Lecture 3 - Digital Communication and Storage I

In this lecture I'm going to look at how signals from a range of inputs, be it voice, video, audio etc. can be converted so that they may be transmitted digitally along an optical fibre network. Optoelectronic communication is reliable and economical; communication technology is playing an increasingly important role in our lives. For example, important discussions now mostly communicated face to face in meetings or conferences, often requiring travel, are increasingly using "teleconferencing". Similarly teleshopping and telebanking will provide services by electronic communication, and the Internet relies completely on fibre technology for its long distance links.

Figure 3.1 shows three examples of communication systems. A typical communication system can be modelled as shown in figure 3.2. The components of a communication system are as follows:

The **source** originates a message. If the data is non-electrical (human voice etc.), it must be converted by an **input transducer** into an electrical waveform referred to as the **baseband signal** or **message signal**.
The **transmitter** modifies the baseband signal for efficient transmission.
The **channel** is a medium—such as a wire, coaxial cable, waveguide or optical fibre—through which the transmitter output is sent.
The **receiver** reprocesses the signal received from the channel. The receiver output is fed to the **output transducer**, which converts the electrical signal back to its original form—the message.
The **destination** is the unit to which the message is communicated.

The channel can filter, attenuate and distort the signal. The signal attenuation increases with the length of the channel. The waveform is distorted because of different amounts of attenuation and phase shift suffered by different frequency components of the signal. This might be the rounding of a square pulse, for example. This type of distortion is known as **linear distortion** and can be partly corrected at the receiver. The channel may also cause **non-linear distortion** through attenuation that varies with the signal amplitude. Such distortion may also be partly corrected by the receiver. The signal is also contaminated by undesirable signals lumped under the broad term **noise**, which are random and unpredictable signals from causes both external and internal. External noise might include interference from nearby channels, faulty equipment, radiation, storms etc. Internal noise results from thermal motion of electrons in conductors, random emission

Figure 3.1 Some examples of communication systems

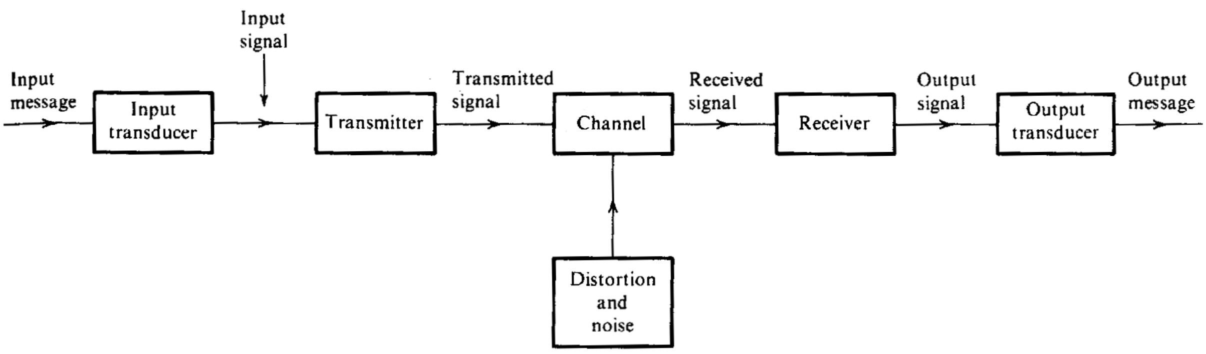

Figure 3.2 A typical communication system

and diffusion or recombination of charged carriers in electronic devices. Noise can be reduced but not eliminated. It is one of the basic factors that sets limits on the rate of communication.

The **signal to noise ratio (SNR)** is defined as the ratio of signal power to noise power. The channel distorts the signal, and noise accumulates along the path. Worse yet, the signal strength decreases while the noise level increases with distance from the transmitter. Thus the SNR is continuously decreasing along the length of the channel. Amplifiers will increase both the signal and the noise, and may indeed introduce more noise of their own.

1 Analogue and Digital Messages

Messages are either digital or analogue. A digital message may be constructed using a number of discrete symbols (e.g. 26 letter of the alphabet). The most important number used is 2, as in binary. Analogue messages, on the other hand, are characterized by data whose values vary over a continuous range. A speech waveform has amplitudes that vary over such a range. Over a given time interval, an infinite number of possible different speech waveforms exist, in contrast to only a finite number of possible digital messages.

1.1 Noise Immunity of Digital Signals

Digital messages are transmitted by using a finite set of electrical or optical waveforms. The task of a receiver is to extract a message from a distorted and noisy signal at the channel output. Message extraction is often easier from digital signals than from analogue signals. Consider the binary case of two signals encoded as rectangular pulses of amplitudes $A/2$ and $-A/2$. The receiver only has to decide between two possible pulses received, not on the details of the pulse shape. This decision can be made with reasonable certainty even if the pulses are distorted and noisy (see figure 3.3). The digital message in figure 3.3(a) is is distorted (figure 3.3(b)) and noise is added(figure 3.3(c)). The data can be recovered as long as the distortion and noise are within limits, because we only need to make a simple binary decision as to whether the pulse is positive or negative. Analogue systems are sensitive to even small amounts of distortion or noise, and hence a digital communication system is more rugged and an analogue one in the sense that it can better withstand noise and distortion.

