modelled by means of the geometrical optics showing that the propagation paths of the axial beam and that of the reflected beams are different.

Mode dispersion can be significantly reduced by using GI (Graded Index) fibres in which the refraction index is decreasing towards the outside of the fibre. The wave velocity in a GI fibre is increasing towards the outside so that the delay of beams with different propagation paths is balanced. The mode dispersion can be entirely eliminated if only the fundamental mode can propagate on the working frequency, i.e. if $\pi \cdot NA \cdot d/\lambda\text{€}<\text{€}2.4$. To achieve this mode, small d (5 to 10 μm) and n_1-n_2 (0.003 to 0.008) are needed.

Chromatic dispersion, however, is present even in the SM fibres. This is due to the values of unmodulated bandwidth $\Delta\omega$ and wavelength $\Delta\lambda$ of the light sources (see Table 8.2) which are so large that the $\beta(\omega)$ of the fundamental mode cannot be considered flat. The $\beta(\omega)$ is curved even if $n(\lambda)$ is constant (waveguide dispersion), in the inflexion point the curvature is zero (see Fig. 8.6.). What is more important for the curvature is that because of the influence of the light on the particles of the material of the fibre, the $n(\lambda)$ is not constant but it has an inflexion at about 1.3 μm. In accordance with Fig. 8.5. this means that $\beta(\omega)$ would be curved even if the refraction index were of the same n_1 along the whole fibre.

As the result of the chromatic dispersion, the inflexion point of $\beta(\omega)$ can be found at about 1.3 μm. To make the analysis of the light emitting diode easier, its light is modelled by manifold of discrete carriers. The difference of the propagation delay of the maximum and the minimum carriers gives the chromatic dispersion $\Delta\tau$€=€$\Delta\lambda\cdot d\tau(\lambda)/d\lambda$€=€$\Delta\lambda\cdot l\cdot D_c(\lambda)$. Here, D_c is the chromatic dispersion

$$D_c(\lambda)\text{€}= \frac{1}{l}\frac{d\tau(\lambda)}{d\lambda} = \frac{1}{l}\frac{d\omega}{d\lambda}\frac{d\tau(\omega)}{d\omega} = -\frac{\omega^2}{2\pi c}\frac{d^2\beta(\omega)}{d\omega^2} \qquad (8.16)$$

which becomes really zero in the inflexion point of the $\beta(\omega)$. Similarly to eq. (8.14), chromatic dispersion can be characterized by the following equation:

$$B_c l = \frac{0.44}{\Delta\lambda\left|D_c(\lambda)\right|} \qquad (8.17)$$

Finally, the total transmitted bandwidth is given by

$$B^{-2}\text{€}=\text{€}B_m^{-2}\text{€}+\text{€}B_c^{-2}\text{€}+\text{€}B_a^{-2}\text{€}+\text{€}B_v^{-2}$$

*(*8.18)

where B_a and B_v are the maximum modulation frequencies of the light transmitter and receiver, respectively. The attenuation and the dispersion determine a maximum usable length of a line section after which an impulse regenerator or the so-called repeater has to be inserted.

8.3.4. Structure of the Dielectric Waveguides

The fibre core and the cladding are made of glass compositions which are directly covered by a soft plastic sheath. The next layer is made of a hard plastic or it is a plastic tube filled with petroleum jelly. The cable consists of several such fibres, kevlar fibres (high-strength strategic plastic) and an additional plastic cover. There is also a pair of copper wire in the cable for the remote power supply of the repeaters. The factory length of the air, ground, river and marine cables is between 1 and 5 km and the cables are delivered wound on a drum. To produce longer cables, the factory-length sections are glued (a = 0.2 dB) or welded (a = 0.5 dB) to each other and protected by a muff. The fibre is connected to the transmitter and the receiver by a connector (a = 0.2 to 1 dB). The connector is mounted on some *m* long piece of fibre which is welded to the first section of the installed cable.

Exercises

1. What is the maximum length of a symmetrical cable without amplification if R€=€54.3 Ω/km, G€=€1€μs/km, L€=€0.7 mH/km, C€=€ 38.5 nF/km, K€=€65 dB and a_n€=€91 dB? (l€=€14.9 km)

2. What is the maximum length of an optical cable between two repeaters if the data of the transmitter, cable, receiver, and channel are as follows:
   Light transmitter (LD): B_t=1€GHz, λ=1.3€μm, Δλ=5€nm
   Optical cable (GI): B_m(l=1km)€=€4.8€MHz, l€=€1€km, ζ€=€0.8, D_c€=5€ps/(km€nm), α€=€0.8 dB/km, a_conn€=€0.9 dB, a_weld€=€0.15 dB
   Light receiver (Ge APD): B_r€=€1€GHz

Channel: *B*€=€0.2€GHz, *a*€=€45€dB (N=42)

Control questions

1. What are the units of α, β and γ in eq. (8.3)?
2. What is the value of the attenuation determined by in the case of short and long 4-wire transmission?
3. What kind of dispersion is in an SI optical fibre?
4. Are there any metal parts in the optical cables?

References

[1] Simonyi K.: Elméleti villamosságtan. Tankönyvkiadó, Budapest, 1976. (published also in English)

[2] Izsák M.: Távközlési kézikönyv. Mûszaki Könyvkiadó, Budapest, 1979. (published also in English)

[3] Cebe L.:Fénytávközlés I. KKVMF, manuscript, 1990.

Notations

B	bandwidth	a	attenuation	α	attenuation constant
C	capacitance/km	b	phase shift	β	phase shift constant
D	dispersion factor	c	velocity of light	γ	propagation constant
G	reluctance/km	e	natural number	δ	angle
K	crosstalk immunity	f	frequency	ε	angle
L	inductance/km	j	imaginary unit	ζ	power of dispersion
M	number of modes	l	length of a section	η	angle
Q	quality factor	n	refraction index	λ	wavelength
R	resistance/km	p	complex frequency	$\Delta\lambda$	spectral width
T	delay	v	wave velocity	ρ	reflexion coefficient
U	voltage	τ	group delay	Z	impedance
ω	angular frequency				

Abbreviations

CCITT	International Telegraph and Telephone Consultative Committee
FDM	Frequency Division Multiplex
TDM	Time Division Multiplex
CDM	Code Division Multiplex
TEM	Transversal Electro-Magnetic
LW	Long Wave
MW	Medium Wave
SW	Short Wave
VHF	Very High Frequency
SI	Step Index

GI	Graded Index
SM	Laser Diode
APD	Avalanche Photo Diode

9. RADIO CHANNELS

As the telecommunication and the radio technique develop, the number of radio systems rapidly increases and newer applications are introduced. Radio systems have to meet, therefore, qualitatively new, enhanced requirements in an electromagnetic environment of growing complexity. For optimum frequency management, the estimation of the parameters of the radio links became more important. In this chapter, radio channel as the medium of the radio links is presented as well as the models of radio channels are overviewed.

9.1. Electromagnetic Spectrum, Frequency Bands

The electromagnetic spectrum used for the radio transmission is extending parallel to the progress in telecommunication. From the early days of radio the spectrum has been divided into ranges in accordance with the similar propagation characteristic and application areas. This division is permanently changed and extended towards the higher frequencies. The actual division of the electromagnetic spectrum together with typical applications is presented in Table 9.1.

Table 9.1. Frequency Ranges Used in Telecommunication

Range	Typical application
3 - 300 kHz	Navigation, beacons, LW broadcasting
300 - 3000 kHz	MW broadcasting, marine radio, navigation
3 - 30 MHz	SW broadcasting and amateur radio
30 - 300 MHz	TV and FM radio broadcasting, air navigation, mobile communication
300 - 3000 MHz	TV broadcasting, satellite communication
3 - 30 GHz	Radar, microwave link, mobile and satellite comm. and broadcasting
30 - 300 GHz	Radar and experimental communications

Models describing the various modes of wave propagation depend on the actual frequency band, the bands, therefore, will be given together with the presentation of the wave propagation modes.

9.2. The Radio Channel

To define a radio channel, the definition of the antenna has to be given first. The antenna is a device for the radiation and the reception of electromagnetic waves. From the system aspect, an antenna can be regarded as a transformer between the transmission line and the free space transforming the energy passed through the transmission line into a radiated electromagnetic wave (transmitter antenna) or transforming the incident electromagnetic wave to a guided wave (receiver antenna).

The radio channel is the medium which determines the parameters (amplitude, phase, polarization, spectrum) of the radio waves propagating between the transmitter and receiver antennas. From the system aspect, a radio channel is a four-pole, with the input of the transmitter antenna and with the output of the receiver antenna, as shown in Fig. 9.1. The attenuation of this four-pole is called *propagation attenuation* and is defined as follows:

$$a_p{}^{\mathrm{dB}}\text{€=€}10\text{€}\lg\text{€}\frac{P_{in}}{P_R} \qquad (9.1)$$

where P_{in} is the input power of the transmitter antenna and P_{R} is the maximum output power of the receiver antenna.

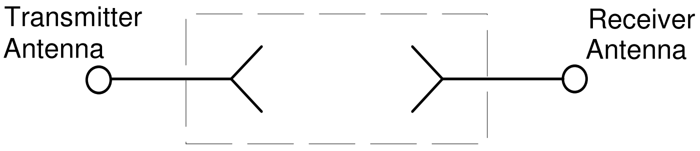


Figure 9.1. Radio Channel

Since the waves are propagating in radio channel without any man-made guide, the propagation attenuation is determined primarily by the properties of the medium between the antennas. To derive precise relations for describing the behaviour of the medium, wave propagation modes should be discussed. First, however, let us consider the properties of the antennas.

9.3. Antennas

9.3.1. Antenna as a Spatial Filter

An important feature of the antennas is their *directivity*. Since the radiation and/or the sensitivity of an antenna is not the same in all directions, this property is described by the *directional characteristic*.

Power radiated by a transmitter antenna is weighted by the directional characteristic and inversely, incident waves are also weighted by the receiver antenna. That is why the antennas are regarded as spatial filters.

9.3.2. Directional Characteristics of the Antennas

The directional characteristics of the antennas are given for the far field since antennas are usually located there. Far field strength at an $\mathbf{r}\text{€=€}(r,\vartheta,\varphi)$ point of the space can be given with the linearly polarized ϑ, φ components as follows:

$$\mathbf{E}(\mathbf{r})\text{€=€}\mathbf{E}(r,\vartheta,\varphi)\text{€=€}E_0(\mathbf{r})\ \mathbf{p}(\mathbf{r})\text{€=€}\frac{e^{-j\beta r}}{r}U_0(\vartheta,\varphi)\mathbf{p}(\vartheta,\varphi), \qquad (9.2)$$

where E_0 is the amplitude of the electrical field strength and $\mathbf{p}\text{€=€}p\vartheta\,\mathbf{e}\vartheta\,\text{€+}p_\varphi\mathbf{e}_\varphi$ is the polarization vector. To introduce the *normalized power characteristics* $P(\vartheta,\varphi)$, let us express the power density by means of eq. (9.2):

$$S(r,\vartheta,\varphi)\text{€=€}\frac{U_0^2(\vartheta,\varphi)}{240\pi r^2}\text{€=€}S_{\max}(r)\ P(\vartheta,\varphi) \qquad (9.3)$$

where S is the maximum power density.

Taking the square root of the $P(\vartheta,\varphi)$, a real function is obtained and it is called by definition the voltage directional characteristic or *amplitude characteristic* and is denoted as $F(\vartheta,\varphi)$.

Instead of the three-dimensional presentation, the directional characteristics use to be given as two-dimensional cross-sections with a fixed parameter φ. Cross-sections which belong to φ€=€0° and φ€=€90° are most frequently used and called *E* and *H*-plane directional diagrams (see Fig. 9.2.).

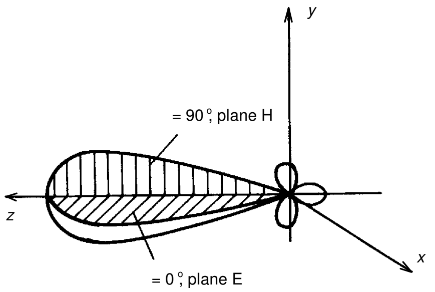


Figure 9.2. Three-Dimensional Directivity Pattern

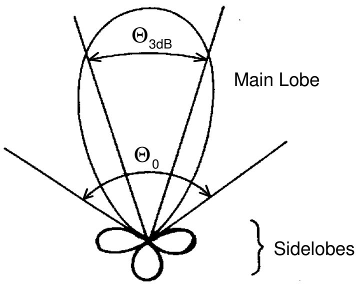


Figure 9.3 Two-Dimensional Directivity Pattern

Sometimes the directivity of an antenna is characterized simply by the 'width' of its main lobe. This is given as the conical angle determined by the zero directions encircling the main lobe and is denoted Θ_o. Usually, two narrower angles are also given: Θ_{3dB} denotes the width of the lobe, carrying half of the power (see Fig. 9.3.).

The directional characteristics of the various antennas are very different. For its importance, the *isotropic* antenna with *F*(ϑ,φ)€=€1 has to be mentioned. Although such an antenna cannot be realized, it is defined and used as a reference.

9.3.3. Unidirectional Effect and the Gain

Directivity of an antenna can be characterized also by its unidirectional effect. This is the ratio of the power density radiated in main direction and the power density of the isotropic antenna radiating the same power P_t:

$$D\text{€=€}S_{max}/S_0 \qquad (9.4)$$

where S_0€=€$P_t/4\pi r^2$.

The gain of an antenna is defined by the ratio of the power density radiated by the main lobe and the power density of the izotropic antenna fed by the same input power P_{in}:

$$G\text{€=€}S_{max}/S_0 \qquad (9.5)$$

where S_0€=€$P_{in}/4\pi r^2$.

So the gain is a transfer parameter which depends on the loss of the antenna. It follows from eq. (9.5) that the loss of the antenna can be characterized by the *efficiency* as follows:

$$\eta\text{€=€}\frac{G}{D}=\frac{P_t}{P_{in}} \qquad (9.6)$$

The terms directional gain and unidirectional effect are also used and defined as $G(\vartheta,\varphi)$€=€$GF^2(\vartheta,\varphi)$ and $D(\vartheta,\varphi)$€=€$D\,F^2(\vartheta,\varphi)$, respectively.

9.3.4. Effective Area of the Receiver Antenna

Receiver antennas can be regarded as active two-poles with inner impedance Z_{in}, open-circuit voltage U_{R} and maximum available effective power P_{R}. To characterize a receiver antenna, parameters describing the relation between the incident wave and the wave propagating on the transmission line have to be defined. One of such conversion parameters is the effective area of the antenna:

$$A_{\mathrm{R}}\text{€}=\text{€}P_{\mathrm{R}}/S \tag{9.7}$$

where S is the incident power density. In eq. (9.7) it is assumed that the polarization of the receiver antenna matches the polarization of the incident wave. Using the reciprocity theorem, it can be shown that between the gain and the effective area of an antenna the following relation exists:

$$\frac{G}{A_R} = \frac{4\pi}{\lambda^2} \tag{9.8}$$

Using eq. (9.8), the reciproque antennas can be sufficiently characterized by any one of the above parameters (usually by the gain).

9.4. Wave Propagation Modes

As it was shown in Chapter 9.2., the properties of the waves propagating between the transmitter and receiver antennas are determined by the radio channel. In the following we will examine the possible wave propagation modes which are resumed in Fig. 9.4. The main propagation modes are as follows: direct (or line-in-sight), reflected, surface, diffractive, tropospherical and ionospherical propagation.

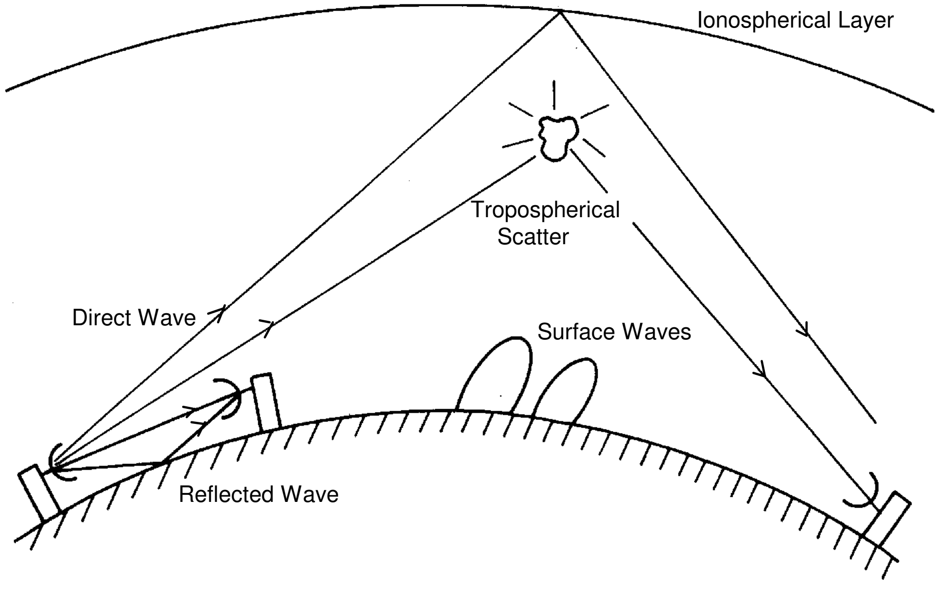


Figure 9.4. Main Wave Propagation Modes

9.4.1. Direct Wave and Free Field Attenuation

When modelling free-space wave propagation, the medium is assumed to be a homogeneous, ideal dielectric, free of charges and of current. In this case, the wave equation can be derived from the Maxwell equations, the general solution of which is a plain wave. Power density at the distance *r* can be determined from the input power of the transmitter antenna using eq. (9.5):

$$S(\vartheta,\varphi)€=€G\cdot F^2(\vartheta,\varphi)\cdot\frac{P_{in}}{4\pi r^2}$$

(9.9)

Free-space attenuation is understood as the attenuation of the radio channel for direct wave propagation. To derive this attenuation, let us express the power density at the receiver antenna by means of eq. (9.9). Suppose the distance from the transmitter antenna is *r* and the receiver antenna is in the main lobe of the transmitter antenna. On the basis of eq. (9.2) and (9.9), the free-space field strength is

$$E_0€=€\frac{\sqrt{60P_{in}G_t}}{r}$$

(9.10)

According to eq. (9.7), the receiver antenna transforms the power density of the incident wave to the power available at the antenna output:

$$P_{\mathrm{R}}€=€P_{\mathrm{in}}\cdot\frac{G_T A_R}{4\pi r^2}. \qquad (9.11)$$

Making use of eq. (9.8), the free-field attenuation can be expressed by the effective area of the transmitter and the receiver antennas

$$a_0=\frac{(r\lambda)^2}{A_T A_R}=\left(\frac{4\pi r}{\lambda}\right)^2\frac{1}{G_T G_R}$$

or it can also be given in the logarithmic form:

$$a_0=20\log\left(\frac{4\pi r}{\lambda}\right)-(G_T^{dB}+G_R^{dB}) \qquad (9.12)$$

9.4.2. Refraction

When speaking about direct wave propagation, *refraction* has also to be taken into consideration. While the wave propagation in vacuum is rectilinear, in the presence of the atmosphere the optical and the radio waves refract because of the varying refractive index of the air.

Air refractive index *n* is slightly greater than one. At the sea level, at normal climate $n€=€1.0003$. This value decreases with increasing height above sea level. To be able to better distinguish the small variations of the refractive index, let us express it as $n€=€1€+€N\ 10^{-6}$. The value of the refractivity *N* is given as

$$N\text{€=€}77.6\,\frac{p}{T}\text{€+€}3.7310^{5}\frac{e}{T^{2}}$$

(9.13)

where p is the air pressure [mbar], e is the partial tension of vaporized water [mbar] and T is the temperature [K]. As the result of numerous measurements, N could be approximated for the normal climate (standard atmosphere) as the function of the height above sea level h in the following form:

$$N(h)\text{€=€}315\text{€}e^{-0.136\mathrm{h}}.$$

(9.14)

Because of refractivity changes, the waves propagating in the atmosphere refract towards the Earth as is shown (in an exaggerated form) in Fig. 9.5.

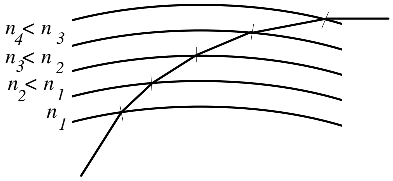


Figure 9.5. Refraction of Radio Waves

Instead of the computation of the curved propagation path, an effective radius of the Earth is defined for practical calculations: $R_{\mathrm{eff}}\text{€=€}kR_{\mathrm{o}}$ (R_{o} is the actual radius) and the path of the wave is regarded to be straight. The radius coefficient k is determined by the actual radius of the Earth (6370 km) and by the gradient of the refractive index:

$$k\text{€=€}\frac{1}{1+R_0\dfrac{dn}{dh}} \qquad (9.15)$$

From eq. (9.14), $k\text{€=€}4/3$ so that $R_{\mathrm{eff}}\text{€=€}8500$ km.

9.4.3. Ground Back-Scatter

To describe the reflection of the radio waves from the ground, a model of plain waves reflecting from a dissipative dielectric is used. Assuming the relative permitivity of the dielectric extended to the infinite half plane to be ε_{r} and its conductivity to be σ, ground reflectivity is defined as the ratio of the reflected and the incident field strengths:

$$\Gamma_f=\frac{E_r}{E_i} \qquad (9.16)$$

Conditions of the ground reflectivity for the horizontal and vertical polarization are shown in Fig. 9.6. It can be seen that for small incident angles ($\vartheta<5^{\circ}$) $\Gamma_{\mathrm{f}}\text{€=€}-1$ regardless of the polarization and the frequency. Another important feature is that for vertical polarization $|\Gamma_{\mathrm{f}}^{\mathrm{v}}|$ has a minimum value at the so-called Brewster-angle and $\left|\,\Gamma_f^V\right|$ =€0 here, if the dielectric is supposed to be ideal.

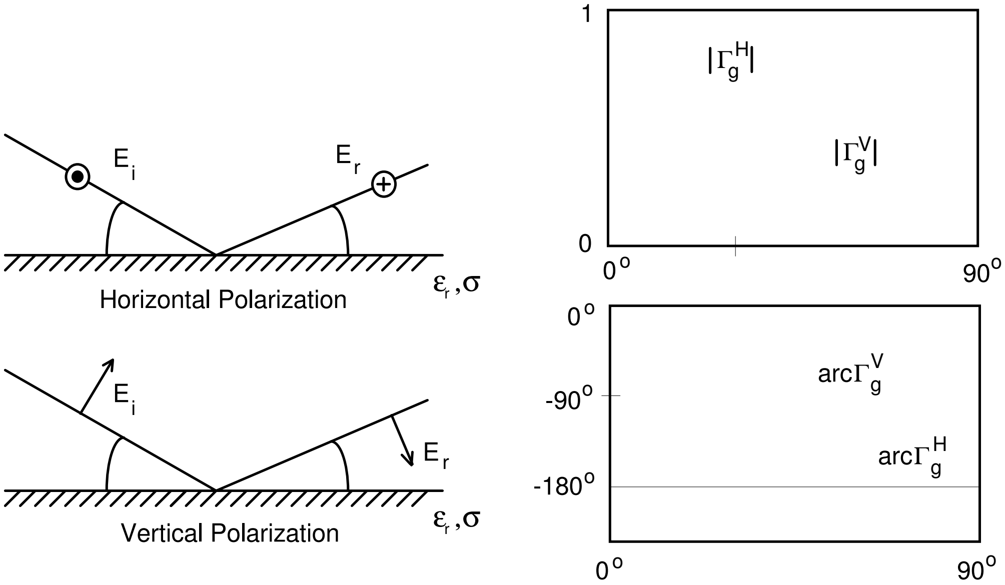


Figure 9.6. Ground Back-Scatter and the Reflection Coefficient

9.4.4. Ground Wave Multipath Propagation

Multipath propagation means that at least one reflected wave is received simultaneously with the direct wave (see Fig. 9.7.) This model is used mainly in the VHF, UHF and microwave range for mobile communications or links established within the horizon.

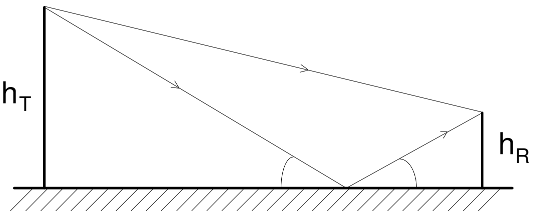


Figure 9.7. Multipath Propagation

Let us denote the length difference of the direct and the reflected path as Δ. Then the resulting field strength E_R at the receiver antenna is obtained by the summation of the direct E_0 and the reflected E_r components:

$$E_R€=€E_0€+€E_r€=€E_0€+€E_0\Gamma_f€e^{j\beta\Delta}$$

(9.17)

Since for small values of ϑ $\Gamma_f€\cong€-1$, i.e.

$$E_R€=€E_0€+€E_r€\cong€E_0€(1€-€e^{j\beta\Delta})$$

(9.18)

The length difference in the direct and the reflected paths can be expressed by the height and the distance of the two antennas:

$$\Delta € = € R_2 € - € R_1 € \cong \frac{2h_T h_R}{r}$$

(9.19)

From eq. (9.18) and (9.19) the absolute value of the field strength at the receiver antenna is

$$|E_\mathrm{R}| € = € 2 € |E_\mathrm{O}| \cdot |\sin\beta \frac{2h_T h_R}{r}| \tag{9.20}$$

thus the attenuation of the multipath propagation is

$$a_\mathrm{p} € = € 20 € \lg[r^2/(h_\mathrm{T} h_\mathrm{R})] € - € (G_T{}^{\mathrm{dB}} € + € G_\mathrm{R}{}^{\mathrm{dB}})$$

(9.21)

9.4.5. Diffraction

According to the laws of geometrical optics, obstacles standing in the path of wave propagation in the free space shadow the receiving antenna. Fortunately, this is the law of wave optics which determines the behaviour of the wave in this case, so that each point of a radiated wavefront is a secondary (Huyghens) source of elementary waves, the radiation of which is summed up (in the correct phase) with the waves of other elementary sources.

To model the obstacles of the terrain, parabolic cylinder or knife edge is used as a model. In the case of the knife edge model the field strength at the receiver depends on the relative height of the knife edge as shown in Fig. 9.8.

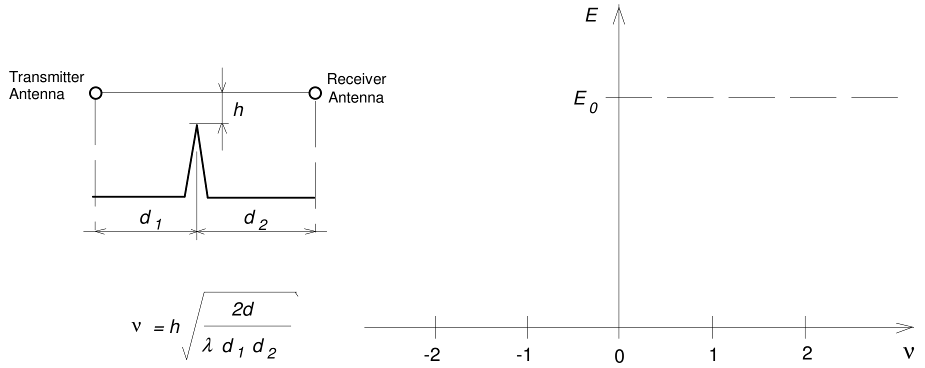


Figure 9.8. Field Strength as a Function of Diffraction

9.4.6. Surface Wave Propagation

Surface waves are generated at the boundary between the well conducting ground and the air. If the height of the antenna is low in comparison with the wavelength, wave generation is efficient since the direct and the reflected waves cancel each other in this case. The conductivity of the soil is good at frequencies from a few kHz to a few MHz so that surface wave propagation is important in this range. The channel attenuation is proportional to the fourth power of the distance. Distances used in practices span to several hundreds of km.

Field strength of vertically polarized surface waves is given by the following equation:

$$E€=€E_0 \cdot A(p) \tag{9.22}$$

where $A(p)=\dfrac{2+0.3p}{2+p+0.6p^2}$ is the surface wave attenuation factor, $p=\dfrac{\pi}{60\lambda\sigma}\left(\dfrac{d}{\lambda}\right)$ is the so-called *numerical distance* and σ is the conductivity of the soil. Using eq. (9.22) the section attenuation can be given as

$$a_s^{dB}=20\log\frac{d\lambda}{4\pi\cdot A(p)}-(G_T+G_R) \tag{9.23}$$

For long distances, $A(p)€\cong€1/2p$ and the attenuation is proportional to the fourth power of the distance. For horizontal polarization the attenuation is much higher wherefore only the vertical polarization is used in practice. The greatest advantage of the surface waves is their ability to follow the curvature of the ground thus propagating to long distances, beyond the horizon.

9.4.7. Tropospherical Scatter

As it was shown, the refraction index of the atmosphere is periodically changing and the change can be described well in a long-term average. Nevertheless, because of the fast local changes of humidity, pressure and the temperature of the air, the refractive index may exhibit sudden changes. These changes are small, however, they may cause significant power-scatter if the transmitted power is large.

Tropospherical links operate in the range between 200 MHz and 10 GHz. The minimum frequency is limited by the size of the high-gain antennas while the attenuation become significant at the high frequencies. A characteristic feature of tropospherical links is the significant fluctuation of the field at the receiver. The typical distance of a tropospherical link is some hundreds of km, usually not more than 800 km. To set up a link through tropospherical scatter, lobes of the transmitting and the receiving antennas have to create a so-called *common scatter volume*. This is generated in the troposphere usually in the height within 10 km.

9.4.8. Ionospherical Propagation

The ionosphere is a layer of ionized gas particles surrounding the Earth in the height between 40 and 100 km. The ionization is caused by the solar ultraviolet and particle-radiations and by the meteorites. Since the main source of the ionization is the Sun, the state of the ionosphere depends primarily on solar activity. Ionospherical layers are characterized by the number of free electrons in the unit of volume. On the base of local maxima of electron density, D, E and F layers are distinguished. At day-time, the layer F splits into layers F_1 and F_2 while at night-time only D and F layers show up.

Since the refractive indices of these ionospherical layers are different, radio waves reflect from the layers. A maximum frequency called the *critical frequency* belongs to each layer. This is defined so that the waves with higher frequencies than the critical frequency will reflect with a probability less than 50 per cent. If the waves reach the layer askew then signals with frequency higher than the critical f_C are also reflected. The relation between the maximum usable frequency (MUF) and the critical frequency is then determined by the skew angle ψ as MUF€=€f_c/cos€ψ.

Control Questions

1. What is the radio channel?
2. Give the definition of section attenuation!
3. How does the antenna act as a spatial filter?
4. Define the normalized power directional characteristics!
5. Define the normalized voltage directional characteristic!
6. What is the directional diagram?
7. What is the definition of the gain and of the directivity?
8. How is the effective area of the antenna defined?
9. List the wave propagation modes!
10. Give the section attenuation formula for the double path propagation!

Exercises

1. Compute the free-space attenuation of the radio link operating on 450 MHz if the distance between the transmitter and the receiver is 10 km and the gain of the antennas is equally 20 dB.
2. Compute the electrical field strength at the receiver if the operating frequency is 145€MHz, the input power of the transmitter is 10€W, the gain of the transmitter antenna is 10€dB and the distance between the 10€m high transmitter and receiver antennas is 5€km. Suppose there is a double-path propagation and that the reflection coefficient of the ground is -1.

Bibliography

[1] Collin R. E.: Antennas and Radiowave Propagation. McGraw-Hill. New York, 1985.

[2] Stutzman W.L.-Thiele G.A.:Antenna Theory and Design. John Wiley. New York, 1981.

List of Notations

a_p	propagation attenuation
P_{in}	input power of the transmitting antenna
P_R	max. output power of the receiving antenna
$P(\vartheta,\varphi)$	normalized power characteristics
$F(\vartheta,\varphi)$	normalized amplitude characteristics
D	unidirectional effect
$D(\vartheta,\varphi)$	directivity of the unidirectional effect
G	gain
$G(\vartheta,\varphi)$	directivity of the gain

λ	wavelength
p	polarization vector
β	phase coefficient
S	power density
S_0	free-space power density of the izotropic antenna
P_t	power radiated by the antenna
Z_in	input impedance of the antenna
U_R	open circiut output voltage of the receiver antenna
A_R,A_T	effective area of the antenna
E_0	free-space field of the antenna
a	free-space attenuation
p	air pressure
e	partial tension of the steam
T	air temperature
n	refractive index
N	refractivity
R_0	radius of the Earth (6370 km)
R_eff	effective radius of the Earth
k	radius coefficient
Γ	groung-reflecivity ratio
$\Gamma_\mathrm{g}{}^\mathrm{H}$,$\Gamma_\mathrm{g}{}^\mathrm{V}$	ground-reflectivity ratios for horizontal and vertical polarization

10. NOISE

Naturally occurring and man-made noises have to be considered generally as the electromagnetic environment of telecommunication systems. The solutions of different telecommunication problems is made more difficult by the fact that noises are superposed to the useful signal during signal processing.

