Telecommunications and RF

Semester 2, 2023

Dr Jacob Coetzee

Tarang Janawalkar

Contents

1   Telecommunications Systems

Telecommunication is the transmission of information over a distance through some technology. Telecommunications systems are designed to transmit information with as little **deterioration** as possible while satisfying design constraints, such as allowable transmittable energy and signal bandwidth.
Signal deterioration commonly measures:

- For Analog Systems: **Signal-to-Noise Ratio** (SNR) at the receiver output — the ratio of the signal power to the noise power

- For Digital Systems: **Bit Error Rate** (BER) at the receiver output — the ratio of the number of bits received in error to the total number of bits transmitted

1.1   Analog and Digital Signals

A system is either **analog** or **digital** based on the possible amplitudes of waveforms it can handle.

- **Analog Information Sources** produce values defined on a continuum:

    - Human voice and other sounds

- **Digital Information Sources** produce a finite set of possible symbols:

    - Computer data
    - MP3 encoder output

1.2   Communications System

A communications system can be summarised by the following block diagram:
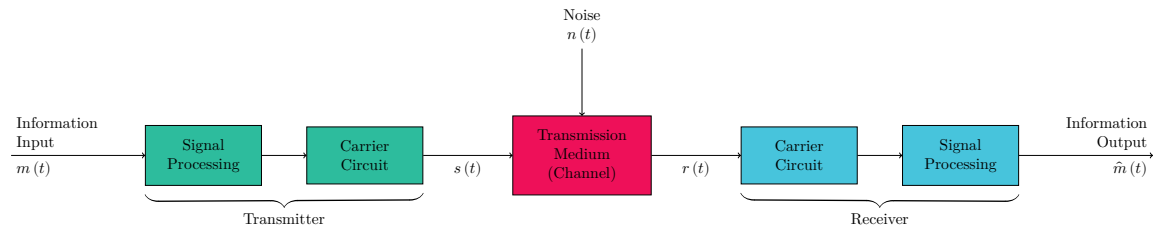
where

- $m(t)$ is the information message signal (prior to conditioning for transmission)

- $s(t)$ is the conditioned signal for transmission

- $n(t)$ contains channel noise and interference

- $r(t)$ is the received signal

- $\hat{m}(t)$ is the reconstructed received message signal (where $\hat{m}(t)$ usually approximates $m(t)$)

1.2.1   Transmitter

The transmitter carries out **signal conditioning** by transforming the signal to a more appropriate form before transmission through the channel. Some examples of signal conditioning techniques are prodived below.

- **Low Pass Filtering** (LPF) restricts signal bandwidth to avoid wasting signal power on frequencies that are not transmitted by the channel and to avoid interference with other signals

- **Analog to Digital Conversion** (ADC) produces a digital word which represents a sample of the analog message waveform

- **Carrier Modulation** transfers the signal to a frequency band that is suitable for transmission through the channel

1.2.2   Channel

A communication channel refers to the physical medium that carries the signal from the transmitter to the receiver. There are two types of channels:

- **Wired**: twisted pair copper telephone lines, waveguides, coaxial cables, fibre-optic cables

- **Wireless**: air, vacuums, sea water, optical fibres

General principles of communications always apply regardless of the type of channel. However, certain conditioning methods are better suited to certain channels.
Channels often **attenuate** signals (reduce their amplitude or strength) through

- random noise

- interference from other sources

and therefore it is a key consideration in the design of a communications system.

1.2.3   Receiver

The receiver acts as the inverse of a transmitter. The receiver:

1. **Demodulates** the received signal by stripping the carrier from the received signal $r(t)$.

2. **Filters** out noise and interference from the demodulated signal.

3. **Reconstructs** an estimate of the original message signal $\hat{m}(t)$.

Due to the finite nature of the SNR, the estimated output of an analog signal can never be exactly equal to the original signal[1]. However, it is often possible to reconstruct a digital signal exactly using error detection and correction techniques at the receiver.

---

[1]A perfect reconstruction requires an infinite SNR which is impractical.

1.2.4 Information Sources

As discussed before, an information source can be classified as either **analog** or **digital**. Analog signals can be modulated or transmitted directly, or converted to digital data and transmitted using digital modulation techniques.

An analogue signal to be transmitted is called the **message signal** and is denoted $m(t)$. The spectral components of this signal lie within a finite bandwidth W, such that $M(f) = 0$ for $|f| > W$, where $M(f)$ is the Fourier Transform of $m(t)$. This signal's bandwidth is limited to prevent interference with other signals.

Many kinds of message signals can be considered:

- Audio

- Video

- Computer data

- Telemetry (measurements)

- Soundings (RADAR, SONAR)

- A mixure of the above (i.e., data over voice)

1.3 Modulation

The process of modulation produces a signal that is suitable for transmission through the channel by transforming the message signal $m(t)$ to a new signal $s(t)$.

Modulation is often performed with respect to another signal, called the **carrier** signal $c(t)$. Here the message *modulates* the carrier to produce the transmitted signal $s(t)$.

1.3.1 Benefits of Modulation

- Modulation shifts the spectral content of a message onto a suitable band. As the size of an antenna is related to the wavelength of a signal, higher carrier frequencies require smaller antennas.

- Modulation facilitates multiplexing, where multiple signals are transmitted over the same spectrum. This is because the frequency spectrum can be divided into non-overlapping frequency bands, each of which can carry a separate signal.

- Modulation provides some control over noise and interference by choosing a bandwidth that is smaller than the allocated channel bandwidth.

1.3.2 Convolution and Modulation

Consider two time domain signals $m(t)$ and $c(t)$ and their Fourier Transforms $M(f)$ and $C(f)$ respectively.

The **Convolution Property** demonstrates:

$$m(t) * c(t) = y(t) \qquad \text{Convolution in Time Domain}$$
$$M(f)C(f) = Y(f) \qquad \text{Multiplication in Frequency Domain}$$

The **Modulation property** demonstrates:

$$m(t)\,c(t) = y(t) \qquad \text{Multiplication in Time Domain}$$
$$M(f) * C(f) = Y(f) \qquad \text{Convolution in Frequency Domain}$$

1.3.3   Carrier Signal

The carrier signal $c(t)$ is a sinusoidal signal of the form:

$$c(t) = A_c \cos(2\pi f_c t + \phi)$$

where A_c is the amplitude, f_c is the frequency and ϕ is the phase of the carrier signal.

1.3.4   Signal Properties

The **energy** of a signal $x(t)$ is defined as:

$$E_x = \int_{-\infty}^{\infty} |x(t)|^2 \, \mathrm{d}t$$

The rate at which energy is transmitted is called the **power** of the signal, and is defined as:

$$P_x = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} |x(t)|^2 \, \mathrm{d}t$$

where T is the time period over which the power is measured.

1.4   Amplitude Modulation Schemes

A message signal $m(t)$ with bandwidth W can be amplitude modulated by mixing (multiplying) the signal with a carrier signal $c(t)$:

$$c(t) = A_c \cos(2\pi f_c t)$$

with amplitude A_c and carrier frequency $f_c \gg W$.
This section will consider:

- Double Side Band — Suppressed Carrier (DSB-SC) modulation

- Double Side Band — Full Carrier (DSB-FC) modulation, or simply conventional Amplitude Modulation (AM)

- Single Side Band (SSB) modulation

1.4.1   Double Side Band — Suppressed Carrier (DSB-SC)

The transmitted signal is defined as:

$$s\left(t\right) = m\left(t\right)c\left(t\right) = A_c m\left(t\right)\cos\left(2\pi f_c t\right).$$

In the frequency domain, the message signal is shifted to the carrier frequency at $\pm f_c$ and the magnitude is halved.

$$S\left(f\right) = \frac{1}{2}A_c\left[M\left(f - f_c\right) + M\left(f + f_c\right)\right]$$

Assuming no channel noise, the received signal can be represented as:

$$r\left(t\right) = s\left(t\right) = A_c m\left(t\right)\cos\left(2\pi f_c t\right)$$

This signal can be demodulated coherently by multiplying a sinusoidal signal of the same frequency as the carrier. Suppose we generate the sinusoidal signal $\cos\left(2\pi f_c t + \phi\right)$ where ϕ is the phase of the sinusoid.

$$
\begin{aligned}
y\left(t\right) &= r\left(t\right)\cos\left(2\pi f_c t + \phi\right) \\
&= A_c m\left(t\right)\cos\left(2\pi f_c t\right)\cos\left(2\pi f_c t + \phi\right) \\
&= \frac{1}{2}A_c m\left(t\right)\left[\cos\left(4\pi f_c t + \phi\right) + \cos\left(-\phi\right)\right] \\
&= \underbrace{\frac{1}{2}A_c m\left(t\right)\cos\left(\phi\right)}_{\text{filter out}} + \underbrace{\frac{1}{2}A_c m\left(t\right)\cos\left(4\pi f_c t + \phi\right)}_{\text{reject}}
\end{aligned}
$$

As the frequency content of the message signal $m\left(t\right)$ is limited to W, a lowpass filter can be used to remove the high frequency component centred at $2f_c$. The output of this filter is then

$$\widehat{m}\left(t\right) = \frac{1}{2}A_c m\left(t\right)\cos\left(\phi\right).$$

Note that $m\left(t\right)$ is multiplied by $\cos\left(\phi\right)$, meaning the power of the message signal is reduced by a factor of $\cos^2\left(\phi\right)$. It is therefore important to choose ϕ such that $\cos\left(\phi\right)$ is as close to 1 as possible. This demonstrates a need for a phase-coherent or synchronous demodulator, i.e., the phase of the locally generated sinusoid should be identical to the received carrier signal.

1.4.2   Double Side Band — Full Carrier (DSB-FC)

This scheme is similar to DSB-SC, however the carrier is sent along with the modulated signal. We begin by defining an envelope signal $g\left(t\right)$ by amplifying and biasing the message signal, so that:

$$g\left(t\right) = 1 + \mu m_n\left(t\right)$$

where the **modulation index** $0 < \mu \leqslant 1$ is chosen to ensure $g\left(t\right) > 0$. $m_n\left(t\right)$ is the normalised message signal defined as

$$m_n\left(t\right) = \frac{m\left(t\right)}{\max\left|m\left(t\right)\right|}$$

such that $|m_n(t)| \leqslant 1$. The transmitted signal is then defined as:

$$s(t) = g(t)c(t) = A_c[1 + \mu m_n(t)]\cos(2\pi f_c t).$$

In the frequency domain, we observe similar behaviour, with impulses at $\pm f_c$.

$$S(f) = \frac{1}{2}A_c\mu[M(f - f_c) + M(f + f_c)] + \frac{1}{2}A_c[\delta(f - f_c) + \delta(f + f_c)]$$

The modulation index can be determined from the AM signal using the following equation:

$$\mu = \frac{A_{\max} - A_{\min}}{A_{\max} + A_{\min}} = \frac{\max|m(t)|}{\max|c(t)|}$$

where $A_{\max}$ and $A_{\min}$ are the maximum and minimum amplitudes of the envelope signal $g(t)$. This value is chosen to be as close to 1 as possible to maximise the power of the transmitted signal. Increasing μ beyond 1 overmodulates the signal, resulting in a full-wave rectified version of the signal.

Modulation efficiency is the percentage of the total power of the modulated signal that conveys information

$$\eta = \frac{\text{sideband power}}{\text{total power}} \times 100\%$$

As the total power comprises of both the sideband power and carrier power, the DSB-FC scheme is very inefficient. This is because the carrier signal does not contain any useful information, and therefore the power of the carrier is wasted. This can be seen when transmitting a sinusoidal message signal,

- Carrier power: $\frac{1}{2}A_c^2$

- Sideband power: $\frac{1}{4}A_c^2\mu^2$

- Total power: $\frac{1}{2}A_c^2\left(1 + \frac{1}{2}\mu^2\right)$

- Modulation efficiency: $\frac{\frac{1}{4}A_c^2\mu^2}{\frac{1}{2}A_c^2\left(1 + \frac{1}{2}\mu^2\right)} = \frac{\mu^2}{2 + \mu^2}$

The maximum efficiency is achieved when $\mu = 1$, i.e., which is 33%. This scheme is often used in AM medium-wave (MW) systems.