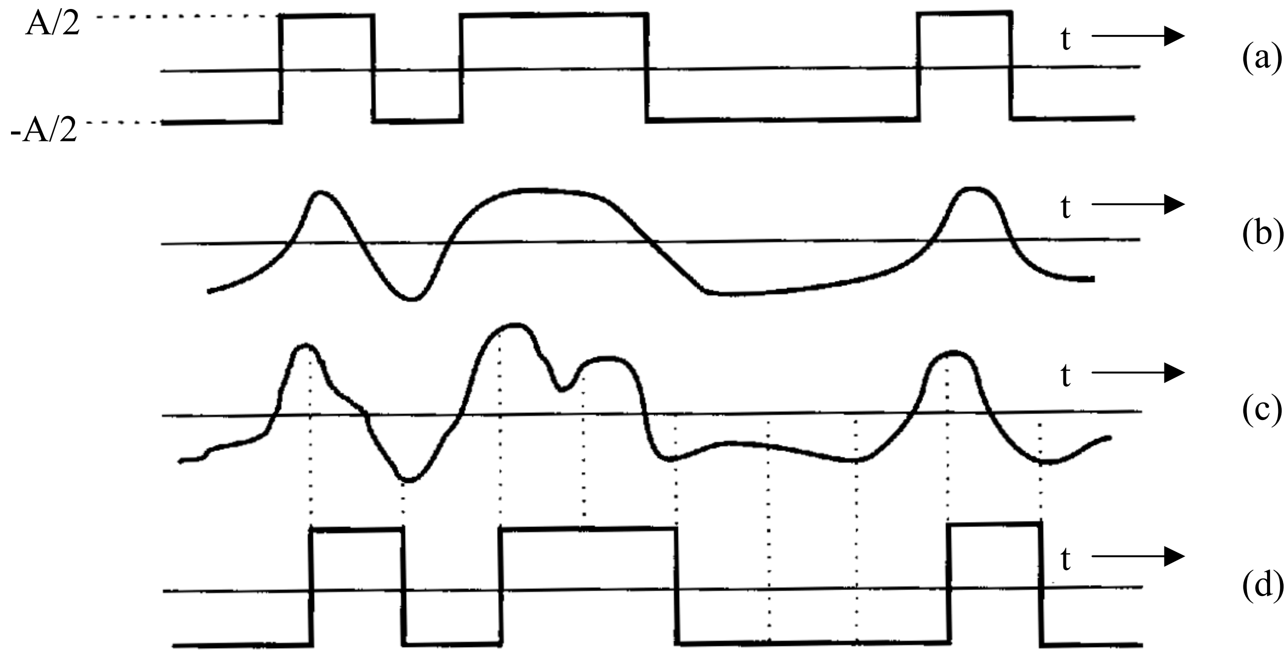

Figure 3.3 (a) Transmitted signal. (b) Received distorted signal (without noise). (c) Received distorted signal (with noise). (d) Regenerated signal (delayed).

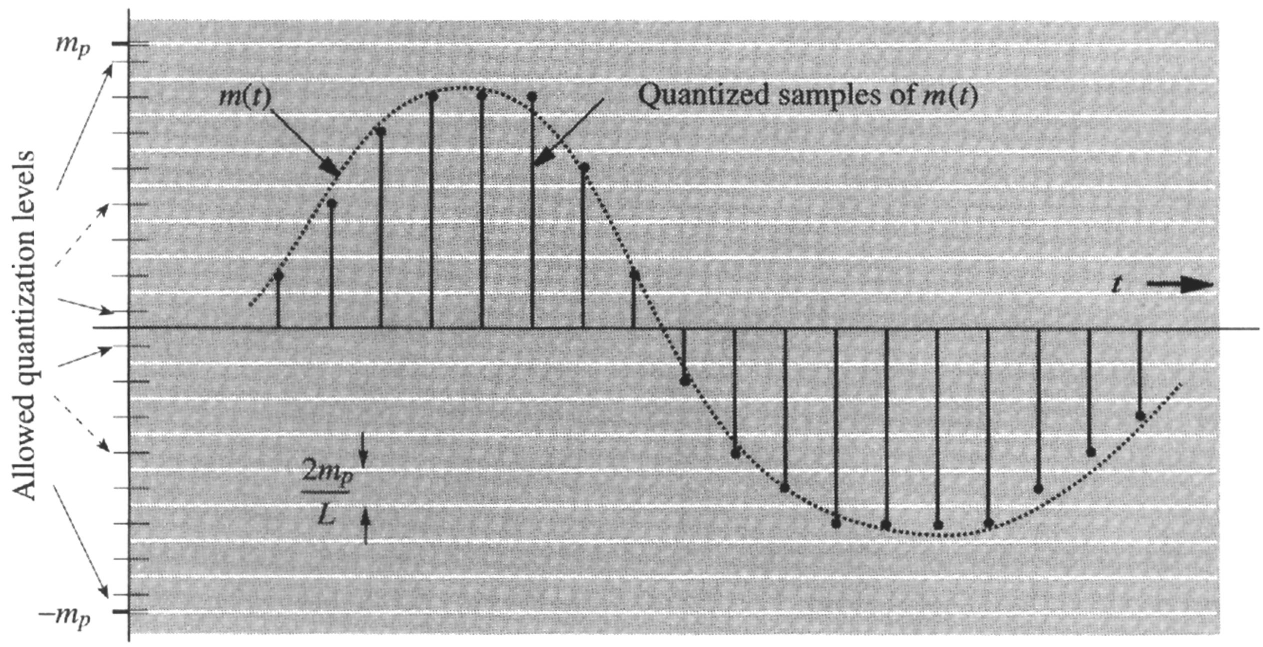

Figure 3.4 Analogue-to-digital conversion of a signal.

1.2 Regenerative Repeaters in Digital Communication

The main reason for the superiority of digital systems over analogue ones is the viability of **regenerative repeaters** in the former. Repeater stations are placed close enough together in a channel to ensure that noise and distortion remain within set limits. At each station the incoming pulses are detected and new clean pulses are transmitted to the next repeater station, thus preventing the accumulation of noise and distortion along the channel. Messages can thus be transmitted over long distances with great accuracy. As mentioned earlier, this is impossible for analogue signals, where the signal degrades and the noise increases with each amplification stage. The advent of optical fibre technology, coupled with cheap digital circuitry, has lead to almost all new communication systems being digital installations.

1.3 Analogue to Digital (A/D) Conversion

In order to transmit analogue signals digitally they need to be converted. The frequency spectrum of a signal indicates the relative magnitudes of various frequency components. The **Nyquist or sampling theorem** (proved later) states that if the highest frequency in the signal spectrum is B (in hertz), the signal can be reconstructed from its samples, taken at a rate not less than $2B$ samples per second. This means that in order to transmit the information in a continuous-time signal, we need only transmit its samples (figure 3.4). The sample values then need to be digitised or quantised where each sample is approximated or rounded off to the nearest quantised level. Amplitudes of the signal $m(t)$ lie in the range $(-m_p, m_p)$, which is partitioned into L intervals, each of magnitude $\Delta v = 2m_p/L$. Each sample amplitude is approximated to the midpoint of the interval in which the sample value falls. The information is thus digitised to one of L levels. The accuracy of the quantised signal can be improved by increasing the number of levels L. For voice only, $L = 16$ is sufficient. For commercial use, $L = 32$ is a minimum, and for telecommunication, $L = 128$ or 256 is commonly used. One way of transmitting the digitised signals would be to transmit each of the L values as a discrete voltage. The second, and preferred alternative is to use binary transmission, where the number of bits is chosen to at least match the quantisation levels (more may be used if error checking is included). The binary case is so important because of its simplicity and ease of detection. Virtually all digital communication today is binary. This scheme of transmitting data by digitising and then using pulse codes to transmit the digitised data is known as **pulse**

code modulation (PCM).

