TCOM 370 NOTES 99-4

BANDWIDTH, FREQUENCY RESPONSE, AND CAPACITY OF COMMUNICATION LINKS

1. Bandwidth:

The bandwidth of a communication link, or in general any system, was loosely defined as the width of the frequency interval such that input sinusoidal frequencies within this interval will appear at the output without significant amplitude or phase change.

Bandwidth is related to maximum pulse transmission rate and hence data transmission rate for the link, and is clearly an important characteristic.

- A more precisely defined characteristic incorporating bandwidth information is the **frequency response** of the link or system.

In defining the frequency response, we will first take a more careful look at sinusoidal signal transmission.

Basic Questions:

To begin with, how do we know that a single sinusoid $\cos(2\pi f_0 t)$ at the input to a communication channel will produce a *sinusoid,* and not some other output, and moreover a sinusoid at the *same frequency* as the input frequency?

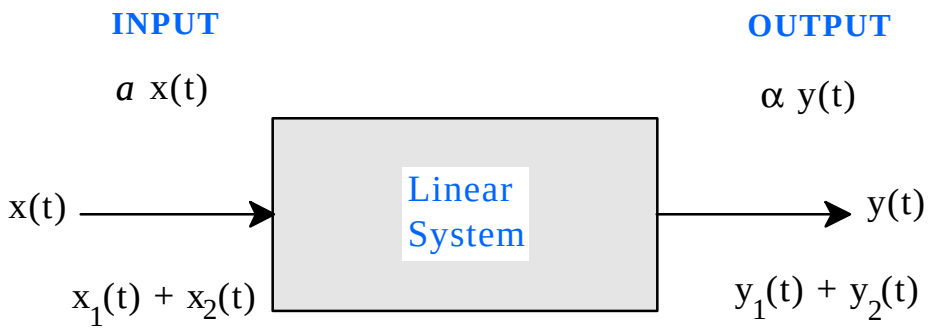
Even if this is so, how do we know that a *sum of* individual *sinusoids* at the input will appear at the output as a *sum of sinusoids,* each output sinusoid depending only on the corresponding input sinusoid?

Sinusoids through Linear Systems:

When signals travel through fixed (time-invariant) electrical systems (transmission media, amplifiers, filters, etc.) they can get changed by the action of the system on the signal.

Often a good mathematical model for this type of system is the time-invariant ***linear system*** model. Linearity means that if an input signal x(t) produces output y(t), and $x_1(t)$ results in output $y_1(t)$ and another input $x_2(t)$ yields output $y_2(t)$, then always

- input *a* x(t) produces output *a* y(t), (*a* is any amplitude scale factor)
- input $x_1(t) + x_2(t)$ yields output $y_1(t) + y_2(t)$ *(superposition)*



Further, ***time-invariance*** means that

- the behavior of the system does not change with time, it is fixed.

These are reasonable behaviors, obtained by design or found to hold for *communication channels under normal conditions.* Furthermore, electronic systems that are used to process input and output signals in communications are usually designed to operate as fixed linear systems.

- Now it can be ***proved*** that *a sinusoidal input to a fixed real linear system is always reproduced at the output as a sinusoid at the same frequency (with amplitude and phase possibly altered).*

A proof is given below for those interested. *It may be skipped.*

Proof:

Consider the complex exponential time-function $e^{j2\pi f t}$ [a component of $\cos(2\pi f t)$] at the input of a time-invariant (fixed) linear system, and suppose the output is $y_f(t)$. The sub f indicates that the output time function depends on f.

Now consider the input $e^{j2\pi f (t-d)}$, a version of the original input delayed by d. Because of time-invariance, the output is now the delayed original output, $y_f(t-d)$. But $e^{j2\pi f (t-d)}$ is also $e^{-j2\pi f d} e^{j2\pi f t}$, an amplitude scaled version of the original input. Thus the output is also $e^{-j2\pi f d}\, y_f(t) = y_f(t-d)$.