To demodulate a DSB-FC signal, we can use an envelope detector. In this circuit, we need to calculate the minimum and maximum frequency that the tuned circuit can respond to:

$$f_0 = \frac{1}{2\pi\sqrt{LC_1}}$$

and calculate the break frequency of the lowpass filter:

$$f_b = \frac{1}{2\pi RC_2}$$

1.4.3 Single Side Band (SSB)

As the double side band schemes produce a signal with twice the bandwidth of the message signal, consider either the lower or upper sidebands of the modulated signal. We can attempt to use a filter to remove the unwanted sideband, however there are two issues with this approach:

1. The filter is particularly difficult to implement when $m(t)$ has a large concentration of power close to $f = 0$

2. The filter must have a very sharp cutoff in the vicinity of the carrier frequency

Instead, we can use a phase shift method through the use of a Hilbert transform. The Hilbert transform of a signal $m(t)$ is defined as

$$\hat{m}(t) = m(t) * \frac{1}{\pi t}$$

The result of this operation is more easily understood in the frequency domain.

$$\hat{M}(f) = -j\,\mathrm{sgn}(f)M(f) = \begin{cases} -jM(f) & f > 0 \\ jM(f) & f < 0 \end{cases}$$

so that negative frequency components of m are phase shifted by $+90°$ $(+\pi/2)$ and positive frequency components are phase shifted by $-90°$ $(-\pi/2)$. The magnitude remains unchanged.
Using this transform, we can define SSB signals as:

$$s_{\mathrm{USB}}(t) = A_c\left[m(t)\cos(2\pi f_c t) - \hat{m}(t)\sin(2\pi f_c t)\right]$$
$$s_{\mathrm{LSB}}(t) = A_c\left[m(t)\cos(2\pi f_c t) + \hat{m}(t)\sin(2\pi f_c t)\right]$$

or compactly,

$$s(t) = m(t)\,c(t) \pm \hat{m}(t)\,\hat{c}(t)$$

These signals can be demodulated coherently using the same process as the DSB-SC scheme. Again assuming that there is no channel noise,

$$r(t) = s(t) = A_c m(t)\cos(2\pi f_c t) \pm A_c \hat{m}(t)\sin(2\pi f_c t)$$

$$\begin{aligned} y(t) &= r(t)\cos(2\pi f_c t + \phi) \\ &= \left[A_c m(t)\cos(2\pi f_c t) \pm A_c \hat{m}(t)\sin(2\pi f_c t)\right]\cos(2\pi f_c t + \phi) \\ &= A_c m(t)\cos(2\pi f_c t)\cos(2\pi f_c t + \phi) \pm A_c \hat{m}(t)\sin(2\pi f_c t)\cos(2\pi f_c t + \phi) \\ &= \frac{1}{2}A_c m(t)\left[\cos(\phi) + \cos(4\pi f_c t + \phi)\right] \pm \frac{1}{2}A_c \hat{m}(t)\left[-\sin(\phi) + \sin(4\pi f_c t + \phi)\right] \\ &= \underbrace{\frac{1}{2}A_c\left[m(t)\cos(\phi) \mp \hat{m}(t)\sin(\phi)\right]}_{\text{filter out}} + \underbrace{\frac{1}{2}A_c\left[m(t)\cos(4\pi f_c t + \phi) \pm \hat{m}(t)\sin(4\pi f_c t + \phi)\right]}_{\text{reject}} \end{aligned}$$

This gives the output

$$\hat{m}(t) = \frac{1}{2}A_c\left[m(t)\cos(\phi) \mp \hat{m}(t)\sin(\phi)\right]$$

Note the $\hat{m}(t)$ term on the right hand side of the equation is referring to the Hilbert transform of $m(t)$, not the output of the demodulator.

2   Angle Modulation

AM suffers from poor noise performance, as amplitude variations in the received signal cannot be removed from the demodulated signal. Angle modulation schemes overcome this issue by modulating the phase or frequency of the carrier signal.

2.1   Carrier Signal

In these schemes, the message signal is modulated onto the angle $\theta(t)$ of the carrier signal:

$$s(t) = A_c \cos(\theta(t))$$
$$\theta(t) = 2\pi f_c t + \phi(t)$$

In **phase modulation** (PM), variations in the message signal are encoded into the phase:

$$\phi(t) = k_p m(t)$$

where k_p is the **phase deviation** constant.
Frequency modulation (FM) considers the frequency deviation from the modulation frequency f_c:

$$f_i - f_c = k_f m(t)$$

where f_i is the instantaneous frequency of the carrier signal and k_f is the **frequency deviation** constant. To determine the phase ϕ, consider the time derivative of the angle $\theta(t)$ with arbitrary frequency f

$$\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(2\pi f t + \phi)$$

$$\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = 2\pi f_i$$

$$\theta(t) = 2\pi \int_0^t f_i \,\mathrm{d}\tau$$

$$\theta(t) = 2\pi \int_0^t f_c + k_f m(\tau) \,\mathrm{d}\tau$$

$$\theta(t) = 2\pi f_c t + 2\pi k_f \int_0^t m(\tau) \,\mathrm{d}\tau$$

2.1.1   Phase Modulation

A phase modulated signal is defined as:

$$s(t) = A_c \cos\left(2\pi f_c t + k_p m(t)\right)$$

with

$$\theta(t) = 2\pi f_c t + k_p m(t)$$

where k_p is the **phase deviation** constant.

2.1.2   Frequency Modulation

A frequency modulated signal is defined as:

$$s(t) = A_c \cos\left(2\pi f_c t + 2\pi k_f \int_0^t m(\tau)\,\mathrm{d}\tau\right)$$

with

$$\theta(t) = 2\pi f_c t + 2\pi k_f \int_0^t m(\tau)\,\mathrm{d}\tau$$

where k_f is the **frequency deviation** constant.

2.1.3   Instantaneous Frequency

The instantaneous frequency f_i may be determined from these signals by considering the time derivative of θ.
In PM,

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = 2\pi f_c + k_p \frac{\mathrm{d}}{\mathrm{d}t} m(t)$$
$$= 2\pi\left(f_c + \frac{k_p}{2\pi}\frac{\mathrm{d}}{\mathrm{d}t} m(t)\right) \implies f_i = f_c + \frac{k_p}{2\pi}\frac{\mathrm{d}}{\mathrm{d}t} m(t)$$

In FM,

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = 2\pi f_c + 2\pi k_f m(t)$$
$$= 2\pi\left(f_c + k_f m(t)\right) \implies f_i = f_c + k_f m(t)$$

2.1.4   Modulation Index

The modulation index β is defined as the ratio of the maximum frequency deviation to the modulating frequency of the message signal. In PM, if we rewrite the phase $\phi(t)$ as,

$$\phi(t) = k_p \max|m(t)|\frac{m(t)}{\max|m(t)|}$$
$$= \beta_p m_n(t)$$

then β_p is the modulation index for PM. By definition, FM has a modulation index of $\beta_f = \frac{k_f \max|m(t)|}{B}$ where B is the bandwidth of the message signal.

2.1.5   Maximum Phase and Frequency Deviation

The maximum phase deviation Δp and maximum frequency deviation Δf are defined as:

$$\Delta p = k_p \max|m(t)| \qquad \Delta f = k_f \max|m(t)|$$

2.1.6   Carson's Rule

Phase and frequency modulation generally expand the bandwidth of a signal. The modulated signals bandwidth W may be approximated by Carson's rule:

$$W = 2B\left(\beta + 1\right)$$

2.1.7   Summary of Angle Modulation Definitions

	Phase Modulation	Frequency Modulation
Modulated Signal $s\left(t\right)$	$A_c \cos\left(2\pi f_c t + k_p m\left(t\right)\right)$	$A_c \cos\left(2\pi f_c t + 2\pi k_f \int_0^t m\left(\tau\right)\mathrm{d}\tau\right)$
Phase $\phi\left(t\right)$	$k_p m\left(t\right)$	$2\pi k_f \int_0^t m\left(\tau\right)\mathrm{d}\tau$
Instantaneous Frequency f_i	$f_c + \dfrac{k_p}{2\pi}\dfrac{\mathrm{d}}{\mathrm{d}t}m\left(t\right)$	$f_c + k_f m\left(t\right)$
Modulation Index	$\beta_p = k_p \max\left|m\left(t\right)\right|$	$\beta_f = \dfrac{k_f \max\left|m\left(t\right)\right|}{B}$
Maximum Deviation	$\Delta p = k_p \max\left|m\left(t\right)\right|$	$\Delta f = k_f \max\left|m\left(t\right)\right|$

2.2   Narrowband Angle Modulation

When $\beta \ll 1$, angle modulation is referred to as narrowband angle modulation. The modulated signal can be approximated as:

$$\begin{aligned}
s\left(t\right) &= A_c \cos\left(2\pi f_c t + \phi\left(t\right)\right) \\
&= A_c \cos\left(2\pi f_c t\right)\cos\left(\phi\left(t\right)\right) - A_c \sin\left(2\pi f_c t\right)\sin\left(\phi\left(t\right)\right) \\
&= A_c \cos\left(2\pi f_c t\right) - A_c \phi\left(t\right)\sin\left(2\pi f_c t\right)
\end{aligned}$$

This is equivalent to a DSB-FC signal with a phase modulated carrier, as the message signal is now modulated onto a sine carrier instead of cosine. The modulated signal bandwidth W is approximately $2B$.

Narrowband angle modulation (low-index angle modulation) does not provide better noise immunity than AM, and is seldom used on its own in practical communication systems.

2.3   Wideband Angle Modulation

Assuming the message signal is a sinusoidal signal, $\phi\left(t\right) = \beta \sin\left(2\pi f_m t\right)$,

$$\begin{aligned}
s\left(t\right) &= A_c \cos\left(2\pi f_c t + \beta \sin\left(2\pi f_m t\right)\right) \\
&= A_c \cos\left(2\pi f_c t\right)\cos\left(\beta \sin\left(2\pi f_m t\right)\right) - A_c \sin\left(2\pi f_c t\right)\sin\left(\beta \sin\left(2\pi f_m t\right)\right)
\end{aligned}$$

To simplify this expression, we can use Bessel functions. Notably

$$\cos\left(\beta\sin\left(2\pi f_m t\right)\right) = J_0\left(\beta\right) + 2\sum_{n=1}^{\infty} J_{2n}\left(\beta\right)\cos\left(2\pi\left(2n\right)f_m t\right)$$

$$\sin\left(\beta\sin\left(2\pi f_m t\right)\right) = 2\sum_{n=1}^{\infty} J_{2n-1}\left(\beta\right)\sin\left(2\pi\left(2n-1\right)f_m t\right)$$

where $J_n\left(\beta\right)$ are Bessel functions of the first kind and order n, evaluated at β.
These results can be used to show that a wideband FM modulated signal can be expressed as:

$$s\left(t\right) = \sum_{n=-\infty}^{\infty} A_c J_n\left(\beta\right)\cos\left(2\pi\left(f_c + nf_m\right)t\right)$$

so that the carrier frequency is located at $2\pi f_c$ with magnitude $A_c J_0\left(\beta\right)$, with an infinite number of sidebands at $2\pi\left(f_c \pm nfm\right)$ with magnitudes $A_c J_{\pm n}\left(\beta\right)$.
When deciding the number of sidebands n to include, consider

$$\left|J_{\pm n}\left(\beta\right)\right| \geqslant 0.1$$

For large values of β, $n \approx \beta$ is sufficient.
According to Carson's rule, $W = 2f_m\left(\beta + 1\right)$ contains at least 98% of the signal power.

2.3.1   Bessel Function Properties

The Bessel functions of the first kind $J_n\left(x\right)$ are defined as the solutions to the Bessel differential equation

$$x^2\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + x\frac{\mathrm{d}y}{\mathrm{d}x} + \left(x^2 - n^2\right)y = 0$$

- $J_n\left(\beta\right)$ is a real function

- $J_n\left(\beta\right) = J_{-n}\left(\beta\right)$

- $\sum_{n=-\infty}^{\infty} J_n^2\left(\beta\right) = 1$

For small values of β,

- $J_0\left(\beta\right) = 1$

- $J_1\left(\beta\right) = \beta/2$

- $J_n\left(\beta\right) = 0$ for $n > 2$

2.4   Effect of Bandwidth

Rewriting Carson's rule using β,

$$W = 2B\left(\beta + 1\right) = \begin{cases} 2B\left(k_p \max\left|m\left(t\right)\right| + 1\right) & \text{PM} \\ 2\left(k_f \max\left|m\left(t\right)\right| + B\right) & \text{FM} \end{cases}$$

From these equations

- Increasing the amplitude of the modulating signal has the same effect in both PM and FM.

- Increase the message signal bandwidth B has a greater effect on the bandwidth of a PM signal than for FM.

2.5 Angle Modulator Implementation

FM can be generated using a Voltage Controlled Oscillator (VCO). A varactor diode is a capacitor whose capacitance changes with applied voltage. This capacitor can be used in the tuned circuit of an oscillator. If the message signal is applied to the varactor, the output frequency of the oscillator will change in accordance with the message signal.
The time varying capacitance of the varactor diode is given by

$$C_v(t) = C_a + k_0 m(t)$$

- When $m(t) = 0$, the frequency of the tuned circuit is given by

$$f_i = f_c = \frac{1}{2\pi\sqrt{LC_a}}$$

- When $m(t) \neq 0$, the frequency of the tuned circuit is given by

$$f_i = \frac{1}{2\pi\sqrt{L(C_a + k_0 m(t))}} = \frac{1}{2\pi\sqrt{LC_a(1 + k_0 m(t)/C_a)}} = f_c \frac{1}{\sqrt{1 + k_0 m(t)/C_a}}$$

Narrowband FM can be converted to wideband using a narrowband-to-wideband convertor by multiplying the frequencies of the narrowhand signal.

2.6 Demodulating Angle Modulated Signals

FM demodulators are implemented by generating an AM signal whose amplitude is proportional to the instantaneous frequency of the FM signal. We can then use an AM demodulator to recover the message signal.
In such a circuit, the LTI system is a differentiator whose frequency response is approximately a straight line in the frequency band of the FM signal. $|H| = 2\pi f$.

2.7 AM vs. FM

- FM capture effect: When two FM signals are received, the stronger signal is demodulated and the weaker signal is ignored.

  - The complete suppression of the weaker signal occurs at the receiver limited, where it is treated as noise and rejected.
  - When both signals are of equal strength, the receiver may switch between the two signals.

- FM requires a higher bandwidth $W_{\text{AM}} < W_{\text{FM}}$.

- FM rejects amplitude noise cause by lightning and other man-made noise.

- AM demodulators: envelope detector, product detector.

- FM demodulators: PLL, ratio detector, frequency discriminator, slope detector.

3   Digital Baseband Modulation

A source encoder converts an analog signal into a sequence of binary digits (bits). Digital modulation is used to transmit digital information over an additive white Gaussian noise (AWGN) channel.

Definition 3.1 (Baseband Channel)**.** A baseband channel is a channel that includes the zero frequency $f = 0$, and there is no need to transmit a carrier signal.

Definition 3.2 (Bandpass Channel)**.** A communication channel that is far from the zero frequency is known as a bandpass channel, and such a signal is impressed on a sinusoidal carrier, that shifts the signal to the desired frequency band that is *passed* by the channel.

