Introduction to Digital Communication

This chapter introduces some basic terminology, describes a model for communication systems and networks and describes some characteristics of data.
After this chapter you should be able to: define the terms introduced this chapter; compute information, entropy, bit, symbol, bit error and frame error rates; compute throughput; convert numbers between different number bases and bit and byte orders.

Model of a Communication System

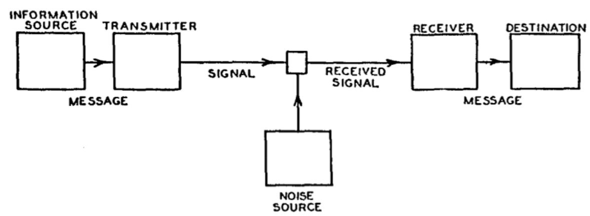


Fig. 1—Schematic diagram of a general communication system.

The diagram above shows a model for a communication system that includes the following[1]:

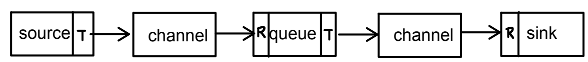
- information source - generates a sequence of "messages," taken from a finite set of possible values. The messages might be voltage levels that convey a sound or image. Or the messages might be more abstract information called "data" which could represent, for example, the characters or words in a document or perhaps numbers whose meaning is unknown ("opaque") to the communication system
- transmitter - a device that converts the messages into a time-varying voltage or current (a "signal") that can be carried over the channel
- channel - carries the signal from the transmitter to the receiver. The channel often distorts the signal and adds another, random, signal to it called "noise"
- receiver - a device that attempts to recover the messages that were transmitted from the received signal
- destination - (sometimes called a "sink") such as a person or computer that makes use of the information

[1]The diagram is from Claude Shannon's fundamental paper, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656, July, October, 1948.

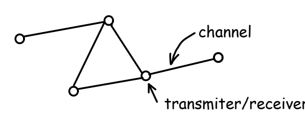
Networks

Since the number of messages is finite, we can represent each one as a number represented as a sequence of digits. These are called digital communication systems.

Information often travels over more than one channel, each channel having an associated transmitter and receiver. Digital communication systems typically operate in a "store-and-forward" manner where messages are grouped into "frames" or "packets" and each received frame is stored ("queued") before being transmitted (forwarded) over another channel:



A collection of these is a network:



Exercise 1: Give an example of a communication network. What are the information source and sink? What channels does it operate over? What transmitters and receivers do you think are used on each channel?

Overview

In this course we will try to answer the following questions about communication systems:

- How do we encode information into messages?
- How do we convert binary digits (bits) to signals? (line codes)

- What are some useful ways to measure channels? (e.g. bandwidth, linearity, noise)
- How fast can we transmit data through a channel? (information rates and capacity)
- How often do errors happen? (error rates) How can we detect and recover from these errors? (error detection and correction)
- How do we group bits for transmission over a channel? (framing)
- How can several users share the same channel? (multiple access and multiplexing)

Digital Communications

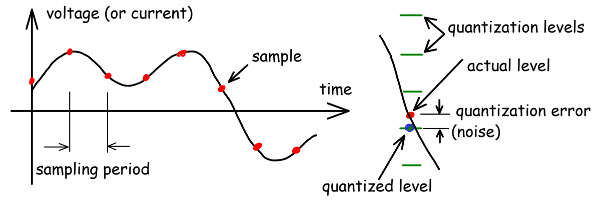
Digital communication systems communicate messages by encoding them as sequences of **b**inary dig**its** which are called bits.

Digital communication systems can be used to transmit digitized speech or video as well as more abstract information such as text or computer software, usually called "data".

The terms "data communications" and "digital communications" are often used interchangeably because today the same communication systems carry both data and digitized waveforms.

