



BUILDING COMMUNICATION THEORY

PARUL PATNAIK

Building Communication Theory

Building Communication Theory

Parul Patnaik



Published by The InfoLibrary,
4/21B, First Floor, E-Block,
Model Town-II,
New Delhi-110009, India

© 2022 The InfoLibrary

Building Communication Theory
Parul Patnaik
ISBN: 978-93-5496-565-4

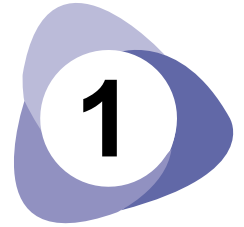
This book contains information obtained from authentic and highly regarded sources. All chapters are published with permission under the Creative Commons Attribution Share Alike License or equivalent. A wide variety of references are listed. Permissions and sources are indicated; for detailed attributions, please refer to the permissions page. Reasonable efforts have been made to publish reliable data and information, but the authors, editors and publisher cannot assume any responsibility for the validity of all materials or the consequences of their use.

Trademark Notice: All trademarks used herein are the property of their respective owners. The use of any trademark in this text does not vest in the author or publisher any trademark ownership rights in such trademarks, nor does the use of such trademarks imply any affiliation with or endorsement of this book by such owners.

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy. Furthermore, the publisher ensures that the text paper and cover boards used have met acceptable environmental accreditation standards.

Table of Contents

| | | |
|------------------|------------------------|------------|
| Chapter 1 | AMPLITUDE MODULATION | 1 |
| Chapter 2 | ANGLE MODULATION | 30 |
| Chapter 3 | RANDOM PROCESS | 48 |
| Chapter 4 | NOISE CHARACTERIZATION | 88 |
| Chapter 5 | INFORMATION THEORY | 119 |



AMPLITUDE MODULATION

Generation and detection of AM wave-spectra - DSBSC, Hilbert Transform, Pre-envelope & complex envelope - SSB and VSB - comparison - Superheterodyne Receiver.

1.1 Generation and detection of AM wave-spectra

Generation of AM waves:

In amplitude modulation, the amplitude of the carrier signal is varied by the modulating/message/information/baseband signal in accordance with the instantaneous values of the message signal. That is, amplitude of the carrier signal is made proportional to the instantaneous values of the modulating signal.

If $m(t)$ is the information signal and $c(t) = A_c \cos(2\pi f_c t + \phi)$ is the carrier, the amplitude of the carrier signal is varied proportional to $m(t)$.

The peak amplitude of the carrier after modulation at any instant is given by $[A_c + m(t)]$. The carrier signal after modulation or the modulated signal is represented by the equation.

$$s(t) = [A_c + m(t)]\cos(2\pi f_c t + \phi)$$

$$s(t) = A_c [1 + k_a m(t)]\cos(2\pi f_c t + \phi)$$

Where, $k_a = \frac{1}{A_c}$ is known as the amplitude sensitivity of the modulator.

Let $m(t) = A_m \cos(2\pi f_m t)$ be the message signal of frequency f_m and peak amplitude A_m . Then single tone modulated signal is given by the equation.

$$s(t) = A_c [1 + k_a A_m \cos(2\pi f_m t)] \cos(2\pi f_c t + \phi)$$

$$s(t) = A_c \left[1 + \frac{A_m}{A_c} \cos(2\pi f_m t) \right] \cos(2\pi f_c t + \phi)$$

$$s(t) = A_c [1 + m \cos(2\pi f_m t)] \cos(2\pi f_c t + \phi)$$

Where, $m = A_m/A_c$ is called the modulation index or depth of modulation.

We have the modulation index given by,

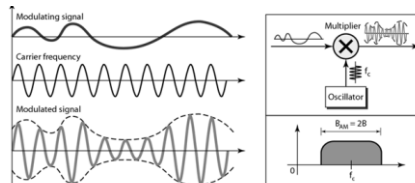
$$m = \frac{A_m}{A_c} \dots\dots\dots(1)$$

$$A_m = \frac{A_{\max} - A_{\min}}{2} \dots\dots\dots(2)$$

$$A_c = A_{\max} - A_m \dots\dots\dots(3)$$

$$A_c = A_{\max} - \frac{A_{\max} - A_{\min}}{2}$$

$$A_c = \frac{A_{\max} + A_{\min}}{2} \dots\dots\dots(4)$$



Message and amplitude modulated signal

Dividing equation (2) by (4) we get,

$$m = \frac{A_m}{A_c} = \frac{A_{\max} - A_{\min}}{A_{\max} + A_{\min}}$$

Where,

A_{\max} = Maximum amplitude

A_{\min} = Minimum amplitude of the modulated signal

Modulation index, m has to be governed such that, it is always less than unity, otherwise it results in a situation known as over modulation ($m > 1$). Over modulation occurs when the magnitude of the peak amplitude of the modulating signal exceeds the magnitude of the peak amplitude of the carrier signal. The signal gets distorted due to over modulation, because of this limitation on modulation index the system clarity is also limited.

Consider the standard expression for AM wave $s(t) = A_c[1 + k_a A_m \cos(2\pi f_m t)]\cos(2\pi f_c t)$. The carrier frequency f_c is much greater than the highest frequency component W of the message signal.

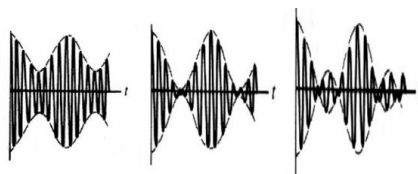
(i.e.) $f_c \gg W$

W is called as the message bandwidth.

The Fourier transform $S(f)$ of the AM wave $s(t)$ is given by,

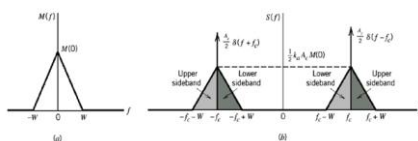
$$s(f) = \frac{A_c}{2} [\delta(f - f_c) + \delta(f + f_c)] + \frac{k_a A_c}{2} [M(f - f_c) + M(f + f_c)]$$

The AM waveforms for different values of modulation index m , are as shown in the below figure.



AM waves for different values of m

Frequency spectrum of AM waves:



Spectrum of message and AM waves

Suppose that the baseband signal $m(t)$ is band limited to the interval $-W \leq f \leq W$ for $f_c > W$. This spectrum consists of two delta functions weighted by the factor $A_c/2$ and occurring at $\pm f_c$ and two versions of the baseband spectrum translated into frequency by $\pm f_c$. From the spectrum, the following points are noted.

(i) For positive frequencies, the highest frequency component of the AM wave is $f_c + W$ and the lowest frequency component is $f_c - W$. The difference between these two frequencies defines the transmission bandwidth B_T for an AM wave, which is exactly twice the message bandwidth W .

$$B_T = 2W$$

(ii) For positive frequencies, the portion of the spectrum of an AM wave lying above the carrier frequency f_c , is referred to as the Upper Side Band (USB), whereas the symmetric portion below f_c , is called as the Lower Side Band (LSB).

For negative frequencies, USB is the portion of the spectrum below $-f_c$ and LSB is the portion above $-f_c$. The condition $f_c > W$ ensures that, the side bands do not overlap.

The AM wave $s(t)$ is a voltage or current wave. In either case, the average power delivered to 1Ω resistor by $s(t)$ comprises of three components.

$$\text{Carrier power} = \frac{A_c^2}{2}$$

$$\text{Upper side-frequency power} = \frac{m^2 A_c^2}{8}$$

$$\text{Lower side-frequency power} = \frac{m^2 A_c^2}{8}$$

Spectrum power:

The power spectrum gives a plot of the portion of a signal's power falling within the given frequency bins. The common way of generating a power spectrum is by using a discrete Fourier transform, but other techniques such as the maximum entropy method can also be used.

Amplitude modulators:

Two basic amplitude modulation principles are,

- Square law modulation
- Switching modulation

Square law modulator:

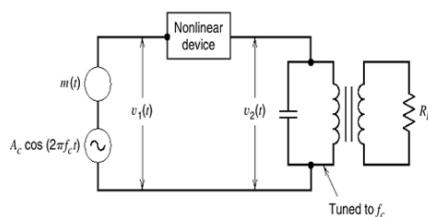
When the output of a device is not directly proportional to the input throughout the operation, the device is said to be non-linear. The input, output relation of a non-linear device can be expressed as,

$$V_0 = a_0 + a_1 V_{in} + a_2 V_{in}^2 + a_3 V_{in}^3 + a_4 V_{in}^4 + \dots$$

When the input is very small, the higher power terms can be neglected. Hence the output is approximately given by,

$$V_0 = a_0 + a_1 V_{in} + a_2 V_{in}^2$$

When the output is considered up to square of the input, the device is called as a square law device and the square law modulator is shown in the below figure.



Square law modulator

Consider a non-linear device to which a carrier signal $c(t) = A_c \cos(2\pi f_c t)$ and an information signal $m(t)$ are fed simultaneously as shown in above figure. The total input to the device at any instant is given by,

$$V_{in} = c(t) + m(t)$$

$$V_{in} = A_c \cos(2\pi f_c t) + m(t)$$

As the level of the input is very small, the output can be considered up to square of the input,

$$(i.e.) V_o = a_0 + a_1 V_{in} + a_2 V_{in}^2$$

$$V_o = a_0 + a_1 [A_c \cos(2\pi f_c t) + m(t)] + a_2 [A_c \cos(2\pi f_c t) + m(t)]^2$$

$$V_o = a_0 + a_1 A_c \cos 2\pi f_c t + a_1 m(t) + \frac{a_2 A_c^2}{2} (1 + \cos 4\pi f_c t) + a_2 [m(t)]^2 + 2a_2 m(t) A_c \cos 2\pi f_c t$$

$$V_o = a_0 + a_1 A_c \cos 2\pi f_c t + a_1 m(t) + \frac{a_2 A_c^2}{2} \cos 4\pi f_c t + a_2 m^2(t) + 2a_2 m(t) A_c \cos 2\pi f_c t$$

Taking Fourier transform on both sides, we get,

$$V_o(f) = (a_0 + \frac{a_2 A_c^2}{2}) \delta(f) + \frac{a_1 A_c}{2} [\delta(f - f_c) + \delta(f + f_c)] + a_1 M(f) + \frac{a_2 A_c^2}{4} [\delta(f - 2f_c) + \delta(f + 2f_c)] + a_2 M(f) + a_2 A_c [M(f - f_c) + M(f + f_c)]$$

Therefore, the output V_o of the square law device consists of,

- A DC component at $f = 0$.
- The information signal ranging from 0 to W Hz and its second harmonics.
- Signal at f_c and $2f_c$.
- Frequency band centered at f_c with a deviation of $\pm W$ Hz.

The AM signal with a carrier frequency of f_c can be separated using a band pass filter at the output of the square law device. The filter should have a lower cut-off frequency ranging between $2W$ and $(f_c - W)$ and upper cut-off frequency between $(f_c + W)$ and $2f_c$.

Therefore the filter output is given by,

$$s(t) = a_1 A_c \cos(2\pi f_c t) + 2a_2 A_c m(t) \cos(2\pi f_c t)$$

$$s(t) = a_1 A_c \left[1 + 2 \frac{a_2}{a_1} m(t) \right] \cos 2\pi f_c t$$

If $m(t) = A_m \cos(2\pi f_m t)$ we get,

$$s(t) = a_1 A_c \left[1 + 2 \frac{a_2}{a_1} A_m \cos 2\pi f_m t \right] \cos 2\pi f_c t$$

Comparing this with the standard representation of AM signal we get,

$$s(t) = A_c [1 + k_a m(t)] \cos(2\pi f_c t)$$

Therefore, modulation index of the output signal is given by,

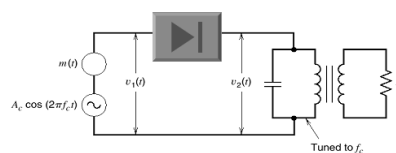
$$m = 2 \frac{a_2}{a_1} A_m$$

The output AM signal is free from attenuation and distortion only when,

$$(f_c - W) > 2W \text{ or } f_c > 3W$$

Switching modulator:

Consider a semiconductor diode, used as an ideal switch to which the carrier signal $c(t) = A_c \cos(2\pi f_c t)$ and information signal $m(t)$ are applied simultaneously as shown in the figure.



Switching modulator

The total input to the diode at any instant is given by,

$$v_1 = c(t) + m(t)$$

$$v_1 = A_c \cos(2\pi f_c t) + m(t)$$

When the peak amplitude of $c(t)$ is maintained more than that of an information signal, the operation is assumed to be dependent on only $c(t)$ irrespective of $m(t)$. When $c(t)$ is positive, $v_2 = v_1$, since the diode is forward biased. Similarly, when $c(t)$ is negative, $v_2 = 0$, since the diode is reverse biased. Based upon the above operation, switching response of the diode is a periodic rectangular wave with an amplitude unity and is given by,

$$p(t) = \frac{1}{2} + \frac{1}{\pi} \sum_{n=-\infty}^{\infty} \frac{(-1)^{n-1}}{2n-1} \cos(2\pi f_c t(2n-1))$$

$$p(t) = \frac{1}{2} + \frac{2}{\pi} \cos(2\pi f_c t) - \frac{2}{3\pi} \cos(6\pi f_c t) + \dots$$

Therefore, the diode response v_2 is a product of switching response $p(t)$ and input v_1 .

$$v_2 = v_1 * p(t)$$

$$v_2 = [A_c \cos 2\pi f_c t + m(t)] \left[\frac{1}{2} + \frac{2}{\pi} \cos 2\pi f_c t - \frac{2}{3\pi} \cos 6\pi f_c t + \dots \right]$$

Applying Fourier transform we get,

$$\begin{aligned} v_2(f) &= \frac{A_c}{4} [\delta(f - f_c) + \delta(f + f_c)] + \frac{M(f)}{2} + \frac{A_c}{\pi} \delta(f) \\ &+ \frac{A_c}{2\pi} [\delta(f - 2f_c) + \delta(f + 2f_c)] + \frac{1}{\pi} [M(f - f_c) + M(f + f_c)] \\ &- \frac{A_c}{6\pi} [\delta(f - 4f_c) + \delta(f + 4f_c)] - \frac{A_c}{3\pi} [\delta(f - 2f_c) + \delta(f + 2f_c)] \\ &- \frac{1}{3\pi} [M(f - 3f_c) + M(f + f_c)] \end{aligned}$$

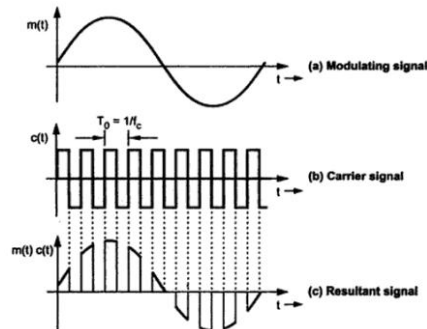
The diode output v_2 consists of,

- A DC component at $f = 0$.
- Information signal ranging from 0 to W Hz and infinite number of frequency bands centered at f , $2f_c$, $3f_c$, $4f_c$,

The required AM signal centered at f_c can be separated using BPF. The lower cut-off frequency for the band pass filter should be between W and $f_c - W$ and the upper cut-off frequency between $f_c + W$ and $2f_c$.

The filter output is given by the equation,

$$S(t) = \frac{A_c}{2} \left[1 + \frac{4}{\pi} \frac{m(t)}{A_c} \right] \cos 2\pi f_c t$$



Waveforms for switching modulator

For a single tone information, let,

$$m(t) = A_m \cos(2\pi f_m t)$$

$$S(t) = \frac{A_c}{2} \left[1 + \frac{4}{\pi} \frac{A_m}{A_c} \cos 2\pi f_m t \right] \cos 2\pi f_c t$$

Therefore the modulation index is given by,

$$m = \frac{4}{\pi} \frac{A_m}{A_c}$$

The output AM signal is free from distortions and attenuations only when $f_c - w > w$ or $f_c > 2w$.

Detection of AM waves:

Demodulation of AM:

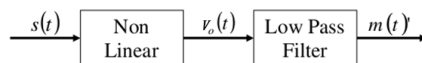
Demodulation is the process of recovering the information or baseband signal from the incoming modulated signal at the receiver.

There are two methods. They are,

- Square law detector
- Envelope detector

Square law demodulator:

Consider a non-linear device to which the AM signal, $s(t)$ is applied. When the level of $s(t)$ is very small, output can be considered up to square of the input.



Demodulation of AM using square law device

Therefore, $V_o = a_0 + a_1V_{in} + a_2V_{in}^2$

If $m(t)$ is the information signal (0 - ω Hz) and $c(t) = A_c \cos(2\pi f_c t)$ is the carrier, input AM signal to the non-linear device is given by,

$$s(t) = A_c[1 + k_a m(t)] \cos(2\pi f_c t)$$

$$V_o = a_0 + a_1 s(t) + a_2 [s(t)]^2$$

$$V_o = a_0 + a_1 A_c \cos(2\pi f_c t) + a_1 A_c k_a m(t) \cos(2\pi f_c t) + a_2 [A_c \cos(2\pi f_c t) + A_c k_a m(t) \cos(2\pi f_c t)]^2$$

Applying Fourier transform on both sides we get,

$$\begin{aligned} V_o(f) = & \left[a_0 + \frac{a_2 A_c^2}{2} \right] \delta(f) + \frac{a_1 A_c}{2} [\delta(f - f_c) + \delta(f + f_c)] \\ & + \frac{a_1 A_c K_a}{2} [M(f - f_c) + M(f + f_c)] + \frac{a_2 A_c^2 K_a^2}{4} [M(f - 2f_c) + M(f + 2f_c)] \\ & + \frac{a_2 A_c^2 K_a^2}{2} \left[M(f) \right] + \frac{a_2 A_c^2 K_a^2}{2} [M(f - 2f_c) + M(f + 2f_c)] \\ & + \frac{a_2 A_c^2}{4} [\delta(f - 2f_c) + \delta(f + 2f_c)] + a_2 A_c^2 K_a [M(f)] \end{aligned}$$

The required information can be separated using low pass filter with cut-off frequency ranging between ω and $f_c - \omega$.

The filter output is given by,

$$m'(t) = \left(a_0 + \frac{a_2 A_c^2}{2} \right) + a_2 A_c^2 K_a m(t) + \frac{a_2 A_c^2 K_a^2 m^2(t)}{2}$$

(DC component) (Message signal) (Second harmonic)

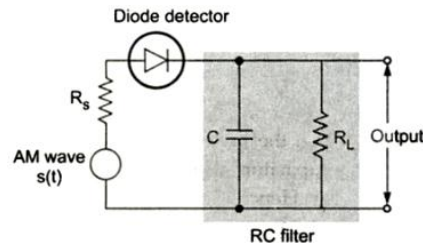
The DC component can be eliminated by using a coupling capacitor or a transformer. The effect of second harmonics of the information signal can be reduced by maintaining its level very low. When $m(t)$ is very low, the filter output is given by,

$$m^1(t) = a_2 A_c^2 K_a m(t)$$

When the information level is very low, the effect of noise increases at the receiver, hence the system clarity is very low.

Envelop detector:

It is a simple and highly effective system. This method is used in most of the commercial AM radio receivers. This circuit is also known as diode detector. An envelop detector is shown below.



Envelope detector

During the positive half cycle of the input signal, the diode D is forward biased and the capacitor C charges up rapidly to the peak of the input signal. When the input signal falls below this value, the diode gets reverse biased and the capacitor C discharges through the load resistor R_L .

The discharge process continues until the next positive half cycle. When the input signal becomes greater than the voltage across the capacitor, the diode conducts again and the process is repeated.

The charging time constant $R_s C$ must be short when compared with the carrier period $1/f_c$. The capacitor charges rapidly and when the diode is conducting it follows the applied voltage up to the positive peak.

The charging time constant shall satisfy the condition,

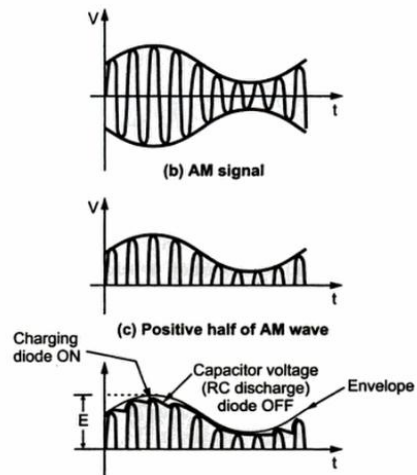
$$R_s C \ll \frac{1}{f_c}$$

On the other hand, the discharging time constant $R_L C$ must be long enough to ensure that the capacitor discharges slowly through the load resistor R_L , between the positive peaks of the carrier wave, but not so long, that the capacitor voltage will not discharge at the maximum rate of change of the modulating wave.

The discharge time constant shall satisfy the condition,

$$\frac{1}{f_c} \ll R_L C \ll \frac{1}{W}$$

Where, W is bandwidth of the message signal.



Envelope and voltage across capacitor

The result is that, the capacitor voltage or detector output is nearly the same as the envelope of AM wave.

1.2 DSBSC

Double Side band Suppressed Carrier (DSB-SC):

Double side band suppressed carrier transmission (DSB-SC) is a transmission in which frequencies produced by amplitude modulation are symmetrically spaced above and below the carrier frequency and the carrier level is reduced to the lowest practical level, ideally suppressed completely.

In DSB, there are two side bands in the frequency spectrum, USB and LSB, but the carrier component in full AM or DSB-LC (Double Side band Large Carrier) does not convey any information, it may be removed or suppressed during the modulation process to attain a higher power efficiency, hence Double Side Band Suppressed Carrier (DSB-SC) Modulation is used.

DSB-SC modulation-Time domain:

A DSB-SC signal is obtained by multiplying the modulating signal with a high frequency carrier wave. The modulated wave $s(t)$ is given by,

$$s(t) = m(t) \times c(t) = m(t) \cdot A_c \cos(\omega_c t + \phi) = A_c m(t) \cos(\omega_c t) \dots \dots \dots (1)$$

Assume the phase angle (ϕ) to be zero.

In amplitude modulation, the general form of modulated signal is given by,

$$s(t) = A(t) \cos(\omega_c t + \phi) \dots \dots \dots (2)$$

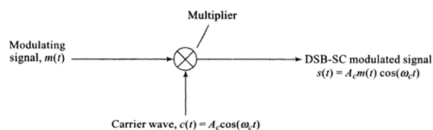
If we compare equation (1) and (2), we observe that $s(t)$ is indeed an amplitude modulated carrier with a time varying amplitude given by,

$$A(t) = A_c m(t)$$

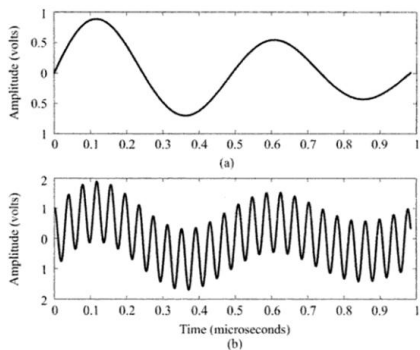
This time varying carrier envelope is proportional to $m(t)$.

The below figure shows the generation of DSB-SC signal. The multiplication operation is usually performed using a product modulator that could be in the form of a ring modulator or a balanced modulator. Integrated circuits are also available which can accomplish the multiplication process.

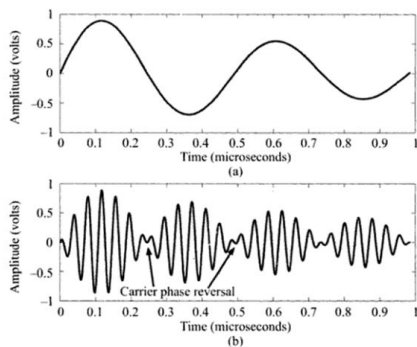
Let us now consider the time domain plot of a DSB-SC waveform corresponding to the same modulating signal as in figure (i).



DSB-SC modulated signal generation

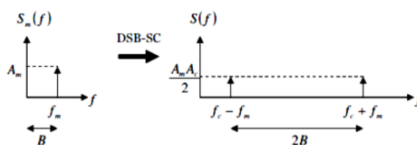


(i) Time domain waveforms of (a) The modulating signal, $m(t)$ and (b) Message plus carrier wave, $m(t) + c(t)$



(ii) Time domain waveforms of (a) The modulating signal, $m(t)$ and (b) The DSB-SC signal

The frequency description of the AM signal-DSB-SC:



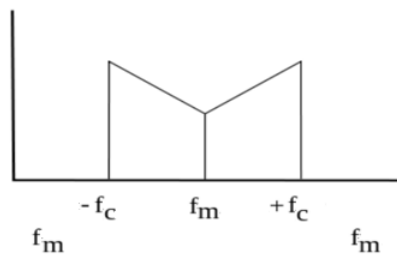
Note that there is no carrier frequency.

From the above analysis, we found that the frequency spectrum of AM waveform, DSB-SC has the following characteristics.

- No component of carrier frequency f_c .
- An upper side band (USB), whose highest frequency component is at $f_c + f_m$.
- A lower side band (LSB), whose highest frequency component is at $f_c - f_m$.
- The bandwidth is twice the modulating signal bandwidth.
- Since there are two side bands in the frequency spectrum, without carrier frequency, it is often called as Double Side band with Suppressed Carrier (DSB-SC).

Spectrum:

DSB-SC is basically an amplitude modulated wave without carrier, therefore reducing wastage of power, giving it a 50% efficiency. This is an increase when compared to normal AM transmission (DSB), which has a maximum efficiency of 33.333%, since 2/3 of the power is in the carrier and each side band carries the same information. Single Side Band Suppressed Carrier is 100% efficient.



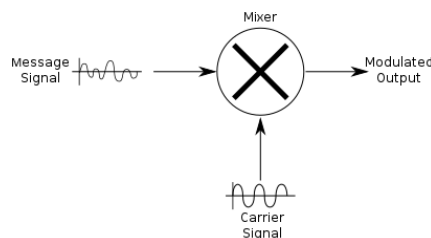
Spectrum plot of an DSB-SC signal

Generation:

DSB-SC is generated by a mixer. This consists of a message signal multiplied by the carrier signal. The mathematical representation of this process is shown below, where the product to sum trigonometric identity is used.

$$\frac{V_m \cos(\omega_m t)}{\text{Message}} \times \frac{V_c \cos(\omega_c t)}{\text{Carrier}} = \frac{V_m V_c}{2} [\cos((\omega_m + \omega_c) t) + \cos((\omega_m - \omega_c) t)]$$

Modulated Signal



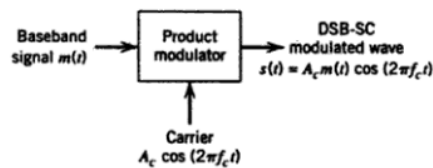
Generation of DSB-SC signal

DSB-SC modulated wave:

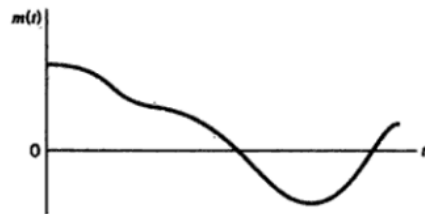
This form of linear modulation is generated by using a product modulator that simply multiplies the message signal $m(t)$ by the carrier wave $A_c \cos(2\pi f_c t)$ as illustrated in the below figure.

We can write as,

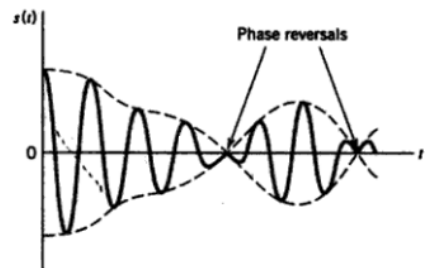
$$s(t) = A_c m(t) \cos(2\pi f_c t) \dots \dots \dots (3)$$



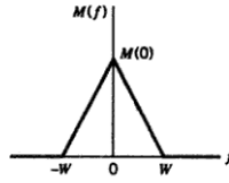
Block diagram of product modulator



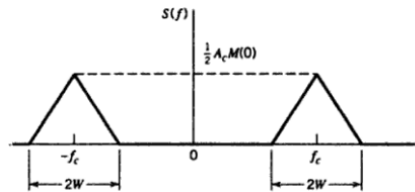
Baseband signal



DSB-SC modulated wave



Spectrum of baseband signal



Spectrum of DSB-SC modulated wave

The modulated signal $s(t)$ undergoes a phase reversal whenever the message signal $m(t)$ crosses zero. Consequently, the envelope of a DSB-SC modulated signal is different from the message signal, this is unlike the case of an AM wave that has a percentage modulation less than 100 percent.

From equation (3), the Fourier transform of $s(t)$ is obtained as,

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

Except for a change in scale factor, the modulation process simply translates the spectrum of the baseband signal by $\pm f_c$. The transmission bandwidth required by DSB-SC modulation is the same as that required for amplitude modulation, namely $2W$.

Synchronous detection of DSB-SC signal:

Demodulation is done by multiplying the DSB-SC signal with the carrier signal, just as modulation process. The resultant signal is then passed through a low pass filter to produce a scaled version of the original message signal.

