Modulation

for

EE311

by

A. DAHIMENE

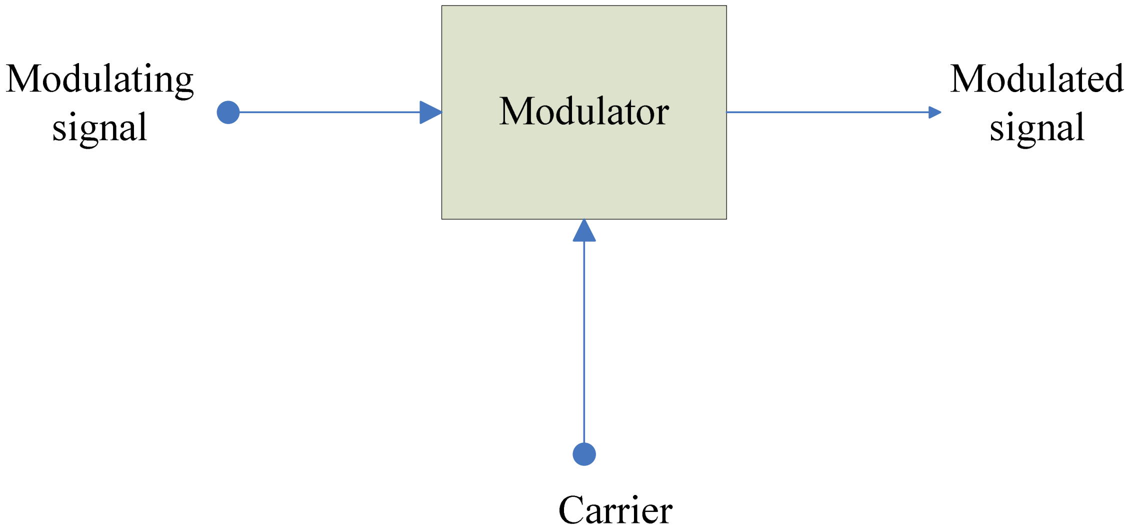
Institut de Génie Electrique et Electronique

Université M'hamed Bougara

Boumerdes

Modulation

In this chapter, we are going to learn basic principles of modulation. Modulation can be defined as the process of impressing information from a modulating signal onto another signal called the carrier. The resulting signal is called the modulated signal (in general, it is a bandpass signal).



Typical modulation system

The reverse process is called demodulation. When the modulator and the demodulator are located in the same apparatus, the system is called a MODEM (MOdulator, DEModulator).

When we do not use modulation, the system is called a "baseband communication" system. At that time, the baseband signal is transmitted directly.

Before we go on in the development of modulation theory, we have first to answer the following question: Why modulate?

Historically, modulation has been introduced in order to use reasonably sized antennas in radio communication. We know that the physical size of an antenna is a fraction of the wavelength.

The wavelength is $\lambda = \frac{c}{f}$, where $c = 3\times10^8$ m/s is the speed of light, f is the frequency of the signal. So, if we want to transmit a baseband signal of 3 kHz by radio, the required wavelength is 100 km. It is evident that it is very hard to build an antenna having many kilometers of length. If we can transfer the information to a bandpass signal with a carrier of 30 MHz, we obtain a wavelength of 10 meters. A quarter wave antenna will be 2.5 meters. This is much more reasonable. Modulation is also used to make the information fit the communication channel. Sophisticated modulation schemes are commonly used nowadays to transmit information. Techniques like OFDM, Trellis Coding, CDMA, etc. are commonly used in everyday communication systems.

Distortionless Communication

If we consider the whole communication system from the baseband source signal to the baseband destination signal, all communication systems can be considered as baseband. In this case, a good communication system must be "*distortionless*". This means that the destination signal must be a scaled (and maybe delayed) replica of the source signal. If $x(t)$ is the source signal and $y(t)$ is the destination one, we must have:

$y(t) = kx(t-\tau)$. k is a constant, τ is a time delay.

In the frequency domain, we obtain:

$Y(f) = ke^{-j2\pi f\tau}X(f)$. This means that the overall communication system must behave like a filter (LTI system) with a transfer function:

$$H(f) = \frac{Y(f)}{X(f)} = ke^{-j2\pi f\tau}$$

So, distortionless communication implies that the amplitude response |*H*(*f*)| must be constant and that the phase response Arg[*H*(*f*)] must be a linear function of the frequency. This means that all frequencies must be delayed by the same amount. If the transfer between the input and output signal is linear and time invariant but without satisfying the above conditions, we say that the communication system is subjected to "linear distortion". This distortion can come from the amplitude response which is not constant or from the phase response which is not linearly related to frequency (phase or delay distortion).

This type of distortion can be cured or minimized by using a filter called an "*equalizer*" at the output of the communication channel. When the transfer function between the input and output is nonlinear, we are in presence of "*nonlinear distortion*".

Harmonic distortion:

When we apply a pure sinewave at a frequency f_0 to a linear system, the output will be a sinewave at the same frequency. However, if the system is nonlinear, the output will be a periodic waveform at the same frequency, but it will not be sinusoidal anymore. So, we observe harmonics at the output.

Let the input be $x(t) = A\cos\omega_0 t$, the output will be

$$y(t) = c_0 + \sum_{n=1}^{\infty} a_n \cos\left(n\omega_0 + \theta_n\right),\ a_n = 2\left|c_n\right| \text{and } \theta_n = \text{Arg}\left[c_n\right].$$

The total harmonic distortion coefficient measures how far the signal $y(t)$ is from a sinewave. It is evaluated as:

$$d = 100 \times \sqrt{\frac{\sum_{n=2}^{\infty} a_n^2}{a_1^2}} \%$$

It is the ratio of the rms value of all the harmonics of the signal $y(t)$ over the rms value of the fundamental.

Classification of modulation systems.

Depending on the modulating signal, we distinguish two different types of modulation systems:

- *Digital modulation systems*: they are used to transmit digital information through physical channels.
- *Analog modulation systems*: the modulating signal in this case is a baseband analog signal.

We can also classify modulation according to the type of carrier used (and therefore the modulated wave produced).

- *Continuous wave* (CW) *modulation*: The carrier is a sinewave and the modulated signal is a narrow bandpass signal.
- *Pulse modulation*: The carrier is a periodic train of pulses. The modulated signal will carry information about samples of the signal.

We are going to analyze first analog CW modulation.

Analog CW modulation.

The modulating signal $\tilde{s}(t)$ is assumed to be bounded. This means that there exists a peak value $\left|\tilde{s}(t)\right|_{\max}$ such that: $\left|\tilde{s}(t)\right| \le \left|\tilde{s}(t)\right|_{\max}$ for all t. We can thus define a normalized signal $s(t) = \dfrac{\tilde{s}(t)}{\left|\tilde{s}(t)\right|_{\max}}$ and we have $\left|s(t)\right| \le 1$.

The signal is also assumed to have an average value of zero. This means that there is no delta impulse at the origin in its spectrum. It is also assumed to be bandlimited to a maximum frequency W. In other words, if $S(f) = \mathcal{F}\left[s(t)\right]$ then $S(f) = 0$ for $\left|f\right| > W$.

In CW modulation, the modulated signal $x(t)$ is a narrow bandpass signal. This means that it can be expressed in either quadrature form:

$$x(t) = a(t)\cos\omega_0 t - b(t)\sin\omega_0 t$$

or in modulus/phase form:

$$x(t) = r(t)\cos\left(\omega_0 t + \varphi(t)\right)$$

$\omega_0 = 2\pi f_0$ being the carrier frequency. Depending on the modulation method, the information (signal $s(t)$) can affect $a(t)$, $b(t)$, $r(t)$ or $\varphi(t)$.

I. Linear modulations

We are going to study in this part modulation methods where the quadrature components $a(t)$ and $b(t)$ are linearly dependent on the baseband signal $s(t)$. A linear modulation method must satisfy the superposition principle. If $x_1(t)$ is produced by $s_1(t)$ and $x_2(t)$ is produced by $s_2(t)$, then $a_1x_1(t)+a_2x_2(t)$ is produced by $a_1s_1(t)+a_2s_2(t)$.

Before proceeding in the analysis of the different types of linear modulation, we are going to study an "*almost linear*" one: The Amplitude Modulation (AM).

Amplitude Modulation (AM)

In AM, the information $s(t)$ is carried by the modulus $r(t)$ of the signal $x(t)$. Since we have the constraint that $r(t)$ must remain positive all the time, we cannot simply make it proportional to $s(t)$. We have to add a constant in order to satisfy the above constraint.

$$r(t) = A_0 + k_a\tilde{s}(t)$$

A_0 is a positive dc signal added to make $r(t) \geq 0$, k_a is a proportionality constant and $\tilde{s}(t)$ is the unnormalized signal. The phase of the carrier $\varphi(t)$ is constant and we use the value of zero. If we introduce the normalized signal $s(t)$, we can re-express $r(t)$ as:

$$r(t) = A_0 + k_a s(t)\left|\tilde{s}(t)\right|_{\max} = A_0\left(1 + ms(t)\right)$$

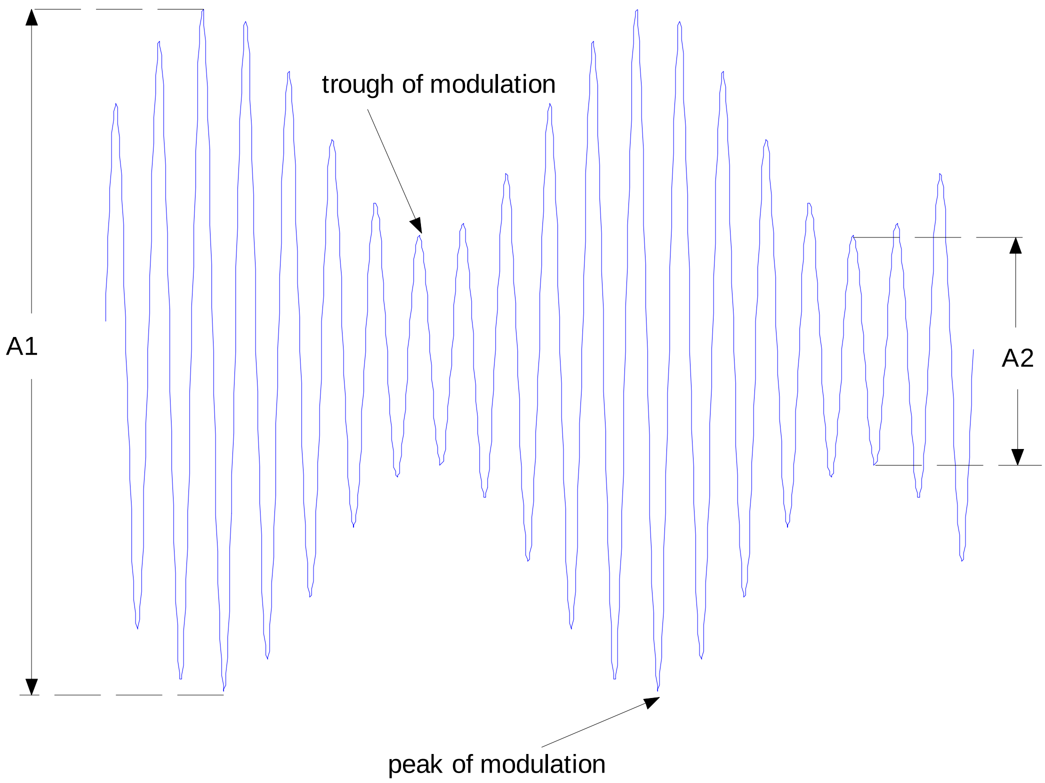
$$m = \frac{k_a\left|\tilde{s}(t)\right|_{\max}}{A_0}$$

m is called the modulation index.

Since $r(t)$ must be positive, we see that we must have $0 \leq m \leq 1$. If it happens that m exceeds 1, we say that we have overmodulation. The AM signal is:

$$x(t) = A_0\left(1 + ms(t)\right)\cos\omega_0 t$$

Historically, amplitude modulation is the first modulation system put into practice. It was used essentially because of the simplicity of the receiver structure. It is easily verified that AM is not linear since it does not satisfy the superposition principle. It is as linear as the function $f(x) = ax + b$. This function is not linear however it is incrementally linear, i.e. an increment of the input is linearly related to an increment of the output.



Sinewave modulated waveform

The signal $s(t)$ can be an energy or a power type of signal. In the first case, we can compute easily the spectrum of the modulated signal as a function of the baseband modulating one.

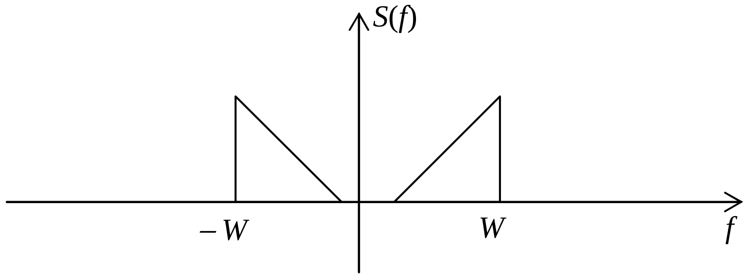
Starting from $x(t)$, we obtain

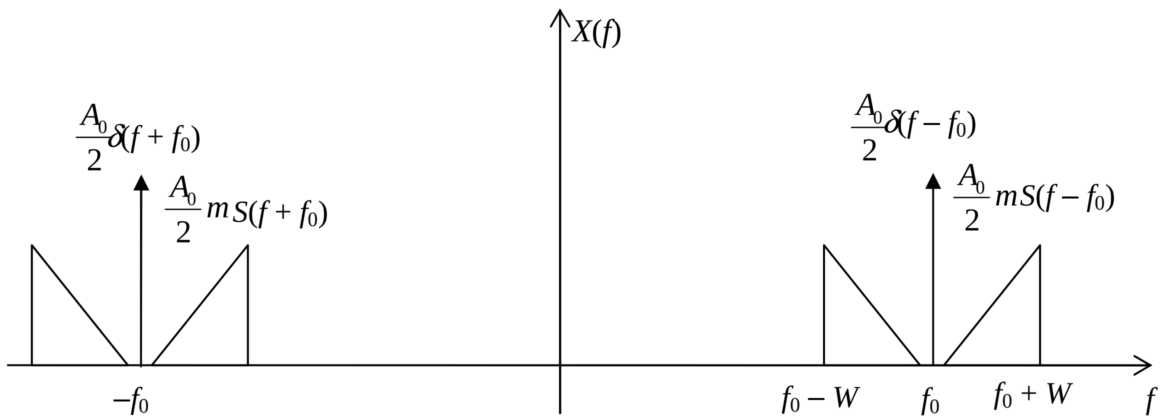
$$x(t) = A_0\left(1+ms(t)\right)\cos\omega_0 t = A_0\cos\omega_0 t + A_0 ms(t)\cos\omega_0 t$$

giving

$$X(f) = \frac{A_0}{2}\delta(f-f_0) + \frac{A_0}{2}\delta(f+f_0) + \frac{A_0}{2}mS(f-f_0) + \frac{A_0}{2}mS(f+f_0)$$

This relation is shown graphically below.





The above sketch shows the transformation from the baseband signal $s(t)$ to the bandpass signal $x(t)$. We can also remark that if we do not want to have a superposition of shifted spectra (aliasing), we must have $f_0 > W$. It is also apparent that the bandwidth B of the modulated signal is twice the bandwidth of the baseband signal.

$$B = 2W$$

If we consider only the positive half of the spectrum, we remark that the Hermitian symmetry of $S(f)$ is translated to f_0. So, the positive half is composed of two halves: *the upper sideband* above the carrier frequency and *the lower sideband* below the carrier frequency.