Sampled Waveforms

To represent analog (e.g. speech) waveforms as binary data, the analog signal – a voltage that is continuous in time and level – is sampled (measured) at a regular rate called the "sampling rate." Each sample is then converted into a binary number with a fixed number of bits. Thus the signal becomes discrete in both time and voltage. In other courses you will learn why the sampling rate (samples per second or Hz) for a signal must be twice the highest bandwidth. You'll also learn that quantizing a signal to B bits per sample introduces an apparent "quantization noise" with a power about $6B$ dB less than the signal power. This results in a signal-to-noise power ratio (SNR) of $6B$ dB.



Exercise 2: Speech is intelligible even if only frequencies below about 4 kHz are transmitted. What is the minimum sampling rate that should be used to sample speech if we first remove frequencies above 4 kHz?

Exercise 3: A signal-to-noise power ratio of about 48 dB is considered "toll quality" (the SNR conventional telephone networks provide). How many bits per sample are required to obtain a quantization SNR equivalent to "toll quality" speech?

Exercise 4: What if the signal was a video signal with a 5 MHz bandwidth and required a quantization SNR of 40 dB?

Exercise 5: How many bits per second need to be transmitted in these two examples?

Words, Byte and Bit Orders

Bits are often organized into "words". Words of 8 bits are called bytes (or "octets" in many standards documents). Words composed of multiple bytes (2, 4 and 8 bytes) are also common.

It is important that communication systems preserve the order of the bytes and the order of the bits within the bytes.

Byte Order

Values stored in computer memory can only be addressed by byte. Thus byte order (but not bit order) is also a concern when multi-byte words are stored in memory or in files.

The bytes in a word can be ordered as the bytes are written, from "most significant byte" (MSB) to "least significant byte" (LSB). This is sometimes called "big endian" order. The reverse order is called "little endian". Most Internet protocols use "big endian" byte order which is sometimes called "network order."

Bit Order

Data is often transmitted over a communication system as individual bits. In this case the bits in a byte can be transmitted least-significant bit first (lsb-first) or most-significant-bit first (msb-first).

Many communication systems (e.g. Ethernet) transmit data with bytes ordered MSB-first but with the bits ordered lsb-first.

Exercise 6: Write the sequence of bits that would be transmitted if the 16-bit value 525 was transmitted with the bytes in little-endian order and the bits lsb-first. Write the sequence of bits that would be transmitted in "network order" and the bits msb-first.

Binary Numbers

When a sequence of bits represents a number, the number may be "unsigned" (non-negative) or signed, typically in two's-complement format. You should be familiar with these encodings from previous courses.

Binary numbers are often written in hexadecimal notation because it allows 8-bit values (bytes) to be written using two hexadecimal digits. Hexadecimal digits are `0` to `9` and `a` through `f` (representing values from 10 through 15). The prefix, "0x" as used in the C programming language, will be used to indicate hexadecimal notation.

Files

Data is often stored in files. A file is an ordered sequence of bytes. Files often have associated metadata such as a file name, length, date, type, etc. The bytes stored in a file can represent text or they may have other interpretation such as a computer program, sound or video.

Characteristics of Data Sources, Channels and Sinks

Information Rate

We can model sources as generating one of a limited number of messages. For example, the messages might be letters, words, pixel values, or measurements. Different messages will often have different probabilities. The probability of a particular message is the fraction of messages of that type.

Exercise 7: We observe a source that outputs letters. Out of 10,000 letters 1200 were 'E'. What would be a reasonable estimate of the probability of the letter 'E'?

We define the information that is transmitted by a message that occurs with a probability P as:

$$I = -\log_2(P) \quad \text{bits}$$

For example, a message with a probability of $\frac{1}{2}$ conveys 1 bit of information. While one with a probability of $\frac{1}{4}$ carries 2 bits of information. Thus, less likely messages carry more information.

The information rate (also known as the "entropy") of a source in units of bits per message can be computed as the average information generated by the source:

$$H = \sum_i (-\log_2(P_i) \times P_i) \quad \text{bits/message}$$

where P_i is the probability of the i'th message.

