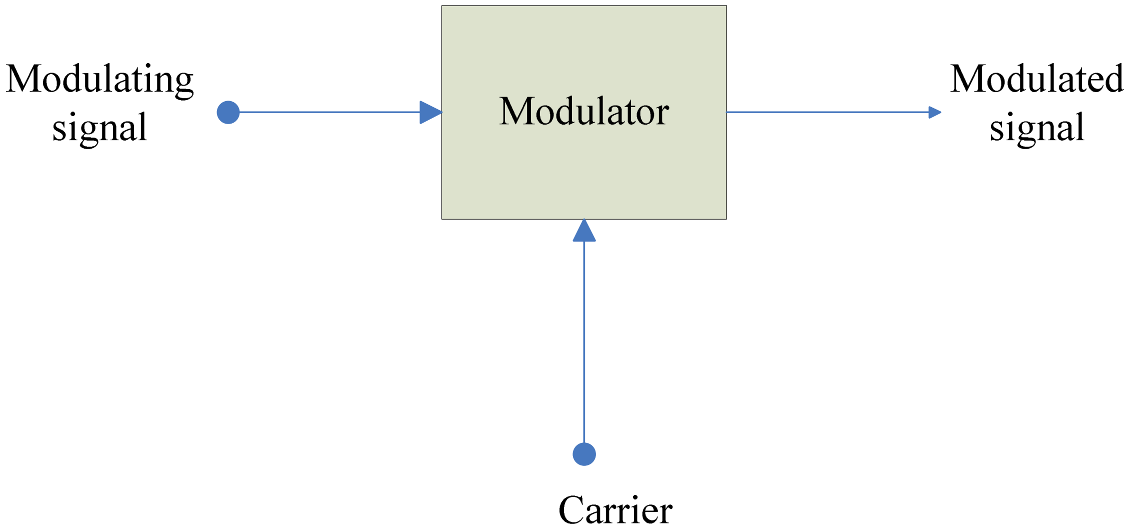
Modulation

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 $c = 3 \times 10^8$ 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 $|\tilde{s}(t)|_{\max}$ such that: $|\tilde{s}(t)| \le |\tilde{s}(t)|_{\max}$ for all t. We can thus define a normalized signal $s(t) = \dfrac{\tilde{s}(t)}{|\tilde{s}(t)|_{\max}}$ and we have $|s(t)| \le 1$.

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 $S(f) = \mathcal{F}[s(t)]$ then $S(f) = 0$ for $|f| > W$.

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:

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

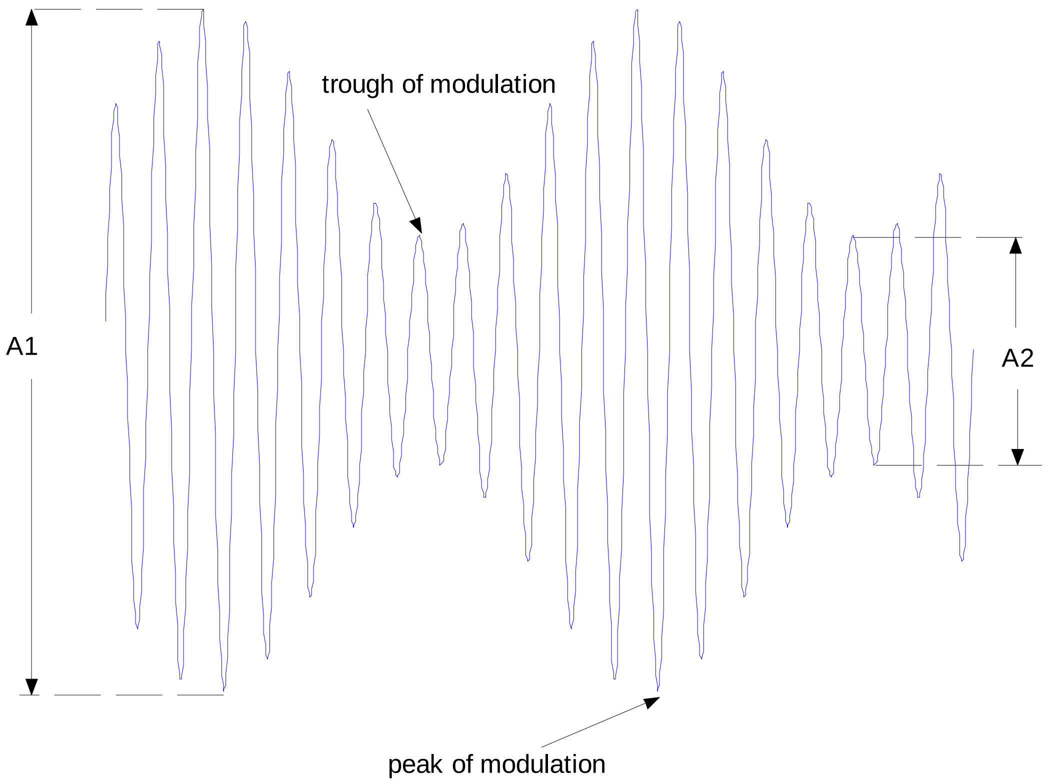
$$m = \frac{k_a \left|\tilde{s}(t)\right|_{\max}}{A_0}$$