When considering a distorted binary signal, such as that shown in figure 3.3, then if A is sufficiently large compared to typical noise amplitudes, the receiver can still distinguish correctly between the two pulses. If the pulse amplitude is $\sim 5-10$ times the noise amplitude, the probability of error at the receiver is less than 10^{-6}. The effect of random channel noise and distortion is thus practically eliminated. One error or uncertainty in the signal still remains however, quantisation noise. This can be reduced by increasing L, the penalty for this is paid in terms of increased bandwidth of transmission.

Although PCM was invented by P.M. Rainey in 1926 and rediscovered by A.H. Reeves in 1939, it was not until the early sixties that Bell Laboratories installed the first communication link using PCM. It was the transistor that made PCM practicable.

1.4 Signal-to-Noise Ratio and Channel Bandwidth

The **bandwidth** B of a channel is the range of frequencies that it can transmit with reasonable fidelity. The number of pulses per second that can be transmitted over a channel is directly proportional to its bandwidth.

The **signal power** S plays a dual role in information transmission. First, S is related to the quality of transmission. Increasing S reduces the effect of channel noise. A larger signal-to-noise ratio (SNR) also allows transmission over a longer distance. Thus a certain minimum SNR is necessary for communication. The second role of the signal power relies on the fact that the channel bandwidth B and the signal power S are interrelated. Thus one may reduce the requirement for S by increasing B, and vice-versa. The rigorous proof of this statement is beyond the scope of this lecture course, but may be found in chapter 15 of Lathi. Since SNR is proportional to S, we can also say that SNR and bandwidth are exchangeable. The exact relation is that, to transmit information at a given channel bandwidth B_1 and signal-to-noise ratio SNR_1, then the same information can be transmitted over a channel bandwidth B_2 with signal-to-noise ratio SNR_2, when

$$\mathrm{SNR}_2 \simeq \mathrm{SNR}_1^{B_1/B_2}$$

Therefore, a relatively small increase in channel bandwidth buys a large advantage in terms of reduced transmission power. This is an upper bound, and in real systems the performance is usually much worse. PCM, however, comes close (within 10dB) to realising this performance.

The limitation imposed on communication by the channel bandwidth and the SNR is highlighted by an equation derived by Shannon relating the rate of error-free information transmission per second C to the bandwidth B and the SNR. The equation is

$$C = B \log_2(1 + \mathrm{SNR}) \qquad \text{bit/s}$$

If there were no noise on the channel ($\mathrm{SNR} = \infty$) then $C = \infty$, and communication would cease to be a problem. Of course a new limit on C would then appear, related to the accuracy with which signal levels could be determined.

1.5 Simultaneous Transmission of Several Signals

If several audio signals are to be transmitted simultaneously then the sensible (and practical) solution is to add these signals as sidebands on multiple carrier waves (either AM or FM). If the various carriers are chosen sufficiently far apart in frequency, the spectra of the modulated signals will not overlap and thus will not interfere with each other. At the receiver one can use a tunable bandpass filter to select the desired station or signal. This method of transmitting several signals simultaneously is known as **frequency-division multiplexing (FDM)**. here the bandwidth of the channel is shared by various signals without any overlapping.

Another method of multiplexing several signals is known as **time-division multiplexing (TDM)**. This method is suitable when a signal is in the form of a pulse train (as in PCM). The pulses are made narrower, and the spaces that are left between them are used for pulses from other signals. Thus, in effect, the transmission time is shared by a number of signals by interleaving the pulse trains of various signals in a specified order. At the receiver, the pulse trains corresponding to various signals are separated. This is illustrated in figure 3.5.

2 Sampling

2.1 Proof of Sampling Theorem

Let $g(t)$ be a signal function which is periodic with period T_0 seconds. Then $g(t)$ may be expressed as a Fourier series over any interval of duration T_0 seconds as

$$g(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega_0 t + b_n \sin n\omega_0 t \qquad t_1 \leq t \leq t_1 + T_0$$