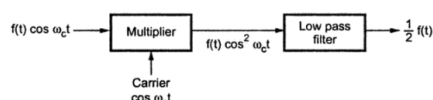
DSB-SC can be demodulated if modulation index is less than unity.

$$\begin{aligned} & \overbrace{\frac{V_m V_c}{2} [\cos((\omega_m + \omega_c)t) + \cos((\omega_m - \omega_c)t)]}^{\text{Modulated Signal}} \times \overbrace{V_c' \cos(\omega_c t)}^{\text{Carrier}} \\ &= \left(\frac{1}{2} V_c V_c' \right) \underbrace{V_m \cos(\omega_m t)}_{\text{original message}} + \frac{1}{2} V_c V_c' V_m [\cos((\omega_m + 2\omega_c)t) + \cos((\omega_m - 2\omega_c)t)] \end{aligned}$$

The equation above shows that by multiplying the modulated signal with the carrier signal, the result is a scaled version of the original message signal plus a second term.

Since $\omega_c \gg \omega_m$, this second term is much higher in frequency than the original message. Once this signal passes through a low pass filter, the higher frequency component is removed, leaving just the original message.

The DSB-SC signal can be detected with the help of synchronous detector as shown in the figure. The modulated signal $f(t) = \cos \omega_c t$ is given to the multiplier. The modulated signal is multiplied with locally generated carrier $\cos \omega_c t$.



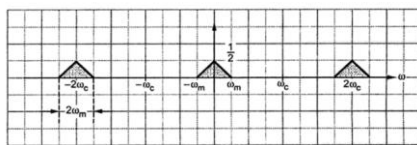
Synchronous detection

Thus, the multiplier output signal is $f(t)\cos^2\omega_c t$. This signal can be expressed mathematically as,

$$\begin{aligned} f(t) \cos^2 \omega_c t &= f(t) \left[\frac{1 + \cos 2\omega_c t}{2} \right] \\ &= \frac{1}{2} f(t) + \frac{1}{2} f(t) \cos 2\omega_c t \end{aligned}$$

The spectrum of the above signal will be,

$$f(t) \cos^2 \omega_c t \xrightarrow{FT} \frac{1}{2} F(\omega) + \frac{1}{4} [F(\omega + 2\omega_c) + F(\omega - 2\omega_c)]$$



Spectrum of signal at multiplier output

Note that, there is a spectrum $\frac{1}{2} F(\omega)$ located at $-\omega_m$ to $+\omega_m$. The two spectra are located at $+2\omega_c$ and $-2\omega_c$.

The low pass filter has a cut-off frequency just higher than ω_m . Hence it passes out the signal $\frac{1}{2} f(t)$. Since $2\omega_c \gg 2\omega_m$, the signal $\frac{1}{2} f(t) \cos 2\omega_c t$ is not passed by the low pass filter. Thus at the output of the low pass filter,

$$f_o(t) = \frac{1}{2} f(t)$$

Thus, the modulating signal is obtained back at the output of a synchronous detector.

Distortion and attenuation:

For demodulation, the demodulation oscillator's phase and frequency must be exactly the same as modulation oscillator, otherwise, distortion and attenuation will occur.

Consider the following conditions,

- Message signal to be transmitted: $f(t)$.
- Modulation signal: $V_c \cos(\omega_c t)$.
- Demodulation signal (with small frequency and phase deviations from the modulation signal): $V_c' \cos[(\omega_c + \Delta\omega)t + \theta]$

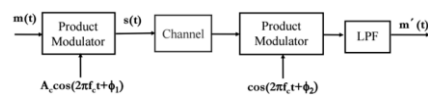
The resultant signal can be given by,

$$\begin{aligned} f(t) \times V_c \cos(\omega_c t) \times V_c' \cos[(\omega_c + \Delta\omega)t + \theta] \\ = \frac{1}{2} V_c V_c' f(t) \cos(\Delta\omega \cdot t + \theta) + \frac{1}{2} V_c V_c' f(t) \cos[(2\omega_c + \Delta\omega)t + \theta] \\ \xrightarrow{\text{After low pass filter}} \frac{1}{2} V_c V_c' f(t) \cos(\Delta\omega \cdot t + \theta) \end{aligned}$$

The $\cos(\Delta\omega \cdot t + \theta)$ terms results in distortion and attenuation of the original message signal. In particular, if the frequencies are correct, but the phase is wrong, contribution from θ is a constant attenuation factor, also $\Delta\omega \cdot t$ represents a cyclic inversion of the recovered signal, which is a serious form of distortion.

Coherent detection of DSB-SC:

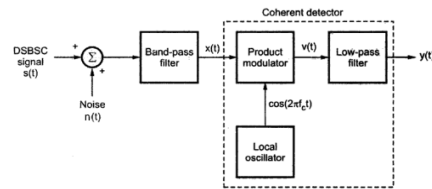
Detector uses another product modulator.



Demodulated signal: $m'(t) = 0.5A_c \cos(\phi_2 - \phi_1) m(t)$.

- Phase offset: If $\phi_2 - \phi_1 = \pm\pi/2$, $m'(t) = 0$.
- Coherent detection ($\phi_2 \approx \phi_1$) is required.
- Synchronization is important.

Derivation for the figure of merit-DSB-SC/AM:



Block diagram of DSB-SC/AM

Let, $s(t) = A_c m(t) \cos(2\pi f_c t)$

The DSB-SC signal is combined with noise and it is passed into BPF. The noise at the input of the receiver is white and Gaussian.

The noise signal $n(t)$ is given by,

$$n(t) = n_I(t) \cos(2\pi f_c t) - n_Q(t) \sin(2\pi f_c t)$$

The signal power at the channel is given by,

$$P_{SC} = \frac{A_c^2 P_m}{2}$$

Where, P_m is the power of the message signal.

The noise power at channel is given by,

$$P_{NC} = P_n$$

Where, P_n is the power of the noise signal, having bandwidth W for DSB-SC.

$$P_n = \int_{-W}^W \delta_N(f) df$$

For white noise,

$$\delta_N(f) = \frac{N_0}{2}$$

$$P_n = \int_{-W}^W \frac{N_0}{2} df = \frac{N_0}{2} \int_{-W}^W df = \frac{N_0}{2} [f]_{-W}^W$$

$$= \frac{N_0}{2} \cdot 2W \Rightarrow WN_0$$

The signal to noise at channel is given by,

$$\left(\frac{S}{N}\right)_c = \frac{P_{sc}}{P_{nc}} = \frac{A_c^2 P_m / 2}{WN_0} = \frac{A_c^2 P_m}{2WN_0}$$

The output of BPF is given by,

$$\begin{aligned} x(t) &= s(t) + n(t) \\ &= A_c m(t) \cos(2\pi f_c t) + n_i(t) \cos(2\pi f_c t) - n_o(t) \sin(2\pi f_c t) \end{aligned}$$

The output of product modulator is given by,

$$\begin{aligned} v(t) &= x(t) \cos(2\pi f_c t) \\ &= [A_c m(t) \cos(2\pi f_c t) + n_i(t) \cos(2\pi f_c t) - n_o(t) \sin(2\pi f_c t)] \cos(2\pi f_c t) \\ &= [A_c m(t) \cos^2(2\pi f_c t) + n_i(t) \cos^2(2\pi f_c t) - n_o(t) \sin(2\pi f_c t)] \cos(2\pi f_c t) \\ &= [A_c m(t) + n_i(t)] \cos^2(2\pi f_c t) - n_o(t) \sin(2\pi f_c t) \cos(2\pi f_c t) \\ &= [A_c m(t) + n_i(t)] \left(\frac{1 + \cos 4\pi f_c t}{2} - n_o(t) \sin \frac{4\pi f_c t}{2} \right) \end{aligned}$$

The output of LPF is given by,

$$\begin{aligned} y(t) &= \frac{A_c m(t) + n_i(t)}{2} \\ &= \frac{1}{2} A_c m(t) + \frac{1}{2} n_i(t) \end{aligned}$$

Signal power at the output is given by,

$$P_{SO} = \frac{A_c^2 P_m}{4}$$

Noise power at the output is given by,

$$\begin{aligned} P_n &= \int_{-2W}^{2W} S_N(f) df \\ P_n &= \int_{-2W}^{2W} \frac{N_0}{2} df = \frac{N_0}{2} (f) \Big|_{-2W}^{2W} = \frac{N_0}{2} [4W] = 2N_0W \\ P_{NO} &= \frac{P_n}{4} = \frac{2N_0W}{4} = \frac{N_0W}{2} \end{aligned}$$

Signal to noise at the output is given by,

$$(S/N)_0 = \frac{P_{SO}}{P_{NO}} = \frac{A_c^2 P_m / 4}{N_0 W / 2}$$
$$= \frac{A_c^2 P_m}{2N_0 W}$$

$$= \frac{(S/N)_O}{(S/N)_C} = \frac{A_c^2 P_m}{2N_0 W} = 1$$

Figure of merit

Thus, figure of merit = 1 for DSB-SC/AM system.

1.3 Hilbert Transform, Pre-envelope & complex envelope

$\hat{x}(t)$ of a signal $x(t)$ is defined by the equation,

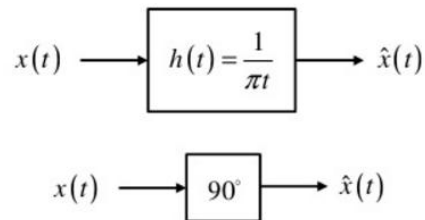
$$\hat{x}(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(s)}{t-s} ds$$

Where the integral is the Cauchy principal value integral.

$$x(t) = -\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\hat{x}(s)}{t-s} ds$$

The reconstruction formula defines the Hilbert inverse transform.

Hilbert transform:



Block diagram of Hilbert transform pair

The pair $x(t)$, $\hat{x}(t)$ is called as Hilbert transform pair, which is an LTI system whose transfer function is $H(v) = -j \cdot \text{sgn } v$, because $\hat{x}(t) = (1/\pi) * x(t)$ which by taking Fourier transform implies, $\hat{X}(v) = -j (\text{sgn } v) X(v)$

Hilbert transformer produces a -90° phase shift for positive frequency components of the input $x(t)$, the amplitude does not change.

Properties of Hilbert transform:

A signal $x(t)$ and its Hilbert transform $\hat{x}(t)$ have,

1. The same auto-correlation function.
2. The same amplitude spectrum.
3. The Hilbert transform of $\hat{x}(t)$ is $-x(t)$.

4. $x(t)$ and $\hat{x}(t)$ are orthogonal.

Pre-envelope:

The pre-envelope of a real signal $x(t)$ is a complex function.

$$x_+(t) = x(t) + j \hat{x}(t)$$

The pre-envelope is useful in treating band pass signals and systems. This is due to the result,

$$X_+(v) = \begin{cases} 2 X(v), & v > 0 \\ X(0), & v = 0 \\ 0, & v < 0 \end{cases}$$

Complex envelope:

The complex envelope of a band pass signal $x(t)$ is given by,

$$\tilde{x}(t) = x_+(t) e^{-j2\pi f_c t}$$

1.4 SSB and VSB - comparison

Single side band amplitude modulation (SSB-AM):

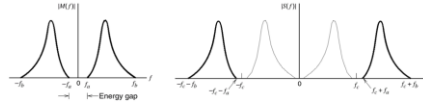
In linear AM, the message is contained in both the upper and lower side band, but the transfer of only one side band is sufficient. This saves the bandwidth in the transfer system. Additionally, the demodulation of SSB-AM is simpler than DSB-AM without carrier.

Generation of SSB-AM:

For the generation of SSB-AM signal, DSB-AM is used, where one of the side bands of the modulated signal is filtered out. Since the filters are available only with a finite edge steepness, SSB-AM can be implemented only for the signal having a lower cut-off frequency not equal to zero. This is the case with speech signals, where the frequency range spans from $0.3 \text{ kHz} < f < 3.4 \text{ kHz}$.

For the possible generation of an SSB modulated signal, the message spectrum must have an energy gap centered at the origin.

Different filter methods can be used for the suppression of unwanted side band. If a Nyquist filter is used instead of a filter with very good edge steepness, the modulation method is called as vestigial side band AM. This enables the transfer of signals with only a slightly higher bandwidth than with SSB-AM. The advantage is that, it can also carry DC voltage signals like TV video signal.



(a) Spectrum of a message signal $m(t)$ with an energy gap centered around the origin (b) Spectrum of SSB signal containing the upper side band

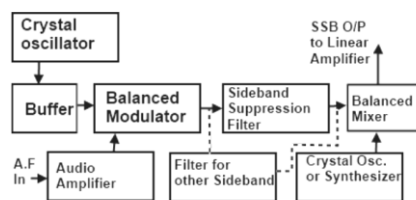
SSB modulation can also be performed by Hilbert transformer, which flips the signal phase by -90° for positive frequencies and $+90^\circ$ for negative frequencies.

SSB transmission:

There are two methods used for SSB transmission.

- Filter method
- Phase shift method

Filter method:



Filter method

A crystal controlled master oscillator produces a stable carrier frequency f_c . This carrier frequency is then fed to the balanced modulator through the buffer amplifier which isolates these two stages.

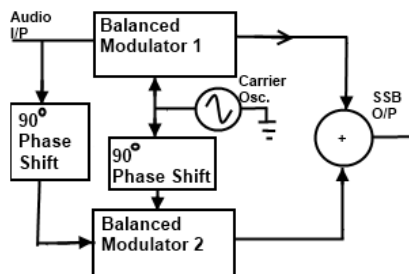
The audio signal from the modulating amplifier modulates the carrier in the balanced modulator. Audio frequency range is from 300 to 2800 Hz. The carrier is also suppressed but allows both the side bands to pass through.

This side band is then heterodyned in the balanced mixer stage with 12 MHz frequency, which is produced by the crystal oscillator or synthesizer, depending upon the requirements of our transmission. So in the mixer stage, the frequency of the crystal oscillator or synthesizer is added to SSB signal. Thus, the output frequency is raised to a value desired for transmission.

Then this band is amplified in the driver and power amplifier stages and then fed to the aerial for transmission.

Phase shift method:

The phasing method of SSB generation uses a phase shift technique that causes one of the side bands to be cancelled out.



Phase shift method

It uses two balanced modulators. The balanced modulators effectively eliminate the carrier. The carrier oscillator is applied directly to the upper balanced modulator along with the audio modulating signal. Then both the carrier and the modulating signal are shifted in phase by 90° and applied to the second lower balanced modulator.

The two balanced modulator outputs are then added together algebraically. The phase shifting action causes one side band to be cancelled out when the two balanced modulator outputs are combined together.

Advantages:

1. It allows better management of the frequency spectrum. More transmission can fit into the given frequency range than it would be possible with double side band DSB signals.
2. All the transmitted power is message power. None is dissipated as carrier power.

Disadvantages:

1. The cost of a single side band receiver is higher than the double side band DSB counterpart, be a ratio of about 3:1.
2. Single side band receivers require several precise frequency control settings to minimize the distortion and may require continual readjustment during the use of the system.

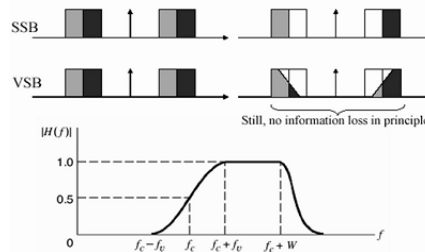
Vestigial Side Band Modulation (VSB):

The following are the drawbacks of SSB signal generation.

- Generation of an SSB signal is difficult.
- Selective filtering should be done to get the original signal back.
- Phase shifter should be tuned exactly to 90° .

To overcome these drawbacks, VSB modulation is used.

Instead of transmitting only one side band as SSB, VSB modulation transmits a partially suppressed side band and a vestige of the other side band.



VSB signal

Advantages:

- VSB is a form of amplitude modulation, intended to save bandwidth over a regular AM. Portions of one of the redundant side bands are removed to form a vestigial side band signal.
- The actual information is transmitted in the side bands, rather than the carrier, both side bands carry the same information, because LSB and USB are essentially the mirror images of each other, one can be discarded or used for the second channel or for diagnostic purposes.

Disadvantages:

- VSB transmission is very similar to SSB transmission, in which one of the side bands is completely removed. However, in VSB transmission, the second side band is not completely removed, but is filtered to remove all but the desired range of frequencies.

Comparison between SSB and VSB:

| Parameter | SSB | VSB |
|-------------------|---------------------|-------------------|
| Power | Less | High |
| Bandwidth | f_m | $f_m < Bw < 2f_m$ |
| Modulating inputs | 1 | 1 |
| Use for | Radio communication | Television |

| | | |
|-------------------------|--|--|
| Carrier suppression | Complete | Not complete |
| Side band suppression | One side band completely | One side band suppressed partially |
| Transmission efficiency | Maximum | Moderate |
| Advantages | <ol style="list-style-type: none"> Better management of the frequency spectrum. Low power consumption. | <ol style="list-style-type: none"> It is a compromise between DSB and SSB. Therefore it is easier to generate than SSB-SC. |
| Disadvantages | <ol style="list-style-type: none"> The generation of exact SSB is difficult. Complex detection. | <ol style="list-style-type: none"> Demodulation system is still complex. Its bandwidth is about 25% greater than SSB-SC. |
| Applications | <ol style="list-style-type: none"> Two way radio. Frequency division multiplexing. | <ol style="list-style-type: none"> Analog TV broadcast systems. |

1.5 Superheterodyne Receiver

To heterodyne means to mix. Heterodyne reception stands for radio reception after converting the modulated carrier voltage into modulated voltage at a different carrier frequency. Thus, the heterodyning process involves a simple change or translation of carrier frequency.

This, change in carrier frequency is achieved by mixing or heterodyning the modulated carrier voltage with a locally generated high frequency voltage in a non-linear device to obtain at the output a modulated carrier voltage at different carrier frequency.

Superheterodyne reception is a form of heterodyne reception in which frequency conversion takes place one or more times before the modulated carrier voltage is fed to the detector to recover the original modulation frequency voltage. The receiver in which frequency conversion takes place twice before detection is called as double superheterodyne receiver or triple detection receiver.

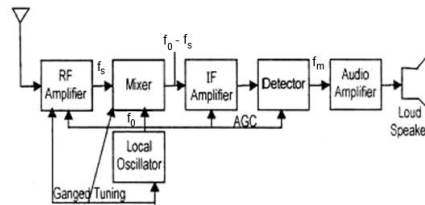
In a simple superheterodyne receiver, the modulated signal of carrier frequency f_s is fed to the mixer to which the voltage of frequency f_0 generated by the local oscillator is also fed. As a result,

output of the mixer is a voltage of frequency f_i , which is the difference of f_o and f_s . This difference is called as intermediate frequency (IF).

The IF voltage obtained at the output of the mixer is exactly similar to the modulated carrier voltage, except for change in carrier frequency.

Constituent stages of superheterodyne receiver:

The figure below gives the block diagram of superheterodyne receiver.



Block diagram of superheterodyne receiver

Functions of different stages are given below.

(i) Antenna (or) aerial:

It intercepts EM waves, voltages induced in the antenna are communicated to the receiver input by means of feeder wire or lead-in-wire. A parallel tuned circuit at the input of the receiver responds only to voltage at desired carrier frequency and rejects voltages at all other frequencies. The voltage thus picked up is fed to the input of the RF amplifier.

(ii) RF amplifier:

This is a small signal voltage amplifier which is tuned to a desired carrier frequency. The main functions of RF amplifier stage are,

- a) To amplify the input voltage to a suitably high level before feeding it to the frequency mixer which contributes large noise. Thus, SNR is improved.
- b) To provide selectivity against image frequency signal and intermediate frequency signal.

(iii) Frequency converter or mixer:

This consists of a local oscillator and frequency mixer. To the frequency mixer both the local oscillator voltage as well as signal voltage are fed. The mixer being a non-linear device, produces at its output the various intermodulation terms.

The difference frequency voltage is picked up by the tuned circuit in the output circuit of the mixer. This difference in frequency is called as intermediate frequency, which is a constant value for the receiver. Only one transistor can function as both local oscillator and frequency mixer.

(iv) IF amplifier stage:

It consists of two or more stages of fixed frequency tuned voltage amplifier having a 3 dB bandwidth of 10 kHz for AM broadcast. This IF amplifier provides most of the receiver amplification and selectivity.

(v) Second detector:

Output of the last IF amplifier stage is fed to this second detector which is generally a linear diode detector. Output of the detector is original modulated frequency voltage for linear detection, the carrier voltage fed to it should be at least 1 volt, for the weakest signal desired to be received by the receiver.

(vi) Audio frequency amplifier:

Audio frequency output from the second detector is fed to the AF amplifier which provides additional amplification. Usually one stage of audio voltage amplifier is used followed by one or more stages of audio power amplifier.

(vii) Loudspeaker:

Amplified audio output voltage is fed to the loudspeaker through impedance matching transformer. The loudspeaker reproduces the original programme.

Application:

- Radio broadcasting.
- TV broadcasting.
- Garage door open keyless remote.
- Transmits TV signal.
- Short wave radio communication.
- Two way radio communication.



ANGLE MODULATION

Phase and frequency modulation - Narrow Band and Wide band FM - Spectrum - FM modulation and demodulation - FM Discriminator - PLL as FM Demodulator - Transmission bandwidth.

2.1 Phase and frequency modulation

Angle modulation:

Angle modulation is a method of analog modulation in which either the phase or frequency of the carrier wave is varied according to the message signal. In this method of modulation the amplitude of the carrier wave is maintained constant.

In general form, an angle modulated signal can be represented as,

$$s(t) = A_c \cos[\theta(t)] \dots \dots \dots (1)$$

Where,

A_c = Amplitude of the carrier wave

$\theta(t)$ = Angle of the modulated carrier and also the function of message signal

The instantaneous frequency of the angle modulated signal, $s(t)$ is given by,

$$f_i(t) = \frac{1}{2\pi} \frac{d\theta(t)}{dt} \dots \dots \dots (2)$$

The modulated signal, $s(t)$ is normally considered as a rotating phasor of length A_c and angle $\theta(t)$. The angular velocity of such a phasor is $d\theta(t)/dt$, measured in radians per second.

An unmodulated carrier has an angle $\theta(t)$ defined by,

$$\theta(t) = 2\pi f_c t + \phi_c \dots \dots \dots (3)$$

Where,

f_c = Carrier signal frequency

ϕ_c = Value of $\theta(t)$ at $t = 0$

The angle modulated signal has an angle $\theta(t)$ defined by,

$$\theta(t) = 2\pi f_c t + \phi(t) \dots \dots \dots (4)$$

Methods of angle modulation:

There are two commonly used methods of angle modulation. They are,

1. Frequency modulation

2. Phase modulation

Phase Modulation (PM):

In phase modulation, the angle is varied linearly with message signal $m(t)$ as,

$$\theta(t) = 2\pi f_c t + k_p m(t) \dots \dots \dots (5)$$

Where, k_p is the phase sensitivity of the modulator in radians per volt.

Thus, the phase modulated signal is defined as,

$$s(t) = A_c \cos [2\pi f_c t + k_p m(t)] \dots \dots \dots (6)$$

Frequency Modulation (FM):

In frequency modulation, the instantaneous frequency $f_i(t)$ is varied linearly with message signal $m(t)$ as,

$$f_i(t) = f_c + k_f m(t) \dots \dots \dots (7)$$

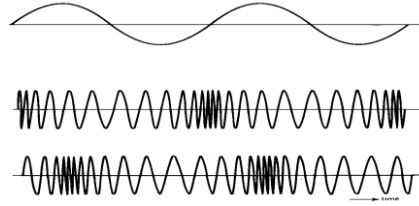
Where, k_f is the frequency sensitivity of the modulator in hertz per volt.

The instantaneous angle can be defined as,

$$\theta(t) = 2\pi f_c t + 2\pi k_f \int_0^t m(t) dt \dots \dots \dots (8)$$

Thus, the frequency modulated signal is given by,

$$s(t) = A_c \cos \left[2\pi f_c t + 2\pi k_f \int_0^t m(t) dt \right] \dots \dots \dots (9)$$



PM and FM waveforms with message signal

Problem:

1. Let us determine the instantaneous frequency of the following waveforms,

(a) $S_1(t) = A_c \cos[100\pi t + 0.25\pi]$

(b) $S_2(t) = A_c \cos[100\pi t + \sin(20\pi t)]$

(c) $S_3(t) = A_c \cos[100\pi t + (\pi t^2)]$

Solution:

Given:

(a) $S_1(t) = A_c \cos[100\pi t + 0.25\pi]$

(b) $S_2(t) = A_c \cos[100\pi t + \sin(20\pi t)]$

(c) $S_3(t) = A_c \cos[100\pi t + (\pi t^2)]$

Using equations (1) and (2),

(a) $f_i(t) = 50$ Hz, instantaneous frequency is constant.

(b) $f_i(t) = 50 + 10 \cos(20\pi t)$, maximum value is 60 Hz and minimum value is 40 Hz.

Hence, instantaneous frequency oscillates between 40 Hz and 60 Hz.

(c) $f_i(t) = (50 + t)$

The instantaneous frequency is 50 Hz at $t = 0$ and varies linearly at 1 Hz/sec.

Single tone frequency modulation:

Consider a sinusoidal modulating signal.

$$m(t) = A_m \cos(2\pi f_m t) \dots \dots (10)$$

Substituting equation (10) in (7), the instantaneous frequency of the FM signal is given by,

$$f_i(t) = f_c + k_f A_m \cos(2\pi f_m t) = f_c + \Delta f \cos(2\pi f_m t) \dots \dots (11)$$

Where, Δf is called the frequency deviation which is given by $\Delta f = k_f A_m \dots \dots (11a)$

The instantaneous angle is given by,

$$\begin{aligned} \theta(t) &= 2\pi \int_0^t f_i(t) dt \\ &= 2\pi f_c t + \frac{\Delta f}{f_m} \sin(2\pi f_m t) \end{aligned}$$

$$= 2\pi f_c t + \beta \sin(2\pi f_m t)$$

Where, $\beta = \frac{\Delta f}{f_m}$, modulation index

The resultant FM signal is given by,

$$s(t) = A_c \cos[2\pi f_c t + \beta \sin(2\pi f_m t)] \dots \dots (12)$$

The frequency deviation factor indicates the amount of frequency change in the FM signal from the carrier frequency f_c on either side of it. Thus, FM signal will have the frequency components between $(f_c - \Delta f)$ to $(f_c + \Delta f)$.

The modulation index, β indicates the phase deviation of the FM signal and is measured in radians. Depending on the value of β , FM signal can be classified into two types.

1. Narrow band FM ($\beta \ll 1$)
2. Wide band FM ($\beta \gg 1$)

Problem:

1. A sinusoidal wave of amplitude 10 volts and frequency 1 kHz is applied to an FM generator that has a frequency sensitivity constant of 40 Hz/volt. Let us determine the frequency deviation and modulation index.

Solution:

Given:

Message signal amplitude, $A_m = 10$ volts

Frequency $f_m = 1000$ Hz

Frequency sensitivity, $k_f = 40$ Hz/volt

Frequency deviation, $\Delta f = k_f A_m = 400$ Hz

Modulation index,

$\beta = \Delta f/f_m = 0.4$ (indicates a narrow band FM)

Bessel's function:

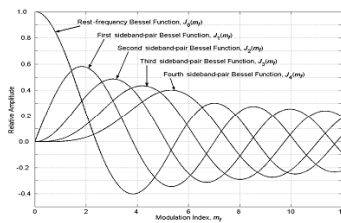
Bessel's function is an useful function to represent FM wave spectrum. Some of the useful properties of Bessel functions are given below.