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.

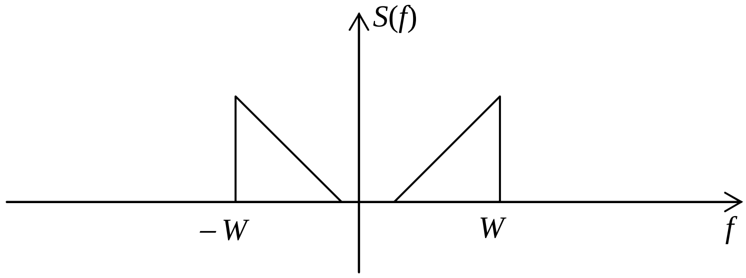
Starting from $x(t)$, we obtain

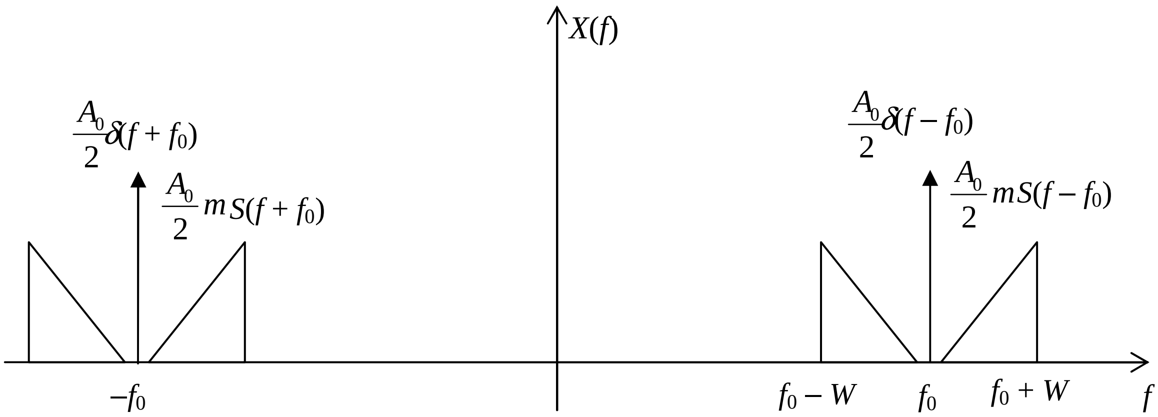
$$x(t) = A_0 \left(1 + ms(t)\right)\cos\omega_0 t = A_0 \cos\omega_0 t + A_0 ms(t)\cos\omega_0 t$$

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 ($P_s = \frac{1}{T_m}\int_{T_m} s^2(t)dt \le \frac{1}{T_m}\int_{T_m} \left|s(t)\right|_{\max}^2 dt = 1$). 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:

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

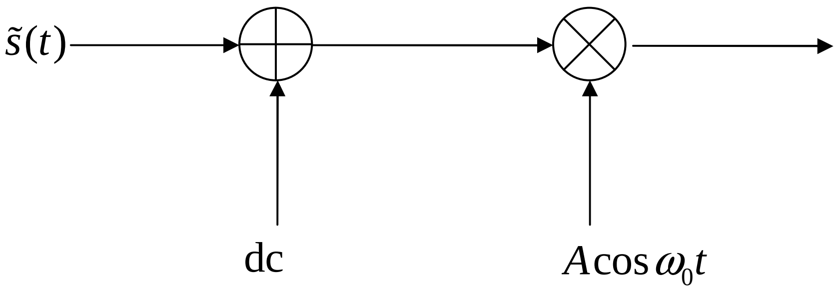
η cannot exceed the value of 1/2.

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.