In the last equation, set d=t and take $e^{-j2\pi f t}$ to the other side, yielding

$y_f(t) = y_f(0)\, e^{j2\pi f t}$.

Thus we see that the output due to $e^{j2\pi f t}$ at the input produces an amplitude scaled version of $e^{j2\pi f t}$, where the amplitude scaling is some constant $[y_f(0)]$ which is generally complex valued $[Ae^{j\phi}]$, depends on f, and effects an amplitude and phase change.

The cosine function $\cos(2\pi f t)$ is itself a combination of two complex exponentials with oppositely signed frequencies. The component $e^{-j2\pi f t}$ of the cosine will also similarly produce output $y_{-f}(t) = y_{-f}(0)\, e^{-j2\pi f t}$, and because the system is real the output this time is a complex conjugate of the original output since the input is a complex conjugate of the original input. Thus $y_{-f}(0) = Ae^{-j\phi}$ and it follows that $y_f(t)$ is now $A\cos(2\pi f t + \phi)$]

The result that was stated (and proved) above for a fixed linear system, together with the second condition defining linearity, addresses the basic questions asked on page 1.

We now know how to analyze for the effect of any linear communication medium (e.g. twisted-wire pair) or system on a signal ; decompose the input into its different frequencies using Fourier series or its limiting form, then each frequency is reproduced at the output either essentially unaltered (within the transmission bandwidth) or attenuated (outside the transmission bandwidth) or somewhat altered (intermediate situation). Larger system bandwidths will allow narrower pulses to be preserved at the output. If we use narrow pulses at the input of a low-bandwidth system, only some of the frequencies will pass through and the output pulse will get wider. This ***distortion*** of the pulse shape is caused by the limited bandwidth.

Bandwidth depends only on the width of the supported frequency band, not on its absolute location. A medium allowing frequencies between 1 and 2 GHz to be transmitted has the same bandwidth as one allowing frequencies between 20 and 21 GHz. (We will see later how the use of *modulation* allows any frequency band to be used for pulse sequence transmission.) In practice, high bandwidth is obtained by operating at high frequencies. The ratio of actual max. to min. operating frequencies for a fixed bandwidth at a low center frequency is much larger than at higher operating frequencies (2-to-1 vs. 1.05-to-1 in the examples above). The smaller range (max./min. ratio) of frequencies at which equipment has to operate makes larger bandwidths more feasible at higher frequencies.

2. Frequency Response of a Linear Transmission Channel:

The *frequency response* characteristic of a transmission medium or any other system modeled as a linear system gives a more detailed picture of the action of the channel on input sinusoids.

- It is a *ratio of output to input as a function of frequency* for sinusoids.

Since for sinusoids the output frequency is the same as the input frequency, the quantities that appear in the ratio are the amplitudes and the phases.

When $x(t)=\cos(2\pi f t)$ (at frequency f) is applied at the input, let the output be $A\cos(2\pi f t + \phi)$. Note that A and ϕ depend on the frequency f, and are more explicitly written as functions A(f) and $\phi(f)$.

- *The two functions* A(f) *and* $\phi(f)$ *constitute the frequency response of the linear system.* (They are called the amplitude response and phase response, respectively).

- More generally *the frequency response is written as a complex-valued function of f, denoted by* H(f). Its amplitude part is A(f), the output-to-input amplitude ratio as a function of f, and the phase or angle part $\phi(f)$ is the phase of the output relative to the phase of the input as a function of frequency (output phase minus input phase). Thus $H(f)=A(f)e^{j\phi(f)}$