Pulse code modulation (PCM) is a method used to digitally represent a sampled analog signal. To transmit these digits, we can use electrical pulses to represent binary digits.

Definition 3.3 (Bit)**.** A bit is a binary digit, i.e., a digit that can take on one of two values; 0 or 1.

Definition 3.4 (Digital Symbol)**.** A digital symbol refers to a sequence of several bits. For a sequence of k bits, there are $M = 2^k$ unique symbols.

3.1   Waveform Representation of Digital Signals

Symbol rate R_s is the number of symbols transmitted per second, measured in symbols/s.
Bit rate R_b is the number of bits transmitted per second, measured in bits/s.
Symbol rate and bit rate are related by

$$R_b = kR_s$$

where k is the number of bits per symbol. When a symbol contains only one bit ($k = 1$), $R_b = R_s$. The duration of a single bit, or symbol can be determined using the reciprocal of each quantity:

$$T_s = \frac{1}{R_s} \qquad T_b = \frac{1}{R_b}$$

These quantities are related by

$$T_b = \frac{1}{k}T_s$$

To effectively transmit digital signals through an analog channel, each pulse must carry sufficient energy either by increasing the amplitude or duration of each pulse.

3.1.1   Antipodal and Orthogonal Signalling

Antipodal signalling is a method of transmitting digital signals where the information bit 1 is represented by a pulse $p(t)$ of duration T_b, and the information bit 0 is represented by $-p(t)$.
In orthogonal signalling, information is represented by two orthogonal pulses $p_1(t)$ and $p_2(t)$ such that

$$\int_0^{T_b} p_1(t)\, p_2(t)\, \mathrm{d}t = 0$$

Each of the above methods require precise synchronisation between the receiver and transmitter.

3.1.2   Not-Return To Zero

A non-return to zero (NRZ) **line code** is a binary code in which bits are represented by two different levels of voltage.
For bipolar NRZ, the bit 1 is represented by a positive voltage, and the bit 0 is represented by a negative voltage.
For unipolar NRZ, the bit 1 is represented by a positive voltage, and the bit 0 is represented by a zero voltage.

3.1.3   Pulse Amplitude Modulation

In pulse amplitude modulation (PAM), the amplitude of a pulse is varied between various voltage levels to represent digital symbols of k bits. The number of voltage levels is given by $M = 2^k$.

3.2   Signal Bandwidth

The bandwidth of a signal provides a measure of the extent of significant spectral content of the signal for positive frequencies. There are various definitions of bandwidth,

- **Null-to-null bandwidth**: range of frequencies between zeros in the magnitude spectrum

- **3-dB bandwidth**: range of frequencies where the magnitude spectrum falls no lower than $1/\sqrt{2}$ of the maximum value of the magnitude spectrum

- **Equivalent noise bandwidth**: width of a fictitious rectangular spectrum such that the power in the rectangular band is equal to the power associated with the actual spectrum over positive frequencies

3.2.1   Pulse Dilemma

The pulse dilemma is the trade-off between the bandwidth and the duration of a pulse. A narrow pulse has a large bandwidth, and a wide pulse has a small bandwidth.

- A perfectly band-limited pulse implies an infinite duration, which is not realisable.

- A precisely duration-limited pulse implies an infinite bandwidth, which is also not realisable.

When pulses are filtered by a communications system to reduce bandwidth, they are **spread** in time. This causes pulses of adjacent symbols to overlap in time, causing **intersymbol interference** (ISI).
In the following sections, we will consider various pulse shapes and analyse their bandwidths.

Definition 3.5 (Spectral Efficiency)**.** The spectral efficiency of a digital signal is the number of bits per second that can be supported by a hertz of bandwidth:

$$\eta = \frac{R_b}{B} \text{bits s}^{-1}\,\text{Hz}^{-1}$$

3.2.2 Rectangular Pulse

If a symbol is represented by a rectangular pulse, it's spectrum is the sinc function:

$$\Pi\left(\frac{t}{T_s}\right) \stackrel{\mathscr{F}}{\Longleftrightarrow} T_s \operatorname{sinc}\left(fT_s\right)$$

By using the null-to-null bandwidth definition, the bandwidth of this pulse is

$$B = \frac{1}{T_s} = R_s.$$

For binary modulation using bipolar NRZ, the spectral efficiency is given by

$$\eta = \frac{R_b}{R_s} = \frac{R_b}{B} = \frac{R_s}{R_s} = 1$$

3.2.3 Nyquist Pulse

A Nyquist pulse is a pulse that satisfies the **Nyquist criterion**, which states that the pulse must be zero at the sampling times of adjacent pulses. Assuming a sampling frequency of $R_s = 1/T_s$, the Nyquist pulse is defined

$$p\left(t\right) = \frac{\sin\left(\pi R_s t\right)}{\pi R_s t}$$

where each pulse is sampled at $t = nT_s$. From this definition, it can be seen that

$$p\left(nT_s\right) = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases}$$

The spectrum of this function is a rectangular function,

$$P\left(f\right) = \frac{1}{R_s}\Pi\left(\frac{f}{R_s}\right)$$

with bandwidth:

$$B \geqslant \frac{R_s}{2}$$

For binary modulation, the spectral efficiency of this pulse is given by

$$\eta = \frac{R_b}{B} = \frac{R_s}{R_s/2} = 2$$

In practise, this pulse shape is not physically realisable due to the following:

- infinite pulse in time domain

- sharp roll-off in frequency domain

- requirement for perfect synchronisation to ensure that the pulse is zero at the sampling times of adjacent pulses

3.2.4   Raised Cosine Pulse

If we relax the filtering and clock timing requirements of the Nyquist pulse, we can use a raised cosine (RC) pulse. The RC pulse has the form

$$p\left(t\right) = \frac{\sin\left(\pi R_s t\right)}{\pi R_s t} \frac{\cos\left(\pi r R_s t\right)}{1 - \left(2r R_s t\right)^2}$$

where r is the roll-off factor. The spectrum of this function is a finite bandwidth piecewise function of the form,

$$P\left(f\right) = \begin{cases} 1 & |f| < \frac{1-r}{2T_s} \\ \frac{1}{2}\left[1 + \cos\left(\frac{\pi T_s}{r}\left[|f| - \frac{1-r}{2T_s}\right]\right)\right] & \frac{1-r}{2T_s} < f < \frac{1+r}{2T_s} \\ 0 & |f| \geqslant \frac{1+r}{2T_s} \end{cases}$$

The bandwidth of this pulse is

$$B = \frac{R_s}{2}\left(1 + R\right)$$

The spectral efficiency of this pulse is given by

$$\eta = \frac{R_b}{B} = \frac{R_s}{R_s/2\left(1 + R\right)} = \frac{2}{1 + R}$$

3.3   Energy of a Digital Symbol

The symbol energy E is the instantaneous power integrated over the duration of one pulse T_s. For a symbol waveform $s\left(t\right)$:

$$E_s = \int_0^{T_s} s^2\left(t\right)\mathrm{d}t$$

such that E is the area under the squared symbol waveform.

If there are M symbols, each with energy E_i, the average symbol energy (assuming all M symbols are equiprobable) is the average of the symbol energies:

$$E_s = \frac{1}{M}\sum_{i=1}^{M} E_i$$

3.4   Additive White Gaussian Noise

Additive white Gaussian noise (AWGN) is a model for the effect of external noise on a signal. The probability density function of the AWGN process n is given by

$$f_N\left(n\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(n - \mu\right)^2}{2\sigma^2}\right]$$

where μ is the mean of the process, and σ^2 is the variance of the process.

The term *white* denotes a process in which all frequency components appear with equal power, i.e., the power spectral density is constant.

This is a result of the Wiener-Khinchin theorem, which states that the power spectral density of a wide sense stationary random process is the Fourier transform of the autocorrelation function of the process. For the above AWGN process, the autocorrelation function is given by

$$R_N\left(\tau\right) = \frac{N_0}{2}\delta\left(\tau\right)$$

where $\delta\left(\tau\right)$ is the Dirac delta function. The power spectral density is then

$$S_N\left(f\right) = \frac{N_0}{2}$$

N_0 refers to the thermal noise power, which is defined

$$N_0 = kT$$

where $k = 1.380\,649 \times 10^{-23}\,\mathrm{J\,K^{-1}}$ is Boltzmann's constant and T is the temperature in Kelvin. At room temperature, $20°C$, $N_0 = 4 \times 10^{-21}\,\mathrm{W} = -174\,\mathrm{dBm}$.

3.5 Binary Transceiver

In a simple binary transciever, bits are represented by impulses, where each pulse is one bit duration T_b apart. These pulses are then converted into an appropriate waveform for transmission, i.e., as a rectangular, since, raised cosine, Gaussian, etc., pulse shape.
The receiver can then detect these pulses by sampling the received signal at the center of each bit period, where the voltage of the signal is either positive or negative, corresponding to a 1 or 0. However, this fails in the presence of noise, and we must consider other approaches.

3.6 Optimum Receiver

The **correlator** is an optimum receiver for binary signals in an AWGN channel. The correlator performs waveform recovery in preparation for the next step of detection.

$$\int_0^{T_s} s_i\left(t\right) s_i\left(t\right) \mathrm{d}t$$

The correlator integrates the product of the received signal and a replica of the transmitted symbol (called the reference signal) for T_s seconds (symbol period), and dumps the result to the decision circuit by closing the switch every T_s seconds.

$$T_s = T_b \qquad\qquad \text{for binary modulation}$$
$$T_s = T_b \log_2\left(M\right) \qquad\qquad \text{for M-ary modulation}$$

A decision is then made regarding the symbol that was transmitted, based on the test statistic r. As an example, the received signal may have the following form:

$$r\left(t\right) = s_i\left(t\right) + n\left(t\right)$$

where we transmit symbol i. By performing integration with s_j, we obtain

$$r_j(t) = \begin{cases} \displaystyle\int_0^{T_s} s_j^2(t)\,\mathrm{d}t + \int_0^{T_s} n(t)\,s_j(t)\,\mathrm{d}t & j = i \\ \displaystyle\int_0^{T_s} s_i(t)\,s_j(t)\,\mathrm{d}t + \int_0^{T_s} n(t)\,s_j(t)\,\mathrm{d}t & j \neq i \end{cases}$$

$$= \begin{cases} E_j + \hat{n}(t) & j = i \\ \displaystyle\int_0^{T_s} s_i(t)\,s_j(t)\,\mathrm{d}t + \hat{n}(t) & j \neq i \end{cases}$$

where $\hat{n}(t)$ is also a Gaussian random process.

- When the correlator is matched to the transmitted symbol, the output is distributed about the autocorrelation of the transmitted symbol.

- When the correlator is matched to the wrong symbol, the output is distributed about the cross-correlation of the transmitted and reference symbol.

Using the test statistic r, we can then make a decision regarding the transmitted symbol, by comparing r with a particular threshold.

3.7   Matched Filter

The filter corresponding to the corellator is called the matched filter. The matched filter is a linear filter that maximises the SNR at a particular sampling time (typically at the end of the bit period).

$$\left(\frac{S}{N}\right) = \frac{|x(t_0)|^2}{\mathrm{E}\left[N_0^2(t)\right]}$$

where $x(t_0)$ is the signal at the sampling time t_0 and $N_0(t)$ is the noise process. The transfer function that achieves this is given by

$$H(\omega) = \frac{1}{2\pi C}\frac{X^*(\omega)}{S_N(\omega)}e^{-j\omega t_o}$$

where $X^*(\omega)$ is the complex conjugate of the Fourier transform of the signal, $S_N(\omega)$ is the power spectral density of the noise, and C is an arbitrary real constant.
As the noise is assumed to be AWGN, $S_N(\omega) = N_0/2$. The impulse response then becomes

$$h_{\mathrm{opt}}(t) = Kx^*(t_0 - t)$$

where K is a constant, and $x^*(t)$ is the complex conjugate of the signal.
For a symbol shape $s_i(t)$, this tells us that the impulse response of the matched filter is $s_i(T_s - t)$.

3.7.1   Error Performance

The bit error rate is given by the ratio of the number of bit errors to the total number of bits transmitted. The theoretical bit error rate is a function of the bit energy E_b and the noise power

spectral density N_0. For binary signals

$$P_b = \frac{1}{2} \operatorname{erfc} \left(\sqrt{\frac{d_{01}^2}{N_0}} \right)$$

where d_{01} is the Euclidean distance between the two symbols,

$$d_{01}^2 = \int_0^{T_b} (s_0 - s_1)^2 \, \mathrm{d}t$$

4   Information Theory

Information is a measure of the uncertainty of an event.

- An event that is 100% probably is unsurprising and yields no information.

- A less probable event is more surprising and therefore yields more information.

The study of information theory therefore applies to random events.

Definition 4.1 (Alphabet). An alphabet $\mathcal{A}$ is a finite set of M symbols $\{a_1, a_2, ..., a_M\}$ used to represent information from a source. For a source $\{X_n\}_{n=-\infty}^{\infty}$, the alphabet $\mathcal{X} = \{x_1, x_2, ..., x_M\}$ represents the set of all possible values that X_n can take.

Definition 4.2 (Probability of a symbol). The probability of a symbol $x \in \mathcal{X}$ is the probability that the random variable X takes the value x:

$$p(x) = \Pr(X = x).$$

When all M symbols in this alphabet are equiprobable, $p(x) = 1/M$ for all $x \in \mathcal{X}$.

4.1   Measure of Information

The informational content associated with a symbol x is inversely proportional to the probability of the occurrence of that symbol:

$$I(x) = \log_2 \left(\frac{1}{p(x)} \right) = -\log_2 (p(x)).$$

This suggests that an event with low probability carries more information.