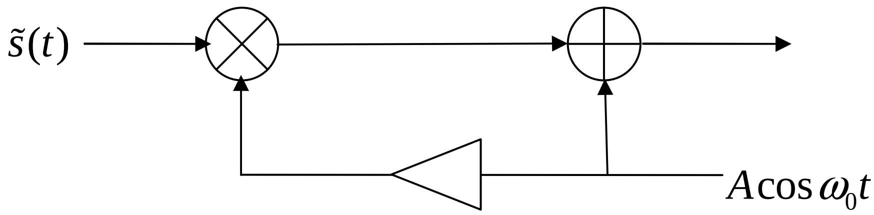
AM production:

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

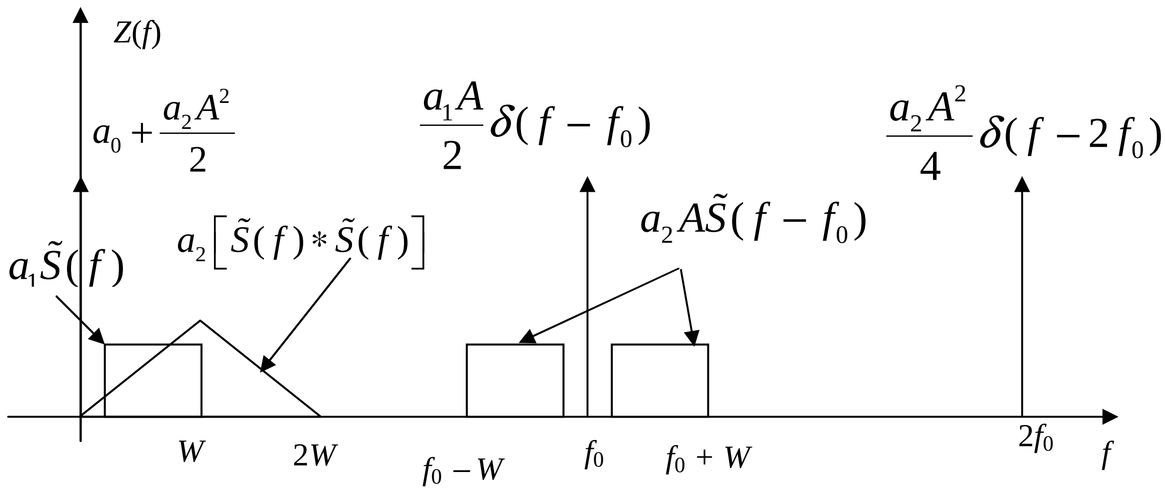
Component at $2\omega_0$: $\frac{a_2 A^2}{2}\cos 2\omega_0 t$

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>

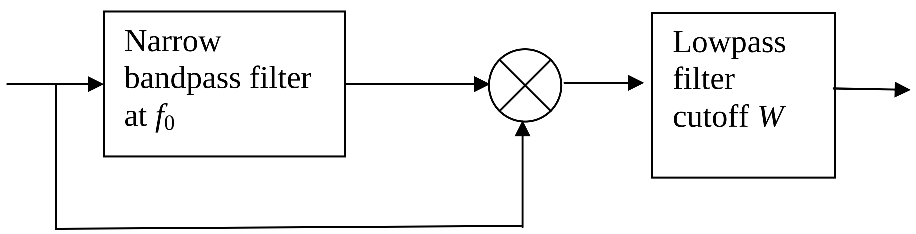
There exist several techniques for AM modulation. They fall into two different classes: Homodyne (Synchronous, Coherent) demodulation and Non Coherent demodulation.

<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: $y(t) = B\cos\left[\left(\omega_0 + \Delta\omega\right)t + \varphi_0\right]$. $\Delta\omega$ is a frequency error and φ_0 is a phase error. The result of the product is:

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

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.

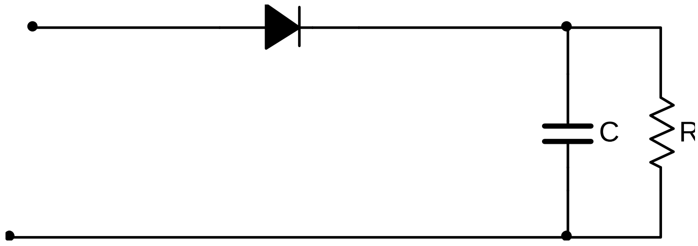


The above figure shows a typical synchronous demodulator.