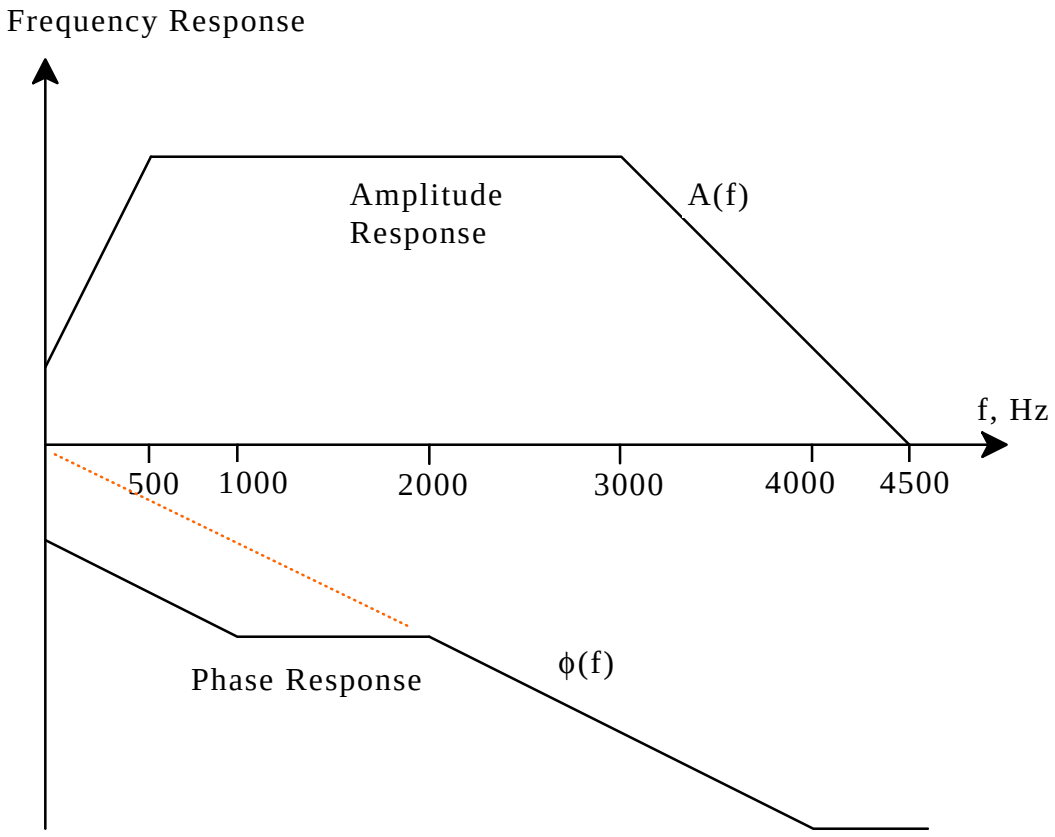
  (This compact form of the frequency response is seen to be the ratio of output $A(f)\, e^{j[2\pi f t + \phi(f)]}$ to the input complex frequency $e^{j2\pi f t}$).

- An ***amplitude response*** A(f)=|H(f)| that is ***flat*** over a band of frequencies gives rise to ***no amplitude distortion*** from selective attenuation over this band of frequencies. All input frequency components in this band are *equally affected in amplitude.*

- A ***phase function*** $\phi(f)=\arg\{H(f)\}$ that is a ***linear function*** of f [say $\phi(f)= -\beta f$] results in ***no delay distortion*** through different relative phase shifts imparted to different frequency components.

  This is because higher frequencies are shifted by larger phases proportionally, so that all frequencies are delayed by the *same time delays.* Not having delay distortion means that any delay that exists is the same for all frequencies.

Note very carefully that for no delay distortion of a signal due to different phase shifts imparted to its different frequency components, the phase should be a linear function of frequency of the form $\phi(f)=$ (constant) x f



Example Frequency Response

On the previous page an example is shown of an amplitude and phase response of a low-frequency channel (the amplitude response is approximately that of a single telephone channel). For this channel, we find that there is no amplitude distortion over the band [500, 3000] Hz, and there is no delay distortion over the band [2000, 4000] Hz. Note that in the other portions of the phase response, even though the phase function has straight line segments, none of the lines pass through the origin. They are not "linear phase".)