Power Computation

In order to analyze power signals, we may assume that $s(t)$ is periodic. We can start the analysis with the simplest real periodic signal: the sinewave. So, let us assume that $s(t) = \cos\omega_m t$ where $\omega_m < \omega_0$.

$$x(t) = A_0\left(1 + m\cos\omega_m t\right)\cos\omega_0 t = A_0\cos\omega_0 t + A_0 m\cos\omega_m t\cos\omega_0 t$$

Using trigonometric identities, we obtain;

$$x(t) = A_0\cos\omega_0 t + \frac{A_0}{2}m\cos\left[\left(\omega_0 - \omega_m\right)t\right] + \frac{A_0}{2}m\cos\left[\left(\omega_0 + \omega_m\right)t\right]$$

The signal in this case is composed of 3 sinewaves: the carrier with amplitude A_0 and the two sidebands with amplitude $\frac{A_0}{2}m$ each. The spectrum consists of only Dirac impulse functions. A more general case is the one of a bandlimited periodic signal. We can express $s(t)$ as:

$$s(t) = \sum_{k=1}^{N} a_k\cos\left(k\omega_m t + \theta_k\right)$$

The signal has a zero dc value and the number of harmonics N is given by $N = \left\lfloor \frac{W}{f_m} \right\rfloor$. The notation $\lfloor\ \ \rfloor$ stands for the *floor* (i.e. the integer just below) of the number written inside. The modulated signal is now given by:

$$x(t) = A_0 \cos \omega_0 t + A_0 m \sum_{k=1}^{N} a_k \cos\left(k\omega_m t + \theta_k\right) \cos \omega_0 t$$

$$x(t) = A_0 \cos \omega_0 t + \frac{A_0}{2} m \sum_{k=1}^{N} a_k \cos\left[\left(\omega_0 - k\omega_m\right)t + \theta_k\right]$$
$$+ \frac{A_0}{2} m \sum_{k=1}^{N} a_k \cos\left[\left(\omega_0 + k\omega_m\right)t + \theta_k\right]$$

The above formula is general enough to allow us to compute the power of the modulated signal. If we assume that the different sinewaves are independent, the total power will be given by the sum of the power of the different components.

$$P_x = \frac{A_0^2}{2} + 2 \times \frac{A_0^2}{8} m^2 \sum_{k=1}^{N} a_k^2$$

In the above relation, we can recognize the carrier power $P_c = \frac{A_0^2}{2}$ and the sideband power $P_{sb} = \frac{A_0^2}{8} m^2 \sum_{k=1}^{N} a_k^2$. So, the total power of the signal is: $P_x = P_c + 2P_{sb}$. The sideband power can also be expressed as a function of the power of the normalized baseband signal $P_s = \sum_{k=1}^{N} \frac{a_k^2}{2}$, i.e. $P_{sb} = \frac{A_0^2}{4} m^2 P_s$. So, in terms of the carrier power and the sideband power, we obtain:

$$P_x = P_c + \frac{A_0^2}{2} m^2 P_s = \frac{A_0^2}{2} + \frac{A_0^2}{2} m^2 P_s$$

Given that the signal $s(t)$ is normalized with a maximum value of 1, its power is less than 1 ($P_s = \frac{1}{T_m}\int_{T_m} s^2(t)dt \le \frac{1}{T_m}\int_{T_m} |s(t)|^2_{\max} dt = 1$). The modulation index m is also less or equal to 1. This means that the power transmitted by the two sidebands is smaller than the power used to transmit a carrier that conveys no information. More than 50% of the total power is used to transmit the carrier. We can evaluate the efficiency of the system using the following efficiency coefficient:

$$\eta = \frac{2P_{sb}}{P_c + 2P_{sb}} = \frac{m^2 P_s}{1 + m^2 P_s}$$

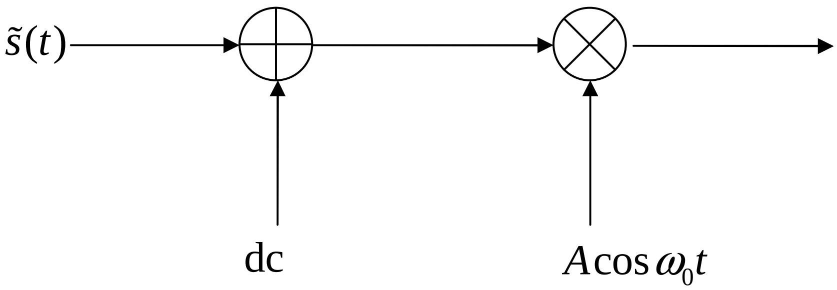
η cannot exceed the value of 1/2.

Example: if $s(t) = \cos\omega_m t$, $P_s = \frac{1}{2}$, then $\eta = \frac{m^2}{2+m^2}$. If we use $m = 1$, the efficiency becomes $\eta = \frac{1}{3}$. In general, AM systems use a modulation index smaller than 1. The typical value is 30%. AM is used to transmit audio signals that have an average power $P_s << 1$. In this case, the efficiency drops to a very small value.

Broadcast AM radio use carrier frequencies in the medium waves (530 kHz to 1.71 MHz) or in short waves (3 to 30 MHz). We can encounter stations transmitting in long waves (148.5 kHz to 283.5 kHz). The bandwidth *B* allocated to every channel is 10 kHz. This means that each sideband occupies a bandwidth *W* of 5 kHz.

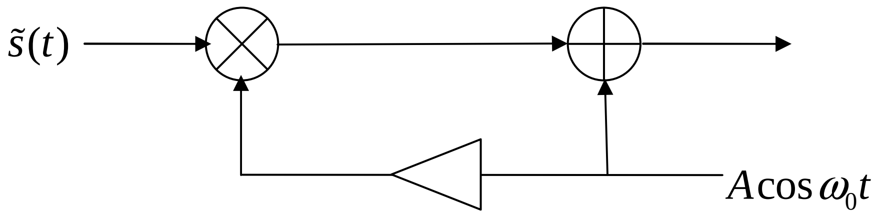
AM production:

The following block diagram reproduces the defining equation of AM.



Another structure can be derived by manipulating the defining equation.

$$x(t) = A_0\left(1 + ms(t)\right)\cos\omega_0 t = A_0\cos\omega_0 t + A_0 ms(t)\cos\omega_0 t$$



These block diagrams are essentially theoretical. In practice, radically different methods are used. A common technique is to use a saturating class C amplifier. The theory behind this type of amplifier will be covered in a future course. Another method for AM production is to use a memoriless nonlinear amplifier.

Consider a system with the following input output transfer:

$$z = a_0 + a_1 w + a_2 w^2$$

Let $w = s_1 + s_2$, then $z = a_0 + a_1 s_1 + a_1 s_2 + a_2 s_1^2 + a_2 s_2^2 + 2a_2 s_1 s_2$

If $s_1(t) = A\cos\omega_0 t$ and $s_2(t) = \tilde{s}(t)$, then

$$z(t) = a_0 + a_1 A\cos\omega_0 t + a_1\tilde{s}(t) + a_2 A^2\cos^2\omega_0 t + a_2\tilde{s}^2(t) + 2a_2 A\tilde{s}(t)\cos\omega_0 t$$

Now, $\cos^2\theta = \frac{1}{2} + \frac{1}{2}\cos 2\theta$, so $z(t)$ is the sum of four different components:

Dc component: $a_0 + \frac{a_2 A^2}{2}$

Baseband component: $a_1\tilde{s}(t) + a_2\tilde{s}^2(t)$

Component around ω_0: $a_1 A\cos\omega_0 t + 2a_2 A\tilde{s}(t)\cos\omega_0 t$

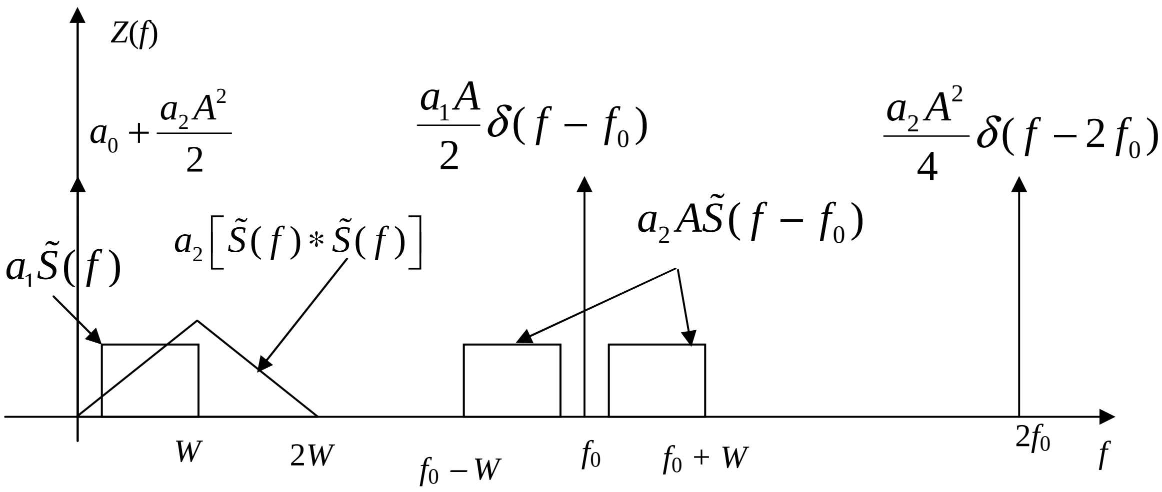
Component at $2\omega_0$: $\frac{a_2 A^2}{2}\cos 2\omega_0 t$

If we use a bandpass filter tuned at f_0, we can select the component around ω_0. This component is:

$$\begin{aligned} x(t) &= a_1 A\cos\omega_0 t + 2a_2 A\tilde{s}(t)\cos\omega_0 t \\ &= a_1 A\left(1 + \frac{2a_2}{a_1}\tilde{s}(t)\right)\cos\omega_0 t \\ &= A_0\left(1 + ms(t)\right)\cos\omega_0 t \end{aligned}$$

In the above expression, $A_0 = a_1 A$ and $m = \frac{2a_2\left|s(t)\right|_{\max}}{a_1}$. In order to specify the filter, we have to compute the spectrum of the signal $z(t)$.

$$\begin{aligned} Z(f) = &\left[a_0 + \frac{a_2 A^2}{2}\right]\delta(f) \\ &+ a_1\tilde{S}(f) + a_2\left[\tilde{S}(f) * \tilde{S}(f)\right] \\ &+ \frac{a_1 A}{2}[\delta(f - f_0) + \delta(f + f_0)] + a_2 A\left[\tilde{S}(f - f_0) + \tilde{S}(f + f_0)\right] \\ &+ \frac{a_2 A^2}{4}[\delta(f - 2f_0) + \delta(f + 2f_0)] \end{aligned}$$



Positive frequency spectrum of *Z(f)*

The convolution term is the Fourier transform of the square $a_2\tilde{s}^2(t)$. If we look at the above spectrum, we remark that the AM signal is present around f_0. It has a bandwidth $B = 2W$. We notice also that we must have $f_0 > 3W$ if we want to avoid overlap of spectra. If the nonlinear system contains a higher degree, then this condition will change.

There exist many other methods for producing AM signals. They are better analyzed in an electronic circuit course.

<u>AM demodulation:</u>

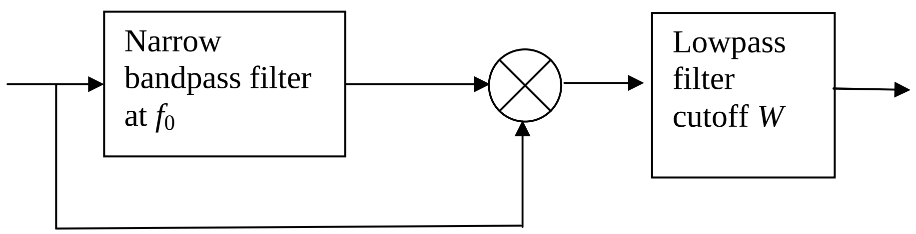
There exist several techniques for AM modulation. They fall into two different classes: Homodyne (Synchronous, Coherent) demodulation and Non Coherent demodulation.

<u>Coherent demodulation:</u>

In this technique, we multiply the received AM signal by a carrier generated at the receiver. The local carrier must have the same frequency and same phase as the one of the AM signal. Let the received signal be: $x(t) = A_0[1 + ms(t)]\cos\omega_0 t$ and the locally generated carrier: $y(t) = B\cos\left[(\omega_0 + \Delta\omega)t + \varphi_0\right]$. $\Delta\omega$ is a frequency error and φ_0 is a phase error. The result of the product is:

$$\begin{aligned} z(t) &= A_0 B[1 + ms(t)]\cos\omega_0 t\cos\left[(\omega_0 + \Delta\omega)t + \varphi_0\right] \\ &= \frac{A_0 B}{2}[1 + ms(t)]\cos\left((\Delta\omega)t + \varphi_0\right) + \frac{A_0 B}{2}[1 + ms(t)]\cos\left(2\omega_0 t + (\Delta\omega)t + \varphi_0\right) \end{aligned}$$

It is composed of two components. If we use a lowpass filter with a bandwidth *W*, we will recover the first term. If we want to demodulate the signal, the frequency error $\Delta\omega$ must be zero and the phase error φ_0 must be as small as possible (far from the value of $\pi/2$). The recovered signal will be a signal proportional to *s*(*t*) plus a dc component. A capacitor in series is enough to eliminate the dc component. The locally generated carrier on the other hand must be exactly synchronized to the received carrier. In most implementations, we can use the received carrier if the baseband signal is itself bandpass. This is the case of most audio[1] and speech[2] signals.

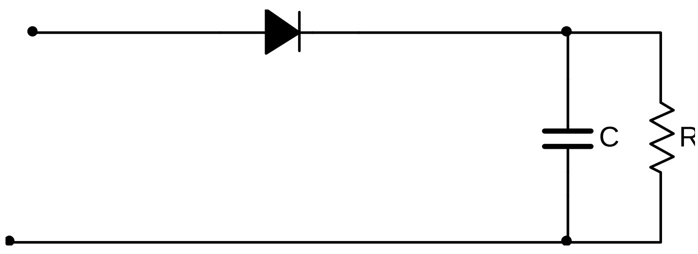


The above figure shows a typical synchronous demodulator.

Non coherent demodulator:

[1] Audio signals are usually bandpass between 50 Hz and 15 kHz.
[2] Speech signals are bandlimited between 300 and 3400 Hz.

The envelop detector is commonly use in AM receivers and is in fact the first demodulator in the history of radio communication.



The envelop detector is composed of a diode, a resistor and a capacitor. It is essentially a peak detector. This means that the time constant of the circuit (RC) must be much larger than the period of the carrier ($1/f_0$). On the other hand, the circuit should not distort the information signal $s(t)$. This implies that the time constant must be smaller than the period of the highest frequency present in the signal (W).

$$\frac{1}{W} > \mathrm{RC} >> \frac{1}{f_0}$$

You will have the occasion to experiment this circuit in the lab. If the above condition is satisfied, the signal obtained at the output will be proportional to the envelop of the AM signal $r(t)$. Here again a dc blocking capacitor is needed to eliminate the dc value present in the demodulated signal.

Double sideband suppressed carrier modulation (DSB-SC):

In AM, we spend more than half of the total power transmitting a carrier that conveys no information. The following method transmits just the sidebands without transmitting the carrier. The DSB-SC signal is then:

$$x(t) = A_0 s(t) \cos \omega_0 t$$

We see immediately than this method is a linear modulation scheme. Furthermore, from the defining equation, the DSB-SC modulated signal is a bandpass signal written in quadrature form. $a(t) = A_0 s(t)$ and $b(t) = 0$. In the frequency domain, the spectrum of the DSB-SC signal is obtained by a straightforward application of the modulation theorem.

$$X(f) = \frac{A_0}{2}[S(f - f_0) + S(f + f_0)]$$

We see that all the power is in the transmitted sidebands. There is no transmitted carrier. The transmitted power is:

$$P_x = 2P_{sb} = \frac{1}{2} A_0^2 P_s$$

Most transmitters are limited by the *peak power* they can transmit. The peak power is given by the square of the maximum of the envelop $P_{peak} = |r(t)|^2_{\max}$. For DSB-SC, the maximum of the envelop is A_0 while for AM this maximum is $A_0(1 + m)$. So, for AM, $P_{peak} = A_0^2 (1+m)^2$ and for DSB-SC, $P_{peak} = A_0^2$. The ratio of sideband power over the peak power for the two modulations is given by:

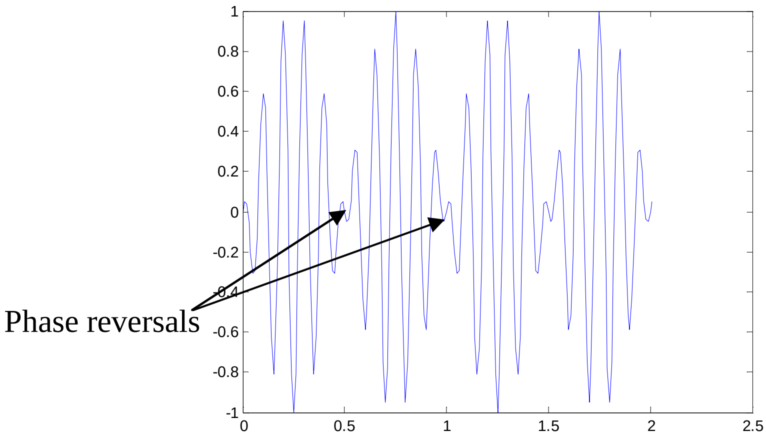
$$\frac{P_{sb}}{P_{peak}} = \begin{cases} \dfrac{P_s}{4} & \text{DSB-SC} \\ \dfrac{m^2 P_s}{4(1+m)^2} & \text{AM} \end{cases}$$

So, for a given peak power, a DSB-SC transmitter produces more than four times the sideband power of an AM transmitter.