<u>Non coherent demodulator:</u>

[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:

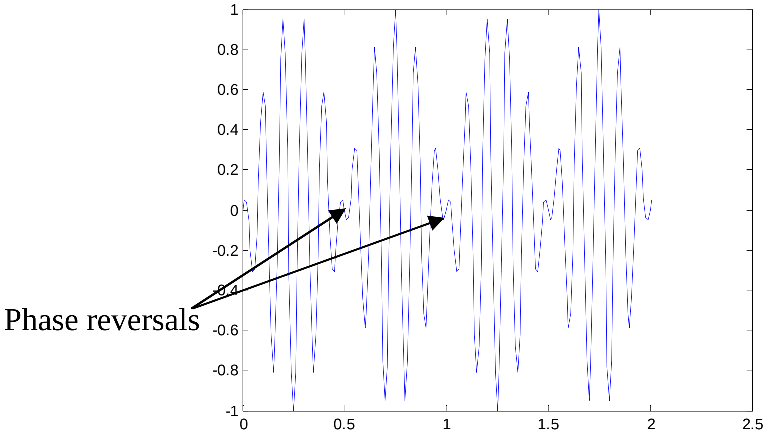
$$\frac{P_{sb}}{P_{peak}} = \begin{cases} \dfrac{P_s}{4} & \text{DSB-SC} \\ \dfrac{m^2 P_s}{4(1+m)^2} & \text{AM} \end{cases}$$

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:

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

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.

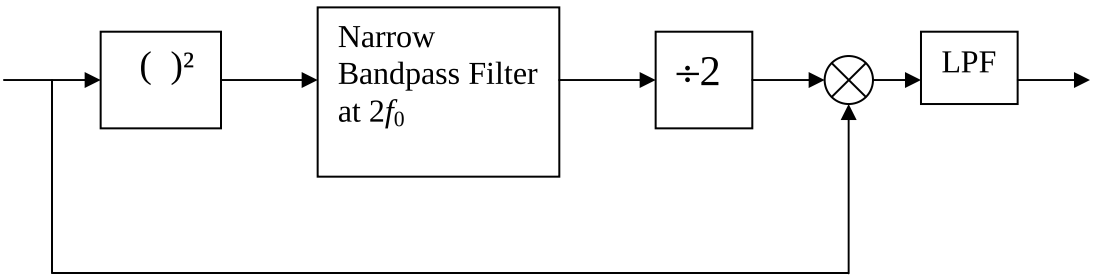
The most commonly used demodulator for DSB is the homodyne one. However, since there is no carrier transmitted along with the signal, the local carrier generation is more complex.

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:

$x^2(t) = A_0^2 s^2(t) \cos^2 \omega_0 t = \frac{A_0^2 s^2(t)}{2} \left(1 + \cos 2\omega_0 t\right)$. 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).



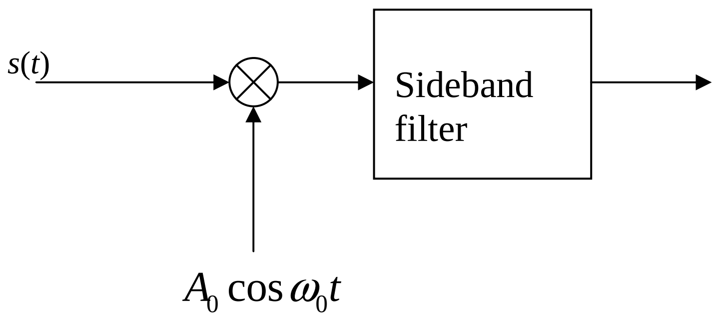
Squaring loop demodulator

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.