(a) $J_n(\beta) = (-1)^n J_{-n}(\beta)$ for all n

(b) $J_{n+1}(\beta) + J_{n-1}(\beta) = \frac{2n}{\beta} J_n(\beta)$

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

(d) For smaller values of β , $J_0(\beta) \cong 1$, $J_1(\beta) \cong \frac{\beta}{2}$ and $J_n(\beta) \cong 0$, for $n > 2$



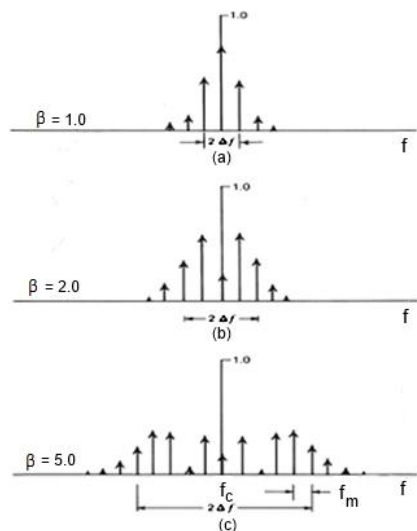
Plots of Bessel functions

Values of Bessel function co-efficients:

| Modulation Index | Sideband | | | | | | | | | | | | | | | | | |
|------------------|----------|-------|-------|-------|-------|-------|-------|-------|------|------|------|------|------|------|------|------|------|--|
| | Carrier | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | |
| 0.00 | 1.00 | | | | | | | | | | | | | | | | | |
| 0.25 | 0.98 | 0.12 | | | | | | | | | | | | | | | | |
| 0.5 | 0.94 | 0.24 | 0.03 | | | | | | | | | | | | | | | |
| 1.0 | 0.77 | 0.44 | 0.11 | 0.02 | | | | | | | | | | | | | | |
| 1.5 | 0.51 | 0.56 | 0.23 | 0.06 | 0.01 | | | | | | | | | | | | | |
| 2.0 | 0.22 | 0.58 | 0.35 | 0.13 | 0.03 | | | | | | | | | | | | | |
| 2.41 | 0 | 0.52 | 0.43 | 0.20 | 0.06 | 0.02 | | | | | | | | | | | | |
| 2.5 | -0.05 | 0.50 | 0.45 | 0.22 | 0.07 | 0.02 | 0.01 | | | | | | | | | | | |
| 3.0 | -0.26 | 0.34 | 0.49 | 0.31 | 0.13 | 0.04 | 0.01 | | | | | | | | | | | |
| 4.0 | -0.40 | -0.07 | 0.36 | 0.43 | 0.28 | 0.13 | 0.05 | 0.02 | | | | | | | | | | |
| 5.0 | -0.18 | -0.33 | 0.05 | 0.36 | 0.39 | 0.26 | 0.13 | 0.05 | 0.02 | | | | | | | | | |
| 5.53 | 0 | -0.34 | -0.13 | 0.25 | 0.40 | 0.32 | 0.19 | 0.09 | 0.03 | 0.01 | | | | | | | | |
| 6.0 | 0.15 | -0.28 | -0.24 | 0.11 | 0.36 | 0.36 | 0.25 | 0.13 | 0.06 | 0.02 | | | | | | | | |
| 7.0 | 0.30 | 0.09 | -0.30 | -0.17 | 0.16 | 0.35 | 0.34 | 0.23 | 0.13 | 0.06 | 0.02 | | | | | | | |
| 8.0 | 0.17 | 0.23 | -0.11 | -0.29 | -0.10 | 0.19 | 0.34 | 0.32 | 0.22 | 0.13 | 0.06 | 0.03 | | | | | | |
| 8.65 | 0 | 0.27 | 0.06 | -0.24 | -0.23 | 0.03 | 0.26 | 0.34 | 0.28 | 0.18 | 0.10 | 0.05 | 0.02 | | | | | |
| 9.0 | -0.09 | 0.25 | 0.14 | -0.18 | -0.27 | -0.06 | 0.20 | 0.33 | 0.31 | 0.21 | 0.12 | 0.06 | 0.03 | 0.01 | | | | |
| 10.0 | -0.25 | 0.04 | 0.25 | 0.06 | -0.22 | -0.23 | -0.01 | 0.22 | 0.32 | 0.29 | 0.21 | 0.12 | 0.06 | 0.03 | 0.01 | | | |
| 12.0 | 0.05 | -0.22 | -0.08 | 0.20 | 0.18 | -0.07 | -0.24 | -0.17 | 0.05 | 0.23 | 0.30 | 0.27 | 0.20 | 0.12 | 0.07 | 0.03 | 0.01 | |

The spectrum of FM signals for three different values of β are shown in the below figure. In this spectrum the amplitude of the carrier component is kept as a unity constant. The variation in the amplitudes of all the frequency components is indicated.

For $\beta = 1$, the amplitude of the carrier component is more than the side band frequencies as shown in figure (a). The amplitude level of the side band frequencies is decreasing. The dominant components are $(f_c + f_m)$ and $(f_c + 2f_m)$. The amplitude of the frequency components $(f_c + nf_m)$ for $n > 2$ are negligible.



Plots of spectrum for different values of modulation index

For $\beta = 2$, the amplitude of the carrier component is considered as unity. The spectrum is shown in above figure. The amplitude level of the side band frequencies is varying. The amplitude levels of the components $(f_c + f_m)$ and $(f_c + 2f_m)$ are more than the carrier frequency component, whereas the amplitude of the component $(f_c + 3f_m)$ is lower than the carrier amplitude. The amplitude of frequency components $(f_c + nf_m)$ for $n > 3$ are negligible.

The spectrum for $\beta = 5$, the amplitude of the carrier component is considered as unity. The amplitude level of the side band frequencies are varying. The amplitude levels of the components $(f_c + f_m)$, $(f_c + 3f_m)$, $(f_c + 4f_m)$ and $(f_c + 5f_m)$ are more than the carrier frequency component, whereas the amplitude of the component $(f_c + 2f_m)$ is lower than the carrier amplitude. The amplitude of frequency components $(f_c + nf_m)$ for $n > 8$ are negligible.

Problem:

1. A carrier wave is frequency modulated using a sinusoidal signal of frequency f_m and amplitude A_m . In a certain experiment conducted with $f_m = 1$ kHz and increasing A_m , starting from zero, it is found that the carrier component of FM wave is reduced to zero for the first time when $A_m = 2$ volts. Let us determine the frequency sensitivity of the modulator and also the value of A_m for which the carrier component is reduced to zero for the second time.

Solution:

Given:

The carrier component will be zero when its co-efficient, $J_0(\beta)$ is zero.

From the table,

$$J_0(x) = 0 \text{ for } x = 2.44, 5.53, 8.65$$

$$\beta = \Delta f/f_m = k_f A_m/f_m \text{ and } k_f = \beta f_m/A_m = (2.40)(1000)/2 = 1.22 \text{ kHz/V}$$

Frequency sensitivity,

$$k_f = 1.22 \text{ kHz/V}$$

The carrier component will become zero for the second time when $\beta = 5.53$.

Therefore,

$$A_m = \beta f_m/k_f = 5.53(1000)/1220 = 4.53 \text{ volts}$$

2.2 Narrow Band and Wide band FM - Spectrum

Frequency spectrum of FM:

The expression for phase is given by the integral,

$$\phi(t) = \int \omega(t) dt = \int (\omega_c + k_{osc} A \cos(\omega_a t)) dt = \omega_c t + \frac{k_{osc} A}{\omega_a} \sin(\omega_a t) + \phi_0$$

Using the above equation (neglecting ϕ_0) and using the usual expansion for $\cos(a + b)$, the complete FM signal for single tone modulation becomes,

$$V(t) = \cos(\phi(t)) = \cos(\phi_m \sin(\omega_a t)) \cos(\omega_c t) - \sin(\phi_m \sin(\omega_a t)) \sin(\omega_c t)$$

Where, $\phi_m = k_{osc} A/\omega_a$ is the modulation index.

Narrow band FM:

Narrow band FM is defined as an FM transmission, where the value of β is small enough that the terms in Bessel expansion, (i.e.) side bands are negligible. For this to be the case the modulation index must be less than 0.5.

Narrow band FM is often used for short distance communications using vehicle mount radios or hand carried equipment. Here the narrow band means that the audio or data bandwidth is small, but it is acceptable for this type of communication.

In NBFM $\beta \ll 1$, therefore $s(t)$ reduces as follows.

$$s(t) = A_c \cos(2\pi f_c t + \beta \sin(2\pi f_m t))$$

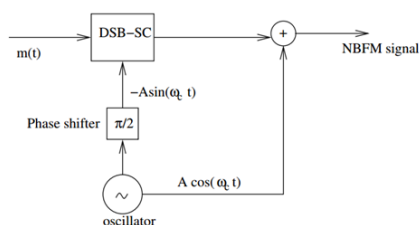
$$= A_c \cos(2\pi f_c t) \cos(\beta \sin(2\pi f_m t)) - A_c \sin(2\pi f_c t) \sin(\beta \sin(2\pi f_m t))$$

Since, β is very small, the above equation reduces to,

$$s(t) = A_c \cos(2\pi f_c t) - A_c \beta \sin(2\pi f_m t) \sin(2\pi f_c t)$$

The above equation is similar to AM. Hence, for NBFM the bandwidth is same as that of AM, (i.e.) $2 \times$ message bandwidth ($2 \times B$).

NBFM signal is generated as shown in figure below.

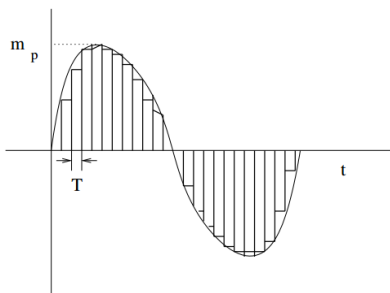


Generation of narrow band FM signal

Wide band FM:

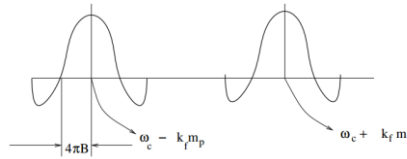
In wide band FM the modulation index is above 0.5. Under these circumstances, the side bands beyond the first two terms are significant. Broadcast FM stations use wide band FM and using this mode they are able to take advantage of the wide bandwidth available to transmit high quality audio as well as services like stereo channel and possibly other services on a single carrier.

Theoretically, a WBFM signal has infinite bandwidth. Spectrum calculation of WBFM signal is a tedious process. Let $m(t)$ be band limited to B Hz and sampled adequately at $2B$ Hz. If time period $T = 1/2B$ is too small, the signal can be approximated by a sequence of pulses as shown in below figure.



Approximation of message signal

If tone modulation is considered and peak amplitude of the sinusoid is m_p , the minimum and maximum frequency deviations will be $\omega_c - k_f m_p$ and $\omega_c + k_f m_p$ respectively. The spread of pulses in frequency domain will be $2\pi/T = 4\pi B$.



Bandwidth calculation of WBFM:

Therefore, total bandwidth is $2k_f m_p + 8\pi B$ and if frequency deviation is considered,

$$BW_{fm} = \frac{1}{2\pi}(2k_f m_p + 8\pi B)$$

$$BW_{fm} = 2(\Delta f + 2B)$$

The bandwidth obtained is higher than the actual value. This is due to the staircase approximation of $m(t)$. The bandwidth should be readjusted. For NBFM, k_f is very small and hence Δf is very small compared to B . This implies, $B_{fm} \approx 4B$.

But the bandwidth for NBFM is the same as that of AM which is $2B$. Therefore, a better bandwidth estimate is given by,

$$BW_{fm} = 2(\Delta f + B)$$

$$BW_{fm} = 2\left(\frac{k_f m_p}{2\pi} + B\right)$$

This is also called as Carson's rule.

2.3 FM modulation and demodulation

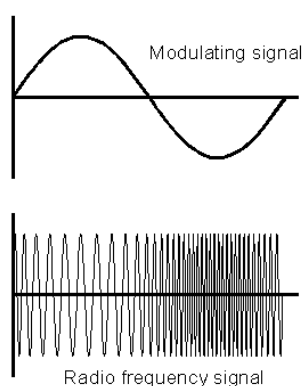
FM modulation:

Frequency modulation is widely used for radio transmission and for a wide variety of applications from broadcasting to general point to point communications.

Frequency modulation offers many advantages, particularly in mobile radio applications, where its resistance to fading and interference is the great advantage. It is also widely used for broadcasting on VHF, where it is able to provide the medium for high quality audio transmissions.

In view of its widespread use, receivers should be able to demodulate these transmissions. There is a variety of different techniques and circuits that can be used including the Foster-Seeley and ratio detectors using discrete components and where the integrated circuits are used, phase locked loop and quadrature detectors are more widely used.

It uses changes in frequency to carry the sound or other information that is required to be placed onto the carrier. As shown in the below figure, when the modulating or baseband signal voltage varies, the frequency of the signal changes in line with it.



Frequency modulating signal

Advantages:

The following advantages mean that FM has been widely used for mobile and broadcasting applications.

This type of modulation brings in several advantages with it such as,

- **Interference reduction:**

When compared to AM, FM offers a marked improvement in the interference. Most of the received noise is amplitude noise, an FM receiver can remove any amplitude sensitivity by driving the IF into limiting.

- **Removal of many effects of signal strength variations:**

FM is widely used for mobile applications, because the amplitude variations do not cause a change in the audio level. As the audio is carried by frequency variations rather than amplitude, under good signal strength conditions this does not manifest itself as a change in audio level.

- **Transmitter amplifier efficiency:**

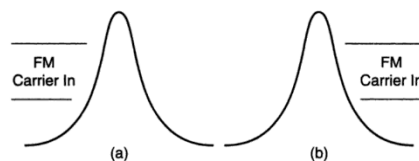
The modulation is carried by frequency variations, this means that the transmitter power amplifiers can be made non-linear. These amplifiers can be made to be far more efficient than the linear ones, thereby saving valuable battery power - a valuable commodity for mobile or portable equipment.

FM demodulation:

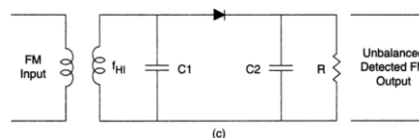
The demodulation circuits for FM are more complex than those for AM. The simplest conceptual FM demodulator would consist of an LC tuned circuit plus the diode and filter part of the AM envelope detector.

For FM demodulation, the parallel resonant circuit would be tuned above or below the frequency of the carrier, so that the incoming FM signal would ride up and down on the resonance curve. As indicated in the below figure, the higher or lower frequency FM excursions would produce more or less output voltage from the detector. The detector would consequently convert the FM to AM and then envelope detect it.

The balanced double tuned or Round-Travis slope detector is a more linear implementation of the slope detector principle. It uses two tuned circuits, one tuned slightly above the carrier frequency and one tuned slightly below the carrier frequency. The slopes of the two tuned circuits are combined to produce a more linear and more stable FM demodulation.



Response curves



Simple tuned circuit FM detector based upon the principles of (a) or (b)

The response settings for an elementary FM demodulator using only one tuned circuit: In figure (a) the resonance peak is set to a frequency slightly higher than the carrier, so that the FM will ride up and down on the left hand side of the resonance curve.

In figure (b) the resonance peak is set to a slightly lower frequency than the carrier center frequency. Therefore, the FM carrier will ride up and down on the right hand side of the resonance curve.

Note that in both the cases the slope of the curve is non-linear. The FM-to-AM-to-envelope detection process is similar to single tuned configuration.

2.3.1 FM Discriminator

There are a number of circuits that can be used to demodulate FM. Each type has its own advantages and disadvantages, some being used when receivers used discrete components and now ICs are widely used.

Below is a list of some of the main types of FM demodulator or detector. In view of the widespread use of FM, even with the competition from digital modes that are widely used today, FM demodulators are needed in many new designs of electronic equipments.

Types of FM demodulation:

(1) Indirect:

These types of demodulator use a phase locked loop (PLL) to match a local oscillator to the modulated carrier frequency.

(2) Direct:

This method employs discriminators, which are devices, that discriminate one frequency from another by transforming frequency changes into amplitude changes.

Types of FM demodulators:

Ratio detector:

This FM demodulator circuit is widely used with discrete components, providing a good level of performance. It is characterized by the transformer with three windings.

Foster-Seeley FM detector:

Like the ratio detector, the Foster Seeley detector or discriminator was used with discrete components, providing excellent performance in many FM radios.

PLL, phase locked loop FM demodulator:

FM demodulators using phase locked loops, can provide high level of performance. They do not require a costly transformer and hence can be incorporated easily within FM radio ICs.

Quadrature FM demodulator:

This form of FM demodulator is very convenient for use within the integrated circuits. It provides high levels of linearity, without the requirement of many external components.

Coincidence FM demodulator:

This form of demodulator has many similarities to the quadrature detector. It uses digital technology and replaces a mixer with a logic NAND gate.

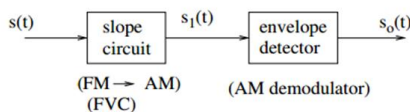
Slope detector:

This form of detector uses the slope of a tuned circuit to convert the frequency variations into amplitude variations. As the frequency of the FM signal varies, it changes its position on the slope of the tuned circuit, so that the amplitude will vary. This signal can then be converted into a baseband signal by using an AM diode detector circuit.

This circuit consists of two units such as,

- Slope circuit and,
- Envelope detector

The slope circuit converts the frequency variations in the FM signal into a voltage signal, which resembles an AM signal. The envelope circuit obtains the output signal proportional to the message signal.

**Block diagram of slope detector**

Consider an FM signal,

$$s(t) = A_c \cos \left[2\pi f_c t + 2\pi k_f \int_0^t m(\tau) d\tau \right]$$

Where, $f_i(t) = f_c + k_f m(t)$

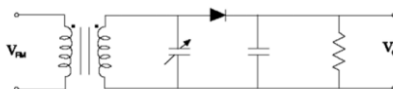
Let the slope circuit be a differentiator.

$$s_1(t) = -A_c \left[2\pi f_c + 2\pi k_f m(t) \right] \sin \left[2\pi f_c t + 2\pi k_f \int_0^t m(\tau) d\tau \right]$$

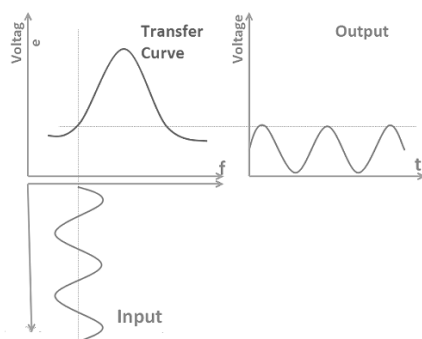
$$s_o(t) \approx -A_c \left[2\pi f_c + 2\pi k_f m(t) \right]$$

Thus, the output of the slope circuit is proportional to the message signal, $m(t)$.

A simple slope detector circuit is shown in the below figure, it consists of a slope circuit and an envelope circuit.



Simple slope detector circuit



Slope detector transfer characteristics

Major limitations:

- As it is linear in very limited frequency range, it is inefficient.
- It reacts to all the amplitude changes.
- It is relatively difficult to tune, as tuned circuit must be tuned to a different frequency rather than carrier frequency.

2.3.2 PLL as FM Demodulator

Phase Locked Loop is the best frequency demodulator. It is an electronic device with a voltage or current driven oscillator, that is constantly adjusted to match in phase with the frequency of an input signal.

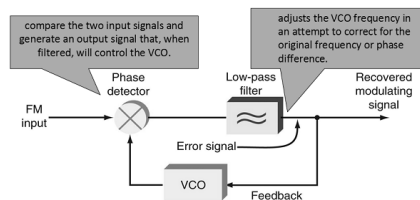
A basic Phase Locked Loop consists of three components such as,

1. Phase discriminator: It compares the phase of two signals and generates the voltage according to the phase difference of the two signals.

2. Loop filter: It is a low pass filter, which is used to filter the output of a phase discriminator.

3. Voltage Controlled Oscillator (VCO): It generates the RF signals, whose frequency depends upon the voltage generated by the phase discriminator.

PLL is a negative feedback system, which can be used for indirect frequency demodulation. This important circuit finds application in both analog and digital communication.



Phase Locked Loop FM detector

Control voltage is directly proportional to the rate of input frequency change. Hence, this signal can be directly used as output. PLL must have low time constant so that it can follow the modulating signal.

Free running frequency of VCO is set equal to the carrier frequency of the FM wave. The lock range must be at least twice the maximum deviation of the signal.

Linearity is governed by voltage to frequency characteristics of VCO. As it swings over the small portion of its bandwidth, the characteristics can be made relatively linear. Thus, the distortion level of PLL demodulators are normally very low.

Amplitude and frequency modulation: comparison

| | AM | FM |
|------------|--|--|
| Stands for | AM stands for Amplitude Modulation | FM stands for Frequency Modulation |
| Origin | AM method of audio transmission was first carried out successfully in the mid 1870s. | FM radio was developed in the United states in 1930s, mainly by Edwin Armstrong. |

| | | |
|-----------------------------------|--|--|
| Modulating differences | In AM, a radio wave known as the carrier wave is modulated in amplitude by the signal that is to be transmitted. The phase and frequency remain the same. | In FM, a radio wave known as the carrier wave is modulated in frequency by the signal that is to be transmitted. The phase and amplitude remain the same. |
| Pros and cons | AM has poorer sound quality when compared with FM, but is cheaper and can be transmitted over long distances. It has a lower bandwidth so that it can have more stations in any frequency range. | FM is less prone to interference than AM. However, FM signals are impacted by physical barriers. Due to higher bandwidth, FM has better sound quality. |
| Frequency range | AM radio ranges from 535 to 1705 KHz or up to 1200 bits per second. | FM radio ranges in a higher spectrum from 88 to 108 MHz or 1200 to 2400 bits per second. |
| Bandwidth requirements | Twice the highest modulating frequency. In AM radio broadcasting, the modulating signal has a bandwidth of 15 kHz and hence the bandwidth of an amplitude modulated signal is 30 kHz. | Twice the sum of the modulating signal frequency and the frequency deviation. If the frequency deviation is 75 kHz and the modulating signal frequency is 15 kHz, then the bandwidth required is 180 kHz. |
| Zero crossing in modulated signal | Equidistant. | Not equidistant. |
| Complexity | Transmitter and receiver are simple but synchronization is needed in case of SSBSC-AM carrier. | Both transmitter and receiver are more complex as variation of modulating signal has to be converted and detected from the corresponding variation in frequencies, (i.e. voltage to frequency and frequency to voltage conversion must be done). |

| | | |
|-------|---|---|
| Noise | AM is more susceptible to noise, because noise affects amplitude, which is where information is stored in an AM signal. | FM is less susceptible to noise, because information in FM is transmitted by varying the frequency and not the amplitude. |
|-------|---|---|

2.4 Transmission bandwidth

An FM wave contains an infinite number of side frequencies, so that the bandwidth required to transmit such a signal is similarly infinite in extent. However in practice, we find that the FM wave is effectively limited to a finite number of significant side frequencies which are compatible with a specified amount of distortion. Therefore, we may specify an effective bandwidth required for the transmission of an FM wave.

First consider the case of an FM wave generated by a single tone modulating wave of frequency f_m . In such an FM wave, the side frequencies that are separated from the carrier frequency, f_c by an amount greater than the frequency deviation, Δf decreases rapidly toward zero, so that the bandwidth always exceeds the total frequency excursion.

Specifically, for large values of modulation index β , the bandwidth approaches and is only slightly greater than the total frequency excursion $2\Delta f$. On the other hand, for small values of the modulation index β , the spectrum of FM wave is effectively limited to the carrier frequency f_c and one pair of side frequencies at $f_c \pm f_m$, so that the bandwidth approaches $2f_m$.

Thus, we may define an approximate rule for the transmission bandwidth of an FM wave, generated by a single tone modulating wave of frequency f_m as follows,

$$B_T \approx 2\Delta f + 2f_m = 2\Delta f \left(1 + \frac{1}{\beta}\right)$$

This relation is known as Carson's rule.

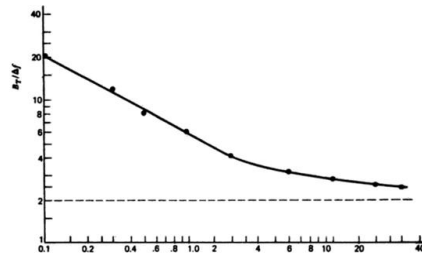
$$s_c(t) = A\{J_0(\beta)\cos(2\pi f_c t) - J_1(\beta)[\cos 2\pi(f_c - f_m)t - \cos 2\pi(f_c + f_m)t] + J_2(\beta)[\cos 2\pi(f_c - 2f_m)t + \cos 2\pi(f_c + 2f_m)t] - J_3(\beta)[\cos 2\pi(f_c - 3f_m)t - \cos 2\pi(f_c + 3f_m)t] + \dots\}$$

Where,

B = Bandwidth of the modulating signal

β = Either the phase modulation index or the frequency modulation index

From the above equation, we observe that the spectrum consists of a carrier component at f_c plus an infinite number of side band components at $f_c + nf_m$ ($n = 1, 2, \dots$). In fact, 98% of the normalized total signal power is contained in the bandwidth.



Universal curve for evaluating the 1 percent bandwidth of an FM wave

The bandwidth of the angle modulated signal with sinusoidal modulating signal depends on β and B . This is known as Carson's rule. It gives a rule-of-thumb expression and an easy way to evaluate the transmission bandwidth of the angle modulated signals. When $\beta \ll 1$, the signal is the narrow band angle modulated signal and its bandwidth is approximately equal to $2B$.



RANDOM PROCESS

Random variables, Central limit Theorem, Random Process, Stationary Processes, Mean, Correlation & Covariance functions, Power Spectral Density, Ergodic Processes, Gaussian Process, Transmission of a Random Process Through a LTI filter.

3.1 Random variables

Sample space:

Consider an experiment of throwing a coin twice. The outcomes $S = \{HH, HT, TH, TT\}$ constitute the sample space.

Random variable:

In this sample space, each of these outcomes can be associated with a number by specifying a rule of association. Such a rule of association is called as random variables.

Example: Number of heads

We denote random variable by the letter ($X, Y, \text{etc.}$) and any particular value of the random variable by x or y .

$$S = \{HH, HT, TH, TT\}$$

$$X(S) = \{2, 1, 1, 0\}$$

Thus, a random variable X can be considered as a function, that maps all the elements in the sample space S into points on the real line. The notation $X(S) = x$ means that x is the value associated with the outcome S by the random variable X .

Discrete random variables:

A discrete random variable is one, which may take on only a countable number of distinct values such as 0, 1, 2, 3, 4, If a random variable can take only a finite number of distinct values, then it must be discrete.

Examples of discrete random variables include the number of children in a family, the number of patients in a hospital, the number of defective light bulbs in a box of ten, etc.

The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values. It is also called as the probability function or the probability mass function.

Suppose a random variable X may take k different values, with the probability that $X = x_i$ defined to be $P(X = x_i) = p_i$.

The probabilities p_i must satisfy the following.

1. $0 < p_i < 1$ for each i .
2. $p_1 + p_2 + \dots + p_k = 1$.

Distribution function of the random variable X or cumulative distribution of the random variable X :

Definition:

The distribution function of a random variable X defined in $(-\infty, \infty)$ is given by,

$$F(x) = P(X \leq x) = P\{s : X(s) \leq x\}$$

Note:

Let the random variable X take the values x_1, x_2, \dots, x_n with probabilities P_1, P_2, \dots, P_n and let $x_1 < x_2 < \dots < x_n$.

Then we have,

$$F(x) = P(X < x_1) = 0, \quad -\infty < x < x_1$$

$$F(x) = P(X < x_1) = 0, \quad P(X < x_1) + P(X = x_1) = 0 + p_1 = p_1$$

$$F(x) = P(X < x_2) = 0, \quad P(X < x_1) + P(X = x_1) + P(X = x_2) = p_1 + p_2$$

$$F(x) = P(X < x_n) = P(X < x_1) + P(X = x_1) + \dots + P(X = x_n)$$

$$= p_1 + p_2 + \dots + p_n = 1$$

Properties of distribution function:

Property 1: $P(a < X \leq b) = F(b) - F(a)$, where $F(x) = P(X \leq x)$

Property 2: $P(a \leq X \leq b) = P(X = a) + F(b) - F(a)$

Property 3: $P(a < X < b) = P(a < X \leq b) - P(X = b)$
 $= F(b) - F(a) - P(X = b)$

Continuous random variables:

A continuous random variable is one, which takes an infinite number of possible values. Continuous random variables are usually measurements. Examples include height, weight, the time required to run a mile, etc.

Suppose a random variable X may take all the values over an interval of real numbers, then the probability that X is in the set of outcomes A , $P(A)$ is defined to be the area above A and under the curve.

The curve which represents a function $p(x)$, must satisfy the following:

1. The curve has no negative values ($p(x) > 0$ for all x).
2. The total area under the curve is equal to 1.

If $f(x)$ is a PDF of continuous random variable X , then the function,

$$F_X(x) = F(x) = P(X \leq x) = \int_{-\infty}^x f(x) dx$$

$$F_X(x) = F(x) = P(X \leq x) = \int_{-\infty}^x f(x) dx, \quad -\infty < x < \infty$$

Properties of CDF of a random variable X:

(i) $0 \leq F(x) \leq 1, \quad -\infty < x < \infty$

(ii) $\lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow \infty} F(x) = 1$

(iii) $P(a \leq X \leq b) = \int_a^b f(x) dx = F(b) - F(a)$

(iv) $F'(x) = \frac{dF(x)}{dx} = f(x) \geq 0$

(v) $P(X = x_i) = F(x_i) - F(x_i - 1)$

Example:

In the experiment of throwing a coin twice the sample space S is $S = \{HH, HT, TH, TT\}$. Let X be a random variable chosen, such that $X(S) = x$ (the number of heads).

Note:

Any random variable whose only possible values are 0 and 1 is called as Bernoulli random variable.

Probability density function:

Consider a continuous R.V. X specified on a certain interval (a, b) (which can also be a infinite interval $(-\infty, \infty)$)

If there is a function, $y = f(x)$ such that,

$$\lim_{\Delta x \rightarrow 0} \frac{P(x < X < x + \Delta x)}{\Delta x} = f(x)$$

Then this function $f(x)$ is termed as the probability density function or simply density function of the R.V. X .

It is also called the frequency function, distribution density or the probability density function.

The curve, $y = f(x)$ is called as the probability curve of the distribution function.