Except for a missing impulse, the spectrum of DSB-SC and the one of AM look alike, however, in the time domain, the is a fundamental difference. The DSB-SC envelop and phase are given by:

$$r(t) = A_0 \left| s(t) \right| \qquad \varphi(t) = \begin{cases} 0 & s(t) > 0 \\ \pi & s(t) < 0 \end{cases}$$

Every time the signal $s(t)$ changes sign, the modulated signal undergoes a phase reversal.



We see that we cannot demodulate a DSB-SC signal with a simple envelop detector. We need a more sophisticated demodulator.

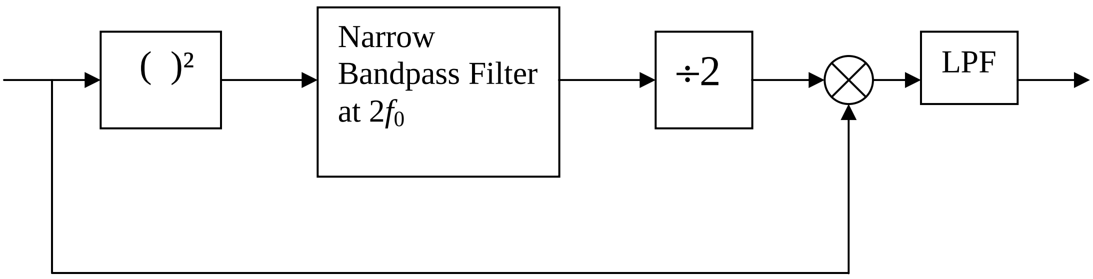
The most commonly used demodulator for DSB is the homodyne one. However, since there is no carrier transmitted along with the signal, the local carrier generation is more complex.

Let $x(t) = A_0 s(t) \cos \omega_0 t$. The signal $s(t)$ is assumed to be a power signal with zero average. If we square the signal $x(t)$, we obtain:

$x^2(t) = A_0^2 s^2(t) \cos^2 \omega_0 t = \dfrac{A_0^2 s^2(t)}{2}\left(1 + \cos 2\omega_0 t\right)$. The signal $s^2(t)$ is completely positive. This means that it has an average value that is different from zero. We can express it as: $s^2(t) = \langle s^2(t) \rangle + s_1(t)$. The signal s_1(t) has a zero average. In the frequency domain, we obtain:

$$Z(f) = \frac{A_0^2 < s^2(t) >}{2}\delta(f) + \frac{A_0^2}{2}S_1(f)$$
$$+ \frac{A_0^2 < s^2(t) >}{4}[\delta(f - 2f_0) + \delta(f + 2f_0)]$$
$$+ \frac{A_0^2}{4}[S_1(f - 2f_0) + S_1(f + 2f_0)]$$

We observe a spectrum around $2f_0$ that is practically the one of an AM signal with a carrier area of $\frac{A_0^2 < s^2(t) >}{4}$. So, we can use a narrow bandpass filter tuned at $2f_0$ to extract a carrier. The filter will be followed by a frequency divider by 2 (a simple D flip-flop).



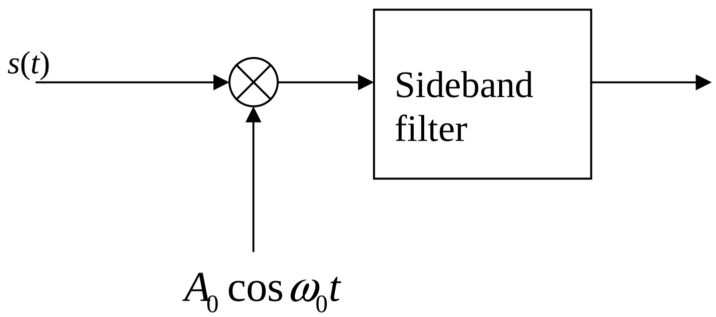
Squaring loop demodulator

In the above system, there remains a small problem. When we divide a frequency by two, we have an ambiguity of π in the phase. This is due to the fact that $\frac{(2\omega_0 t + 2k\pi)}{2} = \omega_0 t + k\pi$. This means that we can have a signal reversal at the demodulator. If the destination of the demodulator is the human ear, this reversal will not be noticed by the auditor. However, if the system is used to transmit data, and if we assign "1" to a positive value and "0" to a negative value, then the data will be negated. One way to prevent this is to send a *prefix* word known to the receiver. If it is received correctly, we keep the output of the squaring loop. Otherwise, we invert the output carrier from the squaring loop.

One way to avoid problems in carrier recovery is to send a subcarrier at a frequency related to the one we want to recover.

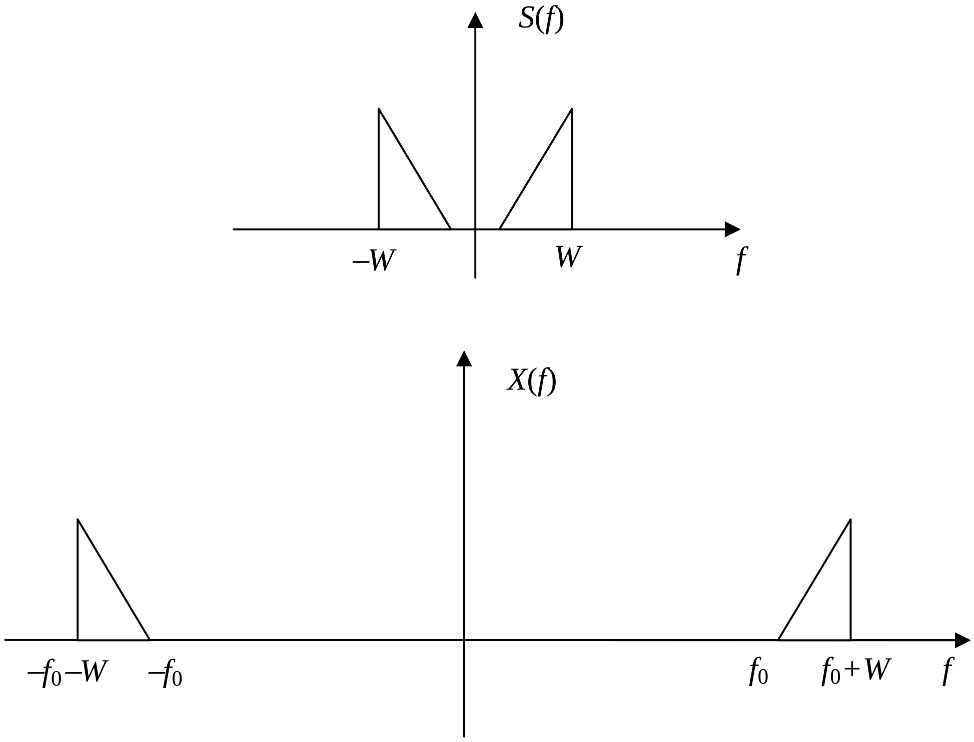
Single Sideband Modulation (SSB):

When we are transmitting real signals in DSB-SC, the two sidebands are related and if we know one, we can deduce the other. So, this modulation method transmits only one of the two sidebands, either the upper sideband (USB-SSB) or the lower sideband (LSB-SSB). Basically, an SSB modulator can be implemented using a DSB-SC one followed by a sideband filter.



SSB modulator

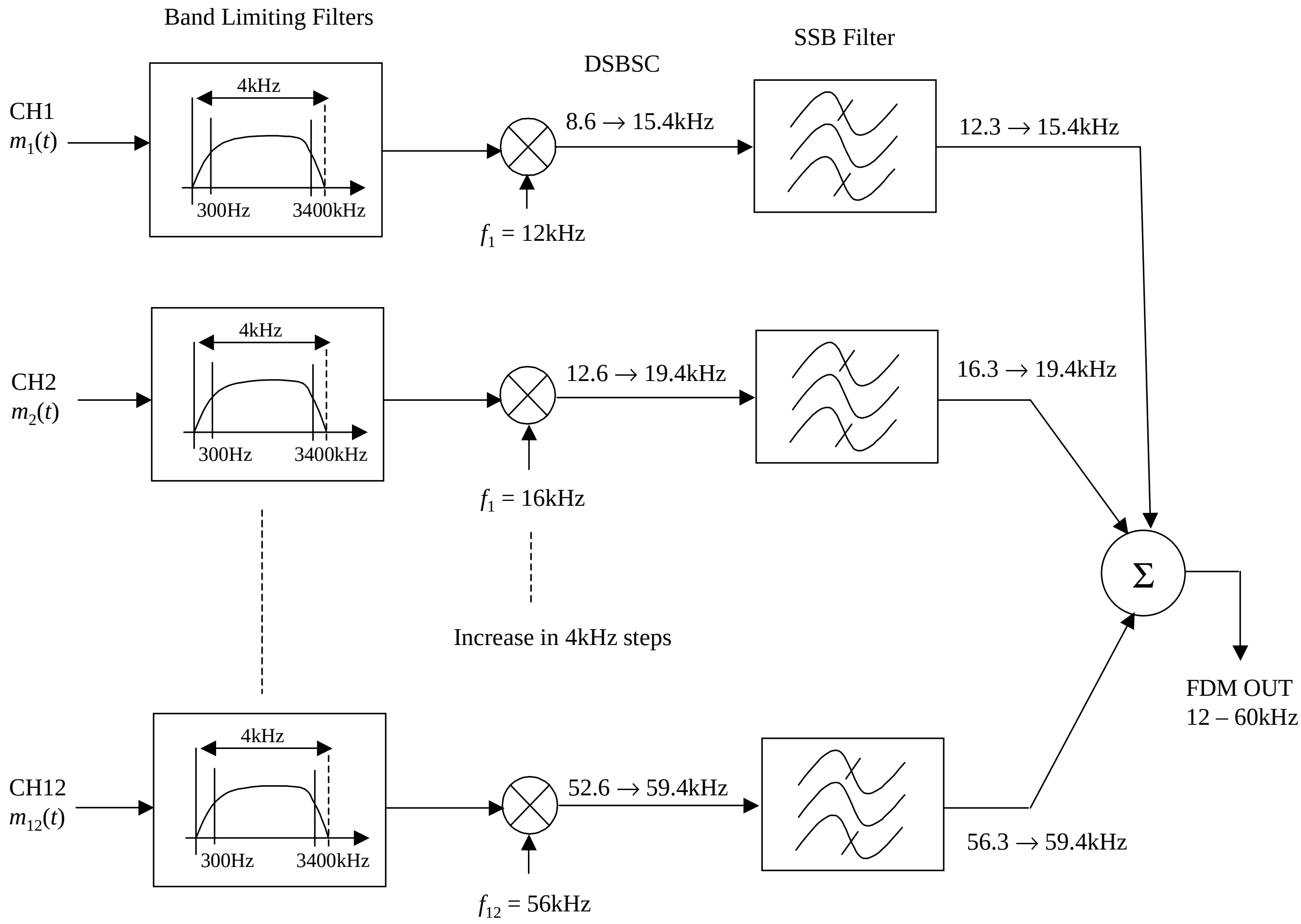
It is quite simple to represent the different operations in the frequency domain. The following sketch shows a USB-SSB signal in the frequency domain.



USB-SSB Spectrum

We can observe from the above sketch that the bandwidth of the SSB signal is the same as the one of the baseband signal. So, for the same information, the SSB modulated signal uses half the bandwidth of the DSB modulated signal. This is why SSB is used in crowded spectrum environment such as amateur radio. It has been used also in Frequency Division Multiplexing (FDM) systems to transmit different voiceband signals[3]. If we observe the following figure, we can observe that the different shifted spectra do not overlap. They can be transmitted using a single wire. To avoid any problem in carrier recovery, a subcarrier is usually transmitted in a separate channel.

[3] A voiceband signal is a signal that conveys human speech. Its spectrum is essentially different from zero in a band between 300 and 3400 Hz.



12 Channel FDM system

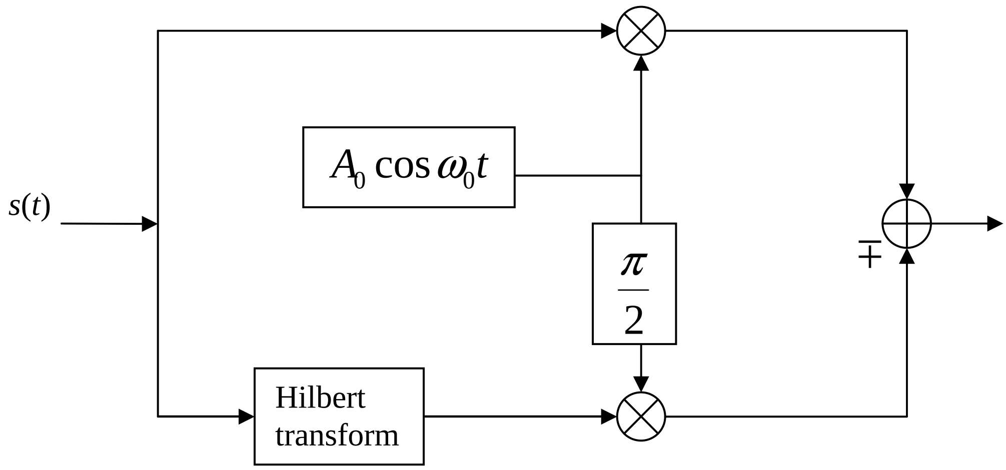
To see how SSB is demodulated, we have to express the SSB signal in the time domain. In order to do so, it is simpler to obtain first the corresponding analytic signal $X_+(f)$ and then use the fact that $x(t) = \mathrm{Re}\left[x_+(t)\right]$. When we observe the USB-SSB spectrum, we notice that the analytic signal $X_+(f)$ is just the analytic signal $S_+(f)$ shifted and scaled in the frequency domain. So, $X_+(f) = A_0 S_+(f - f_0)$. In the time domain, it gives $x_+(t) = A_0 s_+(t)\exp\left[j\omega_0 t\right]$. Replacing $s_+(t)$ by its expression $s(t) + j\hat{s}(t)$, we obtain:

$$\begin{aligned} x(t) &= A_0 \,\mathrm{Re}\left[\left(s(t) + j\hat{s}(t)\right)\left(\cos\omega_0 t + j\sin\omega_0 t\right)\right] \\ &= A_0\left[s(t)\cos\omega_0 t - \hat{s}(t)\sin\omega_0 t\right] \end{aligned}$$

The above expression is the one of a USB-SSB modulated signal. It is a simple matter to show that the expression of the LSB-SSB modulated signal is:

$$x(t) = A_0 \left[s(t) \cos \omega_0 t + \hat{s}(t) \sin \omega_0 t \right]$$

The above two expressions suggest that SSB modulators can be implemented using the following block diagram:



The minus sign is for USB-SSB while the plus is for LSB-SSB. This method of SSB production is called the *Phasing Method*.

SSB Demodulation:

We are going to consider only the coherent demodulation method. A general SSB signal is: $x(t) = A_0 \left[s(t) \cos \omega_0 t \mp \hat{s}(t) \sin \omega_0 t \right]$. We multiply this signal by a carrier $y(t) = B \cos\left[(\omega_0 + \Delta\omega) t + \varphi_0 \right]$. Let us consider first the case of zero frequency offset ($\Delta\omega = 0$).

The result of the product contains terms at low frequency and terms around $2f_0$. The low frequency component is:

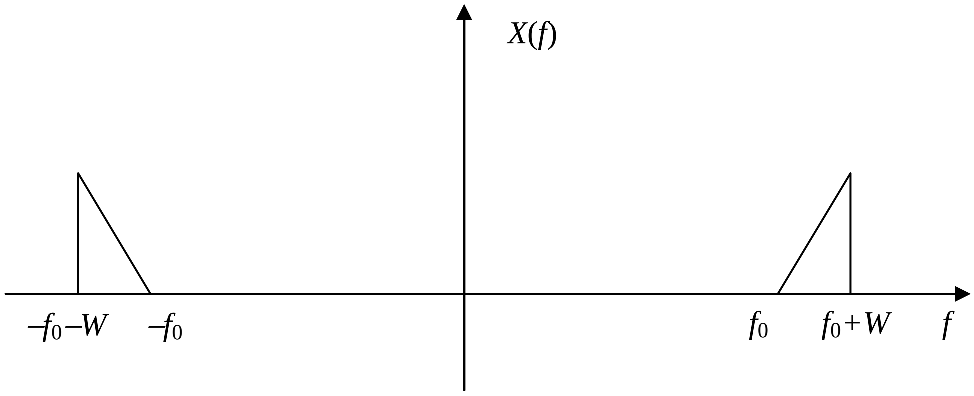
$$\frac{A_0 B}{2} \left[s(t) \cos \varphi_0 \mp \hat{s}(t) \sin \varphi_0 \right]$$

So, the output of the coherent demodulator will contain a linear combination of $s(t)$ and $\hat{s}(t)$. If the end destination is the human ear, this signal will sound exactly as $s(t)$ alone. This is due to the fact that the human ear is insensitive to phase shifts in the signal. In other cases, the phase error cannot be tolerated.