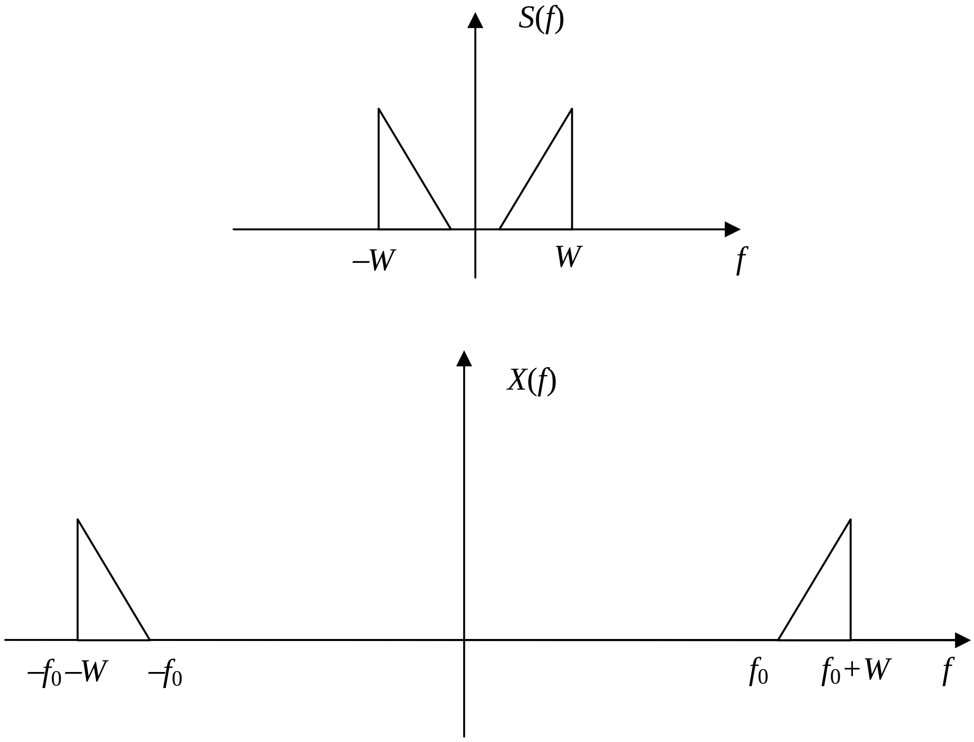
Single Sideband Modulation (SSB):

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

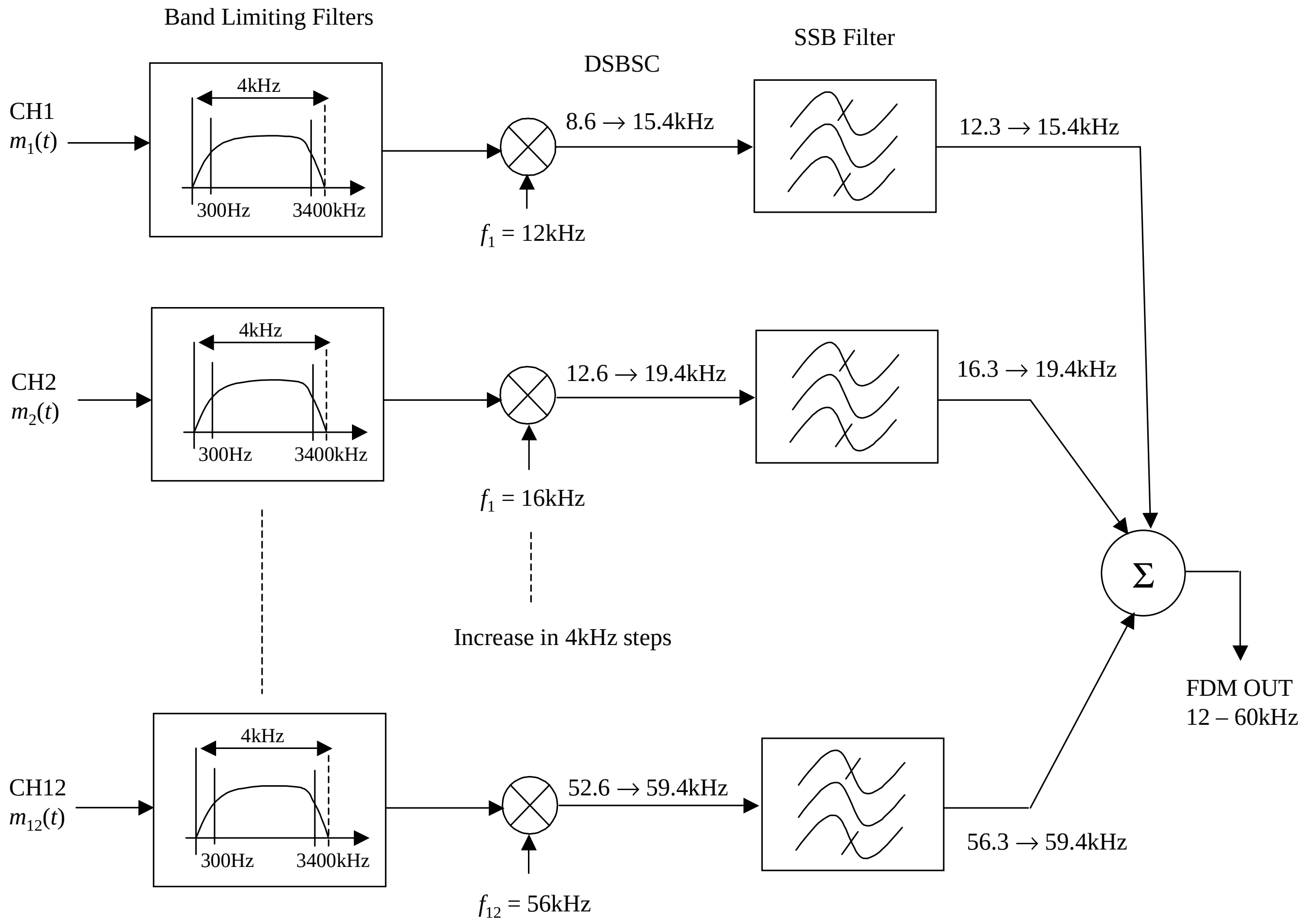
It is quite simple to represent the different operations in the frequency domain. The following sketch shows a USB-SSB signal in the frequency domain.