For this example the only range of frequencies over which there is no amplitude or delay distortion is the band from 2000 to 3000 Hz.

(It is possible to have frequency responses with multiple separate bands over which distortionless transmission is possible.)

- A channel that is ***non-distorting*** over a certain band of frequencies has a **flat amplitude response and a linear phase response** over the band.

Linear Phase and Constant Delay

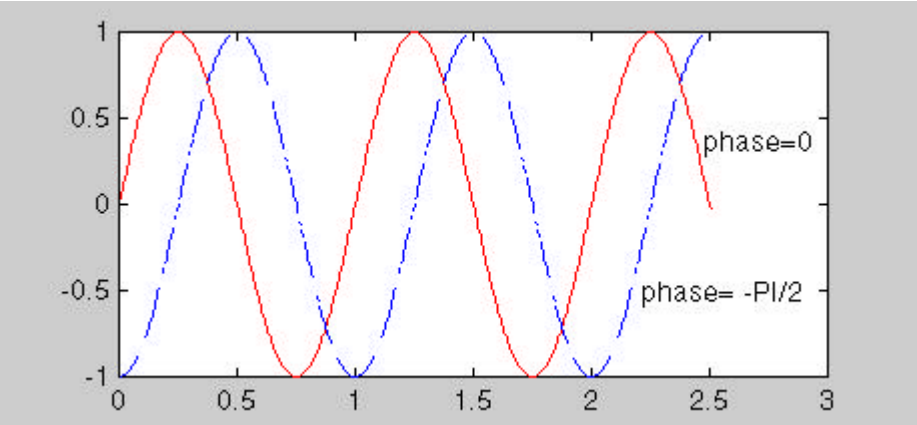


Figure 1

Frequency =1 Hz

Showing time delay caused by -π/2 phase shift

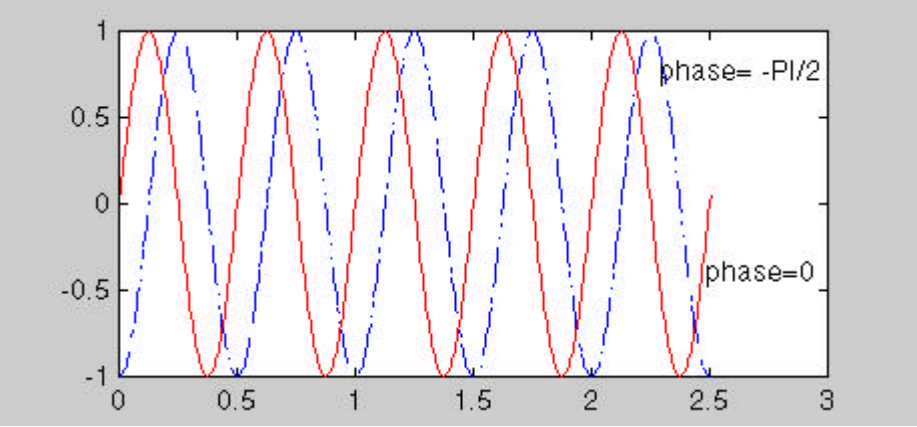


Figure 2.
Frequency =2 Hz

Showing time delay caused by -π/2 phase shift (delay is half that in Figure 1)

Consider a sinusoid frequency $x(t)=\cos(2\pi f\, t)$ and its phase shifted version $y(t)=\cos(2\pi f\, t-\phi)$. We can write $y(t)$ as

$$y(t)= \cos\left(2\pi f \left[t - \frac{\phi}{2\pi f}\right]\right)$$

and we see that it is a delayed version $x\left(t-\frac{\phi}{2\pi f}\right)$ of $x(t)$. The amount of delay is $\frac{\phi}{2\pi f}$. Note that this delay is inversely proportional to frequency for fixed phase shift $-\phi$. We find that to get a constant delay for all frequencies, we must have phase shifts that increase linearly with f.