In the following chapters, basic noise terms and relations will be discussed. We will concentrate almost exclusively on the thermal noise, for other types of noises we are confined to the conceptual introduction of the phenomena.

10.1. Thermal noise

Noises can be classified by their stochastic behaviour (by amplitude distribution and frequency dependence), by their physical nature, etc. There is an important class of noises originating from the incidental fluctuation around a thermal balance. The properties of thermal noise have been exactly derived by physicists, using quantum mechanics approach. We will simply quote the appropriate results in the following discussion. Output noise power of an arbitrary passive network with absolute temperature *T* is characterized by

$$S(f) = \frac{hf}{e^{\frac{\mathrm{hf}}{\mathrm{kT}}} - 1} \tag{10.1}$$

where *S*(*f*) is the power spectral density, *f* is the frequency, *h* is the Plank's constant and *k* is the Boltzman constant. Equation (10.1.) is called Plank's law and it determines the maximum available output noise power of a physical system as the function of frequency.

At low frequencies, the exponent in the denominator is so small that the exponential function can be approximated by the first two elements of its Taylor-series

$$S(f) = \frac{hf}{e^{\frac{\mathrm{hf}}{\mathrm{kT}}} - 1} = \frac{hf}{\left(1 + \frac{hf}{kT} + \ldots\right) - 1} \cong k\,T \tag{10.2}$$

The lower the frequency and the higher the temperature are, the more precise is the approximation above. E.g., if the frequency is 30 GHz and the temperature 30 K, the error caused by the approximation is less than 0.1 dB. In the optical region, conditions are quite different, e.g. at 200 THz frequency and 2000€K temperature the noise is only one thousandth part of that computed from the approximation. That is as far as frequency response of the noise is concerned, the 'radio' range can be distinguished from the `optical' range. In the radio range, the power spectral density of thermal noise is practically constant (*kT*) while in the optical region the thermal noise can be neglected.

So far we have not discussed the precise meaning of 'temperature', supposing that it is equivalent to the physical absolute temperature. In the most simple case, it really is, but in more complex systems the temperature may be different so that an average should be defined to obtain the resulting conditions. Let us postpone the detailed description of this definition and let the so called *equivalent noise temperature* be used as the quantity which characterizes the noise source as if it were the real source temperature.

10.2. Noise Characteristics of a Transmission System

Because the transmission system adds its own noise to the processed signal, there are three components appearing at the output: the useful signal, the noise of the input source and the system's self noise . A normalization is used for system computations to obtain the instantaneous power by squaring the normalized quantity.

E.g. if voltage u is measured on a resistance R then

$$s = \frac{u}{\sqrt{R}} \qquad (10.3)$$

is given as the normalized value of the signal. Suppose the frequency response is constant in the transmission range, then

$$s_{\mathrm{o}} = A{\cdot}s_{\mathrm{i}} + A{\cdot}n_{\mathrm{i}} + n_1 \qquad (10.4)$$

where s_{o} is the output signal, A is the transfer function of the system, s_{i} is the input signal, n_{i} is the input noise and n_1 is the noise generated by the system. The output power is then

$$P_0 = \overline{s_0^2} = \overline{\left(A \cdot s_i + A \cdot n_i + n_1\right)^2} \qquad (10.5)$$

Since the components are independent, the expected value of their products is zero, i.e.

$$P_{\mathrm{o}} = A^2 \cdot \overline{s_i^2} + A^2 \cdot \overline{n_i^2} + n_1^2 \qquad (10.6)$$

Taking into account that the square of the transfer factor is identical with the power gain factor G, we receive

$$P_{\mathrm{o}} = G{\cdot}P_{\mathrm{si}} + G{\cdot}P_{\mathrm{ni}} + P_{\mathrm{n}} \qquad (10.7)$$

Eq (10.7.) seems to be obvious as it states that the output power is the sum of three powers (that of the amplified input signal and amplified input noise, and that of the system noise). This is true only if the three sources are independent otherwise their mutual relation shall be taken into consideration as well.

Output power can also be expressed by the source noise temperature

$$P_{\mathrm{o}} = G{\cdot}P_{\mathrm{si}} + G{\cdot}B{\cdot}k{\cdot}T + P_{\mathrm{n}} \qquad (10.8)$$

or

$$P_{\mathrm{o}} = G{\cdot}P_{\mathrm{si}} + G{\cdot}B{\cdot}k{\cdot}\left(T + \frac{P_n}{GBk}\right) \qquad (10.9)$$

As it turns out of eq. (10.9), the influence of the system can be taken into consideration as if the temperature had risen

$$P_\mathrm{o} = G{\cdot}P_\mathrm{si} + G{\cdot}B{\cdot}k{\cdot}\left(T + T_\mathrm{red}\right) \tag{10.10}$$

thus the second temperature is called *effective (input) noise temperature*. This is a very important quantity as it unanimously defines the system's contribution to the resulting noise.

Besides the effective noise temperature, *noise figure* is one of the most commonly used terms to characterize the noise properties of a block. Noise figure is defined as the ratio of the output noise power related to the amplified input noise power, provided the temperature of the input noise source is T_o = 290 K. The noise figure formula is

$$F = \frac{P_{n0}}{GP_{ni}}\bigg|_{T_0} \tag{10.11}$$

As the output power of the useful signal is G times the value of the input signal power, the noise figure shows also the degradation of the signal-to-noise ratio when the temperature of the input noise source is supposed to be 290 K. Noise figure is usually given in dB (as 10·lg€F). It should be emphasized that 290€K as temperature reference is an important part of the definition of the noise figure, without it the noise figure would not be unanimously defined. The chosen reference value represents a good average for terrestrial conditions. Although laboratory temperature might be somewhat greater than 290€K, this deviation is often not taken into consideration because the resulting error can usually be neglected. For instance, if the laboratory temperature is 310€K and the temperature of the equipment acting as the noise source is greater by 20€K than the ambient temperature, the error is about 0.6 dB.

Both the equivalent noise temperature and the noise figure are unambiguous noise characteristics of a block. The relation between them can be derived from eq. (10.11.)

$$F = \frac{P_{n0}}{GP_{ni}}\bigg|_{T_0} = \frac{GkB\left(T_0 + T_{red}\right)}{GkBT_0} = 1 + \frac{T_{red}}{T_0} \tag{10.12}$$

or

$$T_\mathrm{red} = (F - 1){\cdot}T_\mathrm{o} \tag{10.13}$$

10.3. Noise Figure of Multistage Systems

It is a common task to determine the resulting noise figure of a system consisting of more than one block cascaded as shown on Fig. 10.1.

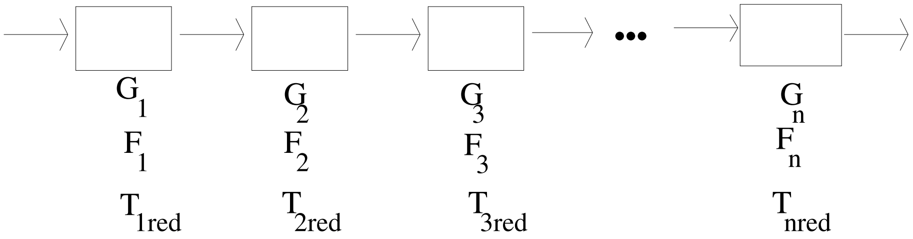


Figure 10.1. Parameters of a Multistage System

The resulting gain is the product of the individual gains

$$G = G_1{\cdot}G_2{\cdot}G_3 \ldots{\cdot}G_{\mathrm{n}} \qquad (10.14)$$

To compute the resulting equivalent noise temperature, the output noise power has to be determined first. For easier calculation, suppose that n=2, i.e. there are only two stages. Output noise power of the first block is

$$P_{1\mathrm{o}} = G_1{\cdot}P_{\mathrm{ni}} + G_1{\cdot}B{\cdot}k{\cdot}T_{1\mathrm{red}} \qquad (10.15)$$

Since the output of the first stage acts as the input of the second one, using eq. (10.10.) again

$$P_{2\mathrm{o}} = G_2{\cdot}P_{1\mathrm{o}} + G_2{\cdot}B{\cdot}k{\cdot}T_{2\mathrm{red}} \qquad (10.16)$$

Substituting eq (10.15.) into (10.16.)

$$P_{2\mathrm{o}} = G_1{\cdot}G_2{\cdot}[P_{\mathrm{ni}} + B{\cdot}k{\cdot}\left(T_{1\mathrm{red}} + \frac{T_{2red}}{G_1}\right)] \qquad (10.17)$$

Resulting equivalent noise temperature can also be read out from the above formula

$$T_{\mathrm{red}} = T_{1\mathrm{red}} + \frac{T_{2red}}{G_1} \qquad (10.18)$$

Results obtained for the two-stage case can be generalized for more stages as

$$T_{\mathrm{red}} = T_{1\mathrm{red}} + \frac{T_{2red}}{G_1} + \frac{T_{3red}}{G_1G_2} + \ldots + \frac{T_{nred}}{G_1G_2\ldots G_{n-1}} \qquad (10.19)$$

Using eq. (10.12.), the resulting noise figure is

$$F = F_1 + \frac{F_2-1}{G_1} + \frac{F_3-1}{G_1G_2} + \ldots + \frac{F_n-1}{G_1G_2\ldots G_{n-1}} \qquad (10.20)$$

The last two equations point to an important conclusion: If the gains are sufficiently great, only the first stage is dominating from the point of the noise. Thus if the noise of a system has to be optimized, the parameters of the preamplifier are critical.

The performance of the preamplifier can, however, be remarkably degraded if a lossy block (e.g. a long cable) is inserted between the signal source (e.g. antenna) and the input, as shown in Fig 10.2.

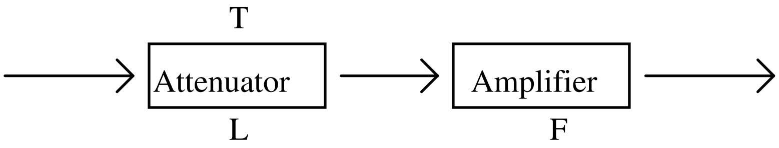


Figure 10.2. Two-Stage System with a Passive Input Block

Interpreting the attenuation *L* of the first stage as

G_1€=€1/L, (10.21)

the relations obtained for the two-stage case can be applied.

To compute the noise figure and the equivalent noise temperature of the attenuator, let us complete the attenuator with an input block of the same temperature, as shown in Fig. 10.3.

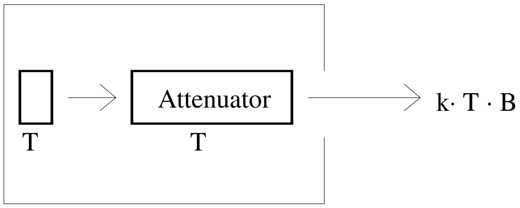


Figure 10.3. The Passive Block as a Source of Noise

The temperature of the resulting system is *T*, so that the output power in the radio range is *k·T·B*. On the other hand, the general rule defined by eq. (10.10) can also be applied

$$k{\cdot}T{\cdot}B = \frac{1}{L}{\cdot}B{\cdot}k{\cdot}\left(T + T_{\mathrm{red}}\right) \qquad (10.22)$$

so that

$$T_{\mathrm{red}} = \mathrm{T}{\cdot}(L - 1) \qquad (10.23)$$

and

$$F_{\mathrm{L}} = 1 + \frac{T}{T_0}{\cdot}(L - 1) \qquad (10.24)$$

The above results are especially simple if the attenuator temperature is close to the reference. In this case

$$F_{\mathrm{L}} = L \qquad (10.25)$$

so the noise figure of the attenuator (cable) is the same as its attenuation (provided its temperature may be taken as 290€K). Furthermore, using eq. (10.20), the resulting noise figure will be

$$F_r = L \cdot F \tag{10.26}$$

Since multiplication means addition of the corresponding values in dB, one may say that the noise figure is increased by as many dB-s by as many dB-s the signal has been attenuated before being amplified.

It was shown that if the noise temperature of a source (e.g. antenna) is equal to the reference temperature then the degradation of the signal-to-noise ratio is directly given by the attenuation. If the noise temperature of the source is lower, the signal-to-noise ratio can be degraded much more.

10.4. Effective Noise Temperature of Composite Sources

Let us examine a composite system consisting of a source and two attenuators each having different temperature as shown on Fig. 10.4.

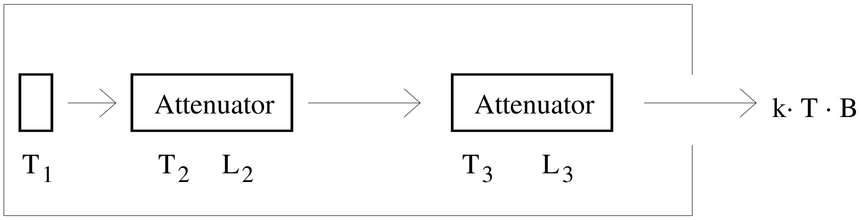


Figure 10.4. Composite Source of Noise

As the first step, a system consisting of only two elements will be discussed. The simplified system is shown on Fig. 10.5.

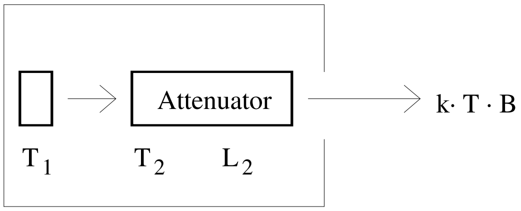


Figure 10.5. Complex System as the Source of Noise

According to eq. (10.23), equivalent noise temperature at the attenuator input is $T_2(L_2$-1), the corresponding 'gain' is $1/L_2$ so that

$$P = \left(T_1 + T_2(L_2 - 1)\right) k \cdot B \cdot \frac{1}{L_2} \tag{10.27}$$

As it can be read out of the equation

$$T = T_1 \cdot \frac{1}{L_2} + T_2 (1 - \frac{1}{L_2}) \tag{10.28}$$

The above results can be applied for the system with two attenuators so that the resulting noise temperature of the complete system will be

$$T = T_1 \frac{1}{L_2} \cdot \frac{1}{L_3} + T_2 (1 - \frac{1}{L_2}) \cdot \frac{1}{L_3} + T_3 (1 - \frac{1}{L_3}) \tag{10.29}$$

For further discussion it is necessary to interpret the above result from the physical point of view. Let us examine the second member of the right side of the equation, as it seems to be a general one, preceded and followed by other terms.

We can see that the temperature is decreased, on one hand, because the following block (3) attenuates the noise generated by block 2 by a factor $1/L_3$. On the other hand, the temperature is increased by a factor which is proportional to the inner loss of the block 2. It is important to mention that a 'transparent' block (i.e. one without loss) plays no role in the resulting loss, noise is generated only by passive elements.

As mentioned in chapter 10.2., the structure of a system is indifferent from the point of view of noise, the output power can always be computed from eq. (10.1) if all its elements have the same temperature. For the time being we have used this equation for guided-wave structures but it is valid for systems containing radiating elements as well. Thus, if we measure noise power at the output of an antenna this will also satisfy eq. (10.1), and in the radio range the simplified version (10.2) may be used.

From the physical aspect, the noise of such a system is not generated by the antenna because the lossless elements do not contribute to the resulting noise figure. The reason of the noise appearing at the antenna output is that there are noise sources in the radiation field picked up by the antenna (being weighted by the gain belonging to the corresponding direction). Weighted average temperature of differently located noise sources with different temperatures can be computed as

$$T_a = \frac{1}{4\Pi} \cdot \iint_{(4\pi)} T(\varphi, \vartheta) \cdot G(\varphi, \vartheta) d\Omega \tag{10.30}$$

where T_a is the noise temperature of the antenna, $G(\varphi,\psi)$ is the direction-dependent gain of the antenna (referred to the izotrophic antenna) and $d\Omega$ is the infinitesimal range of the space angle.

The real conditions are even more complicated because of additional electric effects which cause considerable noise increase. Noise temperature of an antenna located on the Earth surface is determined by the following components:

- Cosmic background noise: 2.78 K, physical origin is unknown, according to the cosmological theory of the 'big bang', the cooled remains of the once hot universe might be the possible reason.
- Galaxian noise: radio radiation of our galaxy
- Tropospherical noise: radio noise caused by the atmosphere.

- Noise caused by Earth surface.
- Noise caused by the 'near' celestial bodies (Sun, Moon).
- Noise caused by the antenna loss.

10.5. Signal to Noise Balance of Radio Communications

As a restrictive factor, noise plays an important role in telecommunications, broadcasting, radio-astronomy, radar and navigation systems. The signal-to-noise ratio, i.e. the power of the useful signal compared to the noise power is an essential qualitative factor of systems. Generally, the amplification is also the function of frequency, thus the noise power is

$$P_{\mathrm{n}} = \int_0^{\infty} G(f) \cdot S(f)\, \mathrm{d}f \qquad (10.31)$$

Practically, spectral power density of the noise can be assumed as constant in the used frequency range (white noise), i.e.

$$S(f) = N_0 \qquad (10.32)$$

Extracting the constant from the integral and introducing G_0 gain measured in the middle of the bandwidth

$$P_{\mathrm{n}} = N_0 \cdot G_0 \cdot \frac{1}{G_0} \cdot \int_0^{\infty} G(f)\mathrm{d}f \qquad (10.33)$$

On the right side, the term of the noise bandwidth

$$B = \frac{1}{G_0} \cdot \int_0^{\infty} G(f)\mathrm{d}f \qquad (10.34)$$

has been introduced which can be used as if the gain were constant over the whole frequency range.

In the following, let us determine the signal-noise balance of a radio communication. We will concentrate only to the so-called RF signal-to-noise ratio neglecting the gain or the loss caused by the demodulator. Let us suppose the following quantities as given: P_{t} transmitting power, G_{t} gain of the transmitting antenna, G_{r} gain of the receiving antenna, D transmitter-receiver distance, T resulting noise temperature of the receiver, L additional attenuation, B noise bandwidth, and λ wavelength of the transmitter.

It was shown in Chapter 9 that free-space attenuation between two anisotrophic antennas is

$$\frac{P_t}{P_r} = \frac{16 \cdot \Pi^2 \cdot D^2}{G_t \cdot G_r \cdot \lambda^2} \tag{10.35}$$

Since the effective input noise can be expressed as

$$P_\mathrm{n} = k{\cdot}T{\cdot}B, \tag{10.36}$$

the RF signal-to-noise ratio results in the following:

$$\left(\frac{S}{N}\right) = \frac{P_t \cdot G_t \cdot \lambda^2}{16 \cdot \Pi^2 \cdot D^2 \cdot L \cdot k \cdot B} \cdot \frac{G_r}{T} \tag{10.37}$$

It can be seen that signal-to-noise ratio is directly proportional to the factor G_r/T if all the other quantities are fixed. This factor is a basic parameter of a receiver station and it is usually given in dB/K (10 log of G_r/T).

10.6. Quantization Noise

From the physical point of view, information flow is not continuous but it is realized by elementary quanta of electrons or photons. If the signal is carried by the current then the minimal quantum is given by the charge of one electron, in the case of high frequency radiation the lower energy limit is $h{\cdot}f$.

In classical communication, thermal white noise played the most important role among the factors limiting sensitivity. Today's communication uses not only radio ranges but runs some five decades higher into the optical range which shows a rather quantized than wave character. Of course, the electromagnetic field, can neither be regarded as a continuous spatial wave or as a quantized flood of massive particles; both models reflect more or less just one side of the sides of the electromagnetic effect.

If the physical process of an information flow is quantized, it is a matter of chance how many elementary units will be observed over a selected time interval; the instantaneous value of the signal fixes only the expected value belonging to a certain number of elements. The actual number of elements varies accidentally around the expected value, the actual distribution can be described by the Poisson's distribution.

In communication systems, statistical fluctuation of incoming electrons or photons appears as noise. We are speaking about *quantum noise* when it is generated by photons while noise caused by current quantization is called *shot noise*. Since the physical basis of both the quantum noise and the shot noise is the same, quantum noise can be called shot noise of the photons.

For the shot noise

$$\overline{i^2} = 2{\cdot}B{\cdot}q{\cdot}I_\mathrm{o} \tag{10.38}$$

can be derived, where I_o is the expected value of the current. Let us assume that the signal of a light source is received by an ideal demodulator, i.e. each incoming photon generates exactly one electron so that the output current of a signal having the power P is

$$I_o = \frac{P \cdot q}{h \cdot f} \tag{10.39}$$

Interpreting the signal as a quantity proportional to the square of the expected value and taking the noise as the mean square of the fluctuation, signal-to-noise ratio can be written as

$$\left(\frac{S}{N}\right) = \frac{I_0^2}{\overline{i^2}} = \frac{P}{2 \cdot B \cdot h \cdot f} \tag{10.40}$$

As it will be shown in the discussion of the modulation theory, baseband bandwidth *B* will be doubled around the carrier frequency, so that eq. (10.41) might formally be interpreted as if the power spectral density of the quantum noise was

$$S(f) = h{\cdot}f \tag{10.41}$$

Although equation (10.41) might have been derived even more exactly, one thing must not be forgotten. There is an essential physical difference between the quantum noise and the thermal noise: the later is present even if the signal is switched off while the quantum noise is generated only in the presence of the signal and because of its 'granular' nature, its replacement by a noise with constant power spectral density is only approximative.

Control Questions

1 How can the 'optical' range and the 'radio' range be distinguished?
2. How can a transmission system be characterized from the point of view of noise?
3. How can the resulting noise figure be determined?
4. What is the definition of the noise bandwidth?
5. What is the reason of the quantum noise?

Exercises

1. Compute the resulting noise figure of a system consisting of a cable and a preamplifier for both the cable or the preamplifier is connected first.

Data: cable length is 15 m, specific attenuation of the cable 1 dB/m, gain of the preamplifier 20 dB and its noise figure 3 dB.

2. How much degraded is the noise temperature of an antenna if it is connected to the preamplifier through a cable which has 1 dB attenuation and temperature 290 K, and the original noise temperature of the antenna is 20 K? Give the value in dB.

3. How much increased is the noise temperature of an antenna with the beam angle of 5 degrees if the Moon appears in the received beam? Data of the Moon: temperature 300 °K, diameter 3476 km, distance from the Earth 384000 km.

References

[1] Ambrózy A.: Electronic Noise, Akadémia Kiadó, Budapest, 1982.
[2] Freeman R.L.: Radio System Design for Telecommunications, John Wiley & Sons, 1987.
[3] Morgan W.L., Gordon G.D.: Communication Satellite Handbook, John Wiley & Sons, 1989.

11. ANALOG MODULATION SCHEMES

The general block diagram of an analog modulation system is shown in Fig. 11.1. The modulated signal $s(t)$ is generated by a modulator using the modulating source signal $s_{\mathrm{m}}(t)$ to modulate a carrier of frequency f_{c}. The modulated signal is then passed through the channel where it is affected by different interferences and distortions (e.g. additive noise, linear and nonlinear distortion, etc.). The signal appearing at the channel output is demodulated and the resulting signal $s_{\mathrm{d}}(t)$ is processed by the sink.

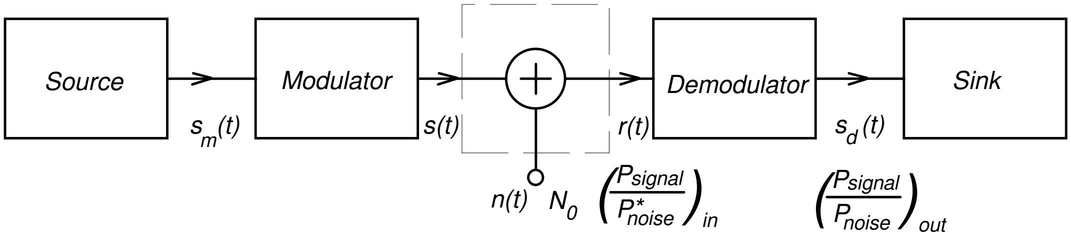

Source
Modulator
Demodulator
Sink
$s_m(t)$
$s(t)$
$r(t)$
$s_d(t)$
$n(t)$
N_0
$\left(\frac{P_{signal}}{P^*_{noise}}\right)_{in}$
$\left(\frac{P_{signal}}{P_{noise}}\right)_{out}$

Figure 11.1. General Block Diagram of Analog Modulation Systems with Additive Gaussian Noise in the Channel

If the signal is affected only by additive Gaussian white noise then the system quality is characterized by the signal-to-noise ratio of the signal $s_{\mathrm{d}}(t)$ which is defined as the ratio of the power of the useful signal to that of the noise. For the sake of simple comparison of different systems, let us define the signal-to-noise ratio at the demodulator input. Moreover, let it be defined so that it depends only on the power density of the signal and that of the noise but is not in direct relation to the total bandwidth of the modulated signal. For that purpose, let us introduce the so-called reference noise power: $P_n^* = 2f_M \cdot N_o / 2 = f_M \cdot N_o$ where f_{M} is the bandwidth of the modulating signal and N_{o} is the single side power density of the Gaussian white-noise.

Since sinusoidal signals play a dominant role in analog modulation, let us start our discussion with systems which use a *sinewave* as the carrier. Some general questions have to be answered first and the possible solutions are then presented.

A modulated sinewave can be expressed in the following general form:

$$s_{\mathrm{c}}(t)\text{€=€}a(t)\cdot\cos[\Theta(t)], \tag{11.1}$$

where $s_{\mathrm{c}}(t)$ is value of the modulated signal in time t, $a(t)$ is the instantaneous amplitude of the carrier, $\Theta(t)$ is the instantaneous phase of the carrier.

Since either the amplitude or the phase of the carrier -€or both of them€- may vary simultaneously, it is necessary to introduce the instantaneous values beside the time-average values normally used. The instantaneous value of the frequency (f_{i}) can then be defined as the time-derivative of the instantaneous phase:

$$f_{\mathrm{i}}\text{€=€}\frac{1}{2\pi}\frac{d\Theta}{dt}, \qquad \text{i.e.} \qquad \omega_{\mathrm{i}}(t)\text{€=€}\frac{d}{dt}\left[\Theta(t)\right] \tag{11.2}$$

It can be seen from eq. (11.1) that either $a(t)$ or $\Theta(t)$ or both can be modulated by the source signal. *Amplitude modulation* and *angle modulation* are the terms used to distinguish which parameter is modulated:

amplitude modulation: $a(t)$€≠€const., f_i€=€const., (11.3)
angle modulation: $a(t)$€=€const., f_i ≠€const. (11.4)

In the following we review the different sorts of amplitude modulation and the most important angle modulation schemes.

11.1. Amplitude Modulation (AM)

As given by the name, the *amplitude* of the AM signals carries the information, i.e. the modulating signal is encoded somehow into the amplitude function $a(t)$. Let us see first, how an AM signal can be described in the time and in the frequency domain. Since ω_i is constant, $\Theta(t)$ can be obtained from (11.2) by simple integration:

$$\Theta(t)=\int_{-\infty}^{0}\omega_i\ \Theta(t)=\int_{-\infty}^{t}\omega_i\cdot d\sigma=\omega_i\int_{0}^{t}d\sigma+\varphi,\qquad \omega_i=2\pi\cdot f_i, \qquad (11.5)$$

where φ is a constant which represents the phase in t€=€0. Since f_i is constant, it is equal to the average carrier frequency, i.e.:

Θ_AM€=€ω_c t€+€φ€ (11.6)

Substituting eq. (11.6) into (11.1), the general expression of the AM signal is

$s_\mathrm{AM}(t)$€=€$a(t)\cdot\cos[\Theta(t)]$€=€$a(t)\cdot\cos[\omega_\mathrm{c}\,t$€+€$\varphi]$ (11.7)

Since the initial phase of an AM signal is usually indifferent, let us simply suppose that φ€=€0, thus

$\mathrm{s}_\mathrm{AM}(t)$€=€$a(t)\cdot\cos(\omega_\mathrm{c}\,t)$€, (11.8)

which is the general time-domain representation of AM signals.

Suppose that the amplitude function $a(t)$ containing the modulating signal is a band-limited signal with the highest frequency f_M, i.e. the spectrum of the $a(t)$ function extends from ($-f_\mathrm{M}$) to ($+f_\mathrm{M}$) in the complex frequency domain (see part (a) of Fig. 11.2.)

Let $A(f)$ be the Fourier transform (spectrum) of $a(t)$ and let us examine how it is influenced by modulation. The Fourier transform of the modulated signal, $S_\mathrm{AM}(f)$, can be written as follows:

$$S_{AM}(f) = \int_{-\infty}^{+\infty} s_{AM}(t)\cdot e^{-j\omega\cdot t}\cdot dt = \int_{-\infty}^{+\infty} a(t)\cos(\omega_c t)\cdot e^{-j\omega\cdot t}\cdot dt =$$

$$= \frac{1}{2}\int_{-\infty}^{+\infty} a(t)\, e^{-j(\omega-\omega_c)t}\cdot dt + \frac{1}{2}\int_{-\infty}^{+\infty} a(t)\, e^{-j(\omega+\omega_c)t}\cdot dt. \qquad (11.9)$$

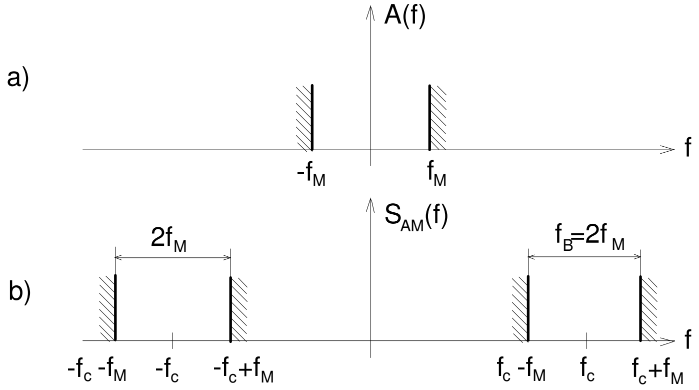


Figure 11.2. The Amplitude Function and the Modulated Signal in the Frequency Domain

Two integral expressions of eq.(11.9.) can be considered as if the spectrum $A(f)$ had been shifted along the frequency axis to $(+f_c)$ and $(-f_c)$ and its amplitude decreased by half of the original:

$$s_{AM}(f) = \frac{1}{2}A(f-f_c) + \frac{1}{2}A(f+f_c) \qquad (11.10)$$

The graphic form of eq. (11.10) is presented in part b) of Fig. 11.2. It is important to notice that AM is a *linear modulation* since the shape of the spectrum has been affected by linear operations, i.e. it has been shifted to $(-f_c)$ and $(+f_c)$ and multiplied by 0.5. It can be also seen that to avoid the spectrum aliasing, the carrier frequency must be at least the double of the maximum modulating frequency.