Definition 4.3 (Entropy). The entropy of a discrete random variable X is the average information carried by each symbol:

$$H(X) = \mathrm{E}[I(x)] = \mathrm{E}[-\log(p(X))] = -\sum_{i=1}^{M} p(x_i) \log_2 (p(x_i)).$$

The entropy of a source $\{X_n\}_{n=-\infty}^{\infty}$ reflects the average amount of uncertainty that is resolved by using the alphabet.

For a binary source, entropy is measured in bits per symbol (bit/symbol). The bit rate R_b is given by the product of the entropy $H(X)$ and the symbol rate R_s:

$$R_b = H(X) R_s$$
$$\mathrm{bit/s} = \mathrm{bit/symbol} \times \mathrm{symbol/s}$$

4.2   Multiple Sources of Information

Given two discrete random variables X and Y with alphabets $\mathcal{X}$ and $\mathcal{Y}$, the joint probability of $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ is given by

$$p(x, y) = \Pr(X = x \cap Y = y).$$

When X and Y are independent, the joint probability is given by the product of the individual (marginal) probabilities:

$$p(x, y) = p(x)\, p(y).$$

The conditional probability of x given y is given by

$$p(x \mid y) = \frac{p(x, y)}{p(y)} \implies p(x, y) = p(x \mid y)\, p(y)$$

and the conditional probability of y given x is given by

$$p(y \mid x) = \frac{p(x, y)}{p(x)} \implies p(x, y) = p(y \mid x)\, p(x).$$

4.2.1   Joint Entropy

The joint entropy of X and Y is given by

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2(p(x, y)).$$

Joint entropy measures the amount of uncertainty associated with the joint distribution of X and Y.

Theorem 4.2.1 (Chain Rule for Entropy)**.** *The above result can be expressed as*

$$H(X, Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y).$$

This is known as the chain rule for entropy.

Proof. We can prove this result by expanding the joint entropy:

$$
\begin{aligned}
H\left(X,\,Y\right) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p\left(x,\,y\right)\log_2\left(p\left(x,\,y\right)\right) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p\left(x,\,y\right)\log_2\left(p\left(x\right)p\left(y\mid x\right)\right) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p\left(x,\,y\right)\left[\log_2\left(p\left(x\right)\right) + \log_2\left(p\left(y\mid x\right)\right)\right] \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p\left(x,\,y\right)\log_2\left(p\left(x\right)\right) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p\left(x,\,y\right)\log_2\left(p\left(y\mid x\right)\right) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p\left(x\right)p\left(y\mid x\right)\log_2\left(p\left(x\right)\right) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p\left(x\right)p\left(y\mid x\right)\log_2\left(p\left(y\mid x\right)\right) \\
&= -\sum_{x\in\mathcal{X}} p\left(x\right)\log_2\left(p\left(x\right)\right)\sum_{y\in\mathcal{Y}} p\left(y\mid x\right) - \sum_{x\in\mathcal{X}} p\left(x\right)\sum_{y\in\mathcal{Y}} p\left(y\mid x\right)\log_2\left(p\left(y\mid x\right)\right) \\
&= H\left(X\right) + \sum_{x\in\mathcal{X}} p\left(x\right)H\left(Y\mid X\right) \\
&= H\left(X\right) + \mathrm{E}_X\left[H\left(Y\mid X\right)\right] \\
&= H\left(X\right) + H\left(Y\mid X\right)
\end{aligned}
$$

A similar proof can be used to show that $H\left(X,\,Y\right) = H\left(Y\right) + H\left(X\mid Y\right)$. $\square$

4.3 Conditional Entropy

The conditional entropy of Y given X is given by

$$
H\left(Y\mid X\right) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p\left(x,\,y\right)\log_2\left(p\left(y\mid x\right)\right).
$$

Conditional entropy measures the amount of information needed to describe the uncertainty of Y given $X = x$.

Proof. As $H\left(Y\mid X\right)$ is a function of Y, we can write

$$
H\left(Y\mid X\right) = \sum_{x\in\mathcal{X}} p\left(y\mid x\right)\log_2\left(p\left(y\mid x\right)\right).
$$

Additionally, the expected value of $H\left(Y\mid X\right)$ with respect to X is given by

$$
\mathrm{E}_X\left[H\left(Y\mid X\right)\right] = \sum_{x\in\mathcal{X}} p\left(x\right)H\left(Y\mid X\right).
$$

which is equivalent to the definition of $H\left(Y\mid X\right)$. Therefore, by combining the two equations, we

obtain

$$
\begin{aligned}
H\left(Y \mid X\right) &= \mathrm{E}_X\left[H\left(Y \mid X\right)\right] \\
&= \sum_{x \in \mathcal{X}} p\left(x\right) H\left(Y \mid X\right) \\
&= -\sum_{x \in \mathcal{X}} p\left(x\right) \sum_{y \in \mathcal{Y}} p\left(y \mid x\right) \log_2\left(p\left(y \mid x\right)\right) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p\left(x\right) p\left(y \mid x\right) \log_2\left(p\left(y \mid x\right)\right) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p\left(x,\, y\right) \log_2\left(p\left(y \mid x\right)\right)
\end{aligned}
$$

$\square$

4.4   Mutual Information

The mutual information between X and Y is given by

$$
\begin{aligned}
I\left(X,\, Y\right) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p\left(x,\, y\right) \log_2\left(\frac{p\left(x,\, y\right)}{p\left(x\right) p\left(y\right)}\right) \\
&= H\left(X\right) - H\left(X \mid Y\right) \\
&= H\left(Y\right) - H\left(Y \mid X\right)
\end{aligned}
$$

Mutual information measures the amount of information that X and Y share. It is also the reduction in uncertainty of X when Y is known, or the reduction in uncertainty of Y when X is known.

5   Source Coding

Given a source X with n symbols, the minimum number of bits per symbol required to transmit the source is given by the entropy $H\left(X\right)$.

5.1   Source Coding Theorem

Given a source X, let $C : \mathcal{X} \to \mathcal{A}$ be any coding function that maps the source alphabet $\mathcal{X} = \{x_1,\, x_2,\, ...,\, x_M\}$ to the set of sequences $\mathcal{A} = \{a_1,\, a_2,\, ...,\, a_M\}$, i.e.,

$$
C\left(x_i\right) = a_i.
$$

Then, the smallest possible expected code word length $L = |C\left(X\right)|$ is the entropy of X.

$$
\mathrm{E}\left[L\right] = \sum_{x=1}^{M} L_i p_i \geqslant H\left(X\right).
$$

For a sequence of n symbols, the expected code word length is at least $nH\left(X\right)$.

Therefore, for **lossless encoding**, a source with entropy $H(X)$ can be encoded with an arbitrarily small error probability at a rate R bits per symbol, as long as

$$R > H(X).$$

When all symbols in the source alphabet are equiprobable, the entropy is maximised, and the minimum number of bits per symbol is

$$R > \log_2(M).$$

This is used in Pulse Code Modulation (PCM).

5.2   Designing a Code

When designing the optimal code C, the following criteria must be satisfied:

1. **Uniqueness**: each code word must be uniquely decodable, i.e., there must be no ambiguity in decoding a code word from a binary sequence.

2. **Instantaneous**: each code word must be recognisable without the need to look ahead at the next code word in the sequence.

3. **Prefix**: no code word can be a prefix of another code word.

Additionally, the code must be **efficient**, i.e., the average code word length $\mathrm{E}[L]$, must be minimised (close to the entropy of the source). The efficiency of a code is given by

$$\eta = \frac{H(X)}{\mathrm{E}[L]} = \frac{H(X)}{\sum_{i=1}^{M} L_i p_i}.$$

5.3   Shannon-Fano Coding

The Shannon-Fano coding algorithm is an algorithm that constructs a prefix code by recursively splitting the source alphabet into two sets, such that the entropy of each set is almost equal.
The steps for the algorithm are as follows:

1. Order the source alphabet $\mathcal{X} = \{x_1, x_2, ..., x_M\}$ in descending order of probability.

2. Split the alphabet into two sets A and B such that

$$\sum_{x \in A} p(x) \approx \sum_{x \in B} p(x).$$

3. Assign a 0 to all symbols in set A and a 1 to all symbols

4. Repeat steps 2 and 3 for each set until each set contains only one symbol.

5. The code word for each symbol x_i is the sequence of 0s and 1s assigned to that symbol.

This algorithm does not always generate an optimal code however. Instead, we can use the Huffman coding algorithm, which generates codes in reverse, i.e., from the leaf to the root.

5.4   Huffman Coding

The Huffman coding algorithm is an algorithm that constructs a prefix-free code by recursively combining the two least probable symbols into a single symbol, until only one symbol remains. The steps for the algorithm are as follows:

1. Order the source alphabet $\mathcal{X} = \{x_1, x_2, \ldots, x_M\}$ in descending order of probability.

2. Combine the two least probable symbols x_i and x_j into a single symbol x_k, where $k > j > i$.

3. Assign a 0 to x_i and a 1 to x_j.

4. Repeat steps 2 and 3 until only one symbol remains.

5. The code word for each symbol x_i is the sequence of 0s and 1s assigned to that symbol.

In this algorithm, step 3 may be performed in multiple ways, for example,

- x_i is assigned 1 and x_j is assigned 0

- x_i is assigned 0 and x_j is assigned 1

- the assignment changes at each iteration

and therefore the Huffman algorithm is not unique. Additionally, when the least probable symbol x_i has a probability p_i, and the next least probable symbol x_j shares a probability p_j with another symbol, the algorithm is ambiguous.

5.5   Variance of Code Word Length

When two codes share the same average code word length, the code with the smaller variance of code length is preferred. The variance of a code length is given by

$$\sigma^2 = \sum_{i=1}^{M} \left(L_i - \mathrm{E}\left[L\right] \right)^2 p_i.$$

A high variance typically indicates that the number of bits per symbol is high. Therefore, it is beneficial to minimise the variance of a code word length.

6   Digital Passband Modulation

Similar to the baseband case, we can use digital modulation to transmit information over a passband channel. For bandpass modulation, the bandwidth of the transmitted signal is in general $B_{\mathrm{Passband}} = 2B_{\mathrm{Baseband}}$. Therefore, in a passband channel,

- $B = 2R_s$ for rectangular pulses

- $B = R_s$ for Nyquist (sinc) pulses

- $B = R_s\left(1 + R\right)$ for raised cosine pulses

A digital signal can be modulated by varying the

- amplitude

- frequency

- phase

- combination of the above

of a sinusoidal carrier wave. If the modulating wave consists of NRZ pulses, the modulated parameter will be switched or **keyed** from one value to another. This is known as **keying**. There are three types of keying:

- **Amplitude Shift Keying (ASK)**: the amplitude of the carrier wave is switched between two values. For example,

    - 0: A_0
    - 1: A_1

- **Frequency Shift Keying (FSK)**: the frequency of the carrier wave is switched between two values with continuous phase. For example,

    - 0: f_0
    - 1: f_1

- **Phase Shift Keying (PSK)**: the phase of the carrier wave is switched between two values. For example,

    - 0: ϕ_0
    - 1: ϕ_1

Quadrature Amplitude Modulation (QAM) is a combination of ASK and PSK.

6.1 Representing the Carrier Wave

The carrier wave is represented by

$$s\left(t\right) = A\left(t\right)\cos\left(2\pi f_c t + \phi\left(t\right)\right)$$

By using the RMS value of this wave, $A_{\mathrm{rms}} = A/\sqrt{2}$, we can say

$$s\left(t\right) = \sqrt{2}A_{\mathrm{rms}}\cos\left(2\pi f_c t + \phi\left(t\right)\right)$$

Assuming a $1\,\Omega$ resistive load, the power of the carrier wave is given by $P = E_s/T_s = V_{\mathrm{rms}}^2/R = A_{\mathrm{rms}}^2$, where E_s is the energy per symbol, and T_s is the symbol period, then

$$s\left(t\right) = \sqrt{\frac{2E_s}{T_s}}\cos\left(2\pi f_c t + \phi\left(t\right)\right).$$

This can be rewritten as

$$s\left(t\right) = \sqrt{\frac{2E_s}{T_s}} \cos\left(2\pi f_c t + \phi\left(t\right)\right)$$

$$= \sqrt{E_s}\sqrt{2R_s}\left[\cos\left(2\pi f_c t\right)\cos\left(\phi\left(t\right)\right) - \sin\left(2\pi f_c t\right)\sin\left(\phi\left(t\right)\right)\right]$$

$$= \underbrace{\sqrt{E_s}\cos\left(\phi\left(t\right)\right)}_{a}\underbrace{\sqrt{2R_s}\cos\left(2\pi f_c t\right)}_{\Psi_1\left(t\right)} - \underbrace{\sqrt{E_s}\sin\left(\phi\left(t\right)\right)}_{b}\underbrace{\sqrt{2R_s}\sin\left(2\pi f_c t\right)}_{\Psi_2\left(t\right)}$$

6.2   Signal Space Representation

In the representation:

$$s\left(t\right) = a\Psi_1\left(t\right) - b\Psi_2\left(t\right),$$

the functions $\Psi_1\left(t\right)$ and $\Psi_2\left(t\right)$ form an orthonormal basis in the **signal space** so that the following inner products over a symbol periods hold. The inner product of the two basis functions is 0:

$$\left\langle \Psi_1\left(t\right),\ \Psi_2\left(t\right)\right\rangle = \int_0^{T_s} \Psi_1\left(t\right)\Psi_2\left(t\right)\mathrm{d}t = 0,$$

and the norm of each basis function is 1:

$$\left\langle \Psi_i\left(t\right),\ \Psi_i\left(t\right)\right\rangle = \int_0^{T_s} \Psi_i^2\left(t\right)\mathrm{d}t = 1.$$

These functions can be represented on the I-Q plane, where I is the **in-phase** component and Q is the **quadrature phase** component of the signal. On this plane, the Cartesian coordinates of s represent:

- The in-phase component $I = a$

- The quadrature phase component $Q = b$

which are obtained by correlating s with the basis functions Ψ_1 and Ψ_2, respectively:

$$a = \left\langle s\left(t\right),\ \Psi_1\left(t\right)\right\rangle = \int_0^{T_s} s\left(t\right)\Psi_1\left(t\right)\mathrm{d}t$$

$$b = -\left\langle s\left(t\right),\ \Psi_2\left(t\right)\right\rangle = -\int_0^{T_s} s\left(t\right)\Psi_2\left(t\right)\mathrm{d}t$$

and in polar coordinates,

- The **amplitude** $\sqrt{a^2 + b^2} = \sqrt{E_s}$ measures the transmitted signal energy during a symbol period.