3. Attenuation

This refers to the ***overall*** decrease in signal amplitude as it passes through a transmission medium or system. Attenuation *by itself* only results in lower signal amplitudes; these can be brought back up by ordinary amplifiers. Any overall attenuation factor in a transmission medium does not distort the signal; all frequencies are affected in the same way, and *signal shapes are preserved.* However from a practical point of view, since input power is constrained, we cannot tolerate too much attenuation because once the signal gets weaker than the random noise and interference that are always adding to it, amplification will not work (noise is also amplified). Thus if the medium has large ***overall attenuation per unit of transmission distance,*** it cannot be used for physically long links without the use of ***line repeaters.*** These are inserted at intervals along a long link to bring back signal amplitudes to reasonable levels and regenerate the data before further transmission.

- Signal attenuation is measured in units called *"decibels".* If over a transmission link the ratio of output power to input power is P_o/P_i , the attenuation is said to be $-10 \log_{10}(P_o/P_I) = 10 \log_{10}(P_i/P_0)$dB. In cascaded links, the attenuation in dB is simply a sum of the individual attenuations in dB. Attenuation per unit length is obtain by dividing overall dB attenuation by the length of the link.

Note that ***selective*** attenuation of some frequencies relative to others results in signal ***distortion,*** as we have discussed. Sometimes we can use ***equalizers*** to correct such signal distortion across a band of frequencies in a transmission band. In addition to amplitude distortion due to selective attenuation, different frequencies may be reproduced at the output with different ***delays*** because of non-ideal phase response. An equalizer is a device that compensates for the amplitude and delay distortions in a band of frequencies. It is usually inserted at the receiving end. Equalizer are ubiquitous in modems allowing data transmission over non-ideal telephone channels. They can be designed to *automatically compensate* for non-ideal channel characteristics.
(Note that the band of frequencies over which equalizers have a chance of working effectively is that band over which the amplitude attenuation is not very severe, otherwise we face the problem of low signal-to-noise power ratio.)
Repeaters may also include some form of equalization.

4. Noise and Interference

Noise is always present as an impediment to achieving reliable (i.e. error free) communication. Noise is present in the form of random motion of electrons in conductors, devices and electronic systems (due to thermal energy), and can also be picked up from external sources (atmospheric disturbances, ignition noise, etc.). Interference generally refers to the unwanted, stray signals picked up by a communication link due to other transmissions taking place in adjacent frequency bands or in physically adjacent transmission lines.

Noise is often modeled as a random process having a ***fixed power in each Hz*** of the transmission band. (This model is known as ***white noise***). We use the notation N_0 to denote the amount of noise in each Hz of the transmission band; the units of N_0 are watts/Hz. In practice, N_0 values of 10^{-7} to 10^{-21} watts/Hz may be encountered.

Noise is the other main limiting factor (in addition to bandwidth) in obtaining high performance from a communication system. It is actually the ratio of signal power to noise power (S/N ratio or SNR) that we normally use in addition to bandwidth as the two main determinants of performance.

Signal-to-noise power ratio is often expressed in decibels (dB's):

$$\text{SNR} = 10 \log_{10}\left(\frac{S}{N}\right) \text{ dB}$$

where $\frac{S}{N}$ is the actual signal power to noise power ratio.

Summary:

The key considerations for data transmission media are:

- **Bandwidth and Frequency Response**
- **Attenuation and Distance Limit**
- **Interference and Noise**

(these lead to data-rate/capacity limitations, and minimum signal power requirement)

5. Signaling Rate and Bit Rate

In signaling with pulses to represent data, the rate at which pulses are transmitted is known as the ***signaling rate*** or the ***baud rate.*** The transmitted ***bit rate*** may be different if each pulse can carry more than one bit of data (or sometimes less than one bit of data). For example, if two-level (bipolar) pulses are used, each pulse carries one bit of information. However, we may use one of four different amplitudes for each pulse, in which case 2 bits of information are carried on each pulse and the bit rate is twice the signaling rate.