11.1.1. Sinewave Modulated AM Signals

Further properties of the AM modulation will be examined by using a simple cosine waveform as the modulating signal $s_m(t)$:

$$s_m(t) = U_m\cdot\cos(\omega_m t) \qquad (11.11)$$

To transmit the above signal by AM, the $a(t)$ has to include somehow the modulating signal $s_m(t)$. At first sight it seems obvious to make $s_m(t)$ equal with $a(t)$. Because of a practical reason let us choose, however, a more general relation:

$$a(t) \overset{\Delta}{=} U_c + s_m(t), \qquad (11.12)$$

where U_v is constant and represents the amplitude of the unmodulated carrier (when $s_\mathrm{m}(t)\equiv 0$). By substituting (11.11) into (11.12)

$$a(t) = \mathrm{U}_\mathrm{c} + \mathrm{U}_\mathrm{m}\cdot\cos(\omega_\mathrm{m}\, t) \qquad (11.13)$$

The simple form of the modulating signal enables us to examine the shape of the AM signal for different ratios of amplitudes U_c and U_m. Starting with eq. (11.8), the time function of the AM signal is

$$\begin{aligned} s_\mathrm{AM}(t) &= a(t)\cdot\cos(\omega_\mathrm{c}\, t) = \left(U_\mathrm{c} + U_\mathrm{m}\cdot\cos\left(\omega_\mathrm{m}\, t\right)\right)\cdot\cos\left(\omega_\mathrm{c}\, t\right) = \\ &= U_\mathrm{c}\cdot\cos\left(\omega_\mathrm{c}\, t\right) + U_\mathrm{m}\cdot\cos\left(\omega_\mathrm{m}\, t\right)\cdot\cos\left(\omega_\mathrm{c}\, t\right). \end{aligned} \qquad (11.14)$$

which can be rewritten as

$$s_\mathrm{AM}(t) = U_\mathrm{c}\cdot\cos\left(\omega_\mathrm{c}\, t\right) + \frac{U_m}{2}\cos\left[\left(\omega_\mathrm{c} + \omega_\mathrm{m}\right) t\right] + \frac{U_m}{2}\cos\left[\left(\omega_\mathrm{c} - \omega_\mathrm{m}\right) t\right] \quad (11.15)$$

The last equation is suitable to present the three basic types of amplitude modulation:

- AM-DSB (Double Sideband Amplitude Modulation): All the three components of eq. (11.15) are present in the signal. The bandwidth of this modulation is $f_\mathrm{B} = 2\cdot f_\mathrm{M}$, its characteristic term is the *modulation depth* defined as $m_\mathrm{a} = U_\mathrm{m}/U_\mathrm{c}$ which can change between 0 and 1. The vector diagram and the time and frequency representation of the AM-DSB signal are shown in Fig 11.3.

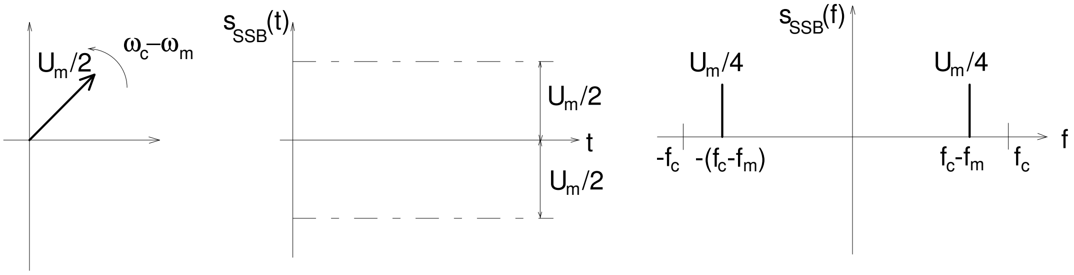


Figure 11.3. Vectorial Diagram and Time and Frequency Domain Representations of the AM-DSB Signal

- AM-DSB/SC (Double Sideband/Suppressed Carrier Amplitude Modulation): The first member of eq. (11.15) is eliminated (e.g. suppressed by a filter or by a balanced multiplier), i.e. carrier frequency is absent in the modulated signal. The bandwidth of the AM-DSB/SC signal is $f_\mathrm{B} = 2\cdot f_\mathrm{M}$, the vector diagram and the time and frequency representation are shown in Fig 11.4.

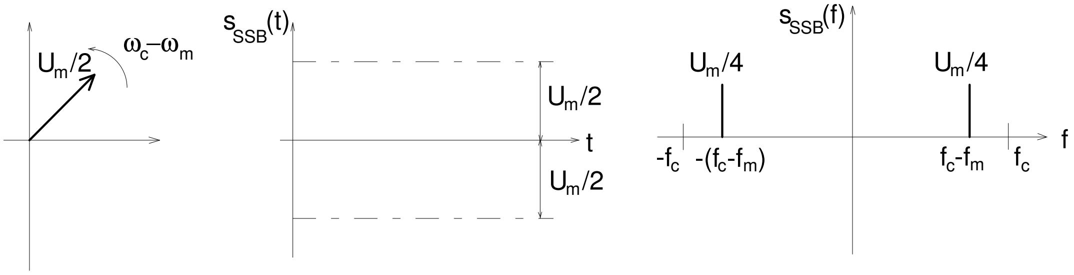


Figure 11.4. Vectorial Diagram and Time and Frequency Domain Representations of the AM-DSB/SC Signal

- AM-SSB/SC (Single Sideband/Suppressed Carrier Amplitude Modulation): Here the first and the second (or the third) member of eq. (11.15) is zero thus only components above (or under) the carrier frequency appear in the modulated signal. The bandwidth of the SSB signal is f_B€=€f_M, the vector diagram and the time and frequency representation are shown in Fig 11.5.

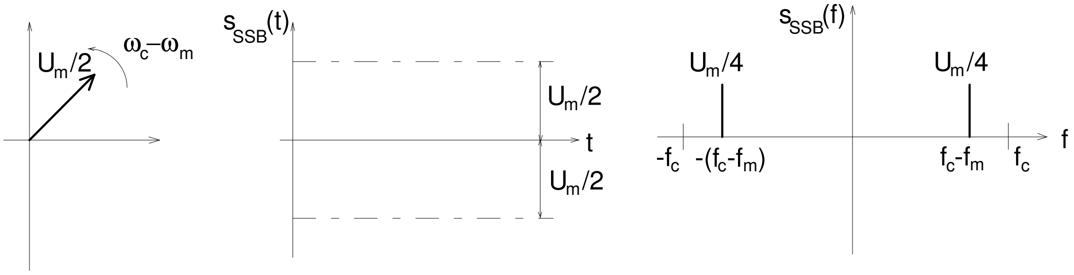


Figure 11.5. Vectorial Diagram and Time and Frequency Domain Representations of the AM-SSB/SC Signal

11.1.2. AM Signal Demodulation

AM signals are generally demodulated by product detectors (multipliers). AM-DSB is an exception since it can be demodulated also by the so-called envelope detector. In the following, the demodulation by multipliers is discussed.

Let us examine what the output of an ideal multiplier will be if one of its inputs is driven by an AM signal and the other by a sinewave of the same frequency as the carrier, shifted by φ in phase. The product of the two signals denoted as $s_d(t)$ is as follows:

$$\begin{aligned} s_d(t) &= s_{AM}(t)\cdot\cos(\omega_c\cdot t+\varphi) = a(t)\cdot\cos(\omega_c\cdot t)\cdot\cos(\omega_c\cdot t+\varphi) = \\ &= \frac{a(t)}{2}\cos(\varphi)+\frac{a(t)}{2}\cos(2\omega_c\cdot t+\varphi). \end{aligned} \tag{11.16}$$

Suppressing the second member of the sum by a filter, the desired baseband signal $a(t)$ is obtained almost exactly (regardless of the 0.5 factor and provided that φ€=€0).

Let us determine the S/N ratio at the demodulator output if the signal has been passed through an additive noisy channel. The signal at the demodulator input is

r(t)€=€s_{AM}(t)€+€n*(t), (11.17)

where $s_{\mathrm{AM}}(t)$ is the AM signal, $n^*(t)$ is that part of the Gaussian white-noise $n(t)$ with $N_o/2$ double-side power-density which falls into the range of the useful signal.

Since the bandwidth of an AM-DSB signal is $4{\cdot}f_{\mathrm{M}}$, the entire power of the $n(t)$ is $P_z=$€$4{\cdot}f_{\mathrm{M}}{\cdot}N_o/2$. It is known that $n(t)$ can be decomposed into modulation form as

$$n^*(t)\text{€}=n_c^*(t)\cdot\cos(\omega_c t)\text{€}+\text{€}n_s^*(t)\cdot\sin(\omega_c t), \tag{11.18}$$

where $n_c^*(t)$, $n_s^*(t)$ is the independent baseband Gaussian noise pair with double side power density N_o and bandwidth f_{M}.

Provided the demodulator works under noisy conditions also according to eq. (11.16), then (if φ€=€0):

$$s_{\mathrm{d}}(t)\text{€}=\text{€}r(t)\cdot\cos(\omega_c\, t)\text{€}=\text{€}a(t)\cdot\cos^2(\omega_c\, t)\text{€}+\text{€}n_c^*(t)\ \cos^2(\omega_c\, t)\text{€}+$$
$$+\text{€}n_s^*(t)\cdot\sin(\omega_c\, t)\cdot\cos(\omega_c\, t)\text{€}\overset{\Delta}{=}\frac{a(t)}{2}+\frac{n_c^*(t)}{2}, \tag{11.19}$$

where =€denotes the baseband part of the signal. Introducing the input reference signal-to-noise ratio:

$$\left(\frac{P_S}{P_N^*}\right)_{in}=\frac{\dfrac{\mathrm{M}\{a^2(t)\}}{2}}{f_M N_o}=\frac{\mathrm{M}\{a^2(t)\}}{2f_M N_o} \tag{11.20}$$

while at the demodulator output

$$\left(\frac{P_S}{P_N}\right)_{out}=\frac{\dfrac{1}{4}\mathrm{M}\{a^2(t)\}}{\dfrac{1}{4}\mathrm{M}\{n_o^{*2}(t)\}}=\frac{\mathrm{M}\{a^2(t)\}}{2f_M N_o}, \tag{11.21}$$

which means that in the case of AM-DSB the input reference S/N ratio is the same as the output S/N ratio. Obviously, this is true only if the entire power of the $a(t)$ function carries information (as it is the case with the AM-DSB/SC). Normally, the AM-DSB uses only a part of the amplitude function $a(t)$ (see eq. 11.6), thus the output S/N ratio decreases if the modulation depth is reduced.

Let us note that for AM-SSB/SC signal the input reference and the output S/N ratios of the demodulator are equal:

$$\left(\frac{P_S}{P_N}\right)_{out} = \left(\frac{P_S}{P_N^*}\right)_{in} = \frac{\frac{\mathrm{M}\left\{a^2(t)\right\}}{4}}{f_M N_o} = \frac{\mathrm{M}\left\{a^2(t)\right\}}{4 f_M N_o}, \tag{11.22}$$

which means that only half of the S/N ratio of AM-DSB can be achieved by the AM-SSB/SC.

11.2 Angle Modulation

As earlier defined, in the case of angle modulation the amplitude of the carrier is constant while the instantaneous frequency -and so the instantaneous phase- is changing with the modulating signal (see eq. 11.4). Similarly as for the AM systems, the relation between the modulating signal $s_{\mathrm{m}}(t)$ and the frequency (or the phase) of the modulated signal has to be determined first. Obviously, the simpler the relation is, the easier it is to modulate and to demodulate the signal. Since the linear relation is the simplest, two types of angle modulations are used: *frequency modulation* and *phase modulation*. The frequency modulation (FM) is defined by

$$f_{\mathrm{i}}\text{€}= \frac{1}{2\pi}\frac{d\Theta}{dt}\text{€}=\text{€}k_{\mathrm{FM}}\cdot s_{\mathrm{m}}(t)\text{€}+\text{€}f_{\mathrm{c}}\text{€}, \tag{11.23}$$

while the phase modulation (PM) is defined by

$$\Theta(t)\text{€}=\text{€}k_{\mathrm{PM}}\cdot s_{\mathrm{m}}(t)\text{€}+\text{€}\omega_{\mathrm{c}}\, t \tag{11.24}$$

where k_{FM} and k_{PM} are constants with different units and f_{c} is the frequency of the unmodulated carrier (also constant). Using eq. (11.1) and (11.24) the general form of the FM signal is as follows:

$$s_{\mathrm{FM}}(t)\text{€}=\text{€}a(t)\cdot\cos(\Theta(t))\text{€}=\text{€}U_c\cos\left(2\pi\int_0^t f_i\cdot d\sigma\right)\text{€}=\text{€}U_c\cos\left[2\pi\left(f_c\cdot t + k_{FM}\int_0^t s_m(\sigma)d\sigma\right)\right]\text{€}=\text{€}$$

$$= U_c\cos\left(\omega_c\cdot t + 2\pi\cdot k_{FM}\int_0^t s_m(\sigma)d\sigma\right)$$

while the same for the PM signal:

$$s_{\mathrm{PM}}(t)\text{€}=\text{€}a(t)\cdot\cos(\Theta(t))\text{€}=\text{€}U_{\mathrm{c}}\cdot\cos\big(\omega_{\mathrm{c}}\, t+k_{\mathrm{PM}}s_{\mathrm{m}}(t)\big) \tag{11.25}$$

Instead of the general form of the modulating signal,

$$s_{\mathrm{m}}(t)\text{€}=\text{€}U_{\mathrm{m}}\cdot\cos(\omega_{\mathrm{m}}\, t) \tag{11.26}$$

will be used for further discussion, similarly as for the AM. Substituting (11.26) into (11.25):

$$s_{FM}(t) = U_c \cos\left(\omega_v t + 2\pi \cdot k_{FM} \int_0^t U_m \cos(\omega_m \sigma)\mathrm{d}\sigma\right) \text{€}=$$

$$=\text{€}U_\mathrm{c}\cos\left[\ \ \omega_\mathrm{v} t\text{€}+\frac{k_{FM}\, 2\pi U_m}{\omega_m}\sin(\omega_\mathrm{m} t)\ \ \right]$$

$$=\text{€}U_\mathrm{c}\cos\left[\ \ \omega_\mathrm{v} t\text{€}+\text{€}\frac{k_{FM} U_m}{\mathrm{f}_m}\text{€}\sin(\omega_\mathrm{m} t)\ \ \right]$$

$$s_\mathrm{PM}(t)\text{€}=\text{€}U_\mathrm{c}\cdot\cos\big(\omega_\mathrm{v} t+k_\mathrm{PM}U_\mathrm{m}\cdot\cos(\omega_\mathrm{m} t)\big)\text{€}. \qquad (11.27)$$

So the information represented by the modulating signal is encoded into the FM signal in the form of frequency changes of the carrier around a central value (f_v). The amplitude of the modulating signal corresponds to the maximum difference or the *deviation* of the carrier frequency from f_v while the frequency of the modulating signal is equal to the frequency the instantaneous carrier frequency is changing around the average f_v. Let us denote the maximum frequency deviation as f_D, i.e.:

$$f_\mathrm{D}\text{€}=\text{€}k_\mathrm{FM}\cdot U_\mathrm{m}\text{€}. \qquad (11.28)$$

and then substituting (11.26) into (11.23):

$$f_\mathrm{p}=k_\mathrm{F}\cdot s_\mathrm{m}(t)+f_\mathrm{c}=k_\mathrm{FM}\cdot U_\mathrm{m}\cos(\omega_\mathrm{m} t)\text{€}+\text{€}f_\mathrm{c}=f_\mathrm{c}+f_\mathrm{D}\cdot\cos(\omega_\mathrm{m} t). \qquad (11.29)$$

which shows that f_D is the *maximum* deviation from the unmodulated center frequency f_c. (To distinguish the two deviations, the instantaneous value is denoted as f_d).

Similarly to the definition of the m_a, the ratio $k_\mathrm{FM}\cdot U_\mathrm{m}/f_\mathrm{m}$ in eq. (11.27) is called the FM modulation factor and is denoted as m_f, i.e.:

$$m_\mathrm{f}\text{€}=\text{€}\frac{k_{FM} U_m}{\mathrm{f}_m}\text{€}=\text{€}\frac{f_D}{f_m} \qquad (11.30)$$

and the product $k_\mathrm{PM}\cdot U_\mathrm{m}$ is called the PM modulation factor and is denoted as m_p, i.e.:

$$m_\mathrm{p}\text{€}=\text{€}k_\mathrm{PM}\cdot U_\mathrm{m}\text{€}. \qquad (11.31)$$

Both the m_f and the m_p have clear physical meanings which can be read out from eq. (11.27): they represent the maximum phase deviation of the modulated carrier with respect to the phase of the unmodulated one. For that reason they are also called phase deviations. The time-domain waveform of an FM signal modulated by a sinewave is shown in Fig. 11.6.

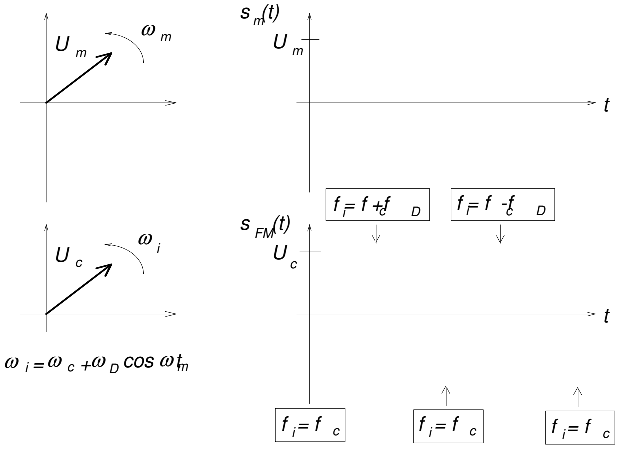


Figure 11.6 Time Domain Representations of the Modulating and the FM Signal

It can be shown by a detailed analysis that the bandwidth of an FM signal is: $f_B = 2 \cdot \alpha \cdot f_m$€, where

$$\alpha = \begin{cases} & \text{if} \quad m_f \langle 0{,}1 \\ m_f & \text{if} \quad m_f \rangle 10 \quad \textit{and} \\ + m_f + \sqrt{m_f} & \quad \textit{otherwise} \end{cases} \tag{11.32}$$

and f_m is the frequency of the modulating sinewave.

11.2.1. FM Signal Demodulation

To demodulate an FM signal, a circuit with output voltage proportional to the instantaneous frequency of the input signal is needed. The amplitude response of the ideal FM demodulator is shown in Fig. 11.7. Circuits with such characteristics are called *frequency discriminators*.

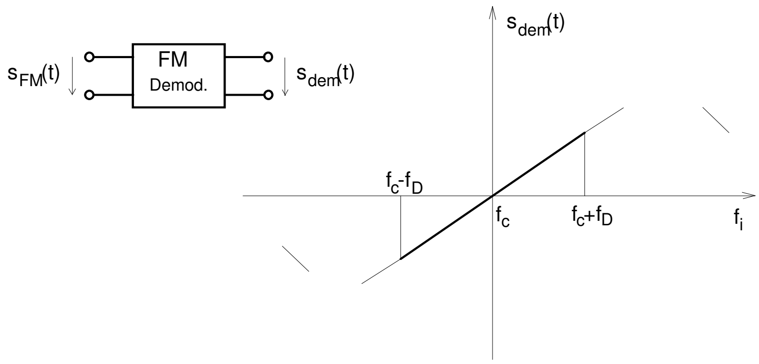


Figure 11.7. The Ideal FM Demodulator

Suppose the input of a frequency discriminator is driven by an FM signal given by eq. (11.25):

$$s_{FM}(t) = U_c \cos\left(\omega_c t + 2\pi\cdot k_{FM}\int_0^t s_m(\sigma)d\sigma\right) \tag{11.33}$$

If the discriminator is ideal, the output signal will be

$$s_{\mathrm{dem}}(t)\text{€}=\text{€}k_{\mathrm{discr}}\cdot 2\pi f_{\mathrm{i}}\text{€}=\text{€}k_{\mathrm{disc}}\cdot 2\pi\left(f_{\mathrm{c}}+\text{€}k_{\mathrm{FM}}\cdot s_{\mathrm{m}}(t)\right) \tag{11.34}$$

since the instantaneous frequency is determined by the time-derivative of the argument of the cosine function ($\Theta(t)$).

The ideal discriminator can be approximated by deriving the signal and then demodulating the output by an envelope detector. Namely, if the FM signal is derived in time

$$\frac{d}{dt}\left[s_{FM}(t)\right] = U_c\cdot 2\pi\cdot\left(f_c + k_{FM}\cdot s_m(t)\right)\cdot\sin\left[2\pi\left(f_c\cdot t + k_{FM}\int_0^t s_m(\sigma)d\sigma\right)\right] \tag{11.35}$$

the result is an FM signal, the *amplitude* of which changes proportionally to the modulating signal $s_{\mathrm{m}}(t)$. Demodulation of this AM-FM signal by an envelope detector will lead to a voltage proportional with the amplitude, i.e. with the modulating signal.

If noise is also present, then the S/N ratio at the demodulator input can be determined as follows:

$$\left.\left(\frac{P_s}{P_n^*}\right)\right|_{in} = \frac{U_c^2/2}{f_M\cdot N_o} = \frac{U_c^2}{2f_M\cdot N_o} \tag{11.36}$$

The S/N ratio at the demodulator output can be calculated for low-level noise as follows: From eq. (11.34), power of the useful signal -provided that k_{d}€=€1- is obtained from the following formula:

$$P_s\big|_{out} = (2\pi)^2 k_{FM}^2\cdot M[s_M^2(t)] \tag{11.37}$$

Suppose that a band-limited Gaussian white noise (see eq. 11.18) is added the modulated signal (eq. 11.33). The sinusoidal baseband component of this noise ($n_s^*(t)$) produces a phase-noise or the so-called *jitter* which can be defined by

$$\varepsilon(t) = \frac{n_s^*(t)}{U_c}; \qquad \text{if} \qquad |\varepsilon(t)|\text{€}<<\text{€}1 \tag{11.38}$$

This 'phase' can be determined by computing the phase change caused by the noise given by eq. (11.25). The derivative of such 'phase-noise' is then added to the useful signal and the resulting sum can be considered as the instantaneous 'frequency-noise':

$$\frac{d}{dt}\left[\varepsilon(t)\right] = \frac{1}{U_c}\frac{d}{dt}\left[n_s^*(t)\right] \tag{11.39}$$

The entire baseband noise power in the frequency range f_M is given by the following expression:

$$P_n\big|_{out} = \frac{1}{U_c^2}\int_{-\omega_M}^{+\omega_M}\frac{N_o}{2\pi}\omega^2 d\omega = \frac{N_o}{2\pi}\frac{2}{U_c^2}\frac{\omega_M^3}{3} = (2\pi)^2\frac{2}{3U_c^2}N_o f_M^3 \qquad (11.40)$$

so that for the output S/N ratio

$$\left(\frac{P_s}{P_n}\right)_{out} = \frac{(2\pi)^2 k_{FM}^2\cdot M\left[s_m^2(t)\right]}{(2\pi)^2\frac{2}{3U_c^2}N_o f_M^3} = 3\frac{k_{FM}^2\cdot M\left[s_m^2(t)\right]}{f_M^2}\cdot\frac{U_c^2}{2f_M N_o} \qquad (11.41)$$

is obtained. It can be seen from eq. (11.41) that the S/N ratio at the demodulator output is

$$\left(\frac{P_s}{P_n}\right)_{out} = 3\cdot\left(\frac{P_s}{P_n^*}\right)_{in}\frac{k_{FM}^2\cdot M\left[s_m^2(t)\right]}{f_M^2}, \qquad (11.42)$$

where $k_{FM}^2 \cdot \{s_m^2(t)\}$ is the square of the frequency deviation (see eq. (11.28)). So if the radio frequency power is kept constant and the noise level is small, S/N ratio can be improved by increasing the frequency deviation of the FM.

Control questions

1. Give the general structure of analog modulation systems and define the amplitude and the angle modulation.
2. Draw the vector diagram, the time function and the spectrum of the AM-DSB, AM-DSB/SC and AM-SSB/SC signals in the case of sinusoidal modulation.
3. What is the reference noise power?
4. Determine the signal-to-noise ratio of the AM-DSB modulation system for the Gaussian white-noise of one sided power density N_o.
5. What is the frequency and the phase deviation and how is the modulation factor defined in FM and PM systems?
6. How can the bandwidth of an FM signal modulated by a sinewave be approximately computed?
7. Determine the signal-to-noise ratio of the FM modulation system for the Gaussian white-noise of one sided power density N_o.

Exercises

1. Draw the spectrum of an AM-DSB signal (f_c = 20 kHz), if

   (a) the time function of the modulating signal is

   $s_m(t) = U_{m1}\cos(\omega_1 t) + U_{m2}\cos(\omega_2 t)$,

   U_c = 1V, U_{m1} = 0,2V, U_{m2} = 0,5 V,

   $\omega_1 = 2\pi\cdot 10^3$ rad/sec, $\omega_2 = 2\pi\cdot 2\cdot 10^3$ rad/sec

   (b) the Fourier-transform of the modulating signal is

   $$S_m(f) = \begin{cases} C\left(1-\dfrac{|f|}{f_M}\right) & ;\text{if } |f| \le f_M, \\ 0 & ;\text{otherwise,} \end{cases}$$

   U_c = 1 V ; $C = 0{,}5\cdot 10^{-3}\left[\dfrac{V}{Hz}\right]$; f_M = 1 kHz

2. Draw the spectrum of the AM-DSB/SC and of the AM-SSB/SC signal, if

   (a) the time function of the modulating signal is

   $s_m(t) = U_{m1}\cos(\omega_1 t)+U_{m2}\cos(\omega_2 \mathrm{t})$

   (b) the spectrum of the modulating signal is

   $$S_m(f) = \begin{cases} C\left(1-\dfrac{|f|}{f_M}\right) & ;if\ |f| \le f_M, \\ 0 & ;\text{otherwise,} \end{cases}$$

and the data are the same as in Exercise 1.

3. Determine the output signal-to-noise ratio of the AM-DSB if the data are as follows:

   f_M = 3 kHz; $N_o = 10^{-6}\left[\dfrac{W}{Hz}\right]$;

   $a(t) = U_c+U_m\cos(\omega_m t)$; U_m = 0,5 V,

   (a) U_c = 1 V; (b) U_c = 0 V

   The reference resistance is 1 Ω.

4. What are the values of ω_D and m_p of a PM system if

   $s_{PM}(t) = U_c\cos(\omega_c t + cU_m\cos(\omega_m t))$ and

   U_c = 1 V; $\omega_c = 2\pi\cdot 10^6\dfrac{rad}{\text{sec}}$; $c = 0{,}1\dfrac{1}{V}$; $\mathrm{U_m}$ = 1 V; $\omega_m = 2\pi\cdot 10^3\dfrac{rad}{\text{sec}}$

5. What is the maximum phase deviation and m_f of an FM system if

   $s_{FM}(t) = U_c\cos(\omega_c t+c\cdot U_m\sin(\omega_m \mathrm{t}))$ and

   U_c = 1 V; $\omega_c = 2\pi\cdot 10^6\dfrac{rad}{\text{sec}}$; $c = 0{,}2\ \dfrac{1}{V}$; U_m = 1 V; $\omega_m = 2\pi\cdot 10^2\ \dfrac{rad}{\text{sec}}$

6. What is the approximate value of the bandwidth of an FM system if its parameters are as follows:

   $s_{\mathrm{FM}}(t)$=$U_{\mathrm{v}}\cos\left(\omega_{\mathrm{v}}t+\frac{k_{FM}U_m}{f_m}\sin(\omega_{\mathrm{m}}t)\right)$; k_{FM}= $10^3\,\frac{Hz}{V}$; $U_{\mathrm{m}} = 1$ V

   (a) $f_{\mathrm{m}} = 10^3$ Hz;   (b) $f_{\mathrm{m}} = 10$ Hz

7. Draw the ratio of the bandwidth of a sinewave-modulated FM signal to the frequency of the modulating signal as the function of m_{f}.

8. Compute the output signal-to-noise ratio of a sinewave-modulated FM system if:

   $s_{\mathrm{m}}(t) = U_{\mathrm{m}}\cos(\omega_{\mathrm{m}}t)$ and

   $U_{\mathrm{c}} = 1$ V ; $f_{\mathrm{m}} = f_{\mathrm{M}} = 10^3$ Hz ; $N_{\mathrm{o}} = 10^{-5}\frac{W}{Hz}$; $U_{\mathrm{m}} = 1$ V ; $k_{\mathrm{FM}} = 10^3\,\frac{Hz}{V}$;

   The reference resistance is 1 Ω.

References

[1] Lucky,€R.W.€-€Salz, J.€-€Weldon,€E.J.: Principles of Data Communication, McGraW-Hill, New York, 1968.

[2] Schwartz,€M.: Information Transmission, Modulation and Noise, McGraw-Hill, New York, 1990.

[3] Papoulis,€A.: Signal Analysis, McGraw-Hill, New York, 1977.

12. DIGITAL MODULATION SCHEMES

Digital modulation systems can be divided into two big groups. The baseband signals are transmitted in channels with lowpass character while in the case of modulated signals the channel has a bandpass character.

A general block diagram of the digital modulation systems is shown in Fig. 12.1.

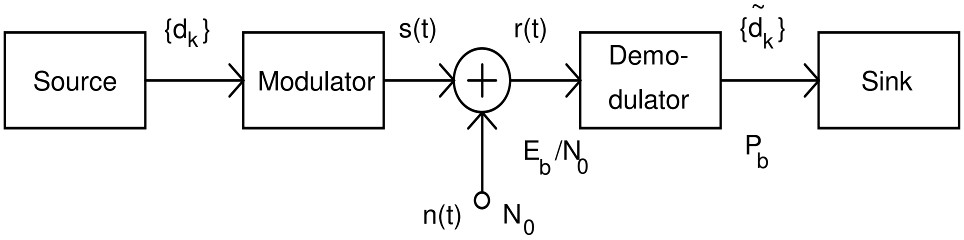


Fig. 12.1 General block diagram of the digital modulation systems

Symbols d_k transmitted by the source at time T feed the modulator which generates the modulated signal $s(t)$. In the channel the signal is exposed to various disturbing effects (additive noise, linear and nonlinear distortion, etc.). The distorted signal $r(t)$ comes into the demodulator which produces a series of estimated symbols $\tilde{d}_k$.

The system quality can be characterized by the bit-error ratio P_b related to the series of estimated symbols. If only Gaussian white noise is added to the signal in the channel, the bit-error ratio P_b depends firstly on the signal-to-noise ratio related to the signal $r(t)$ which in fact is determined by the ratio E_b/N_o where E_b is the signal energy carried by one bit and N_o is the single side power density of the additive Gaussian white noise.

In the following the possibilities of the modulation by a digital signal will be examined similarly as it was done for the analog modulating signal in the previous chapter. It has to be kept in mind, however, that true-to-form transmission of digital signals is not as important as the possible smallest probability of error of reproducing the original digital data from the transmitted signal. Anyway, if the ratio of mistaken decisions can be kept low then the shape of the transmitted signals is indifferent.

12.1 Baseband Modulation

In the case of digital *baseband modulation*, the information can be encoded in the amplitude, in the duration or in the position of the impulse. Accordingly, they are called the Pulse Amplitude Modulation (PAM), Pulse Duration Modulation (PAM) and Pulse Position Modulation (PPM) systems. PAM is the system with the most efficient use of the power and bandwidth thus we will concentrate to this kind of pulse modulation.

A general block diagram of the baseband PAM system is shown in Fig. 12.2.

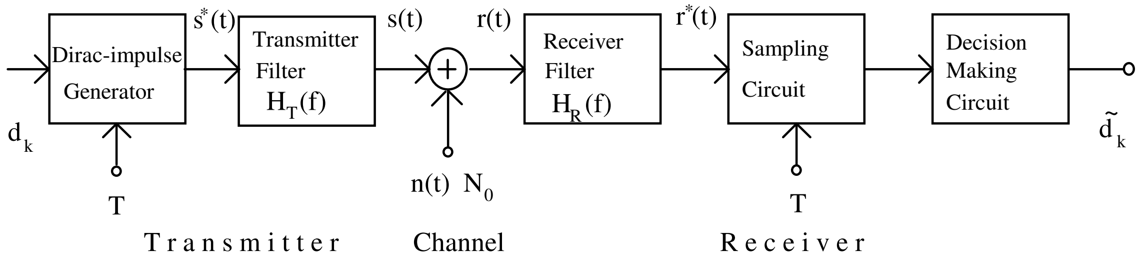


Fig. 12. 2. Block Diagram of Baseband Digital Transmission Systems

Source symbols $\{d_k\}$ control the Dirac-impulse generator at each time T. Suppose that the symbols can have M different values within the set $\{-(M-1),...,-3,-1,1,3,...,(M-1)\}$ and that the output signal of the Dirac-impulse generator is as follows:

$$s^*(t) = \sqrt{PT}\sum_{k=-\infty}^{+\infty} d_k \delta(t-kT) \qquad (12.1)$$

where P is power of the signal. This signal controls the transmitting filter at the output of which the modulated signal is obtained in the following form:

$$s(t) = \sqrt{PT}\sum_{k=-\infty}^{+\infty} d_k h_T(t-kT) \qquad (12.2)$$

where $h_T(t)$ is the impulse response of the transmitting filter with the frequency response $H_T(f)$. Additive Gaussian white noise added to the signal results in

$$r(t)€=€s(t)€+€n(t) \qquad (12.3)$$

appearing at the input of the receiver. It is important to note that the noise $n(t)$ has infinite power thus the signal $r(t)$ has to be filtered in the receiver. This is carried out by the receiving filter with frequency response $H_R(t)$ which simultaneously shapes the signal passed to the sampling and decision circuits.