- The **angle** ϕ measures the relation of the signal to another.

Using this representation, each symbol s_i is expressed as a vector of its coordinates:

$$s_i = \left[a_i,\ b_i\right]$$

29

6.3   Signal Constellation

The set of all possible signal vectors s is called the **signal constellation** of the modulation scheme. This set is used to represent the information transmitted by the modulated signal.

6.4   Binary Modulation Schemes

6.4.1   Binary Amplitude Shift Keying (BASK)

In BASK, the amplitude of the carrier wave is switched between two values. One such technique is to use **On Off Keying** (OOK). The signal constellation for BASK is given below

$$s_1 = [0,\ 0]$$
$$s_2 = \left[\sqrt{E_b},\ 0\right]$$

6.4.2   Binary Phase Shift Keying (BPSK)

In BPSK, the phase of the carrier wave is switched between two values. The signal constellation for BPSK is given below

$$s_1 = \left[\sqrt{E_b},\ 0\right]$$
$$s_2 = \left[-\sqrt{E_b},\ 0\right]$$

6.4.3   Binary Frequency Shift Keying (BFSK)

In BFSK, the frequency of the carrier wave is switched between two values f_1 and f_2. The signal constellation for BFSK is given below

$$s_1 = \left[\sqrt{E_b},\ 0\right]$$
$$s_2 = \left[\sqrt{E_b},\ 0\right]$$

where the in-phase component has a different frequency for each symbol.

6.4.4   Summary of Binary Modulation Schemes

Name	E_s	ϕ	Signal Constellation	
BASK	$(0,\ E_b)$	$(0,\ 0)$	$s_1\left(t\right) = 0\cos\left(0\right)\sqrt{2R_b}\cos\left(2\pi f_c t\right)$	$= 0$
			$s_2\left(t\right) = \sqrt{E_b}\cos\left(0\right)\sqrt{2R_b}\cos\left(2\pi f_c t\right)$	$= \sqrt{2E_b R_b}\cos\left(2\pi f_c t\right)$
BPSK	$(E_b,\ E_b)$	$(0,\ \pi)$	$s_1\left(t\right) = \sqrt{E_b}\cos\left(0\right)\sqrt{2R_b}\cos\left(2\pi f_c t\right)$	$= \sqrt{2E_b R_b}\cos\left(2\pi f_c t\right)$
			$s_2\left(t\right) = \sqrt{E_b}\cos\left(\pi\right)\sqrt{2R_b}\cos\left(2\pi f_c t\right)$	$= -\sqrt{2E_b R_b}\cos\left(2\pi f_c t\right)$
BFSK	$(E_b,\ E_b)$	$(0,\ 0)$	$s_1\left(t\right) = \sqrt{E_b}\cos\left(0\right)\sqrt{2R_b}\cos\left(2\pi f_1 t\right)$	$= \sqrt{2E_b R_b}\cos\left(2\pi f_1 t\right)$
			$s_2\left(t\right) = \sqrt{E_b}\cos\left(0\right)\sqrt{2R_b}\cos\left(2\pi f_2 t\right)$	$= \sqrt{2E_b R_b}\cos\left(2\pi f_2 t\right)$

6.5   M-ary Modulation Schemes

6.5.1   M-ary Amplitude Shift Keying (M-ASK)

In M-ASK, the amplitude of the carrier wave is switched between M values. The signal constellation for M-ASK is given below

$$s_i = \left[\left(2i - 1 - M\right)\sqrt{E_i},\ 0\right]$$

where E_i is the energy per symbol given by

$$\sqrt{E_i} = \left(2i - 1 - M\right)\sqrt{E_0}.$$

where E_0 is the energy of the symbol closest to the origin. The average energy per symbol is given by

$$E_s = \frac{M^2 - 1}{3}E_0$$

The error performance of M-ASK is given by

$$P_b = \frac{2\left(M - 1\right)}{M\log_2\left(M\right)}Q\left(\sqrt{\frac{6\log_2\left(M\right)}{M^2 - 1}\left(\frac{E_b}{N_0}\right)}\right)$$

6.5.2   M-ary Phase Shift Keying (M-PSK)

In M-PSK, the phase of the carrier wave is switched between M values. The signal constellation for M-PSK is given below

$$s_i = \left[\sqrt{E_s}\cos\left(\frac{2\pi}{M}i\right),\ \sqrt{E_s}\sin\left(\frac{2\pi}{M}i\right)\right]$$

where E_s is the energy per symbol, which is constant for all symbols. The error performance of M-PSK is given by

$$P_b = \frac{1}{\log_2{(M)}} Q \left(\sqrt{\log_2{(M)} \sin^2{\left(\frac{\pi}{M} \right)} \frac{2E_b}{N_0}} \right)$$

6.5.3 Quadrature Phase Shift Keying (QPSK)

In QPSK, the phase of the carrier wave is switched between four values. The signal constellation for QPSK is given below

$$s_i = \left[\sqrt{E_s} \cos{\left(\frac{\pi}{4} (2i - 1) \right)}, \ \sqrt{E_s} \sin{\left(\frac{\pi}{4} (2i - 1) \right)} \right]$$

where E_s is the energy per symbol, which is constant for all symbols. As opposite symbols incur a phase shift of π, alternative modulation schemes are used to reduce the phase shift to $\pi/2$.
One such technique is **Offset QPSK** (OQPSK), where the quadrature component of a signal is offset by T_b to ensure only one bit change at any time. $\pi/4$-QPSK is another modulation scheme where the phase difference varies for each symbol. This allows symbols to be mapped to two square constellations, which are rotated by $\pi/4$. The phase difference is given by

$$\phi_k = \phi_{k-1} + \theta_k$$

6.5.4 M-ary Quadrature Amplitude Modulation (M-QAM)

In M-QAM, both the amplitude and phase of the carrier wave are varied. The general expression for the signal constellation of M-QAM is given by

$$s_i = \left[\sqrt{E_i} \cos{(\phi_i)}, \ \sqrt{E_i} \sin{(\phi_i)} \right]$$

7 Radio Frequency (RF) Systems

7.1 S-Parameters

Two port networks consist of:

- **Input port**: port 1

- **Output port**: port 2

with a voltage and current at each port:

- V_1 — voltage across port 1

- I_1 — current into port 1

- V_2 — voltage across port 2

- I_2 — current into port 2

Such networks can be represented using various parameters:

- z-**parameters**: open-circuit impedance

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \end{bmatrix}$$

  where $z_{ij} = V_i/I_j$ is measured in Ohms (Ω).

- y-**parameters**: short-circuit admittance

$$\begin{bmatrix} I_1 \\ I_2 \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$$

  where $y_{ij} = I_i/V_j$ is measured in Siemens (S).

- $ABCD$-**parameters**: chain or transmission parameters

$$\begin{bmatrix} V_1 \\ I_2 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} V_2 \\ -I_1 \end{bmatrix}$$

  where $AD - BC = 1$.

- h-**parameters**: hybrid parameters

$$\begin{bmatrix} V_1 \\ I_2 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} I_1 \\ V_2 \end{bmatrix}$$

- g-**parameters**: inverse hybrid parameters

$$\begin{bmatrix} I_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} \begin{bmatrix} V_1 \\ I_2 \end{bmatrix}$$

7.2   Scattering Parameters

Scattering parameters are defined in terms of incident and reflected waves at ports, as it is often difficult to measure voltages and currents directly at UHF and microwave frequencies. The scattering matrix **S** is defined

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

where a_i are the incident wave amplitudes and b_i are the reflected wave amplitudes, defined:

$$a_i = \frac{V_i^+}{\sqrt{Z_{0i}}} = \frac{1}{2} \left(\frac{V_i}{\sqrt{Z_{0i}}} + \sqrt{Z_{0i}} I_i \right)$$

$$b_i = \frac{V_i^-}{\sqrt{Z_{0i}}} = \frac{1}{2} \left(\frac{V_i}{\sqrt{Z_{0i}}} - \sqrt{Z_{0i}} I_i \right)$$

where V_i^+ and V_i^- are the incident and reflected voltages, respectively, and Z_{0i} is the reference impedance of port i. Typically this is assumed to be $50\,\Omega$ for every port.

The voltage V_i across port i is given by

$$V_i = \sqrt{Z_{0i}}\,(a_i + b_i)$$

and the current I_i into port i is given by

$$I_i = \frac{1}{\sqrt{Z_{0i}}}\,(a_i - b_i)\,.$$

The average power can be calculated using

$$P_i = \frac{1}{2}\Re\left\{V_i I_i^*\right\} = \frac{1}{2}\left(|a_i|^2 - |b_i|^2\right)$$

These four parameters have the following physical interpretations:

- S_{11} — voltage reflection coefficient at port 1

- S_{12} — voltage transmission coefficient from port 2 to port 1

- S_{21} — voltage transmission coefficient from port 1 to port 2

- S_{22} — voltage reflection coefficient at port 2

The voltage reflection coefficient is given by

$$S_{11} = \Gamma_{\text{in}} = \frac{b_1}{a_1} = \frac{Z_1 - Z_0}{Z_1 + Z_0}$$
$$S_{22} = \Gamma_{\text{out}} = \frac{b_2}{a_2} = \frac{Z_2 - Z_0}{Z_2 + Z_0}$$

7.3 Oscillators

Oscillators are used to generate a sinusoidal signal at a desired frequency. Oscillators are generated using a gain device such as an amplifier or transistor, and a resonator. An example is the LC network, which consists of an amplifier with a feedback path consisting of a capacitor and inductor in series. Under proper conditions, this device will be unstable, and resonate. The particular frequency at which the resontator oscillates is called the **resonant frequency**. For the LC network, this is

$$\omega_0 = \frac{1}{\sqrt{LC}}.$$

Ideally, a perfect resonator will resonate precisely at ω_0, however, in practice, the oscillating frequency is spread around the resonant frequency. By performing **spectral analysis** (power vs. frequency) on the output of the oscillator, we can see that energy is spread over a range of frequencies centered around ω_0.
The **bandwidth** of the output spectrum relates to the **quality** of the resonator.

- A **high-Q** resonator has a narrow bandwidth, and therefore a high quality factor.

- A **low-Q** resonator has a wide bandwidth, and therefore a low quality factor.

Generally, low-Q resonators are **lossy** and high-Q resonators exhibit **low loss**. LC networks are low-Q.

7.3.1 Resonators

Some examples of high-Q resonators include:

- **Crystal resonators**: dependent on geometry and lattice structure

- **Dielectric resonators**: dependent on cavity geometry and dielectric constant. Can be made very small

- **Transmission line resonators**: made using transmission line sections, (stripline or microstrip)

7.3.2 Oscillator Tuning

The above resonators have a fixed frequency, and therefore cannot be tuned. To tune an oscillator, we can use a **varactor diode**. This is a diode with a capacitance that varies with the applied voltage. The resonant frequency is given by

$$\omega_0 = \frac{1}{\sqrt{LC\left(V_D\right)}}.$$

7.3.3 Harmonics and Spurs

Oscillators also produce other frequencies not at the resonant frequency. These occur at integer multiples of ω_0 called **harmonics** or aribtrarily at other frequencies called spurious frequencies (**spurs**).

As the power of these frequencies is typically much lower than the resonant frequency, we often measure them in **dBc** (decibels relative to the carrier power P_c).

$$P(\text{dBc}) = 10\log_{10}\left(\frac{P}{P_c}\right)$$

7.3.4 Phase Noise

Consider the signal

$$\cos\left(\phi\left(t\right)\right) = \cos\left(\omega_0 t + \phi_0\right)$$

where $\phi\left(t\right) = \omega_0 t + \phi_0$ is the total or absolute phase of the signal. Note that this is a linearly increasing function of time. If we define the relative phase as $\phi_r\left(t\right) = \phi\left(t\right) - \omega_0 t$, then

$$\cos\left(\phi\left(t\right)\right) = \cos\left(\omega_0 t + \phi_r\left(t\right)\right).$$

This allows us to define the **instantaneous frequency** as

$$\omega\left(t\right) = \frac{\mathrm{d}\phi\left(t\right)}{\mathrm{d}t} = \omega_0 + \omega_r\left(t\right).$$

Therefore the frequency of the signal changes over time. If $\phi_r\left(t\right)$ is a random process, then the frequency of the signal is also a random process. This is called **phase noise**.

The spectrum of this noise can be quantified using its average spectral power density measured in W/Hz.

- For white noise, the spectral power density is constant with respect to frequency.

- For an oscillator, the spectral power density changes with respect to frequency. The average spectral power density increases as ω approaches ω_0.

Phase noise is also measured using **dBc**, where power is measured in a 1Hz bandwidth, for dimensions to be consistent.