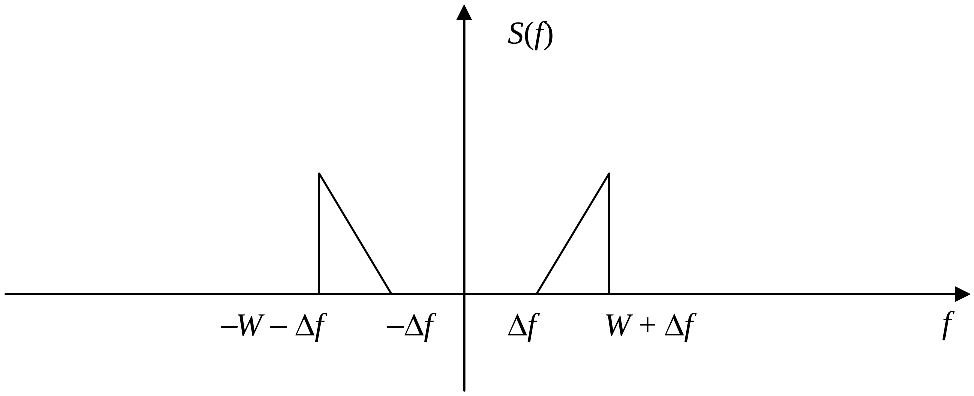
The analysis of the frequency error is easier to study in the frequency domain. Using the modulation theorem, the result of the SSB signal multiplied by a carrier is:

$$\mathscr{F}\left[x(t)\cos\left(\omega_0+\Delta\omega\right)t\right]=\frac{1}{2}X(f-f_0-\Delta f)+\frac{1}{2}X(f+f_0+\Delta f)$$

Starting from a USB-SSB signal with the spectrum shown below:



We obtain the following spectrum after elimination of the components around $2f_0$:

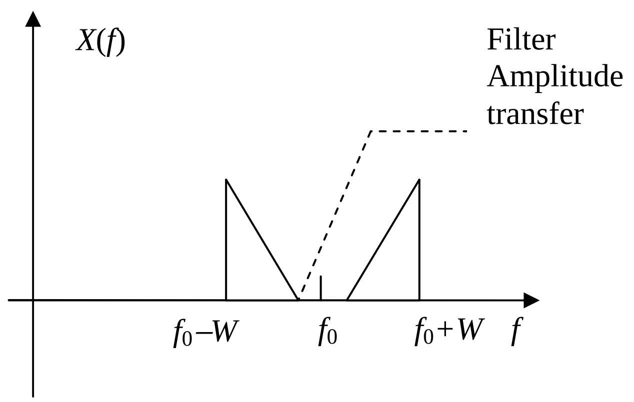


We remark that all the frequencies of the message are translated by a constant shift. This constant shift does not make the speech unintelligible. However, when Δf is positive, it makes everybody sound like "Donald Duck", hence the name; *Donald Duck* distortion. On the other hand, music will be completely distorted since the harmonic relations between notes will disappear.

Advantages and disadvantages of SSB:

We see that SSB is a linear modulation system that saves on bandwidth. The transmission bandwidth is equal to the signal one. However, in order to achieve this result we need very complex hardware.

In the filtering method, we have to transmit completely one sideband and eliminate completely the other. This means that the transition region of the filter is zero. The only way to achieve reasonable filters is to use this method for signals that have no energy around zero frequency.

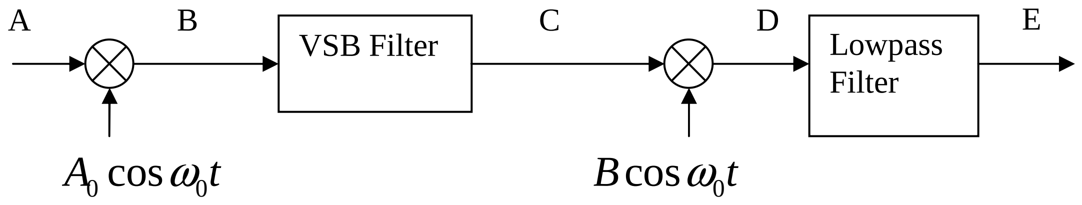


If we want to apply the phasing method, we will encounter the same problem. It is impossible to build a filter that phase shifts all frequencies from zero to *W*. This means that the phase response of the filter has also a zero transition band. We can approximate the Hilbert transformer if the signal $s(t)$ has the same character as above. It must not have any energy around zero. So, SSB is useful if we want to transmit speech. Audio signals can be transmitted at the expense of a quite complicated hardware. Data cannot be transmitted in the shape of a sequence of pulses. This signal possesses power at zero frequency. If we want to transmit signals that have spectra that are different from zero around dc, one solution is to use Vestigial Sideband.

Vestigial Sideband (VSB):

In VSB modulation, we use filter that transmit most of one sideband and a very small amount of the other (a vestige).

In order to determine the filter characteristics, we must analyze a complete modulation and demodulation system. The demodulation method is always coherent. We multiply the received signal by a carrier $B\cos\omega_0 t$ and we lowpass filter the result to eliminate terms around $2\omega_0$.



In the above block diagram, we must determine the signals at different points.

At A, we have the baseband signal $x_A(t) = s(t)$ with spectrum $X_A(f) = S(f)$.

At B, we obtain the DSB-SC signal $x_B(t) = A_0 s(t)\cos\omega_0 t$ with spectrum

$X_B(f) = \frac{A_0}{2} S(f - f_0) + \frac{A_0}{2} s(f + f_0)$.

At C, we have the VSB signal obtained by filtering the DSB-SC signal. We are going to characterize it in the frequency domain only: $X_C(f) = \frac{A_0}{2}\left[H(f)S(f - f_0) + H(f)S(f + f_0)\right]$

At D, we use the modulation theorem of Fourier transforms and we obtain:

$$X_D(f) = \frac{A_0 B}{4}[H(f - f_0)S(f - 2f_0) + H(f - f_0)S(f) + H(f + f_0)S(f) + H(f + f_0)S(f + 2f_0)]$$

The lowpass filter eliminates all the terms around $\pm 2f_0$. So, the signal at E will be: $X_E(f) = \frac{A_0 B}{4}\left[H(f - f_0)S(f) + H(f + f_0)S(f)\right]$

If we want to have a distortionless transmission, this signal must be proportional to $s(t)$. This means that:

$$H(f + f_0) + H(f - f_0) = \text{constante}$$

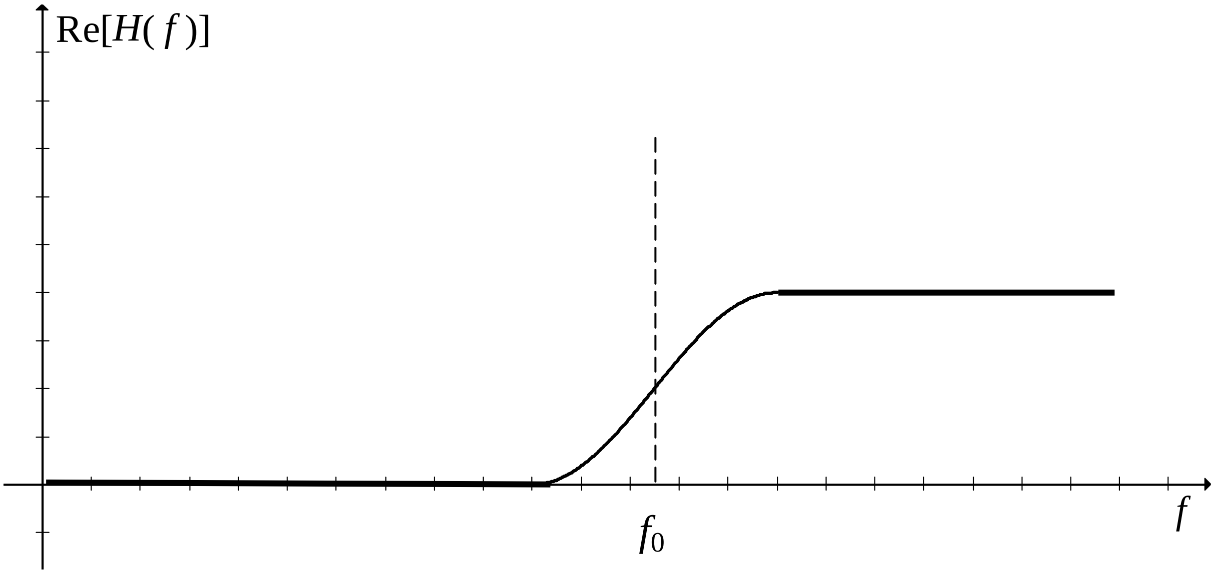
After some manipulations, we obtain that the transfer function of the filter must satisfy:

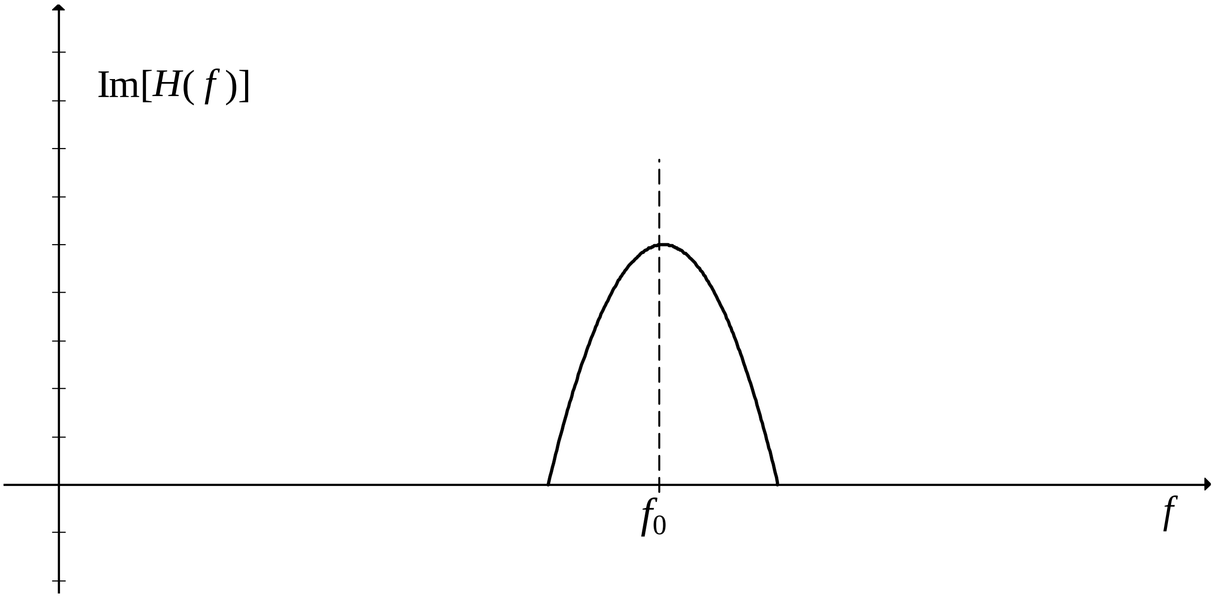
$$H(f_0 + x) + H^*(f_0 - x) = 2\,\mathrm{Re}\left[H(f_0)\right] \text{ for } f \text{ around } f_0.$$

If $H(f) = R(f) + jX(f)$, then

$$R(f_0 + x) + R(f_0 - x) = 2R(f_0)$$

$$X(f_0 + x) = X(f_0 - x)$$





The above graph shows the different symmetries that the real and imaginary part of the transfer function must satisfy. The real part must show an odd symmetry with respect to the point (f_0, Re[$H(f_0)$]) while the imaginary part must have the vertical line passing by f_0 as a symmetry axis.

The VSB signal has been characterized in the frequency domain. We have seen that it can be demodulated using coherent demodulation. We can also have the expression of the VSB signal in the time domain.

Being a general bandpass signal, it can be expressed in quadrature form and it is completely described by its complex envelop. The VSB signal appears at point C in our block diagram. Its complex envelop is given by the filtering of the complex envelop of the DSB-SC signal at B by the lowpass equivalent filter.

The complex envelop of the signal at B is $m_{xB}(t) = A_0 s(t)$ with a spectrum $M_{xB}(f) = A_0 S(f)$. The lowpass equivalent filter $H_{lp}(f)$ is the filter $H(f)$ translated down. Using the symmetries derived above, we can express the equivalent lowpass filter transfer function as:

$$H_{lp}(f) = R(f_0) + A(f) + jX_{lp}(f) = R(f_0)\left[1 + jQ(f)\right]$$

In the above expression, $A(f)$ is an odd function ($R(f)$ translated down to $f = 0$ and shifted down by $R(f_0)$) while $X_{lp}(f)$ is an even function. This implies that $Q(f)$ satisfies $Q(f) = Q^*(-f)$. The complex envelop of the VSB signal is then $M_x(f) = A_0 R(f_0) S(f)\left[1 + jQ(f)\right]$ or in the time domain: $m_x(t) = A_0 R(f_0) s(t) + jA_0 R(f_0)\mathcal{F}^{-1}\left[Q(f)S(f)\right]$. The function $Q(f)S(f)$ satisfies the condition of Hermitian symmetry. This implies that its inverse Fourier transform is real. Let $q(t) = \mathcal{F}^{-1}\left[Q(f)S(f)\right]$ then the complex envelop of the VSB signal is: $m_x(t) = A_0 R(f_0)\left[s(t) + jq(t)\right]$ and the VSB signal can be written as:

$$x(t) = A_0 R(f_0)\left[s(t)\cos\omega_0 t - q(t)\sin\omega_0 t\right].$$

The signal $q(t)$ is the output of the filter with transfer function $Q(f)$.

$$Q(f) = \frac{A(f)}{jR(f_0)} + \frac{X_{lp}(f)}{R(f_0)}$$

Two extreme cases are interesting:

If we want to keep the upper sideband and eliminate completely the lower one, we must have $H(f)=\begin{cases}2R(f_0) & f>f_0\\ 0 & 0<f<f_0\end{cases}$

This implies that $Q(f)=-j\,\text{sgn}(f)$ and $q(t)=\hat{s}(t)$. The VSB signal in this case is a USB-SSB signal.

The other extreme case is when we want to keep both sidebands. At that time, $Q(f)=0$ and the signal is just a DSB-SC one.

In our analysis, we have assumed that we favor the upper sideband. We can obtain the same results for the lower sideband. The modulated signal bandwidth is intermediate: $W < B < 2W$.

Envelop demodulation of linear modulation + carrier:

If we add a large amplitude carrier to the inphase component of a bandpass signal (DSB, SSB, VSB) we obtain:

$$x(t)=B\cos\omega_0 t+A_0 s(t)\cos\omega_0 t\pm A_0 q(t)\sin\omega_0 t$$

The envelop of the signal is:

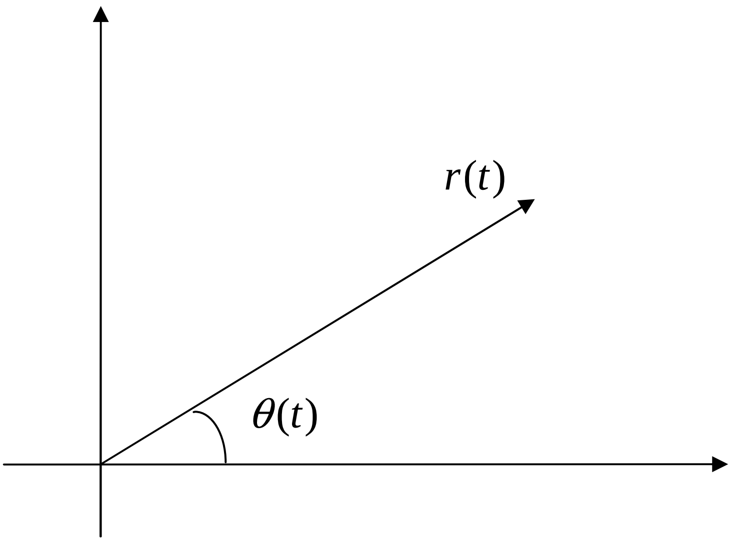
$$r(t)=\sqrt{\left(B+A_0 s(t)\right)^2+A_0^2 q^2(t)}=\left|B+A_0 s(t)\right|\sqrt{1+\frac{A_0^2 q^2(t)}{\left(B+A_0 s(t)\right)^2}}$$

If $B >> A_0$, the expression inside the absolute value will always be positive and $r(t)\approx\left(B+A_0 s(t)\right)+\dfrac{A_0^2 q^2(t)}{2\left(B+A_0 s(t)\right)}\approx B+A_0 s(t)$ .This technique is used in the transmission of analog television where the TV signal is transmitted in VSB+Carrier.