The signal $r(t)$ appearing at the output of the receiving filter has to fulfil two following requirements:

- the power of the filtered additive noise ($n^*(t)$) has to be as low as possible,
- the signal samples appearing periodically every time T has to be dependent on only one input symbol.

The first condition can be satisfied by matching the transmitting and the receiving filters i.e. by choosing their transfer functions so that $H_R(f)€=€H_T^*(f)$. (Matched filters best suppress the noise without significant distortion of the useful signal.) To satisfy the second condition, the transfer function $H(f)$ of the entire transmission chain given by

$$H(f)€=€H_T(f)\cdot H_R(f) \qquad (12.4)$$

must be chosen in such a way that the impulse response $h(t)$ has special properties, namely: among the samples of $h(t)$ taken every time T only one sample may differ from zero, all the others must be zero. Such a system is said to have zero *Inter Symbol Interference* (ISI). The selection of the baseband signal shape is discussed in the next chapter.

12.1.1. Baseband Shaping of Impulses

Let us examine the case when the channel is ideal, i.e. free of noises so that the output signal $r^*(t)$ is determined only by the transfer functions of the filters. Let $h(t)$ be the impulse response corresponding to the transfer function $H(t)$ and let us determine the signal at the output of the receiving filter in the absence of noise:

$$r^*(t)=\sqrt{PT}\sum_{k=-\infty}^{+\infty} d_k h_T(t-kT)\ . \qquad (12.5)$$

According to the so-called Nyquist criterium, the ISI-free condition can be satisfied with impulse responses satisfying the above condition:

$$\sum_{k=-\infty}^{+\infty} H\left(f+\frac{\ell}{T}\right)=T, \qquad if \quad |f|\le\frac{1}{2T} \qquad (12.6)$$

where in the case of matched filters:

$$H(f)=H_R(f)\cdot H_T(f)=|\ H_R(f)|^2=|\ H_T(f)|^2$$

The Nyquist criterion defined by eq. (12.6.) can also be expressed in graphic form as shown in the upper part of Fig. 12.3. As can be seen the frequency response of a Nyquist filter is specified at frequency 1/2*T* as 50 per cent of the maximum value, furthermore it is point-symmetrical to the so-called Nyquist point. Since the Nyquist criterium specifies only one point and the symmetry of the frequency response, the number of possible responses is practically infinite.

To make the choice easier, the Fig. 12.3 b) shows the shapes of the pulses in the time domain corresponding to the same parameter α (a parameter characteristic for the slope of the spectrum rounding). It can be noticed that the greater is the slope of the rounding, the greater are the ripples, the 'overshots' of the time functions. It can also be seen that if the bandwidth is less than 1/2*T* then the ISI cannot be eliminated even theoretically since the spectrum cannot be shaped as point-symmetrical. On the other hand it is not necessary to have the bandwidth greater than 1/*T* since -€because of the point-symmetry€- the spectrum does not exceed this value.

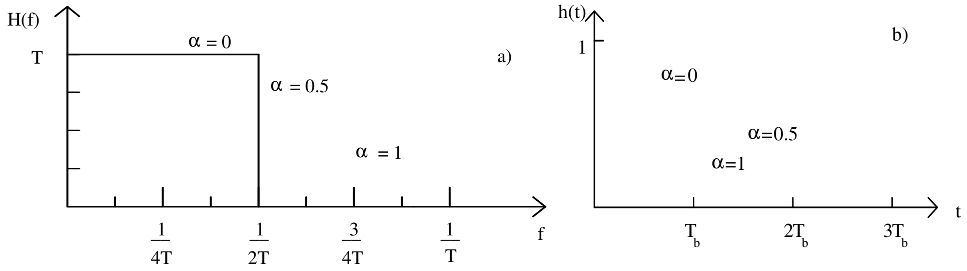


Fig. 12.3 Frequency (a) and Time-Domain (b) Representations of the Nyquist Criterium

To sum up the conclusions made for the received pulses: A minimum of 1/2*T* bandwidth is required for the zero-ISI data transmission. If a greater bandwidth is available then data can be transmitted with a considerably greater reliability by proper shaping of the impulse *h*(*t*), i.e. by using the band up to 1/*T*. In practice the so-called 'raised cosine' function with soft transition of *H*(*t*) from *T* to zero is used (α=1) since it is 'gently sloping' and it is an acceptable compromise between the bandwidth increase (with respect to 1/2*T*) and the time function free of overshots. (Overshots are dangerous since the probability of wrong decisions may increase if the timing is not proper in the receiver.)

12.1.2. The Error Ratio

In the case of PAM the error ratio is determined by the amount of the noise power within the resulting signal $r^*(t)$. Let us suppose that the signal is binary (d_k may be +1 or -1) and that the $H(f)$ satisfies the Nyquist criterium. In the absence of noise the signal appearing at times *nT* at the output of the receiving filter might have the value $\pm\sqrt{PT}$ (see Fig. 12.3). The Gaussian white noise *n*(*t*) filtered by the receiving filter is denoted as $n^*(t)$. The expected value of the filtered noise is zero, the standard deviation is (using also Fig. 12.3.) as follows:

$$\sigma_{n^*}^2 = \int_{-\frac{1}{T}}^{\frac{1}{T}} \frac{N_0}{2} \left| H_R(f) \right|^2 df = \int_{-\frac{1}{T}}^{\frac{1}{T}} \frac{N_0}{2} H(f)\, df = \frac{N_0}{2} \int_{-\frac{1}{T}}^{\frac{1}{T}} H(f)\, df = \frac{N_0}{2} \qquad (12.7)$$

Probability density functions of the received binary symbols disturbed by the noise are shown in Fig 12.4.

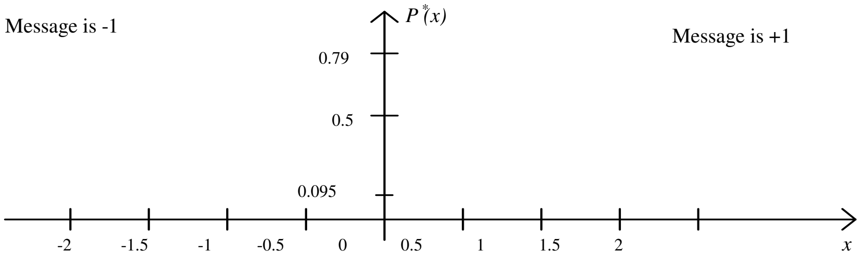


Fig. 12.4 Probability Density Function of the Signal at the Output of the Receiver Filter if $\sigma^2 = 0.25$ and $\sqrt{PT} = 1$

As it can be seen, instead of ideal values (+1, -1) any signal may practically appear with a certain chance (distribution) at the output of the receiving filter. To decide whether the sent signal was a +1 or a -1, the sign of the sampled signal $r^*(t)$ has obviously to be examined. The probability of error can then be given as

$$P_b = P\{r^*(nT) \geq 0 \mid d_n = -1\} \tag{12.8}$$

which can be calculated from the following expression:

$$\begin{aligned} P_b &= \int_0^\infty \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x+\sqrt{PT})^2}{2\sigma^2}\right)dx = \int_{\sqrt{PT}}^\infty \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{y^2}{2\sigma^2}\right)dy = \\ &= \int_{\frac{\sqrt{PT}}{\sigma}}^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)dy = Q\left(\frac{\sqrt{PT}}{\sigma}\right) = Q\left(\sqrt{\frac{2PT}{N_o}}\right) \end{aligned} \tag{12.9}$$

where Q is the so-called Gaussian error function defined as follows:

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)dy \tag{12.10}$$

Since $P \cdot T = E_b$ is the energy of the useful signal per one bit, eq. (12.9) is usually given as

$$P_b = Q\left(\sqrt{\frac{2E_b}{N_o}}\right) \tag{12.11}$$

The above function is shown in double logarithmic scale in Fig. 12.5.

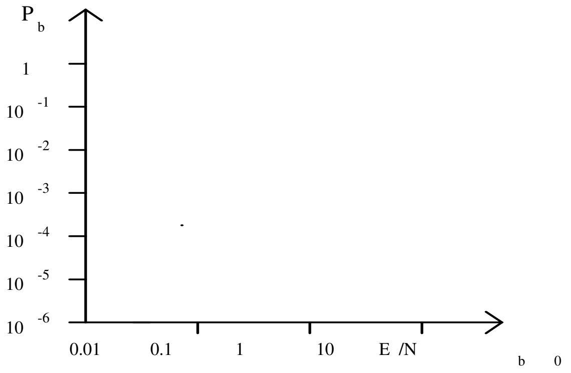


Fig. 12.5 Probability of Bit Error of the PAM System as the function of the Signal-to-Noise Ratio

12.2. Carrier Modulations

Four important *carrier modulation* methods suitable for the transmission of binary data through a bandpass channel are shown in Fig. 12.6.(a)...(d). Figure (a) shows the discrete amplitude modulation with the corresponding binary data shown below. As it can be seen, in this modulation the carrier is switched on if the data are equal to 1 and switched off if the data are equal to 0. This modulation is called the *amplitude shift keying* and is abbreviated as ASK.

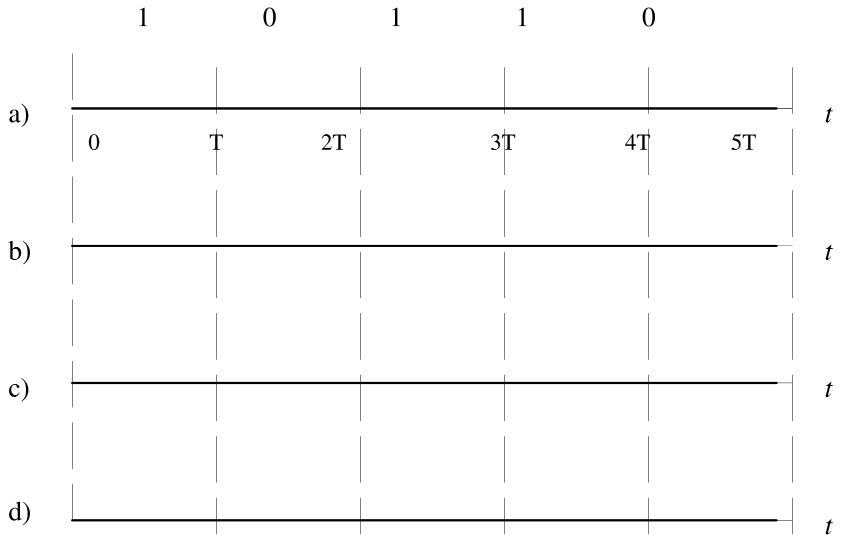


Fig. 12.6. Waveshapes of Digitally Modulated Carrier

Similarly, the carrier frequency can also be switched between two different values corresponding to the binary data. This modulation shown in Figure (b) is called the *frequency shift keying* and is abbreviated as FSK. As it can be seen, it is the frequency of the FSK signal which changes in the rhythm of the binary data.

Figure (c) shows a signal both the amplitude and the frequency of which are constant; it is only the phase which changes according to the modulation. Therefore it is quite logical to call this modulation as *phase shift keying* and abbreviate it as PSK.

Finally, Figure (d) shows the case when the sinusoidal carrier is modulated in amplitude by such a discrete PAM signal which was previously `smoothed' i.e. filtered as shown in Chapter 12.2. Among the four procedures presented, this modulation -which in fact is AM-DSB- requires the minimum bandwidth although the equipment generating, transmitting, and demodulating such signals are very complex. On the contrary, ASK, FSK and PSK can be implemented with simpler and much cheaper devices. The price we have to pay for it is the greater bandwidth and the greater required transmitting power. If the bandwidth is not the main aspect of the design then the digital procedures can be well used since they have relatively good parameters, e.g. high noise immunity against different interferences.

12.2.1. Structure of Binary Modulation Systems

The essential task of the ASK, FSK and PSK receivers is to recognize binary data, i.e. to be able to make difference between the $s_1(t)$ and the $s_2(t)$. The quality of the receiver is determined by the probability of error and the structure of the receiver is considered optimal if the probability of error is minimal. In this chapter the structure of an optimal receiver suitable to receive ASK, FSK and PSK signals will be presented.

If the input noise of a receiver is a Gaussian white noise then it can be shown that the most important part of the receiver is the matched filter. It can be also shown that this filter can be realized by a correlator consisting of a multiplier and an integrator. The receiver is synchronized to the input signal, i.e. its local oscillator generates a sinewave whose frequency and phase are exactly the same as those of the input signal (This feature is called *coherence*).

Binary ASK, FSK and PSK signals can be demodulated also by non coherent methods. Although their quality is not optimal, noncoherent receivers are much simpler and thus widely used in low-speed data transmission systems.

The general block diagram of a binary data transmission system using digital modulation is shown in Fig. 12.7.

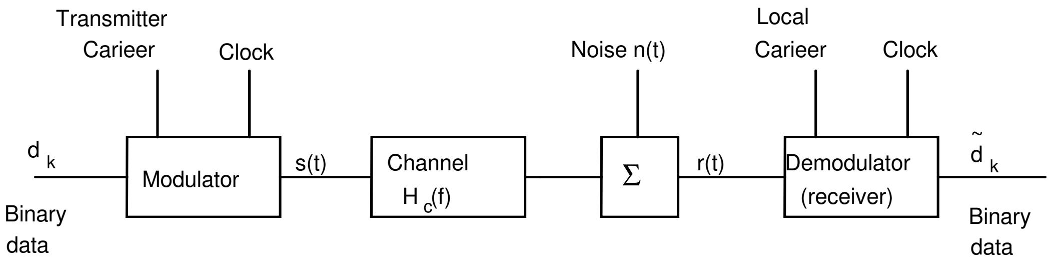


Fig. 12.7 Block diagram of Modulated Binary Data Transmission System

The input signal is a sequence of bits d_k, the bitrate (the speed) of which is $1/T$ and the duration of the bits -the time-slot- is T. At the time-slot k the output signal of the demodulator depends on the value of the d_k. The signal $s(t)$ generated by the modulator in the time-slot k is one of the two possible waveforms, $s_1(t)$ or $s_2(t)$ shifted to the time of the k_{th} bit. Thus the $s(t)$ is a stochastic process determined as follows:

$$s(t) = \begin{cases} s_1(t - kT), & if \quad d_k = -1 \\ s_2(t - kT), & if \quad d_k = \ \ 1 \end{cases} \qquad (12.12)$$

provided that $k\cdot T \le t < (k+1)\cdot T$.

The duration of both signals is T and both are finite energy signals since both $s_1(t)$ and $s_2(t) \equiv 0$ if t lies outside the period 0...T; inside the period, however, the time integral of the square of both functions is finite.

Table 12.1. Signalling Waveforms of ASK, PSK and FSK

$s_1(t), \ 0 \le t \le T$	$s_2(t), \ 0 \le t \le T$	modulation
0	$A \cdot \cos(\omega_c \cdot t)$ $or \quad A \cdot \sin(\omega_c \cdot t)$	Amplitude shift-keying (ASK)
$-A \cdot \cos(\omega_c \cdot t)$ $or \quad -A \cdot \sin(\omega_c \cdot t)$	$A \cdot \cos(\omega_c \cdot t)$ $or \quad A \cdot \sin(\omega_c \cdot t)$	Phase shift-keying (PSK)
$A \cdot \cos[(\omega_c - \omega_d)t]$ $or \quad A \cdot \sin[(\omega_c - \omega_d)t]$	$A \cdot \cos[(\omega_c + \omega_d)t]$ $or \quad A \cdot \sin[(\omega_c + \omega_d)t]$	Frequency shift-keying (FSK)

The signal shape depends on the actual modulation as summarized in Table 12.1. The output signal of the modulator goes through a bandpass channel the transfer function of which is $H_c(f)$. Suppose the channel is ideal, i.e. the transmission is free of distortions, except a finite time delay and a zero-mean, stationary Gaussian noise with known double-side spectral power density ($N_o/2$). The received signal can thus be written as follows:

$$r(t) = \begin{cases} s_1(t - kT - \tau) + n(t) & if \quad d_k = -1, \\ s_2(t - kT - \tau) + n(t) & if \quad d_k = \ \ 1, \end{cases} \qquad (12.13)$$

provided that $k\cdot T \le t < (k+1)\cdot T$, where τ is the time delay of the transmission which can be considered to be zero without restricting the generality.

The block diagram of the receiver is shown in Fig. 12.8. The task of the receiver is to decide which one of the $s_1(t)$ and $s_2(t)$ functions is present at the input. The actual receiver consists of a filter, a sampler and a threshold detector (comparator). First, the signal $r(t)$ goes through a filter and is sampled at the end of each bit time-slot. The sample is then compared with a previously determined threshold and the generated bit is decoded as 1 or -1 depending on the relation between the $r^*(t)$ and the threshold.

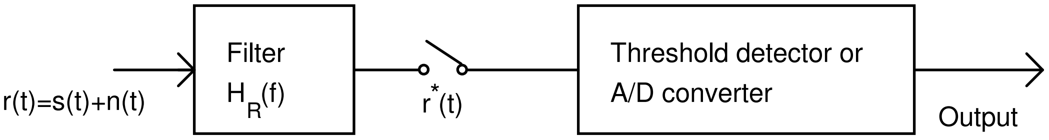


Fig. 12.8. Receiver of the Binary Data Transmission System

Owing to the noise, the receiver makes sometimes false decisions. The probability of errors depends on the signal power and spectral power density of the noise at the receiver's input, on the signalling frequency (bitrate) and on the parameters of the receiver, such as the transfer function $H_R(t)$ of the filter and the threshold value (similarly to the PAM type transmission).
(a) Binary ASK Modulation

Being very simple to realize, the binary ASK was used in wireless telegraph communication (spark telegraph) at the beginning of this century. Although ASK has been almost entirely replaced by the much more effective FSK and PSK, it is instructive to get acquainted with the most important parameters of the ASK since the ASK is a very clear model of the binary signalling systems. The actual form of the ASK signal -defined generally by eq. (12.12)- is $s_2(t)\in=\in A\cdot\cos(\omega_c t)$, if $0\leq\in t\in\leq\in T$ and $s_1\in=\in 0$. Suppose that the 2π multiple of the carrier frequency is $\omega_c\in=\in 2\cdot\pi\cdot n/T$ where n is an integer.

The time function of the modulated signal can be written as

$$s(t)\in=\in d(t)\cdot[A\cdot\cos(\omega_c t)] \qquad (12.14)$$

where $d(t)$ represents a baseband pulse train. Suppose that $d(t)$ is a random squarewave with period T. It follows from eq. (12.14) that the ASK signal can be generated by a multiplier which means that the $d(t)$ signal can be used to switch the carrier on and off. The relation between the spectral power density of the modulated signal and that of the pulse train $d(t)$ is as follows:

$$G_s(f)\in=\frac{A^2}{4}\cdot\left(G_d(f\text{-}f_c)+G_d(f+f_c)\right). \qquad (12.15)$$

The time function $d(t)$ is a random binary signal with only two levels: 0 and 1. Using the autocorrelation function it can be shown that the spectrum of the modulated signal is as follows:

$$G_s(f)=\frac{A^2}{16}\left[\delta(f-f_c)+\delta(f+f_c)+\frac{\sin^2[\pi T(f-f_c)]}{\pi^2T^2(f-f_c)^2}+\frac{\sin^2[\pi T(f+f_c)]}{\pi^2T^2(f+f_c)^2}\right] \qquad (12.16)$$

The graphic form of the above spectrum is shown in Fig. 12.9. The Dirac functions at frequencies f_c and $-f_c$ represent the carrier while the $(\sin x)/x$ components are the sidebands. As can be seen, the bandwidth is theoretically infinite but the spectrum decreases rapidly with the distance from the carrier frequency; neglecting the small components a finite bandwidth can be specified. It can be shown that if the spectrum is limited so that less than 5 per cent of the whole signal energy is that of the neglected part, the bandwidth of the ASK signal is about 3/*T*.

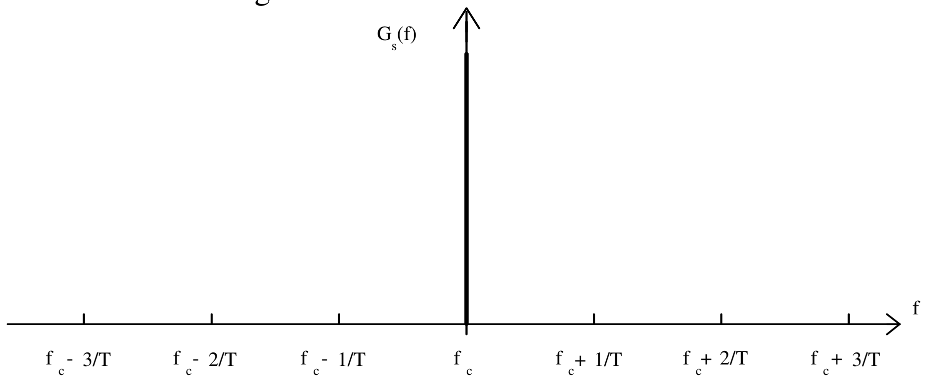


Fig. 12.9 Spectral Power Density Function of a Random Binary ASK Signal

The bandwidth can be reduced even more if the modulating signal is not a square wave but the *d*(*t*) is smoothed before the modulation. So the bandwidth decreases to about 2/*T* thus for such an ASK signal a channel with bandwidth 2/*T*...3/*T* is appropriate.

The ASK signal can be demodulated both in a coherent and a non-coherent way. The coherent method requires the information about frequency and the phase of the input signal and in principle consists of an integration and a decision while the non-coherent one uses an envelope detector.

(b) Binary PSK Modulation

PSK is a discrete phase modulation in which two signals of opposite phases: $s_1(t)€=€-A\cdot\cos(\omega_c t)$ and $s_2(t)€=€A\cdot\cos(\omega_c t)$ are assigned to the binary data -1 and +1. As usually, the modulated PSK signal *s*(*t*) can be written as

$$s(t)€=€d(t)\cdot[A\cdot\cos(\omega_c t)] \qquad (12.17)$$

where *d*(*t*) is a random binary sequence of -1's and 1's with period *T*. It is interesting to notice that the only difference between the ASK and the PSK signal is that if the data are -1, the ASK multiplies the carrier by zero while the PSK multiplies it by -1. It can be shown that the spectral power density of the PSK signal is

$$G_s(f)€=€\frac{A^2}{4}€\left(G_d(f\text{-}f_c)€+€G_d(f\text{+}f_c)\right)€, \qquad (12.18)$$

where

$$G_d(f)€=€\frac{\sin^2(\pi f\cdot T)}{\pi^2 f^2T^2}, \qquad (12.19)$$

Comparing eqs (12.19) and (12.16), the spectral power densities seem to be similar. The only difference is that PSK does not contain the Dirac functions which means that no discrete spectral component is present at the carrier frequency. It is also a logical conclusion that the bandwidth of the PSK is the same as for the ASK.

The PSK signal, however, can be demodulated only by a coherent demodulator, i.e. the carrier generated in the receiver has to be synchronized in frequency and in phase to the carrier of the input signal. As can be exactly shown, PSK requires half of the power (-3€dB) required by the (also coherent) ASK for the same probability of error.

(c) Binary FSK Modulation

FSK is used mainly in low-speed data transmission systems. The FSK receiver can be realized without coherent demodulation since two different frequencies can be detected by quite simple circuits. The generation of the FSK signal is also simple. Nevertheless it has to be mentioned that as far as the power requirement and the bandwidth are concerned, the efficiency of the FSK is not as good as that of the PSK.

Two shapes of the binary FSK signal corresponding to +1 and -1 data are as follows:

$$s_1(t) = A\cdot\cos(\omega_c t - \omega_d t), \qquad \text{and} \qquad s_2(t) = A\cdot\cos(\omega_c t + \omega_d t), \qquad (12.20)$$

where ω_d is the frequency deviation. Thus the information transmitted by an FSK signal is given by its actual frequency.

The binary FSK is an FM signal with constant amplitude and continuous phase. One of the possible mathematical descriptions of such a signal is given by the following formula:

$$s(t) = A\cdot\cos\left(\omega_c t + \omega_d \int_{-\infty}^{t} d(\tau)\,d\tau + \varphi\right), \qquad (12.21)$$

where $d_k(t)$ is a random binary pulse train with the value +1 if d_k=1 and -1 if d_k=-1. The instantaneous frequency of the binary FSK signal is -according to the definition- the time derivate of the phase of the signal $s(t)$, i.e.:

$$f_p = \frac{d}{dt}\left(\omega_c t + \omega_d \int_{-\infty}^{t} d(\tau)\,d\tau + \varphi\right) = \omega_c + \omega_d\cdot d(t). \qquad (12.22)$$

Since $d(t)=\pm 1$, the instantaneous frequency has also only two values: $\omega_c \pm \omega_{d.}$

The analysis of the digital FM signal is rather complex. Instead of a detailed discussion only the results of the calculations are given in Fig. 12.10. If the frequency deviation is small (less than 25 per cent in comparison with the carrier frequency) then the main part of the spectrum is concentrated around the carrier frequency (which is the reciproque value of the bitrate T). The bandwidth is in the order of $2/T$ Hz which corresponds to the value of the PSK. If the deviation, however, increases above 25 per cent, then two peaks are shifted towards the signalling frequencies (f_c+f_d) and (f_c-f_d) thus the bandwidth is greater than $2/T$ Hz. If the deviation is even greater yet (above 75 per cent), then the spectrum looks like two interlaced ASK spectra with two carrier frequencies at about (f_c+f_d) and (f_c-f_d).

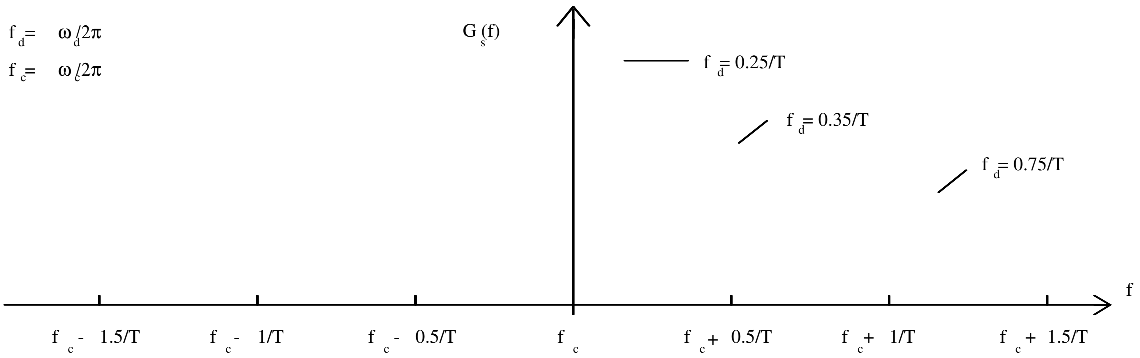


Fig. 12.10. a) Spectral Power Density Function of the Binary FSK Signal

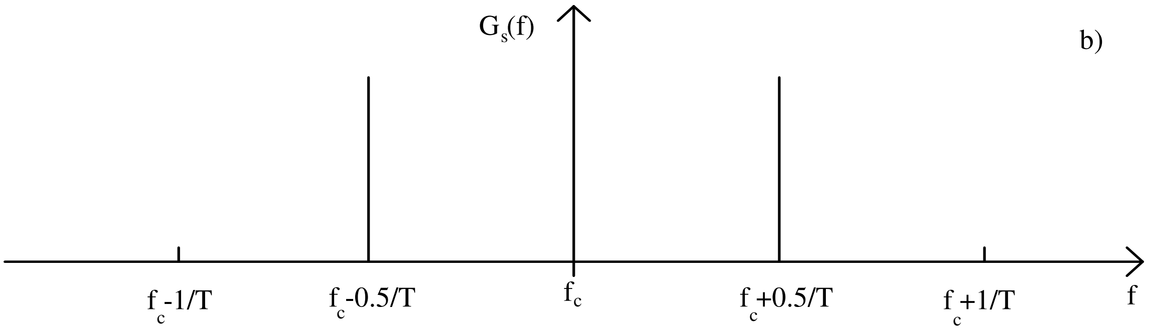


Fig. 12.10. b) Spectrum of binary FSK signal when $2f_d = 1/T$

Further, if the frequency deviation is such that $2{\cdot}f_{\mathrm{d}} = m/T$ where m is an integer, then the spectrum contains also two discrete spectrum lines as shown in Fig. 12.10.b. It can be said generally that the bandwidth of an r signal is greater than either that of the ASK or of the PSK.

As it was mentioned above the FSK signal given by eq. (12.21) is continuous in phase, i.e. the signal actually keyed is of the same phase as it was before keying.

Control Questions

1. What does the general structure of the baseband digital modulation systems look like?. Define the term of matched filtering in a channel with Gaussian white noise.
2. What is the condition of ISI-free transmission in a baseband digital modulation system?
3. How can the probability of error of a baseband binary digital modulation system be determined?
4. What are the characteristic types of digital carrier modulating schemes? Draw the specific waveforms of these schemes.
5. Draw the general block diagram of digital carrier modulation systems.
6. Give the spectral power density of the output signals of binary ASK, PSK and FSK systems.

Exercises

1. Determine the frequency response of the receiving filter for the baseband binary digital modulation to have an ISI-free transmission, if

$$H_T(f) = \begin{cases} & \mathit{if} \quad |f| < \dfrac{1}{T} \\ & \mathit{otherwise.} \end{cases}$$

2. Determine the error ratio of a binary baseband digital transmission if $H_T(f)$€=€$H_R(f)$;

$$P = 10^{-9}\ \text{W}; \qquad T = 10^{-3}\ \text{s}; \qquad N_o = 4 \cdot 10^{-11}\ \frac{W}{Hz} \qquad \text{and}$$

$$H(f) = \begin{cases} & \mathit{if} \quad |f| < \dfrac{1}{2T} \\ & \mathit{otherwise.} \end{cases}$$

3. Draw the spectral power density of the output signal of a binary PSK system if A€=€1 V ; T€=€10^{-3}€s and the reference resistance value is 1€Ω.

References

[1] Lucky, R.W. - Salz, J. - Weldon, E.J.: Adatátvitel, Mûszaki Könyvkiadó, Budapest, 1973.

[2] Proakis, J.G.: Digital Communication, McGraw-Hill, New York, 1983.

[3] Viterbi, A.J.: Principles of Coherent Communication, McGraw-Hill, New York, 1966.

Abbreviations

ASK	Amplitude Shift Keying
BER	Bit Error Ratio
FSK	Frequency Shift Keying
ISI	Intersymbol Interference
PAM	Pulse Amplitude Modulation
PDM	Pulse Duration Modulation
PPM	Pulse Position Modulation

13. CHANNEL ALLOCATION

When we discussed the *properties of information sources* in the previous chapters, we also determined the bandwidth required to transmit the information through a telecommunication channel. We know that for *analog sources*, the *highest spectral component* determines the frequency range occupied in the *baseband* while different *modulations* used for transposition of the signal spectra can significantly extend this range (e.g. AM-DSB requires twice as large bandwidth as that of the baseband and FM may use many times as much). For *digital* sources, the required bandwidth is the function of the bitrate, the spectrum-shaping coding and digital modulation methods. Efficiency of a given method is usually measured on the base *of unity bandwidth bit rate* (bit/s/Hz). *Typical values* are between 0.5 (for simple methods) and 5 (for complex, multi-level modulation procedures).

Earlier we have seen that because of physical reasons the frequency range of *wirebound* communications is limited in frequency. In *radio systems*, important legal regulations play also role beside the technical aspects. The bandwidth of *optical cables* is by many orders higher than that of the first two. Nevertheless, bandwidth or transfer capacity of telecommunication channel is generally much greater than the bandwidth required by the individual source signals. There are many cases when signals of several sources of information are to be transmitted together through a common channel. Methods used for solving problems of this kind are called multiplexing procedures.

13.1. Frequency and Time Division Multiplexing

Common channel can generally be divided in the *frequency* or in the *time* domain. (There is also a third method, called *code division* which will be discussed latter.) Frequency division multiplexing is abbreviated as FDM, similarly the short for time division multiplexing is TDM.

Let us overview these techniques on simple but frequently used procedures applied in *speech transmission*. When using FDM, baseband spectra of different speech signals are shifted to a higher frequency by single side AM-modulation and placed side by side to a common range. Details of such an FDM procedure for a base group containing 12 voice-grade channels is shown in Fig. 13. Fig. 13.1.a) shows the baseband of the analog speech signal. Fig. 13.1.b) shows the symbolic notation of a standardized telephone voice-grade channel limited to the frequency range from 300 Hz to 3400 Hz. Ramp-shaped notation of the spectrum shown in the figure has nothing to do with its actual shape, it might just as well be represented by a rectangle, however this notation is well suitable to show how the spectrum position varies in consequence of different frequency transpositions (modulations). Fig. 13.1.c) shows the spectrum allocation of 12 channels in the 60 ... 108 kHz range.