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

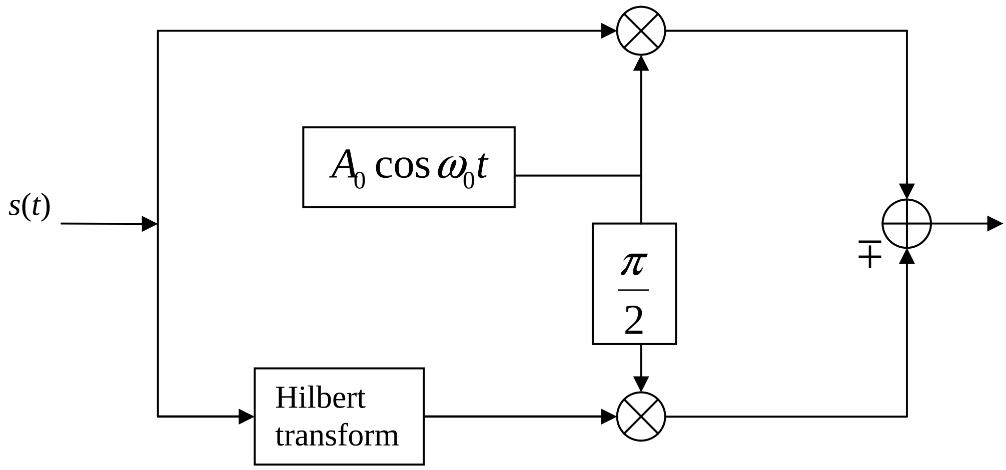
To see how SSB is demodulated, we have to express the SSB signal in the time domain. In order to do so, it is simpler to obtain first the corresponding analytic signal $X_+(f)$ and then use the fact that $x(t) = \mathrm{Re}\left[x_+(t)\right]$. When we observe the USB-SSB spectrum, we notice that the 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 $y(t) = B\cos\left[\left(\omega_0 + \Delta\omega\right)t + \varphi_0\right]$. 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:

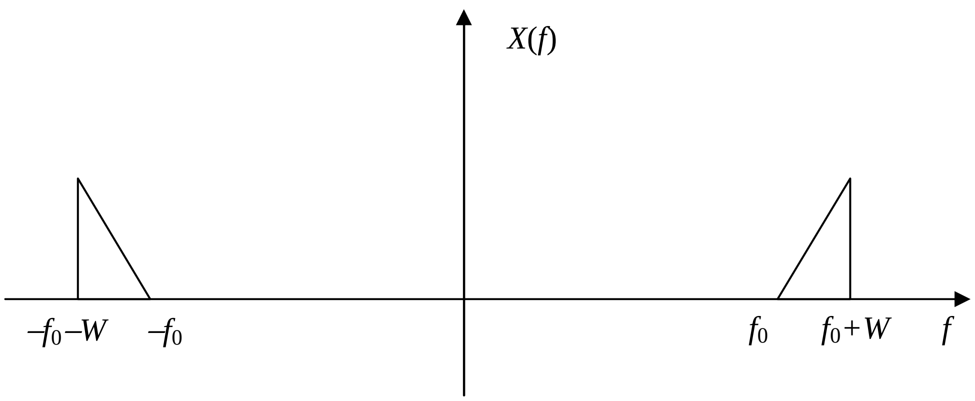
$$\frac{A_0 B}{2}\left[s(t)\cos\varphi_0 \mp \hat{s}(t)\sin\varphi_0\right]$$

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.