Exponential modulation

Instantaneous frequency

Up to now, we have defined the frequency as the speed of rotation of a phasor (constant frequency phasor) $\phi(t) = A_0 \exp\left[j\left(\omega_0 t + \theta_0\right)\right]$. We are going to generalize this definition to general complex functions of the real variable *t*. Consider such function: $z(t) = r(t)\exp\left(j\theta(t)\right)$. It is obvious that it is a generalization of the constant frequency phasor. It can be represented graphically as a vector with modulus (amplitude) $r(t)$ and argument $\theta(t)$.



The argument $\theta(t)$ of $z(t)$ is called the instantaneous phase. Since this phase varies, the generalized phasor is going to rotate. However, it is not going to rotate at a constant speed. We can thus define an instantaneous speed of rotation for this function. It is the instantaneous frequency:

$$\omega(t) = \frac{d\theta}{dt} \text{ rd/s}$$

We can measure this frequency in Hertz:

$$f(t) = \frac{\omega(t)}{2\pi} = \frac{1}{2\pi}\frac{d\theta}{dt}$$

In this section, we are interested in constant amplitude phasors. Furthermore, we assume that the instantaneous frequency has an average value f_0 with a deviation around it $d(t)$:

$$f(t) = f_0 + d(t)$$

The average value of $d(t)$ is zero. This means that the instantaneous phase can be expressed as:

$$\theta(t) = 2\pi f_0 t + \phi(t) = \omega_0 t + \phi(t)$$

$\phi(t)$ is called the instantaneous phase deviation and we have:

$$d(t) = \frac{1}{2\pi}\frac{d\phi}{dt}$$

To this generalized phasor, we can associate the following signal:

$$x(t) = \mathrm{Re}\left[w(t)\right] = r(t)\cos\left(\omega_0 t + \phi(t)\right)$$

This signal has the general shape of a bandpass signal. In order for $x(t)$ to be bandpass, its quadrature components must be bandlimited to a frequency W < f_0. The quadrature components are:

$$a(t) = r(t)\cos\phi(t) \text{ and } b(t) = r(t)\sin\phi(t).$$

This condition is not always satisfied. However, in practical exponential modulation, the carrier f_0 is usually very high (hundreds of MHz). So, we can consider that the obtained modulated signals are bandpass signals.

Frequency and Phase Modulation (FM & PM)

For both phase and frequency modulation, the modulated signal must have constant amplitude. The information is carried in the phase deviation. These modulations are called "*exponential Modulation*" because the signal has always the following shape:

$$x(t) = \mathrm{Re}\left[A_0 \exp\left(j\phi(t) \right) \exp\left(j\omega_0 t \right) \right]$$

Phase modulation is a modulation process that makes the phase deviation $\phi(t)$ proportional to the baseband signal $s(t)$.

$$\phi(t) = k_\phi \tilde{s}(t) = (\Delta\phi) s(t)$$

The constant $\Delta\phi = k_\phi \left| \tilde{s}(t) \right|_{\max}$ is called the maximum phase deviation. In order to avoid phase ambiguity, this constant cannot exceed π.

$$0 \le \Delta\phi \le \pi$$

The expression of a real phase modulated signal is:

$$x(t) = A_0 \cos\left(\omega_0 t + (\Delta\phi) s(t) \right)$$

Frequency modulation, on the other hand, is a modulation system where the frequency deviation is made proportional to the information signal.

$$d(t) = k_f \tilde{s}(t) = (\Delta f) s(t)$$

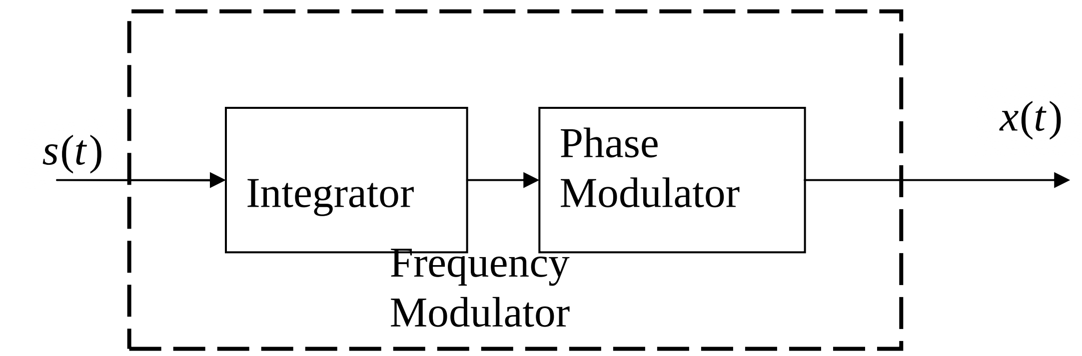
The constant $\Delta f = k_f \left| \tilde{s}(t) \right|_{\max}$ is called the maximum frequency deviation. The instantaneous frequency is:

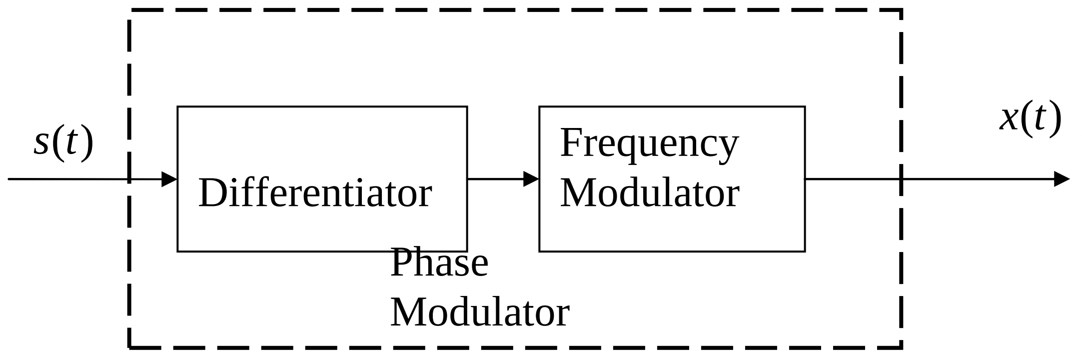
$$f(t) = f_0 + (\Delta f) s(t)$$

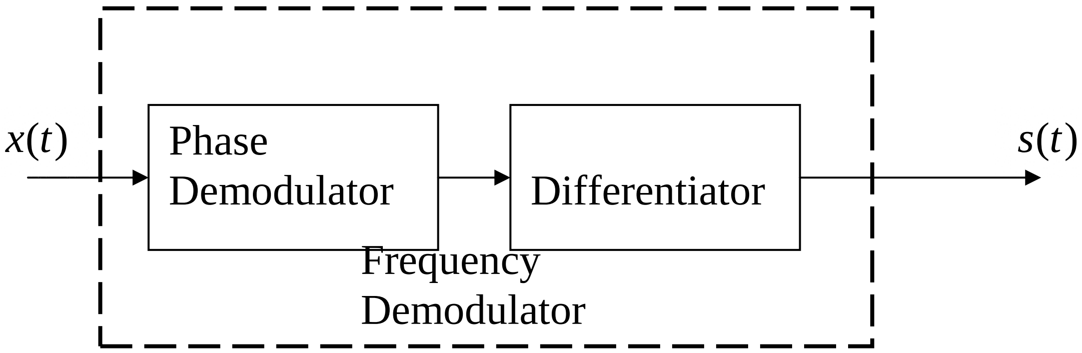
In order to have always a positive frequency, we must have $\Delta f \leq f_0$. The instantaneous phase deviation is: $\phi(t) = 2\pi(\Delta f)\int^t s(\lambda)d\lambda$. The lower bound of the integral is not indicated to take into account any initial phase. Sometimes, the lower bound is assumed to be $-\infty$. So, the expression of a frequency modulated signal is:

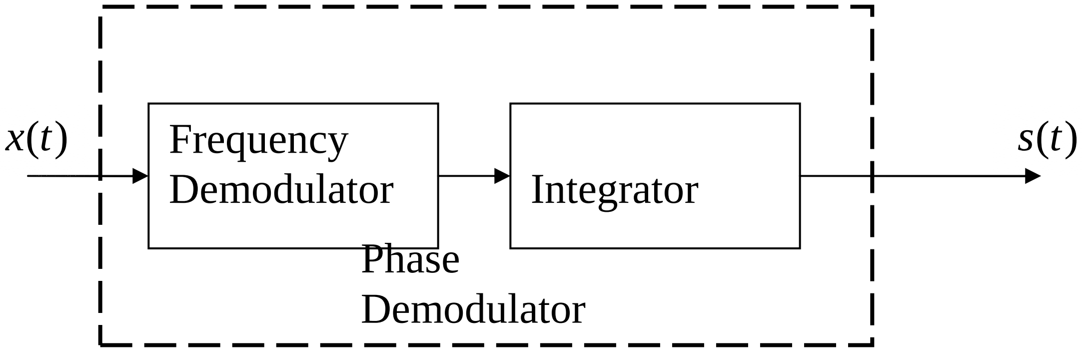
$$x(t) = A_0 \cos\left(\omega_0 t + 2\pi(\Delta f)\int^t s(\lambda)d\lambda\right)$$

If we look at the relations that exist between the phase and the frequency, we remark that the two modulations are related. In fact, we can built a frequency modulator using a phase modulator, a frequency demodulator using also a phase demodulator and vice versa.









The above four figures show how we can build one type of modulator or demodulator using the other.

Exponential modulation is a highly nonlinear modulation. This means that it is very hard to relate the spectrum of the modulated waveform with the one of the baseband as we did with the linear modulations. So, an analysis of the modulated signal in the frequency domain is quite difficult in the general case.

There are two special cases where this analysis is not very complicated: the narrowband phase and frequency modulation where the phase deviation is very small and the sinusoidal modulation where the baseband signal is sinusoidal.

Narrowband phase and frequency modulation

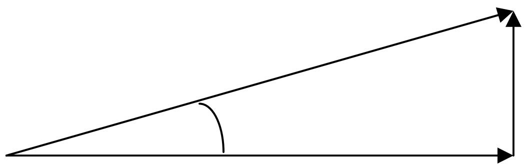
The modulated signal in this case has the general shape of:

$$x(t) = A_0 \cos\left(\omega_0 t + \phi(t)\right) \text{along with } \left|\phi(t)\right|_{\max} << 1$$

Developing the cosine, $x(t)$ becomes:

$$x(t) = A_0 \cos\phi(t)\cos\omega_0 t - A_0 \sin\phi(t)\sin\omega_0 t$$

Using the fact that $\left|\phi(t)\right|_{\max} << 1$, we have: $\cos\phi(t) \approx 1$ and $\sin\phi(t) \approx \phi(t)$, giving: $x(t) = A_0 \cos\omega_0 t - A_0\phi(t)\sin\omega_0 t$.



The above phasor diagram illustrates that the signal x(t) is the projection on the real axis of the sum of two phasors rotating at the same speed ω_0 and making an angle of 90° between them. We see that this phase shift produces the phase modulation. Furthermore, the Fourier transform of the expression of $x(t)$ can be evaluated. So, if $\Phi(f) = \mathcal{F}\left[\phi(t)\right]$, then

$$X(f) = \frac{A_0}{2}\left[\delta(f - f_0) + \delta(f + f_0)\right] - \frac{A_0}{2j}\left[\Phi(f - f_0) - \Phi(f + f_0)\right].$$

If the signal is PM, then $\phi(t) = (\Delta\phi)s(t)$ giving $\Phi(f) = (\Delta\phi)S(f)$. So, if the signal is bandlimited to W, then the PM signal will be limited to a bandwidth $B = 2W$.

If the signal is FM, then $\phi(t) = 2\pi(\Delta f)\int^{t} s(\lambda)d\lambda$ giving $\Phi(f) = \frac{(\Delta f)}{jf}S(f)$. Here also the bandwidth of the FM signal is $2W$.

Sinusoidal modulation

The other case that has a simple analytic expression is when the modulating signal is sinusoidal. When a signal is sinusoidal, its derivative is also sinusoidal. So, we can use the same analysis for both frequency and phase modulation. The modulated signal in both cases is $x(t) = A_0 \cos\left(\omega_0 t + \phi(t)\right)$. $\phi(t) = (\Delta\phi)s(t)$ for PM and $\phi(t) = 2\pi(\Delta f)\int^{t} s(\lambda)d\lambda$ for FM.

For FM modulation, we assume that $s(t) = \cos\omega_m t$. This gives:

$\phi(t) = \frac{2\pi(\Delta f)}{\omega_m}\sin\omega_m t$. $\beta = \frac{\Delta f}{f_m}$ is called the modulation index. So, $x(t) = A_0 \cos\left(\omega_0 t + \beta\sin\omega_m t\right)$.

For PM modulation, the modulating signal is $s(t) = \sin\omega_m t$. The instantaneous phase deviation becomes: $\phi(t) = (\Delta\phi)\sin\omega_m t$. In this case, the modulation index is $\beta = (\Delta\phi)$ and we obtain the same expression. So, for both cases, the expression of the modulated signal is:

$$x(t) = A_0 \cos\left(\omega_0 t + \beta\sin\omega_m t\right) = A_0 \operatorname{Re}\left[\exp\left(j\beta\sin\omega_m t\right)\exp\left(j\omega_0 t\right)\right]$$

In the above expression, the function $\exp\left(j\beta\sin\omega_m t\right)$ is periodic with a period $T_m = \frac{2\pi}{\omega_m}$. It can be developed in Fourier series. The development is:

$$\exp\left(j\beta\sin\omega_m t\right) = \sum_{n=-\infty}^{+\infty} J_n(\beta)\exp\left(jn\omega_m t\right)$$

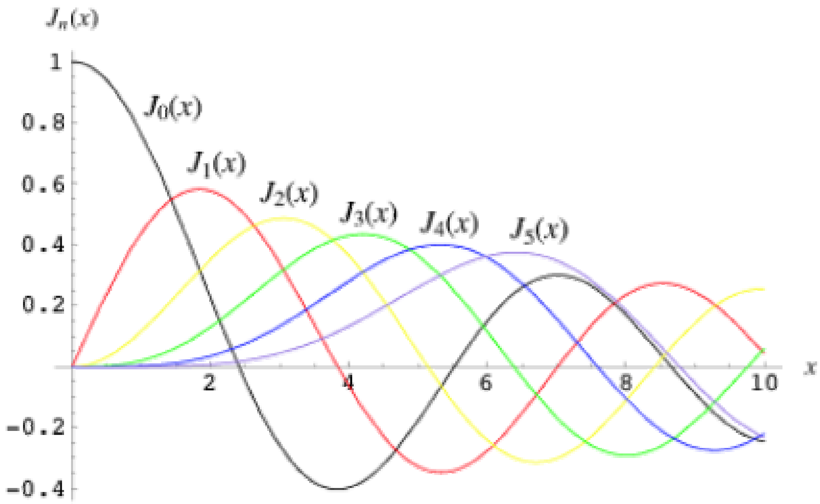
The Fourier coefficients $J_n(\beta)$ are the Bessel functions of the first kind of order *n* and argument β. These functions are tabulated and can be easily computed. They appear as solutions of differential equations. For positive order, we can use the following Mc Lauren series:

$$J_n(\beta) = \left(\frac{\beta}{2}\right)^n \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(n+k)!} \left(\frac{\beta}{2}\right)^k$$

and when *n* is negative, we use the following relation:

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

The following figure shows the behavior of the first 6 Bessel functions.



They look like damped sinewaves. From the Mc Lauren series we can deduce their properties for β around 0. For very small values of β, we have the following approximations:

$$J_0(\beta) \approx 1$$

$$J_n(\beta) \approx \frac{1}{n!}\left(\frac{\beta}{2}\right)^n \text{ for } n > 0$$

This means that the only functions that we should consider around zero are J_0 and J_1. So, for $\beta < 0.1$, $J_0(\beta) \approx 1$ and $J_1(\beta) \approx \frac{\beta}{2}$.

The following table gives the value of the first Bessel functions.