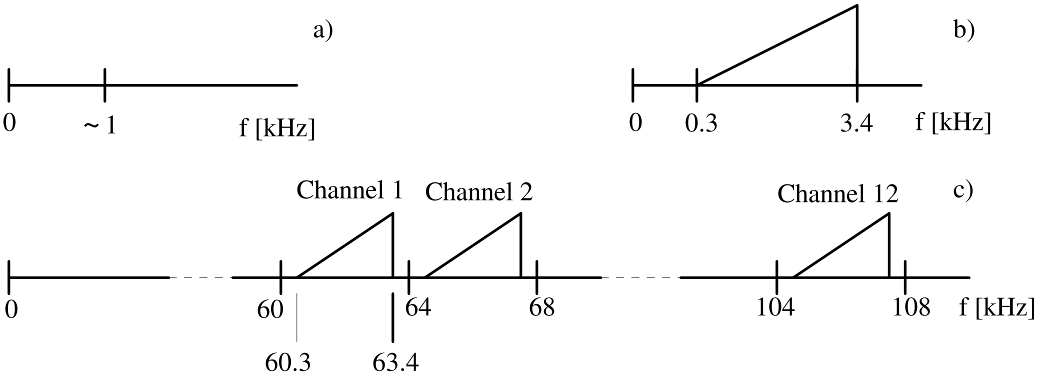


Fig. 13.1 Frequency Division Multiplexing of Voice Channels

FDM signal is composed according to the block diagram of Fig. 13.2.a). Each pair of multiplier and filter shifts one of the channels up to the appointed band. For instance, 60 kHz sinewave carrier $c_1(t)$ of the Channel 1 is cut off by a lowpass filter together with the lower sideband of the AM-DSB signal appearing at the multiplier output. Decomposition of the FDM signal is shown on Fig. 13.2.b). Each input filter selects one channel which is 'shifted back' by a multiplier and a lowpass filter to the baseband.

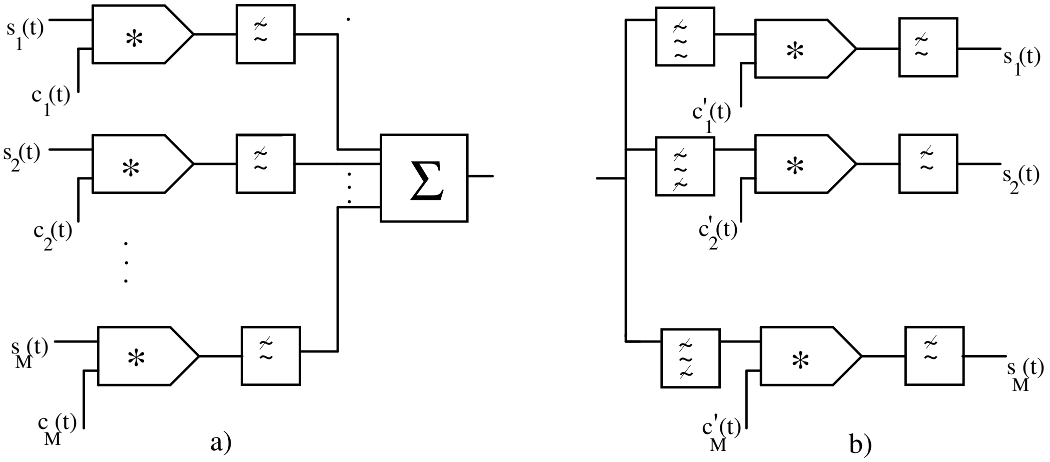


Fig. 13.2 FDM Signal Multiplexing and Demultiplexing

In the case of time division multiplexing we make use of the fact that the data rate can be much greater in the channel that the data rate of a digitized speech signal. Remember that telecommunication standards fixed 8 kHz as sampling frequency and 8 bit resolution for each sample, i.e. source data rate is 64 kbit/s. If 32 such sources are put together as it is done in primary data rate TDM generally used in telecommunication networks, the resulting data rate is 2048 kbit/s.

Summing up, frequency and time division multiplexing methods are used if more source signals are simultaneously present at the same point of a telecommunication system and they are collected to form a single FDM or TDM signal to be transmitted in one common broadband channel. However, the multiplexing of a common channel is needed also in cases when the source signals are not available at the same point (at the multiplex input). Methods described above can be used even in this case and they are denoted as FDMA and TDMA (Frequency-Division Multiple Access/Time-Division Multiple Access).

The practical use of the FDMA is limited since the division of a broadband channel into narrow subchannels -especially in radio systems- is not always possible because of legal and technical reasons and the application of individual narrowband transmitters and receivers can be expensive.

In TDMA, each individual station uses the same broadband channel but only in the time slot allocated for the station. If there is no information to be transmitted, the slot remains empty. If there is more information than what can be transmitted in a single time slot then the excess can be transmitted only in the following slot(s). Time division control may be centralized or distributed. Further details of TDMA systems are given in Chapter 18.

13.2. Code Division

Code-Division Multiple Access (CDMA) or Spread Spectrum Multiple Access (SSMA) is based on a special modulation technique called spectrum spreading. In the one of the main versions of the *spectrum spreading* technique, a transmitter-receiver pair uses a specific l bit long binary codeword and its inverse to represent "0"-s and "1"-s of the binary source (l is typically 16, 32 or 64). This procedure is called direct sequence spread spectrum.

In another main version called *frequency hopping* spread spectrum, sine-bursts of different frequencies are selected from a given set and inserted into the time-slots accordingly to a special code series. (The frequency hopping can take place slowly, i.e. the same frequency is transmitted during more time-slots or quickly, i.e. the frequency changes several times in one time-slot.)

In both cases, bandwidth of the modulated signal is significantly spread over a wider spectrum so that the transmitted power is spread to the same extent. Originally, spread spectrum modulation was developed for the protection of telecommunication systems against noise, since the signal spread over a wide frequency range cannot be disturbed by a signal with power concentrated into a narrower band or very high power would be required for the efficient disturbance.

As the basic principle of the spread spectrum multiple access, different code-series are assigned to different sources so that they can operate in the same frequency range using either the direct-sequence or the frequency hopping method. If the codeword length is l, 2^l different codewords can theoretically be assigned to, practically much less codewords are used (only those which can be "well distinguished", see chapter 7.)

So the spread spectrum technique can be used for multiplexing and for multiple access as well. As a characteristic example of CDMA, let us mention the part of Radio LAN-s (Local Area Network). Here, spectral density of the relatively wideband signal is kept at a low value so that several systems spread over a greater area can operate in the same frequency band without disturbing each other.

Let us go back now to the time division multiple access to discuss some of the efficient channel allocation techniques based on the TDMA methods.

13.3. Dynamic Channel Allocation. Polling Methods.

An obvious drawback of the so called *static TDMA* methods discussed above is that the time slot assigned to one station (user) is occupied even if the user is idle. Assuming that it is possible to exchange a message about the user's status between the TDMA system controller and the user, the time slot of the quiescent user can be used by the next user etc. Such a questioning of the user is called *polling* and time gained in this way is shortened by the time needed for the exchange of message.

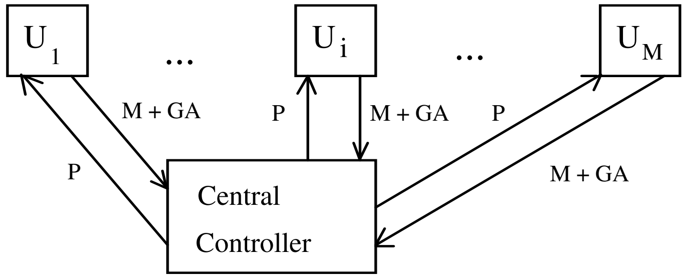


Fig. 13.3. Roll-Call Polling

One possible arrangement for polling is shown in Fig. 13.3. If the control messages P (poll) and GA (go ahead) are short in comparison to the useful message (M) and the propagation times between the central controller and the stations are also short, this system called "roll-call polling" works well. If the propagation time is not negligible, it is more advantageous to use the so-called "hub-polling" shown in Fig. 13.4. Here the time lost by the propagation can be much smaller since the right for go-ahead is passed directly from station to station.

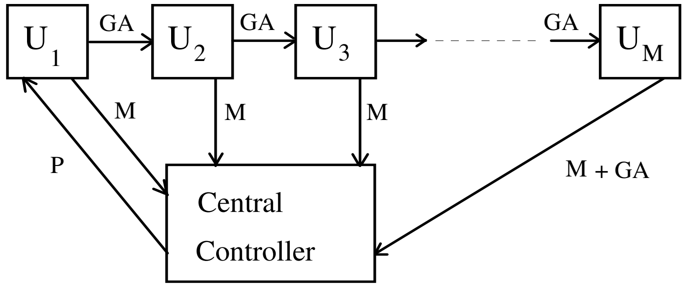


Fig. 13.4 Hub Polling

Polling methods are widely used in telecommunication systems. In a terminal network which is the oldest and most known such an application, local or remote terminals are connected to a controller which polls the terminals and passes the terminal messages towards a central equipment. Polling is also used in local area networks where this type of access is called *token-protocol.* 'Token' is the term used for short control messages the stations are passing to hand over the right of transmission. Central controller in not used in such a system, the stations fix up specific rules on the base of distributed algorithms, e.g. who has the token first or who starts when the token is 'lost' (faulty). Token is the principle of bus and ring local networks which will be discussed later.

Procedures discussed so far (FDMA, TDMA, CDMA, polling) belong to *scheduled* access methods. There is another group of so called *random* access procedures, although the term *free* access would be more adequate to express the principle of the method which will be discussed in the following.

13.4. Random Access Time Division Procedures

The principle of the random access based on the observation that the activity of the majority of the users is discontinuous, i.e. the duty cycle of their operations is small. The term 'duty cycle' has been borrowed from electronics to illustrate that the users actually engage the communication channel just for a small fraction of time while the majority of time is spent on the preparation of the messages. As an example, let us see the *dialog management* of an interactive terminal using 1200 bit/s (relatively low) data rate. Suppose the length of the message to be one row of text, program or any other data, the time needed for the transmission of the message is about 0.5 s while its preparation (typing in, checking, etc.) takes about 0.5-1 min., thus the duty cycle is in the order of 10^{-2}.

Communication between such discontinuous users flows on two common channels at most, one of them being of multiple access and the other a data broadcasting channel or just one multiple access/data broadcasting channel. (Data broadcast is like the program broadcast so that the transmitted information is received by all user stations although the message may be addressed only to one of them.) In the following part, the most important applications of effective and flexible multiple channel access will be discussed.

It is often possible in practical communication networks to create a common channel with multiple access, and even more, there are cases where this is an obvious solution. Let us overview now the most important cases and let us sum up the most important features of the corresponding multiple access and data broadcasting channels. The three most characteristic examples are:

- satellite system used as repeater,
- terrestrial radio network,
- local cable network.

Geostationary satellite is an obvious solution for telecommunication among user population widely spread over a certain area of the Earth (see Fig. 13.5.)

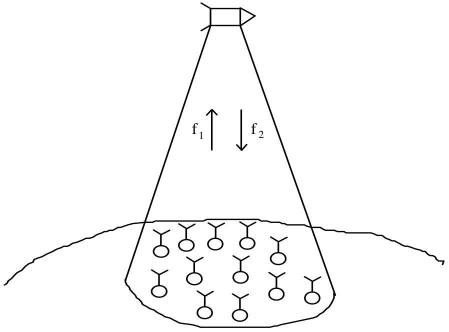


Fig. 13.5 Satellite Channel with Multiple Access

Satellite stations are working as repeaters, i.e. they receive users' messages on carrier frequency f_1 and transmit them back on frequency f_2. One of the specific features of the satellite channel is that everybody can listen to its own transmission so that he is automatically informed whether the packet (message) has successfully arrived to its destination or whether it collided during the multiple access with the packets of other users and it was therefore damaged. (The packet may of course be damaged by noise however this cannot be concluded from the backward information.) Another specific feature is that the propagation time is great in comparison with the packet time. The entire (two-way) propagation time is between 0.24-0.27 s. Supposing the channel data rate to be 32 kbit/s and taking a packet of 1000 bits, the propagation time is almost ten times longer than the packet time.

Terrestrial radio networks are usually built of one central (base) station and several users (often mobile) as is shown on Fig. 13.6. In the direction from users to the base station, there is a multiple access radio channel operating at f_1, while in the opposite direction a data broadcast channel on f_2 is used. Unlike in satellite systems, the users are not automatically provided with information about the success or fail of the transmission of their packets, this must be confirmed by the base station e.g. by so called positive acknowledgement.

The other characteristic is that the propagation time is short as compared to the packet time. Let us show a numerical example: suppose we want to transfer a 1000 bit packet to a distance of 32 km with 9600 bit/s data rate (which is a considerably high rate used in telephone networks and radiotelephone channels, as well). The propagation time for that distance will be 10^{-4} s which is about three decades less than the 0.1 s packet time. This may be important when the network topology provide the users to listen to each others transmission and thus making use of such an information. Methods of sensing the transmission (so called carrier sense) will be discussed in chapter 13.4.2.

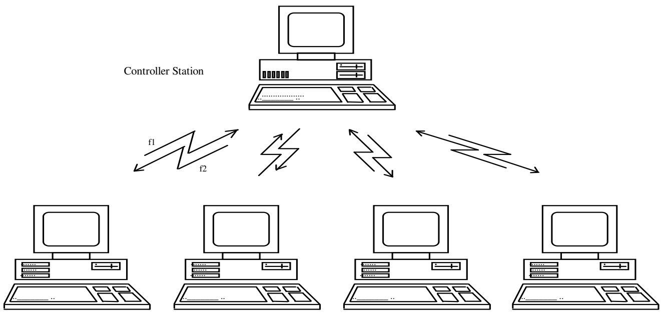


Fig. 13.6 Terrestrial Radio Data Transmission System

The third example is taken from the local area networks. A bus-structured local network is shown on Fig. 13.7. Here, the communication channel is a coaxial cable terminated at both ends with matching resistors. Stations are connected to the cable via line drivers and receivers.

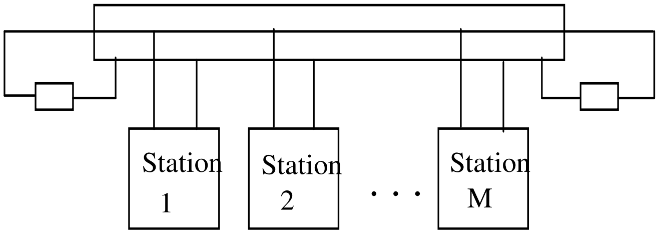


Fig. 13.7 Multiple Access via Wired Bus Structure

Main features of the cable channel are the high data rate (some Mbits/s) and low propagation time because of the short distance (it is reasonable to build the cable networks inside one building or between some buildings standing near each other). Thanks to the bus structure, users can check whether the channel is idle before sending their own packet and this information is quite reliable because of the short propagation time. The users can also listen to their own transmission and verify whether the transmission was correct or not. Thus the advantages of the satellite and terrestrial systems seem to be unified in this application.

13.4.1 Aloha-Type Random Access Procedures

Aloha is the most known and the simplest random access method. It has two versions, one of them is called *pure* and the other *slotted* Aloha. ('Aloha' is a Hawaiian greeting and it is used since the procedure has been worked out and tested first at the University of Honolulu.)

Any Aloha-user sends his packet immediately after it is ready. The packet starts with a 'header' containing the terminal address (identifier) and is protected by an error-detecting code (some kind of cyclic code). The central controller checks whether the packet has not collided (overlapped) with the packet of another user(s). Note that random bit errors result also in failed transmission.

Successful transmission is acknowledged by the central station or the lack of acknowledgement informs the user(s) that there was a collision. In the latter case, the user tries to repeat the transmission after a random delay. A possible process flow is illustrated in Fig. 13.8.

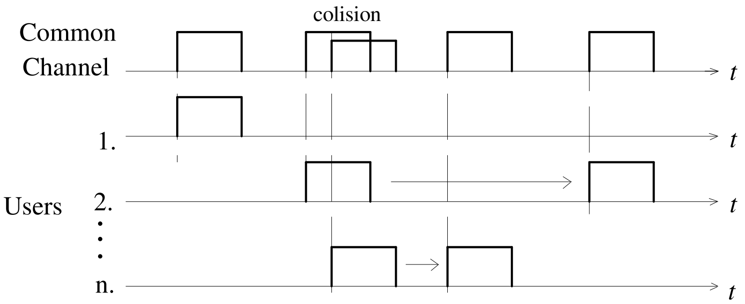


Fig. 13.8 An Example of Aloha Procedure

The advantage of the pure Aloha lies in its simplicity and acceptable performance in the case of low number of terminals. Slotted Aloha is an improved version which defines a time grid equal to the packet time. The length and the position of the time slots are known to the users, i.e. there is a kind of synchronism but still they do not co-operate with each other. The only difference compared to the pure Aloha is that the user has to send his packet(s) precisely inserted into the time slot(s). It is easy to see that the sending will be either successful or there shall be a full overlap with other packet(s). As a result, the channel efficiency will be higher since the 'vulnerable' time interval (when the packet may collide because of the start of another transmission) is halved.

13.4.2 Carrier Sense Procedures

Carrier sensing procedures form another group of random access methods. Short term for this group is CSMA (carrier sensing multiple access).

Suppose a network with star topology and (like in the Aloha) users communicating just with a central station. However, the users can listen to others or at least they are able to detect whether a transmission is on. Under such a condition, a significant number of collisions is avoided since the users examine the channel status before sending their packets and attempt the transmission only after they found the channel empty (no carrier detected).

Depending on what the user do after having examined the channel, several procedures may be followed. The simplest of them is the so called *non persistent* procedure which goes on according to the following rules:

a) If the channel is idle at the time the message is ready for transmission, the packet is sent.
b) If the channel is busy, user checks it later and repeats the procedure above.

It can be seen that carrier sensing cannot completely remove all the collisions since -due to the finite propagation delay and detection time- the users may detect the channel idle although it is already busy. If the packet time is denoted as T and the sum of the propagation delay and detection time as a**,** it is obvious that the lower is the ratio a/T, the less will be the number of collisions and thus the better will be the channel utilization.

13.4.3 Conflict Resolving Algorithms

When distributed control is necessary, conflict resolving methods are used. The basic idea of these methods is the definite effort to resolve any conflict caused by simultaneously occurring demands, i.e. that it is a primary aim that channel has to be allocated to the users taking part in the collision. For this purpose, new users are not allowed to transmit until the conflict is resolved. The specific feature of conflict resolving algorithms is that if channel utilization is lower than the value specified for a given procedure then the operation is stable. (For the Aloha-type procedures, when the input traffic increases, the system may become unstable: propagation delay of packets increases, channel utilization decreases, and the channel may be 'blocked'.)

Control questions

1. What is the procedure for frequency division multiplexing ?
2. Why is polling advantageous as compared with static TDMA ?
3. What are the main practical cases of multiple channel access and what are the main features of these cases ?
4. Why is carrier sensing multiple access more advantageous than the Aloha method ?

References

[1] Dallos Gy.- Szabó Cs.: Hírközlõ csatornák véletlen hozzáféréses módszerei. Akadémiai kiadó, 1984.

[2] Tanenbaum A.: Computer Networks. Prentice Hall, 1989.

Abbreviations

AM-DSB	Amplitude Modulation - Double Sideband
FDM	Frequency Division Multiplex
FM	Frequency Modulation
TDM	Time Division Multiplex
TDMA	Time Division Multiple Access
CDMA	Code Division Multiple Access
SSMA	Spread Spectrum Multiple Access
CSMA	Carrier Sense Multiple Access

14. PUBLIC SWITCHED TELEPHONE NETWORK

Recently used public communication services (telephone, telex, data transmission, etc.) are provided by specific network. Characteristic common task of this network is to establish upon subscribers' demand a great number of simultaneous connections among end-equipment connected to the network. The general model of such a network contains end-equipment connected by the subscribers' lines to switched nodes (exchanges) and channels interconnecting the network nodes (see Fig. 14.1).

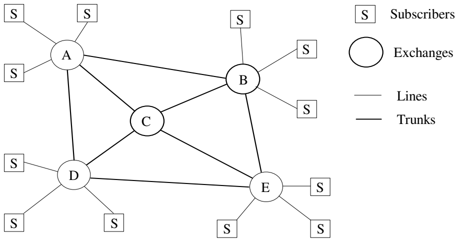


Figure 14.1. Theoretical Model of the Telephone Network

While the subscribers connected to *mobile* telephone network (see Chapter 17) can arbitrarily change their position even during communication, position of end-equipment connected to wired networks is fixed and it is rarely changed. According to the type of the end-equipment, the network is either *circuit switched* or *packet switched*. In the case of circuit switching, a communication path is established between two end-equipment by serial connection of channels. During the whole interval of the communication this part of the network is used exclusively by the two end-equipment. Since the communication path is continuously assigned to the link, the delay of the communication depends only on the parameters of the established path. Circuit switching is used in networks facing strict real-time specifications (e.g. telephone).

In the case of packet switching, there is not an individual path established for each communication, instead of, an identifier unanimously describing the source-to-destination path is assigned to each transfer. The end-equipment are sending their information to the node in *packets* with identifier placed in the *header* of the packet. The header is decoded by the node and according to its content, the packet is passed to the next node provided the channel is free. If the channel is engaged the collided packet is stored in a temporary buffer (e.g. in a FIFO register) and transmitted later.

Packet switching increases the channel efficiency, the delay time is however increased by waiting for the channel access. The delay can even fluctuate and packets can get lost when the buffer is full. Packet switching is therefore used in such areas where the real-time demand is not so strict and the time of the actual information transfer is significantly shorter than the interval of connection.

Old-established public switched services work on the base of circuit switching. This is due partly to the above mentioned aspects but also for the technological limits of the age these networks were established.

14.1. Structure of the Telephone Network

Public switched telephone network (PSTN) is the greatest automate of the Earth able to establish speech communication of good quality between any two points of the Earth. As the result of more than 100 years of evolution, the newest digital systems have to co-operate with several obsolete systems based on old technology and old techniques. The greatest challenge of operating telephone systems are therefore solutions for the co-operation of these greatly differing systems.

Telephone network is a classic example of circuit switched network and telephone exchanges are their principal elements necessary for switching functions. The general model of telephone exchanges is shown in Fig 14.2. The task of the *switch matrix* is the switching of the voice channels on the users demand. *Space-division* and *time-division* switching networks are used. The space-division switching network separates the simultaneous connections in space while the time-division switching network separates them in time.

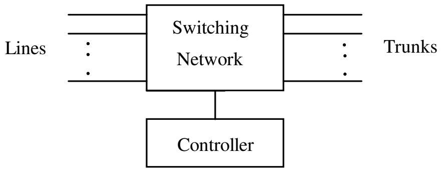


Figure 14.2 Principal Block Diagram of the Telephone Exchange

The task of the controller is to establish the connection between the caller and the called subscriber with respect to the actual state of the called line and that of the switching network. In manually operated exchanges, the switching network and the controller were separated indeed as shown in Fig. 14.2. In automatic exchanges the switching and the control functions have been partly or entirely merged. Latter these functions gradually split again and became fully independent in the Stored-Program-Control (SPC) exchanges.

In Figure 14.3 two versions of subscriber services are presented for an idealized area: a one-exchange and a four-exchange version. The latter reduces the average line length to one customer but new lines, called the *trunks* appear in the network. The total cost is then given by the costs of the subscriber lines and by the cost of the trunks. This cost is less, however, than it is in the case of the single exchange since the number of trunks is significantly smaller than the number of subscribers. A trunk can also pass any call of its subscribers, if it is not engaged by another call.

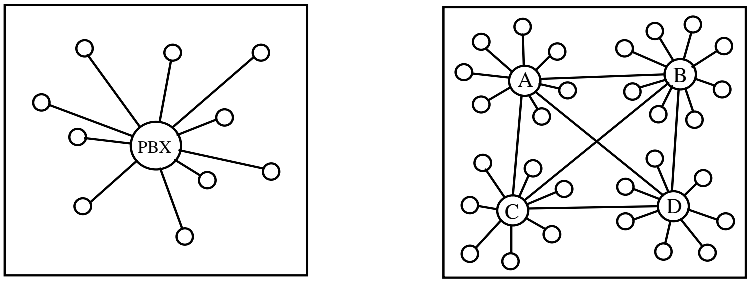


Figure 14.3 Local Networks

The arrangement shown in Fig. 14.3 b) is called the *meshed network* since all the exchanges are interconnected. Also such arrangements exist where some (or all) exchanges are connected to other exchanges via two serial trunks separated by so-called *tandem* exchange(s).

The subscriber lines are the most expensive but the least utilized parts of the telephone networks. The average traffic of a subscriber line is only about 10 per cent even in the busy hours of the day. Further cost-reducing techniques are used, therefore, besides the multi-exchange structure. The simplest of them is the party-line where two subscribers are connected to the exchange by the same cable pair.

If a greater group of subscribers is near to each other, *Remote Switching Units* (RSU) are used to bring a part of the switching equipment of the exchange closer to the subscribers (see Fig. 14.4.). The number of the connecting trunks *G* is nearly the same as the number of connections between the RSU and the rest of the exchange. The trunks have to provide also for the functional interfacing between the RSU and the exchange besides the simple interconnection.

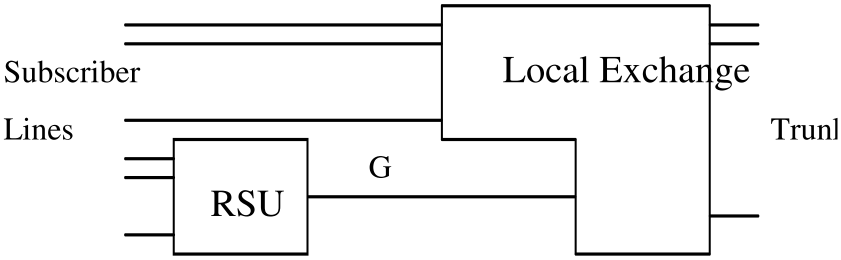


Figure 14.4 Exchange with Remote Switching Unit

An RSU can also play the role of a local exchange. In this case subscribers connected to the RSU can call each other even if the link between the RSU and the exchange is interrupted. Remote Switching Units are widely used in digital exchanges since they reduce network costs, operate reliably (with appropriate redundancy, if necessary) and do not need special buildings.

Reliable operation of the network devices is very important. Maintenance costs should be low otherwise the savings achieved by the economical network arrangement are spent on maintenance and the grade of the service deteriorates.

Local networks are connected into national networks. Joining the national networks, international, continental and intercontinental networks are formed (see Fig. 14.5.). Only a small portion of the originated traffic goes outside the local network. This traffic is passed to the so-called primary exchange where outgoing traffic of several local networks is collected and according to the destination address, it is directed to one of the local networks held together by the primary exchange or to a further (secondary) exchange. The secondary exchange holds together several primary exchanges and ternary exchanges may exist, as well.

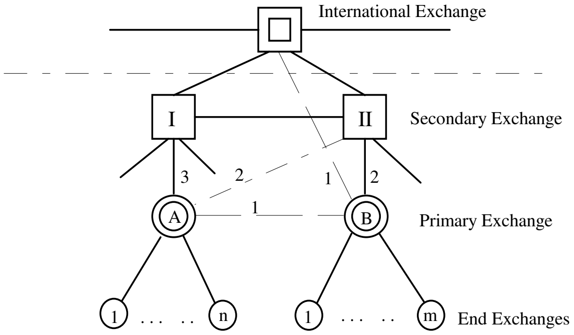


Figure 14.5 Hierarchical Telephone Network

Pure radial networks are easy to control but they are very sensitive to break-downs and not optimal from the point of view of the costs. The meshed network among secondary exchanges is more reliable since it enables to establish connections between two secondary exchanges even if the direct path is interrupted or there is a congestion in the direct path.

To explain the principle of the multipath routing on the level of primary exchanges, two examples are given in Fig. 14.5. The first example shows how the primary exchanges belonging to different secondary exchanges are controlled. Traffic from exchange A to exchange B is attempted first through the direct route. If this fails then A-II-B route is tried and if this also fails, then A-I-II-B route is attempted. The second example presents two possible routes of the international traffic outgoing of B.

As a general rule, always the shortest route, i.e. that containing the minimum of cascaded trunks has to be used. For instance, if direct route does not exists between the exchanges A and B, then the first selected route shall be the A-II-B route. Primary, secondary and international exchanges are the so-called transit exchanges which do not switch local lines, only trunks.

14.2. Telephone Sets

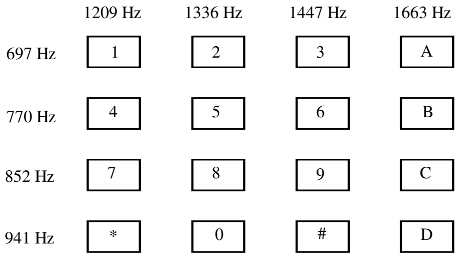
Before dealing with the exchanges, telephone sets have to be discussed briefly. The important parts of a telephone set are as follows:

- electroacoustical transducers: the microphone and the receiver,
- signalling device (dial or push-button type),
- cradle,
- ringer,
- speech circuit (hybrid), matching the 4-wire input/output of the telephone set to the 2-wire local line so that the signal of the microphone is attenuated towards its own receiver (sidetone reduction).

Carbon microphones have been exclusively used for almost 100 years. This type of microphone has to be fed by an external battery. In the beginning, a local battery (LB) was used, later it was substituted by a central battery (CB), which is still in use. CB-feeding has technical and economical advantage since the direct current can be used for the subscribers signalling towards the exchange.

As a signalling device, the dial was used exclusively for a long time. In this case the line current is chopped by the dialler at a speed of 10 impulses/s as many times as given by the dialled number. The signalling (ringing) tone is a 25 Hz, 75 V_{eff} AC signal transmitted to the called party by the exchange. For the customer's information other voice-band signals are also generated by the exchange (dial tone, busy tone, ringing tone, etc.)

The low-voltage, low-power microelectronic devices made it possible to develop modern telephone sets based entirely on the solid state technology which is very important because of the CB feeding. Here we discuss only the so-called dual tone multi-frequency (DTMF) dialler (see Fig. 14.6.) introduced instead of the slow and uncomfortable dial.



	1209 Hz	1336 Hz	1447 Hz	1663 Hz
697 Hz	1	2	3	A
770 Hz	4	5	6	B
852 Hz	7	8	9	C
941 Hz	*	0	#	D

Figure 14.6. The DTMF Keyboard

Two groups of four frequencies each are assigned to the rows and the columns of the keyboard. When a button is pressed, two frequencies are generated and passed to the exchange, one of them identifying the row and the other the column of the pushed button. (The buttons of the fourth column are not used in telephone sets.)

With a DTMF keyboard the digits can be sent to the exchange much faster and since the used frequencies fall into the voice-band, information and control commands can also be sent to the called partner.

14.3. Automatic Telephone Exchanges

Telephone exchanges are seldom used in the full extent of their capacity. For economical reasons, only the minimum required amount of equipment should be installed and during their lifetime, the exchanges should be expandable to full capacity in several steps without disturbing the actual traffic. It is also important to use a structure which is able to cover the possible widest area of application. To meet these demands, a part of the network switching the local lines is formed in groups which are connected through a group selector which can also be gradually expanded (see Figure 14.7.)

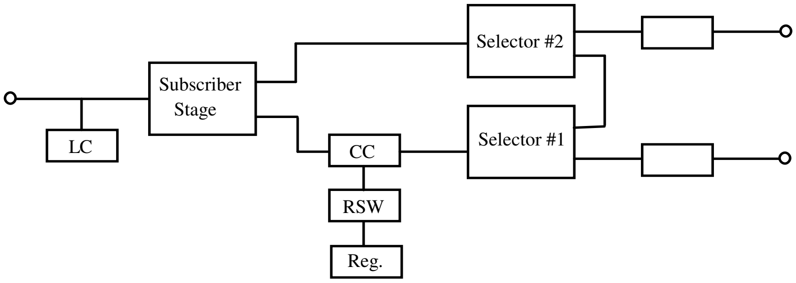


Figure 14.7. Block Diagram of a High Capacity Local Exchange

Functions which have to be performed in the course of a simple call can be classified as follows:

- functions to be performed per local lines,
- functions to be performed during the call set-up,
- functions to be performed during the call.