7.3.5   Frequency Pulling and Pushing

The output of an oscillator is always attached to some load impedance Γ_L. The impedance of this load can affect the operating frequency of the oscillator. This is called **frequency pulling**. Therefore an oscillator is designed assuming that the load is matched, i.e., $\Gamma_L = 0$. Frequency pulling is specified as the maximum deviation from the nominal frequency given some worst case load.
Frequency pulling can be minimsed by isolating the oscillator from the load, using an **isolation amplifier**.
Frequency pushing is similar to frequency pulling, where the supply voltage causes the operating frequency to change. Frequency pushing can be reduced by using a high-Q resonator, and by regulating the supply voltage.

7.4   Mixers

A mixer is a device that multiplies two signals to produce a third signal. It is a three-port, non-linear device, where the ports are referred to as:

- **RF port**: the radio frequency input

- **LO port**: the local oscillator input

- **IF port**: the intermediate frequency output

Ideally, the output of a mixer is given by

$$
\begin{aligned}
v_{\text{IF}} &= v_{\text{RF}} v_{\text{LO}} \\
&= \cos\left(\omega_{\text{RF}} t\right) \cos\left(\omega_{\text{LO}} t\right) \\
&= \frac{1}{2} \cos\left(\left(\omega_{\text{RF}} + \omega_{\text{LO}}\right) t\right) + \frac{1}{2} \cos\left(\left(\omega_{\text{RF}} - \omega_{\text{LO}}\right) t\right)
\end{aligned}
$$

so that the output band is located at the sum and difference of the RF and LO frequencies.

7.4.1   Diode Mixers

A non-ideal mixer can be implemented as the sum of the RF and LO voltages across a non-linear device. This is called a **diode mixer**. The current across a diode is given by

$$
i_D\left(t\right) = i_s\left(A e^{v_D(t)} - 1\right)
$$

for some constant A. If we consider the Taylor series expansion of the exponential term and group all constants, this becomes

$$
i_D\left(t\right) = a_1 v_D\left(t\right) + a_2 v_D^2\left(t\right) + a_3 v_D^3\left(t\right) + \ldots
$$

As the voltage across the diode is the sum of the RF and LO voltages, we can write

$$v_D(t) = v_{\mathrm{RF}}(t) + v_{\mathrm{LO}}(t)$$

so that

$$
\begin{aligned}
i_D(t) = a_1 \left(v_{\mathrm{RF}}(t) + v_{\mathrm{LO}}(t) \right) + \\
a_2 \left(v_{\mathrm{RF}}^2(t) + 2 v_{\mathrm{RF}}(t) v_{\mathrm{LO}}(t) + v_{\mathrm{LO}}^2(t) \right) + \\
a_3 \left(v_{\mathrm{RF}}^3(t) + 3 v_{\mathrm{RF}}^2(t) v_{\mathrm{LO}}(t) + 3 v_{\mathrm{RF}}(t) v_{\mathrm{LO}}^2(t) + v_{\mathrm{LO}}^3(t) \right)
\end{aligned}
$$

By considering the output from each coefficient, we can can identify the freqencies produced at the output of the mixer.

- 1st order terms:

$$v_{\mathrm{RF}}(t) + v_{\mathrm{LO}}(t) = \cos(\omega_{\mathrm{RF}}t) + \cos(\omega_{\mathrm{LO}}t)$$

  produce frequencies at ω_{RF} and ω_{LO}.

- 2nd order terms:

$$
\begin{aligned}
v_{\mathrm{RF}}^2(t) + 2 v_{\mathrm{RF}}(t) v_{\mathrm{LO}}(t) + v_{\mathrm{LO}}^3(t) = \\
\frac{1}{2} \left[\cos(2\omega_{\mathrm{RF}}t) + \cos((\omega_{\mathrm{RF}} - \omega_{\mathrm{LO}})t) + \cos((\omega_{\mathrm{RF}} + \omega_{\mathrm{LO}})t) + \cos(2\omega_{\mathrm{LO}}t) \right]
\end{aligned}
$$

  produce frequencies at $2\omega_{\mathrm{RF}}$, $|\omega_{\mathrm{RF}} - \omega_{\mathrm{LO}}|$, $\omega_{\mathrm{RF}} + \omega_{\mathrm{LO}}$, and $2\omega_{\mathrm{LO}}$.

- 3rd order terms:

$$
\begin{aligned}
v_{\mathrm{RF}}^3(t) + 3 v_{\mathrm{RF}}^2(t) v_{\mathrm{LO}}(t) + 3 v_{\mathrm{RF}}(t) v_{\mathrm{LO}}^2(t) + v_{\mathrm{LO}}^3(t) = \\
\frac{1}{4} \left[3\cos(\omega_{\mathrm{RF}}t) + \cos(3\omega_{\mathrm{RF}}t) + \right. \\
6\cos(\omega_{\mathrm{LO}}t) + 3\cos((2\omega_{\mathrm{RF}} - \omega_{\mathrm{LO}})t) + 3\cos((2\omega_{\mathrm{RF}} + \omega_{\mathrm{LO}})t) + \\
6\cos(\omega_{\mathrm{RF}}t) + 3\cos((2\omega_{\mathrm{LO}} - \omega_{\mathrm{RF}})t) + 3\cos((2\omega_{\mathrm{LO}} + \omega_{\mathrm{RF}})t) + \\
\left. 3\cos(\omega_{\mathrm{LO}}t) + \cos(3\omega_{\mathrm{LO}}t) \right]
\end{aligned}
$$

  produce frequencies at $3\omega_{\mathrm{RF}}$, $|2\omega_{\mathrm{RF}} - \omega_{\mathrm{LO}}|$, $2\omega_{\mathrm{RF}} + \omega_{\mathrm{LO}}$, $|2\omega_{\mathrm{LO}} - \omega_{\mathrm{RF}}|$, $2\omega_{\mathrm{LO}} + \omega_{\mathrm{RF}}$, and $3\omega_{\mathrm{LO}}$.

The ideal mixer output sets all terms to zero, except:

$$v_{\mathrm{IF}}(t) = 2a_2 v_{\mathrm{RF}}(t) v_{\mathrm{LO}}(t)$$

which produces frequencies at $|\omega_{\mathrm{RF}} - \omega_{\mathrm{LO}}|$ and $\omega_{\mathrm{RF}} + \omega_{\mathrm{LO}}$.

7.5   RF Amplifiers

Low noise amplifiers (LNAs) are used to amplify weak signals from antennas. They usually consist of a single stage RF transistor amplifier. The purpose of an LNA is to improve the system noise temperature by

- providing useful gain

- having a very low noise figure

As this amplifier receives a high frequency signal, it requires impedance matching at the input and output stages of the RF transitor (to obtain the minimal noise figure), and therefore high gain is usually not achieved.

7.5.1   Impedance Matching

To maximise the power transfer from a source to a load, it is important to match the impedance of the source to the impedance of the load. This corresponds to achieving $\Gamma = 0$, so that no power is reflected back to the source. This occurs precisely when $Z_L = Z_0$.

When this is not possible, we can use a matching network to achieve impedance matching. This is a network of capacitors and inductors that modify the impedance of the load. Note that using resistors will reduce the efficiency of the network, as power will be dissipated.

7.5.2   LNA Noise Figure

Several noise parameters are published to characterise the optimum noise figure operating conditions. These include:

- $F_{\min}$: the minimum noise figure when the source admittance $Y_s = Y_{\mathrm{opt}}$

- Γ_{opt}: the optimum reflection coefficient of the source impedance for minimum noise figure $F_{\min}$

- R_N: the equivalent noise resistance of the transistor

7.5.3   Dynamic Range

The dynamic range of an amplifier is the range of input power levels over which the amplifier can operate. This sits above a **noise floor** which is the minimum input power level that can be detected by the amplifier. This noise floor is dependent on the operating temperature of the amplifier and the construction of the amplifier. The upper limit of the dynamic range is the **1dB compression point** at which the amplifier output power is 1dB below the typical linear output gain. The amplifer enters the saturation region beyond this point.

7.5.4   Amplifier Non-Linearity

Ideally amplifiers should provide the same gain for a wide range of frequencies, so that in the ideal case,

$$v_{\mathrm{out}} = A v_{\mathrm{in}}$$

however, actual amplifiers will contain several non-linear terms, which can be described by a Taylor series expansion:

$$v_{\mathrm{out}} = A v_{\mathrm{in}} + B v_{\mathrm{in}}^2 + C v_{\mathrm{in}}^3 + ...$$

Generally it is possible to design an amplifier with small B and C, such that

$$v_{\mathrm{out}} \approx A v_{\mathrm{in}}$$

however, as v_{in} becomes large, v_{in}^2 and v_{in}^3 become significantly large, and the terms Bv_{in}^2 and Cv_{in}^3 become significant, which results in the output being distorted, rather than amplified. This phenomenon is known as **intermodulation distortion**.

7.5.5 Harmonics

In addition to this, the amplifier will also produce harmonics of the input signal, which is known as **harmonic distortion**. If we assume the input signal to be sinusoidal with amplitude a, then

$$
\begin{aligned}
v_{\text{out}} &= Av_{\text{in}} + Bv_{\text{in}}^2 + Cv_{\text{in}}^3 + ... \\
&= Aa\cos(\omega t) + Ba^2\cos^2(\omega t) + Ca^3\cos^3(\omega t) + ... \\
&= Aa\cos(\omega t) + \frac{1}{2}Ba^2\left(1 + \cos(2\omega t)\right) + \frac{1}{4}Ca^3\left(3\cos(\omega t) + \cos(3\omega t)\right) + ...
\end{aligned}
$$

so that the second order term produces a harmonic at 2ω,

$$
\frac{Ba^2}{2}\cos(2\omega t)
$$

and the third order term produces harmonics at 3ω,

$$
\frac{Ca^3}{4}\cos(3\omega t).
$$

The **harmonic power** of this amplifier is defined as,

- 1st order:
$$
P_1^{\text{out}} = A^2 P_{\text{in}} = G_1 P_{\text{in}}
$$

- 2nd order:
$$
P_2^{\text{out}} = \frac{B^2}{4} P_{\text{in}}^2 = G_2 P_{\text{in}}^2
$$

- 3rd order:
$$
P_3^{\text{out}} = \frac{C^2}{16} P_{\text{in}}^3 = G_3 P_{\text{in}}^3
$$

where G_i is the gain of the ith order harmonic. Expressed in decibels, this becomes

$$
\begin{aligned}
P_1^{\text{out}}(\text{dBm}) &= G_1(\text{dB}) + P_{\text{in}}(\text{dBm}) \\
P_2^{\text{out}}(\text{dBm}) &= G_2(\text{dB}) + 2P_{\text{in}}(\text{dBm}) \\
P_3^{\text{out}}(\text{dBm}) &= G_3(\text{dB}) + 3P_{\text{in}}(\text{dBm})
\end{aligned}
$$

Plotting each of these signals on a power input to power output graph, we can see the effect of increasing the input power on the output of the higher order harmonics. The output power where the output of the second harmonic intercepts the output power of the first harmonic is known as the **second-order intercept point**,

$$
P_2^{\text{int}} := P_1^{\text{out}} = P_2^{\text{out}}
$$

this occurs when

$$G_1 + P_\text{in} = G_2 + 2P_\text{in} \implies P_\text{in} = G_1 - G_2 \implies P_2^\text{int} = 2G_1 - G_2$$

Similarly, the output power where the output of the third harmonic intercepts the output power of the first harmonic is known as the **third-order intercept point**,

$$P_3^\text{int} := P_1^\text{out} = P_3^\text{out}$$

this occurs when

$$G_1 + P_\text{in} = G_3 + 3P_\text{in} \implies P_\text{in} = \frac{G_1 - G_3}{2} \implies P_3^\text{int} = \frac{3G_1 - G_3}{2}$$

When choosing an amplifier, a higher intercept point is desirable, as this means that the amplifier can handle higher input powers before harmonic distortion becomes significant. Note that decreasing G_2 and G_3 increases the intercept points.

7.5.6 Intermodulation Distortion

Typically only the third order intercept is important to consider, as it plays a significant role in two-tone intermodulation distortion. Consider the input signal with two tones,

$$v_\text{in} = a \cos(\omega_1 t) + a \cos(\omega_2 t)$$

The second order harmonic is given by

$$B\left[\frac{a^2}{2} \cos(2\omega_1 t) + a^2 \cos((\omega_1 - \omega_2) t) + a^2 \cos((\omega_1 + \omega_2) t) + a^2 + \frac{a^2}{2} \cos(2\omega_2 t) \right]$$

and the third order harmonic is given by

$$\begin{aligned} C\Bigg[&\frac{3}{4} \cos(\omega_1 t) + \frac{1}{4} \cos(3\omega_1 t) + \\ &\frac{6}{4} \cos(\omega_2 t) + \frac{3}{4} \cos((2\omega_1 - \omega_2) t) + \frac{3}{4} \cos((2\omega_1 + \omega_2) t) + \\ &\frac{6}{4} \cos(\omega_1 t) + \frac{3}{4} \cos((2\omega_2 - \omega_1) t) + \frac{3}{4} \cos((2\omega_2 + \omega_1) t) + \\ &\frac{3}{4} \cos(\omega_2 t) + \frac{1}{4} \cos(3\omega_2 t) \Bigg] \end{aligned}$$

When $\omega_1 \approx \omega_2$, the third order harmonic produces frequencies at $2\omega_1 - \omega_2 \approx \omega_1$ and $2\omega_2 - \omega_1 \approx \omega_2$, which are close to the input frequencies. During the filtering stage it is nearly impossible to filter out these frequencies without also filtering out the input frequencies. Hence, we must choose an amplifier with a high third order intercept point, so that the effect of these harmonics is minimised.