				$J_n(x)$					
x / n	0.5	1	2	3	4	6	8	10	12
0	0.9385	0.7652	0.2239	- 0.2601	-0.3971	0.1506	0.1717	-0.2459	0.0477
1	0.2423	0.4401	0.5767	0.3391	-0.0660	-0.2767	0.2346	0.0435	-0.2234
2	0.0306	0.1149	0.3528	0.4861	0.3641	- 0.2429	-0.1130	0.2546	-0.0849
3	0.0026	0.0196	0.1289	0.3091	0.4302	0.1148	-0.2911	0.0584	0.1951
4	0.0002	0.0025	0.0340	0.1320	0.2811	0.3576	-0.1054	-0.2196	0.1825
5	—	0.0002	0.0070	0.0430	0.1321	0.3621	0.1858	-0.2341	-0.0735
6		—	0.0012	0.0114	0.0491	0.2458	0.3376	-0.0145	-0.2437
7			0.0002	0.0025	0.0152	0.1296	0.3206	0.2167	-0.1703
8			—	0.0005	0.0040	0.0565	0.2235	0.3179	0.0451
9				0.0001	0.0009	0.0212	0.1263	0.2919	0.2304
10				—	0.0002	0.0070	0.0608	0.2075	0.3005
11					—	0.0020	0.0256	0.1231	0.2704
12						0.0005	0.0096	0.0634	0.1953
13						0.0001	0.0033	0.0290	0.1201
14						—	0.0010	0.0120	0.0650

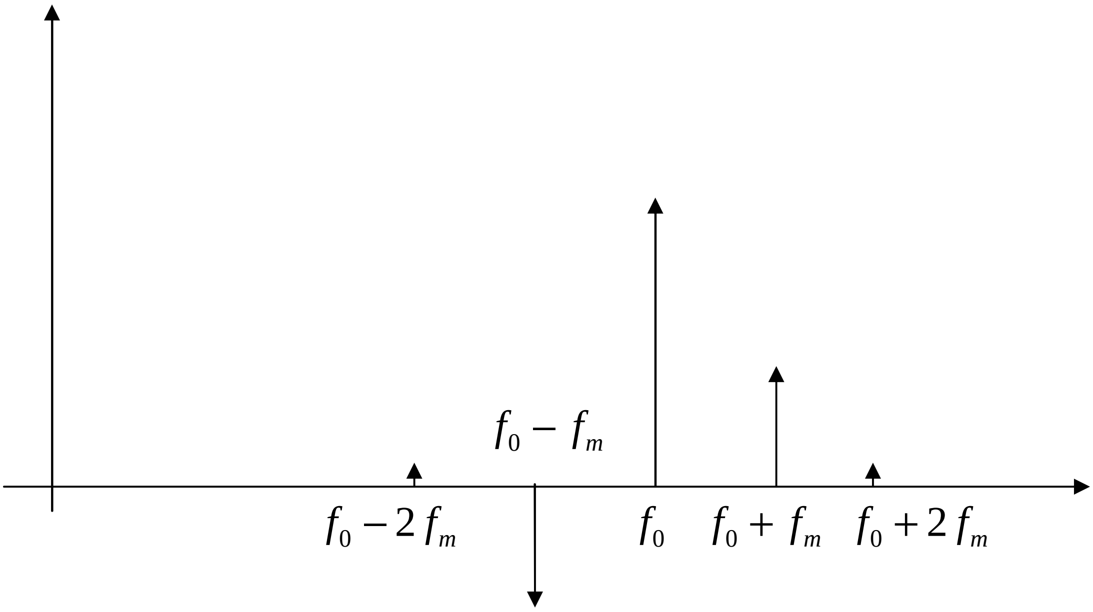
The exponentially modulated signal is:

$$x(t) = A_0 \sum_{n=-\infty}^{+\infty} J_n(\beta)\cos\left[(\omega_0 + n\omega_m)t\right]$$

We see that the signal contains a large number of components around the carrier frequency. For example, let us assume $\beta = 1$. Then only J_0, J_1 and J_2 are significantly different from zero. We can then write:

$$x(t) \approx A_0 \sum_{n=-2}^{2} J_n(1)\cos[(\omega_0 + n\omega_m)t]$$

The modulated signal is the sum of five sinewaves. The spectrum is displayed below. The values of the different amplitudes are read from the above table.



Single sided spectrum of the signal

The above spectrum is approximately bandlimited. For the exponentially modulated signal, we can use as transmission bandwidth the band of frequency that contains most of the power of the signal. The total power of the transmitted waveform can be computed as follows. From the series development of $x(t)$, we obtain:

$$P = < x^2(t) > = < A_0^2 \left(\sum_{n=-\infty}^{+\infty} J_n(\beta) \cos\left[(\omega_0 + n\omega_m) t \right] \right)^2 >$$

If the different sinewave are independent, the total power is the sum of the individual powers. We obtain:

$$P = \frac{A_0^2}{2} \sum_{n=-\infty}^{+\infty} J_n^2(\beta) = \frac{A_0^2}{2}$$

We have used the following property of the Bessel functions:

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

If we keep n components on each side of the carrier, we obtain:

$$P(n,\beta) = \frac{A_0^2}{2} \sum_{k=-n}^{n} J_k^2(\beta)$$

The ratio of this power to the total power is:

$$\frac{P(n,\beta)}{P} = \sum_{k=-n}^{n} J_k^2(\beta)$$

This ratio is very close to 1 (≈ 0.95) for $n = \lfloor \beta + 1 \rfloor$. So, the band of frequency components containing the sidebands from $f_0 - nf_m$ to $f_0 + nf_m$ can be used as an approximation for the bandwidth of the exponentially modulated waveform for sinusoidal modulating waveform. So the transmission bandwidth is approximately:

$$B = 2nf_m = 2(\beta + 1)\, f_m$$

For FM signal, $\beta = \dfrac{\Delta f}{f_m}$, we obtain:

$$B = 2(\Delta f) + 2\, f_m$$

The above rule is called the "*Carson's Rule*". It has been found empirically that this rule can be applied even if the modulating signal is not sinusoidal. In this case, we can define the "*Deviation Ratio*";

$\delta = \dfrac{\Delta f}{W}$ where W is the bandwidth of the baseband modulating signal.

Carson's rule generalizes to $B = 2(\delta + 1)\, f_m = 2(\Delta f) + 2W$. For example, broadcast FM uses a value of Δf = 75 kHz for transmitting a baseband signal having a bandwidth W = 15 kHz. This gives a transmission bandwidth B = 2×75 + 2×15 = 180 kHz. The normalized bandwidth is 200 kHz. Carson's rule underestimates it, but the error is small.

A general remark about FM and PM is that the bandwidth of the resulting signal is in general much larger than $2W$. This is due to the fact that these modulations are highly non-linear. Another remark about FM is that it is very resistant to perturbations induced by noise and interference. So, we can say that FM protects the information of the signal at the expense of a bandwidth increase.

When β is very small, we can use $J_0(\beta) \approx 1$ and $J_1(\beta) \approx \frac{\beta}{2}$ to express the modulated signal:

$$x(t) = A_0 \cos \omega_0 t + A_0 \frac{\beta}{2} \cos\left[(\omega_0 + \omega_m)t\right] - A_0 \frac{\beta}{2} \cos\left[(\omega_0 - \omega_m)t)\right]$$

giving $x(t) = A_0 \left[\cos \omega_0 t - \beta \sin \omega_m t \sin \omega_0 t\right] = A_0 \left[\cos \omega_0 t - \beta s(t) \sin \omega_0 t\right]$ which is the narrow band approximation.

Filtering the FM signal

Being a non linear modulation, the usual method of filtering the complex envelop of the FM signal by the equivalent lowpass filter does not work for general filter shapes. In some specific cases, this technique can be used. In order to use it, this FM signal must be bandpass. In this case, the complex envelop is easily extracted.

$x(t) = A_0 \cos\left(\omega_0 t + \phi(t)\right)$ giving a complex envelop $m_x(t) = A_0 \exp\left(j\phi(t)\right)$. One case where this technique can be used is the case of a filter with an amplitude response of the type:

$H(f) = M(f)\exp\left(j\theta(f)\right)$ where $M(f) = M_0 + k(f - f_0)$ for $f > 0$ and $\theta(f) = -2\pi\tau_g(f - f_0) + \theta_0$. The equivalent lowpass filter is:

$H_{lp}(f) = [M_0 + kf]\exp\left(-j2\pi\tau_g f\right)\exp(j\theta_0)$. In the frequency domain, the complex envelop of the output is:

$M_y(f) = M_x(f)H_{lp}(f) = [M_0 + kf]\exp\left(-j2\pi\tau_g f\right)\exp(j\theta_0)M_x(f)$

So:

$$\begin{aligned} M_y(f) &= M_0\exp\left(j\theta\right)M_x(f)\exp\left(-j2\pi\tau_g f\right) \\ &\quad + kf\exp\left(j\theta\right)M_x(f)\exp\left(-j2\pi\tau_g f\right) \end{aligned}$$

Given that $\mathscr{F}\left[\frac{dx(t)}{dt}\right] = 2\pi jfX(f)$, the multiplication by f in the frequency domain becomes a differentiation in the time domain. So,

$$m_y(t) = \left[M_0 m_x(t-\tau_g) + \frac{k}{2\pi j}\frac{dm_x(t-\tau_g)}{dt}\right]\exp\left(j\theta\right)$$

From the expression of the input complex envelop, we obtain:

$\frac{dm_x(t-\tau_g)}{dt} = jA_0\frac{d\phi(t-\tau_g)}{dt}\exp\left(j\phi(t)\right)$ and finally:

$$m_y(t) = A_0\left[M_0 + \frac{k}{2\pi}\frac{d\phi(t-\tau_g)}{dt}\right]\exp\left(j\phi(t-\tau_g)\right)\exp\left(j\theta\right)$$

The output signal is then:

$$y(t) = A_0\left[M_0 + \frac{k}{2\pi}\frac{d\phi(t-\tau_g)}{dt}\right]\cos\left(\omega_0 t + \phi(t-\tau_g) + \theta\right)$$

If the signal is FM, $\frac{d\phi(t)}{dt} = 2\pi(\Delta f)s(t)$. In this particular case:

$$y(t) = A_0\left[M_0 + k(\Delta f)s(t-\tau_g)\right]\cos\left(\omega_0 t + 2\pi(\Delta f)\int^{t-\tau_g} s(\lambda)d\lambda + \theta\right)$$

We remark that the output signal has two different modulations: FM and AM. The information signal is contained in the envelop of the output signal. So, an envelop detector will demodulate the signal.

If the filter has a different transfer function, we can use the concept of "*quasi-static*" approximation. If the carrier frequency is much higher than the baseband modulating frequency, then we can safely assume that the frequency is constant over a quite long time. The FM signal will behave almost like a constant frequency sinewave. We know that if the input of a filter with transfer function $H(f)$ is a pure sinewave $A_0\cos\left(2\pi f_0 t\right)$, the output is also a sinewave at the same frequency: $y(t) = A_0\left|H(f_0)\right|\cos\left[2\pi f_0 t + \text{Arg}\left[H(f_0)\right]\right]$. In the quasi- static approximation, we replace f_0 by the instantaneous frequency $f(t)$.

So, given $f(t) = f_0 + (\Delta f)s(t)$, we obtain:

$$y(t) = A_0\left|H(f_0 + (\Delta f)s(t)\right|\cos\left[2\pi f_0 t + 2\pi(\Delta f)\int^{t} s(\lambda)d\lambda + \text{Arg}\left[H\left(f_0 + (\Delta f)s(t)\right)\right]\right]$$

Example:

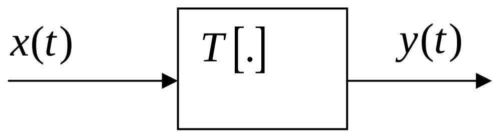
Using $H(f) = M(f)\exp\left(j\theta(f)\right)$ with $M(f) = M_0 + k(f - f_0)$ for $f > 0$ along with $\theta(f) = -2\pi\tau_g(f - f_0) + \theta_0$, we obtain:

$\left|H(f(t))\right| = M_0 + k(\Delta f)s(t)$ and $\text{Arg}\left[H(f(t))\right] = -2\pi\tau_g(\Delta f)s(t) + \theta_0$.

We see that we obtain the same amplitude variation as in the previous case (except for a time delay).

FM through nonlinear system

Consider the following memoriless nonlinear system.



If the input is $x(t) = A_0 \cos\theta(t)$, the output will be $y(t) = T\left[A_0 \cos\theta(t)\right]$. The output signal $T\left[A_0 \cos\theta(t)\right]$ is not periodic as a function of the variable *t*, it is periodic, with a period 2π if we consider it function of θ. We can thus develop the output *y* in Fourier series.

$$y(t) = \frac{a_0}{2} + \sum_{n=1}^{+\infty} a_n \cos n\theta(t)$$

The Fourier series is a cosine series because the input is $A_0 \cos\theta$. It is even. So, $T\left[A_0 \cos\theta\right]$ is also even. The coefficients are given by:

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} T\left[\cos\theta\right]\cos n\theta d\theta$$

Replacing the argument θ by its value leads to:

$$y(t) = \frac{a_0}{2} + \sum_{n=1}^{+\infty} a_n \cos\left[n\omega_0 t + n\phi(t)\right]$$

The above signal is a superposition of an infinite number of exponentially modulated waveforms. For the FM case, we can write:

$$y(t) = \frac{a_0}{2} + a_1 \cos\left(\omega_0 t + 2\pi(\Delta f)\int^t s(\lambda)d\lambda\right) + a_2 \cos\left(2\omega_0 t + 2\pi(2\Delta f)\int^t s(\lambda)d\lambda\right) + \cdots$$

We can remark that the output is a sum of FM signals at carrier frequencies nf_0 each one with a maximum deviation $n\Delta f$. If the different spectra do not intersect, we can select one of them using a bandpass filter tuned at nf_0 and obtain at the output:

$$z(t) = a_n \cos\left(n\omega_0 t + 2\pi(n\Delta f)\int^t s(\lambda)d\lambda\right)$$

FM Generation

Direct Method

The FM signal can be generated directly using a Voltage Controlled Oscillator (VCO) like the one used in the lab generators (GW-Instek GFG8255A). The output signal is a sinewave with an instantaneous frequency given by $f = f_v + k_m v_{in}$. The frequency f_v is called the free running frequency and the constant k_m is called the VCO gain (It is measured in Hz/Volts). In general, these VCOs can use a variable reactance in a parallel RLC circuit used to tune an oscillator. We can use "*varactors*" for example. The output frequency of this type of oscillator is the resonant frequency of the RLC circuit.

$f = \dfrac{1}{\sqrt{LC}}$ where $C = C_0 - cv_{in}$. The input voltage is proportional to $s(t)$. $v_{in} = |V_{in}|_{\max} s(t)$, so: $f(t) = \dfrac{1}{\sqrt{LC}} = \dfrac{1}{\sqrt{LC_0}}\left[1 - \dfrac{c}{C_0}v_{in}(t)\right]^{-\frac{1}{2}}$.

If $\left|\dfrac{c}{C_0}v_{in}(t)\right|_{\max} << 1$, we can use the approximation $[1-\varepsilon]^{-\frac{1}{2}} \approx 1 + \dfrac{1}{2}\varepsilon$.

The instantaneous frequency is given by: $f \approx f_v\left[1 + \dfrac{c}{2C_0}v_{in}\right]$. The free running frequency is $f_v = \dfrac{1}{\sqrt{LC_0}}$ and the VCO gain is $k_m = \dfrac{cf_v}{2C_0}$. The maximum frequency deviation is $\Delta f = |V_{in}|_{\max} k_m$. There exist a large number of VCO circuits. The most common ones (the ones that are found in integrated circuits and in signal generators) generate triangular waves using integrators or capacitors charged by controlled current sources. A good reference is "K. K. Clarke & D. T. Hess, *Communication Circuits: Analysis and Design*, 2nd ed. Krieger Pub Co, 1994" which is the textbook for the communication circuit course.

(The part between the two ♦ is for reading only)

Frequency mixing

In this part, we introduce an important technique used in receiver and transmitter design: Frequency mixing. The mixer is a device capable of changing a carrier frequency for any type of modulation. It is based on the frequency translation theorem of Fourier theory. We start first with real signal mixing.