For economical reasons the above functions are realized by a separated equipment and are switched onto the local line only when it is necessary. There is a line circuit permanently assigned to each local line, the task of which is to signal the idle state, the call attempts, the busy state and the line-blocking (due to an error or to any other reason).

The number of junctor circuits is significantly less than that of the subscriber lines (about 10% of the latter) since they share them jointly. The subscriber stage connects a free junctor circuit to the calling line. When the call is ended the line becomes free thus being able to serve another call. The main functions of the junctor circuit are as follows:

- to supervise the calling and the called line (to detect the call-answer and the disconnection from both sides),
- to switch the ringing current for the called party,
- to switch the ringing tone for the caller,
- to supply the current for both microphones,
- to maintain the busy status of the calling and the called party for the time of the connection.

The register which controls the call set-up is switched by the register-switch to the junctor circuit for the time while the connection to the called party is set up. The main functions of the unit are as follows:

- to set the line circuit of the caller to the busy state,
- to send a dial tone to the caller,
- to receive, store and analyze the called address,
- to set-up the connection to the called customer,
- to supervise the status of the caller's line (early disconnection).

The above grouping of the control functions is characteristic for the so-called electromechanical exchanges used in the first generation of automatic exchanges. The particular feature of these exchanges is that the local line is always connected (galvanically) to the circuit which responds to the subscriber's 'logical' act -€expectable in the given phase of the call€- and generates the next state of the call.

If the called subscriber belongs to another exchange than the calling one then the group selector I. connects the junctor circuit to one of the free trunks leading to the called subscriber's exchange. The other end of the output trunk is connected as input to the group selector II. of the called subscriber's exchange. Information necessary to control the GS II. and the SS are sent via the established connection by the register of the caller's exchange. It is the task of the input trunk circuit to ring the called subscriber set and to feed its microphone.

If exchanges with different signalling systems take part in communication, it become more complex to handle the incoming and outgoing calls. Generally, the exchange newly entering the network has to be adapted to the actual signalling scheme which is a -frequently disadvantageous- constraint for a new system.

14.4. Stored Program Control Exchanges

In spite of the intensive research work, the development of the telephone switching and control equipment has been much slower and more protracted than that of the electronic computers. A spectacular break-through was achieved May 30. 1965 when No.1 ESS, the first SPC exchange, was installed in Succasunna (USA).

The essence of the stored program control is that all functions of the switching system are realized by programs stored in a memory. Processors, similar to those used in computers have been developed for this purpose. The operating programs as well as the parameters of the controlled equipment and the data of its actual state are stored in the processor memory.

Using the SPC the structure of the exchange does not determine its functions thus by means of appropriate programs an exchange can theoretically be adapted to any service and to the co-operation with any environment. By loading a new program into the SPC exchange, it can be simply adapted to the demands of the newly introduced services.

On the other hand, a wide variety of controlling, supervisory and administrating programs had to be developed and difficulties caused by the handling and application of the programs had to be taken also into account. The program packet of the first No.1 ESS exchange contained more than 100.000 instructions, and today's up-to-date systems of exchanges contain at least one million instructions. The control programs are only a small part of all programs, the supervisory, diagnostic, managing, etc. programs forming their greater part.

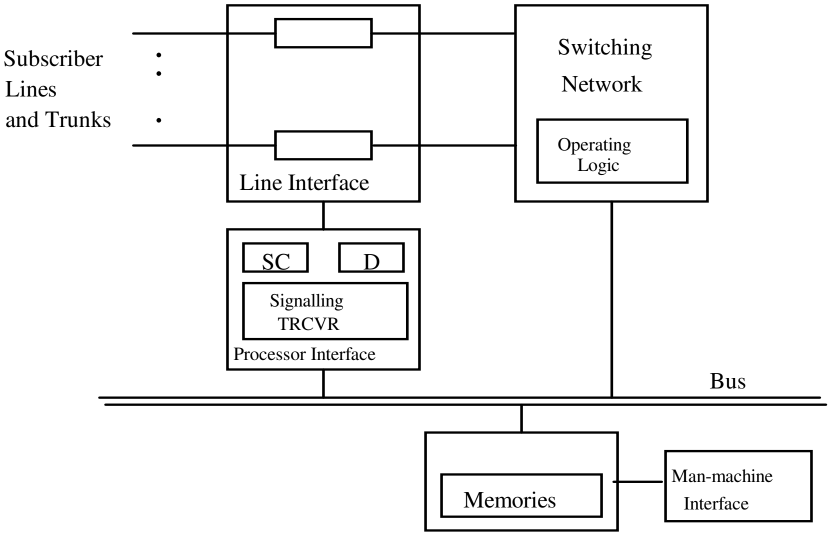


Figure 14.8. Functional Blocks of the SPC Exchange

The main blocks of the SPC exchanges are shown in Fig.€14.8. This block diagram shows both exchanges with space-division and time-division switching networks. (Exchanges exist where after replacing the analog switch matrix with a digital one, the block diagram remains the same as before.)

The change of the inner structure can, of course, influence its environment. For instance, the functions of the junctor circuit used in exchanges operating with space-division switching jointly by more subscribers, have to be moved to the line interfacing circuits individually assigned to the local lines if replaced by the time-division digital switching network.

The first generation SPC exchanges have been equipped with hermetically sealed cross-points operating with fast, low-power galvanic switches (reed and ferreed relays). These switches were much slower than the electronic switches. At that time, however, semiconductor devices with the switching parameters necessary for the switching network were not known, yet.

An important block of the SPC exchange is the unit interfacing the processor to the environment via a network interface. The processor interrogates the status of the lines and that of the trunk by a scanner, compares them with the previous line status to find out the changes, and stores the data until the next scan. The status change represents an event which is processed by the appropriate program and the required operation is then passed to the interface via the distributor circuit. According to different signalling systems, events or information (e.g. the address of the calling or the called party, etc.) are transferred among the processor and the trunks by signalling receivers (such as the DTMF signalling receiver) and signalling transmitters.

14.5. Digital Exchanges

The development of the highly integrated semiconductor devices made it possible to realize digital speech encoding, switching and transmission at an acceptable price which was a decisive factor to use the *time-division* digital switching network. Functions which had to be realized when interfacing the subscriber lines to the digital switching network are symbolized by the acronym BORSCHT which stands for:

B:€Battery feeding
O:€Overvoltage-protection
R:€Ringing
S:€Signalling (or Supervision)
C:€Coding (A/D conversion, see Chapter 3)
H:€Hybrid (2/4-wire conversion, see Chapter 8)
T:€Test (switching to the equipment testing the line and the line circuit)

The aim of the manufacturers of the equipment is to integrate the maximum of the above functions since the circuit implementing them represents the 30-40% of the cost of the local exchange. E.g., codecs with programmable time-slot assignment realize a part of the switching functions of the subscriber stage. Due to the rapid development of the semiconductor technology, all functions have been integrated except the T and the R function which are carried out by relays even at the beginning of the '90-s. The logical relations of the functions are shown in Figure 14.9.

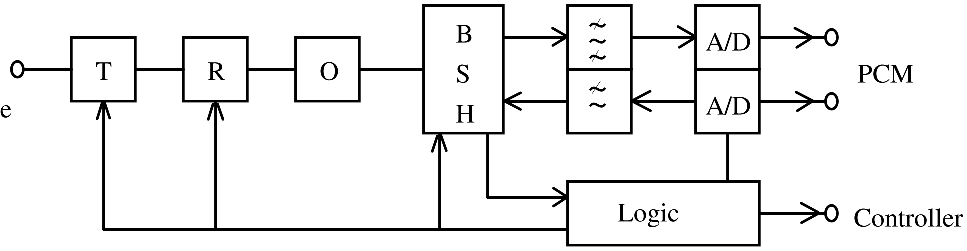


Figure 14.9. Relations Among the BORSCHT Functions

The time-division digital switching and transmission mutually enhance each other's efficacy, even more, certain functions of the switching and the transmission (e.g. multiplexing and concentration) can be unified. The wide-spread use of microprocessors made it more economical to decentralize the earlier centralized stored-program control used together with the above mentioned integrated switching and transmission technique (IDN€-€Integrated Digital Network). This resulted in bringing the units concentrating the traffic closer to the subscribers thus reducing the cost of the local network. (On top of that, these small, reliable operating units do not require special buildings.)

In the latest stage of development, the digital transmission has been extended to the subscribers' sets enabling thus the telephone network to be used as a single common medium for the speech and data transmission (ISDN€-€Integrated Services Digital Network, see Chapter 18). Besides the basic services, users can subscribe to a wide variety of additional services. To implement such a system, a highly sophisticated signalling system is necessary between the telephone set and the exchange, and among the individual exchanges, respectively. Several signalling systems have been developed by the CCITT for this purpose, the most important of which are the DSS 1 (Digital Subscriber Signalling System No.1) and the CCS€(Common Channel Signalling System No.7).

Control Questions

1. Economic design of subscriber networks.
2. Switching networks, types and structure aspects.
3. Classification of control functions.
4. Multi-path traffic control.
5. Functional blocks of the SPC exchanges.

References

[1]€Távközlõ hálózatok forgalmi tervezése. (Szerkesztette: Dr. Sallai Gyula) Közlekedési Dokumentációs Vállalat, 1980

[2]€T. H. Flowers, Introduction to Exchange Systems, John Wiley and Sons, London, 1976

[3]€GRINSEC, Electronic Switching, Elsevier Science Publishers B. V. 1983

[4]€John Bellamy, Digital Telephony, John Wiley & Sons, 1982

Abbreviations

CB	Central Battery
CCITT	International Telegraph and Telephone Consultative Committee
CCS	Common Channel Signalling
DSS1	Digital Subscriber Signalling No. 1
DTMF	Dual Tone Multi-Frequency
IDN	Integrated Digital Network
ISDN	Integrated Services Digital Network
LB	Local Battery
RSU	Remote Switching Unit
SPC	Stored Program Control

15. TELETRAFFIC THEORY

The telecommunication traffic theory is the application of the mass-servicing theory for telecommunication systems. The theory of the telecommunication traffic was founded by the Danish A. K. Erlang and published between 1909€-€1928. The traffic theory applied for practical cases is based upon the condition of the statistical equilibrium.

Telecommunication systems are built upon sets of different resources (switching units, transmission channels, etc.). Due to economical reasons, the number of such resources is limited in a system thus the customers have to share them somehow. This may happen so that a servicing unit is engaged during a call and it is disengaged, i.e. `given back' to the common resource at the end of the call. This method implies the chance that it is impossible to set-up a call when all servicing devices of the system are engaged simultaneously. Such unpleasant event is called *congestion*, and from this point of view, services are qualified by the so-called GOS (grade of service). The aim of the traffic design is to provide for a sufficient number of servicing devices and to maximize their utilization.

15.1. Terms and Definitions

To make the following discussion easier, some terms and definitions are mentioned in advance:

-€*call intensity* (λ)**:** the sum of the demands directed to a switching unit, group of circuits, or customers per unit of time.

-€*holding time* (h): the time for which the servicing device is engaged by an acknowledged demand.

-€*traffic intensity*: the sum of the holding times of calls simultaneously in progress during a particular period of time. Let us take first the sum of the holding times:

$$\sum_{i=1}^{z} h_i = z \cdot \overline{h} \tag{15.1}$$

where h_i is time of the i-th call, z is the number of the calls and $\overline{h}$ is the average holding time. If the period of time during which the holding times were counted is T then the average of the occupations is:

$$\frac{1}{T}\sum_{i=1}^{z} h_i = \frac{1}{T} \cdot z \cdot \overline{h} = \lambda \cdot \overline{h} \tag{15.2}$$

where λ is the call intensity and

$$\lambda \cdot \overline{h} = A \tag{15.3}$$

gives the average of the simultaneous occupations for a given period of time, generally called traffic intensity or simply traffic. Traffic is a quantity with no dimension, the word Erlang is, however, used with the value to indicate that the number is a term used in *telecommunication*.

-€*offered traffic (A)*: amount of traffic offered for a group of devices corresponding to the theoretical description of the given traffic situation (a presumed quantity).

-€*carried traffic (Y)*: traffic carried (transmitted) by a certain group. It can be used both for theoretical description and -since it can be measured- for actual situation as well.

-€*busy hour*: a daily period of one hour in which the traffic is the greatest. The time of the busy hour generally depends on the calendar days.

-€*time-consistent busy hour*: period of one hour starting at the same time each day in which the traffic intensity is the maximum for the group of devices and the days examined. In all likelihood, the customers would like to have a satisfying GOS even in the busiest hours of the year. This is, however, impossible to determine, and what is more, it is growing, further such a system would not be well utilized in the other hours thus it would be too expensive. Therefore it is more reasonable to choose and to use for the design such a traffic value which is exceeded only in a few days of the year.

15.1.1. Mathematical Model of Telephone Traffic

The *input process*, the *service procedure* and the *servicing rules* are the characteristic features of a model. The input process is defined by the distribution of time passing between the arrivals of two consecutive call demands. The service procedure is determined by the number of the service units, by the distribution of the service (holding) times and by the access mode to the service units. The service rules dispose of the congested demands. In *loss* systems the congestion is resolved by clearing the congested demands while in *delay* systems the calls form a queue and are serviced e.g. in the order of arrivals.

Two terms of the congestion are used: *Time congestion* is the proportion of the time during which all accessible service units are simultaneously engaged. *Call congestion* is the proportion of those calls which were rejected in a loss system or had to wait in a delay system.

The different cases are based upon the following conditions:

1.)€The principle of the statistical equilibrium may be used.
2.)€The operation of individual traffic sources is independent of the state of the other sources.
3.)€Time between two consecutive calls has a negative exponential distribution.
4.)€Time of the individual occupations is independent of other occupations.
5.)€Duration of the individual occupations has a negative exponential distribution.
6.)€The fate of the unsuccessful calls is regulated by deterministic rules.

The traffic is considered as events generated by individual sources each capable to initiate simultaneously only one call. The number of the sources can be finite or infinite, the traffic offered by them, however, is always finite!

The distribution of arrival times has a negative exponent and their average is 1/λ:

$$F(t) = 1 - e^{-\lambda t} = P(\le t) \qquad (15.4)$$

The distribution of times between an arbitrary point of time and a call follows the same exponential function as the distribution of the times between the calls and it is independent of *t* (i.e. memoryless). The number of arriving calls is described by Poisson distribution.

The holding time distribution is given by the holding time average *h* in the negative exponent:

$$F(t) = 1 - e^{-\frac{i}{\overline{h}}} = P(\le t) \qquad (15.5)$$

If the distribution of the holding times is exponential and provided the number of independent occupations is *i*, then the number of ending calls within time *t* is

$$\mu_t(t) = i \cdot t / \overline{h} \qquad (15.6)$$

The number of the call occurrences can be calculated from the call intensity of the free traffic sources (λ'):

$\lambda_i = (S\text{-}i)\cdot\lambda'$ (Bernoulli, Engset) (15.7)

$\lambda = \lambda_i$ (Poisson,Erlang) (15.8)

where *i* is the number of engaged traffic sources, *S* is the number of traffic sources and λ_i is the frequency of call intensity (occurrence).

An important element of the service mechanism is the mode of accessing (grouping) the service units which can be as follows:

1.)€*Full-availability group* in which any input has access to any output; a free output can thus always be accessed by a given input regardless of the occupations between other inputs and outputs. As shown in Fig. 15.1. a.), inputs and outputs are interconnected through the switching elements located at the cross-points of vertical and horizontal lines. Because of this arrangement, the array of switches is called the switch matrix.

To analyze the traffic behaviour, the model shown in Fig 15.1.c) is used. Here the small circles arranged in a straight line represent the corresponding inputs and outputs. The presentation most frequently used for the full-availability switch matrix is shown in Fig. 15.1. b).

2.)€*Limited-availability* (or grading) group was used in space division exchanges to increase the throughput of the circuits.

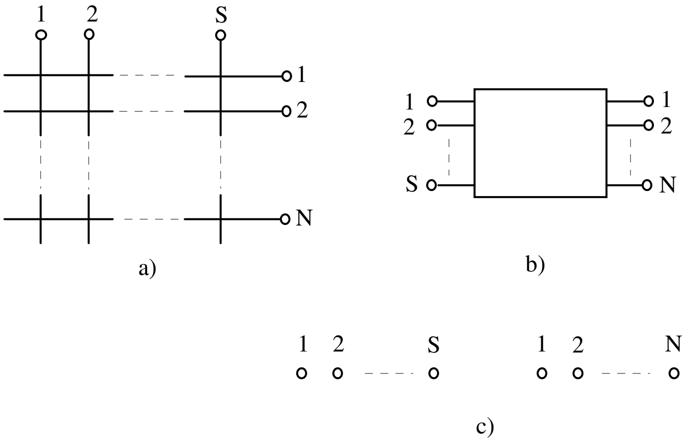


Fig 15.1. Different Presentation of Full-Availability Groups

3.)€*Link system* in which the input/output interconnections are realized by two or more serially connected full-availability switch matrices with a small number of cross-points (see Fig. 15.2.). The link system minimizes the number of cross-points, its application began in cross-bar systems using precious metal contacts.

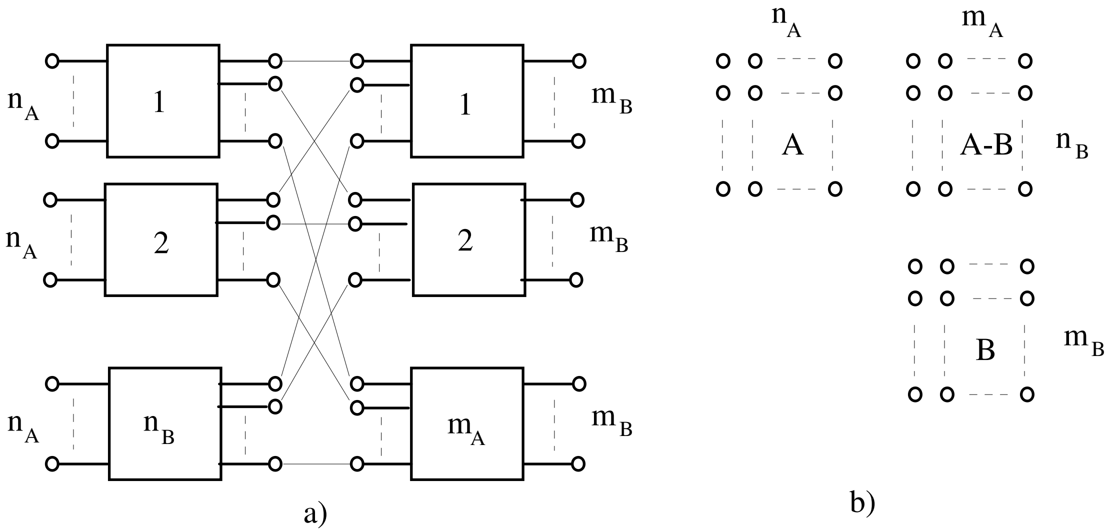


Fig 15.2. Different Presentation of Two-Stage Link Systems

15.2. Loss Systems

The principle of statistical equilibrium can be applied for the busy hour since during this time the traffic is neither increasing nor decreasing but it is fluctuating near the average value. This is only possible if the frequency of transitions from the (i-1) occupations to i occupations is the same as inversely.

15.2.1. Full-Availability Group

Let *P*(*i*) be the probability of *i* simultaneous occupations out of *N* service units (*i*€=€0, 1, ... *N*). P(*i*) represents also the proportion of the time during which *i* simultaneous occupations exist. Let λ_i denote the call intensity, μ_i the call terminations per unit of time in a system with *i* occupations.

Using the principle of the statistical equilibrium:

$$P(i+1)\cdot\mu_{i+1} = P(i)\cdot\lambda_i \quad \text{from which} \quad P(i+1) = P(i)\cdot\frac{\lambda_i}{\mu_{i+1}} \qquad (15.9)$$

(Note that λ_N€=€0 in loss systems!)

a.) *Erlang-type system*

If the call intensity is constant (*S* >>€*N*) then λ_i€=€λ. Substituting μ_{i+1}€=€*h*/(*i*+1) and $\lambda\cdot h$€=€*A* into eq. (15.9):

$$P(i+1)€=€P(i)\,\frac{A}{i+1} \qquad (15.10)$$

With recursion from *i*€=€0 and the full series of events:

$$P(i) = \frac{\dfrac{A^i}{i!}}{\displaystyle\sum_{i=0}^{i=N}\frac{A^i}{i!}} \qquad (15.11)$$

Time congestion is defined as:

$$E = \sum_{i\geq N} P(i) \qquad (15.12)$$

so that the time congestion of Erlang-type systems is

$$E_N(A) = \frac{\dfrac{A^N}{N!}}{1 + A + \dfrac{A^2}{2!} + \ldots + \dfrac{A^N}{N!}} \qquad (15.13)$$

The call congestion (*B*) is defined as

$$B = \frac{\displaystyle\sum_{i\geq N}\lambda_i P(i)}{\displaystyle\sum_{i\geq 0}\lambda_i P(i)} \qquad (15.14)$$

Substituting λ_i =€λ we obtain

$$B_N(A) = P_N(A) = E_N(A) \qquad (15.15)$$

which is called the first Erlang formula. Tables or diagrams are used for the quick evaluation of the Erlang formula B, as is shown in Fig. 15.3. Usually, loss is given as the parameter but any other variable can be used instead.

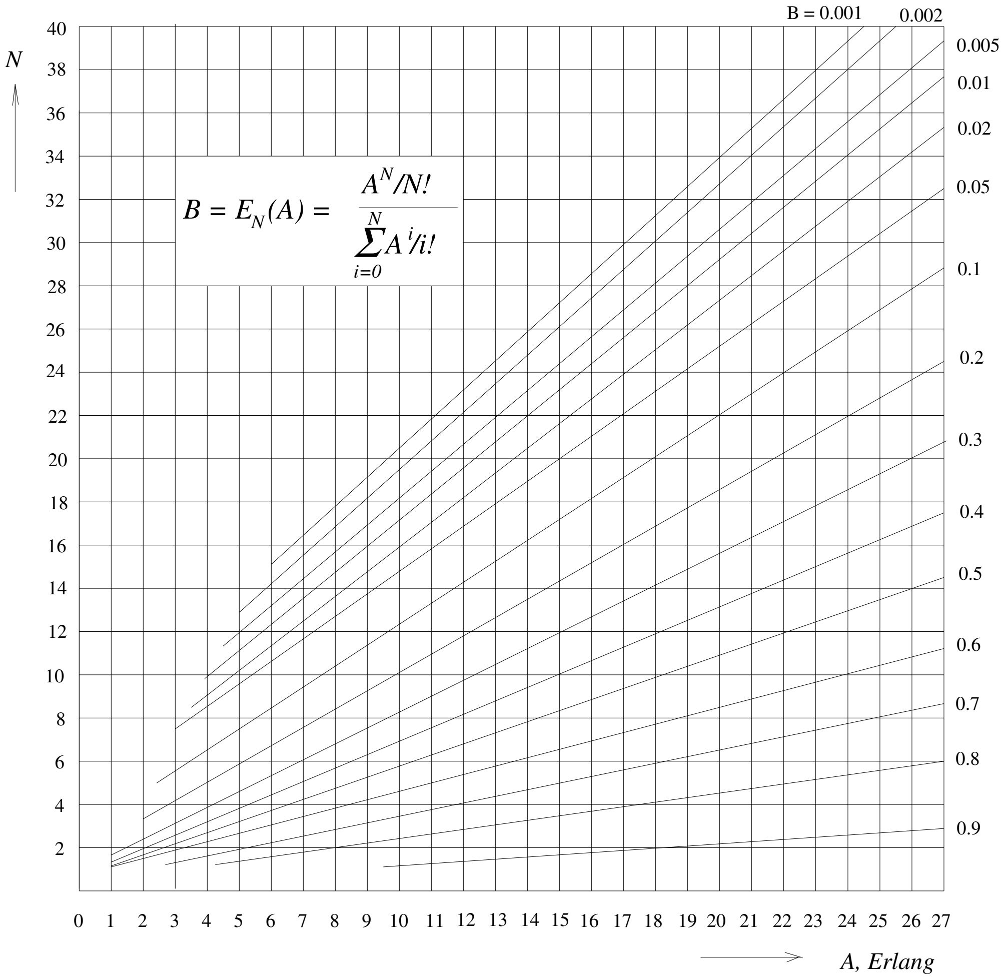


Figure 15.3 Graphical Representation of the Erlang Formula

The carried traffic: $Y = \sum_{i=1}^{N} i \cdot \mathrm{P}(i)$ (15.16)

can be expressed also as

$$Y = A \cdot [1 - \mathrm{E}_N(A)] \qquad (15.17)$$

The average usage in the case of random hunting of the service units is

$$a = \frac{A}{N}\cdot[1\text{-}\mathrm{E}_{\mathrm{N}}(A)] \tag{15.18}$$

The usage of the *i*-th service unit in the case of sequential hunting:

$$a_{\mathrm{i}} = A\cdot[\mathrm{E}_{\mathrm{i-1}}(A) - \mathrm{E}_{\mathrm{i}}(A)] \tag{15.19}$$

b.) *Varying call intensity.*

If *S* is not much greater than *N*, the call intensity depends on the number of the engaged service units. Substituting λ_{i} from (15.7) into eq. (15.9) representing the principle of the statistical equilibrium:

$$\mathrm{P}(i+1) = P(i)\cdot(S\text{-}i)\cdot\frac{\lambda\cdot\bar{h}}{i+1} \tag{15.20}$$

Using notation $\lambda'\cdot h = \alpha$ (offered traffic of the free traffic-source in a unit of time!) and starting recursion with *i*€=€0 we have

$$\mathrm{P}(i) = \mathrm{P}(0)\cdot\binom{S}{i}\cdot\alpha^{\mathrm{i}} \tag{15.21}$$

P(0) can be determined from the total event and the well-known binomial distribution is obtained for *S*€=€*N*:

$$\mathrm{P}(i) = \binom{S}{i}\cdot\frac{\alpha^{i}}{(1+\alpha)^{S}} = \binom{S}{i}\cdot a^{i}(1-a)^{S-i} \tag{15.22}$$

where a€=€$\frac{\alpha}{1+\alpha}$ is the traffic offered by the traffic source in the unit of time.

The call congestion (*B*) is zero here and the time congestion is:

$$\mathrm{E} = \mathrm{P}(N) = a^{\mathrm{S}} \tag{15.23}$$

15.2.2. Link Systems

The exact calculation of the traffic situations of a link system is very complex. The approximation which can be practically used are based almost exclusively on the theory worked out by Jacobaeus.

Suppose that the state of occupation of the links and that of the outputs is independent, and that every free pair of link and output can be engaged with the same probability (random search) and that the entire congestion is small. The two stage link system is the simplest way to demonstrate the Jacobeous' theory but it can be extended to three or more stages as well.

According to the actual parameters of the input switch matrix (**A** stage), the link system can be basic (*n*=*m*), expanded (*n*<*m*) or concentrated (*n*>*m*). Let *a*, *b* and *c* denote the occupations of an input, a link and an output, respectively. Let *G*(*p*) be the probability that *p* outputs out of *m* are engaged and *H*(*m-p*) the probability that *m-p* links leading to free outputs are engaged (see Fig. 15.4.). (The positions of *G*(*p*) and *H*(*m-p*) can, of course, be swapped).

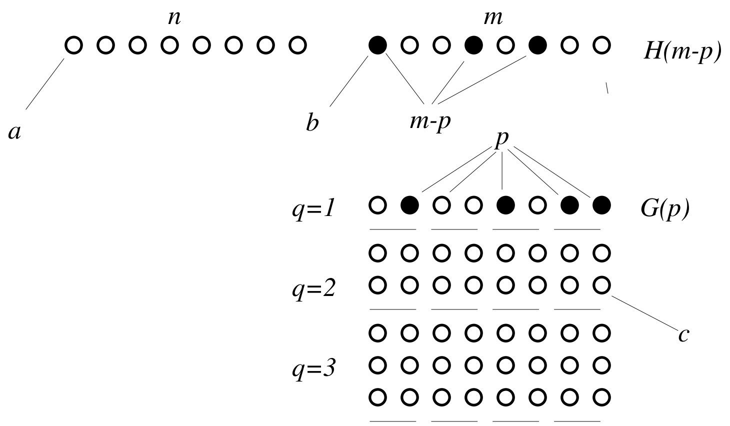


Figure 15.4. Traffic Model of the Two-stage Link System

Obviously, time congestion occurs if the two events are simultaneous. The probability of congestion for *n*€=€*m* is:

$$E = \sum_{p=0}^{p=m} G(p) \cdot H(m-p) \qquad (15.24)$$

The probability of the output occupation can be given either by the Erlang or by the Bernoulli distribution while for the link occupation, the Bernoulli distribution is used. If the Erlang distribution is used for *G*(*p*) and the Bernoulli distribution for *H*(*m-p*), then

$$E = \sum_{p=0}^{p=m} \frac{\frac{(m \cdot c)^p}{p!}}{\sum_{i=0}^{i=m} \frac{(m \cdot c)^i}{i!}} \cdot b^{m-p} \qquad (15.25)$$

where *m·c* is the outgoing traffic (*m*). The expression can be rewritten as

$$E = \frac{E_m(m \cdot c)}{E_m\left(m \cdot \frac{c}{b}\right)} \qquad (15.26)$$

where the Erlang loss of the *m* lines in the case of *m·c* amount of offered traffic is in the numerator while the loss of the same number of lines for the fictive traffic *m·c*/*b* is in the denominator.

Generally, the individual directions can be accessed from the matrices *B* also from more than one output (q≠1). The congestion is then:

$$E = \frac{E_{mq}\, mqc}{E_{mq}\, mq\,\frac{c}{b}} \qquad (15.27)$$

Describing the occupation of the outputs also by the Bernoulli distribution:

$$\mathrm{E} = (b + c^{q} - bc^{q})^{m} \qquad (15.28)$$

15.3. Delay Systems

Besides the probability of waiting, delay systems are characterized by the average waiting time, by the probability of waiting more than a given time interval, and by the expected length of the waiting queue. The exact solution can be given for negative-exponent service time distribution or for constant holding time system with one service unit.

The analysis is given for the so-called Erlang-type system. The previous conditions are thus extended by the further ones:

- each customer who has to wait keeps on waiting for the service,
- the amount of the offered traffic is less than the number of the service units ($A<N$),
- service is done in the order of the arrivals,
- the queue is not limited (infinite number of waiting positions).

The analysis of the delay system is based also on the principle of the statistical equilibrium but the possible states of the system do not end at *N* but in the case of the simultaneous occupation of *N* service units, there may be *j* customers in the queue (j=0,1...∞).

In eq. (15.9) representing the principle of the statistical equilibrium, the rate of increase exists also for $i \geq N$ (λ), the rate of decrease exist always for $i \geq N$ (N/h) because it is assumed that only serviced calls leave the system. Hence

$$\mathrm{P}(i+1) = \mathrm{P}(i)\cdot\frac{A}{i+1} \qquad \text{if } i \leq N\text{-}1 \qquad (15.29)$$

$$\mathrm{P}(i+1) = \mathrm{P}(i)\cdot\frac{A}{N} \qquad \text{if } i > N\text{-}1 \qquad (15.30)$$

With the help of the recursion and using the condition of the full event:

$$P(i)=\frac{\frac{A^i}{i!}}{\sum_{i=0}^{N-1}\frac{A^i}{i!}+\frac{A^N}{N!}\frac{N}{N-A}} \qquad \text{if } i \le N\text{-}1$$

and

$$P(i)=\frac{\frac{A^N}{N!}\left(\frac{A}{N}\right)^{i-N}}{\sum_{j=0}^{N-1}\frac{A^j}{j!}+\frac{A^N}{N!}\frac{N}{N-A}} \qquad \text{if } i > N\text{-}1$$

Using eq.(15.12) and (15.14), it turns out that the time congestion and the call congestion are of the same value. This is stated by the second (or *D*) Erlang formula, which gives the probability of waiting *P*(*t*>0) as:

$$D_N(A)=\frac{\frac{A}{N!}\frac{N}{N-A}}{\sum_{j=0}^{N-1}\frac{A^j}{j!}+\frac{A^N}{N!}\frac{N}{N-A}} \tag{15.31}$$

The relation between the Erlang B and D formulae is:

$$D_N(A)=\frac{N\cdot E_N(A)}{N-A\cdot\left[1-E_N(A)\right]} \tag{15.32}$$

As it follows from the above conditions, *Y* = *A*, i.e. the carried traffic is equal to the offered traffic. Without derivation, the distribution function of the waiting times is as follows:

$$\mathrm{P}(>t)=\mathrm{D_N}(A).e^{-\frac{N-A}{\bar{h}}t} \tag{15.33}$$

If a call has to wait then the probability of waiting longer than a given period of time is:

$$\mathrm{P_v}(>t)=\frac{P(>t)}{P(>0)}=e^{-\frac{N-A}{\bar{h}}t} \tag{15.34}$$

The expected value of waiting times is given by eq. (15.35) for all calls and by eq (15.36) for the waiting calls:

$$\tau€=\frac{\bar{h}}{N-A}\mathrm{D}€(A) \tag{15.35}$$

$$\tau_{\mathrm{v}}\text{€}=\text{€}\frac{\bar{h}}{N-A} \qquad (15.36)$$

The expected length of the waiting queue is:

$$(q)\text{€}=\frac{A}{N-A}\mathrm{D}_{\mathrm{N}}(A) \qquad (15.37)$$

For systems which can be characterized by constant holding times (*h*), an exact solution can be given if N€=€1. The values are halves of those obtained for the exponential holding time:

$$\tau=\frac{h}{2}\frac{\alpha}{1-\alpha} \qquad (15.38)$$

$$\tau_{\mathrm{v}}=\frac{h}{2}\frac{1}{1-\alpha} \qquad (15.39)$$

where *a*€=€offered traffic€=€carried traffic <1.