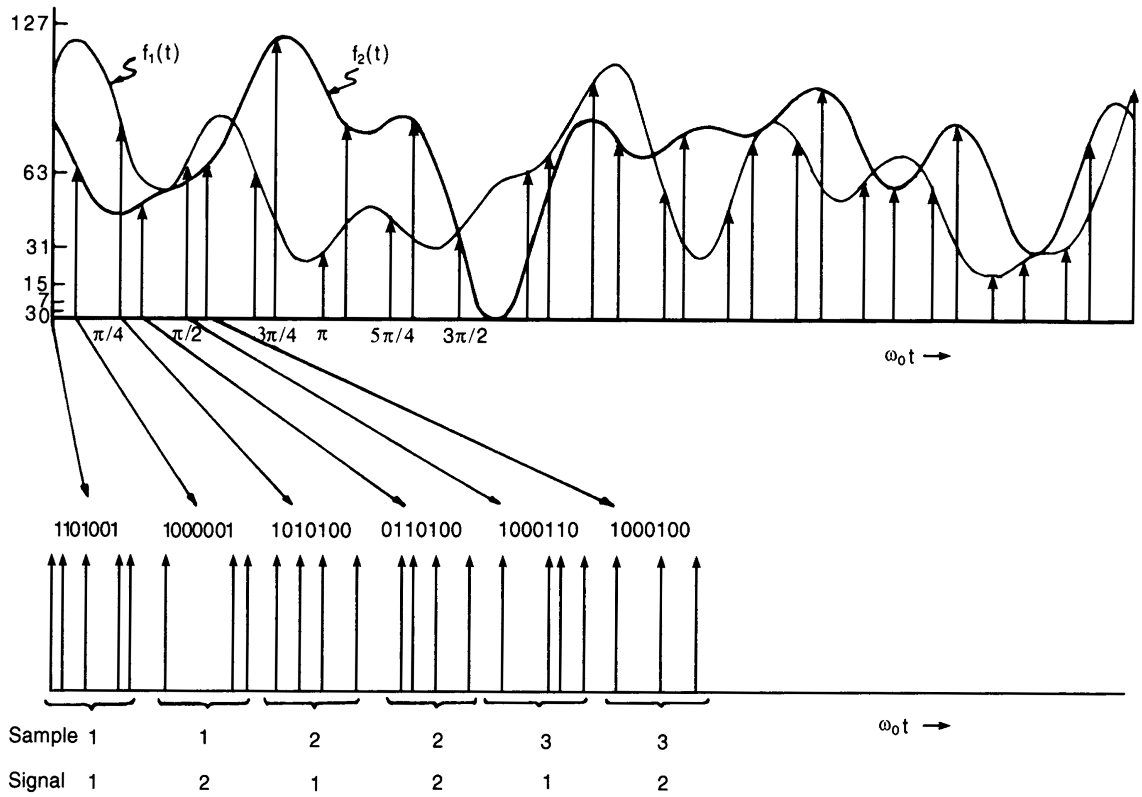

Figure 3.5 Curves 1 and 2 correspond to two different signals to be transmitted along the same channel simultaneously. Each signal is first sampled, then coded into a binary code, and finally intermingled so that the two signals appear at different times. The interspersed series of pulses is then sent along the channel. This is known as time division multiplexing (TDM).

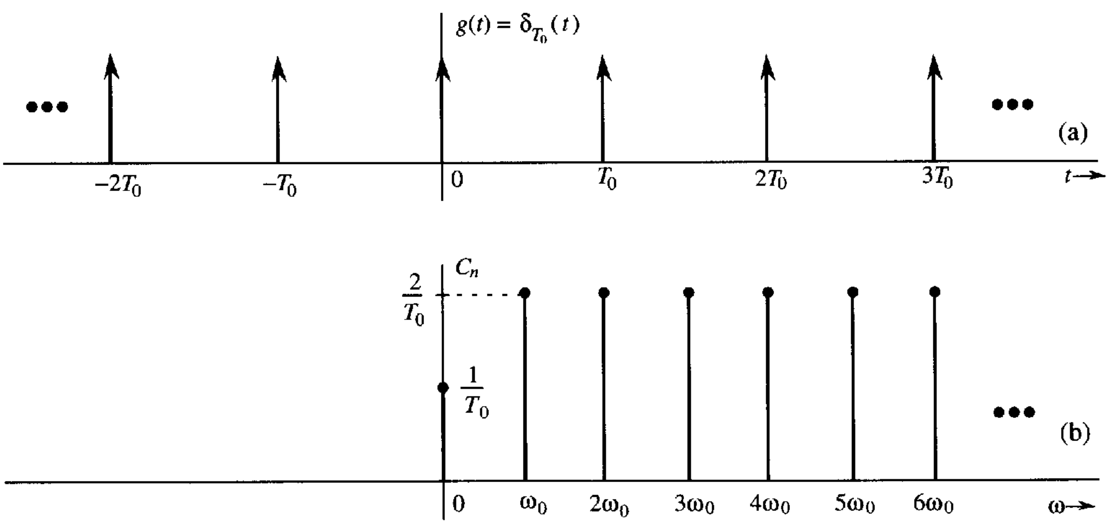

Figure 3.6 Impulse train and its Fourier spectrum.

where

$$\omega_0 = \frac{2\pi}{T_0}$$

and

$$a_n = \frac{2}{T_0}\int_{t_1}^{t_1+T_0} g(t)\cos n\omega_0 t \,\mathrm{d}t \qquad n = 1, 2, 3, ...$$

and similarly for b_n with sine functions. The Fourier series contains sine and cosine terms of the same frequency. These can be combined into a single term of the same frequency using the trigonometric identity

$$a_n \cos n\omega_0 t + b_n \sin n\omega_0 t = C_n \cos(n\omega_0 t + \theta_n)$$

where

$$\begin{aligned} C_n &= \sqrt{a_n^2 + b_n^2} \\ \theta_n &= \tan^{-1}\left(\frac{-b_n}{a_n}\right) \\ C_0 &= a_0 \end{aligned}$$

Thus, $g(t)$ may be expressed in the **compact form** of the trigonometric Fourier series as

$$g(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \theta_n) \qquad t_1 \le t \le t_1 + T_0$$

Now consider the Fourier series for a periodic train of delta functions $\delta_{T_0}(t)$ as shown in figure 3.6 (a). We have

$$\delta_{T_0}(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \theta_n) \qquad \omega_0 = \frac{2\pi}{T_0}.$$

We first compute a_0, a_n, and b_n:

$$\begin{aligned} a_0 &= \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} \delta(t)\,\mathrm{d}t = \frac{1}{T_0} \\ a_n &= \frac{2}{T_0}\int_{-T_0/2}^{T_0/2} \delta(t)\cos n\omega_0 t\,\mathrm{d}t = \frac{2}{T_0} \end{aligned}$$

Similarly, using the sampling property of the delta function, we obtain

$$b_n = \frac{2}{T_0}\int_{-T_0/2}^{T_0/2} \delta(t)\sin n\omega_0 t\,\mathrm{d}t = 0$$

Therefore, $C_0 = 1/T_0$, $C_n = 2/T_0$ and $\theta_n = 0$. Thus

$$\delta_{T_0}(t) = \frac{1}{T_0}\left(1 + 2\sum_{n=1}^{\infty} \cos n\omega_0 t\right).$$

Figure 3.6 (b) shows the amplitude spectrum. The phase spectrum is zero.

Now we can show that a signal whose spectrum is bandwidth limited to B Hz can be reconstructed exactly (without any error) from its samples taken uniformly at a rate $R > 2B$ Hz (samples per second). In other words, the minimum sampling frequency is $f_s = 2B$ Hz. Consider a signal $g(t)$ (figure 3.7 (a)) whose spectrum is bandwidth limited to B Hz (figure 3.7 (b)). Sampling $g(t)$ at a rate of f_s Hz can be accomplished by multiplying $g(t)$ by an impulse train $\delta_{T_s}(t)$ (figure 3.7 (c)), consisting of delta functions repeating periodically every T_s seconds, where $T_s = 1/f_s$. This results in the sampled signal $\overline{g}(t)$ shown in figure 3.7 (d). The sampled signal consists of impulses spaced every T_s seconds. The nth impulse, located at $t = nT_s$, has strength $g(nT_s)$, the value of the $g(t)$ at $t = nT_s$. Thus,

$$\overline{g}(t) = g(t)\delta_{T_s}(t) = \sum_n g(nT_s)\delta(t - nT_s)$$

From above, the Fourier series for $\delta_{T_s}(t)$ is

$$\delta_{T_s}(t) = \frac{1}{T_s}[1 + 2\cos\omega_s t + 2\cos 2\omega_s t + 2\cos 3\omega_s t + ...] \qquad \omega_s = \frac{2\pi}{T_s} = 2\pi f_s.$$