7.6 RF Filters

RF (microwave) filters are typically a passive, reciprocal, 2-port linear device used to remove unwanted frequencies from a signal of interest. If port 2 is terminated with a matched load, then the incident (input) and output power are related by

$$P_\text{out} = |S_{21}|^2 P_\text{in}.$$

Here we can define the **power transmission coefficient** T as

$$T = \frac{P_{\text{out}}}{P_{\text{in}}} = |S_{21}|^2$$

where for a passive filter, $0 \leqslant T \leqslant 1$, or $P_{\text{out}} \leqslant P_{\text{in}}$.

The power reduction of a filter can either be absorbed by the filter P_{abs} (dissipated as heat), or reflected back to the source (P_r). Then by the conservation of energy, we have

$$P_{\text{in}} = P_{\text{out}} + P_{\text{abs}} + P_r$$

Ideally, microwave filters are lossless, so that $P_{\text{abs}} = 0$, and

$$P_{\text{in}} = P_{\text{out}} + P_r \implies 1 = \frac{P_{\text{out}}}{P_{\text{in}}} + \frac{P_r}{P_{\text{in}}} = T + R$$

where R is the **power reflection coefficient** given by

$$R = \frac{P_r}{P_{\text{in}}} = |S_{11}|^2.$$

This tells us that

$$1 = |S_{21}|^2 + |S_{11}|^2$$

For a microwave filter, R and T are functions of ω, so that for some signal frequency ω_s, we have

$$R(\omega_s) \approx 0 \qquad \text{and} \qquad T(\omega_s) \approx 1$$

where ω_s lies in the **passband** and all frequencies **pass through** the filter. Or

$$R(\omega_s) \approx 1 \qquad \text{and} \qquad T(\omega_s) \approx 0$$

where ω_s lies in the **stopband** and all frequencies are **reflected** at the input of the filter.

7.6.1 Low Pass Filters

A low pass filter is a filter that passes low frequencies and rejects high frequencies. For this filter,

$$T(\omega) \approx \begin{cases} 0 & \omega > \omega_c \\ 1 & \omega < \omega_c \end{cases} \qquad \text{and} \qquad R(\omega) \approx \begin{cases} 1 & \omega > \omega_c \\ 0 & \omega < \omega_c \end{cases}$$

The cut-off frequency ω_c is defined as the frequency where $T = \frac{1}{2}$. For a lossless filter, this also implies $R = \frac{1}{2}$.

7.6.2 High Pass Filters

A high pass filter is a filter that passes high frequencies and rejects low frequencies. For this filter,

$$T(\omega) \approx \begin{cases} 1 & \omega > \omega_c \\ 0 & \omega < \omega_c \end{cases} \qquad \text{and} \qquad R(\omega) \approx \begin{cases} 0 & \omega > \omega_c \\ 1 & \omega < \omega_c \end{cases}$$

The cut-off frequency ω_c is defined as the frequency where $T = \frac{1}{2}$. For a lossless filter, this also implies $R = \frac{1}{2}$.

7.6.3 Band Pass Filters

A band pass filter is a filter that passes frequencies within a certain bandwidth $\Delta\omega$ and rejects frequencies outside this bandwidth. For this filter,

$$T(\omega) \approx \begin{cases} 1 & |\omega - \omega_0| < \frac{\Delta\omega}{2} \\ 0 & |\omega - \omega_0| > \frac{\Delta\omega}{2} \end{cases} \quad \text{and} \quad R(\omega) \approx \begin{cases} 0 & |\omega - \omega_0| < \frac{\Delta\omega}{2} \\ 1 & |\omega - \omega_0| > \frac{\Delta\omega}{2} \end{cases}$$

where ω_0 is the centre frequency of the filter.

7.6.4 Band Stop Filters

A band stop filter is a filter that rejects frequencies within a certain bandwidth $\Delta\omega$ and passes frequencies outside this bandwidth. For this filter,

$$T(\omega) \approx \begin{cases} 0 & |\omega - \omega_0| < \frac{\Delta\omega}{2} \\ 1 & |\omega - \omega_0| > \frac{\Delta\omega}{2} \end{cases} \quad \text{and} \quad R(\omega) \approx \begin{cases} 1 & |\omega - \omega_0| < \frac{\Delta\omega}{2} \\ 0 & |\omega - \omega_0| > \frac{\Delta\omega}{2} \end{cases}$$

where ω_0 is the centre frequency of the filter.

7.6.5 Filter Bandwidth

The vast majority of microwave filters are bandpass filters. These filters are characterised by their **percentage bandwidth** defined as

$$\%\text{BW} = \frac{\Delta f}{f_0} \times 100\%.$$

where Δf is the bandwidth and f_0 is the centre frequency.
Bandpass filters are classified as either narrowband or wideband filters, depending on either the percentage bandwidth or the technology used to construct the filters.

- **Percentage bandwidth**: A percentage bandwidth less than 10%, is considered a **narrowband** filter.

- **Technology used to construct the filter**: If the filter is constructed using lumped and/or distributed elements (e.g., inductors, capacitors, transmission lines), it is considered a **narrowband** filter. Note that it is difficult to achieve small percentage bandwidths (less than 3%) using these elements.

For very narrowband filters, we must use high-Q resonators such as dielectric cavities, crystals, or SAWs, though achieving a percentage bandwidth of less than 0.1% is difficult.

7.6.6 Filter Phase Function

The phase function of a filter is defined as the phase of the scattering parameter S_{21} as a function of frequency,

$$\angle S_{21} = \arctan\left(\frac{\Im(S_{21})}{\Re(S_{21})}\right).$$

42

Given an incident wave $V_1^+ (z_1) = V_{01}^+ e^{-j\beta z}$, the output wave is given by

$$
\begin{aligned}
V_2^- (z_2) &= V_{02}^- e^{j\beta z_2} \\
&= S_{21} V_{01}^- e^{j\beta z_2} \\
&= |S_{21}| V_{01}^- e^{j(\beta z_2 + \angle S_{21})}
\end{aligned}
$$

This phase shift is created by the propagation delay of the filter. The propagation delay $\tau(\omega)$ and phase delay $\angle S_{21}(\omega)$ are related by

$$
\tau(\omega) = \frac{\partial \angle S_{21}(\omega)}{\partial \omega}
$$

which implies that a constant phase shift corresponds to a constant delay.
If different frequencies propagate at different velocities through a microwave filter (each frequency has a different delay), the output signal will be distored. This phenomenon is known as **dispersion**, and this can be calculated using the phase delay $(\Delta \angle S_{21})$ and signal bandwidth $2\pi B_s$. When

$$
B_s \Delta \angle S_{21} \ll 1
$$

the dispersion is insignificant. Typically a filter is designed so that

$$
B_s \Delta \angle S_{21} \leqslant 0.1
$$

7.6.7 Normalised Frequency Parameter

We can use normalised attenuation charts to predict the attenuation of frequencies in the stopband. The normalised frequency parameter is defined as

- **Low pass filters**:
$$
\alpha = \frac{f}{f_c} - 1
$$

- **High pass filters**:
$$
\alpha = \frac{f_c}{f} - 1
$$

where f_c is the cut-off frequency of the filter. For bandpass and bandstop filters:

- **Band pass filters**:
$$
\alpha = \left| \frac{1}{\Delta} \left(\frac{f}{f_0} - \frac{f_0}{f} \right) \right| - 1
$$

- **Band stop filters**:
$$
\alpha = \left| \frac{1}{\Delta} \left(\frac{f_0}{f} - \frac{f}{f_0} \right)^{-1} \right| - 1
$$

where $f_0 = \sqrt{f_L f_H}$ is the centre frequency, and $\Delta = \frac{f_H - f_L}{f_0}$ is the percentage bandwidth.

7.7   Receiver Design

A receiver is a device that receives a signal from an antenna and converts it to a form suitable for further processing.

7.7.1   General Requirements

A receiver must meet the following general requirements:

- **High Gain**: A gain of around $100\,\text{dB}$ used to amplify $10\,\mu\text{V}$ signals to $1\,\text{V}$. In a $50\,\Omega$ system, this corresponds to a gain of $-87\,\text{dBm}$ to $13\,\text{dBm}$.

- **Good Selectivity**: Ability to reject adjacent channel signals, image frequencies and interference. For example, AM requires a channel spacing of $9\,\text{kHz}$, while FM requires a channel spacing of $200\,\text{kHz}$.

- **Good Sensitivity**: Ability to discern weak signals. Receiver noise must be less than $1\,\mu\text{V}$, with an SNR of at least $20\,\text{dB}$.

- **Large Dynamic Range**: Ability to operate properly when the received signal amplitude varies.

and the following application specific requirements:

- **EMC Compliance**: Minimal RFI emissions and adequate immunity.

- **Compliance to Specific Standards**: Varies depending on application.

- **Other Requirements**: Physical, environmental, power, etc.

7.7.2   Superheterodyne Receivers

A superheterodyne receiver is a receiver that tunes the received signal frequency to a fixed intermediate frequency (IF) using a tunable local oscillator (LO). The advantages of this receiver are that amplification and filtering occurs at a fixed IF frequency, which sits lower than the signal frequency allowing for high gain and sharp cut-off characteristics when filtering.
We use the following terms to describe the superheterodyne receiver:

- **Down Conversion**: The IF frequency is lower than the received signal frequency.

- **Up Conversion**: The IF frequency is higher than the received signal frequency.

- **Single Conversion**: Only one mixer stage, giving one IF frequency.

- **Double Conversion**: Two mixer stages, giving two IF frequencies.

- **Triple Conversion**: Three mixer stages, giving three IF frequencies.

The IF filter must be able to filter adjacent channels with an attenuation between $30\,\text{dB}$ and $60\,\text{dB}$. When down converting a signal,

$$|f_s - f_{\text{LO}}| = f_{\text{IF}}$$

or in other words,

- **Low-side tuning**: The LO frequency is lower than the desired signal frequency.

$$f_{\text{LO}} = f_s - f_{\text{IF}}$$

which has a lower cost, higher output power, and better frequency accuracy.

- **High-side tuning**: The LO frequency is higher than the desired signal frequency.

$$f_{\text{LO}} = f_s + f_{\text{IF}}$$

which requires a smaller percentage bandwidth, and results in lower phase noise.

7.8   Spurious Signals

As discussed in the previous section, the mixer will produce many spurious frequencies. These frequencies are of particular importance, as they will pass through the IF filter. Let us consider first, second, and third-order spurious frequencies.

7.8.1   First-Order Frequencies

The first-order frequency that can pass through the IF filter is the IF frequency itself $f_{\text{RF}} = f_{\text{IF}}$.

7.8.2   Second-Order Frequencies

Second-order spurious frequencies are called **image frequencies**, $f_{\text{image}} \neq f_s$ that satisfy

$$f_{\text{IF}} = \left| f_{\text{image}} - f_{\text{LO}} \right|$$

where f_{IF} is fixed, and f_{LO} is either high side or low side tuned. This leads to the following two cases:

- In **low-side tuning**

$$f_{\text{LO}} = f_s - f_{\text{IF}}$$

  so that

$$f_{\text{IF}} = \begin{cases} f_{\text{image}} - (f_s - f_{\text{IF}}) & f_{\text{image}} > f_{\text{LO}} \\ (f_s - f_{\text{IF}}) - f_{\text{image}} & f_{\text{image}} < f_{\text{LO}} \end{cases}$$

  which results in an image frequency at $f_{\text{image}} = f_s - 2f_{\text{IF}}$.

- In **high-side tuning**

$$f_{\text{LO}} = f_s + f_{\text{IF}}$$

  so that

$$f_{\text{IF}} = \begin{cases} f_{\text{image}} - (f_s + f_{\text{IF}}) & f_{\text{image}} > f_{\text{LO}} \\ (f_s + f_{\text{IF}}) - f_{\text{image}} & f_{\text{image}} < f_{\text{LO}} \end{cases}$$

  which results in an image frequency at $f_{\text{image}} = f_s + 2f_{\text{IF}}$.

In general, this frequency satisfies

$$\left| f_{\text{image}} - f_s \right| = 2f_{\text{IF}}$$

This is particularly problematic as second order terms have a significant effect on the output. In addition to this, we must never select the LO frequency to be equal to the IF frequency, as it will pass unimpeded through the IF filter.

7.8.3   Third-Order Frequencies

If we consider third-order terms, we also find a similar problem. The third-order terms of particular significance are the ones nearest to the desired signal frequency, namely

$$|2f_{\mathrm{RF}} - f_{\mathrm{LO}}| = f_{\mathrm{IF}} \qquad \text{and} \qquad |2f_{\mathrm{LO}} - f_{\mathrm{RF}}| = f_{\mathrm{IF}}$$

Again, we can consider the two cases of low-side and high-side tuning:

- In **low-side tuning**

$$f_{\mathrm{LO}} = f_s - f_{\mathrm{IF}}$$

  so that

$$f_{\mathrm{IF}} = \begin{cases} 2f_{\mathrm{RF}} - (f_s - f_{\mathrm{IF}}) & 2f_{\mathrm{RF}} > f_{\mathrm{LO}} \\ (f_s - f_{\mathrm{IF}}) - 2f_{\mathrm{RF}} & 2f_{\mathrm{RF}} < f_{\mathrm{LO}} \\ 2(f_s - f_{\mathrm{IF}}) - f_{\mathrm{RF}} & 2f_{\mathrm{LO}} > f_{\mathrm{RF}} \\ f_{\mathrm{RF}} - 2(f_s - f_{\mathrm{IF}}) & 2f_{\mathrm{LO}} < f_{\mathrm{RF}} \end{cases}$$

  which results in four frequencies at

$$f_{\mathrm{RF}} = \frac{f_s}{2}$$
$$f_{\mathrm{RF}} = \frac{f_s - 2f_{\mathrm{IF}}}{2}$$
$$f_{\mathrm{RF}} = 2f_s - 3f_{\mathrm{IF}}$$
$$f_{\mathrm{RF}} = 2f_s - f_{\mathrm{IF}}$$

  where the most significant are the first and third frequencies.