Remark:

If $f(x)$ is the PDF of the R.V. X , then the probability that a value of the R.V. X will fall in some interval (a, b) is equal to the definite integral of the function $f(x)$ in the interval a to b .

$$P(a < x < b) = \int_a^b f(x) dx$$

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

Some important definitions in terms of PDF:

If $f(x)$ is the PDF of a random variable X , which is defined in the interval (a, b) then,

| | |
|--------------------------|--------------------------------|
| Arithmetic mean | $\int_a^b x f(x) dx$ |
| Harmonic mean | $\int_a^b \frac{1}{x} f(x) dx$ |
| Geometric mean 'G' log G | $\int_a^b \log x f(x) dx$ |

| | |
|--------------------------------------|--|
| Moments about origin | $\int_a^b x^r f(x) dx$ |
| Moments about any point A | $\int_a^b (x - A)^r f(x) dx$ |
| Moment about mean μ_r | $\int_a^b (x - \text{mean})^r f(x) dx$ |
| Variance μ_2 | $\int_a^b (x - \text{mean})^2 f(x) dx$ |
| Mean deviation about the mean is M.D | $\int_a^b x - \text{mean} f(x) dx$ |

Mathematical expectations:

Definition: Let X be a continuous random variable with probability density function f(x).

Then the mathematical expectation of X is denoted by E(X) and is given by,

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

It is denoted by,

$$\mu_r' = \int_{-\infty}^{\infty} x^r f(x) dx$$

Thus,

$$\mu_1' = E(X) \text{ (}\mu_1' \text{ about origin)}$$

$$\mu_2' = E(X^2) \text{ (}\mu_2' \text{ about origin)}$$

$$\therefore \text{Mean} = \bar{x} = \mu_1' = E(X)$$

And,

$$\text{Variance} = \mu_2' - \mu_1'^2$$

$$\text{Variance} = E(X^2) - [E(X)]^2 \dots \text{(a)}$$

r^{th} moment (about mean),

Now,

$$E\{X - E(X)\}^r = \int_{-\infty}^{\infty} \{x - E(X)\}^r f(x) dx$$

$$= \int_{-\infty}^{\infty} \{x - \bar{X}\}^r f(x) dx$$

Thus,

$$\mu_r = \int_{-\infty}^{\infty} \{x - \bar{X}\}^r f(x) dx \dots \text{(b)}$$

Where, $\mu_r = E[X - E(X)]^r$

This gives the r^{th} moment about mean and it is denoted by μ_r .

Substitute $r = 1$ in equation (b), we get,

$$\mu_1 = \int_{-\infty}^{\infty} \{x - \bar{X}\} f(x) dx$$

$$= \int_{-\infty}^{\infty} x f(x) dx - \int_{-\infty}^{\infty} \bar{X} f(x) dx$$

$$= \bar{X} - \bar{X} \int_{-\infty}^{\infty} f(x) dx \quad \left[\because \int_{-\infty}^{\infty} f(x) dx = 1 \right]$$

$$= \bar{X} - \bar{X}$$

$$\mu_1 = 0$$

Substitute $r = 2$ in equation (b), we get,

$$\mu_2 = \int_{-\infty}^{\infty} (x - \bar{X})^2 f(x) dx$$

$$\text{Variance} = \mu_2 = E[X - E(X)]^2$$

Which gives the variance in terms of expectations.

Note:

Let $g(x) = K$ (Constant), then,

$$E[g(X)] = E(K) = \int_{-\infty}^{\infty} K f(x) dx$$

$$= K \int_{-\infty}^{\infty} f(x) dx \quad \left[\because \int_{-\infty}^{\infty} f(x) dx = 1 \right]$$

$$= K \cdot 1 = K$$

Thus, $E(K) = K \Rightarrow E[\text{a constant}] = \text{constant}$.

Problems:

1. A random variable X has the following probability function,

| | | | | | | | | | |
|------------------|---|----|----|----|----|-----|-----|-----|-----|
| Values of X | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Probability P(X) | a | 3a | 5a | 7a | 9a | 11a | 13a | 15a | 17a |

(i) Let us determine the value of a.

(ii) Let us determine $P(X < 3)$, $P(X \geq 3)$, $P(0 < x < 5)$.

(iii) Let us also determine the distribution function of X.

Solution:

Given:

| | | | | | | | | | |
|------------------|---|----|----|----|----|-----|-----|-----|-----|
| Values of X | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Probability P(X) | a | 3a | 5a | 7a | 9a | 11a | 13a | 15a | 17a |

(i) We know that, if $p(x)$ is the probability of mass function then,

$$\sum_{i=0}^8 p(x_i) = 1$$

$$p(0) + p(1) + p(2) + p(3) + p(4) + p(5) + p(6) + p(7) + p(8) = 1$$

$$a + 3a + 5a + 7a + 9a + 11a + 13a + 15a + 17a = 1$$

$$81a = 1$$

$$a = 1/81$$

Substitute $a = 1/81$ in above table, we get,

| | | | | | | | | | |
|--------------|------|------|------|------|------|-------|-------|-------|-------|
| X = x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| P(x) | 1/81 | 3/81 | 5/81 | 7/81 | 9/81 | 11/81 | 13/81 | 15/81 | 17/81 |

$$(ii) P(X < 3) = p(0) + p(1) + p(2)$$

$$= 1/81 + 3/81 + 5/81 = 9/81$$

$$P(X \geq 3) = 1 - p(X < 3)$$

$$= 1 - 9/81 = 72/81$$

$$P(0 < x < 5) = p(1) + p(2) + p(3) + p(4) \text{ (Here 0 and 5 are not included)}$$

$$= 3/81 + 5/81 + 7/81 + 9/81$$

(iv) To find the distribution function of X.

| X = x | F(X) = P(x ≤ x) |
|--------------|---|
| 0 | $F(0) = p(0) = 1/81$ |
| 1 | $F(1) = P(X \leq 1) = p(0) + p(1)$ $= 1/81 + 3/81 = 4/81$ |
| 2 | $F(2) = P(X \leq 2) = p(0) + p(1) + p(2)$ $= 4/81 + 5/81 = 9/81$ |
| 3 | $F(3) = P(X \leq 3) = p(0) + p(1) + p(2) + p(3)$ $= 9/81 + 7/81 = 16/81$ |

| | |
|---|---|
| 4 | $F(4) = P(X \leq 4) = p(0) + p(1) + p(2) + \dots p(4)$ $= 16/81 + 9/81 = 25/81$ |
| 5 | $F(5) = P(X \leq 5) = p(0) + p(1) + p(2) + \dots p(4) + p(5)$ $= 25/81 + 11/81 = 36/81$ |
| 6 | $F(6) = P(X \leq 6) = p(0) + p(1) + p(2) + \dots p(6)$ $= 36/81 + 13/81 = 49/81$ |
| 7 | $F(7) = P(X \leq 7) = p(0) + p(1) + p(2) + \dots + p(6) + p(7)$ $= 49/81 + 15/81 = 64/81$ |
| 8 | $F(8) = P(X \leq 8) = p(0) + p(1) + p(2) + \dots + p(7) + p(8)$ $= 64/81 + 17/81 = 81/81 = 1$ |

2. Let us determine $\text{Var}(3X + 8)$ when $\text{Var}(X) = 4$, where X is a random variable.

Solution:

Given:

$$\text{Var}(X) = 4$$

Formula to be used:

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

Substitute $a = 3$ and $\text{Var}(X) = 4$, we get,

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

$$\text{Var}(3X + 8) = 3^2(4) = 36.$$

3. X and Y are independent random variables with variance 2 and 3. Let us determine the variance of $3X + 4Y$.

Solution:

Given:

Variance of $X = 2$ and variance of $Y = 3$.

$$V(3X + 4Y) = 9\text{Var}(X) + 16\text{Var}(Y) + 24\text{Cov}(XY)$$

$$= 9 \cdot 2 + 16 \cdot 3 + 0 \quad (\because X \text{ and } Y \text{ are independent, } \text{Cov}(XY) = 0)$$

$$= 18 + 48 = 66.$$

4. The cumulative distribution function of a R.V X is given by,

$$F(x) = \begin{cases} 1 - \frac{4}{x^2}, & x > 2 \\ 0, & x \leq 2, \end{cases}$$

Let us determine,

(1) $P(x < 3)$

(2) $P(4 < X < 5)$

(3) $P(X \geq 3)$

Solution:

$$F(x) = \begin{cases} 1 - \frac{4}{x^2}, & x > 2 \\ 0, & x \leq 2, \end{cases}$$

Given:

$$(1) P(X < 3) = F(3) = 1 - \frac{4}{3^2} = 1 - \frac{4}{9} = \frac{5}{9}$$

$$(2) P(4 < X < 5) = F(5) - F(4) \\ = \left(1 - \frac{4}{25}\right) - \left(1 - \frac{4}{16}\right) \\ = \frac{9}{100} = 0.09$$

$$(3) P(X \geq 3) = 1 - P(x < 3) \\ = 1 - \frac{5}{9} = \frac{4}{9}$$

$$P(X \geq 3) = \frac{4}{9}.$$

5. Let X be a R.V. with $E(X) = 1$ and $E[X(X - 1)] = 4$. Let us determine the $\text{Var}(X/2)$ and $\text{Var}(2 - 3X)$.

Solution:**Given:**

$$E(X) = 1, E[X(X - 1)] = 4$$

$$\text{(i.e.) } E[X^2] - E[X] = 4$$

$$\Rightarrow E(X^2) - 1 = 4$$

$$\Rightarrow E(X^2) = 5$$

Formula to be used:

$$\text{Var}(X) = E(x^2) - (E(x))^2$$

$$\text{Var}(aX + b) = a^2\text{Var}(X)$$

$$\text{Var}(X) = E(x^2) - (E(x))^2$$

$$= 5 - 1 = 4$$

$$\text{Now } \text{Var}\left(\frac{X}{2}\right) = \frac{1}{4} \text{Var}(X) \text{ W.K.T}$$

$$\text{Var}(X) = 4$$

$$\text{Var}(aX + b) = a^2\text{Var}(X)$$

$$\text{Var}\left(\frac{X}{2}\right) = \frac{1}{4}(4) = 1$$

$$\text{Var}(2 - 3X) = 9\text{Var}(X)$$

$$= 9(4)$$

$$\text{Var}(2 - 3X) = 36.$$

$$f(x) = \begin{cases} x, & 0 \leq x < 1 \\ \frac{3}{2}(x-1)^2, & 1 \leq x < 2 \\ 0, & \text{otherwise.} \end{cases}$$

6. If X is a continuous R.V. with PDF

Let us determine the cumulative distribution function F(x) of X and use it to determine $P\left(\frac{3}{2} < X < \frac{5}{2}\right)$.

Solution:

$$f(x) = \begin{cases} x, & 0 \leq x < 1 \\ \frac{3}{2}(x-1)^2, & 1 \leq x < 2 \\ 0, & \text{otherwise.} \end{cases}$$

Given:**Formula to be used:**

$$F(x) = \int_{-\infty}^x f(x) dx$$

We know that, $F(x) = \int_{-\infty}^x f(x) dx$

When x lies in $0 \leq x < 1$ we get,

$$F(x) = \int_{-\infty}^x f(x) dx = \int_0^x x dx = \left(\frac{x^2}{2}\right)_0^x = \frac{x^2}{2}$$

When x lies in $1 \leq x < 2$

$$\begin{aligned} F(x) &= \int_{-\infty}^x f(x) dx = \int_{-\infty}^0 f(x) dx + \int_0^1 f(x) dx + \int_1^x f(x) dx \\ &= \int_0^1 x dx + \int_1^x \frac{3}{2}(x-1)^2 dx \end{aligned}$$

$$= \left(\frac{x^2}{2}\right)_0^1 + \frac{3}{2} \left[\frac{(x-1)^3}{3}\right]_1^x$$

$$F(x) = \frac{1}{2} + \frac{1}{2}[(x-1)^3 - (0)]$$

$$F(x) = \frac{1}{2} + \frac{(x-1)^3}{2}$$

When x lies in $x \geq 2$

$$\begin{aligned}
 F(x) &= \int_{-\infty}^0 f(x) dx + \int_0^1 f(x) dx + \int_1^2 f(x) dx + \int_2^{\infty} f(x) dx \\
 &= 0 + \int_0^1 x dx + \int_1^2 \frac{3}{2}(x-1)^2 dx + 0 \\
 &= \left(\frac{x^2}{2}\right)_0^1 + \frac{3}{2} \left(\frac{(x-1)^3}{3}\right)_1^2 + 0 \\
 &= \frac{1}{2} + \frac{1}{2}
 \end{aligned}$$

$$F(x) = 1$$

∴ Cumulative distribution function F(X) of x

$$F(X) = \begin{cases} 0 & ; x < 0 \\ \frac{x^2}{2} & ; 0 \leq x \leq 1 \\ \frac{1}{2} + \frac{(x-1)^3}{2} & ; 1 \leq x < 2 \\ 1 & ; x \geq 2 \end{cases}$$

$$\begin{aligned}
 P\left(\frac{3}{2} < X < \frac{5}{2}\right) &= F(5/2) - F(3/2) \\
 &= 1 - \left(\frac{1}{2} + \frac{(0.5)^3}{2}\right) \\
 &= 1 - 1/2 - \frac{1}{16} \\
 P\left(\frac{3}{2} < X < \frac{5}{2}\right) &= \frac{7}{16}
 \end{aligned}$$

7. Let X be a random variable with PDF $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $-\alpha_1 < x < \alpha_1$. Let us determine the PDF of the R.V, $Y = X^2$.

Solution:

Given:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, -\alpha_1 < x < \alpha_1.$$

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y)$$

$$= P(-\sqrt{y} \leq X \leq \sqrt{y}) \text{ if } y \geq 0$$

$$F_Y(y) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) \dots \mathbf{(1)}$$

$$\text{And } F_Y(y) = 0 \text{ if } y > 0$$

[Since $X^2 = Y$ has no roots, when $y > 0$].

Differentiating equation (1) with respect to y.

$$f_Y(y) = \frac{1}{2\sqrt{y}} \{f_X(\sqrt{y}) + f_X(-\sqrt{y})\}; \text{ if } y \geq 0$$

$$= 0 \text{ if } y < 0 \quad \dots (2)$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}; \quad -\infty < x < \infty$$

$$\therefore f_X(\sqrt{y}) = \frac{1}{\sqrt{2\pi}} e^{-y/2} \text{ and } f_X(-\sqrt{y}) = \frac{1}{\sqrt{2\pi}} e^{-y/2}$$

Substituting in equation (2), we get,

$$f_Y(y) = \frac{1}{2\sqrt{y}} \cdot 2 \frac{1}{\sqrt{2\pi}} e^{-y/2}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2}; \quad y > 0.$$

8. If the density function of a continuous random variable X is given by,

$$f(x) = \begin{cases} ax & ; & 0 \leq x \leq 1 \\ a & ; & 1 \leq x \leq 2 \\ 3a - ax & ; & 2 \leq x \leq 3 \\ 0 & ; & \text{otherwise} \end{cases}$$

(i) Let us determine a.

(ii) Let us also determine the PDF of X.

Solution:

$$f(x) = \begin{cases} ax & ; & 0 \leq x \leq 1 \\ a & ; & 1 \leq x \leq 2 \\ 3a - ax & ; & 2 \leq x \leq 3 \\ 0 & ; & \text{otherwise} \end{cases}$$

Given:

Formula to be used:

$$\int_{R_x} f(x) dx = 1$$

(i) Since $f(x)$ is a PDF, $\int_{R_x} f(x) dx = 1$

$$\text{i.e., } \int_0^3 f(x) dx = 1$$

$$\text{i.e., } \int_0^1 ax dx + \int_1^2 adx + \int_2^3 (3a - ax) dx = 1$$

$$\text{i.e., } 2a = 1$$

$$\therefore a = \frac{1}{2}$$

(ii) $F(x) = P(X \leq x) = 0$, when $x < 0$

$$F(x) = \int_0^x \frac{x}{2} dx = \frac{x^2}{4}, \text{ when } 0 \leq x \leq 1$$

$$= \int_0^1 \frac{x}{2} dx + \int_1^x \frac{1}{2} dx$$

$$= \frac{x}{2} - \frac{1}{4} \text{ when } 1 \leq x \leq 2$$

$$= \int_0^1 \frac{x}{2} dx + \int_1^2 \frac{1}{2} dx + \int_2^x \left(\frac{3}{2} - \frac{x}{2} \right) dx$$

$$= \frac{3}{2}x - \frac{x^2}{4} - \frac{5}{4}, \text{ when } 2 \leq x \leq 3$$

$$= 1, \text{ when } x > 3.$$

3.2 Central limit Theorem

In probability theory, the Central Limit Theorem states that, under certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and variance, will be approximately normally distributed.

The Central Limit Theorem describes the characteristics of the "population of means", which has been created from the means of an infinite number of random population samples of size (N), all of them drawn from the given "parent population".

The Central Limit Theorem predicts the following, regardless of the distribution of parent population.

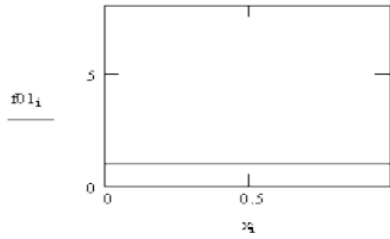
- (1) The standard deviation of the population of means is always equal to the standard deviation of the parent population, divided by the square root of the sample size (N).
- (2) The distribution of means will approximately increase the normal distribution, as the size N of the sample increases.
- (3) The mean of population of means is always equal to the mean of parent population from which the population samples were drawn.

The consequence of Central Limit Theorem is that, if we average the measurements of a particular quantity, the distribution of our average tends toward a normal one. In addition, if a measured variable is actually a combination of several other uncorrelated variables, all of them "contaminated" with the random error of any distribution, our measurement tends to be contaminated with the random error, which is normally distributed as the number of these variables increases.

Thus, the Central Limit Theorem explains the ubiquity of the famous bell-shaped "Normal distribution" in the measurements domain.

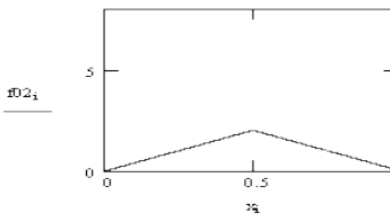
Examples:

- 1/X distribution
- Parabolic distribution
- Uniform distribution
- Triangular distribution



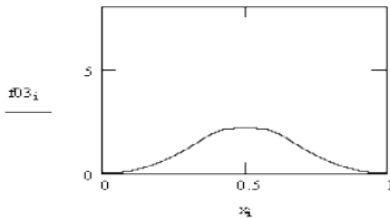
Non-normal distributions

The uniform distribution shown in the above figure is obviously non-normal. Hence, it is called as parent distribution.



Distributions of Xbar when the sample size is 2

To compute an average, \bar{X} , two samples are drawn at random, from the parent distribution and averaged. Then another sample of two is drawn and another value of \bar{X} is computed. This process is repeated, over and over and averages of two are computed. The distribution of averages of two is shown in the above figure.

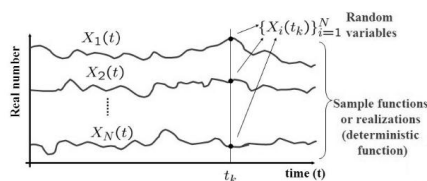


Distributions of Xbar when the sample size is 3

Repeatedly taking three from the parent distribution and computing the averages, produces the probability density as shown in the above figure.

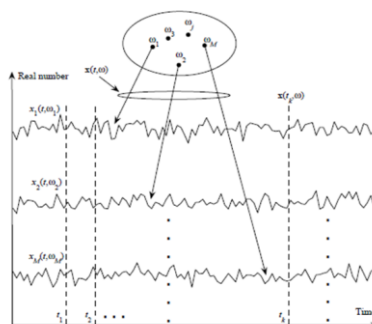
3.3 Random Process

A random process is a collection of time functions or signals, corresponding to various outcomes of a random experiment. For each outcome there exists a deterministic function which is called as a sample function or realization.



Ensemble:

- A set of possible time functions that one sees.
- This set is denoted by $x(t)$, where the time functions $x_1(t, \omega_1)$, $x_2(t, \omega_2)$, $x_3(t, \omega_3)$, . . . are specific members of the ensemble.
- At any two time instants, say t_1 and t_2 , we have two different random variables $x(t_1)$ and $x(t_2)$.
- At any time instant, $t = t_k$, we have random variable $x(t_k)$.
- The relationship between any two random variables is called Joint PDF.



Mapping from a sample space to a set of time functions

Classifications of random process:

Based on whether its statistics changes with time, the process is stationary or non-stationary.

Different levels of stationary:

- **Strictly stationary:** The joint PDF of any order is independent of a shift in time.
- **N^{th} order stationary:** The joint PDF does not depend on the time shift.

3.4 Stationary Processes

A stationary random process is a random process, $X(\zeta, t)$, whose statistics are independent of time. For a stationary random process,

$$\mu_x(t_1) = E\{x(t_1, \zeta)\} \neq f(t)$$

$$V(t) = \sigma_x^2(t_1) = E\{[x(t_1) - \mu_x(t_1)]^2\} = \sigma_x^2$$

$$R_{xx}(t, \zeta) = R_{xx}(\zeta) \neq f(t)$$

$$V(t) = R(t, 0) = V \neq f(t)$$

The statistics or expectations of a stationary random process are not necessarily equal to the time averages. However for a stationary random process, whose statistics are equal to the time averages is said to be ergodic.

A common assumption in many time series techniques is that the data is stationary. A stationary process has the property that, the autocorrelation, mean and variance structure do not change over time.

First order strictly stationary process:

Stationary Process or Strictly Stationary Process or Strict Sense Stationary Process [SSS Process]:

A random process $X(t)$ is said to be stationary in the strict sense, (i.e.) its statistical characteristics do not change with time.

Stationary Process:

Formula:

$$E[X(t)] = \text{Constant}$$

$$\gamma[X(t)] = \text{Constant}$$

Second order and wide sense stationary process:

A process is said to be second order stationary, if the second order density function is stationary.

$$f(x_1, x_2 : t_1, t_2) = f(x_1, x_2 : t_1 + \delta, t_2 + \delta), \forall x_1, x_2 \text{ and } \delta$$

If a random process $X(t)$ is WSS, then it must also be covariance stationary.

If $X(t)$ is WSS then,

$$i) E[X(t)] = \mu = a \text{ (constant) for all } t$$

ii) $R(t_1, t_2) = a$ function of $(t_1 - t_2)$

The autocorrelation function is given by,

$$C(t_1, t_2) = R(t_1, t_2) - E[X(t_1) X(t_2)]$$

$$= R(t_1 - t_2) - E[X(t_1) E[X(t_2)]]$$

$$= R(t_1 - t_2) - \mu(\mu)$$

$$= R(t_1 - t_2) - \mu^2$$

This depends only on the time difference. Hence $X(t)$ is covariance stationary.

Example:

Take some random process defined by $y(t, \zeta)$, which can be expressed as,

$$y(t, \zeta) = a \cos(\omega_0 t + \theta(\zeta))$$

$$y_i(t) = a \cos(\omega_0 t + \theta_i)$$

Where, $\theta(\zeta)$ is a random variable which lies within the interval 0 to 2π , with a constant uniform PDF such that,

$$f_\theta(\theta) = \begin{cases} 1/2\pi; & \text{for } (0 \leq \theta \leq 2\pi) \\ 0; & \text{else} \end{cases}$$

Statistical average:

The statistical mean is not a function of time.

$$E\{y(t_o, \zeta)\} = \int_0^{2\pi} \frac{1}{2\pi} a \cos(\omega_0 t_o + \theta) d\theta = 0$$

Statistical variance:

The statistical variance is also independent of time.

$$V(t_o) = R(\tau = 0) = \frac{a^2}{2}$$

Statistical correlation:

This correlation is not a function of t , where τ is a constant.

$$\begin{aligned}
 E\{y(t_o, \zeta)y(t_o + \tau, \zeta)\} &= R(t_o, \tau) \\
 &= \int_0^{2\pi} \frac{1}{2\pi} a^2 \cos(\omega_o t_o + \theta) \cos(\omega_o [t_o + \tau] + \theta) d\theta \\
 &= \frac{1}{2} a^2 \cos \omega_o \tau
 \end{aligned}$$

Since statistics are independent of time, it is a stationary process.

Problems:

1. Consider a RP $X(t) = \cos(\omega_o t + \theta)$, where θ is uniformly distributed in the interval $-\pi$ to π . Let us determine whether $X(t)$ is stationary or not and let us also determine the first and second moments of the process.

Solution:

Given:

$X(t) = \cos(\omega_o t + \theta)$, where θ is uniformly distributed in $(-\pi, \pi)$

$$f(\theta) = \frac{1}{\pi - (-\pi)} = \frac{1}{2\pi}, \quad -\pi < \theta < \pi$$

To prove:

(i) $X(t)$ is SSS process

(ii) $E[X(t)] = \text{Constant}$

(iii) $\text{Var}[X(t)] = \text{Constant}$

Formula to be used:

$$E[X(t)] = \int_{-\infty}^{\infty} X(t) f(\theta) d\theta$$

$$\text{Var}[X(t)] = E[X^2(t)] - [E[X(t)]]^2$$

$$\begin{aligned}
 E[X(t)] &= \int_{-\infty}^{\infty} X(t) f(\theta) d\theta \\
 &= \int_{-\pi}^{\pi} \cos(\omega_o t + \theta) \cdot \frac{1}{2\pi} d\theta \\
 &= \frac{1}{2\pi} [\sin(\omega_o t + \theta)]_{-\pi}^{\pi}
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{2\pi} [\sin(w_0 t + \pi) + \sin(\pi - w_0 t)] \\
 &= \frac{1}{2\pi} [-\sin w_0 t + \sin w_0 t] = 0 \\
 E[X^2(t)] &= E[\cos^2(w_0 t + \theta)] \\
 &= \frac{1}{2} E[1 + \cos(2w_0 t + 2\theta)] \\
 E[1] &= \int_{-\pi}^{\pi} \frac{1}{2\pi} d\theta = 1
 \end{aligned}$$

$$\begin{aligned}
 E[\cos(2w_0 t + 2\theta)] &= \int_{-\pi}^{\pi} \cos(2w_0 t + 2\theta) \cdot \frac{1}{2\pi} \\
 &= \frac{1}{2\pi} \left[\sin \frac{(2w_0 t + 2\theta)}{2} \right]_{-\pi}^{\pi} \\
 &= \frac{1}{4\pi} [0] = 0 \\
 \therefore E[X^2(t)] &= \frac{1}{2}(1) + 0 = \frac{1}{2} \\
 \text{Var}[X(t)] &= E[X^2(t)] - [E[X(t)]]^2 \\
 &= \frac{1}{2} - 0 \\
 &= \frac{1}{2} = \text{const}
 \end{aligned}$$

$\therefore X(t)$ is a SSS process.

2. If $X(t) = A \cos \lambda t + B \sin \lambda t$, where A and B are two independent normal random variable with $E(A) = E(B) = 0$, $E(A^2) = E(B^2) = \sigma^2$, where λ is a constant. Let us prove that $\{X(t)\}$ is a Wide Sense Stationary process of order 2.

Solution:

Given:

$$X(t) = A \cos \lambda t + B \sin \lambda t \dots (1)$$

$$E(A) = 0 = E(B) = 0 \dots (2)$$

$$E(A^2) = \sigma^2 = k, E(B^2) = \sigma^2 = k$$

$$E[AB] = E[A]E[B] \text{ [A and B are independent]}$$

$$= 0$$

To prove:

$X(t)$ is WSS process

(i.e.)

(i) $E X(t) = \text{Constant}$

(ii) $R(t_1, t_2) = \text{a function of } (t_1 - t_2)$

$$E[X(t)] = E[A \cos \lambda t + B \sin \lambda t]$$

$$= \cos \lambda t E[A] + \sin \lambda t E[B]$$

$$= 0$$

$$R(t_1, t_2) = E[X(t_1)X(t_2)]$$

$$= E[(A \cos \lambda t_1 + B \sin \lambda t_1)(A \cos \lambda t_2 + B \sin \lambda t_2)]$$

$$= E[A^2 \cos \lambda t_1 \cos \lambda t_2 + B^2 \sin \lambda t_1 \sin \lambda t_2 + AB (\cos \lambda t_1 \sin \lambda t_2 + \sin \lambda t_1 \cos \lambda t_2)]$$

$$= \cos \lambda t_1 \cos \lambda t_2 E[A^2] + \sin \lambda t_1 \sin \lambda t_2 E[B^2] + E[AB][\sin(\lambda t_1 + \lambda t_2)]$$

$$= K \cos \lambda t_1 \cos \lambda t_2 + K \sin \lambda t_1 \sin \lambda t_2 + 0$$

$$= K \cos \lambda (t_1 + t_2)$$

$$= \text{a function of } (t_1 + t_2)$$

Both the conditions are satisfied. Hence $X(t)$ is a WSS process.

3. Let us consider a random process defined by $X(t) = U \cos t + (V + 1)\sin t$, where U and V are independent random variables for which $E(U) = E(V) = 0$, $E(U^2) = E(V^2) = 1$. Let us determine whether $X(t)$ is WSS.