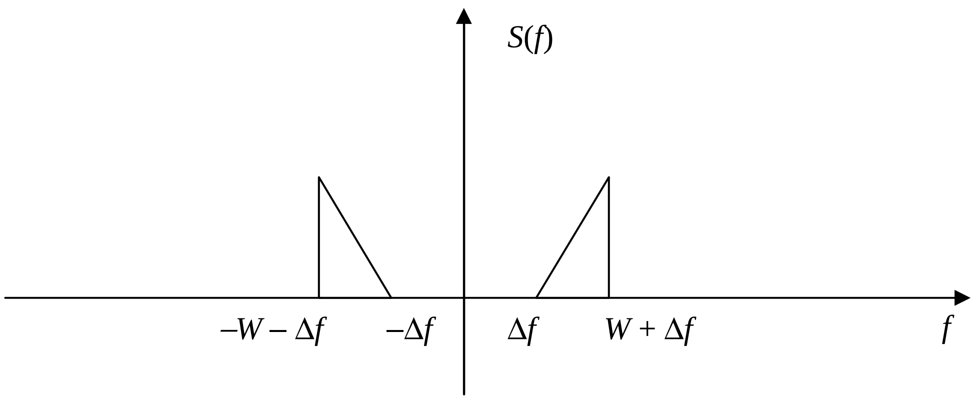
The analysis of the frequency error is easier to study in the frequency domain. Using the modulation theorem, the result of the SSB signal multiplied by a carrier is:

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

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



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

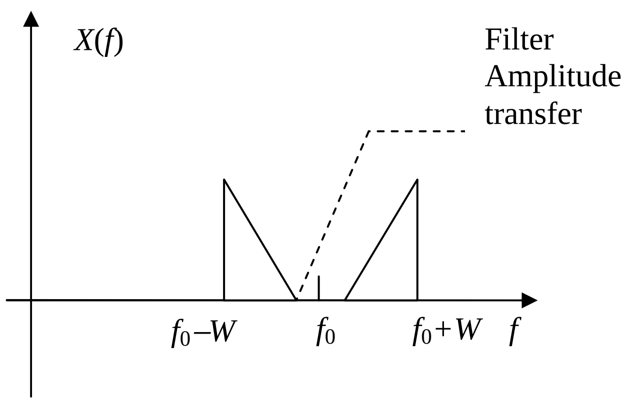


We remark that all the frequencies of the message are translated by a constant shift. This constant shift does not make the speech unintelligible. However, when Δf is positive, it makes everybody sound like "Donald Duck", hence the name; *Donald Duck* distortion. On the other hand, music will be completely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

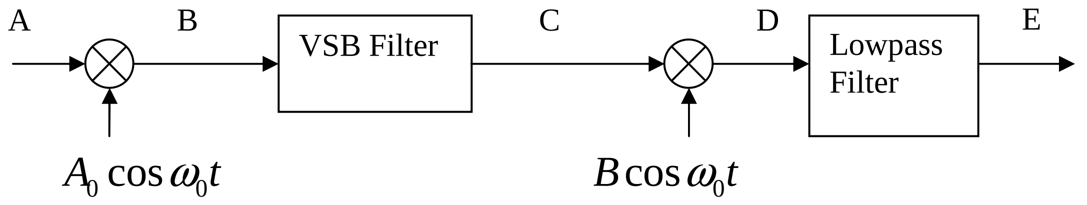


If we want to apply the phasing method, we will encounter the same problem. It is impossible to build a filter that phase shifts all frequencies from zero to W. This means that the phase response of the filter has also a zero transition band. We can approximate the Hilbert transformer if the signal $s(t)$ 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:

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

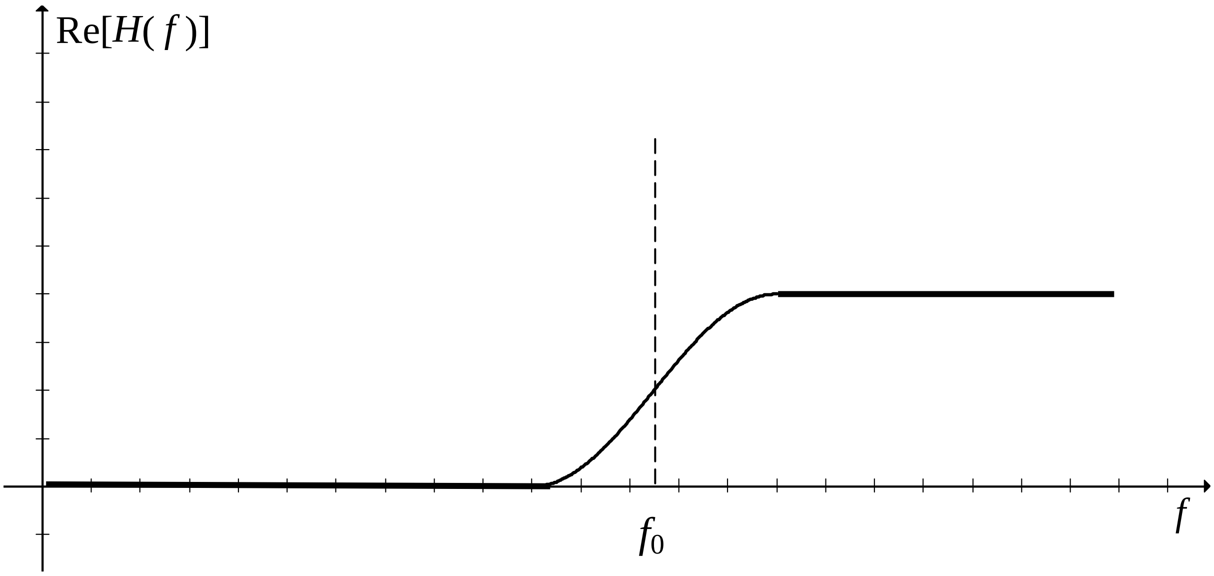
After some manipulations, we obtain that the transfer function of the filter must satisfy:

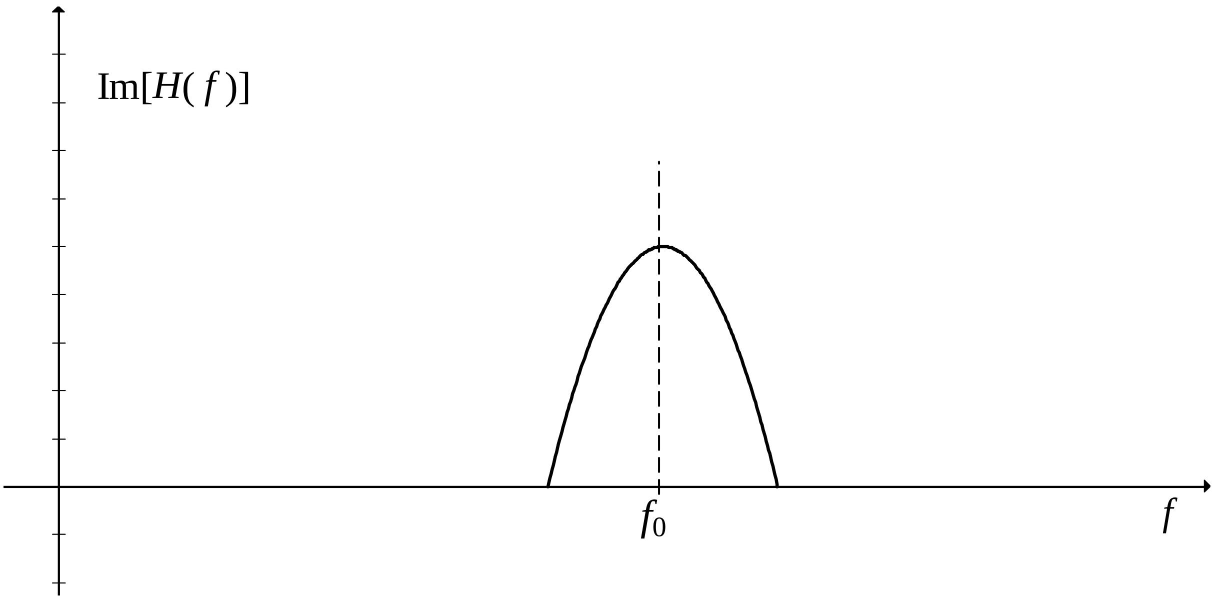
$$H(f_0 + x) + H^*(f_0 - x) = 2\,\mathrm{Re}\left[H(f_0)\right] \text{ for } f \text{ around } f_0.$$

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 $Q(f)=-j\,\mathrm{sgn}(f)$ 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.