Exercise 8: A source generates four different messages. The first three have probabilities 0.125, 0.125, 0.25. What is the probability of the fourth message? How much information is transmitted by each message? What is the entropy of the source? What is the average information rate if 100 messages are generated every second? What if there were four equally-likely messages?

We will see later in the course that there is a limit, called the "capacity," (C) for the information rate that can be transmitted over a given channel.

Data and Bit Rates

The rate at which bits, symbols or messages are generated by a particular system, independent of their probabilities, is the "data rate".

Data rates can be specified in different units. When the data rate is measured in bits per second (bps or b/s) it is also called the bit rate. Unfortunately, the same acronym is often used for "bytes per second". To avoid confusion it's best to spell out the units if the meaning is not clear from the context.

Some people use the convention that a capital 'B' indicates bytes and a lower-case 'b' indicates bits, but you should not assume this convention is universally understood.

In this course, and in almost all technical documents (equipment specifications, interoperability standards, data sheets, etc.), "bps" or "b/s" means "bits per second".

Computer storage is often measured in units with prefixes that are powers of 2 (e.g. kilo means 2^{10} or 1024 rather than 1000 or 10^3). However, communication system data rates always use units that are powers of 10 (e.g. 1 kb/s is 1000 b/s).

Exercise 9: How long will it take to transfer 1 MByte at a rate of 10 kb/s?

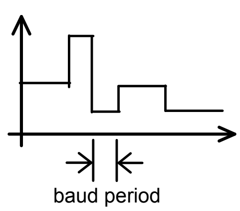
Symbol and Baud Rates

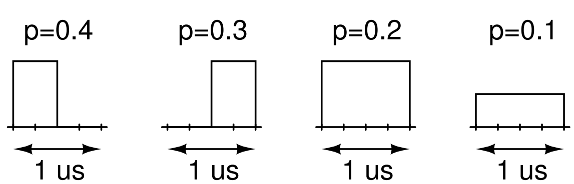
A signal is a voltage or current that varies with time; typically to convey information. The purpose of a transmitter is to convert data, in the form of bits, to a signal that carries information for transmission over a channel.

The IEEE's standards dictionary[2] defines a symbol as *"The smallest unit of data transmission on the medium."*

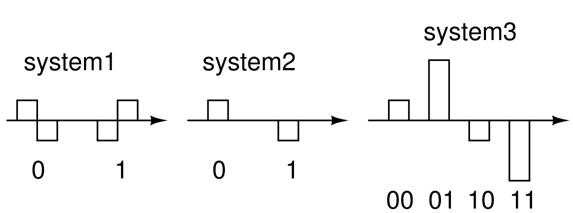
Often the transmitter uses a group of n bits to select from one of 2^n different "symbols", to transmit. The rate at which these signals are transmitted is called the "symbol rate". The symbol rate is equal to or lower than the bit rate(*why?*).

The IEEE dictionary defines a baud as *"A unit of signaling speed equal to the number of discrete conditions or signal events per second, or the reciprocal of the time of the shortest signal element in a character."*. For practical purposes we can consider "baud rate" and "symbol rate" to be synonymous.





Exercise 10: A communication system transmits one of the symbols above each microsecond. The probability of each symbol being transmitted is given above each symbol. What are the bit rate, the symbol rate, and the information rate?



Exercise 11: Another system, as shown above, encodes each bit using two pulses of opposite polarity (H-L for 0 and L-H for 1). A second system encodes bits using one pulse per bit (H for 0 and L for 1). A third system encodes two bits per pulse by using four different pulse levels (-3V for 00, -1V for 01, +1V for 10 and +3V for 11). How many different symbols are used by each system? What are the symbol rates? Assuming each system transmits at 1000 bits per second, what are the symbol rates in each case? Assuming each symbol is equally likely, what are the information rates?