Consider a general bandpass signal $x_{rf}(t) = r(t)\cos\left[\omega_{rf} t + \phi(t)\right]$. If we multiply this signal by a sinewave $x_{lo}(t) = B\cos\omega_{lo} t$, the result is:

$$z(t) = x_{rf}(t) \times x_{lo}(t) = Br(t)\cos\left[\omega_{rf} t + \phi(t)\right]\cos\omega_{lo} t$$

Using trigonometric identities, we see that this signal is the sum of two bandpass signals:

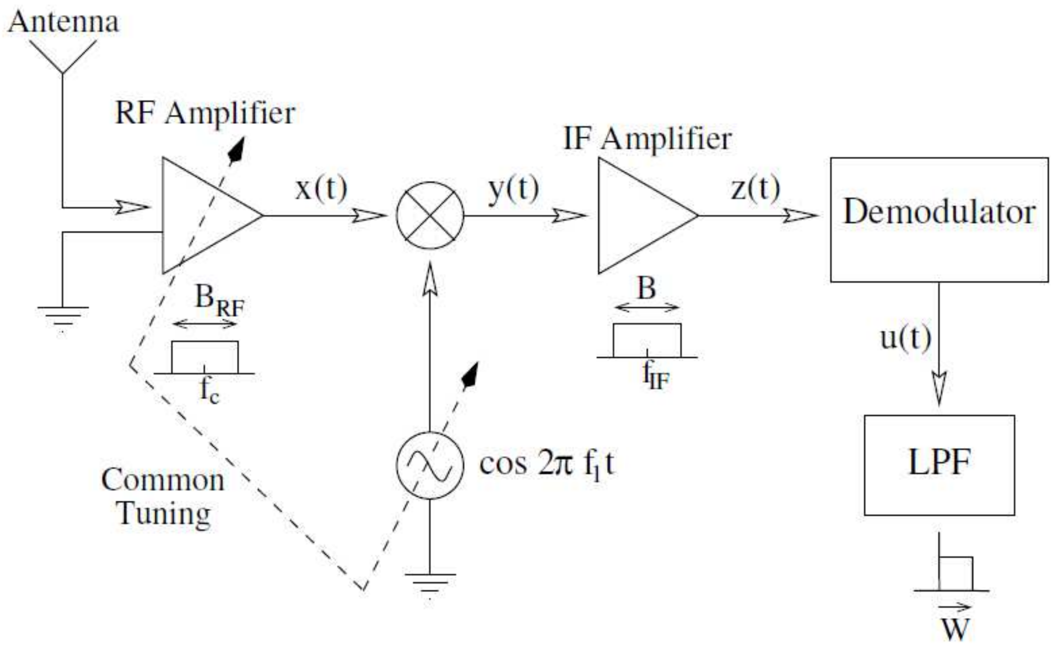
$$z(t) = \frac{B}{2} r(t)\cos\left[\left(\omega_{rf} + \omega_{lo}\right)t + \phi(t)\right] + \frac{B}{2} r(t)\cos\left[\left(\omega_{rf} - \omega_{lo}\right)t + \phi(t)\right].$$

Using a bandpass filter tuned at either the sum or the difference frequency, we obtain a bandpass signal having the same complex envelop (i.e. the same information) but a different carrier frequency. The new frequency is usually called "*intermediate frequency*" f_{if}.

If $f_{if} = f_{rf} + f_{lo}$, we say that we are doing "*up mixing*". On the other hand, if $f_{if} = f_{rf} - f_{lo}$ or $f_{if} = f_{lo} - f_{rf}$, we say that we are performing down mixing.

♦The "*mixer*" is an important electronic subsystem in any communication receiver or transmitter. It is the basic building block of the "*superheterodyne*" receiver. This concept of receiver was introduced in order to solve the very complex problem of amplifying and selecting one radio station among a large number of stations transmitting at different frequencies.

The first solution that comes to mind is to use a "*tunable*" bandpass filter. However, the construction of a very selective tunable bandpass filter is very complex. Furthermore, due to component aging, such system is prone to random changes and mistuning after a certain time. It is much easier to build a fixed frequency very selective filter. So, instead of translating the center frequency of a tunable filter before the different signals, it is much easier to translate the frequency of the signals before the center frequency of a fixed bandpass filter. This is the concept of the super heterodyne receiver. The superheterodyne receiver is composed of a tunable local oscillator ganged with a wide band tunable RF amplifier, a mixer and a fixed frequency IF amplifier. It is built using the block diagram shown below.



Let us assume we are using down mixing. The rf amplifier pre-selects a band of frequencies containing a small number of stations around the station at frequency f_c. The bandwidth B_{RF} is large compared to the bandwidth B required by the modulation used (FM, AM, any linear one) but smaller than $2f_{IF}$, the intermediate frequency. Using down mixing, we must have:

$f_{IF} = f_c - f_1$ giving $f_c = f_1 + f_{IF}$ or $f_{IF} = f_1 - f_c$ giving $f_c = f_1 - f_{IF}$.

From the above two relations, we see that if the rf filter does not exist, then we can receive two different stations if we simply use the local oscillator for tuning. These two stations are separated by $2f_{IF}$. These two frequencies are called "*image frequencies*". The job of the tunable rf amplifier is to eliminate one of them so that it will not interfere with the station that we want to receive. Intermediate frequency for broadcast receivers has been standardized to the values of 455 kHz for AM and 10.7 MHz for FM.

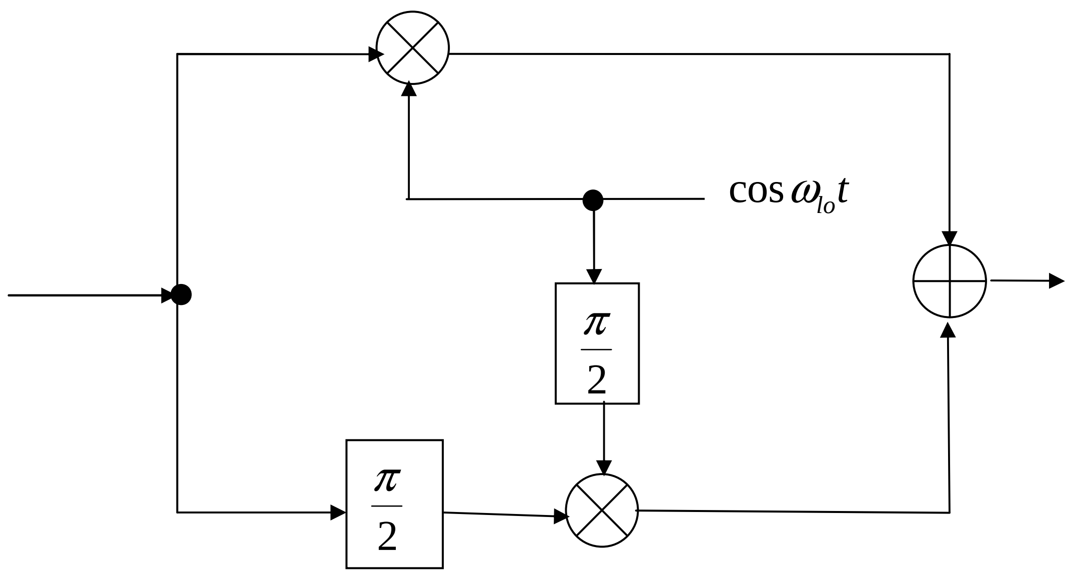
Another technique is used to avoid this problem of image frequencies. It is based on frequency translation using complex phasors. Consider the high frequency signal: $x_{rf}(t) = r(t)\cos\left[\omega_{rf} t + \phi(t)\right]$. The associated analytic signal is: $x_{rf+}(t) = x_{rf}(t) + j\hat{x}_{rf}(t) = r(t)\exp\left(j\phi(t)\right)\exp\left(j\omega_{rf} t\right)$. If we multiply this signal by the phasor: $\exp\left(-j\omega_{lo} t\right)$, we obtain the signal: $z(t) = r(t)\exp\left(j\phi(t)\right)\exp\left(j\left(\omega_{rf} - \omega_{lo}\right)t\right)$. The real part is:

$x_{if}(t) = \text{Re}\left[z(t)\right] = r(t)\cos\left[\left(\omega_{rf} - \omega_{lo}\right)t + \phi(t)\right]$. This is the correct translated signal. So, the process of performing the above operation is:

$x_{if}(t) = \text{Re}\left[\left(x_{rf}(t) + j\hat{x}(t)\right)\left(\cos\omega_{lo} t - j\sin\omega_{lo} t\right)\right]$ giving:

$$x_{if}(t) = x_{rf}(t)\cos\omega_{lo} t + \hat{x}_{rf}(t)\sin\omega_{lo} t$$

This leads to the following block diagram:



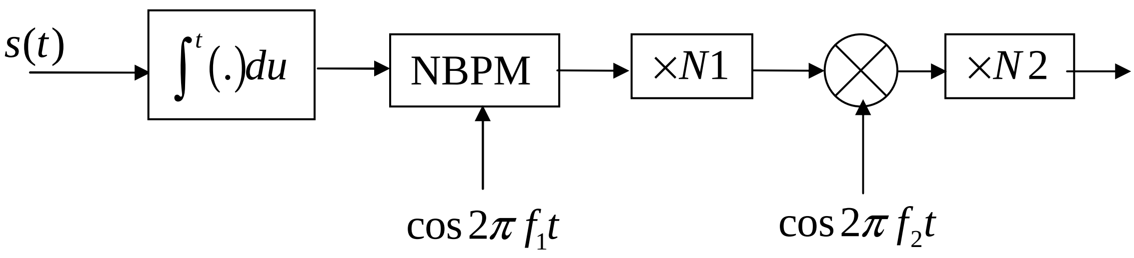
Imageless Mixer

The above circuit can be used without any image rejection filter before.

♦

Indirect FM generation

This technique of FM generation is the one that is commonly used in FM transmitters. This is due to the fact that the carrier frequency and the maximum frequency deviation can be set with high precision. It is based on a Narrow Band Frequency modulator cascaded with nonlinear amplifiers that are used as frequency multipliers. Mixers are also used to translate the carrier because frequency multiplication leads usual to impractically high carrier frequencies. A general block diagram of such system is:



The frequency multipliers are implemented using a nonlinear amplifier (Class C) followed by a narrow bandpass filter tuned at the proper harmonic.

If we consider the above block diagram, the carrier frequency is given by $f_0 = N2\left(N1\times f_1 - f_2\right)$ or $f_0 = N2\left(f_2 - N1\times f_1\right)$. If the frequency deviation at the output of the NBFM is Δf_1, the final deviation is $\Delta f = N1\times N2\times\Delta f_1$. In general the frequency multiplication is achieved by a cascade of frequency doublers and triplers. It is impossible to achieve an efficient amplifier if the multiplication factor is larger than three.

FM Demodulation

FM demodulation by differentiation (FM to AM conversion):

If we compute the derivative of an FM signal, we obtain:

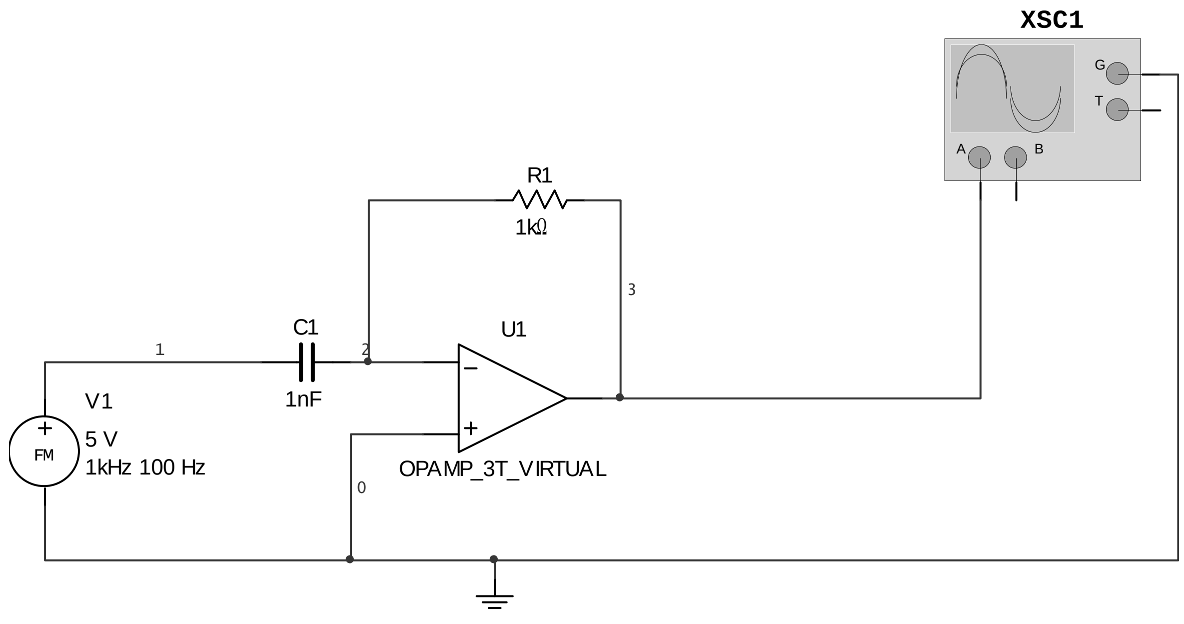
$$\frac{d}{dt}\left[A_0\cos\left(\omega_0 t+\phi(t)\right)\right]=-A_0\left(\omega_0+\frac{d\phi}{dt}\right)\sin\left(\omega_0 t+\phi(t)\right)$$

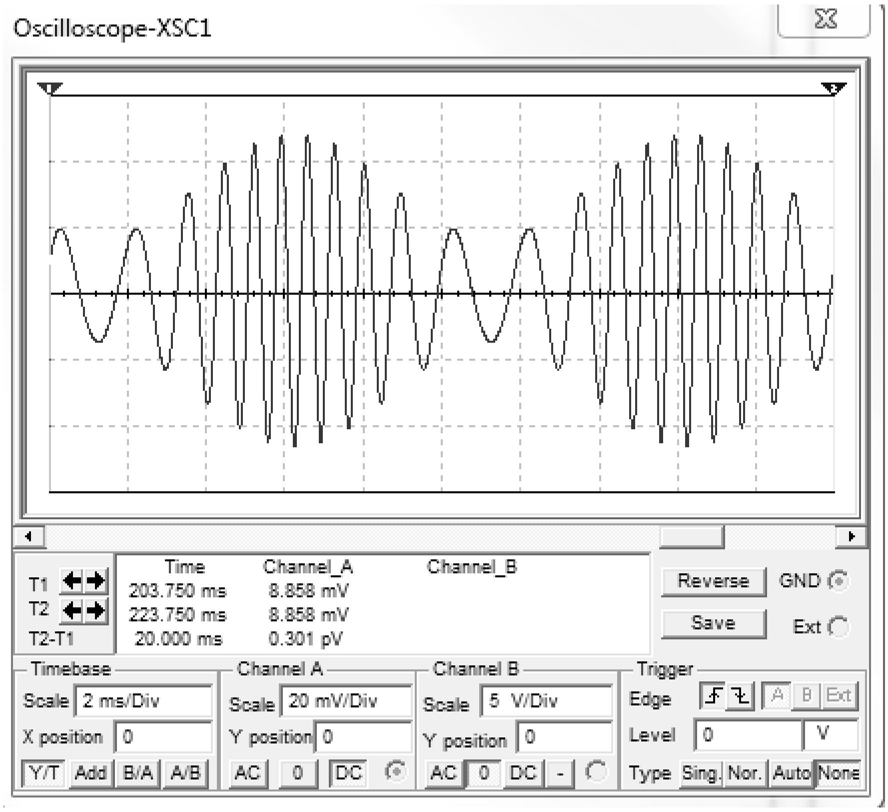
Since the signal is an FM one, $\frac{d\phi}{dt}=2\pi(\Delta f)s(t)$, we get:

$$\frac{d}{dt}\left[A_0\cos\left(\omega_0 t+\phi(t)\right)\right]=-A_0\left(\omega_0+2\pi(\Delta f)s(t)\right)\sin\left(\omega_0 t+2\pi(\Delta f)\int^t s(\lambda)d\lambda\right)$$

We can see that the output of a differentiation circuit will produce a modulation of the amplitude. This modulation can be detected using any AM demodulator.

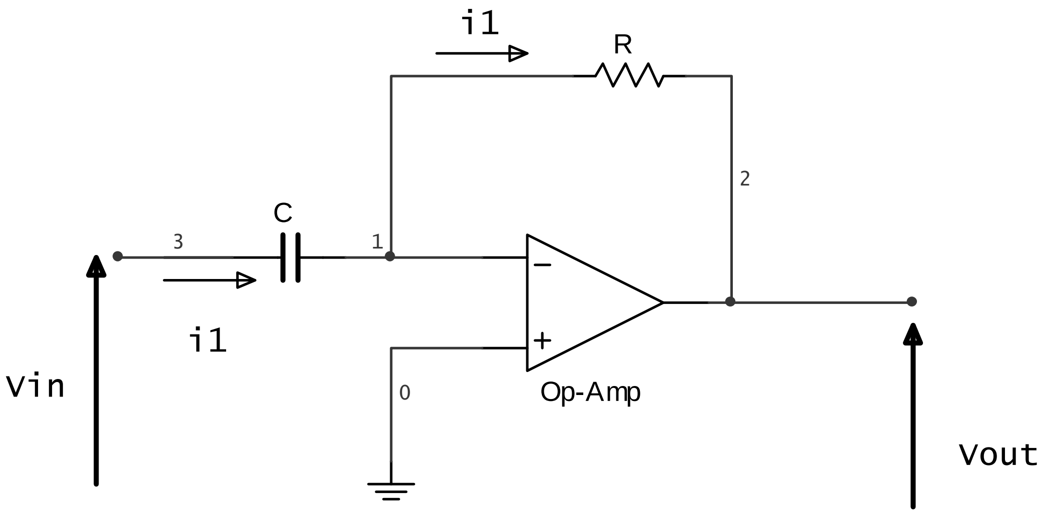
The following circuit is a differentiator built using an OP-Amp simulated using Multisim.