Solution:

Given:

$$X(t) = U \cos t + (V + 1) \sin t$$

$$E(U) = E(V) = 0$$

$$E(U^2) = E(V^2) = 1$$

$$E(UV) = E(U)E(V) = 0$$

$$E[X(t)] = E[U \cos t + (V + 1)\sin t]$$

$$= E[(U)\cos t] + E[(V)\sin t + \sin t]$$

$$= 0 + 0 + \sin t$$

$$= \sin t$$

$$\neq \text{a constant}$$

$\Rightarrow X(t)$ is not a WSS process.

4. If $X(t)$ is a wide sense stationary process with autocorrelation $R(\tau) = Ae^{-\alpha|\tau|}$, let us determine the second order moment of the random variable $X(8) - X(5)$.

Solution:

Given:

$$R(t_1, t_2) = Ae^{-\alpha|t_1 - t_2|}$$

$$R(\tau) = Ae^{-\alpha|\tau|}$$

$$E[X^2(t)] = R(t, t) = A$$

$$E[X^2(8)] = A$$

$$E[X^2(5)] = A$$

$$E[X(8)X(5)] = R|8, 5| = Ae^{-\alpha|8-5|}$$

$$= Ae^{-3\alpha}$$

The second moment of $X(8) - X(5)$ is given by,

$$E[X(8) - X(5)]^2 = E[X^2(8)] + E[X^2(5)] - 2E[X(8)X(5)]$$

$$= A + A - 2Ae^{-3\alpha}$$

$$= 2A(1 - e^{-3\alpha})$$

5. Let us show that the RP $X(t) = A \cos(w_0 t + \theta)$ is a WSS process, if A and w_0 are constants and θ is a uniformly distributed random variable in $(0, \pi)$.

Solution:

Given:

$$X(t) = A \cos(w_0 t + \theta), \text{ in } (0, \pi)$$

Formula to be used:

$$E[X(t)] = \int_{-\infty}^{\infty} X(t)f(\theta)d\theta$$

$$f(\theta) = \frac{1}{\pi - 0} = \frac{1}{\pi} \quad 0 < \theta < \pi$$

$$E[X(t)] = \int_{-\infty}^{\infty} X(t)f(\theta)d\theta$$

$$= \int_0^{\pi} A \cos(\omega_0 t + \theta) \frac{1}{\pi} d\theta$$

$$= \frac{A}{\pi} [\sin(\omega_0 t + \theta)]_0^{\pi}$$

$$= 0 = \text{Constant}$$

Therefore, $X(t)$ is a WSS process.

3.5 Mean, Correlation & Covariance functions, Power Spectral Density

In statistics, dependence is the statistical relationship between two random variables or two sets of data. Correlation refers to any of the broad class of statistical relationships involving its dependence.

Its familiar examples include the correlation between the physical statures of parents and their offspring and the correlation between the demand for product and its price. Correlations are very useful, because they indicate the predictive relationship, which is exploited in practice.

For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. In this example, there is a causal relationship, because extreme weather causes people to use more electricity for heating or cooling.

Formally, dependence refers to any situation in which the random variables do not satisfy a mathematical condition of probabilistic independence. In loose usage, correlation can refer to any departure of two or more random variables from independence, but technically it refers to any of the specialized types of relationship between mean values.

There are several correlation co-efficients which are often denoted as ρ or r , which is used for measuring the degree of correlation. The most common of these is the Pearson correlation co-efficient, which is sensitive only to a linear relationship between two variables.

Other correlation co-efficients have been developed to be more robust than the Pearson correlation co-efficient, which is more sensitive to non-linear relationships. Mutual information can also be applied to measure the dependence between two variables.

Pearson's correlation co-efficient:

The most familiar measurement of dependence between two quantities is the Pearson product-moment correlation co-efficient or Pearson's correlation co-efficient, commonly known as correlation co-efficient. It is obtained by dividing the covariance of two variables by the product of their standard deviations. Karl Pearson developed the co-efficient from a similar but slightly different idea by Francis Galton.

The population correlation co-efficient $\rho_{X,Y}$ between two random variables X and Y with the expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as,

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Where,

E = Expected value operator

cov = Covariance

corr is a widely used alternative notation for correlation co-efficient.

The Pearson correlation is defined only if both the standard deviations are finite and non-zero. It is the corollary of the Cauchy-Schwarz inequality for which the correlation cannot exceed 1 in the absolute value. The correlation co-efficient is symmetric: $\text{corr}(X, Y) = \text{corr}(Y, X)$.

The Pearson correlation is +1 in the case of a perfect direct linear relationship, -1 in the case of a perfect decreasing linear relationship and some value between -1 and +1 in all other cases, indicating the degree of linear dependence between the variables. As it approaches to zero, there is less of a relationship which is closer to the coefficient of either -1 or 1, stronger the correlation between the variables.

If the variables are independent, Pearson's correlation co-efficient is 0, but the converse is not true, because the correlation co-efficient detects only linear dependencies between two variables.

For example, suppose the random variable X is symmetrically distributed about zero and $Y = X^2$. Then Y is completely determined by X, so that X and Y are perfectly dependent, but their correlation is zero and they are uncorrelated. However, when X and Y are jointly normal, its uncorrelatedness is equivalent to independence.

If we have a series of n measurements of X and Y written as x_i and y_i where $i = 1, 2, \dots, n$, then the sample correlation co-efficient can be used to estimate the population Pearson correlation, r between X and Y.

This can also be written as,

$$r_{xy} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{(n-1) s_x s_y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

Where,

s_x and s_y = Sample standard deviations of X and Y

\bar{x} and \bar{y} = Sample means of X and Y

If x and y are the results of measurements which contains the measurement error, its realistic limits on the correlation co-efficient are not -1 to +1, but a smaller range.

Co-variance functions:

In probability theory and statistics, covariance is a measure of how much two variables change together and covariance function or kernel, describes the spatial covariance of a random variable process or field. For the random field or stochastic process $Z(x)$ on a domain D, its covariance function $C(x, y)$ gives the covariance of the values of the random field at two locations x and y .

$$C(x, y) := \text{cov}(Z(x), Z(y))$$

The same $C(x, y)$ is called as the auto covariance function in two instances such as in time series and in multivariate random fields.

Mean and variance of covariance functions:

For locations $x_1, x_2, \dots, x_N \in D$ the variance of every linear combination is given by,

$$X = \sum_{i=1}^N w_i Z(x_i)$$

$$\text{var}(X) = \sum_{i=1}^N \sum_{j=1}^N w_i C(x_i, x_j) w_j.$$

A function is a valid covariance function if and only if this variance is non-negative for all possible choices of N and weights w_1, \dots, w_N . A function with this property is called as positive definite.

Power Spectral Density:

The power spectrum is transformed when the process is filtered by a linear time invariant filter. Let $Y(t)$ be the output of a linear time invariant filter with frequency response $H(f)$ and $X(t)$ is the input signal, then the power spectral density of the output process $Y(t)$ is given as,

$$S_Y(f) = S_X(f) |H(f)|^2$$

If $\{X(t)\}$ is a stationary process with autocorrelation function $R_{XX}(t)$, then the Fourier transform of $R_{XX}(t)$ is called as the power spectral density function of $\{X(t)\}$ or simply spectral density of $X(t)$ and is given as,

$$S_{XX}(\omega) = \text{Fourier Transform of } R_{XX}(\tau)$$

$$= \int_{-\infty}^{\infty} R_{XX}(\tau) e^{-i\omega\tau} d\tau$$

Thus,

$$S_{XX}(f) = \int_{-\infty}^{\infty} R_{XX}(\tau) e^{-i2\pi f\tau} d\tau$$

Power Spectral Density and Autocorrelation:

The energy spectral density and autocorrelation function of energy signals are important tools for the characterization of energy signals. Power signals are infinite in time. It remains finite as time t approaches infinity and so their energy is infinite.

Power signals are very important in telecommunications, since they include,

- Periodic signals
- Channel white noise

Most transmitted signals are either analog or digital signals.

Therefore for power signals, it is desirable to have a counter part of the energy spectral density and autocorrelation function of energy signals. They are called as power spectral density and autocorrelation function of power signals.

In time domain, we define average power as,

$$P_x = \lim_{T_0 \rightarrow +\infty} \frac{1}{2T_0} \int_{-T_0}^{+T_0} |x(t)|^2 dt$$

Which is finite but non-zero.

In frequency domain, the description of power might not be possible because the signal is infinite and may not have Fourier transform.

The simplest way to define the PSD is by assuming that our infinite duration signal is the limit for a proper finite duration signal, (i.e.),

$$x(t) = \lim_{T_0 \rightarrow +\infty} x_{T_0}(t)$$

with $x_{T_0}(t)$ given by,

$$x_{T_0}(t) = \begin{cases} x(t), & |t| < T_0 \\ 0, & \text{otherwise} \end{cases}$$

So, the above equation for average power becomes,

$$P_x = \lim_{T_0 \rightarrow +\infty} \frac{1}{2T_0} \int_{-\infty}^{+\infty} |x_{T_0}(t)|^2 dt$$

From the Rayleigh theorem,

$$\int_{-\infty}^{+\infty} |x_{T_0}(t)|^2 dt = \int_{-\infty}^{+\infty} |X_{T_0}(f)|^2 df$$

With,

$X_{T_0}(f) = \int_{-\infty}^{+\infty} x_{T_0}(t) \exp(-j2\pi ft) dt$ denoting the Fourier transform of $x_{T_0}(t)$. Note that $x_{T_0}(t)$ is finite and has Fourier transform. Now, the average power can be written in terms of frequency as,

$$P_x = \lim_{T_0 \rightarrow +\infty} \frac{1}{2T_0} \int_{-\infty}^{+\infty} |X_{T_0}(f)|^2 df$$

The convergence of the integral will allow us to include the limit inside the integrand and therefore, the PSD of $x(t)$ is given as,

$$S_x(f) = \lim_{T_0 \rightarrow +\infty} \frac{1}{2T_0} |X_{T_0}(f)|^2$$

In general, the energy of the signal $x_{T_0}(t)$ associated to a power signal grows approximately linearly with T_0 . Therefore, the limit associated to PSD exists and is finite.

However, this approach for computing the PSD has two difficulties.

- For most power signals it is difficult to obtain $X_{T_0}(f)$ and the limit associated to PSD. This is especially true for random power signals.
- The limit can be infinite for some power signals. For instance, if $x_{T_0}(t) = A$ then $X_{T_0}(f) = AT_0 \text{sinc}(fT_0)$, which means that $S_x(0) = +\infty$.

To overcome these difficulties, we can define an autocorrelation function of power signals and then relate it with PSD. Essentially, we can define the PSD of $x(t)$ as,

$$S_x(f) = \int_{-\infty}^{+\infty} R_x(\tau) \exp(-j2\pi f\tau) d\tau$$

Where, $R_x(\tau)$ denotes the autocorrelation of $x(t)$ and is defined as,

$$R_x(\tau) = \lim_{T_0 \rightarrow +\infty} \frac{1}{2T_0} \int_{-T_0}^{T_0} x(t)x^*(t-\tau) dt$$

This means that the PSD of $x(t)$ is the Fourier transform of its autocorrelation.

The average power of $x(t)$ is given by,

$$P_x = \int_{-\infty}^{+\infty} S_x(f) df$$

From the autocorrelation definition,

$$P_x = R_x(0)$$

The average power, PSD and autocorrelation have the following properties:

1. $P_x = \int_{-\infty}^{+\infty} S_x(f) df$

2. $P_x = R_x(0)$

3. The power of the signal associated to the band $[f_1, f_2]$ is given by,

$$\int_{f_1}^{f_2} S_x(f) df$$

4. $S_x(f) = \mathcal{F}\{R_x(\tau)\} = \int_{-\infty}^{+\infty} R_x(\tau) \exp(-j2\pi f\tau) d\tau$

5. $R_x(\tau) = \mathcal{F}^{-1}\{S_x(f)\} = \int_{-\infty}^{+\infty} S_x(f) \exp(j2\pi f\tau) df$

6. $|R_x(\tau)| \leq R_x(0)$

7. $S_x(f) \geq 0$

8. $R_x(-\tau) = R_x^*(\tau)$

9. For real signal, $R_x(-\tau) = R_x(\tau)$ and $S_x(-f) = S_x(f)$

10. If $x(t)$ is submitted to a filter with impulse response $h(t)$ and frequency response $H(f)$, then the resulting signal $y(t) = x(t) * h(t)$ has PSD,

$$S_y(f) = S_x(f) |H(f)|^2$$

Its autocorrelation is expressed as,

$$R_y(\tau) = R_x(\tau) * h(\tau) * h^*(-\tau)$$

Properties of power density spectrum:

The properties of power density spectrum $S_{xx}(\omega)$ for a WSS random process $X(t)$ are given as,

- (1) $S_{xx}(\omega) \geq 0$

Proof:

The expected value of a non-negative function $E[|X_T(\omega)|^2]$ is always non-negative.

Therefore, $S_{xx}(\omega) \geq 0$ hence proved.

(2) The power spectral density at zero frequency is equal to the area under the curve of the autocorrelation $R_{xx}(\tau)$, (i.e.) $S_{XX}(0) = \int_{-\infty}^{\infty} R_{XX}(\tau) d\tau$

Proof:

$$S_{XX}(0) = \int_{-\infty}^{\infty} R_{XX}(\tau) d\tau$$

We know that, $S_{XX}(\omega) = \int_{-\infty}^{\infty} R_{XX}(\tau) e^{-j\omega\tau} d\tau$ at $\omega=0$.

Hence proved .

(3) The power density spectrum of a real process $X(t)$ is an even function, (i.e.) $S_{xx}(-\omega) = S_{xx}(\omega)$.

Proof:

Consider a WSS real process $X(t)$, then,

$$S_{XX}(\omega) = \int_{-\infty}^{\infty} R_{XX}(\tau) e^{-j\omega\tau} d\tau \text{ also } S_{XX}(-\omega) = \int_{-\infty}^{\infty} R_{XX}(\tau) e^{j\omega\tau} d\tau$$

Substitute $\tau = -\tau$ then,

$$S_{XX}(-\omega) = \int_{-\infty}^{\infty} R_{XX}(-\tau) e^{-j\omega\tau} d\tau$$

Since $X(t)$ is real, we know that, $R_{xx}(-\tau) = R_{xx}(\tau)$.

Therefore, $S_{XX}(-\omega) = \int_{-\infty}^{\infty} R_{XX}(\tau) e^{j\omega\tau} d\tau$

$S_{xx}(-\omega) = S_{xx}(\omega)$ hence proved.

(4) $S_{xx}(\omega)$ is always a real function.

Proof:

We know that,
$$S_{XX}(\omega) = \lim_{T \rightarrow \infty} \frac{E[|X_T(\omega)|^2]}{2T}$$

Since the function $|X_T(\omega)|^2$ is a real function, $S_{XX}(\omega)$ is always a real function hence proved.

(5) If $S_{XX}(\omega)$ is a PSD of a WSS random process $X(t)$, then,

$\frac{1}{2\pi} \int_{-\infty}^{\infty} S_{XX}(\omega) d\omega = A \{E[X^2(t)]\} = R_{XX}(0)$ or the time average of the mean square value of a WSS random process equals the area under the curve of the power spectral density.

Proof:

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} S_{XX}(\omega) d\omega \text{ at } \tau=0,$$

We know that, $R_{XX}(\tau) = A \{E[X(t+\tau)X(t)]\}$

$$R_{XX}(0) = A \{E[X^2(t)]\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_{XX}(\omega) d\omega = \text{Area under the curve of the power spectral density.}$$

Hence proved.

(6) If $X(t)$ is a WSS random process with PSD $S_{XX}(\omega)$, then the PSD of the derivative of $X(t)$ is equal to ω^2 times the PSD $S_{XX}(\omega)$.

That is, $S_{\dot{X}\dot{X}}(\omega) = \omega^2 S_{XX}(\omega)$.

Proof:

$$X_T(\omega) = \int_{-T}^T X(t) e^{-j\omega t} dt$$

We know that
$$S_{XX}(\omega) = \lim_{T \rightarrow \infty} \frac{E[|X_T(\omega)|^2]}{2T} \text{ and}$$

$$\begin{aligned}
 &= \dot{X}_T(\omega) = \frac{d X_T(\omega)}{dt} = \frac{d}{dt} \int_{-T}^T X(t) e^{-j\omega t} dt \\
 &= \int_{-T}^T X(t) \frac{d}{dt} e^{-j\omega t} dt \\
 &= \int_{-T}^T X(t) (-j\omega) e^{-j\omega t} dt \\
 &= (-j\omega) \int_{-T}^T X(t) e^{-j\omega t} dt
 \end{aligned}$$

Therefore, $S_{\dot{X}\dot{X}}(\omega) = \lim_{T \rightarrow \infty} \frac{E[|\dot{X}_T(\omega)|^2]}{2T}$

$$\begin{aligned}
 &= \lim_{T \rightarrow \infty} \frac{E[|(-j\omega)X_T(\omega)|^2]}{2T} \\
 &= \lim_{T \rightarrow \infty} \frac{\omega^2 E[|(-j\omega)X_T(\omega)|^2]}{2T} \\
 S_{\dot{X}\dot{X}}(\omega) &= \omega^2 \lim_{T \rightarrow \infty} \frac{E[|X_T(\omega)|^2]}{2T}
 \end{aligned}$$

Therefore, $S_{\dot{X}\dot{X}}(\omega) = \omega^2 S_{XX}(\omega)$ hence proved.

Problems :

1. If $R(\tau) = e^{-2\lambda|\tau|}$ is the autocorrelation function of a random process $\{X(t)\}$, then let us determine the spectral density of $\{X(t)\}$.

Solution:

Given:

$$R(\tau) = e^{-2\lambda|\tau|}$$

Formula to be used:

$$S(\omega) = \int_{-\infty}^{\infty} R(\tau) e^{-j\omega\tau} d\tau$$

$$\begin{aligned}
 S(\omega) &= \int_{-\infty}^{\infty} R(\tau) e^{-i\omega\tau} d\tau \\
 &= \int_{-\infty}^{\infty} e^{-2\lambda|\tau|} (\cos\omega\tau - i\sin\omega\tau) d\tau \\
 &= 2 \int_0^{\infty} e^{-2\lambda\tau} \cos\omega\tau d\tau \\
 &= \left[\frac{2e^{-2\lambda\tau}}{4\lambda^2 + \omega^2} (-2\lambda\cos\omega\tau + \omega\sin\omega\tau) \right]_0^{\infty} \\
 S(\omega) &= \frac{4\lambda}{4\lambda^2 + \omega^2}
 \end{aligned}$$

2. A wide sense stationary noise process $N(t)$ has an autocorrelation function $R_{NN}(\tau) = Pe^{-3|\tau|}$, where P is a constant. Let us calculate its power spectrum.

Solution:

Given:

$$R_{NN}(\tau) = Pe^{-3|\tau|}$$

$$S_{NN}(\omega) = \int_{-\infty}^{\infty} e^{-j\omega\tau} R_{NN}(\tau) d\tau$$

Formula to be used:

$$S_{NN}(\omega) = P \int_{-\infty}^{\infty} e^{-3|\tau|} (\cos\omega\tau - j\sin\omega\tau) d\tau$$

$$= P \int_{-\infty}^{\infty} e^{-3|\tau|} \cos\omega\tau d\tau - Pj \int_{-\infty}^{\infty} e^{-3|\tau|} \sin\omega\tau d\tau$$

$$= 2P \int_0^{\infty} e^{-3\tau} \cos\omega\tau d\tau - 0 \quad (\text{By properties of defining integral})$$

$$= 2P \left[\frac{e^{-3\tau}}{9 + \omega^2} (-3\cos\omega\tau + \omega\sin\omega\tau) \right]_0^{\infty}$$

$$S_{NN}(\omega) = \frac{6P}{9 + \omega^2}$$

3. Let us determine the power spectral density of a random signal with autocorrelation function $e^{-\lambda|\tau|}$.

Solution:

Given:

$$e^{-\lambda|\tau|}$$

$$S(\omega) = \int_{-\infty}^{\infty} R(\tau) e^{-i\omega\tau} d\tau$$

Formula to be used:

$$\begin{aligned} S(\omega) &= \int_{-\infty}^{\infty} R(\tau) e^{-i\omega\tau} d\tau \\ &= \int_{-\infty}^{\infty} e^{-\lambda|\tau|} (\cos \omega\tau - i \sin \omega\tau) d\tau \\ &= 2 \int_0^{\infty} e^{-\lambda\tau} \cos \omega\tau d\tau \end{aligned}$$

$$\begin{aligned} &= 2 \left[\frac{e^{-\lambda\tau}}{\lambda^2 + \omega^2} (-\lambda \cos \omega\tau + \omega \sin \omega\tau) \right]_0^{\infty} \\ &= 2 \left[0 - \frac{1}{\lambda^2 + \omega^2} (-\lambda) \right] = \frac{2\lambda}{\lambda^2 + \omega^2} \end{aligned}$$

3.6 Ergodic Processes

In the event that the distributions and statistics are not available, we can avail ourselves of the time averages from the particular sample function. The mean of the sample function $X\lambda_0(t)$ is referred to as the sample mean of the process $X(t)$ and is defined as,

$$\langle \mu(X)T \rangle = \left(\frac{1}{T} \right) \int_{-T/2}^{T/2} X\lambda_0(t) dt$$

This quantity is actually a random variable by itself, because its value depend on the parameter sample function. The sample variance of the random process is defined as,

$$\langle \sigma^2(X)T \rangle = \left(\frac{1}{T} \right) \int_{-T/2}^{T/2} |X\lambda_0(t) - \langle \mu(X)T \rangle|^2 dt$$

The time averaged sample ACF is obtained via the relation,

$$\langle R_{XX} \rangle T = \left(\frac{1}{T} \right) \int_{-T/2}^{T/2} x(t) * x(t - T) dt$$

In general, these quantities are not the same as the ensemble averages. A random process $X(t)$ is said to be ergodic in the mean, (i.e.) first order ergodic, if the mean of the sample average asymptotically approaches the ensemble mean.

$$\begin{aligned} \lim_{T \rightarrow \infty} E\{\langle \mu(X)T \rangle\} &= \mu_X(t) \\ \lim_{T \rightarrow \infty} \text{var}\{\langle \mu(X)T \rangle\} &= 0 \end{aligned}$$

Similarly, a random process $X(t)$ is said to be ergodic in the ACF, (i.e) second order ergodic if,

$$\lim_{T \rightarrow \infty} E\{\langle R_{XX}(\tau) \rangle\} = R_{XX}(\tau)$$

$$\lim_{T \rightarrow \infty} \text{var}\{\langle R_{XX}(\tau) \rangle\} = 0$$

The concept of ergodicity is also significant from a measurement perspective, because in practical situations, we do not have access to all the sample realizations of a random process.

Ergodic processes are signals for which the measurements based on a single sample function are sufficient to determine the ensemble statistics. Random signal for which this property does not hold are referred to as non-ergodic processes.

3.7 Gaussian Process

A random process $X(t)$ is a Gaussian process, if for all n and all (t_1, t_2, \dots, t_n) , the random variables have a jointly Gaussian density function. For Gaussian processes, the knowledge of mean and autocorrelation, (i.e.) $m_x(t)$ and $R_x(t_1, t_2)$ gives a complete statistical description of the process.

If the Gaussian process $X(t)$ is passed through an LTI system, then the output process $Y(t)$ will also be a Gaussian process. For Gaussian processes, strict stationary and WSS are equivalent.

A Gaussian process is a stochastic process X_t , $t \in T$, for which any finite linear combination of samples has a joint Gaussian distribution. More accurately, any linear function applied to the sample function X_t will give a normally distributed result.

Notation wise, one can write $X \sim GP(m, K)$, which means the random function X is distributed as a GP, with mean function m and covariance function K .

When the input vector t is two or multi-dimensional, Gaussian process might also be known as a Gaussian random field. The sufficient condition for ergodicity of a stationary zero mean Gaussian process $X(t)$ is that,

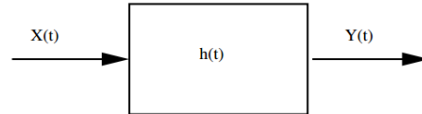
$$\int_{-\infty}^{\infty} R_X(\tau) d\tau < \infty$$

Jointly Gaussian process:

The random processes $X(t)$ and $Y(t)$ are jointly Gaussian, if for all n, m and all (t_1, t_2, \dots, t_n) and $(\tau_1, \tau_2, \dots, \tau_m)$, the random vector $(X(t_1), X(t_2), \dots, X(t_n), Y(\tau_1), Y(\tau_2), \dots, Y(\tau_m))$ is distributed according to an $n + m$ dimensional jointly Gaussian distribution. For jointly Gaussian processes, uncorrelatedness and independence are equivalent.

3.8 Transmission of a Random Process Through a LTI filter

- A random process $X(t)$ is applied as input to a linear time invariant filter of impulse response $h(t)$.
- It produces a random process $Y(t)$ at the filter output as shown in the below figure.



Transmission of a random process through a linear filter

- It is difficult to describe the probability distribution of an output random process $Y(t)$, even if the probability distribution of an input random process $X(t)$ is completely specified for $-\infty \leq t \leq +\infty$.

Mean:

The input to the above system $X(t)$ is assumed to be stationary. The mean of the output random process $Y(t)$ can be calculated as,

$$\begin{aligned}
 m_Y(t) &= E[Y(t)] = E \left[\int_{-\infty}^{+\infty} h(\tau) X(t - \tau) d\tau \right] \\
 &= \int_{-\infty}^{+\infty} h(\tau) E[X(t - \tau)] d\tau \\
 &= \int_{-\infty}^{+\infty} h(\tau) m_X(t - \tau) d\tau \\
 &= m_X \int_{-\infty}^{+\infty} h(\tau) d\tau \\
 &= m_X H(0)
 \end{aligned}$$

Where, $H(0)$ is the zero frequency response of the system.

Autocorrelation:

The autocorrelation function of the output random process $Y(t)$ can be written as,

$$R_Y(t, u) = E[Y(t)Y(u)]$$

Where, t and u denote the time instants at which the process is observed. Therefore, we may use the convolution integral to write,

$$\begin{aligned}
 R_Y(t, u) &= E \left[\int_{-\infty}^{+\infty} h(\tau_1) X(t - \tau_1) d\tau_1 \int_{-\infty}^{+\infty} h(\tau_2) X(u - \tau_2) d\tau_2 \right] \\
 &= \int_{-\infty}^{+\infty} h(\tau_1) d\tau_1 \int_{-\infty}^{+\infty} h(\tau_2) E[X(t - \tau_1)X(u - \tau_2)] d\tau_2
 \end{aligned}$$

When its input $X(t)$ is a wide stationary random process, autocorrelation function of $X(t)$ is only a function of the difference between the observation times $t - \tau_1$ and $u - \tau_2$.

Substituting $\tau = t - u$, we get,

$$R_Y(\tau) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h(\tau_1)h(\tau_2)R_X(\tau - \tau_1 + \tau_2) d\tau_1 d\tau_2$$

$$R_Y(0) = E[Y^2(t)]$$

The mean square value of the output random process $Y(t)$ is obtained by substituting $\tau = 0$ in the above equation.

$$\begin{aligned} E[Y^2(t)] &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h(\tau_1)h(\tau_2)R_X(\tau_2 - \tau_1) d\tau_1 d\tau_2 \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} H(\omega) \exp(j\omega\tau_1) d\omega \right] h(\tau_2)R_X(\tau_2 - \tau_1) d\tau_1 d\tau_2 \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} H(\omega) d\omega \int_{-\infty}^{+\infty} h(\tau_2) d\tau_2 \int_{-\infty}^{+\infty} R_X(\tau_2 - \tau_1) \exp(j2\omega\tau_1) d\tau_1 \end{aligned}$$

Substituting $\tau = \tau_2 - \tau_1$, we get,

$$\begin{aligned} E[Y^2(t)] &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} H(\omega) d\omega \int_{-\infty}^{+\infty} h(\tau_2) \exp(j\omega\tau_2) d\tau_2 \int_{-\infty}^{+\infty} R_X(\tau) \exp(-j2\omega\tau) d\tau \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} H(\omega) d\omega \int_{-\infty}^{+\infty} H^*(\omega) d\omega \int_{-\infty}^{+\infty} R_X(\tau) \exp(-j\omega\tau) d\tau \end{aligned}$$

This is the Fourier transform of the autocorrelation function $R_X(t)$ of the input random process $X(t)$. This transform is denoted by $S_X(f)$.

$$S_X(\omega) = \int_{-\infty}^{+\infty} R_X(\tau) \exp(-j\omega\tau) d\tau$$

$S_X(\omega)$ is called as the power spectrum or power spectral density of the wide sense stationary random process $X(t)$.

$$E[Y^2(t)] = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |H(\omega)|^2 S_X(\omega) d\omega$$

The mean square value of the output of a stable linear time invariant filter in response to the wide sense stationary random process is equal to the integral, over all the frequencies of the power spectral density of the input random process multiplied by the squared magnitude of the transfer function of the filter.



NOISE CHARACTERIZATION

Noise sources and types – Noise figure and noise temperature – Noise in cascaded systems. Narrow band noise – PSD of in-phase and quadrature noise – Noise performance in AM systems – Noise performance in FM systems – Pre-emphasis and de-emphasis – Capture effect, threshold effect.