Control questions

1.€What is the aim of the traffic design in telecommunication?
2.€What is the definition of the time-congestion and that of the call-congestion?
3.€What is the principle of the statistical equilibrium and how is it used?
4.€What are the parameters of link systems?
5.€What are the characteristic parameters of delay systems?

Examples

1.€How many per cent of calls has to pay at least two tariff units if one unit is paid for each commenced 3-minute interval and the average holding time is 2 minutes?
Using the inverse of the eq (15.5): P(t>3)€=€$e^{-3/2}$€=€22%.

2.€Suppose a full-availability loss system containing 5 circuits. What will be the values of the carried traffic for the sequential and for the random hunting, provided the offered traffic is 2 Erlangs?
The traffic of the individual circuits in the case of sequential hunting is given by (15.19) and congestion can be computed from the following recursive relation:

3.€What should be the minimum number of lines if 20 Erlangs of offered traffic has to be serviced at a congestion not greater than 0.002?
From Fig. 15.3.: N≥33.
4.€How can the average carried traffic of the circuits of full-availability lines be evaluated, if the number of lines is increasing and the congestion is 0.005?

Considering eq. (15.18) for the average usage of the circuits, 0.005 can be neglected with respect to 1 so that *A*/*N* approximation may be used. From the diagrams of the Fig. 15.3.:

5.€How can the average usage of circuits for a full-availability group of 10 circuits be evaluated as a function of the offered traffic?
Using diagrams of Fig. 15.3. and the eq. (15.18):

A, Erlang	$E_{10}(A)$	*a*, Erlang
3.1	0.001	0.31
4.5	0.01	0.45
7.2	0.1	0.68
12.0	0.3	0.84
18.3	0.5	0.92

It can be shown that for every N $a \to 1$ if $A \to \infty$.

6.€Suppose an Erlang-type full-availability delay system with $N = 30$ service units, offered traffic of which is 700 calls/hour and the average holding time is 108 s.

€What is the value of the offered traffic?

From (15.3): $A = \lambda \cdot h = \frac{700}{3600} 108 = 21$ Erlang

€What is the probability of waiting?

From the relation (15.32) and from the Fig. 15.3.:

$$D_{30}(21) = \frac{30 \cdot E_{30}(21)}{30 - 21 \cdot [1 - E_{30}(21)]} = \frac{30 \cdot 0.015}{30 - 21 \cdot (1 - 0.015)} = 0.048$$

€What is the average waiting time of the waiting calls?

From the eq. (15.36): $\tau_v = \frac{108s}{30 - 21} = 12s$

€What is the probability for a waiting call that it has to wait longer than 24 s?

From the eq. (15.34): $P_v(t > 24s) = e^{-9 \cdot \frac{24}{108}} = 0.1353$

References

[1] R. Syski, Introduction to Congestion Theory in Telephone Systems, Oliver and Boyd, Edinburg and London, 1960

[2] L. Kleinrock: Sorbanállás, kiszolgálás, Mûszaki Könyvkiadó, Budapest, 1979

[3] A távközlési forgalom tervezése. (CCITT 1984) Közlekedési Dokumentációs Vállalat, Budapest, 1986

16. POINT-TO-POINT TRANSMISSION

As we have seen in Chapter 1, the aim of the transmission is to pass the information from the source to the sink. Transmission systems therefore include the source, the sink and the interconnecting transmission channel. In this chapter we treat the transducers, multiplexers and A/D converters as the part of the source since their properties are irrelevant in the following discussion.

Several tasks of digital processing are carried out in the transmission channel. In fact, the design of the transmission channel is a design of digital processing operations in which two factors has to be considered : the quality requirements for the signal reaching the sink and the model of the channel. Quality requirements are usually fixed in national standards based on international standards or -what is more typical- on recommendations. The properties of the channels are essentially determined by the parameters of the transmission media but they are influenced also by the properties of the circuits and devices used in the channel.

16.1. Quality Requirements

In a transmission system, distortions have to be kept at a acceptable level, i.e. the speech must remain intelligible, the speaker's voice recognizable, the music enjoyable, etc. Interference, distortion and noise components present in the signal transmitted towards the sink are supposed to be given; we do not deal with the criteria these specifications are based upon.

With certain restrictions, analog signals may be of arbitrary waveform. In the analog signal transmission, waveform fidelity is of primer importance; to ensure a 'good' transmission, distortion components may be specified separately. In long haul transmissions, however, the noise at the output is of main importance thus the signal-to-noise ratio is the most important parameter. In digital transmission the number of waveforms is finite, sources usually send binary signals which are apriori known, hence, at the output of the channel it is not necessary to reconstruct precisely the waveform, it is enough to decide which one of the input signals was sent. This decision, of course, can be either right or wrong. Thus the most important quality parameter here is the probability of the right decision. Signal-to-noise ratio and the probability of error is influenced by all listed distortional effects (additive noise, linear distortion, etc.).

16.2 General Block Diagrams of Transmission Systems

As we have seen, transmission systems are built upon sources, sinks and transmission channels interconnecting the former two. The transmission channel is based primarily on the transmission media which can be wirebound or wireless. Wirebound connection is either metal or optical cable, the wireless (radio) transmission can be either terrestrial or satellite.

A part of interfacing equipment depend on the transmission media but is not depending on the properties of the signal. Firstly, let us discuss the radio transmission without dealing with the properties of, the transmitted signals. In Fig. 16.1. basic blocks of a radio transmitter-receiver pair are presented. The receiver shown here is of so called superheterodine type in which the signal is amplified by the so called intermediate frequency (IF) amplifier. This is not the only solution but it is used almost exclusively.

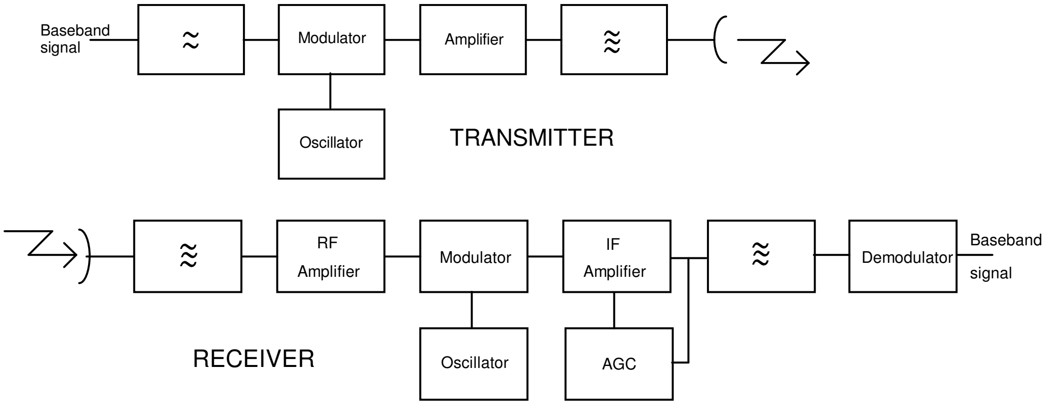


Figure 16.1 Block Diagram of a Transmitter and a Receiver

General block diagram of an optical transmission system is given in Fig. 16.2. The optical modulator is framed in the figure since it is missing in the majority of cases, the light source is modulated by the driver circuit.

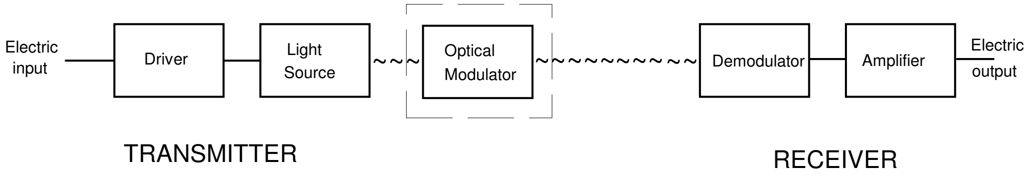


Figure 16.2 Block Diagram of an Optical Transmitter and Receiver

In the following we discuss the microwave and the optical transmission systems. Guided wave system have already been discussed in Chapter 8.

16.3. Microwave Transmission Systems

16.3.1. General Features

Since great bandwidth is required for long-haul transmissions, the frequency range between 1 GHz and 300 GHz is used for this purpose. This range is called the *microwave* range and the region between 30 and 300 GHz is sometimes distinguished as the *millimeter wave* range. High frequency and short wavelength are the most important features of the microwaves. The microwave carrier can be thus modulated by a relatively wideband signal and small antennas can be used to produce a narrow radiation pattern. On the other hand, microwaves do not diffract on great objects such as the Earth, so that only line-of-sight propagation can be used at these frequencies.

A section of a microwave transmission link can therefore extend only to the horizon; of course, at higher locations the horizon is more distant, so that the microwave antennas have to be installed high, on mountain peaks, on towers, etc. Taking into account also terrestrial obstacles, the horizon is hardly farther than 50 km. To communicate to greater distances, either the signal has to be *repeated* every 50 km (i.e. received, amplified and retransmitted) or a single repeater has to be used, located on a *satellite* positioned at high altitude above the Earth. The first solution is known as the terrestrial or radio-relay system while the latter is the satellite communication system. Among the satellite systems, especially the so-called *geostationary* satellites are suitable for microwave transmission. These satellites are positioned about 36000 km above the equator and since the revolution time of the satellite is equal to the time of one turn of the Earth, the satellite seems to be `fixed' above a given point of the Earth.

The ratio of the power of the carrier (*C*) to the additive thermal noise (*N*) is of essential importance in microwave transmission:

$$SNR \hat{=} \frac{C}{N} = \frac{P_a}{a_s \cdot k \cdot T \cdot B} \qquad (16.1)$$

where P_a is the transmitter power, a_s is the free-space loss, $k \cdot T \cdot B$ is the noise power of the resistor having an equivalent absolute temperature *T*. The free-space loss is:

$$a_s = \frac{\left(4\pi D \cdot f^{2}\right)}{G_a \cdot G_v \cdot c^2} \cdot A_f \qquad (16.2)$$

where G_a and G_V are the gains of the transmitter and of the receiver antennas, *D* is the length of the section, *c* is the velocity of the light, *f* is the frequency and A_f is the fading attenuation (power ratio).

The signal-to-noise ratio can thus be given as follows:

$$SNR = \frac{P_a \cdot G_a \cdot G_v \cdot c^2}{\left(4\pi Df\right)^2 \cdot k \cdot T \cdot B} \cdot \frac{1}{A_f\left(t,f\right)} \qquad (16.3)$$

where the time and frequency dependence of the fading attenuation (both random) is indicated. Statistical parameters of the fading attenuation depend on the frequency range and on the kind of the system (terrestrial or satellite), moreover, different parameters are of importance for analog and for digital transmission.

In terrestrial transmission the following frequency ranges are used : 2, 4, 6, 7, 8, 11, 13 and 17 GHz. The ranges used for the satellite communications are 4-6, 12-14, 20-30 GHz (the greater frequency is used in the Earth-satellite direction and the lower one in the backward transmission).

16.3.2. Digital Radio Transmission

The cost of a radio transmission system is determined mainly by the required transmission power, by the required bandwidth, and by the complexity of the signal processing.

Without going into details, let us remark that in the case of digital transmission the above factors can be freely converted among each other by choosing the appropriate modulation system. Further, frequency seems to be the most `expensive' factor, and the expenses of the signal processing strongly depend on the actual technological level. Fortunately, these expenses are continually decreasing with the new developments in technology.

First, let us examine the transmission of a single digital signal disturbed by additive Gaussian noise. Modulating a carrier with binary signals, two different signal shapes can be generated. For the amplitude, frequency and phase modulation (ASK, FSK and PSK) these signal pairs are as follows:

$$\begin{aligned} ASK\!: s_1(t) &= \sqrt{2}A\cdot\cos\omega_c\cdot t & FSK\!: s_1(t) &= \sqrt{2}A\cdot\cos\omega_c\cdot t \\ s_2(t) &= 0 & s_2(t) &= \sqrt{2}A\cdot\cos(\omega_c+\Delta\omega)t \end{aligned} \tag{16.4}$$

$$\begin{aligned} PSK\!: s_1(t) &= \sqrt{2}A\cdot\cos\omega_c\cdot t \\ s_2(t) &= \sqrt{2}A\cdot\cos(\omega_c+\Delta\Phi)t; \quad t\in(0,T) \end{aligned}$$

where T is the bit time, A is the effective amplitude, ω_c is the angular frequency of the carrier and $\Delta\omega$ and $\Delta\Phi$ are the frequency deviation and the phase deviation, respectively. The error ratio is minimum when PSK with $\Delta\Phi$€=€π is used, i.e. if $s_1(t)$ = -$s_2(t)$. With an ideal receiver, the error ratio would be

$$P_e = \frac{1}{2}erfc\left[\sqrt{E/N_0}\right] \sim \frac{1}{2}\sqrt{N_0/\pi\cdot E}\,\exp(-E/N_0), \tag{16.5}$$

where E = $A^2\cdot T$ is the energy of one bit and N_0 = $k\cdot T$ is the noise spectral density. (Mark ~ represents the asymptotical equality.)

While the phase modulation is more or less plausible, demodulation is not so easy: to distinguish the signals (s_2 = -s_1), the initial phase has to be known. Systems requiring a *phase reference* for the demodulation are called *coherent*.

Not only the PSK but any other system where the phase of the carrier is important has to be transmitted coherently and the reference phase is then recovered from the transmitted signal. If there is no such reference, the transmission is called *non-coherent*.

Fig 16.3. illustrates how the decision is made in the case of PSK modulation. Since there is one frequency only, the usual vectorial presentation can be used (let us remark that it can also be used in a more general case). The absolute value of the signal vectors is equal to the square root of the energy. The vertical line is the so-called *decision threshold*.

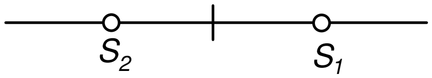


Figure 16.3. Vectorial Representation of the PSK Symbol Pair

If the phase of the received vector falls into the right half-plane it is taken as s_1, if it is in the left plane it is taken as s_2. The probability of error is determined by the probability that a transmitted s_1 is phase-shifted by the noise and received with a phase falling into the left half-plane.

As far as the modulation bandwidth is concerned, it can be shown that it is in the order of $1/T$ around the carrier frequency, more precisely, it is a little bit greater if a single carrier is used the amplitude and phase of which are modulated. Because of the cost of the frequency, more economical modulation methods have to be used, especially for high-speed signal transmissions. For the given source speed, the only possibility is to increase the value of T, i.e. to concentrate more bits into one symbol. If n bit form a symbol and the carrier is not modulated then the bandwidth is reduced to $1/n$-th of its original value. On the contrary, the number of the possible states will have been increased to $M = 2^n$. It can be shown that M-ary PSK (MPSK) is the optimum modulation method if $M \leq 6$; for greater values of n, however, the simultaneous amplitude and phase modulation is a better solution. For higher number of states, M-ary quadrature-amplitude modulation (MQAM) is almost exclusively being used since it is close to the optimum. The MQAM signal can be expressed as follows:

$$s(t)=\mathrm{A}(a\cdot\cos\omega_c t+q\cdot\sin\omega_c t) \quad a,q=\frac{\pm1}{L-1},\frac{\pm3}{L-1},\dots,\pm1; \quad L=\sqrt{M}\,. \qquad (16.6)$$

The vectorial representation of the four-phase PSK (QPSK) and 16QAM signals is shown in Figure 16.4. As the price to be paid for the reduced bandwidth, a much greater power has to be used to achieve the same error ratio as previously.

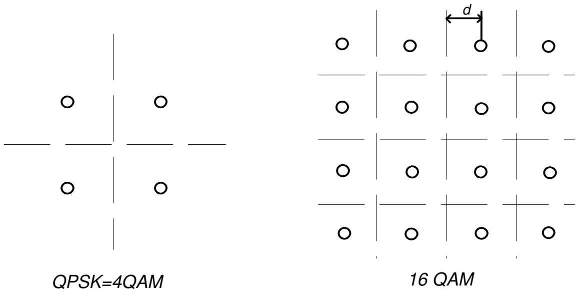


Figure 16.4. Representation of the QPSK and 16QAM Signals

Comparing Fig. 16.4. with Fig. 16.3., it can be seen that while in PSK noise greater than $\sqrt{E}$ could only cause faulty decisions then in 16QAM the decision threshold is closer so even a smaller noise level (d) may cause errors. It can be shown that if M€>€2 then the error ratio is approximately

$$P_{\mathrm{E}} = \frac{1}{2}\cdot\mathrm{erfc}\left[\sqrt{d^2 / N_o}\right] \tag{16.7}$$

where *d* is the minimum distance out of the distances of the individual signals from the decision threshold.

It can be shown that for the same error ratio the power ratio of QAM and PSK signal is

$$\frac{P_{\mathrm{MQAM}}}{P_{\mathrm{PSK}}} = \frac{2\left(2^{n/2}-1\right)^2}{n}; \qquad M = 2^n \tag{16.8}$$

i.e. for the linear reduction of the bandwidth the power has to be increased nearly exponentially. QPSK is used for low-speed transmissions while 64QAM, 256QAM and even 1024QAM systems are used in high-speed communications (for bit rates higher than 100 Mbit/s). Let us notice that QPSK is a certain kind of optimum: the bandwidth is the half of that used by the PSK while the power is the same.

As we have seen in Chapter 12., a so-called Nyquist-filter is used to avoid intersymbol interference. In some cases additional linear distortion is caused by the transmission medium itself.

Since the amplitude of the MPSK signal is constant, the nonlinear distortion of the amplifiers does not increase the error ratio. In MQAM transmission, however, the distortion has to be kept at a low level. As it can be seen from Fig. 16.5. that either the gain or the phase shift depend on the input amplitude, and the signal vectors are shifted closer to the decision threshold (d' < d) so that the error ratio increases. This can be reduced if the signal is *predistorted* so that the resulting response of the amplifier and the predistortion unit are approximately linear.

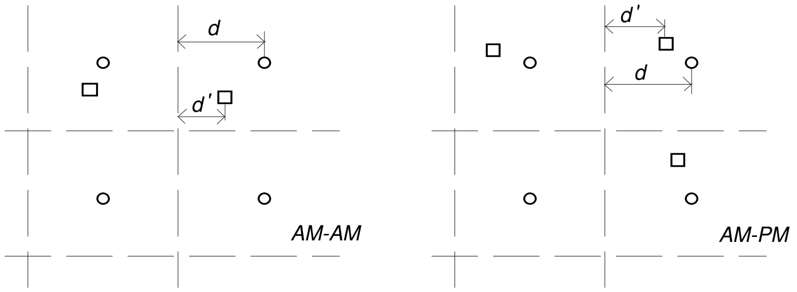


Figure 16.5. Influence of the Nonlinear Amplifier on 16QAM
(o: original vector;   €: distorted vector)

As it was mentioned, a phase reference is needed for coherent demodulation. Since the carrier is suppressed in all modulation systems discussed above, it cannot be simply recovered from the received signal by a linear filter, a nonlinear operation has to be used instead. As an example, carrier recovery used in QPSK system is shown in Figure 16.6.

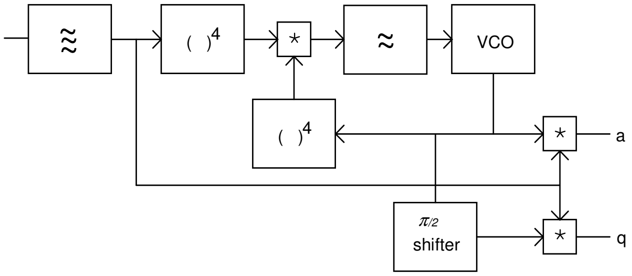


Figure 16.6. QPSK Carrier Recovery and Demodulation

As it turns out from Fig. 16.6., the fourth multiple of the four vectors points to the same direction (π). Multiplying the received signal by the reference, the demodulated baseband signal is obtained. The decision is applied to the demodulated signal. Following the decision, the shape of the original baseband signal is regenerated by the *regenerator*. *Timing* information needed for the regeneration is recovered -like the carrier recovery- by a nonlinear clock-recovery operation.

A part of the channel encoding, e.g. the conversion of the binary signals to multiple states, is done by the so-called line codec which definitely belongs to the transmission equipment.

To summarize the considerations made throughout this chapter, a digital radio section is shown in Fig. 16.7.

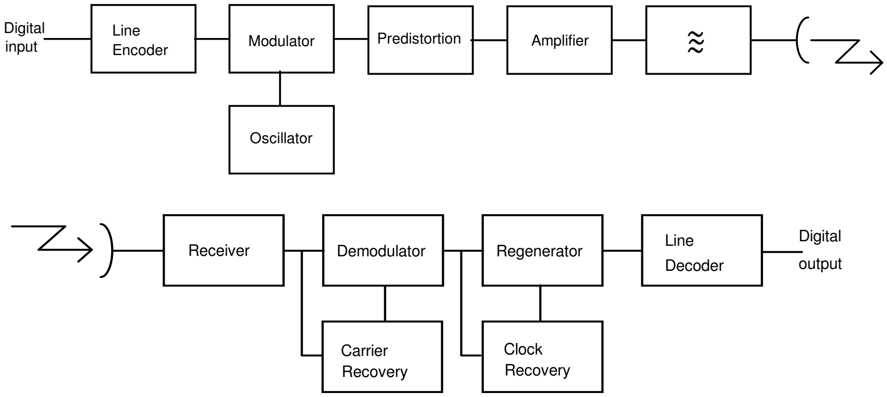


Figure 16.7. Digital Radio Section

16.3.3. Digital Radio Relays

As it was mentioned above, terrestrial radio-relay systems are realized by chains of repeater stations. Factors which have to be considered in such system are as follows:

- kind of the signal to be repeated (RF, IF or baseband),
- choice of the carrier frequencies,
- mode of compensation of the disadvantageous properties of the transmission medium (i.e. of microwaves propagating close to the ground).

Since the subsequential repeaters are independent of each other, the noise powers of individual receivers are summed when the signals received and amplified by nearly linear receivers are retransmitted. It can be seen from equations (16.6) and (16.7) that the error ratio strongly depends on the signal-to-noise ratio. For instance, if the signal-to-noise ratio is reduced by 3 dB, the error ratio increases by about three orders of magnitude. Therefore, it is more advantageous to regenerate the signal at each repeater since in this case only the erroneous decisions are summed up.

When the frequency plan of a radio-relay system is being designed, it has to be taken into account that receiver-transmitter pairs operate together, i.e. that very high and very low level signals are simultaneously present. Because of the finite backward attenuation of the receiver and the transmitter antennas, there is a danger of feedback between the transmitter output and receiver input. To avoid the self oscillation of the repeater, the transmitting and the receiving frequencies should not be too close to each other. In this case, filters tuned to each other's carrier can provide sufficient selectivity to reduce the loop gain of the unwanted loop.

As we have seen, fading (caused by the multipath propagation and by the rain-attenuation at frequencies above 10 GHz) is the most significant disturbing factor. For low bit rate transmissions (below 40 Mbit/s) fading can be considered to be wideband, i.e. the transmission medium is supposed to be a time variant attenuator independent of frequency. Fortunately, the chance of very high attenuations is very small. The attenuation caused by normal fading is compensated by the so-called *fading margin*, i.e. the power of the transmitter has to be determined in such a way the error-rate specifications are satisfied even in the case of fading. More precisely, the percentage of time is specified during which the signal-to-noise ratio may be degraded by the fading so that the error ratio be higher than 10^{-3}. Under these conditions, the communication will be of excellent quality for the greater part of time and will be unacceptably bad only during short (specified) intervals.

In high-speed transmissions, the fading cannot be considered as being independent of frequency, it becomes *selective* thus causing linear distortion. As a consequence of this distortion, the error ratio increases and even more, if the distortion is too great, the error ratio can be unacceptably high even in the absence of noise, merely because of the intersymbol interference. Of course, this effect can not be compensated by increasing the transmitted power.

Adaptive equalization or the so-called *diversity* system (or the combination of the two) are used to eliminate the selective fading. In adaptive equalization, the actual state of the channel is 'measured' in some way, then the response of the adaptive equalizer is adjusted so that the resulting response shall be close to the optimum. The most effective solution is the so-called *decision-feedback* loop containing a FIR and an IIR filter with variable weighting factors.

In the diversity system, the same information is transmitted in two different channels. Two methods can be used: in the *space diversity* system, the signal is received from two directions by two antennas and by two receivers while in the *frequency diversity* system two signals are transmitted simultaneously on two different frequencies. The principle of both systems is that the probability of having bad propagation conditions in both channels is much smaller than the probability that one of the channels will be useless.

16.3.4 Digital Satellite Communication

Since a great part of the hemisphere is visible from a geostationary satellite, it can be used as a repeater even between two very distant points on the Earth. Yet there are some differences between the terrestrial and the satellite radio communications:

a.) the waves from the satellite travel only a short path through the atmosphere thus fading is much smaller (below 10 GHz), and it can be neglected altogether,

b.) a satellite is capable to establish communication with several pairs of terrestrial stations, i.e. the satellite can be used for *multiple access*,

c.) the distance to be bridged is considerably long so that the path attenuation the *propagation delay* are very great. The entire delay of the Earth-satellite-Earth link is about 250 ms; for this reason, not more than one satellite section may be used in one communication.

Generally, QPSK is used for the satellite systems. The on-board satellite repeater is called the *transponder* which in fact may consist of a single linear amplifier. In more sophisticated systems, the transponder is a signal processing equipment which can restore and regenerate the baseband signals as well. A general block diagram of a transponder is shown in Figure 16.8. Here f_1 stands for the uplink frequency and f_2 for the downlink frequency.

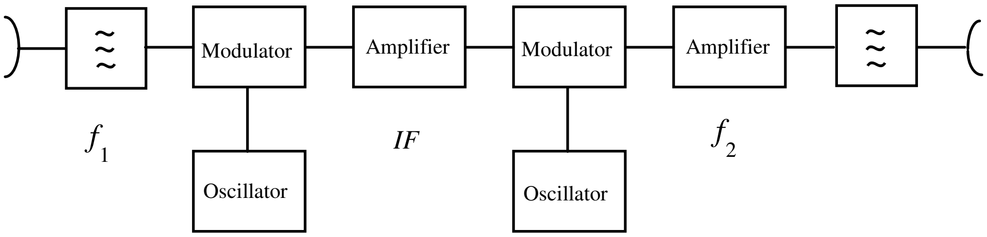


Figure 16.8. Block Diagram of an On-Board Transponder

Both frequency and time division multiple access are used in satellite communications. In the first case, the Earth stations transmit at different frequencies while in the second case different time-slots are used in the transmission. Moreover, *space division* may be used if the transponder antenna has several sidelobes. Space division can be also combined either with FDMA or with TDMA. The multiple access may be fixed, i.e. the time slots may be assigned to individual Earth stations or it may be variable in accordance with the actual demands.

16.3.5. Analog Radio-Relay Systems

Analog systems operate mostly with FDMA channels using NBFM modulation. In accordance with the central limit theorem, FDM signals are modelled by Gaussian noise.

Besides the thermal noise, *intermodulation noise* caused by nonlinear distortion is the main source of the noise. E.g. intermodulation in telephone channels results in crosstalk among the channels. Because of the Gaussian distribution of the modulating signal, the intermodulation has the same effect as the thermal noise thus the equipment have to be design so that *sum* of the thermal and intermodulation noises should be minimal.

16.4. Light-Wave Communication Systems

Signals with frequencies falling into the visible or near infrared range are transmitted almost exclusively by means of dielectric waveguides since the properties of dielectric transmission lines are very good, the attenuation and dispersion of optical waveguides are much smaller than those of the free span. So the light can be used for wide band and long-haul communications.

Almost exclusively digital signals are transmitted by optical fibres in long-haul communications. In short-haul communication, analog signals are also used, e.g. in cable TV systems, in microwave signal transmission, in interconnections among the parts of an equipment, in personal communication, etc. In the following, only the long-haul communication will be dealt with.

16.4.1. Intensity Modulated Optical Transmission

The most simple and nowadays almost exclusively used method of optical transmission is the modulation of the light intensity by a binary signal. In fact, this corresponds to the ASK modulation: zero intensity is assigned to `0' bit and a finite intensity of the light to the binary '1'. The average power of the light emitted by a laser diode is proportional to the current flowing through the diode, thus the current of the diode has to be switched corresponding to the values assigned to '0' and '1'.

This method is called the *direct intensity* modulation which works well up to frequencies of some GHz so that it can be well used practically in all systems operating at present. If the modulating frequency is higher than the so-called *relaxation oscillation* frequency, the light intensity can be varied by means of an *optical modulator*. In both cases a photo diode or an avalanche diode is used for demodulation. The noise properties of photo diodes are better. In a photo diode the number of electrons corresponds to the number of photons, more precisely, it is the η-th multiple of that, where $\eta < 1$.

In light-modulation systems, -€provided the other factors are ideal - an erroneous decision might be caused only by the shot noise due to the quantum nature of light. More precisely, when a '0' bit is being transmitted, the received signal will be zero since no additive noise is present. The optimum decision rule is thus as follows: decide on '0' if the number of the received photons is zero, and decide on '1' if the number of the received photons is greater than zero.

The probability of error can be expressed then as follows:

$$P_{\mathrm{E}}\text{€}=\text{€}\frac{1}{2}\cdot\mathrm{P}[n\geq 1|0]\text{€}+\text{€}\frac{1}{2}\cdot\mathrm{P}[n=0|1]\text{€}=\text{€}\frac{1}{2}\cdot\mathrm{P}[n=0|1]$$

(16.11)

where P[.] is the probability of the event put in the bracket and *n* is the number of detected photons; the condition is given by the binary value of the actually transmitted bit.

The number of photons in a light impulse is given by the Poisson distribution:

$$P_{\mathrm{E}}=\frac{1}{2}\,e^{-\mathrm{n}'}=\frac{1}{2}\cdot e^{-2\mathrm{n}} \qquad (16.12)$$

where n′= 2n is the average number of the photons when a binary '1' is received, while *n* is their average. Furthermore

$$\overline{n}^{\,'}=E_1/h\cdot f_c=2\cdot P\cdot T/h\cdot f_c$$

so that

$$P_E=\frac{1}{2}\cdot\exp\left[\;-2\cdot P\cdot T/h\cdot f_c\;\right] \qquad (16.13)$$

where E_1 is the energy received with binary '1', *P* is the average received power and *T* is the bit time. It can be seen that the optical power required to suppress the error ratio below a given limit is proportional to both the optical frequency and to the bit frequency. For instance, 10 photons should be received on the average (i.e. 20 for each '1') to have the error ratio less than 10^{-9}. This is the so-called *quantum limit*; in practice a power greater by about 15-25 dB is needed because of the noise of the photo diode and that of the following amplifier.

If such a long distance has to be bridged that the optical receiver cannot be provided with sufficient signal (e.g. this is the case of maritime optical cables) then repeater stations have to be inserted. At present, the optical signal has to be converted to an electrical one, regenerated and converted again to an optical signal. Solutions for the pure optical regeneration are being developed. A common drawback of both solutions is that the repeaters have to be supplied with DC power so that additional copper wires have to be included in the optical cable.