Therefore,

$$\begin{aligned}
\overline{g}(t) &= g(t)\delta_{T_s}(t) \\
&= \frac{1}{T_s}[g(t) + 2g(t)\cos\omega_s t + 2g(t)\cos 2\omega_s t + 2g(t)\cos 3\omega_s t + ...]
\end{aligned}$$

To find $\overline{G}(\omega)$, the Fourier transform of $\overline{g}(t)$, we take the Fourier transform of the right-hand side of the above equation, term by term. The transform of the first term in brackets is $G(\omega)$. The transform of the second term $2g(t)\cos\omega_s t$ is $G(\omega - \omega_s) + G(\omega + \omega_s)$. This represents the spectrum $G(\omega)$ shifted to $+\omega_s$ and $-\omega_s$. Similar expressions are found for the third term ($\pm 2\omega_s$ shift), and so on to infinity. This means that the spectrum $\overline{G}(\omega)$ consists of $G(\omega)$ repeating periodically with angular frequency $\omega_s = 2\pi/T_s$, as shown in figure 3.7 (e). Therefore,

$$\overline{G}(\omega) = \frac{1}{T_s}\sum_{n=-\infty}^{\infty} G(\omega - n\omega_s).$$

$G(\omega)$ can only be recovered from $\overline{G}(\omega)$ (and hence $g(t)$) if there is no overlap between successive cycles of $\overline{G}(\omega)$. Figure 3.7 (e) shows that this requires

$$f_s > 2B$$

Also, the sampling interval $T_s = 1/f_s$. Therefore,

$$T_s < \frac{1}{2B}.$$

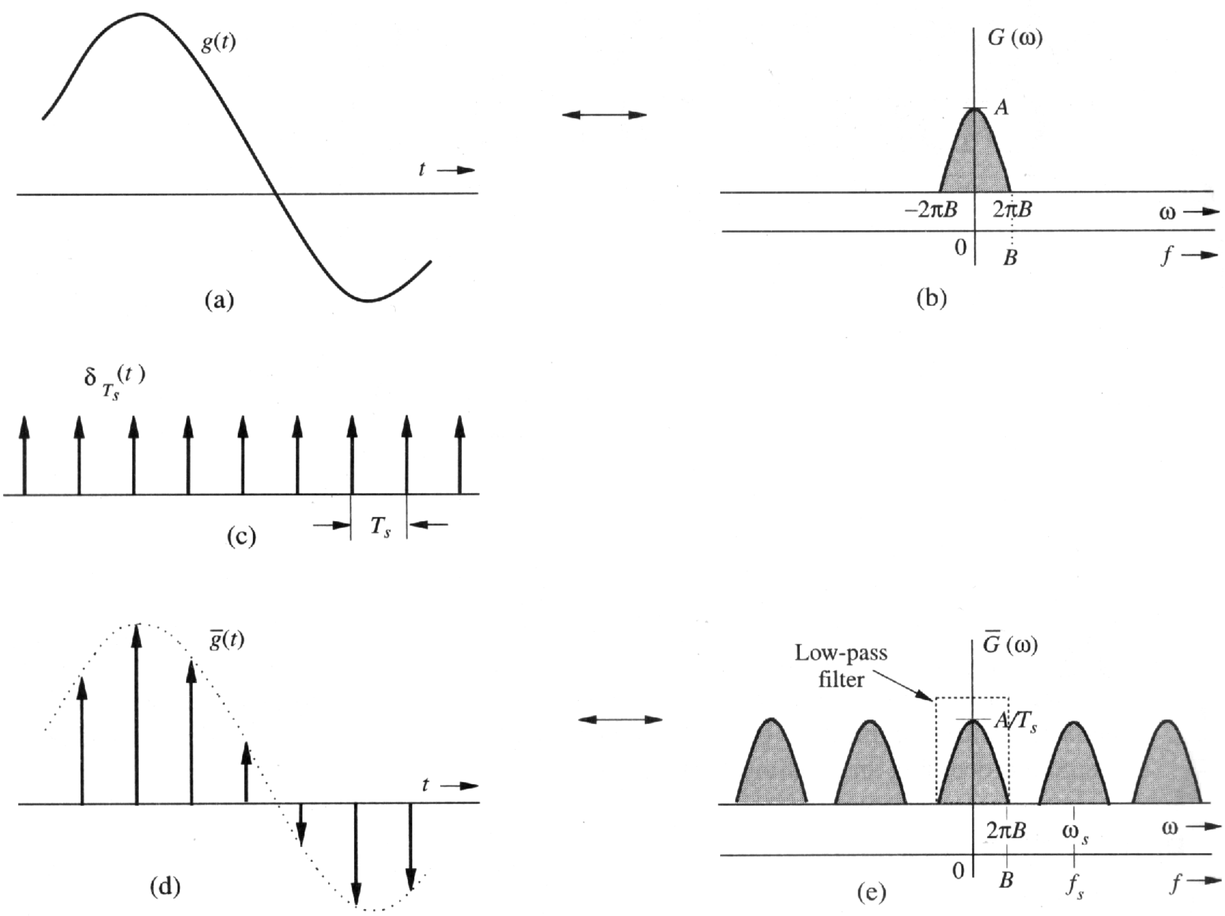

Figure 3.7 Sampled signal and its Fourier spectrum.