4.1 Noise sources and types

Noise is a random, undesirable electrical energy that enters the communication system through the communicating medium and interferes with the transmitted message. However, some noise is also produced in the receiver.

Noise may be defined as any unwanted form of energy which tends to interfere with proper reception and reproduction of wanted signal.

Noise may be classified into two types. They are,

- **Internal noise**
- **External noise**

Internal noise can be easily evaluated mathematically and can be reduced to a great extent by proper design. External noise on the other hand cannot be reduced, except by changing the location of the receiver or the entire system.

Internal noise in communication:

Internal noise is the noise which gets generated within the receiver or communication system. Internal noise may be put into the following four categories.

- **Shot noise**
- **Thermal noise or white noise or Johnson noise**
- **Miscellaneous internal noise**
- **Transit time noise**

Shot noise:

The most common type of noise is referred to as shot noise, which is produced by the random arrival of electrons or holes at the output element, at the plate in a tube or at the collector or drain in the transistor.

Shot noise is also produced by the random movement of electrons or holes across a PN junction. Even though current flow is established by external bias voltage, there will be some random movement of electrons or holes due to the discontinuities in the device.

An example of such discontinuity is the contact between the copper lead and the semiconductor materials. The interface between the two creates discontinuity, that causes random movement of the current carriers.

Thermal noise:

Conductors contain a large number of free electrons and ions that are strongly bounded by molecular forces. The ions vibrate randomly about their normal positions, however, this vibration is a function of temperature. Continuous collisions between the electrons and vibrating ions take place. Hence, there is a continuous transfer of energy between the ions and electrons. This is the source of resistance in a conductor.

The movement of free electrons constitutes a current which is purely random in nature and over a long time, averages zero. There is a random motion of electrons which gives rise to noise voltage called as thermal noise. Hence, the noise generated in any resistance due to the random motion of electrons is called as thermal noise or white noise or Johnson noise.

The analysis of thermal noise is based on the Kinetic theory. It shows that, the temperature of particles is a way of expressing its internal kinetic energy. Therefore, temperature of a body can be said to be equivalent to the statistical RMS value of the velocity of motion of particles in the body.

At -273°C , the kinetic energy of the body becomes zero. Hence, we can relate the noise power generated by the resistor to be proportional to its absolute temperature. Noise power is also proportional to the bandwidth over which noise is measured.

$$P_n \propto TB$$

$$P_n = KTB \dots\dots (1)$$

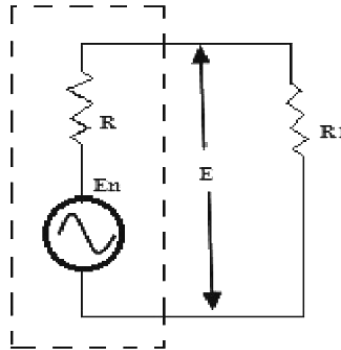
Where,

P_n = Maximum noise power output of the resistor

K = Boltzmann's constant = 1.38×10^{-23} joules/Kelvin

T = Absolute temperature

B = Bandwidth over which noise is measured



$$P_n = \frac{E^2}{R_1} = \frac{E^2}{R} = \frac{\left(\frac{E_n}{2}\right)^2}{R} = \frac{E_n^2}{4R}$$

Therefore, from above figure if $R = R_1$, $E_n = E/2$.

$$E_n^2 = 4RP_n$$

$$E_n^2 = 4RKT B$$

$$E_n = \sqrt{4KTRB} \dots\dots(2)$$

But, when the frequency of operation is high and the signal being processed is the magnitude of transit time, then problem can occur.

Miscellaneous internal noise or flicker noise:

Flicker noise or modulation noise is the one appearing in transistors operating at low audio frequencies. Flicker noise is proportional to the emitter current and junction temperature. However, it is inversely proportional to frequency. Hence, it may be neglected at frequencies above 500 Hz and therefore, possess no serious problems.

Transit time noise:

Another kind of noise that occurs in the transistor is termed as transit time noise. Transit time is the duration of time that it takes for the current carrier such as a hole or electron to move from input to the output.

The devices are very tiny, so the distances involved are minimal. Yet the time it takes for the current carriers to move even short distance is finite.

At low frequencies, this time is negligible, but when the frequency of operation is high and the signal being processed is the magnitude of the transit time, then problem can occur. Then transit

time shows up as a kind of random noise within the device and this is directly proportional to the frequency of operation.

External noises:

External noise may be classified into the following three types.

- **Extraterrestrial noise**
- **Atmospheric noise**
- **Man-made noises or industrial noise**

Atmospheric noise:

Atmospheric noise is caused by the lightning discharges in the thunderstorms and other natural electrical disturbances occurring in the atmosphere. These electrical impulses are random in nature. Thus, the energy is spread over the complete frequency spectrum used for radio communication.

Atmospheric noise consists of spurious radio signals with components spread over a wide range of frequencies. These spurious radio waves constitute the noise, get propagated over the earth in the same fashion as the desired radio waves of the same frequency.

Accordingly at a given receiving point, the receiving antenna picks up not only the signal, but also the static energy from all the thunderstorms, local or remote.

Extraterrestrial noise:

- **Cosmic noise**
- **Solar noise**

Cosmic noise:

Sun is known as the distant star and has high temperatures. Therefore, these stars radiate noise in the same way as sun. The noise received from these distant stars is thermal noise and these get distributed uniformly over the entire sky. We also receive the noise from the center of our own galaxy, other distant galaxies and from other virtual point sources such as quasars and pulsars.

Solar noise:

This is the electrical noise emanating from the sun. Under quite condition, there is a steady radiation of noise from the sun. This results, because sun is a large body which has very high temperature and radiates electrical energy in the form of noise over a wide frequency spectrum including the spectrum used for radio communication. The intensity interference make many frequencies unusable for communications. During other years, the noise is at a minimum level.

Man-made noise or industrial noise:

The electrical noise is produced by sources such as electrical motors and switch gears, automobiles and aircraft ignition, fluorescent lights, leakage from high voltage lines and numerous other heavy electrical machines. Such noises are produced by the arc discharge taking place during the operation of these machines. Such man-made noise is intensive in industrial and densely populated areas.

Man-made noise in such areas far exceeds all other sources of noise in the frequency range extending from 1 MHz to 600 MHz.

4.2 Noise figure and noise temperature - Noise in cascaded systems**Noise figure:**

Noise figure is defined as the ratio of signal to noise power at the input to the signal to noise power at the output. The device under consideration can be the entire receiver or a single amplifier stage.

The noise figure also called as the noise factor can be computed with the expression,

$F = \text{Signal to noise power input} / \text{Signal to noise power output}$.

We can express the noise figure as a number, but mostly it is expressed in decibels.

$$NF_{dB} = 10 \log(F) = 10dB$$

$$F = 10^{1.0} = 10.0 \text{ as } F = 1 + \frac{T_e}{T_0}$$

$$T_0 = 290 \text{ K}, T_e = (F - 1) \times T_0 = 9.0 \times T_0 = 2610 \text{ K}$$

$$k = 1.38 \times 10^{-23} \text{ W/HzK}, B = 2 \times 10^7 \text{ Hz}$$

$$N = kT_e B = 7.2 \times 10^{-13} \text{ Watts} \cong -121 \text{ dBW}$$

Noise temperature:

The noise temperature is a means for specifying the noise in terms of an equivalent temperature. Noise power is directly proportional to the temperature in degree Kelvin and that noise power collapses to zero at absolute zero.

Note that the equivalent noise temperature T_e is not the physical temperature of the amplifier, but rather a theoretical construct, (i.e.) an equivalent temperature which produces that amount of noise power.

Noise temperature is related to the noise factor by,

$$T_e = (F_n - 1)T_o$$

It is also related to the noise figure by,

$$T_e = \left[\text{antilog}\left(\frac{NF}{10}\right) - 1 \right] K T_o$$

We have noise temperature T_e . We can also define the noise factor and noise figure in terms of noise temperature.

$$F_n = \frac{T_e}{T_o} + 1$$

And,

$$NF = 10 \text{ LOG}\left(\frac{T_e}{T_o} + 1\right)$$

The total noise in any amplifier or network is the sum of internally and externally generated noise. In terms of noise temperature,

$$P_{n(\text{total})} = GKB(T_o + T_e)$$

Where,

K = Boltzmann's constant (1.38×10^{-23} J/°K)

$P_{n(\text{total})}$ = Total noise power

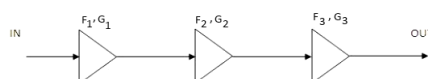
B = Network bandwidth in hertz (Hz)

$T_o = 290^\circ\text{K}$

G = Amplifier gain

Noise in cascade amplifiers:

For an amplifier, noise signal is the valid input signal. Thus in a cascade amplifier, the final stage sees an input signal which consists of the original signal and noise, amplified by each successive stage. Each stage in the cascade chain amplifies the signal and noise from the previous stages and contributes some noise of its own.



Block diagram of cascaded systems

If only loss exists in the cascade, then the cascaded noise figure equals the magnitude of the total loss. Cascaded noise figure is mostly affected by the noise figure of the components, closest to the input of the system as long as some positive gain exists in the cascade.

The overall noise factor for a cascade amplifier can be obtained from the Friis noise equation,

$$F_N = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + \dots + \frac{F_n - 1}{G_1 G_2 \dots G_n}$$

Where,

F_N = Overall noise factor of N stages in cascade

F_1 = Noise factor of stage 1

F_2 = Noise factor of stage 2

F_n = Noise factor of the n^{th} stage

G_1 = Gain of stage 1

G_2 = Gain of stage 2

G_{n-1} = Gain of stage (n - 1)

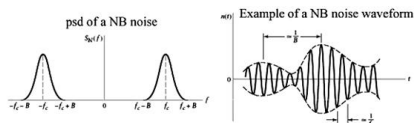
From the above equation, the noise factor of the entire cascade chain is dominated by the noise contributed by the first one or two stages. High-gain, low-noise amplifiers typically use a low noise amplifier circuit, only for the first one or two stages in the cascade chain.

4.3 Narrow band noise

A random process $X(t)$ is band pass or narrow band random process, if its power spectral density $S_X(f)$ is non-zero only in a small neighborhood of some high frequency f_c . Deterministic signals are defined by its Fourier transform. Random processes are defined by its power spectral density.

Narrow band noise representation:

In most of the communication systems, we are dealing with band pass filtering of signals. Wide band noise will be shaped into band limited noise. If the bandwidth of the band limited noise is relatively small compared to the carrier frequency, we refer this as narrow band noise.



Representation of narrow band noise

We can derive the power spectral density $G_n(f)$ and the autocorrelation function $R_{nn}(\tau)$ of the narrow band noise and use them to analyze the performance of linear systems. In practice, we often deal with mixing, which is a non-linear operation and the system analysis becomes difficult. In such a case, it is useful to express the narrow band noise as,

$$n(t) = x(t) \cos(2\pi f_c t) - y(t) \sin(2\pi f_c t)$$

Where,

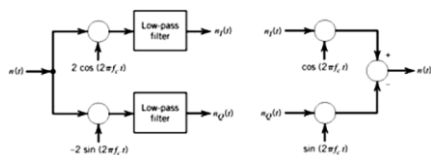
f_c = Carrier frequency within the band occupied by the noise

$x(t)$ and $y(t)$ = Quadrature components of the noise $n(t)$

The Hilbert transform of $n(t)$ is,

$$\hat{n}(t) = H[n(t)] = x(t) \sin(2\pi f_c t) + y(t) \cos(2\pi f_c t)$$

Generating $N_I(t)$ and $N_Q(t)$ from $N(t)$ and vice versa:



Generation of quadrature components of $n(t)$:

$x(t)$ and $y(t)$ have the following properties.

1. $x(t)$ and $y(t)$ have the same means and variances as $n(t)$.
2. $E[x(t) y(t)] = 0$. $x(t)$ and $y(t)$ are uncorrelated with each other.
3. $x(t)$ and $y(t)$ have identical power spectral densities, related to the power spectral density of $n(t)$ by,

$$G_x(f) = G_y(f) = G_n(f - f_c) + G_n(f + f_c) \text{ for } f_c - 0.5B < |f| < f_c + 0.5B$$

Where, B is the bandwidth of $n(t)$.

4. If $n(t)$ is Gaussian, then $x(t)$ and $y(t)$ are also Gaussian.

In-phase and quadrature components:

In-phase and quadrature sinusoidal components,

$$\sin(A + B) = \sin(A)\cos(B) + \cos(A)\sin(B)$$

From the trigonometric identity, we have,

$$x(t) \triangleq A \sin(\omega t + \phi) = A \sin(\phi + \omega t) = [A \sin(\phi)] \cos(\omega t) + [A \cos(\phi)] \sin(\omega t)$$

$$\triangleq A_1 \cos(\omega t) + A_2 \sin(\omega t)$$

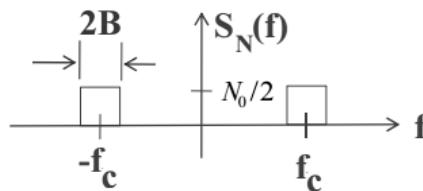
From this we may conclude that, every sinusoid can be expressed as a sum of sine function and cosine function. If the sine part is called as the in-phase component, then the cosine part can be called as the phase quadrature component.

In general, phase quadrature means 90° out of phase. It is also the case that, every sum of an in-phase and quadrature component can be expressed as a single sinusoid at some amplitude and phase. Note that they only differ by a relative phase shift of 90° .

Power spectral density:

$$S_{N_i}(f) = S_{N_q}(f) = \begin{cases} S_N(f - f_c) + S_N(f + f_c), & -B \leq f \leq B. \\ 0, & \text{otherwise.} \end{cases}$$

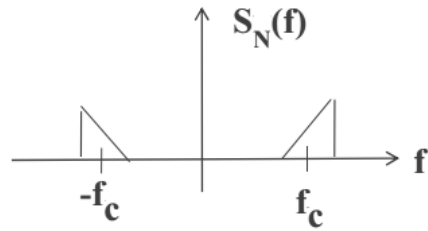
Example:



$$S_{N_i}(f) = S_{N_q}(f) = \begin{cases} S_N(f - f_c) + S_N(f + f_c), & -B \leq f \leq B. \\ 0, & \text{otherwise.} \end{cases}$$

Cautions:

1. $S_N(f)$ need not be symmetric around f_c .
2. f_c need not be at the center of $S_N(f)$.

**Some properties of narrow band noise:**

1. If $N(t)$ is stationary, $N_I(t)$ and $N_Q(t)$ are also stationary.
2. If $N(t)$ is Gaussian, $N_I(t)$ and $N_Q(t)$ are also Gaussian.
3. **Mean:** $E[N_I(t)] = E[N_Q(t)] = 0$.

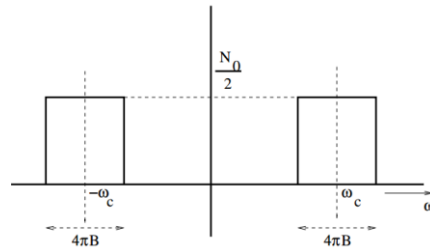
Variance:

$$\text{var} [N_I(t)] = \text{var} [N_Q(t)] = \text{var}[N(t)]$$

4.4 PSD of in-phase and quadrature noise

Power spectral density of noise:

- $N_0/2$ is defined for both positive and negative frequency.
- N_0 is the average power/(unit BW) at the front end of the receiver in AM and DSB-SC.



Band limited noise spectrum

The filtered signal available for demodulation is given by,

$$x(t) = s(t) + n(t)$$

$$n(t) = n_i(t) \cos \omega_c t - n_q(t) \sin \omega_c t$$

Where,

$$n_i(t) \cos \omega_c t = \text{In-phase component}$$

$$n_q(t) \sin \omega_c t = \text{Quadrature component}$$

$$n(t) = \text{Representation for narrow band noise}$$

Different measures are used to define the Figure of Merit of different modulators.

Input SNR:

$$(SNR)_I = \frac{\text{Average power of modulated signal } s(t)}{\text{Average power of noise}}$$

Output SNR:

$$(SNR)_O = \frac{\text{Average power of demodulated signal } s(t)}{\text{Average power of noise}}$$

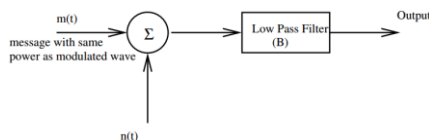
The output SNR is measured at the receiver.

• **Channel SNR:**

$$(SNR)_C = \frac{\text{Average power of modulated signal } s(t)}{\text{Average power of noise in message bandwidth}}$$

Figure of Merit (FOM) of receiver:

$$FOM = \frac{(SNR)_O}{(SNR)_C}$$



Basic channel model

To compare the SNR of different modulators, we assume that,

- The modulated signal $s(t)$ of each system has the same average power.
- Channel noise $w(t)$ has the same average power in the message bandwidth B .

4.5 Noise performance in AM systems

Noise in AM system using coherent detection:

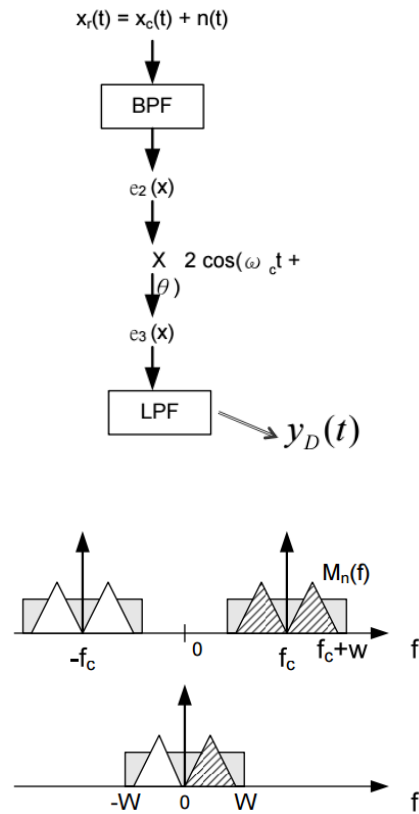
Received signal:

$$x_c(t) = A_c[1 + a m_n(t)] \cos(\omega_c t + \theta)$$

Where,

a = Modulation index

m_n = Normalized message



$$y_D(t) = A_c a m_n(t) + n_c(t) + A_c$$

A_c is removed (DC term).

1) Assume $\overline{m(t)} = 0$, thus, we can remove DC term.

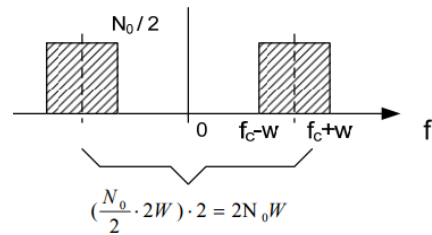
2) In practice, we cannot recover DC term of $m(t)$. Thus, we simply remove it.

Post detection signal power:

$$S_D = \overline{(A_c a m_n(t))^2} = A_c^2 a^2 \cdot \overline{m_n^2}$$

Post detection noise power:

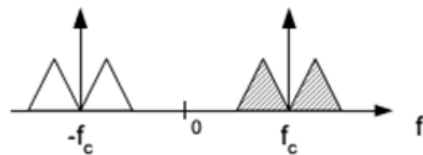
$$N_D = \overline{n_c^2} = 2N_0W$$

**Post detection SNR:**

$$(SNR)_D = \frac{A_c^2 a^2 \overline{m_n^2}}{2N_0W} \quad (\text{For AM})$$

Transmission signal power is given by,

$$\begin{aligned} S_T &= \overline{\{A_c[1 + am_n(t)]\cos(\omega_c t + \theta)\}^2} \\ &= \overline{A_c^2 \cdot \cos^2(\omega_c t + \theta)} + \overline{A_c^2 a^2 m_n^2(t) \cdot \cos^2(\omega_c t + \theta)}. \\ (\overline{m_n^2(t)} = m_n^2, \quad \cos^2(\omega_c t + \theta) &= \frac{1}{2} + \cos(2\omega_c t + \theta).) \end{aligned}$$



$$P_T = S_T = \frac{1}{2} A_c^2 + \frac{1}{2} A_c^2 a^2 \overline{m_n^2}$$

BP noise power:

$$N_T = 2N_0W$$

Pre-detection:

$$(SNR)_T = \frac{1}{2} \cdot \frac{A_c^2 + A_c^2 a^2 \overline{m_n^2}}{2N_0W}$$

$$\frac{(SNR)_D}{(SNR)_T} = \frac{A_c^2 \overline{a^2 m_n^2}}{\frac{1}{2}(A_c^2 + A_c^2 \overline{a^2 m_n^2})} = \frac{2\overline{a^2 m_n^2}}{1 + \overline{a^2 m_n^2}} < 1$$

Efficiency:

$$E_{ff} = \frac{\overline{a^2 m_n^2}}{1 + \overline{a^2 m_n^2}} \Rightarrow \frac{(SNR)_D}{(SNR)_T} = 2E_{ff}$$

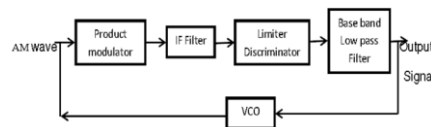
Noise in AM receivers using envelope detection:

In standard AM wave, both side bands and the carrier are transmitted. The AM wave may be written as,

$$s(t) = A_c[1 + k_a m(t)] \cos(2\pi f_c t)$$

$$n(t) = n_1(t) \cos(2\pi f_c t) - n_2(t) \sin(2\pi f_c t)$$

The phase of the composite signal $x(t)$ at the limiter-discriminator input is approximately equal to $n_2(t)/A_c$. This assumes that the carrier to noise ratio is high. The envelope of $x(t)$ is not important, because the limiter removes all variations in the envelope.



AM noise reduction

Thus, the composite signal at the discriminator input consists of a small index phase-modulated wave, with the modulation derived from the component $n_2(t)$ of noise that is in phase quadrature with the carrier. When feedback is applied, the VCO generates a wave that reduces the phase modulation index of the wave at the IF filter output, (i.e.) the quadrature component $n_2(t)$ of noise.

$$s(t) = A_c[1 + k_a m(t)] \cos(2\pi f_c t)$$

Where,

$$A_c \cos(2\pi f_c t) = \text{Carrier wave}$$

$$m(t) = \text{Message signal}$$

K_a = Sensitivity of the modulator, determines the percentage of modulation

The average power of the modulated signal is given by,

$$s(t) = (A_c^2 [1 + K_a^2 P]) / 2$$

Where, P is the average power of the message m(t).

The average noise power in the message bandwidth W is WN_0 .

Then, the channel signal to noise ratio is given by,

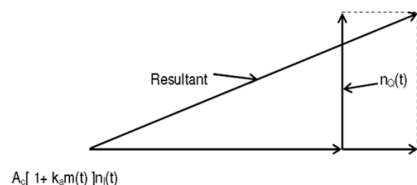
$$(\text{SNR})_c = \frac{A_c^2 [1 + K_a^2 P]}{2WN_0}$$

The received signal x(t) at the envelope detector input consists of a modulated message signal s(t) and narrow band noise n(t). Then,

$$x(t) = [A_c + A_c K_a m(t) + n_i(t)] \cos(2\pi f_c t) - n_o(t) \sin(2\pi f_c t)$$

From the below phasor diagram, the receiver output can be obtained as,

$$y(t) = \text{Envelope of } x(t)$$



Phasor diagram for x(t)

The average noise power is derived from both in-phase and quadrature components.

$$= \{[A_c + A_c K_a m(t) + n_i(t)]^2 + n_o(t)^2\}^{1/2}$$

When the average carrier power is large when compared to average noise power, (i.e.) the signal term $A_c + A_c K_a m(t)$ is large when compared with the terms $n_i(t)$ and $n_o(t)$, then, we may approximate the output y(t) as,

$$y(t) = A_c + A_c K_a m(t) + n_i(t)$$

The term A_c is DC component, therefore, it is neglected.

Signal to noise ratio:

The output signal power is, $A_c^2 K_a^2 P$

The output noise power is, $2WN_0$

The average noise power is given by,

$$\frac{1}{2}[2WN_0] + \frac{1}{2}[2WN_0] = 2WN_0$$

Therefore, the carrier to noise ratio can be written as,

$$\rho = \frac{A_c^2}{4WN_0}$$

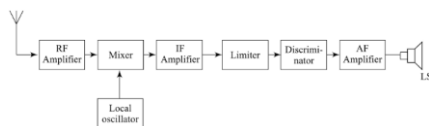
Or

$$(\text{SNR})_0 = 2\rho k_a^2 P$$

The output signal to noise ratio is given by,

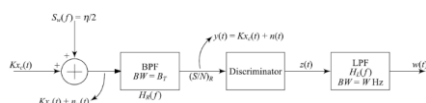
$$(\text{SNR})_0 = \frac{A_c^2 k_a^2 P}{2WN_0}$$

4.5.1 Noise performance in FM systems



Block diagram of a FM broadcast superheterodyne receiver

For the purpose of noise performance evaluation, we model the receiver as shown in below figure.



Receiver model for noise performance evaluation

Additive noise of the channel is modelled as zero mean white Gaussian noise of a two sided power spectral density, $\eta/2$. Where, K represents the channel attenuation.

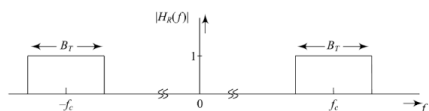
The modulated signal in this case, is given by,

$$x_c(t) = A_c \cos[\omega_c t + \phi(t)], \quad f_c = f_{i.f} \text{ of the receiver} \dots\dots(1)$$

Where,

$$\phi(t) = 2\pi k_f \int_0^t x(\alpha) d\alpha \dots\dots(2)$$

The response characteristic of an ideal BPF is shown in the below figure. This is used to represent the combined effect of both the RF and IF amplifiers.



Response characteristic of BPF

The bandwidth of this BPF is the transmission bandwidth B_T of the modulated signal, $x_c(t)$ and is also the bandwidth of the front end of the receiver. The signal at the input to the filter is $Kx_c(t) + n_w(t)$, (i.e.) the modulated signal and the additive white noise.

Its output, however is $Kx_c(t) + n(t)$, where $n(t)$ is the band pass noise, centered on f_c and is obtained by filtering the white noise using the BPF of bandwidth B_T , with center frequency $f_c = f_{i,f}$, the intermediate frequency of the superheterodyne receiver.

The FM detector, called the discriminator produces an output voltage, which at any instant is proportional to the deviation of the instantaneous frequency of the input signal from the carrier frequency.

The input signal for the discriminator is,

$$Kx_c(t) + n(t)$$

Where,

$x_c(t)$ = FM signal

$n(t)$ = Band pass noise centered on f_c

In case of amplitude modulation, the additive noise would add to the amplitude modulated signal $x_c(t)$ and thus change its envelope, which the envelope detector would extract.

So in case of AM, the additive noise directly affects that parameter of the input signal which the detector tries to extract. So the effect of additive noise is considerable in case of AM, but in case of FM, at each instant, the discriminator extracts the frequency deviation of the carrier of the input signal and produces an output voltage proportional to the instantaneous frequency deviation.

The additive noise does not directly affect the frequency deviation of the incoming FM signal. Thus, in a qualitative way, we say that FM will not be affected by the channel noise to the same extent as AM.

Since the bandwidth of the BPF is B_T and the two sided PSD of the additive white noise in the channel is $\eta/2$, the noise power entering the receiver is given by,

$$\overline{n^2(t)} = \frac{\eta}{2} \times 2B_T = \eta B_T \triangleq N_R \dots\dots(3)$$

The received signal power is equal to the average power of the component $Kx_c(t)$ of $y(t)$, the input to the discriminator. This is denoted by S_R and is given by,

$$S_R = \frac{(KA_c)^2}{2} = \frac{A_R^2}{2} \dots\dots(4)$$

Therefore, the pre-detection SNR is given by,

$$\left(\frac{S}{N}\right)_R = \frac{S_R}{N_R} = \frac{A_R^2}{2} \cdot \frac{1}{\eta B_T} = \frac{A_R^2}{2\eta B_T} \dots\dots(5)$$

We know that, $n(t)$ is the band pass noise centered on f_c and we may represent it by its inphase and quadrature components as,

$$n(t) = n_i(t)\cos \omega_c t - n_q(t) \sin \omega_c t \dots\dots(6)$$

Alternatively, we may use the envelope and phase angle representation,

$$n(t) = R_n(t)\cos[\omega_c t + \phi_n(t)] \dots\dots(7)$$

Where, $R_n(t)$, the envelope is related to $n_i(t)$ and $n_q(t)$ by,

$$R_n(t) = \sqrt{n_i^2(t) + n_q^2(t)} \dots\dots(8)$$

And is Rayleigh distributed.