- In **high-side tuning**

$$f_{\mathrm{LO}} = f_s + f_{\mathrm{IF}}$$

  so that

$$f_{\mathrm{IF}} = \begin{cases} 2f_{\mathrm{RF}} - (f_s + f_{\mathrm{IF}}) & 2f_{\mathrm{RF}} > f_{\mathrm{LO}} \\ (f_s + f_{\mathrm{IF}}) - 2f_{\mathrm{RF}} & 2f_{\mathrm{RF}} < f_{\mathrm{LO}} \\ 2(f_s + f_{\mathrm{IF}}) - f_{\mathrm{RF}} & 2f_{\mathrm{LO}} > f_{\mathrm{RF}} \\ f_{\mathrm{RF}} - 2(f_s + f_{\mathrm{IF}}) & 2f_{\mathrm{LO}} < f_{\mathrm{RF}} \end{cases}$$

  which results in four frequencies at

$$f_{\mathrm{RF}} = \frac{f_s + 2f_{\mathrm{IF}}}{2}$$
$$f_{\mathrm{RF}} = \frac{f_s}{2}$$
$$f_{\mathrm{RF}} = 2f_s + f_{\mathrm{IF}}$$
$$f_{\mathrm{RF}} = 2f_s + 3f_{\mathrm{IF}}$$

  where the most significant are the first and third frequencies.

7.8.4 Rejecting Spurious Signals

As suggested above, we can reject spurious signals, using a **preselector filter** which is a bandpass filter centred around the desired signal frequency. It is important to consider the bandwidth of the preselector filter, as it must be wide enough to pass the signal entirely, but narrow enough to reject the spurious signals.

An important parameter is the IF frequency itself, when rejecting second-order frequencies, we may increase the IF frequency to increase the distance between the desired signal frequency and the image frequency, making it easier to filter out. However, when rejecting some third-order frequencies, it may be better to decrease the IF frequency. Clearly we must choose an IF frequency that is a compromise between these two requirements.

7.8.5 Up Conversion

Instead of down converting the signal, we can also up convert the signal to only produce one image frequency, and fewer third-order spurious frequencies, by considering

$$f_{\text{IF}} = f_{\text{RF}} + f_{\text{LO}}$$

Despite this advantage, we must also consider the cost and complexity implications of designing an IF filter at a higher frequency than the desired frequency.

7.9   Noise

The noise level of a system sets the lower limit on the strength of a signal that can be detected in the presense of noise. Noise can be attributed to several sources:

- **Thermal noise** — caused by thermal vibration of bound charges.

- **Shot noise** — caused by random fluctuations of charge carriers in an electron tube or solid-state device.

- **Flicker noise** — occurs in solid-state components and vacuum tubes. Flicker noise power varies inversely with frequency.

- **Plasma noise** — caused by random motion of charges in an ionized gas, e.g., a plasma, the ionosphere or sparking electrical contacts.

- **Quantum noise** — caused by quantized nature of charge carriers and photons. Usually insignificant relative to other noise sources.

7.9.1 Thermal Noise

Thermal noise is the most significant source of noise in a device. The RMS voltage of thermal noise is given by

$$V_{\text{noise}} = \sqrt{4k_B T R B}$$

where k_B is the Boltzmann constant, T is the temperature in K, R is the resistance in Ω, and B is the bandwidth in Hz. Using a Thevenin equivalent circuit consisting of a noiseless resistor with

resistance R and voltage source V_{noise}, the maxmimum noise power transfer from the noisy resistor to a load resistor $R_L = R$ is given by

$$N_{\text{out}} = P_{\text{noise}} = \frac{V_{\text{noise}}^2}{4R_L} = k_B T B$$

measured as a dimensionless quantity. Therefore noise power is independent of resistance and is proportional to temperature and bandwidth.

7.9.2 Equivalent Noise Temperature

The equivalent noise temperature T_e (K) describes the noise generated by a device relative to the noise generated by a resistor at room temperature. It is the noise power of the device in the absense of noise at the input (input noise power is zero). It is defined as

$$T_e = \frac{N_{\text{out}}}{k_B B}$$

where N_{out} is the output noise power of the device. For an **amplifier** with gain G, the output noise power is given by

$$N_{\text{out}} = G k_B T_e B$$

and therefore the equivalent noise temperature is given by

$$T_e = \frac{N_{\text{out}}}{G k_B B}.$$

An alternative specification for output noise is the **excess noise ratio** (ENR), which is the relative difference between the noise temperature of a device and room temperature T_0 (which is typically specified as $290\,\text{K}$), so that

$$\text{ENR} = \frac{T_e - T_0}{T_0}.$$

7.9.3 Measuring Noise Temperature

To measure the equivalent noise temperature of a device, we must measure the output noise power at a hot ambient temperature T_h, and at a cold ambient temperature T_c. The output noise power is then

$$N_{\text{out,hot}} = G k_B \left(T_h + T_e \right) B$$
$$N_{\text{out,cold}} = G k_B \left(T_c + T_e \right) B$$

Then by calculating the Y-factor, we can determine T_e experimentally.

$$Y = \frac{N_{\text{out,hot}}}{N_{\text{out,cold}}} = \frac{T_h + T_e}{T_c + T_e}$$
$$T_e = \frac{T_h - Y T_c}{Y - 1}$$

This method is accurate when $Y \gg 1$, and hence T_c must be as small as possible, and T_h must be as large as possible.

7.9.4   Noise Figure

The **noise figure** F of a device is the ratio of the input signal-to-noise ratio to the output signal-to-noise ratio, so that

$$F = \frac{\text{SNR}_{\text{in}}}{\text{SNR}_{\text{out}}} = \frac{S_{\text{in}}/N_{\text{in}}}{S_{\text{out}}/N_{\text{out}}}.$$

For a two-port network, the input noise power is given by

$$N_{\text{in}} = k_B T_p B$$

where T_p is the physical temperature of the device. The output noise power is then the sum of the input noise power and the noise generated by the device, so that

$$N_{\text{out}} = G N_{\text{in}} + G k_B T_e B = G k_B \left(T_p + T_e\right) B$$

Noise figure can then be related to equivalent noise temperature as follows,

$$
\begin{aligned}
F &= \frac{S_{\text{in}}/N_{\text{in}}}{S_{\text{out}}/N_{\text{out}}} \\
&= \frac{S_{\text{in}}}{k_B T_p B} \frac{G k_B \left(T_p + T_e\right) B}{G S_{\text{in}}} \\
&= \frac{T_p + T_e}{T_p} \\
&= 1 + \frac{T_e}{T_p}
\end{aligned}
$$

Note that noise figure is typically measured at room temperature, so that $T_p = T_0$, and therefore

$$F = 1 + \frac{T_e}{T_0}.$$

Solving for equivalent noise temperature, we find that

$$T_e = (F - 1) T_0.$$

7.9.5   Noise Figure of Cascaded Devices

In a receiver system, consider some devices i with gain G_i, equivalent noise temperature T_i, and bandwidth B_i. The input of the first device is connected to a source with ambient temperature T_0, so that the input noise power N_{in} is given by

$$N_{\text{in}} = k_B T_0 B_1.$$

The output noise power of the first device N_1 is given by

$$
\begin{aligned}
N_1 &= G_1 N_{\text{in}} + G_1 k_B T_1 B_1 \\
&= G_1 k_B \left(T_0 + T_1\right) B_1.
\end{aligned}
$$

The output noise power of the second device N_2 is then given by

$$
\begin{aligned}
N_2 &= G_2 N_1 + G_2 k_B T_2 B_2 \\
&= G_1 G_2 k_B \left(T_0 + T_1 \right) B_1 + G_2 k_B T_2 B_2 \\
&= G_1 G_2 k_B \left(T_0 B_1 + T_1 B_1 + \frac{T_2 B_2}{G_1} \right).
\end{aligned}
$$

For a system with $n > 0$ cascaded devices, the output noise power is given by

$$
\begin{aligned}
N_n &= \left(\prod_{i=1}^{n} G_i \right) k_B \left(T_0 B_1 + \sum_{i=1}^{n} \frac{T_i B_i}{\prod_{j=1}^{i-1} G_j} \right) \\
&= \underbrace{\left(\prod_{i=1}^{n} G_i \right) k_B T_0 B_1}_{\text{noise from source}} + \underbrace{\left(\prod_{i=1}^{n} G_i \right) k_B \sum_{i=1}^{n} \frac{T_i B_i}{\prod_{j=1}^{i-1} G_j}}_{\text{noise from system}}
\end{aligned}
$$

When the bandwidth B is the same for all devices, the output noise power is given by

$$
\begin{aligned}
N_n &= \left(\prod_{i=1}^{n} G_i \right) k_B \left(T_0 + \sum_{i=1}^{n} \frac{T_i}{\prod_{j=1}^{i-1} G_j} \right) B \\
&= \underbrace{\left(\prod_{i=1}^{n} G_i \right) k_B T_0 B}_{\text{noise from source}} + \underbrace{\left(\prod_{i=1}^{n} G_i \right) k_B \sum_{i=1}^{n} \frac{T_i B}{\prod_{j=1}^{i-1} G_j}}_{\text{noise from system}}
\end{aligned}
$$

The equivalent noise temperature T_e of the system can be found by considering the system as a single device with gain $G = \prod_{i=1}^{n} G_i$ and output noise power N_{out}, so that

$$
N_{\text{out}} = G N_{\text{in}} + G k_B T_e B = G k_B \left(T_0 + T_e \right) B.
$$

Equating this with N_n shows that

$$
G k_B \left(T_0 B_1 + \sum_{i=1}^{n} \frac{T_i B_i}{\prod_{j=1}^{i-1} G_j} \right) = G k_B \left(T_0 + T_e \right) B
$$

$$
\frac{1}{B} \left(T_0 B_1 + \sum_{i=1}^{n} \frac{T_i B_i}{\prod_{j=1}^{i-1} G_j} \right) = T_0 + T_e
$$

$$
T_e = \frac{1}{B} \left(T_0 \left(B_1 - B \right) + \sum_{i=1}^{n} \frac{T_i B_i}{\prod_{j=1}^{i-1} G_j} \right)
$$

and when B is the same for all devices,

$$
T_e = \sum_{i=1}^{n} \frac{T_i}{\prod_{j=1}^{i-1} G_j}.
$$

50

The noise figure of the system is then given by

$$(F-1)T_0 = \frac{1}{B}\left(T_0\left(B_1 - B\right) + \sum_{i=1}^{n}\frac{\left(F_i - 1\right)T_0 B_i}{\prod_{j=1}^{i-1}G_j}\right)$$

$$F-1 = \frac{1}{B}\left(B_1 - B + \sum_{i=1}^{n}\frac{\left(F_i - 1\right)B_i}{\prod_{j=1}^{i-1}G_j}\right)$$

$$F = \frac{1}{B}\left(B_1 + \sum_{i=1}^{n}\frac{\left(F_i - 1\right)B_i}{\prod_{j=1}^{i-1}G_j}\right) = \frac{1}{B}\left(F_1 B_1 + \sum_{i=2}^{n}\frac{\left(F_i - 1\right)B_i}{\prod_{j=1}^{i-1}G_j}\right)$$

and when B is the same for all devices,

$$F = 1 + \sum_{i=1}^{n}\frac{F_i - 1}{\prod_{j=1}^{i-1}G_j} = F_1 + \sum_{i=2}^{n}\frac{F_i - 1}{\prod_{j=1}^{i-1}G_j}$$

7.9.6   Antenna Noise

Similar to the equivalent noise temperature of a resistor, the equivalent noise temperature of an antenna T_A is defined as

$$T_A = \eta T_b + \left(1 - \eta\right)T_p$$

where T_b is the average background brightness temperature of the antenna[2], T_p is the physical temperature of the antenna, and η is the antenna efficiency. The output noise power of an antenna is given by

$$N_A = k_B T_A B = k_B\left(\eta T_b + \left(1 - \eta\right)T_p\right)B.$$

The signal-to-noise ratio at the point where the antenna feeds the receiver is given by

$$\mathrm{SNR} = \frac{P_{\mathrm{Rx}}}{N_A} = \left(\frac{G_{\mathrm{Rx}}}{T_A}\right)\frac{G_{\mathrm{Tx}}P_{\mathrm{Tx}}}{k_B B}\left(\frac{\lambda}{4\pi r}\right)^2$$

using the Friis transmission equation, which states

$$\frac{P_{\mathrm{Rx}}}{P_{\mathrm{Tx}}} = G_{\mathrm{Tx}}G_{\mathrm{Rx}}\left(\frac{\lambda}{4\pi r}\right)^2.$$

This suggests that a receiver with a better G/T ratio performs better.

7.9.7   Noise Figure of a System

Given the noise power from an antenna, the total input noise power is given by

$$N_{\mathrm{in}}^{\mathrm{total}} = N_A + N_{\mathrm{out}} = k_B\left(T_A + T_e\right)B$$

therefore the total output noise power of the antenna and receiver is given by

$$N_{\mathrm{out}}^{\mathrm{total}} = GN_{\mathrm{in}}^{\mathrm{total}} = Gk_B\left(T_A + T_e\right)B.$$

---

[2]The background brightness temperature is the equivalent noise temperature of a resistor required to produce the same noise power as the environment seen by the antenna. It indicates the amount of electromagnetic noise in the environment. As the background brightness temperature is direction dependent, the average background brightness temperature T_b is used.

7.9.8   Transmission Line Noise

Transmission lines are used to connect RF components together and typically degrade the power of a signal. This is specified as a loss L, which is inversely proportional to the gain. Therefore, we can treat transmission lines as a device with gain $G = 1/L$ and noise figure $F = L$.