16.4.2 Coherent Optical Transmission

In optics, all demodulation procedures based on mixing the optical signal with the signal of a local oscillator are called *coherent*. Furthermore, *homodine* or *heterodine* receivers are distinguished depending on whether the frequency of the local oscillator is the same as that of the transmitter or not. The block diagram of a heterodine receiver is shown in Figure 16.9. The only difference in the homodine receiver is that it does not have a demodulator since the intermediate frequency is zero so that the signal is mixed directly to the baseband.

In both receiver types the local laser is voltage (or current) controlled. In the homodine receiver the phase is controlled by an optical PLL so that a coherent system is created in the original sense of the word. In the heterodine receiver, only the frequency of the optical oscillator is controlled and the reference phase is generated at the electrical intermediate frequency.

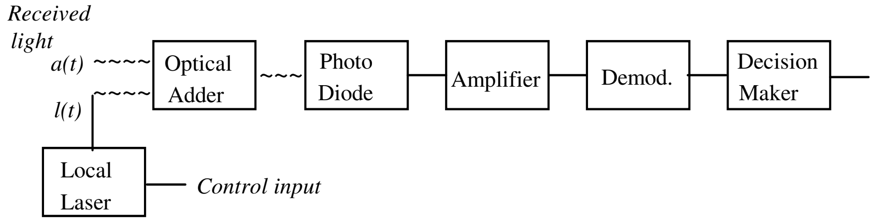


Figure 16.9. Coherent Optical Demodulator (Heterodine)

It can be shown that in coherent optical systems the error ratio of the two-state PSK is (under ideal conditions)

$$P_E = \frac{1}{2}\mathrm{erfc}[(2n\alpha)^{0.5}] \qquad (16.14)$$

Comparing (16.18) with (16.11), it can be seen that for α=1 the error ratio of the coherent and the intensity modulation systems is the same while for α=0.5 a 3€dB higher optical power is needed for the same error ratio.

Finally, let us make some interesting remarks:

- the homodine system is 'better' by 3 dB than the heterodine one; on radio frequencies the two systems are equivalent;
- from the point of view of the average power, the quality of an ideal PSK coherent optical transmission is hardly better than the quality of an intensity modulated system; the required power, however, is 3 dB less.
- the difference is greater if real conditions are considered; it can be seen from eq. (16.16) that the signal current is proportional to the local oscillator power. If the power of the local laser is chosen sufficiently high, the noise generated by the photodiode and by the preamplifier can be neglected.
- it can be assumed that optical systems of the future will apply more complex modulation procedures, similarly to today's microwave transmissions; coherent methods are necessary for such solutions.

Control questions

1. How can the factors determining the task of a transmission system be specified?
2. Draw the block diagram of a radio transmitter and receiver.
3. What are the main properties of microwaves from the aspect of transmission? What are the consequences of these properties? What is selective fading and how can it be eliminated?
4. What are the main cost factors of a digital radio system? How can the required bandwidth be reduced? What is the price of the reduction?
5. What is a coherent system and what is a non-coherent system? How can the information necessary for the coherence be obtained? What is the coherence in optical transmission?
6. What is the quality of an intensity-modulated optical transmission system determined by under ideal conditions?

Exercises

1. A microwave signal has to be transmitted with 2 Mbit/s on 2 GHz to 50 km. The diameters of both antennas are 2 m each, the noise factor of the receiver is 5 dB and 35 dB is supposed as fading margin. What shall be the transmitting power if the error ratio has to be kept below 10^{-3}? (k = 1,38. 10^{-23} V A sec/K).
2. A 25 MHz band is available for the transmission of a 155 Mbit/s signal. What kind of modulation has to be used? What should be the receiver bandwidth and the P_{MQAM}/P_{PSK} ratio?
3. A 2 Gbit/s signal is transmitted by means of intensity modulated light with 1.5 μm wavelength. What should be the peak power of the laser diode if the optical losses are about 10 dB and the noise caused by the photo diode and the preamplifier is 15 dB? (h€=€6,62·10^{-34} VAsec^2.)
4. What should be the power of the transmitter laser for a heterodine PSK system?

References

[1] Frigyes, Szabó, Ványai: Digitális mikrohullámú átviteltechnika, Mûszaki Könyvkiadó

[2] Lajtha, stb: Fénytávközlés, Mûszaki Könyvkiadó

Abbreviations

FIR	Finite Impulse Response	IIR	Infinite Impulse Response
IF	Intermediate Frequency	RF	Radio Frequency
NRZ	Non-Return to Zero	QPSK	Quaternary Phase Shift Keying
PM	Phase Modulation		
QAM	Quadrature Amplitude Modulation		

17. MOBILE COMMUNICATIONS

17.1. Main features of mobile communications

Most deterministic factor of the mobile communication is that at least one of the connected parties is moving or is changing his location in unpredictable way. As a consequence, mobile communication can be realized only by means of wireless transmission. For this purpose, terrestrial radio waves are the most convenient means. Area radiated by the radio waves is determined by the law of physics, state borders do not play any role here. The radio frequency spectrum is one the one hand a natural treasure exposed to quickly growing demand and on the other hand -because of the enormous evolution of electronics- it is more and more polluted by electromagnetic noises.

17.1.1. Regulations of Mobile Services

Soon the above problems were recognized and a wide international and national co-ordination was established for the utilization of radio frequencies. The highest level of this co-ordination is the International Telecommunication Union (ITU), which publishes and regularly updates the International Radio Regulations (IRR). The most important part of the IRR is the frequency table which shows the distribution of frequencies in the range from 9 kHz to 3000 GHz for three regions of the world, managed by about 30 radio services. As far as mobile communication is concerned, terrestrial, aerial, maritime and satellite services are specified by the IRR.

The next level of international co-operation is regional. On the regional level the frequency distribution is further refined, in certain cases the right of frequency usage is even bounded to observance of detailed technical specification. On this level, frequency bands of various *private* mobile services operated for safety, technological etc. reasons and *public* services accessible by anyone are distinguished.

On the national level, responsible authorities work out a National Frequency Band Distribution Table which -based on the rules of the international distribution- allocates the frequency bands for groups of users (e.g. civil, governmental, etc.) accordingly to the domestic conditions. It is also the responsibility of the national authority to allocate operating frequencies for interested private or legal persons. Such a frequency allocation is valid for a determined time and given location and for equipment specified in detail in separate standards. These conditions aim to avoid mutual interferences and help the economical use of the frequencies.

As we have seen, giving licenses for frequencies and for equipment operating at them is a national authority which has essential influence on the interest of the users, operators, sellers and manufacturers of the communication equipment and systems. These interests are substantially influenced by frequency management practice and system standartization. Due to that, frequency distribution and utilization is focused also by those not belonging to the profession and in more important questions decision can be made only following a wide conciliation of interests. Designers and operators of radio equipment must therefore know several stipulation and specification to be able to solve their task. This is especially true if the equipment is the part of the public communication network which is the greatest synchronized machine of the Earth.

In the following we discuss some essential questions of the terrestrial mobile communication and then we describe various systems, mainly from the users' point of view.

17.1.2. Structure and Operation of Terrestrial Mobile Services

Majority of terrestrial mobile communication services operate in the 30-1000 MHz (VHF) range. In Hungary, civil services use 80, 160, 450 and 900 MHz ranges. Channel spacing at these ranges is 12.5, 20 or 25 kHz. The majority of communications is carried out between a highly located fixed base station and moving stations installed in a vehicle, carried on back, in hand or in pocket.

The position of the moving station is dependent on the kind of vehicle which is generally moving on the Earth surface. Operation of portable and hand-held sets must be provided at any places where persons use to move, i.e. inside building, as well. This is, however, not an easy task since the radio waves are strongly attenuated when penetrating through the walls. The operational range of handy sets is even more limited because of the smaller output power given by the size and duration of the battery.

The simplest mobile communication system is used by the private *dispatch services*. The communication is alternating, called simplex, i.e. the antenna is alternately switched to the transmitter and the receiver by a switch. The carrier frequency could be the same in both directions (base to mobile and mobile to base). This is called the *single-frequency simplex* mode. Practically, however, *two-frequency simplex* mode is used in which the two frequencies differ by at least 2 to 5 percent of the operating frequency (Δf_D, the duplex frequency shift).

The reason of two-frequency operation is to avoid interference among the systems of different users located in the same region. If the transmitting frequencies of all base stations have fallen into the same frequency range, it would be impossible to receive the signal while another base station is transmitting. Thus a gap between the transmitting and receiving frequencies is provided for proper operation of filters at the receivers' input. The two-frequency mode also makes possible simultaneous operation of the transmitter and receiver of the same station (*duplex mode*). In duplex mode the transmitter, receiver and the antenna are interconnected by a selective three-port called the *duplex filter*. In the private systems -because of economical reasons- duplex filter is used only at the base station (*half duplex* mode). By using two frequencies, it is also possible to retransmit the received signal by cascading the receiver and transmitter of the base station. This is called the *mobile relay* mode which is frequently used in private systems.

The aim of *mobile public radiotelephone* system is to provide service analog to the wired telephone so that it is necessarily operating in duplex mode.

Let us remark at this point that for the duplex mode, duplex filter is necessary only in such (analog) systems in which the simultaneous bidirectional communication takes place strictly in real time. In the case of digital transmission, duplex mode can be realized in single-frequency case, as well, since the information of the two communicating parties is transmitted in different time slots.

17.1.3. The Mobile Radio Channel

Technical solutions used in mobile radio systems cannot be understood without the knowledge of the radio channel. For this reason, let us follow the way of the signal from the antenna of the base station up to the receiver of the mobile station. During the propagation, transmitted wave is exposed to several effects. Since the mobile station is close to the ground, the propagation is considerably influenced by the relief, by the building up and by the vegetation of the terrain. The first, well separable effect is the multipath propagation. This effect is caused by elevations , hills, high buildings, etc. which reflect the wave so that it arrives to the receiver from several directions with considerable amplitude and group-delay differences. This effect has already been described by equation (6.7)

$$y(t) = \sum_i a_i x(t - \tau_i)\ .$$

Propagation delays are defined here as the *deviation* from the arrival of the first (direct) wave. Plotting the coefficients a_i against the values of $\tau_{i,}$, we obtain the *group delay profile*. This is usually a continuous curve with local maxima corresponding to the dominant paths. The shape of the curve may strongly differ depending on the specific conditions, but on the average, corresponding to basic physical expectation, a_i has a decreasing tendency with growing values of τ_i. Therefore the group delay is usually approximated by a negative exponential function.

Because of the multipath propagation, the transfer function $a(f)$ becomes uneven. For the sake of simplicity, *coherent bandwidth* is defined as the frequency band within which the attenuation does not show any significant difference. Supposing exponential delay profile, the coherent bandwidth B_C can be written as

$$B_C = \frac{1}{2\pi \cdot S} \qquad (17.1)$$

where S is the standard deviation of the propagation delays. According to measurements, the value of S is between 0.25 and 1 μs which corresponds to coherent bandwidth of 160...640 kHz. This value is smaller than the duplex frequency shift Δf_D, consequently considerable difference may exist in the propagation attenuation from base-to-mobile and mobile-to-base transmission, respectively.

Let us assume in the following that an unmodulated carrier is transmitted and the strength of the electric field is observed in a far point where the wave is arriving scattered by multiple reflections from buildings trees and other objects. As the consequence of that, a rather complex standing wave field is formed. Specific feature of this field is that the maxima of the field are following each other every half wavelength on average and that fluctuations as high as 30 to 40 dB within a quarter of a wavelength are quite common (Figure 17.1).

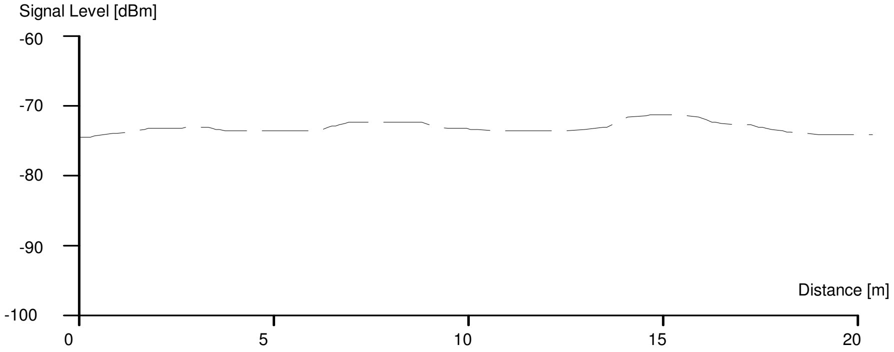


Figure 17.1. Short Distance Field Distribution of a Mobile Communication

Since the physical model used above (the sum of several vectors of incidentally different magnitude and phase) is described by the Rayleigh probability distribution, the signal fading is called the Rayleigh (or fast) fading. The effect has not take place if there is no moving object between the base station and the observed point and slow atmospheric changes are neglected. The field is steady in that case until the receiver antenna starts to move.

If the station is moving, the Doppler effect shall be taken into consideration, too. The Doppler frequency shift is

$$f_D = \frac{v}{c} f_0 \cos\alpha = \frac{v}{\lambda} \cos\alpha = f_{D\max} \cos\alpha\,, \tag{17.2}$$

where f_0 is the transmitted frequency, λ is the wavelength, v is the velocity of the mobile station, α is the angle of wave incidence with respect to the direction of the moving vehicle and $f_{\mathrm{Dmax}}=v/\lambda$.

Since the waves arrive practically from all directions in an urban area, as the consequence of the Doppler effect, spectral component f_0 will broaden to a continuous band having the width $2f_{\mathrm{Dmax}}$.

If, for instance, v = 15 m/s (54 km/h) and f = 900 MHz (λ = 0,33 m), then f_{Dmax} = 45 Hz which is not disturbing in voice quality transmission since the bandwidth of the modulating signal is 300...3400 Hz, so that the spectrum of the modulated signal is initially 300 Hz from the carrier. However, if the velocity of the vehicle is much greater (e.g. aero service) or the carrier frequency is increased then the Doppler and the modulated spectra may overlap which can be avoided only by complex frequency controlling procedures. Besides other, this is the reason why the mobile communication shall not be used above 3 GHz.

As the result of the effects described above, modulated signal appearing on the antenna of the mobile station is given as

$$u(t) = \mathrm{Re}\ [r(t)\ \exp\ (\mathrm{j}\omega_0 t)] \qquad (17.3)$$

where $r(t) = \sum A_\mathrm{m}(t)\ e(t-\tau_\mathrm{m})$
$A_\mathrm{m}(t) = \sum c_\mathrm{mn}\exp[-\mathrm{j}(\Theta_\mathrm{mn}+\omega_\mathrm{mn}t)]$;
$e(t)$ is the complex envelope of the transmitted signal;
$A_\mathrm{m}(t)$ is the amplitude of the m-th wave;
c_mn *is the* n-th spread component of the m-th wave;
Θ_mn is the phase of the spread component;
ω_mn is the angular Doppler-frequency of the spread component.

Eq (17.3) describes well the signal in proximity of a given location but gives no information about the signal changes as functions of the distance from the transmitter, of the heights of antennas and of the operating frequency. This parameters are treated by theoretical models discussed in Chapter 9. (line-in-sight propagation, multipath propagation), these, however, can not take into consideration with adequate accuracy all physical phenomena affecting the propagation. Thus in mobile communication the above mentioned main installation parameters are determined by empirical or semiempirical models.

If the propagation is blocked by such an object or objects which can be geometrically well modelled, then the attenuation can be computed in deterministic way with a relatively good approximation on the base of the diffraction model described in Chapter 9. Mobile stations operate mostly on wavy terrain possibly shadowed by trees or on urban areas. Influence of such factors is impossible to describe by deterministic models, hence statistical approach based on series of measurements is used instead. A typical result of such a measurement of the electrical field strength is shown in Fig. 17.2.

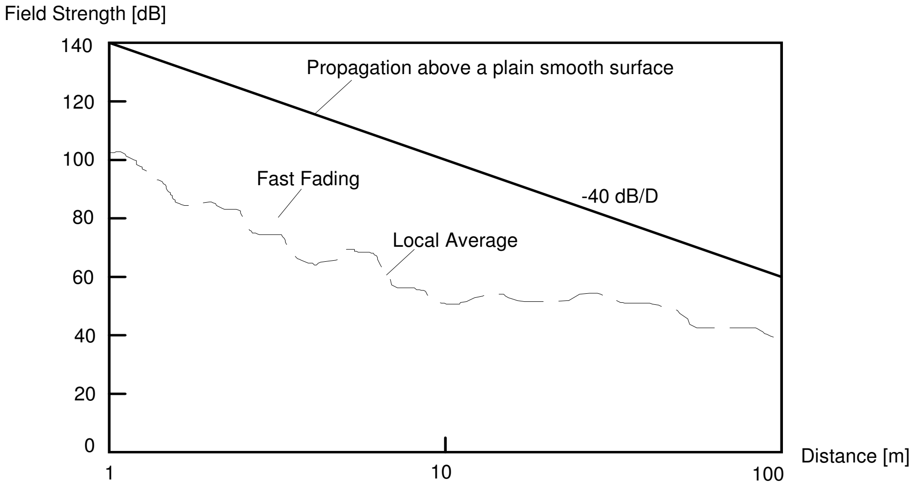


Figure. 17.2. Distribution of the Electrical Field in a Mobile Connection

On Fig. 17.2., three different phenomena causing the change of the electrical field strength when moving off the transmitter can be observed. The fastest change is the Rayleigh fading which, smoothed by averaging, results in the dashed line representing the local average. Ripples in the local average are due to the shadowing effect caused by different obstacles and by the vegetation. Besides the ripples of the local average, a monotonous decrease of 40 dB/decade proportional to the distance can also be noticed. This corresponds to the attenuation slope of multipath propagation described in Chapter 9., with about a 20-30 dB attenuation surplus. The general equation describing this monotonous decrease is

$$E = A - B \lg d \quad \text{dB/}\mu\text{V/m}, \tag{17.4}$$

where, for the above figure, A =140 and B =40. If the dashed curve obtained by measurement is approximated by the line described by Eq. (17.4) then the values of A and B can be computed. Besides that, the deviations of local averages from the line can be determined. Since these deviations -if expressed in dB- are in good accordance with the normal distribution, field changes caused by the shadowing is called *lognormal fading*.

Median value of deviations from the regression line is zero to a certain distance. This means that within this region approximation of the attenuation by a single line is fairly good and that the lognormal fading is characterized by its standard deviation. If we want to keep this property even for greater distances, then two or more regression lines of different slope have to be used.

Values *A* and *B* obtained from the regression depend on the height of the transmitter and that of the receiver antenna, on the waviness and on the degree of urbanization of the area and on the frequency. Since the frequency bands used for mobile communications are relatively narrow, the frequency dependence can be neglected within a given band.

In practical computations the communication is characterized by the section attenuation between the two given antennas. Providing the field is described by Eq. (17.4.), the section attenuation is given as

$$a = C + B \lg d \quad \text{dB}. \tag{17.5}$$

Example: For the frequency range around 900 MHz, antenna heights h_B = 50 m and h_M = 1,5 m in densely urbanized areas the section attenuation for distances from 1 to 40 km is:

$$a(50) = 123{,}3 + 33{,}7 \lg d \quad \text{dB}. \tag{17.6}$$

The same for the 1800 MHz range:

$$a(50) = 133{,}2 + 33{,}8 \lg d \quad \text{dB}. \tag{17.7}$$

In Eqs (17.6) and (17.7) *d* is given in km and the result is the median value of the lognormally distributed section attenuation. The standard deviation of the lognormal distribution at 900 MHz in urban areas is σ = 6...6,5 dB, while in open rural areas σ = 8 dB. It is also worth mentioning that changes caused by meteorological factors can be neglected up to distances of about 30 to 40 km.

17.2. Private Systems

The purpose of a private (or dispatch) service to co-ordinate the operation of a group of people who are generally moving while working. The simplest of such a service can be formed out of a base station and some mobile or handy radio stations. The dispatcher is usually in the proximity of the base station and his control panel is connected to the base station transceiver via a four-wire audio circuit and some further lines required for the control functions. Traffic is carried out generally in half duplex mode which for the mobile stations has the advantage of interrupting the transmission of the base station in the case of emergency. In half duplex mode, the dispatcher's transmission is usually heard by the all mobile stations while transmissions of the mobiles are received only by the dispatcher.

If the dispatcher is located far from the base station, leased line or a dedicated radio connection is used to connect the control panel to the transceiver. Since such solutions are expensive, mobile relay mode is rather used instead, i.e. the dispatcher is equipped by the same mobile set as the mobile stations. Unlike in the half duplex mode, in mobile relay mode all the station are mutually hearing each other which, in certain services (e.g. taxi), can be encountered as an advantage.

To decrease the risk of mutual interferences, it seems reasonable to activate the base station transmitter (carrier) only for the time of the message (modulation). To make the dispatcher's work easier, in half duplex mode this used to be done by a voice-operated automatic switch, while in mobile relay mode carrier incoming to the receiver is sensed.

It may occur that several independent private services are operating on the same area and their individual traffic is much less that what can be carried by one base station. In such a case, it is more economical to assign a single common base station to these services. The drawback of this solution may be that the mobile stations have to listen to messages not concerning them. This can be avoided by selective group-call, the simplest solution of which is assignment of an audio frequency code to each group and to activate by them the (otherwise idle) mobile receivers of the group belonging to the code. Of course, the same method can be used for personal selective calling if the nature of the services requires such a feature.

Installing personal selective code transmitters to the mobile stations, it is possible to automatically identify the mobile station. This can be advantageous not only for security reasons but also for shortening the communication overhead (verbal introduction to the dispatcher and acknowledgements of dispatcher's messages). Instead of, it is enough to send the identification code taking time of about 0.3-0.4 s which, in the case of short messages, results in significant decrease of traffic. As an example to this solution, the average holding time of a taxi service has been decreased by this method from 30 s to 15 s so that the number of cars serviced by one base station could be doubled.

17.3. Public Mobile Telecommunications

17.3.1. Main Features and Considerations

The obvious aim of mobile telecommunication is to provide services similar to the wirebound communication even if the user is moving or changing frequently his location. Public mobile services are sorted according to the nature of their use as follows:

- public mobile radio systems,
- cordless telephone sets,
- radiopaging systems,
- mobile data transmission systems.

The aim of the public mobile radio telephone systems is to connect "anyone to anybody", from "anywhere to anywhere" at "any time". Concerning this specification, former systems had several limitations, especially the area of motion was strongly limited, usually at least by the state borders.

Cordless telephone sets have been developed to provide the user with the freedom of motion within some dozens of meters from the base set connected to the PSTN.

On the contrary, the public radiopaging systems can be used on great area (usually throughout a whole country), if only for limited data transmission.

Depending on the data transmission rate and the network structure (point-to-point, point-to-more points, etc.), several variations of mobile data transmission systems have been developed.

Similarly to the general trends in telecommunications, the mobile telecommunication systems can also be characterized by following tendencies:

- wide usage of digital solutions,
- effort for regional and global unification, respectively,
- emphasize to the personal character of the communication.

Personal character of the telecommunication can be supported one the one hand by the pocket size of the mobile set, on the other hand (when a terminal of greater size is used), by a identifier card (smart card). Word-wide standardization is the clue for further expansion of the area of motion. Not negligible consequences of that shall be the considerable increase of production volume and decrease of prices.

17.3.2. Pan European Mobile Telecommunication System

As a first result of unification, a regional (European) system called the Global System for Mobile Communication (GSM) has been developed by the countries of the European Community and adapted also in several other countries (e.g. Australia, China, India).

To cover the area of telecommunication, GSM uses the well proven cellular principle (see Fig. 17.3.).

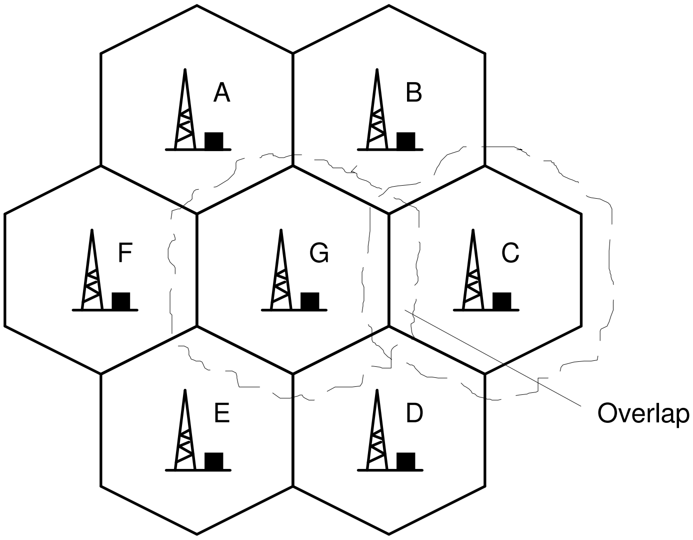


Figure 17.3. Cellular Division of the Communication Area

As can be seen from the figure, the whole area of interest is divided into cells of hexagonal shape each covered by one Base Transceiver Station (BTS), marked as A,B, ... G in the figure.

The whole system (Fig. 17.4.) consists of Base Transceiver Stations, Base Station Controllers (BSC). Mobile Switching Centre (MSC) and Operational and Maintenance Centre (OMC). The MSC is connected to the Public Switched Telephone Network (PSTN).

Every mobile station is assigned to an MSC and its data are stored there in so called Home Location Register (HLR). Data of a roaming mobile station (belonging to another MSC but currently being in the area under consideration) are written through the signalling channel into the Visiting Location Register (VLR).

The security and safety of both the service operator and the user is provided by Authentication Centre (AUC). Another register called the Equipment Identity Register (EIR) prevents from illegal usage in the case the mobile set is stolen or it reports an invalid identification number (a fake).

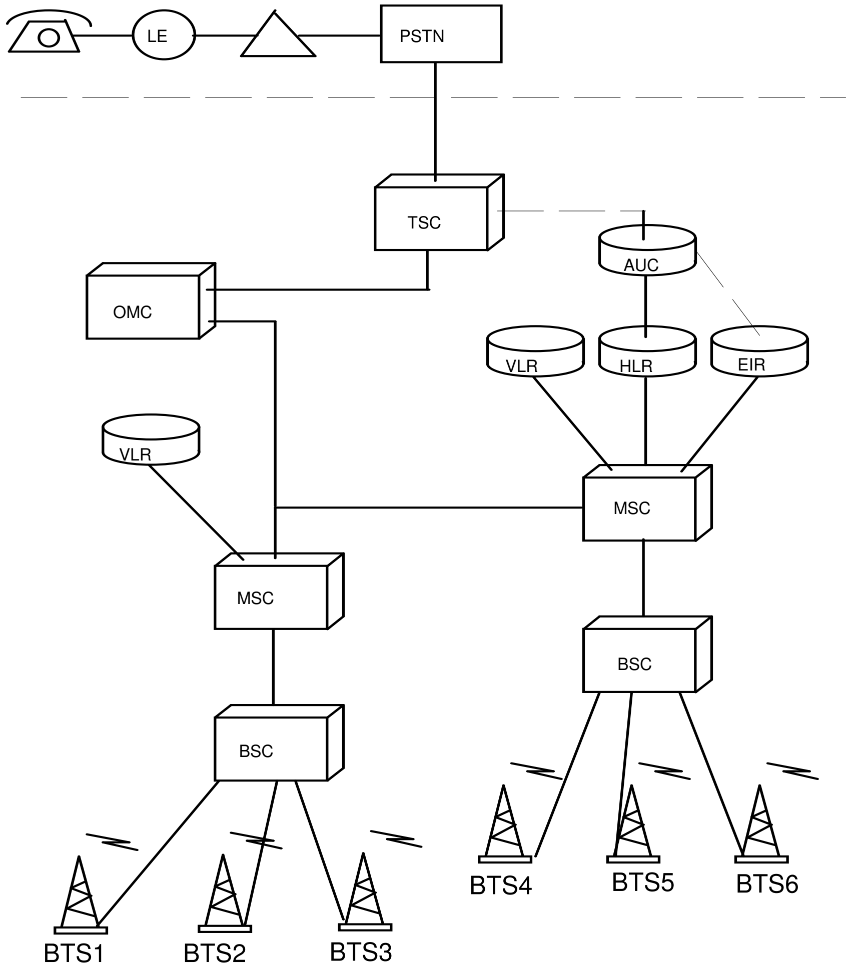


Fig. 17.4. Block Diagram of the GSM System

Main parameters of the GSM system are as follows:

• Operating frequency range,	M - B: 890 ... 915 MHz,
•	B - M: 935 ... 960 MHz
• FDMA channel bandwidth:	200 kHz.
• TDMA in each FDMA channel	8 timeslots, 270,83 kbit/s,
• Modulation scheme:	GMSK.
• Frequency hopping:	270/s.
• Error controlled source coding:	13 kbit/s, 22,8 kbit/s.
• Transmit-to-receive delay:	1,15 ms.
• Transmitter power:	20; 8; 5; 2; 0,8 W.

In parallel to the GSM, a personal mobile radio system working at about 1800 MHz has been developed. This system is much alike the GSM and is called the Digital Cellular System (DCS 1800).

Detailed description of the above systems is beyond the limit of this book. The interested reader may refer to the bibliography.

17.4. Expected Evolution Trends

Recently mobile communication is one of the fastest developing areas of telecommunications and for that it is hard to estimate precisely what can be expected. Just because of the fast progress, current projects are focused on systems in which old and new solutions work together. Digital systems offer a suitable base to that compatibility since the various formats of digital data can easily be converted among each other.

Since 1987 the International Telecommunication Union (ITU) is working on Future Public Land Mobile Telecommunication System (FPLMTS). This system will unify all public systems so far operating separately and probably offer a solution for the private systems, as well ,since being suitable for group calls and other dispatcher services.

CEPT is working out a Universal Mobile Telecommunication System (UMTS), with the aim to put to a common platform

- the Private Mobile Radio (PMR) services,
- public mobile services (e.g. GSM),
- personal radio, services (e.g. DCS 1800),
- cordless telephone sets
- public radio paging services,
- satellite mobile services.

As it turns out from above, the two concepts do not differ to much from each other and although neither of them have been standardized, several elements of the European system are already in use. Therefore, it is important to keep an eye on the European trends when forming domestic systems.

Control questions

1. Which national and international organizations are involved in frequency management and what is their role?
2. What is the meaning of frequency distribution, frequency allocation and frequency licence?
3. What are the mobile services according to International Radio Regulations?
4. What is the structure of a private mobile service?
5. What kind of transmission modes are used in mobile services? What are their advantages and drawbacks?
6. Describe the model of the mobile channel and give its main properties!
7. What types of fading can have a signal of a mobile radio channel?
8. Describe the various public mobile communication systems!
9. Depict the GSM system and name its main elements!

Bibliography

[1] Cox, D. D.: Correlation Bandwidth and Delay Spread Multipath Propagation characteristics for 910 MHz Urban Mobile Radio Channels. IEEE Trans.on Communications, Vol.COM-23, No.11, 1975.

[2] Jakes, W. C.: Microwave Mobile Communications. John Wiley, 1974.

[3] ETSI-SMG: European Digital Cellular Telecommunication System (Phase 2); Radio Network Planning Aspects. GSM 03.30; January 1993.

[4] Okumura, Y. et al.: Field Strength and its Variability in VHF and UHF Land-Mobile Radio Service, Review of the Electrical Communication Laboratory, NTT (Japan) Vol.16, No. 9-10, 1968.

[5] Ökrös Tiborné: Egységes európai digitális mobil távközlõhálózatok. PKI Közlemények, 40.szám, 1992.

[6] Ökrös Tiborné - Dr.Dárdai Árpád: Ultrarövidhullámú rádiótelefon-berendezések és rendszerek. Magyar Posta Központja, 1990.

[7] Ökrös Tiborné: Cellás hálózatok optimális kialakítása. Magyar Távközlés, 11.szám, 1991.

Abbreviations

ITU	International Telecommunication Union
IRR	International Radio Regulations
CEPT	Conference of European Posts and Telecommunication Administrations
PSTN	Public Switched Telephone Network
GSM	Global System for Mobile Communications
BSC	Base Station Controller
MSC	Mobile Switching Center
OMC	Operation and Maintenance Centre
HLC	Home Location Register
VLC	Visiting Location Register
AUC	Authentication Centre
EIR	Equipment Identity Register
GMSK	Gaussian Minimum Shift Keying
DCS	Digital Cellular System
FPLMTS	Future Public Land Mobile Telecommunication System
UMTS	Universal Mobile Telecommunications System