[2] *The Authoritative Dictionary of IEEE Standards Terms*, Seventh Edition, IEEE, 2000.

Error Rates

The bit error rate (BER, P_e) is the average fraction of bits that are received incorrectly.

When these bits are grouped into "frames" we are often interested in the average fraction of the frames that contain one or more errors. This is known as the FER (Frame Error Rate). Sometimes frames include additional bits that allow us to detect most, but not all, errors. We usually want the UEP (Undetected Error Probability) to be very small (e.g. one undetected error per many years).

Exercise 12: You receive 1 million frames, each of which contains 100 bits. By comparing the received frames to the transmitted ones you find that 56 frames had errors. Of these, 40 frames had one bit in error, 15 had two bit errors and one had three errors. What was the FER? The BER?

Throughput

The average rate at which data arrives at the destination can be different than the instantaneous rate at which the transmitter sends data to the channel because:

- the channel may have to be shared between different users
- the transmitter may add (and the receiver remove) "overhead" bits for addressing, error detection, etc
- incorrectly received data may have to be retransmitted
- there may be gaps between frames

The long-term average data rate arriving at the destination is called the "throughput". It can be computed by dividing the total amount of data delivered to the sink by the elapsed time. Each type of communication system involves different overheads and thus requires a different throughput computation.

Exercise 13: A system transmits data at an (instantaneous) rate of 1 Mb/s in frames of 256 bytes. 200 of these bytes are data and the rest are overhead. The time available for transmission over the channel is shared equally between four users. A 200 μs gap must be left between each packet. What throughput does each user see? Now assume 10% of the frames are lost due to errors. What is the new throughput per user?

Compression

When data is not random and we can make use of the redundancy to reduce the amount of data that needs to be transmitted. Both lossless and lossy compression are examples of "source coding."

Lossless. Some types of data contains redundancy such as sequences of bits or bytes that occur more often than others. This type of data can be compressed before transmission and then decompressed at the receiver without loss of information. An example of this "lossless" compression is the 'zip' compression used for computer files.

Another definition of information rate is "the minimum data rate, assuming the best possible lossless compression". Lossless compression does not reduce the information rate but it may reduce the bit rate.

Lossy. Data representing speech and video can often be compressed with little degradation because humans cannot perceive certain details of sounds and images. These details can be removed resulting in lower data rates. Examples of these "lossy" compression techniques include "MP3" for compressing audio and MPEG-4 for video.

Data Rate Variability

The data rate of a source can be:

- constant: "isochronous" sources generate data at a constant bit rate (CBR) and are typical of regularly sampled waveforms such as (uncompressed) audio or video sources.

- variable: variable bit rate (VBR) sources are typical of compressed speech and video because different parts of the speech or video signal have different amounts of redundancy and can be compressed to different bit rates.

- bursty: bursty data sources generate large amounts of information at instants of time in-between pauses where no information is generated. This is typical of systems involving human-machine interaction such as web surfing.

Exercise 14: Plot some sample data rate versus time curves for these three types of sources. Can you think of some characteristics of a video source that might result in a variable bit rate when it is compressed? (*Hint: what types of redundancy are there in video?*)
.

Tolerance To Impairments

Data sinks vary in their tolerance to channel impairments such as errors, delay and variability of delay (delay "jitter").

For example, computer data transmission systems usually require very low undetected error rates (e.g. one undetected error in decades) but can often tolerate high delay and delay variability (seconds). On the other hand telephone systems can tolerate loss of a small percentage of the speech waveform but become difficult to use if delays exceed a significant fraction of a second.

Exercise 15: For each of the following communication systems identify the tolerance it is likely to have to errors and delay: a phone call between two people, "texting", downloading a computer program, streaming a video over a computer network. What do you think might be the maximum tolerable delay for each?

Exercise 16: Highlight or underline each term where it is defined in these lecture notes.