6. Some Fundamental Limits

Nyquist Signaling Criterion

For an ideal (flat amplitude response, no delay distortion) transmission channel of ***bandwidth W,*** with no noise, a ***maximum signaling rate of 2W pulses per second*** is possible (with the right type of pulses) if individual pulse amplitudes are to be recovered perfectly by sampling at the pulse centers at the channel output.

The bandwidth determines the highest pulse rate at which pulse overlapping at pulse centers can be avoided and individual pulse amplitudes can be recovered from the output.

Note that if each pulse can carry one of M different amplitude levels, then each pulse conveys $\log_2 M$ bits of information.

- Thus ***maximum bit rate*** in this case is **$2W \log_2 M$** bps.

As we have stated earlier, a more **practical rate** of pulsing over a bandwidth-W channel is ***W pulses per second*** and therefore the practical data rate is **$W \log_2 M$** bps.

- The ***bandwidth efficiency*** B achieved by a transmission scheme is the ratio of bit rate R it provides to the bandwidth W used. Its maximum value is $2\log_2 M$ if the scheme uses M-level pulses.

Channel Capacity

Note that according to the above, in theory we can get ***arbitrarily high bit rates*** from a ***finite bandwidth*** channel by allowing lots of different amplitude levels for each pulse (large M)! However, if noise is present, then the levels cannot be too close together otherwise it will not be possible to distinguish between adjacent levels with high accuracy.

Thus to be able to use large M, we need to keep the amplitude level spacing above a minimum value and this implies use of large maximum pulse amplitudes. But this requires large signal power! If the signal power is constrained (as is the case in practice), then in the presence of noise, arbitrarily high data rates are not possible over a finite bandwidth channel.

A very famous result due to *Shannon* (1948) states that the *maximum possible transmission rate* or ***capacity*** of a link of bandwidth W Hz and received signal-to-noise ratio $\frac{S}{N}$, is (under certain conditions on the nature of the noise, which often hold)

$$\mathbf{C = W \log_2(1 + \frac{S}{N})\ \ bps}$$

This is a theoretical limit and is not easy to come close to in practice.

Example:

The V.34 telephone line modem transmits on a telephone line bandwidth of approximately 3400 Hz. The signal-to-noise (power) ratio for received signals with such modems is of the order of 35 dB under good conditions.

Thus the actual S/N ratio is, from $10 \log_{10}(S/N) = 35$,

$$S/N = 10^{3.5} = 3162$$

According to Shannon's formula, the capacity should be $3400 \log_2(3163)$ which is approximately 3400x11.7=40,000 bps. The modem achieves a maximum bit rate of 33,400 bps.

7. Performance Measure for Data Communication

In actual systems the communication quality for data transmission is stated as a **probability of bit error** or **bit error rate**. Theoretically, we can get up to C bps with zero error rate. In practice, the actual bit error rate is always non-zero; the typical range is 10^{-3}-10^{-9}. For any particular type of transmission technique over a bandwidth W, the bit error rate P_e is a function of the SNR.

[We can obtain a quantity related to the SNR in the following way:

Let the bit rate achieved be R bps over a bandwidth W Hz in the presence of white noise with power density N_0 watts/Hz, and let the received signal power be S watts . Then each bit takes 1/R secs. to transmit, so the energy expended per bit is E_b=S/R watt-secs or joules (note that power is the rate of expenditure of energy). The total noise power in the communication band is N_0W watts. Thus the signal-to-noise power ratio $\frac{S}{N} = \frac{E_bR}{N_0W}$.

Thus $\frac{S}{N} = \frac{E_b}{N_0}$ x [bandwidth efficiency]. For a given scheme with a given bandwidth efficiency, the signal-to-noise ratio is proportional to $\frac{E_b}{N_0}$ and hence the error probability P_e depends on $\frac{E_b}{N_0}$.]

(You are not required to memorize this result! This is for your overall appreciation of the importance of signal to noise ratio in determining performance.)