The phase angle, $\phi_n(t)$ is given by,

$$\phi_n(t) = \tan^{-1} \left[\frac{n_q(t)}{n_i(t)} \right] \dots\dots(9)$$

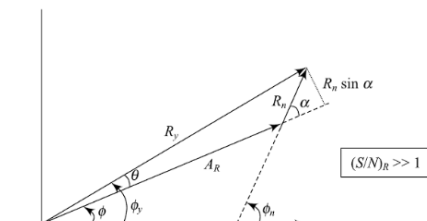
As it is more convenient to use the envelope and phase representation, we shall write $y(t)$, the input to the discriminator as,

$$\begin{aligned} y(t) &= A_R \cos[\omega_c t + \phi(t)] + n(t) \\ &= A_R \cos[\omega_c t + \phi(t)] + R_n(t) \cos[\omega_c t + \phi_n(t)] \dots\dots(10) \end{aligned}$$

We shall make use of the above equation to examine how the noise term $n(t)$ affects the angle $\phi(t)$ of the FM signal and thus, changes its frequency deviation. So, we shall proceed by making the simplifying and reasonable assumption that, the SNR at the input of the discriminator is high.

i.e., $\left(\frac{S}{N} \right)_R \gg 1 \dots\dots(11)$

Under this assumption, we draw the phasor diagram for equation (10), it will appear as shown in



the below figure.

Phasor diagram when $(S/N)_R \gg 1$

For the band pass signal $y(t)$, if $R_y(t)$ is the envelope and $\phi_y(t)$ is the phase angle, we may write $y(t)$ as,

$$y(t) = R_y(t) \cos [\omega_c t + \phi_y(t)] \dots\dots(12)$$

Since $y(t)$ is the input to the discriminator, it produces an output $z(t)$, which at any instant is proportional to the instantaneous frequency deviation given by,

$$f_i(t) = \frac{1}{2\pi} \frac{d}{dt} [\phi_y(t)] \dots\dots(13)$$

So let,

$$z(t) = \frac{1}{2\pi} \frac{d}{dt} [\phi_y(t)] \dots\dots(14)$$

The phasor diagram of above figure shows how the additive noise component $n(t)$ affects the phase angle ϕ and thereby the frequency deviation of the incoming FM signal. $\phi(t)$ is the phase angle of the received FM signal, $\phi_n(t)$ is the phase angle of the band pass noise component $n(t)$.

The sum of the phasors A_R and R_n , gives R_y , the envelope of $y(t)$, the phase angle of which is $\phi_y(t)$. Note that, because of our assumption that the pre-detection SNR is very much greater than 1,

$$P[R_n(t) \ll A_R] \text{ is almost equal to } 1 \dots\dots(15)$$

But from receiver model,

$$\sin \theta(t) = [R_n(t) \sin \alpha(t)] / R_y(t) \dots\dots(16)$$

However, from equation (15), the following small angle approximation can be made so that,

$$\sin \theta(t) \cong \theta(t)$$

And hence, equation (16) can be rewritten as,

$$\theta(t) = \frac{R_n(t) \sin \alpha(t)}{R_y(t)} \dots\dots(17)$$

Thus,

$$\phi_y(t) = \phi(t) + \theta(t) \dots\dots(18)$$

We have,

$$\phi_y(t) = \phi(t) + \frac{R_n(t) \sin \alpha(t)}{R_y(t)} \dots\dots(19)$$

But because of equation (15), we may make the following approximation.

$$R_y(t) \cong A_r \dots\dots\dots(20)$$

Hence, from equations (14) and (18), we have,

$z(t)$ = Discriminator output signal

$$= \frac{1}{2\pi} \frac{d}{dt} [\phi_y(t)] = \frac{1}{2\pi} \frac{d}{dt} \phi(t) + \frac{1}{2\pi} \frac{d}{dt} \theta(t) = k_f x(t) + n_d(t) \dots\dots\dots(21)$$

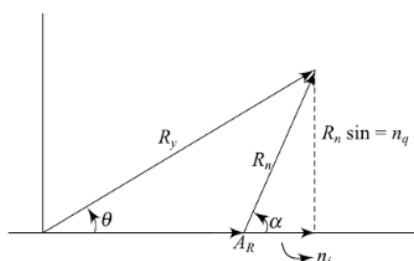
Since $\phi(t)$ is the phase angle caused due to frequency modulation of the carrier by the message signal $x(t)$, from equations (1) and (2), the first term in equation (21) clearly represents the message signal component in the output of the discriminator. Since $\theta(t)$ is the additional phase caused by noise, the second term of equation (21) represents the noise term in the output of the discriminator and is denoted by $n_d(t)$.

To see how much of this noise goes past the low pass filter and reaches the destination, we have to examine the spectrum of the noise term in equation (21). For this purpose, let us rewrite it as follows.

$$\frac{1}{2\pi} \frac{d}{dt} \theta(t) = \frac{1}{2\pi} \frac{d}{dt} \left[\frac{R_n(t) \sin \alpha(t)}{A_r} \right] \text{ (From equations (17) and 20))} \dots\dots\dots(22)$$

From the phasor diagram, we find that,

$$\alpha(t) = \phi_n(t) - \phi(t) \dots\dots\dots(23)$$



Phasor diagram with no modulation $(S/N)_r \gg 1$

This seems to indicate that the post detection noise, $n_d(t)$, is dependent on the modulation angle $\phi(t)$. Now, $\phi_n(t)$ is the phase angle of the band pass noise in its envelope, but, we know that in such a representation, the envelope is Rayleigh distributed while the phase angle $\phi_n(t)$ is uniformly distributed over $-\pi$ to $+\pi$.

If we can assume that $\alpha(t)$, which is $[\phi_n(t) - \phi(t)]$, is itself uniformly distributed over $-\pi$ to $+\pi$, then this coupling between the post detection noise and the modulation angle will be removed and $n_d(t)$ will be independent of modulation.

Such an assumption is justified provided the carrier to noise ratio is large. In that case, we may assume that there is no modulation and only an unmodulated carrier is transmitted. In such a case, the phasor diagram will appear as shown in above figure. ($\phi(t) = 0$ when there is no modulation).

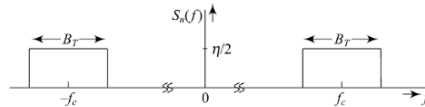
Since $\phi(t) = 0$, $\alpha(t) = \phi_n(t)$ and so,

$$R_n(t)\sin \alpha t = R_n(t) \sin \phi_n(t) = n_q(t) \dots\dots(24)$$

Hence equation (22) may be rewritten as,

$$\frac{1}{2\pi} \frac{d}{dt} \theta(t) = \frac{1}{2\pi} \frac{d}{dt} \left[\frac{R_n(t) \cdot \sin \alpha t}{A_R} \right]$$

$$n_d(t) = \frac{1}{A_R} \frac{1}{2\pi} \frac{d}{dt} [n_q(t)] \dots\dots(25)$$



Power spectral density of $n(t)$

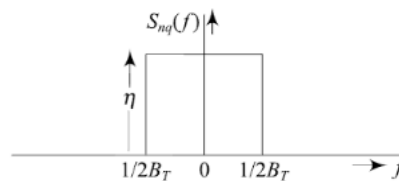
Therefore, to determine how much power of this post detection noise goes past the low pass filter with a cut-off frequency of W Hz, we have to determine the power spectrum of $n_d(t)$. To do this, we first note that $n_q(t)$ is the low pass equivalent of the band pass noise, $n(t)$, that has entered the receiver.

Since the BPF at the front end of the receiver has a transfer function of $H_R(f)$, its output, $n(t)$, will have a power spectrum of,

$$S_n(f) = S_{n_w}(f) |H_R(f)|^2 = \frac{\eta}{2} |H_R(f)|^2 \dots\dots(26)$$

The PSD of the band pass noise, $n(t)$, is shown in the above figure and its low pass equivalent, $n_q(t)$, is shown in the below figure.

$$S_{n_q}(f) = \eta \pi (f/B_T) \dots\dots(27)$$

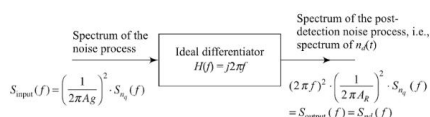


Power spectral density of $n_q(t)$, low pass equivalent of $n(t)$

The power spectrum of $\frac{1}{A_R} \cdot \frac{1}{2\pi} n_q(t)$ is then given by,

$$\frac{1}{(2\pi)^2} \cdot \frac{1}{A_R^2} \cdot S_{nq}(f) \dots\dots(28)$$

To find the power spectrum of post detection noise $n_d(t)$, in view of equation (25), we proceed as in figure below.

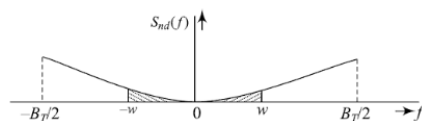


Deriving the spectrum of post detection noise

Substituting for $S_{nq}(f)$ in the expression for $S_{nd}(f)$ and simplifying, we get,

$$S_{n_d}(f) = \left(\frac{\eta f^2}{A_R^2}\right) \Pi\left(\frac{f}{B_T}\right)$$

$$= \left(\frac{\eta f^2}{2S_R}\right) \Pi\left(\frac{f}{B_T}\right) \dots\dots(29)$$



Power spectral density of post detection noise

A sketch of the post detection noise spectrum is given in the above figure. While the message has a bandwidth of only W Hz, this noise process has a bandwidth of $B_T/2$, which is much greater than W . Hence, there is considerable noise outside the message bandwidth. This out-of-band noise has to be removed using a low pass filter having a cut-off frequency of W Hz.

The average power of the noise at the output of the low pass filter = N_D = Destination noise power = Area of the shaded region

$$= \int_{-W}^W \left(\frac{\eta f^2}{2S_R}\right) \Pi\left(\frac{f}{B_T}\right) df$$

$$= \int_{-W}^W \frac{\eta f^2}{2S_R} df = \frac{\eta W^3}{3S_R}$$

$$N_D = \frac{\eta W^3}{3S_R} \dots\dots(30)$$

The message signal component at the output of the discriminator has been found to be $k_f x(t)$. Since this has a bandwidth of W , all of it passes through the low pass filter.

Hence, the destination signal power is given by,

$$S_D = k_f^2 \overline{x^2(t)} \dots\dots(31)$$

∴ Destination signal to noise ratio is given by,

$$\begin{aligned} \left(\frac{S}{N}\right)_D &= \left(\frac{S_D}{N_D}\right) = \frac{k_f^2 \overline{x^2(t)}}{(\eta W^3 / 3S_R)} \\ &= 3 \left(\frac{k_f}{W}\right)^2 \overline{x^2(t)} \left(\frac{S_R}{\eta W}\right) \end{aligned}$$

But, we know that when $x(t)$ is the normalized message signal, k_f denotes the peak frequency deviation. Since we have k_f over W as a factor in the above expression for the destination SNR, let us replace that factor by the deviation ratio denoted by D .

$$\left(\frac{S}{N}\right)_{FM} = 3D^2 \overline{x^2(t)} \gamma \dots\dots(32)$$

Hence, the figure of merit for FM systems may be written as,

$$\text{Figure of Merit}_{(FM)} = \frac{(S/N)_D}{(S/N)_C} = \frac{3D^2 \overline{x^2} \gamma}{\gamma} = 3D^2 \overline{x^2} \dots\dots(33)$$

4.6 Pre-emphasis and de-emphasis

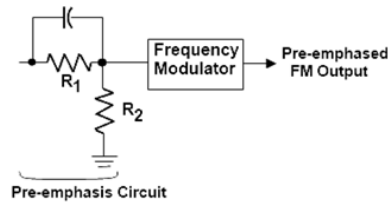
Pre-emphasis refers to boosting the relative amplitudes of the modulating voltage for higher audio frequencies from 2 to approximately 15 KHz.

Pre-emphasis circuit:

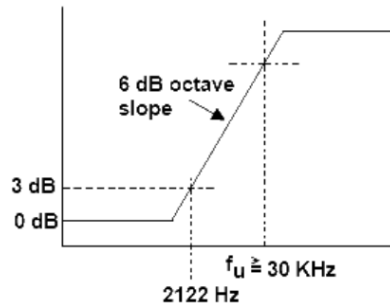
At the transmitter, the modulating signal is passed through a simple network which amplifies the high frequency components more than the low frequency components. The simplest form of such a circuit is a simple high pass filter as shown in figure (a). Specification controls a time constant of 75 μ s, where $t = RC$.

Any combination of resistor and capacitor giving this time constant will be satisfactory. Such a circuit has a cut-off frequency f_c of 2122 Hz. This means that frequencies higher than 2122 Hz will be linearly enhanced. The output amplitude increases with frequency at a rate of 6 dB per octave. The pre-emphasis curve is shown in figure (b).

This pre-emphasis circuit increases the energy content of the high frequency signals, so that they will tend to become stronger than the high frequency noise components. This improves the signal to noise ratio and increases intelligibility and fidelity.



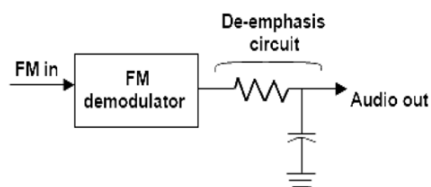
(a) Pre-emphasis circuit



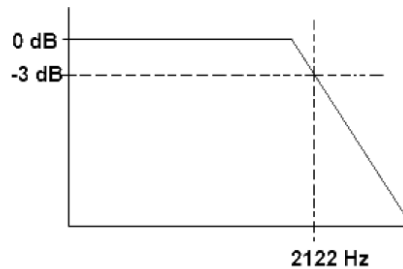
(b) Pre-emphasis curve

De-emphasis:

De-emphasis means attenuating those frequencies by an amount by which they are boosted. However, pre-emphasis is done at the transmitter and de-emphasis is done at the receiver. The purpose is to improve the signal to noise ratio for FM reception. A time constant of $75 \mu\text{s}$ is specified in the RC or L/Z network for pre-emphasis and de-emphasis.

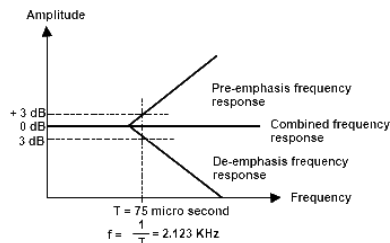
De-emphasis circuit:

(c) De-emphasis circuit



(d) De-emphasis curve

A de-emphasis circuit is used at the receiver, to return the frequency response to its normal level. This is a simple low pass filter with a time constant of $75 \mu\text{s}$. It features a cut-off frequency of 2122 Hz and causes signals above this frequency to be attenuated at the rate of 6 dB per octave.



Combined frequency response

As a result, pre-emphasis circuit at the transmitter is exactly offset by the de-emphasis circuit at the receiver, providing a normal frequency response. The combined effect of pre-emphasis and de-emphasis is to increase the high frequency components during transmission, so that they will be stronger and not masked by noise.

4.7 Capture effect, threshold effect

Capture effect:

Capture effect is a phenomenon, associated with FM reception, in which only the stronger of two signals at or near the same frequency will be demodulated.

The complete suppression of the weaker signal occurs at the receiver limiter, where it is treated as noise and rejected. When both signals are nearly equal in strength or are fading independently, the receiver may switch from one to the other.

In frequency modulation, the signal can be affected by another frequency modulated signal, whose frequency content is close to the carrier frequency of the desired FM wave. The receiver may lock

such an interference signal and suppress the desired FM wave, when the interference signal is stronger than the desired signal.

When the strength of the desired signal and interference signal are nearly equal, the receiver fluctuates back and forth between them, (i.e.) receiver locks the interference signal for some time and the desired signal for some time and this goes on randomly. This phenomenon is known as the capture effect. **Threshold effect:**

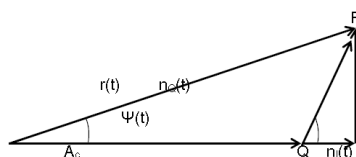
The output signal to noise ratio of FM receiver is valid, only if the carrier to noise ratio measured at the discriminator input is high compared to unity. It is observed that, as the input noise is increased, the carrier to noise ratio is decreased and so the FM receiver breaks. At first, individual clicks are heard in the receiver output and as the carrier to noise ratio decreases further, the clicks rapidly merge into a crackling or sputtering sound.

Near the break point, output SNR begins to fail, predicting the values of output SNR larger than the actual ones. This phenomenon is known as threshold effect. The threshold effect is defined as the minimum carrier to noise ratio, that gives the output SNR not less than the value predicted by the usual signal to noise formula, assuming a small noise power.

For qualitative measurement of the FM threshold effect, consider, when there is no signal present, so that the carrier is unmodulated. Then the composite signal at the frequency discriminator input is given by,

$$x(t) = [A_c + n_i(t)] \cos(2\pi f_c t) - n_q(t) \sin(2\pi f_c t)$$

Where, $n_i(t)$ and $n_q(t)$ are inphase and quadrature phase components of the narrow band noise $n(t)$ with respect to carrier wave $A_c \cos(2\pi f_c t)$.



A phasor diagram interpretation

As the amplitudes and phases of $n_i(t)$ and $n_q(t)$ change randomly with time, the point P wanders around the point Q. When the carrier to noise ratio is large, $n_i(t)$ and $n_q(t)$ are small compared to A_c , so that the point P is always around Q. Thus, the angle $\theta(t)$ is small and within a multiple of 2π radians.

The point P occasionally sweeps around the origin and $\theta(t)$ either increases or decreases by 2π radians, when the carrier to noise ratio is small. The clicks are produced only when $\theta(t)$ changes by $\pm 2\pi$ radians. From the phasor diagram, we may deduce the condition required for the clicks to occur.

A positive going click occurs when the envelope $r(t)$ and phase $\Psi(t)$ of the narrow band noise $n(t)$ satisfy the following conditions.

$$r(t) > A_c$$

$$\Psi(t) < \pi < \Psi(t) + d\Psi(t)$$

$$d\Psi(t)/dt > 0$$

These conditions ensure that $\theta(t)$ changes by 2π radians in the time increment dt , during which the phase of the narrow band noise increases by an amount $d\Psi(t)$.

Similarly, the condition for negative going clicks to occur are,

$$r(t) > A_c$$

$$\Psi(t) < -\pi < \Psi(t) + d\Psi(t)$$

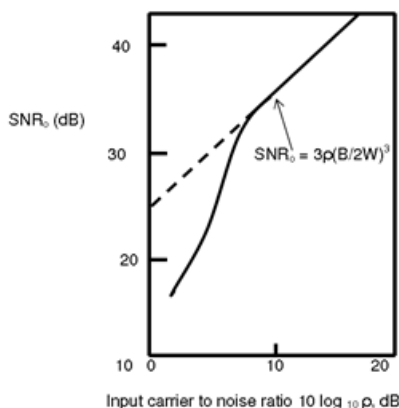
$$d\Psi(t)/dt < 0$$

These conditions ensure that $\theta(t)$ changes by -2π radians in the time increment dt .

As the carrier to noise ratio is decreased, the average number of clicks per unit time increases. When this number becomes large, the threshold is said to occur. Consequently, the output SNR deviates from the linear function of the carrier to noise ratio, when the latter falls below the threshold.

This effect is shown in below figure, this calculation is based on the following two assumptions.

1. The output signal is taken as the receiver output, measured in the absence of noise. The average output signal power is calculated for a sinusoidal modulation, that produces a frequency deviation Δf equal to $1/2$ of the IF filter bandwidth B . Thus, the carrier is enabled to swing back and forth across the entire IF band.
2. The average output noise power is calculated in the absence of signal, (i.e.), the carrier is unmodulated, with no restriction placed on the value of carrier to noise ratio.



Variation of output SNR with carrier to noise ratio, demonstrating the FM threshold effect

The curve in the above figure is plotted for the ratio, $(B/2W) = 5$. The linear portion of the curve corresponds to the limiting value, $3\rho(B/2W)^3$. From the figure, we may observe that the output SNR deviates appreciably from a linear function of the carrier to noise ratio ρ , when ρ becomes less than a threshold of 10 dB.

The threshold carrier to noise ratio ρ_{th} depends on the ratio of IF filter bandwidth to the message bandwidth B/W and ρ_{th} is influenced by the presence of modulation.

We may state that, the loss of message at an FM receiver output is negligible, if the carrier to noise ratio satisfies the condition,

$$\frac{A_c^2}{2BN_0} \geq 10$$

$$(SNR)_c \geq \frac{10B}{W}$$

The IF filter bandwidth B is designed to equal the FM transmission bandwidth, we may use Carson's rule to relate B to the message bandwidth W as follows,

$$B = 2W(1 + D)$$

Where,

D is the deviation ratio.

For sinusoidal modulation, the modulation index is used instead of D . Therefore, for no significant loss of message at an FM receiver, the output can be written as,

$$(SNR)_c \geq 20(1 + D)$$

Or in terms of decibels,

$$10\log_{10}(\text{SNR})_c \geq 13 + 10\log_{10}(1 + D)$$



INFORMATION THEORY

Entropy - Discrete Memoryless channels - Channel Capacity - Hartley-Shannon law - Source coding theorem - Huffman & Shannon-Fano codes

5.1 Entropy

Entropy is defined as the average amount of information contained in each message received. Here the message stands for an event, sample or character drawn from a distribution or data stream.

Thus, entropy characterizes our uncertainty about the source of information. The source is also characterized by the probability distribution of the samples drawn from it.

Formula for entropy:

Informations are strictly in terms of probability of events. Therefore, let us consider that we have a set of probabilities, $P = \{p_1, p_2, \dots, p_n\}$.

We define the entropy of the distribution P as,

$$H(P) = \sum_{i=1}^n p_i * \log(1/p_i)$$

Shannon defined the entropy of a discrete random variable X with possible values $\{x_1, \dots, x_n\}$ and probability mass function $P(X)$ as,

$$H(X) = E(I(X)) = E[-\ln(P(X))]$$

Where,

E = The expected value operator

I = The information content of X

$I(X)$ = Random variable

One may also define the conditional entropy of two events X and Y by taking values x_i and y_j respectively as,

$$H(X|Y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(y_j)}{p(x_i, y_j)}$$

Where, $p(x_i, y_j)$ is the probability, that $X = x_i$ and $Y = y_j$.

Properties:

If X and Y are two independent variables, then knowing the value of Y does not influence our knowledge on the value of X .

$$H(X|Y) = H(X)$$

The entropy of two simultaneous events is not more than the sum of entropies of each individual event and are equal if the two events are independent. More specifically, if X and Y are two random variables on the same probability space and (X, Y) denotes their Cartesian product then it can be written as,

$$H[(X, Y)] \leq H(X) + H(Y)$$

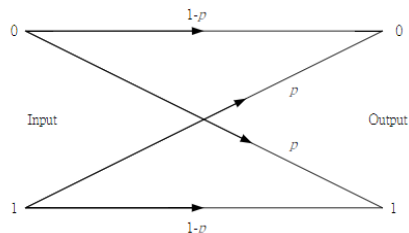
5.2 Discrete Memoryless channels - Channel Capacity

Channel models:

- Binary symmetric channel (BSC)
- Discrete memoryless channels (DMC)
- Discrete input, continuous output channels
- Waveform channels

Binary symmetric channel (BSC):

If the modulator employs binary waveforms and the detector makes hard decision, then the channel has a discrete time binary input sequence and a discrete time binary output sequence.



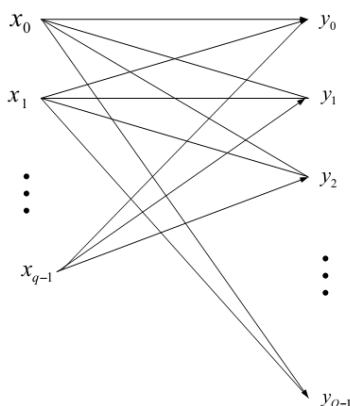
If the channel noise and other interferences cause statistically independent errors in the transmitted binary sequence with average probability p , then the channel is said to be a binary symmetric channel. Since each output bit from the channel depends only on the corresponding input bit, the channel is also memoryless.

Discrete memoryless channels (DMC):

A DMC is the same as that of BSC, but with q -ary symbols at the output of a channel encoder and Q -ary symbols at the output of a detector, where $Q \geq q$.

If the channel and modulator are memoryless, then it can be described by a set of qQ conditional probabilities.

$$P(Y = y_i | X = x_j) \equiv P(y_i | x_j), \quad i = 0, 1, \dots, Q - 1, \quad j = 0, 1, \dots, q - 1$$



If the input to a DMC is a sequence of n symbols u_1, u_2, \dots, u_n selected from the alphabet X and the corresponding output is the sequence of symbols v_1, v_2, \dots, v_n selected from the alphabet Y , the joint conditional probability is given by,

$$P(Y_1 = v_1, Y_2 = v_2, \dots, Y_n = v_n | X_1 = u_1, X_2 = u_2, \dots, X_n = u_n) = \prod_{k=1}^n P(Y_k = v_k | X_k = u_k)$$

The conditional probabilities $P(y_i | x_j)$ can be arranged in the matrix form, $P = [p_{ji}]$. Where, P is called as the probability transition matrix for the channel.

Channel coding theorem for a discrete memoryless channel:

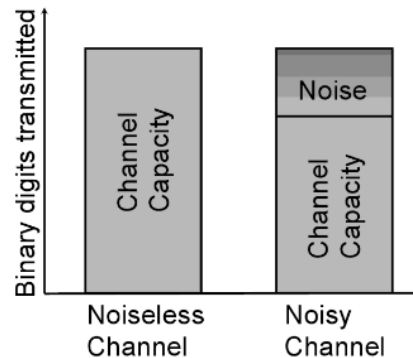
Given a source of M equally likely messages, with $M \gg 1$, which is generating the information at a rate R and the channel capacity is C . Then if $R \leq C$, there exists a coding technique such that the output of the source may be transmitted over the channel with a probability of error in the received message, which can be made arbitrarily small.

This theorem says that if $R \leq C$, it is possible to transmit the information without any error, even if noise is present.

Channel capacity:

Channel capacity is defined as the maximum amount of information which can be communicated from a channel's input to its output. Capacity can be measured in terms of the amount of information per symbol and if a channel communicates n symbols per second, then its capacity can be expressed in terms of information per second (e.g. bits/s).

The capacity of a channel is similar to the capacity of a bucket and the rate is like the amount of water we pour into the bucket. The amount of water (rate) we pour (transmit) into the bucket is up to us and the bucket can hold (communicate or transmit reliably) less than its capacity, but it cannot hold more.



Consider an alphabet of α symbols, where $\alpha = 2$ if the data is binary. If a noiseless channel transmits data at a fixed rate of n symbols/s, then it transmits information at a maximum rate or channel capacity of $n \log \alpha$ bits/s, which equals n bits/s for a binary data.

However, the capacity of a channel is different from the rate at which the information is actually communicated through that channel. The rate is the number of bits of information communicated per second, which depends on the code used to transmit data.

The rate of a given code may be less than the capacity of a channel, but it cannot be greater, the channel capacity is the maximum rate that can be achieved, when over all possible codes are considered. For example, a code for binary data in which 0s and 1s occur equally, often ensures that each binary digit (symbol) conveys one bit of information, but for any other code each binary digit conveys less than one bit.

Thus, the capacity of a noiseless binary channel is numerically equal to the rate at which it transmits binary digits, whereas the capacity of a noisy binary channel is less than this, because it is limited by the amount of noise in the channel.

Channel capacity of a discrete memoryless channel:

The channel capacity of a discrete memoryless channel is given by the maximum average mutual information. The maximization is taken with respect to input probabilities $P(x_i)$.

$$C = \max I(X:Y) P(x_i)$$

5.3 Hartley-Shannon law

This law is named after Ralph Hartley and Claude Shannon. In information theory, the Shannon-Hartley theorem gives us the maximum rate at which the information can be transmitted over a communication channel of a specified bandwidth in the presence of noise.

It is an application of the noisy-channel coding theorem to the archetypal case of a continuous time analog communication channel subject to Gaussian noise.

The theorem establishes Shannon's channel capacity for such a communication link, a bound on the maximum amount of error free digital data that can be transmitted within a specified bandwidth in the presence of noise interference, assuming that the signal power is bounded and the Gaussian noise process is characterized by a known power or power spectral density.

Statement:

Considering all possible multilevel and multiphase encoding techniques, Shannon-Hartley theorem states the channel capacity C , which means that, the theoretical tightest upper bound on the information rate of data which can be communicated at an arbitrarily low error rate using an average received signal power S through an analog communication channel subject to additive white Gaussian noise of power N is given by,

$$C = B \log_2 \left(1 + \frac{S}{N} \right)$$

Where,

C = Channel capacity in bits per second

B = Bandwidth of the channel in hertz

S = Average received signal power over the bandwidth, measured in watts (or volts squared)

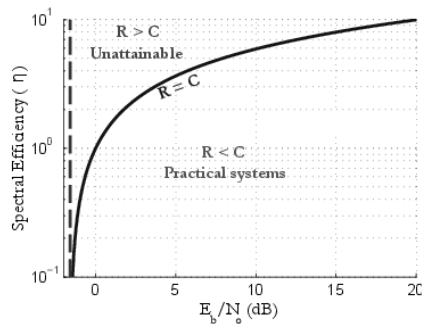
N = Average noise or interference power over its bandwidth, measured in watts

S/N is the signal to noise ratio (SNR) or the carrier to noise ratio (CNR) of the communication signal to the noise and interference at the receiver, expressed as a linear power ratio and not as logarithmic decibels.