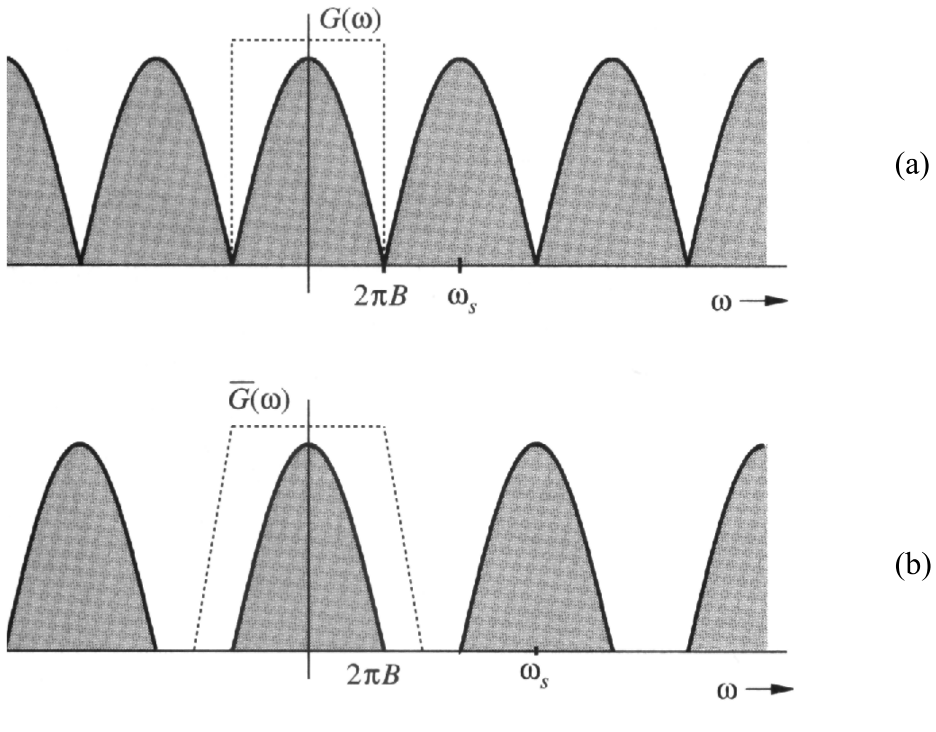

Figure 3.8 Spectra of a sampled signal. (a) At the Nyquist rate. (b) Above the Nyquist rate.

Thus, as long as the sampling frequency f_s is greater than twice the signal bandwidth B, $\overline{G}(\omega)$ will consist of non-overlapping repetitions of $G(\omega)$. When this is true, figure 3.7 (e) shows that $g(t)$ can be recovered from its samples $\overline{g}(t)$ by passing the sampled signal $\overline{g}(t)$ through an ideal low-pass filter of bandwidth B Hz. The minimum sampling rate $f_s = 2B$ required to recover $g(t)$ from its samples $\overline{g}(t)$ is called the **Nyquist rate** for $g(t)$, and the corresponding sampling interval $T_s = 1/2B$ is called the **Nyquist interval** for $g(t)$.

2.2 Aliasing

There are two main problems in reconstructing the signal from its sampled transform. The first arises because it is impossible in practice to isolate the spectrum $\overline{G}(\omega)$ using a real filter if the data is sampled at the Nyquist rate. One solution is to sample faster than this, thus enabling a more realistic frequency filter to be employed. Unfortunately even this solution can only give a partial reconstruction as the frequency filter still has to have a gain of zero beyond the first cycle $G(\omega)$. This is impossible, although as the sampling rate increases, the recovered signal approaches the desired signal more closely. This is illustrated in figure 3.8.

The second fundamental practical difficulty in reconstructing a signal from its samples is due to aliasing. The proof of the sampling theorem above relied on the fact that the signal was bandwidth limited. This is not the case in practical signals, which are almost always time-limited (of finite duration of width). Thus they have a infinite bandwidth, and the spectrum $\overline{G}(\omega)$ consists of overlapping cycles of $G(\omega)$ repeating every f_s Hz (the sampling frequency), as shown in figure 3.9. We thus have spectral overlap as a constant feature, regardless of the sampling rate. Thus it is no longer possible, even theoretically, to recover $g(t)$ from the sampled signal $\overline{g}(t)$. If the sampled signal is passed through an ideal low pass filter, the output is not $G(\omega)$ but a version of $G(\omega)$ distorted as a result of two separate causes:

1. The loss of the tail of $G(\omega)$ beyond $|f| > f_s/2$ Hz.
2. The reappearance of this tail inverted or folded onto the spectrum.

The spectra cross at frequency $f_s/2 = 1/2T_s$ Hz. This frequency is called the **folding frequency**. Thus components with frequencies above $f_s/2$ reappear as components with frequencies below $f_s/2$. This tail inversion, known as **spectral folding** or **aliasing**, is shown shaded in figure 3.9. Compare this with Brillouin zones in condensed matter

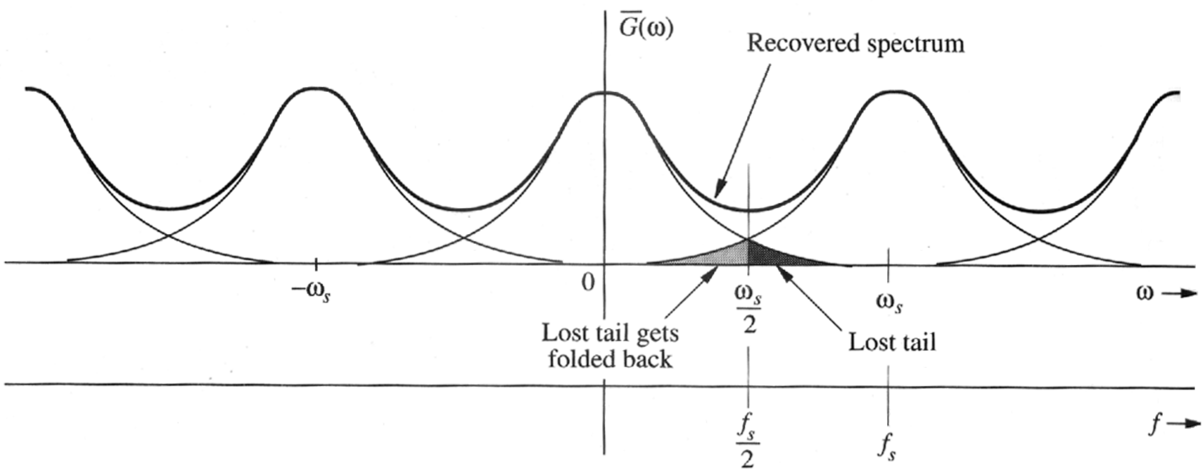

Figure 3.9 Aliasing effect.