The above picture displays the output signal. We clearly see the two modulations: AM and FM. The signal produced by the FM source is a sinewave modulated FM signal with β = 5, carrier frequency = 1 kHz and modulating frequency = 100 Hz.

♦Differentiator using an op-amp



The Operational Amplifier is an amplifier that possesses a very high gain, differential input and very high input impedance. If the voltage at the inverting input is v_-, the input at the non-inverting input is v_+ and the gain of the op-amp is *G*, its output is given by $G\left(v_+ - v_-\right)$. *G* is assumed very large. This means that a very small difference will produce a measurable output. We can safely assume that this difference is zero. At that time, the inverting input is practically at ground in the above schematic. This implies that the capacitor is in parallel with Vin. So, $i1 = C\frac{dVin}{dt}$. The input impedance of the op-amp is very large. The same current will flow though the resistance R. So, $Vout = -Ri1$. After the elimination of i1, we obtain:

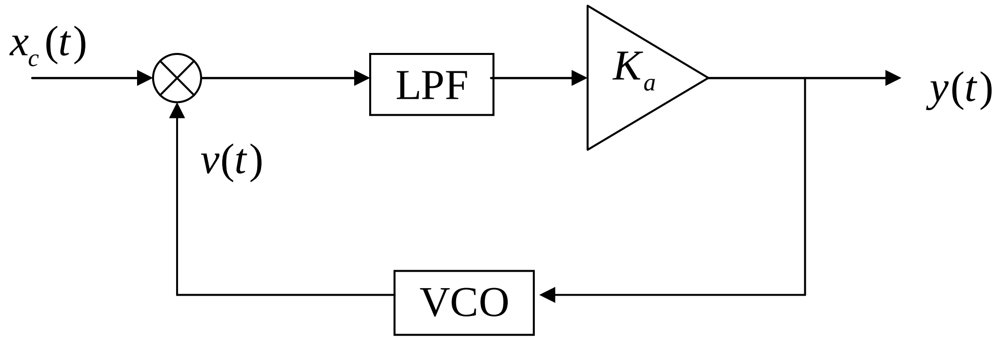
$$Vout = -RC\frac{dVin}{dt}$$

♦

There exist a large number of other FM demodulators. The interested student should consult the previous reference (Clarke & Hess).

The Phase Locked Loop

The phase locked loop (PLL) is a feedback system composed principally of a voltage controlled oscillator (VCO), a phase detector (PD) and a lowpass filter (LPF). The phase detector is usually modeled as a multiplier.



The input signal is $x_c(t) = A_c \cos\theta_c(t)$ and the output of the VCO is $v(t) = A_v \cos\theta_v(t)$. The output of the multiplier is $z(t) = \frac{A_c A_v}{2}\left[\cos\left(\theta_c(t) - \theta_v(t)\right) + \cos\left(\theta_c(t) + \theta_v(t)\right)\right]$. The lowpass filter eliminates the sum term and filters the difference term. So, we can consider that the output of the lowpass filter (after amplification) is

$$y(t) = K_a h(t) * \left\{\frac{A_c A_v}{2}\left[\cos\left(\theta_c(t) - \theta_v(t)\right)\right]\right\}.$$

Let us introduce a variable ε such that: $\theta_v(t) = \theta_c(t) - \varepsilon(t) + \frac{\pi}{2}$. The output of the lowpass filter becomes:

$$y(t) = K_a h(t) * \left\{\frac{A_c A_v}{2}\left[\sin\varepsilon(t)\right]\right\}$$

We have a one to one relationship between $y(t)$ and $\varepsilon(t)$ if $-\frac{\pi}{2} \le \varepsilon(t) \le \frac{\pi}{2}$. The relation becomes linear if $\varepsilon << 1$. To simplify the analysis of the PLL and eliminate the effect of the amplitudes, let us make $A_c = 2$ and $A_v = 1$. The input signal has a carrier frequency f_0. This makes $\theta_c(t) = \omega_0 t + \phi(t)$. The VCO free running frequency f_v is shifted from f_0 by an offset f_Δ: $f_v = f_0 - f_\Delta$. The instantaneous phase of the VCO is: $\theta_v(t) = 2\pi f_v t + \phi_v(t) + \frac{\pi}{2}$. The $\frac{\pi}{2}$ constant is added in order to introduce the variable $\varepsilon(t)$ in the following expressions. The instantaneous phase deviation $\phi_v(t)$ is produced by the output signal

$y(t)$: $\dot{\phi}_v(t) = 2\pi K_v y(t)$[4]. The constant K_v is the VCO gain expressed in Hz/Volts. Using the definition of $\varepsilon(t)$, we can write:

$$\varepsilon(t) = \theta_c(t) - \theta_v(t) + \frac{\pi}{2}$$

Replacing each phase, we obtain:

$$\varepsilon(t) = 2\pi f_\Delta t + \phi(t) - \phi_v(t)$$

After differentiation:

$$\dot{\varepsilon}(t) = 2\pi f_\Delta + \dot{\phi}(t) - 2\pi K_v y(t)$$

However, $y(t)$ is the output of the lowpass filter amplified by K_a. So:

$$y(t) = K_a \int_{-\infty}^{+\infty} h(\tau) \sin\left(\varepsilon(t-\tau)\right) d\tau$$

The PLL is thus governed by the following integro-differential equation:

$$\dot{\varepsilon}(t) + 2\pi K \int_{-\infty}^{+\infty} h(\tau) \sin\left(\varepsilon(t-\tau)\right) d\tau = \dot{\phi}(t)$$

K = $K_a K_v$ is called the "*Loop Gain*". The above equation is a nonlinear equation that is quite complex to solve. In our course, we are going to consider two different cases: The first order loop (with no filtering) to analyze in a simple manner the "locking mechanism" and the linear approximation when $\sin\varepsilon \approx \varepsilon$.

Frequency Acquisition

Consider a filter with a transfer function $H(f) = 1$ over the frequency band of interest. The impulse response will be a Dirac impulse: $h(t) = \delta(t)$ and we will have:

[4] A dot above a function means that the function is differentiated.

$$\int_{-\infty}^{+\infty} h(\tau)\sin\left(\varepsilon(t-\tau)\right)d\tau = \sin\left(\varepsilon(t)\right)$$

We obtain the following first order differential equation:

$$\dot{\varepsilon} + 2\pi K \sin\varepsilon = \dot{\phi}(t) + 2\pi f_{\Delta}$$

Let us assume that we apply the signal $x_c(t) = 2\cos\left(\omega_0 t + \phi_0\right)$ at the time $t = 0$. In this case, $\phi(t) = \phi_0$ and its derivative is zero. We also assume that the free running frequency of the VCO is different from the carrier we want to lock on ($f_{\Delta} \neq 0$). The differential equation becomes $\dot{\varepsilon} + 2\pi K \sin\varepsilon = 2\pi f_{\Delta}$ for $t \geq 0$. This equation can be written as:

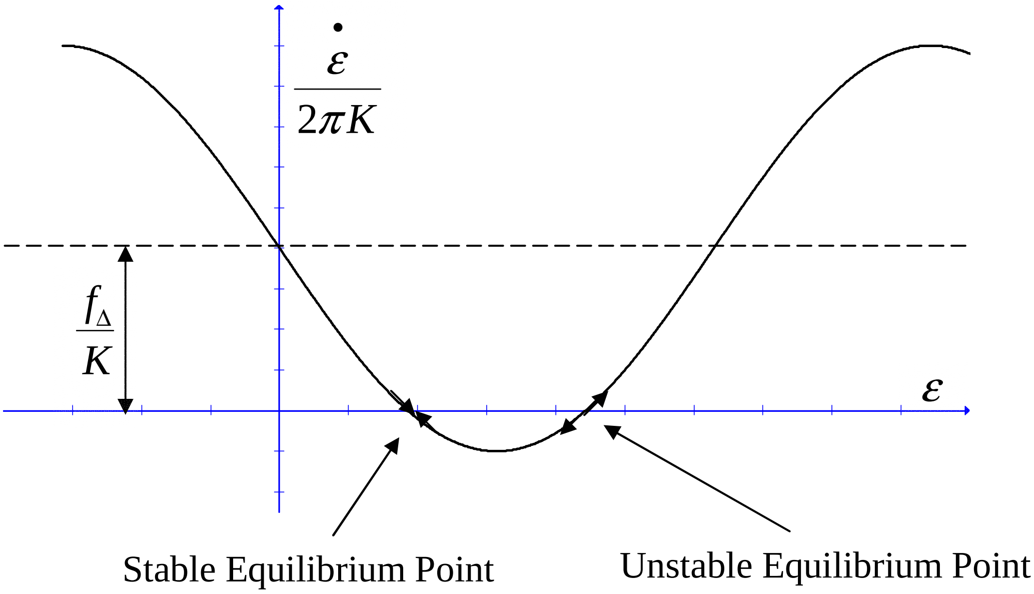
$$\frac{\dot{\varepsilon}}{2\pi K} = -\sin\varepsilon + \frac{f_{\Delta}}{K}$$

This equation relates the derivative of ε with ε. The set of points $\left(\dot{\varepsilon}, \varepsilon\right)$ is called the phase plane. In our case, it is better to use $\left(\frac{\dot{\varepsilon}}{2\pi K}, \varepsilon\right)$. This describes a single curve with t as a parameter. As t varies, the point is going to move along the curve shown below. The points on the curve where $\dot{\varepsilon} = 0$ are called "*equilibrium points*". We distinguish two different types of equilibrium points: A point is called stable if after a small perturbation around the point, the trajectory in the phase plane will go back to the equilibrium point. The point is called unstable if after a small perturbation, the trajectory will go away from the equilibrium point.

The following mechanical analogy will show the difference between the points.

Stable Equilibrium Unstable Equilibrium

If a trajectory starts at an equilibrium point, it will remain there.



Phase Plan Plot

The above curve shows the locus of the points $\left(\frac{\dot{\varepsilon}}{2\pi K}, \varepsilon\right)$ as t goes from zero to infinity. The trajectory starts at an initial point ε_0 and it will move. It is easy to see that for any starting point, the trajectory will move toward a stable equilibrium point. However, we can have equilibrium points if and only if the curve intersects the ε axis. This is possible if $\left|\frac{f_\Delta}{K}\right| \leq 1$. So, if this condition is satisfied, then $\lim_{t\to\infty} \dot{\varepsilon}(t) = 0$ and the steady state value of ε will be:

$$\varepsilon_{st} = \sin^{-1}\frac{f_\Delta}{K}$$

The steady state value of *y*(*t*) will be:

$$y_{st} = K_a \sin\varepsilon_{st} = \frac{f_\Delta}{K_v}$$

and since $\theta_v(t) = \theta_c(t) - \varepsilon(t) + \frac{\pi}{2}$, the steady state output of the VCO will be:

$$v_{st}(t) = \cos\left(\omega_0 t + \phi_0 - \varepsilon_{st} + \frac{\pi}{2}\right)$$

If the loop gain is very large, then ε_{st} will be very small. The VCO output will have exactly the same frequency as the input signal with a phase shift that is practically $\frac{\pi}{2}$.

If $\left|\frac{f_\Delta}{K}\right| > 1$, it is impossible to have a lock. The trajectory will not intersect the ε axis and there is no equilibrium point and no steady state solution. The VCO frequency will keep changing.

In order to have a lock, the VCO free running frequency must be in the following range: $[f_0 - K, f_0 + K]$. This range is called the "*lock range*". The value found is valid for a first order loop only. It will be different for another filter.

FM Demodulation

In order to analyze the PLL when the input is frequency modulated, we assume that the VCO free running frequency is the same as the input carrier frequency. This means that $f_\Delta = 0$. We also assume that ε remains small all the time. We can replace sinε by ε in our analysis. With these hypotheses, we have:

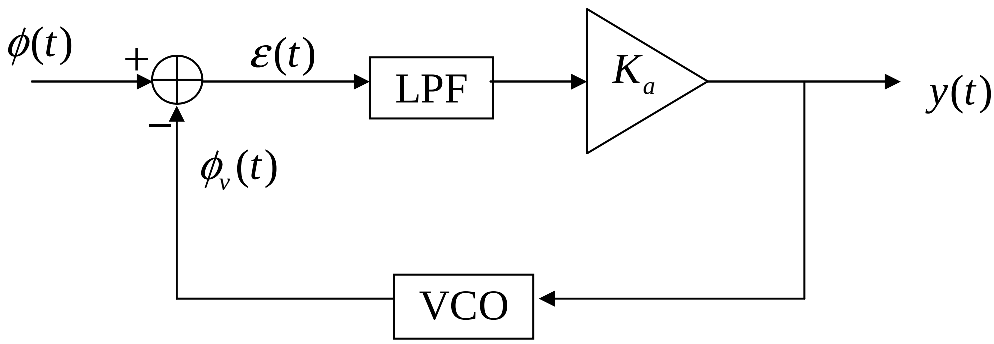
$$x_c(t) = 2\cos\left(\omega_0 t + \phi(t)\right)$$

$$v(t) = \cos\left(\omega_0 t + \phi_v(t) + \frac{\pi}{2}\right)$$

and $y(t) = K_a \int_{-\infty}^{+\infty} h(\tau)\sin\left(\varepsilon(t-\tau)\right)d\tau \simeq K_a \int_{-\infty}^{+\infty} h(\tau)\varepsilon(t-\tau)d\tau$

$$\varepsilon(t) = \phi(t) - \phi_v(t)$$

Since y is linearly related to ε and ϕ_v is also linearly related to y, it is more interesting to use the phase deviations as primary variables instead of using x_c and v. the different relationships are better described by the following block diagram.



The above block diagram is completely linear. The VCO transfer function is given by the following relation:

$$\phi_v(t) = 2\pi K_v \int^t y(u)du$$

This gives the following relation in the frequency domain:

$$\Phi_v(f) = \frac{K_v}{jf} Y(f)$$

The closed loop transfer function (in the frequency domain) is:

$$Y(f) = \frac{1}{K_v} \frac{jfH(f)}{H(f) + j\frac{f}{K}} \Phi(f)$$

where $H(f) = \mathcal{F}\left[h(t)\right]$.

If the input is an FM wave, $\phi(t) = 2\pi\Delta f \int^t s(u)du$, then

$$\Phi(f) = \frac{\Delta f}{jf} S(f)$$

The transfer function becomes:

$$Y(f) = \frac{\Delta f}{K_v} \frac{H(f)}{H(f) + j\frac{f}{K}} S(f)$$

We see that the output signal is the baseband information signal filtered. If the loop gain is very high, the output signal will be proportional to $s(t)$.

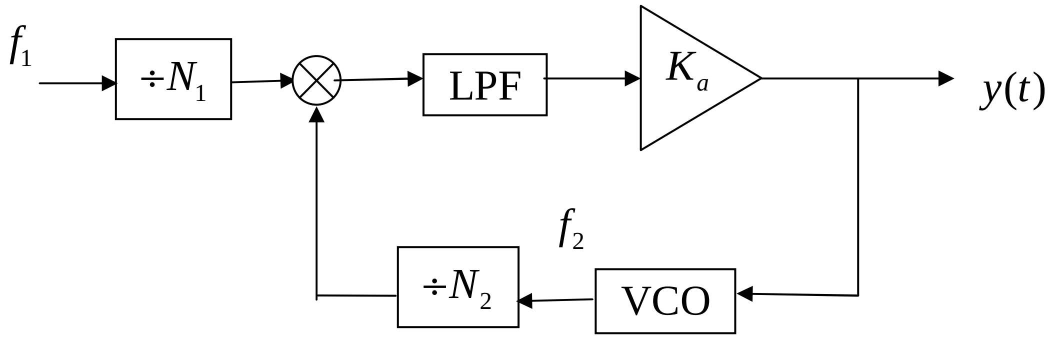
$$y(t) \approx \frac{\Delta f}{K_v} s(t)$$

So, if we can make sure that the error signal is small at all times, the PLL can be used with advantage as a frequency demodulator.

Another important application of the PLL is the implementation of frequency synthesizers.

Frequency Synthesis

When a PLL is locked, the frequencies of the signals arriving at the two inputs of the phase detector (multiplier) are equal.



The blocks labeled $\div N_k$ are frequency dividers (usually implemented by presettable logic counters). At the inputs of the phase detector (assuming that the PLL is locked), we can write:

$$\frac{f_1}{N_1} = \frac{f_2}{N_2}$$

So, the VCO will produce: $f_2 = \frac{N_2}{N_1} f_1$. This means that we can produce a signal with a frequency that can be set using digital hardware and that can be very stable if the reference oscillator producing f_1 is very stable. This technique of frequency generation is commonly used in modern receivers. The PLL is usually built in a microcontroller.