The above expression for channel capacity makes intuitive sense.

- The speed at which the information symbols can be sent over the given channel is limited by the bandwidth.

- The SNR ratio limits how much information we can squeeze in each transmitted symbols. Increasing SNR makes the transmitted symbols more robust against noise. SNR is a function of signal power, signal quality and the characteristics of the channel. It is measured at the receiver's front end.
- To increase the information rate, the signal to noise ratio and the allocated bandwidth have to be traded against each other.
- When there is no noise, the signal to noise ratio becomes infinite and so an infinite information rate is possible at a very small bandwidth.



Channel capacity and power efficiency limit

The Shannon's equation relies on two important concepts.

1. In this principle, the trade-off between SNR and bandwidth is possible.
2. The information capacity depends on both SNR and bandwidth.

In the above graph, the dashed line represents the asymptote of E_b/N_0 as the bandwidth B approaches infinity. The asymptote at $E_b/N_0 = \ln(2) = -1.59\text{dB}$. This value is called as Shannon's Limit or specifically Shannon's power efficiency limit.

5.4 Source coding theorem

The theorem can be stated as follows.

Given a discrete memoryless source of entropy $H(S)$, the average code word length L for any distortionless source coding is bounded as,

$$L \geq H(S)$$

This theorem provides a mathematical tool for assessing data compaction, (i.e.) lossless data compression of data generated by a discrete memoryless source.

The source coding theorem is also known as the noiseless coding theorem, since it establishes the condition for error free encoding to be possible.

Assume a set of N symbols that is to be transmitted through the communication channel. These symbols can be treated as N independent samples of a random variable X with probability $P(X)$ and entropy $H(X) = -\sum_x P(X) \log P(X)$.

For example, $P(X = s)$ is much higher than $P(X = z)$. To minimize the code length and number of bits for these symbols, it is natural to assign a shorter code for symbols of high probabilities. For example, shorter code is assigned for s than for z . The length of the code for a symbol x with $P(X = x)$ can be $-\log P(X = x)$.

Let L be the average number of bits to encode N symbols. Shannon proved that the minimum L satisfies,

$$H(X) \leq L < H(X) + \frac{1}{N}$$

This source coding theorem establishes the limits of data compression.

In practice, true probability $P(X)$ is not available and can be estimated only by $Q(X)$, which is to be used for source coding purpose.

In this case, the minimum L_Q satisfies,

$$H(X) + KL(P||Q) \leq L_Q < H(X) + KL(P||Q) + \frac{1}{N}$$

Where, $KL(P||Q)$ is the relative entropy or the Kullback-Leibler (KL) divergence, which can be defined as,

$$KL(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \geq 0$$

This is a measure of the difference between two probability distributions, (i.e.) from the true probability distribution $P(X)$ to its estimate $Q(X)$. Obviously KL divergence is zero if and only if, $P(x) = Q(x)$. However, the KL divergence is not a distance metric as it is not symmetric.

$$KL(P||Q) \neq KL(Q||P)$$

5.5 Huffman codes

Huffman codes are an important class of prefix codes, used to assign each symbol with a sequence of bits roughly equal in length to the amount of information conveyed by the symbol.

An optimum code is the one that has the highest efficiency. The final result of this coding is a source code, whose average codeword length approaches the fundamental limit provided by entropy, say H .

In Huffman coding, fixed length blocks of the source output are mapped to variable length binary blocks called fixed to variable length coding, (i.e.) more frequent fixed length sequences are mapped to shorter binary sequences, whereas less frequent fixed length sequences are mapped to longer binary sequences.

Encoding process or algorithm:

1. Sort out source symbols in order of decreasing probability.

(a) The two source symbols of lowest probability will be assigned 0 and 1.

The above step is known as splitting stage.

2. Merge the two least probable outputs into a single output, whose probability is the sum of corresponding probabilities.

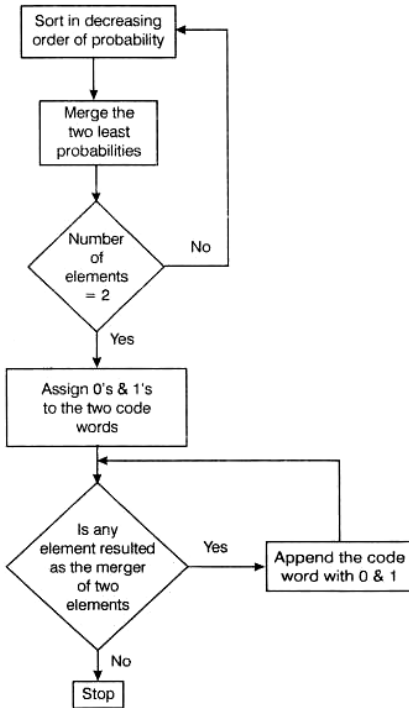
a) The list of source symbols are reduced in size by one.

b) The probability of new symbol is placed in the list in accordance with its value.

3. This procedure is continued till the final list of source symbols of only two, (i.e.) for which 0 and 1 are assigned.

This type of reduction process is continued in a step by step manner, by working backward the code for original symbol is found.

Flow diagram:



Flow diagram for Huffman coding

Examples:

Example of Huffman coding which summarizes the code construction and resulting codewords.

| | |
|---------|----------------|
| 0 | $\frac{1}{2}$ |
| 1 0 | $\frac{1}{4}$ |
| 1 1 0 | $\frac{1}{8}$ |
| 1 1 1 0 | $\frac{1}{16}$ |
| 1 1 1 1 | $\frac{1}{16}$ |

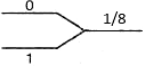
Step 1:

Sort out in the order of decreasing probability and assign 0 and 1.

| | | | |
|---------|----------------|---|-------------------|
| 0 | $\frac{1}{2}$ | | |
| 1 0 | $\frac{1}{4}$ | | |
| 1 1 0 | $\frac{1}{8}$ | | * Splitting stage |
| 1 1 1 0 | $\frac{1}{16}$ | 0 | -0 |
| 1 1 1 1 | $\frac{1}{16}$ | 1 | -1 |

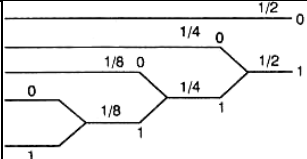
Step 2:

Merge the two least probable outputs.

| | | |
|---------|----------------|---|
| 0 | $\frac{1}{2}$ | |
| 1 0 | $\frac{1}{4}$ | |
| 1 1 0 | $\frac{1}{8}$ | * List reduced in size by one |
| 1 1 1 0 | $\frac{1}{16}$ | * Probability of new symbol placed in list |
| 1 1 1 1 | $\frac{1}{16}$ |  |

Step 3:

Continue the above two steps.

| | | |
|---|---------------|---|
| 0 | $\frac{1}{2}$ |  |
|---|---------------|---|

| | | |
|---------|----------------|--|
| 1 0 | $\frac{1}{4}$ | |
| 1 1 0 | $\frac{1}{8}$ | |
| 1 1 1 0 | $\frac{1}{16}$ | |
| 1 1 1 1 | $\frac{1}{16}$ | |

Examples:

(a) Huffman encoding algorithm:

| Symbol | Stage 1 | Stage 2 | Stage 3 | Stage 4 |
|--------|---------|---------|---------|---------|
| X_0 | | | | |
| X_1 | | | | |
| X_2 | | | | |
| X_3 | | | | |
| X_4 | | | | |

i.e.

(b) Source code:

| Symbol | Probability | Codeword |
|--------|-------------|----------|
| | | |

| | | |
|-------|-----|-------|
| x_0 | 0.4 | 0 0 |
| x_1 | 0.2 | 1 0 |
| x_2 | 0.2 | 1 1 |
| x_3 | 0.1 | 0 1 0 |
| x_4 | 0.1 | 0 1 1 |

(B) The average codeword length is,

$$\bar{l} = 0.4(2) + 0.2(2) + 0.2(2) + 0.1(3) + 0.1(3)$$

$$\bar{l} = 2.2$$

(C) Entropy is given by,

$$H = 0.4 \log_2 \left(\frac{1}{0.4} \right) + 0.2 \log_2 \left(\frac{1}{0.2} \right) + 0.2 \log_2 \left(\frac{1}{0.2} \right) + 0.1 \log_2 \left(\frac{1}{0.1} \right) + 0.1 \log_2 \left(\frac{1}{0.1} \right)$$

$$H = 0.52877 + 0.46439 + 0.46439 + 0.33219 + 0.33219$$

$$H = 2.12193 \text{ bits}$$

Problems:

1. Let us design a Huffman code for the source given in Q_9 . Let us also determine the average code length and coding efficiency.

| Symbol | x_0 | x_1 | x_2 | x_3 | x_4 |
|-------------|-------|-------|-------|-------|-------|
| Probability | 0.55 | 0.15 | 0.15 | 0.10 | 0.05 |

Solution:

Given:

| | | | | | |
|--------------------|-------|-------|-------|-------|-------|
| Symbol | X_0 | X_1 | X_2 | X_3 | X_4 |
| Probability | 0.55 | 0.15 | 0.15 | 0.10 | 0.05 |

Formula to be used:

$$L = \sum_{k=0}^4 P_k I_k$$

$$L = \sum_{k=0}^2 P_k I_k$$

$$\eta = (H/L)$$

(i) Placing the probability as high as possible,

| x_i | Stage 1 | Stage 2 | Stage 3 | Stage 4 |
|-------|---------|---------|---------|---------|
| x_0 | | | | |
| x_1 | | | | |
| x_2 | | | | |
| x_3 | | | | |
| x_4 | | | | |

| | |
|-------|--|
| x_3 | |
| x_4 | |

| Message (Symbol/source) | Probability | Code | Length |
|----------------------------|-------------|------|--------|
| x_0 | 0.55 | 0 | 1 |
| x_1 | 0.15 | 11 | 2 |
| x_2 | 0.15 | 100 | 3 |
| x_3 | 0.1 | 1010 | 4 |
| x_4 | 0.05 | 1011 | 4 |

Average codeword length is,

$$L = \sum_{k=0}^4 P_k I_k$$

$$L = (0.55 \times 1) + (0.15 \times 2) + (0.15 \times 3) + (0.1 \times 4) + (0.05 \times 4)$$

$$L = 1.9 \text{ bits/symbol}$$

| | |
|----------|---|
| x_i | Stage |
| $(x_0)A$ | <p>The diagram shows a probability tree starting with a root node of 0.3. A diagonal line splits this into two branches: an upper branch to 0.41 and a lower branch to 0.25. From the 0.25 node, a horizontal line leads to a vertical line that splits into two branches: an upper branch to 0.3 and a lower branch to 0.16.</p> |

| | |
|----------|--|
| $(x_1)B$ | |
| $(x_2)C$ | |

| Message | Probability | Code | Length |
|---------|-------------|------|--------|
| x_0 | 0.3 | 0 | 1 |
| x_1 | 0.25 | 10 | 2 |
| x_2 | 0.16 | 11 | 2 |

$$H = 0.3 \log_2 \left(\frac{1}{0.3} \right) + 0.16 \log_2 \left(\frac{1}{0.16} \right) + 0.25 \log_2 \left(\frac{1}{0.25} \right)$$

$$H = 1.458 \text{ bits/message}$$

$$L = \sum_{k=0}^2 P_k I_k$$

$$L = (0.3 \times 1) + (0.25 \times 2) + (0.16 \times 2)$$

$$L = 1.12$$

$$\eta = \frac{1.458}{1.12} = 1.30$$

2. Using Huffman code I, encode the following symbols,

$$S = [0.3, 0.2, 0.25, 0.12, 0.05, 0.08]$$

Let us calculate,

(i) Average codeword length

(ii) Entropy of the source

(iii) Code efficiency

(iv) Redundancy

Solution:

Given:

$$S = [0.3, 0.2, 0.25, 0.12, 0.05, 0.08]$$

$$\bar{N} = \sum_{k=0}^{L=1} p_k n_k$$

Formula to be used:

$$H = \sum_{k=1}^M p_k \log_2 \left(\frac{1}{p_k} \right)$$

$$\eta = \frac{H}{\bar{N}}$$

$$\gamma = 1 - \eta$$

| Stage 1 | Stage 2 | Stage 3 | Stage 4 | Stage 5 | Codeword | Number of codes | |
|---------|---------|---------|---------|---------|----------|-----------------|----------|
| | | | | | | 01 | 2 |
| | | | | | | 01 | 2 |
| | | | | | | 11 | 2 |

| | | |
|--|------|---|
| | 101 | 3 |
| | 1000 | 4 |
| | 1001 | 4 |

(i) Average codeword length:

$$\bar{N} = \sum_{k=0}^{L-1} p_k n_k$$

$$= 0.3 \times 2 + 0.25 \times 2 + 0.2 \times 2 + 0.12 \times 5 + 0.08 \times 4 + 0.05 \times 4$$

$$\bar{N} = 2.38$$

(ii) Entropy of the source:

$$H = \sum_{k=1}^M p_k \log_2 \left(\frac{1}{p_k} \right)$$

$$= p_1 \log_2 \left(\frac{1}{p_1} \right) + p_2 \log_2 \left(\frac{1}{p_2} \right) + p_3 \log_2 \left(\frac{1}{p_3} \right) + p_4 \log_2 \left(\frac{1}{p_4} \right) + p_5 \log_2 \left(\frac{1}{p_5} \right) + p_6 \log_2 \left(\frac{1}{p_6} \right)$$

$$= 0.3 \log_2 \left(\frac{1}{0.3} \right) + 0.25 \log_2 \left(\frac{1}{0.25} \right) + 0.2 \log_2 \left(\frac{1}{0.2} \right) + 0.12 \log_2 \left(\frac{1}{0.12} \right) + 0.08 \log_2 \left(\frac{1}{0.08} \right) + 0.05 \log_2 \left(\frac{1}{0.05} \right)$$

$$= 0.521 + 0.5 + 0.4643 + 0.367 + 0.2915 + 0.216$$

$$= 2.3568 \text{ bits of information/message}$$

(iii) Code efficiency:

$$\eta = \frac{H}{\bar{N}}$$

$$= \frac{2.3568}{2.38}$$

$$= 0.99$$

(iv) Redundancy:

$$\gamma = 1 - \eta$$

$$\gamma = 0.01$$

3. Let us determine the Huffman coding for the probabilities $P = \{0.0625, 0.25, 0.125, 0.125, 0, 25, 0.125, 0.0625\}$ and the efficiency of the code.

Solution:

Given:

$$P = \{0.0625, 0.25, 0.125, 0.125, 0, 25, 0.125, 0.0625\}$$

$$\bar{N} = \sum_{k=0}^6 p_k n_k$$

Formula to be used:

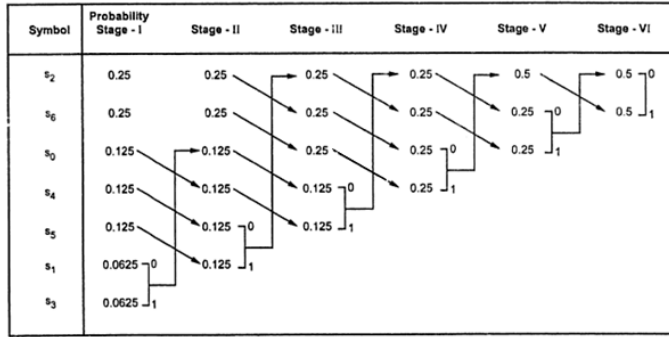
$$H = \sum_{k=0}^6 p_k \log_2 \left(\frac{1}{p_k} \right)$$

$$\eta = \frac{H}{\bar{N}}$$

(i) To obtain codewords:

In Huffman coding, minimum code variance can be obtained by putting the combined symbol probability as high as possible.

Huffman coding:



As per the above table, the codes are listed below.

Codewords:

| Probability | Digits obtained by tracing | Codeword | Number of bits per symbol n_k |
|-------------|----------------------------|----------|---------------------------------|
| 0.125 | 100 | 001 | 3 |
| 0.0625 | 0000 | 0000 | 4 |
| 0.25 | 01 | 10 | 2 |
| 0.0625 | 1000 | 0001 | 4 |
| 0.125 | 010 | 010 | 3 |
| 0.125 | 110 | 011 | 3 |
| 0.25 | 11 | 11 | 2 |

(ii) To obtain average codeword length:

Average codeword length is given as,

$$\bar{N} = \sum_{k=0}^6 p_k n_k$$

$$= 0.125 (3) + 0.0625 (4) + 0.25 (2) + 0.0625 (4) + 0.125 (3) + 0.125 (3) + 0.25 (2)$$

$$= 2.625 \text{ bits/symbol}$$

(iii) To obtain entropy of the source:

Entropy is given as,

$$H = \sum_{k=0}^6 p_k \log_2 \left(\frac{1}{p_k} \right)$$

$$= 0.125 \log_2 \left(\frac{1}{0.125} \right) + 0.0625 \log_2 \left(\frac{1}{0.0625} \right) + 0.25 \log_2 \left(\frac{1}{0.25} \right)$$

$$+ 0.0625 \log_2 \left(\frac{1}{0.0625} \right) + 0.125 \log_2 \left(\frac{1}{0.125} \right)$$

$$+ 0.125 \log_2 \left(\frac{1}{0.125} \right) + 0.25 \log_2 \left(\frac{1}{0.25} \right)$$

$$= 2.5 \text{ information/message}$$

(iv) To obtain code efficiency:

Code efficiency is given as,

$$\eta = \frac{H}{\bar{N}}$$

$$= \frac{2.5}{2.625}$$

$$= 0.9523$$

5.6 Shannon-Fano codes

In Shannon-Fano coding, the symbols are arranged in the order of most probable to least probable and then divided into two sets, whose total probabilities are as close as possible so that they become equal. All symbols then have the first digit of their codes assigned, symbols in the first set receive "0", whereas symbols in the second set receive "1".

As long as any set with more than one member remain, the same process is repeated on those sets, to determine the successive digits of their codes. When a set has been reduced to one symbol, this means that the symbol's code is complete and will not form the prefix of any other symbol's code.

Procedure for Shannon-Fano algorithm:

A Shannon-Fano tree is built according to a specification, designed to define an effective code table. The actual algorithm is simple.

1. For a given list of symbols, develop a corresponding list of probabilities or frequency counts so that each symbol's relative frequency of occurrence is known.
2. Sort the list of symbols according to the frequency, (i.e.) the most frequently occurring symbols at the left and the least common at the right.
3. Divide the list into two parts, with the total frequency counts of the left part being as close as to the total frequency counts of the right.
4. The left part of the list is assigned the binary digit 0 and the right part is assigned the digit 1. This means that, the codes for all the symbols in the first part will start with 0 and all the codes in the second part will start with 1.
5. Recursively apply steps 3 and 4 to each of the two halves, subdivide the groups and add bits to the codes until each symbol has become a corresponding code leaf on the tree.

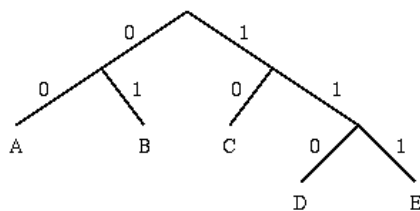
Example of Shannon-Fano algorithm:

| | | | | | |
|---------------|----|---|---|---|---|
| Symbol | A | B | C | D | E |
| Count | 15 | 7 | 6 | 6 | 5 |

Encoding for the Shannon-Fano algorithm:

Top-down approach:

1. Sort symbols according to their frequencies or probabilities, (e.g.) ABCDE.
2. Recursively divide into two parts, each with approximately same number of counts.



Tree diagram

| Symbol | Count | $\log_2(1/p_i)$ | Code | Subtotal (number of bits) |
|--------|-------|-----------------|------|---------------------------|
| A | 15 | 1.38 | 00 | 30 |
| B | 7 | 2.48 | 01 | 14 |
| C | 6 | 2.70 | 10 | 12 |
| D | 6 | 2.70 | 110 | 18 |
| E | 5 | 2.96 | 111 | 15 |

Total (number of bits): 89

Shannon-Fano encoding:

Shannon-Fano encoding is the first established encoding method. This method and the corresponding code were invented simultaneously and independently of each other by C. Shannon and R. Fano in 1948.

This method constructs reasonably efficient separable binary codes for sources without memory. Sources without memory are such sources of information, where the probability of the next transmitted symbol does not depend on the probability of the previously transmitted symbol. Separable codes are those codes for which the unique decipherability holds.

The ensemble of original messages that are to be transmitted have their corresponding probabilities as,

$$[X] = [x_1, x_2, \dots, x_n], [P] = [p_1, p_2, \dots, p_n]$$

Our task is to associate the sequence of binary numbers (b_k) of unspecified length (n_k) to each message (X_k) such that,

- No sequences of employed binary numbers b_k can be obtained from each other by adding more binary digits to the shorter sequence.
- The transmission of the encoded message is efficient, (i.e.) 0 and 1 appear independently and with almost equal probabilities. This ensures the transmission of almost 1 bit of information per digit of the encoded messages.
- Another important general consideration, which was taken into account by C. Shannon and R. Fano is that, a more frequent message has to be encoded by a shorter encoding vector and a less frequent message has to be encoded by a longer encoding vector.
- It is not an optimal code.
- It produces fairly efficient variable length encoding.

Problem:

1. Let us design a Shannon-Fano code for the source given in Q₉. Let us also determine the average code length and efficiency.

| Message | Probability | Codeword | Number of bits |
|---------|-------------|----------|----------------|
| A | 0.3 | 0 | 1 |
| B | 0.16 | 1 0 | 2 |
| C | 0.25 | 1 1 | 2 |

Solution:

Given:

| Message | Probability | Codeword | Number of bits |
|---------|-------------|----------|----------------|
| A | 0.3 | 0 | 1 |
| B | 0.16 | 1 0 | 2 |
| C | 0.25 | 1 1 | 2 |

Formula to be used:

$$H(X) = \sum_{k=1}^3 P_k \log_2 \left(\frac{1}{P_k} \right)$$

$$\bar{N} = \sum_{k=1}^3 P_k n_k$$

$$\eta = \frac{H(X)}{\bar{N}}$$

$$\begin{aligned}
 H(X) &= \sum_{k=1}^3 P_k \log_2 \left(\frac{1}{P_k} \right) \\
 &= 0.3 \log_2 \left(\frac{1}{0.3} \right) + 0.16 \log_2 \left(\frac{1}{0.16} \right) + 0.25 \log_2 \left(\frac{1}{0.25} \right)
 \end{aligned}$$

H(X) = 1.458 bits/message

$$\bar{N} = \sum_{k=1}^3 P_k n_k$$

$$= (0.3 \times 1) + (0.16 \times 2) + (0.25 \times 2)$$

$$\bar{N} = 1.12$$

$$\eta = \frac{H(X)}{\bar{N}} = \frac{1.458}{1.12} = 1.30$$

2. Let us compare the Huffman coding and Shannon-Fano coding algorithms for data compression. For a discrete memoryless source “X” with six symbols x_1, x_2, \dots, x_6 , let us determine a compact code for every symbol if the probability distribution is as follows,

$$p(x_1) = 0.3, p(x_2) = 0.25, p(x_3) = 0.2$$

$$p(x_4) = 0.12, p(x_5) = 0.08, p(x_6) = 0.05$$

Let us also determine (i) Entropy of the source, (ii) Average length of the code, (iii) Efficiency and (iv) Redundancy of the code.

Solution:

Given:

$$p(x_1) = 0.3, p(x_2) = 0.25, p(x_3) = 0.2$$

$$p(x_4) = 0.12, p(x_5) = 0.08, p(x_6) = 0.05$$

$$H = \sum_{k=1}^M p_k \log_2 \left(\frac{1}{p_k} \right)$$

Formula to be used:

$$\bar{N} = \sum_{k=0}^{L-1} p_k n_k$$

$$\eta = \frac{H}{\bar{N}}$$

$$\gamma = 1 - \eta$$

(i) Entropy of the source:

$$H = \sum_{k=1}^M p_k \log_2 \left(\frac{1}{p_k} \right)$$

For six messages the above equation becomes,

$$\begin{aligned} &= p_1 \log_2 \left(\frac{1}{p_1} \right) + p_2 \log_2 \left(\frac{1}{p_2} \right) + p_3 \log_2 \left(\frac{1}{p_3} \right) + p_4 \log_2 \left(\frac{1}{p_4} \right) + p_5 \log_2 \left(\frac{1}{p_5} \right) \\ &\quad + p_6 \log_2 \left(\frac{1}{p_6} \right) \end{aligned}$$

On substituting the values we get,

$$H = 0.3 \log_2 \left(\frac{1}{0.3} \right) + 0.25 \log_2 \left(\frac{1}{0.25} \right) + 0.2 \log_2 \left(\frac{1}{0.2} \right) + 0.12 \log_2 \left(\frac{1}{0.12} \right) + 0.08 \log_2 \left(\frac{1}{0.08} \right) + 0.05 \log_2 \left(\frac{1}{0.05} \right)$$

$$= 0.521 + 0.5 + 0.4643 + 0.367 + 0.2915 + 0.216$$

$$= 2.3568 \text{ bits of information/message}$$

(a) Shannon-Fano coding:

The table below shows the procedure for obtaining Shannon-Fano codes.

Shannon-Fano algorithm:

| Symbol | Probability | Stage-I | Stage-II | Stage-III | Stage-IV | Codeword | No. of bits per message n_k |
|--------|-------------|---------|----------|-----------|----------|----------|-------------------------------|
| x_1 | 0.3 | 0 | 0 | | | 00 | 2 |
| x_2 | 0.25 | 0 | 1 | | | 01 | 2 |
| x_3 | 0.2 | 1 | 0 | | | 10 | 2 |
| x_4 | 0.12 | 1 | 1 | 0 | | 110 | 3 |
| x_5 | 0.08 | 1 | 1 | 1 | 0 | 1110 | 4 |
| x_6 | 0.05 | 1 | 1 | 1 | 1 | 1111 | 4 |

(ii) To obtain average number of bits per message (\bar{N}):

\bar{N} is given as,
$$\bar{N} = \sum_{k=0}^{L-1} p_k n_k$$

On substituting the values in above equation we get,

$$\bar{N} = (0.3)(2) + (0.25)(2) + (0.2)(2) + (0.12)(3) + (0.08)(4) + (0.05)(4)$$

$$\bar{N} = 2.38$$

(iii) To obtain code efficiency:

Code efficiency is given as,

$$\eta = \frac{H}{\bar{N}}$$

$$= \frac{2.3568}{2.38}$$

$$= 0.99$$

(iv) To obtain redundancy of the code:

Redundancy is given as,

$$\text{Redundancy } (\gamma) = 1 - \eta = 1 - 0.99 = 0.01$$

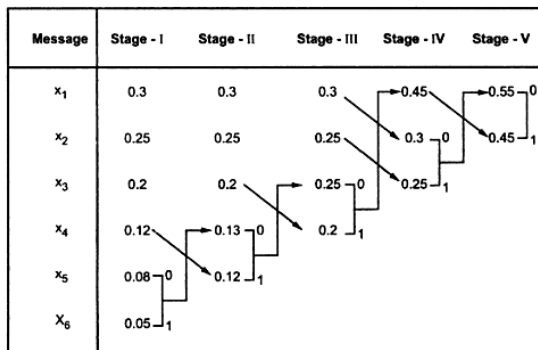
Here, 0.01 indicates that there are 1% of redundant bits in the code.

(b) Huffman coding:

(i) To obtain codewords:

Table below lists the Huffman coding algorithm.

Huffman algorithm:



Based on the above encoding arrangements following codes are generated.

Huffman codes:

| Message | Probability r_k | Digits obtained by tracing $b_3 b_2 b_1 b_0$ | Codeword $b_0 b_1 b_2 b_3$ | Number of digits n_k |
|---------|-------------------|---|-------------------------------|------------------------|
| x_1 | 0.3 | 0 0 | 0 0 | 2 |

| | | | | |
|-------|------|---------|---------|---|
| x_2 | 0.25 | 1 0 | 0 1 | 2 |
| x_3 | 0.2 | 1 1 | 1 1 | 2 |
| x_4 | 0.12 | 1 0 1 | 1 0 1 | 3 |
| x_5 | 0.08 | 0 0 0 1 | 1 0 0 0 | 4 |
| x_6 | 0.05 | 1 0 0 1 | 1 0 0 1 | 4 |

(ii) To obtain average number of bits per message (\bar{N}):

\bar{N} is given as,

$$\bar{N} = \sum_{k=0}^{l-1} p_k n_k$$

On substituting the values in above equation we get,

$$\begin{aligned} \bar{N} &= (0.3)(2) + (0.25)(2) + (0.2)(2) + (0.12)(3) + (0.08)(4) + (0.05)(4) \\ &= \mathbf{2.38} \end{aligned}$$

(iii) To obtain code efficiency:

Code efficiency is given as,

$$\eta = \frac{H}{\bar{N}}$$

$$= \frac{2.3568}{2.38}$$

$$= 0.99$$

Thus, the code efficiency of Shannon-Fano algorithm and Huffman algorithm is the same.

(iv) Redundancy of the code:

Redundancy (γ) is given as,

$$\gamma = 1 - \eta$$

$$\gamma = 1 - 0.99 = 0.01$$