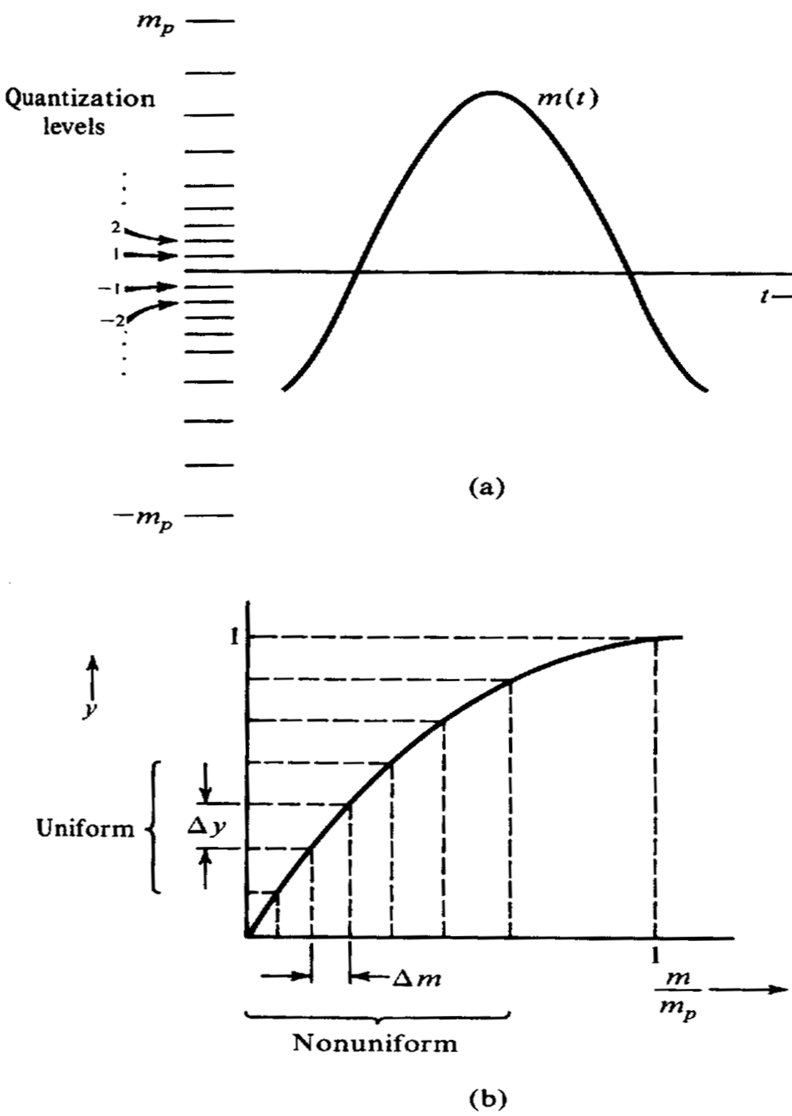

Figure 3.10 Non-uniform quantization.

physics.

The solution is to suppress the components with frequencies above $f_s/2$ from $g(t)$ *before* sampling $g(t)$. This way only the components beyond the folding frequency are lost, and no spurious signal appear below this frequency. This suppression of of higher frequencies can be accomplished by an ideal low-pass filter of bandwidth $f_s/2$ Hz. This filter is called an **antialiasing filter**, although ideal behaviour is not achievable in practice.

2.3 Quantizing

When a uniform digitising or quantizing algorithm is used, say for voltages between 0 and 16 V, then if the signal is broken up into 16 levels, the minimum voltage that can be represented is 1 V. For speech or music, where a large dynamic range is possible, this would mean that soft parts of the signal would be lost. In practical communication systems this problem is overcome by the use of non-uniform quantization. Referring to figure 3.4, it can be shown that the signal to noise ratio in a quantised system can be expressed as:

$$\frac{S}{N} = 3L^2 \frac{\overline{m^2(t)}}{m_p^2}$$

where m_p is the peak amplitude value, and $\overline{m^2(t)}$ is the signal power. This means that the SNR is a linear function of the message signal power. Ideally we would like to have a constant SNR for all values of the message signal power.

The root of this difficulty lies in the fact that the quantizing steps are of uniform value $\Delta v = 2m_p/L$. The quantization noise is, in fact, directly proportional to the square of the step size. The propblem can be solved by using smaller steps for smaller amplitudes (non-uniform quantization). This is illustrated in figure 3.10 (a). The same result is obtained by first compressing signal samples and then using a uniform quantization. The input–output characteristics of a compressor are shown in figure 3.10 (b). The horizontal axis is the normalised input signal (m/m_p), and the vertical axis is the output signal y. An approximately logarithmic compression characteristic yields a quantization noise nearly proportional to the signal power $\overline{m^2(t)}$, thus making the SNR practically independent of the input signal power over a large dynamic range (see figure 3.11).

Two compression laws have been adopted as world standards, the μ-law used in North America and Japan, and the A-law used in Europe and the rest of the world and

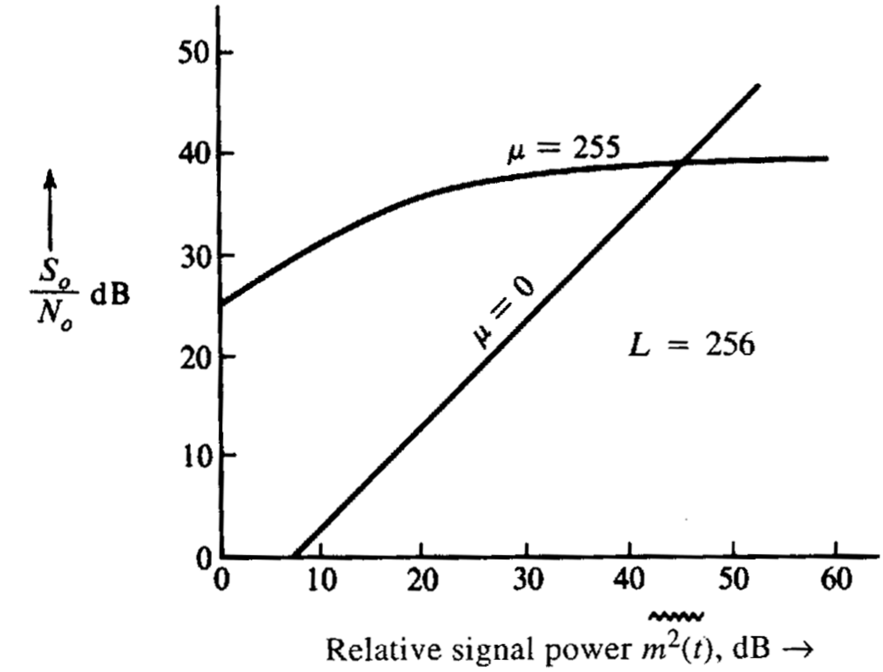

Figure 3.11 Signal-to-quantization noise ratio in PCM with and without compression.

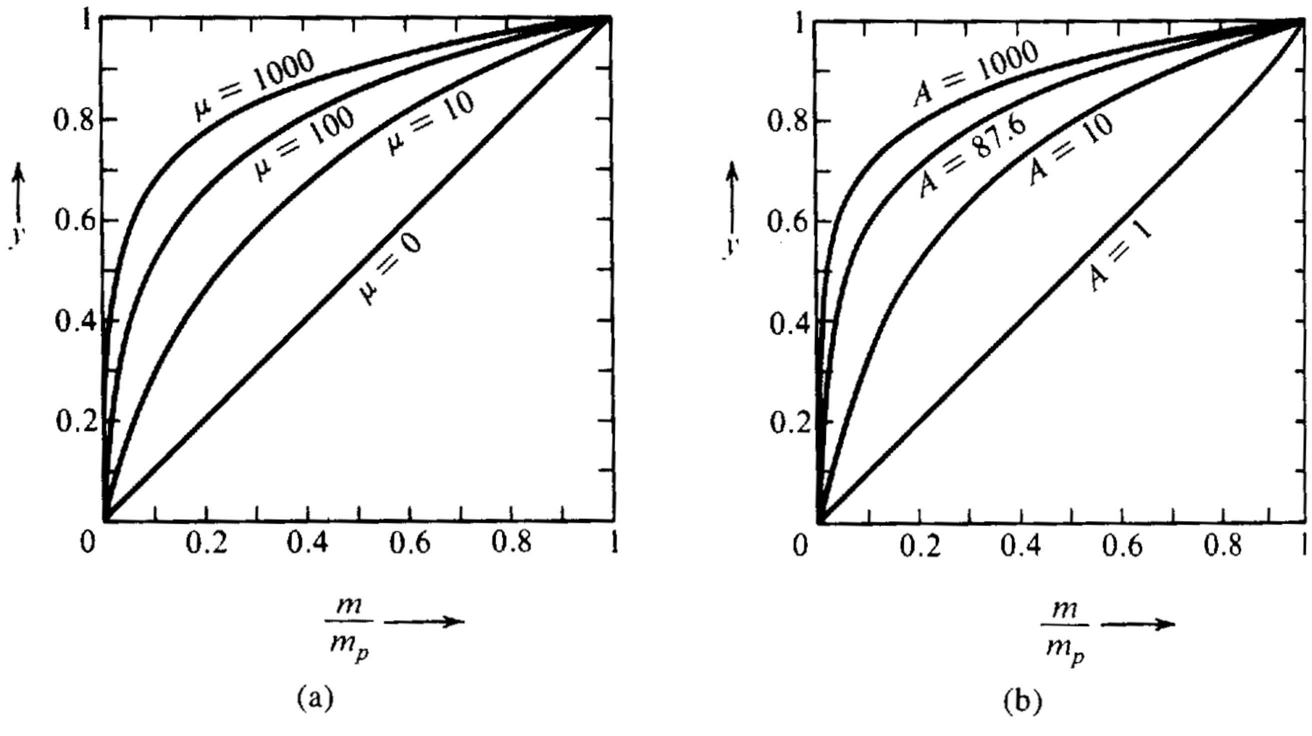

Figure 3.12 (a) μ-law characteristic. (b) A-law characteristic.

international routes. The μ-law (for positive amplitudes) is given by

$$y = \frac{1}{\ln(1+\mu)} \ln\left(1 + \frac{\mu m}{m_p}\right) \qquad 0 \le \frac{m}{m_p} \le 1$$

and the A-law (for positive amplitudes) is

$$\begin{aligned} y &= \frac{A}{1+\ln A}\left(\frac{m}{m_p}\right) && 0 \le \frac{m}{m_p} \le \frac{1}{A} \\ y &= \frac{1}{1+\ln A}\left(1 + \ln\frac{Am}{m_p}\right) && \frac{1}{A} \le \frac{m}{m_p} \le 1 \end{aligned}$$

These characteristics are shown in figure 3.12.

2.4 Historical Note form Lathi's Book

Gottfried Wilhelm Leibnitz (1646-1716) was the first mathematician to work out systematically the binary representation (using **1**'s and **0**'s) for any number. He felt a spiritual significance in this discovery, reasoning that **1**, representing unity, was clearly a symbol for God, while **0** represented the nothingness. Therefore, if all numbers can be represented merely by the use of **1** and **0**, surely this proves that God created the universe out of nothing!

3 Advantages of Digital Communication

Some advantages of digital communication over analogue communication are listed below:
1. Digital communication is more rugged than analogue communication because it can withstand channel noise and distortion much better as long as the noise and distortion are within limits. This is not true of analogue messages; any distortion or noise, no matter how small, will distort the received signal.
2. The greatest advantage of digital communication over analogue communication, however, is the viability of regenerative repeaters. Amplifying noisy weak analogue signals does not improve the SNR, and thus the signal is lost eventually in the noise and distortion. A regenerative amplifier placed in a digital communication system before the noise and distortion gets too bad can reconstruct the signal exactly, thus enabling transmission over great distances with high reliability. The most significant error in PCM comes from quantizing.
3. Digital hardware implementation is flexible and permits the use of microprocessors, digital switching, and large-scale integrated circuits.

4. Digital signals can be coded to yield extremely low error rates and high fidelity as well as privacy.

5. It is easier and more efficient to multiplex several digital signals.

6. Digital communication is inherently more efficient than analogue in realising the exchange of SNR for bandwidth.

7. Digital storage is relatively easy and inexpensive. It also has the ability to search and select information from distant electronic storehouses.

8. Reproduction with digital messages is extremely reliable without deterioration, as in digital imagery and CDs.

9. The cost of digital hardware falls while performance and capacity continues to rise. There seems to be no end to this trend yet.

3.1 Logarithmic Units

A logarithmic unit for the power ratio is the decibel (dB), defined as $10\log_{10}$ (power ratio). Thus, a SNR is x dB, where

$$x = 10\log_{10}\frac{S}{N}$$

Although the decibel is a measure of power ratios, it is often used as a measure of power itself. For instance 100 W of power may be considered as a power ratio of 100 with respect to 1 W, and is expressed in units of dBW as

$$P_{\mathrm{dBW}} = 10\log_{10}100 = 20\,\mathrm{dBW}$$

Thus 100 W of power is 20 dBW. Similarly, power measured with respect to 1 mW is dBm. For instance 100 W is

$$P_{\mathrm{dBm}} = 10\log_{10}\frac{100\,\mathrm{W}}{1\,\mathrm{mW}} = 50\,\mathrm{dBm}.$$